International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:5, 2009

A Novel Prediction Method for Tag SNP Selection using Genetic Algorithm based on KNN

Li-Yeh Chuang, Yu-Jen Hou, Jr., and Cheng-Hong Yang

***Abstract*—**Single nucleotide polymorphisms (SNPs) hold much promise as a basis for disease-gene association. However, research is limited by the cost of genotyping the tremendous number of SNPs. Therefore, it is important to identify a small subset of informative SNPs, the so-called tag SNPs. This subset consists of selected SNPs of the genotypes, and accurately represents the rest of the SNPs. Furthermore, an effective evaluation method is needed to evaluate prediction accuracy of a set of tag SNPs. In this paper, a genetic algorithm (GA) is applied to tag SNP problems, and the K-nearest neighbor (K-NN) serves as a prediction method of tag SNP selection. The experimental data used was taken from the HapMap project; it consists of genotype data rather than haplotype data. The proposed method consistently identified tag SNPs with considerably better prediction accuracy than methods from the literature. At the same time, the number of tag SNPs identified was smaller than the number of tag SNPs in the other methods. The run time of the proposed method was much shorter than the run time of the SVM/STSA method when the same accuracy was reached.

***Keywords*—**Genetic Algorithm (GA), Genotype, Single nucleotide polymorphism (SNP), tag SNPs.

I. Introduction

SINGLE nucleotide polymorphisms (SNPs) are the most common variants amongst species. The number of identified SNPs is very high and is currently estimated to be about 10 million [1].With the genome-wide SNP discovery, many genome-wide association (GWA) studies are likely to identify multiple genetic variants that are associated with complicated diseases [2], [3]. However, genotyping all existing SNPs for a large number of samples remains a challenge. Therefore, it is essential to select informative SNPs representing the original SNP distributions in the genome (tag SNP selection) for genome-wide association studies. These SNPs are usually chosen from haplotype data and are thus called haplotype tag SNPs (htSNPs). Accordingly, the scale and cost of genotyping can be significantly decreased. Recently, some hybrid algorithms, such as HAPLO-IHP [4] and ISHAPE [5], have been developed which are capable of improving the performance of haplotyping.

Many algorithms have been developed to select the most informative tag SNPs. Tag SNP selection can follow two different strategies: the block-based and the block-free methods. Numerous block-free methods are also available [13]–[18]. A block-based method is based on the haplotype block structure of the human genome. The rationale is that the human genome can be partitioned into discrete blocks [6] and that most of the population share a very small subset of common haplotypes within each block. Haplotype diversity is limited and conserved in the haplotype block of the whole genome [7], [8]. Many algorithms first partition genomes into haplotype blocks [8]–[11] and then select the tag SNP subset within each block. This method focuses on finding a set of tag SNPs to distinguish all the common haplotypes [6], [12]. The main problem with the block-based method is that the definition of the blocks is not always straightforward and there is no consensus how the blocks are formed. Moreover, tag SNP selection based only on the local correlations between markers of each block ignores inter-block correlations [13].

In a block-free method, the tag SNPs is regarded as a subset of all SNPs, from which the remaining SNPs can be reconstructed with minimal error [14], [15]. Black-free methods do not assume prior block partitioning or limit the diversity of haplotypes. Block-free tagging SNP methods are based on weak correlations that occur across nearby blocks [15]. They make use of the proximity of potentially predictive SNPs and are less limiting than methods involving rigid notation of haplotype blocks. A natural measure for evaluating the prediction accuracy of a set of tag SNPs was developed for these methods [16]. Researchers developed a novel algorithm called STAMPA (selection of tag SNPs to maximize prediction accuracy) to find a minimum set of tag SNPs and minimize their prediction error. Dynamic programming was applied in STAMPA to select tag SNPs and maximize prediction accuracy. STAMPA was found to provide higher prediction accuracy than ldSelect [19] and HapBlock [20] tested on a variety of data sets [16]. He and Zelikovsky have introduced two novel approaches for informative SNP prediction based on multiple linear regression (MLR-tagging) [21] and support vector machines (SVM/STSA) [22]. These prediction algorithms combined were with a

Li-Yeh Chuang is with the Department of Chemical Engineering, I-Shou University, Kaohsiung, Taiwan (email: chuang@isu.edu.tw)
Yu-Jen Hou is with the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Taiwan (e-mail: 1096305143@cc.kuas.edu.tw).
Cheng-Hong Yang is with the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Taiwan (e-mail: chyang@cc.kuas.edu.tw)

stepwise tag selection algorithm (STSA) to select a tag SNP set of minimal size. In a direct comparison of MLR-tagging and SVM/STSA, SVM/STSA was proved more effective than MLR-tagging, but also more time-consuming.

Yet another method of tag SNP selection is through the calculation of correlation between each pair of SNPs (such as linkage disequilibrium, LD). Linkage disequilibrium describes the correlation between genotypes at a pair of polymorphic sites and is usually higher when pairwise SNPs are closer. Two statistical values are used to describe LD, named D' and r^2 [26]. r^2 is most frequently used for pairwise SNP correlation, because it is directly related to statistical power to detect disease associations [19]. However, some studies try to identify a minimum set of LD bin set in existing SNPs with high-LD (r^2 ≥ 0.8) [19]. To do this, SNPs will be partitioned into different regions according to the relevance of SNPs [19], [23]–[25]. SNPs within a bin are denoted tag SNP, and only one tag would be genotyped per bin. However, the disadvantage of this method is that it can not exclude the possibility that SNPs with a low-LD also enhance prediction accuracy.

In this paper, a genetic algorithm (GA) is applied to the tag SNPs problem and the K-nearest neighbor (K-NN) methods serves as an evaluator of the GA; it is used to evaluate the prediction accuracy of a set of tag SNPs. GAs are a randomized search and optimization techniques that derive their working principles from natural genetics; they have been successfully applied to the optimization of a variety of problems. The results of our study were compared to state-of-the-art studies and indicate that the proposed method can effectively select a minimum number of tag SNPs with higher prediction accuracy.

II. PROBLEM FORMULATION

In a haplotype sequence, SNPs are generally bi-allelic, meaning that there are only two alleles in a single SNP: a major type and a minor. In bi-allelic SNPs, each haplotype can be represented by a binary string set. The allele information value is formed by a sequence of base pairs {A, T, C, G}. Each haplotype can be formalized by binary strings 0 and 1 where 0 represents the major allele and 1 represents the minor allele. Thus, we can represent a haplotype h with m SNPs as $h = \{h_1, h_2, \ldots, h_m\}$, $h_i \in \{0, 1\}$.

$$h_i = \begin{cases} 0 : allele\ of\ ith\ SNP\ is\ major \\ 1 : allele\ of\ ith\ SNP\ is\ minor \end{cases} \quad (1)$$

In a genotype sequence, the allele information value is formed by {A/A, A/T, A/C, A/G…G/C, G/T}. In order to present our method, If a genotype g has m SNPs, it can be represented by $g = \{g_1, g_2, \ldots, g_m\}$, $gi \in \{0, 1, 2\}$. We used 0 and 1 to represent the homozygous types ({0,0} or {1,1}), and 2 to represent the heterozygous types ({0,1} or {1,0}).

$$g_i = \begin{cases} 0 : two\ alleles\ of\ ith\ SNP\ are\ major\ homozygous \\ 1 : two\ alleles\ of\ ith\ SNP\ are\ minor\ homozygous \\ 2 : two\ alleles\ of\ ith\ SNP\ are\ heterozygous \end{cases} \quad (2)$$

A sample S of a population P of genotype (or haplotype) individuals on m SNPs was given. Our goal then was to find a minimum set of tag SNPs $T = \{t_1, t_2, \ldots, t_k\}$, where k represents the number of tag SNPs ($k < m$), which consists of selected SNPs of the genotypes, and can predict the remaining unselected SNPs with minimum error. In order to achieve this goal, we need to find the minimum number of tag SNPs. The two major processes involved are the tag selection algorithm and the SNP prediction algorithm.

III. METHODS FOR TAG SNP SELECTION

The purpose of tag SNP selection is to find a small subset of informative SNPs (tag SNP), which accurately represents the rest of the genome sequence. In this paper, a GA was applied to the tag SNP selection problem, and the K-nearest neighbor (K-NN) method served as an evaluator of the GA.

A. Genetic Algorithm (GA)

Genetic Algorithms (GAs) were developed by Alan Turing in 1950, and further required by John Holland in 1970 [26]. The main components of the GA used in our study are the encoding schemes, population initialization, fitness evaluation, selection, crossover operator, mutation operator, and the amendment chromosome. The flowchart of the proposed method is shown Figure 1. The components are explained in detail below.

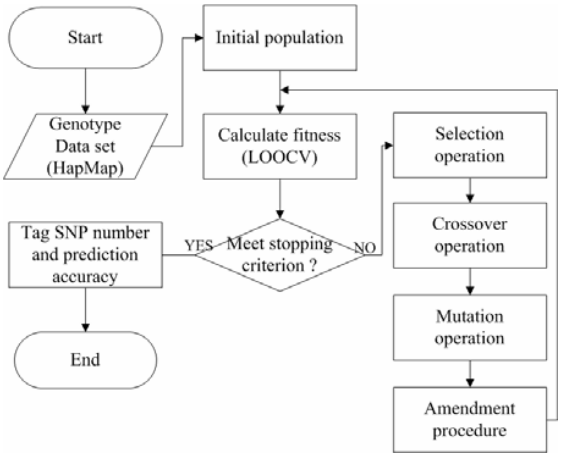

Fig. 1 Flowchart of the proposed method

B. Encoding schemes

Fundamental to the GA's structure is the encoding scheme. In this paper, the binary encoding method used in a chromosome corresponds to the tag SNP selection problem, as shown in Figure 2. Given are p chromosomes of a population, with each chromosome containing m SNPs (dimension). Each chromosome of the length m is a sequence over $\{0, 1\}^m$ (0 represents a non-selected SNP and 1 represents a selected SNP). The binary encoding method used can be described by:

$C_i = \{c_{i1}, c_{i2}, \ldots, c_{im}\}$ and $c_{ij} = \{0, 1\}$, $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, m$, where p represents the size of population. $c_{ij} = 1$ means that the jth SNP on the ith chromosome was selected. For example, assume there is a chromosome represented by $C_i = \{1, 0, 1, 0, 0, 1, 0\}$. In this encoding scheme SNP_1, SNP_3 and SNP_6 are predicted to be tag SNPs.

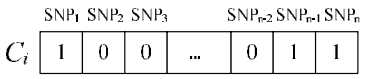

Fig. 2 Chromosome design chart

C. Population initialization

In general, the chromosome initialization model was produced by the stochastic approach, but the SNP quantity of each data set collection was different. Efficiency decreased if the necessary tag SNP quantity was not considered, and the probability of a SNP being selected as a tag SNP by the same value was not set. Thus, the tag SNP quantity has to be chosen appropriately in the initial population.

D. Fitness function

The fitness function is one of the most important parameters in a GA. It is used to determine which chromosome is selected during the selection operation. We used a similar prediction accuracy computing mode as STAMPA, the leave-one-out cross-validation (LOOCV) method. Given was an unknown genotype g_i and a predicted sample *S'*. We used the K-nearest neighbour (KNN) method to identify three genotypes as the nearest neighbor of g_i, and obtained a predicted sample S' by voting on these three neighbours.

KNN was proposed by Fix and Hodges in 1951 [27]. Given a test document *d* (whose class is unknown), the system finds the k nearest neighbours among the training corpus, and uses the classes of the *k* nearest neighbors to weight candidates. In this study, the distance was defined as the Hamming distance between two SNP loci. The *k*-closest neighbours (*k* representing the number of neighbors) between genotypes had to be determined. In the GA method, we used 3-NN (k = 3) to determine 3 neighbors of a genotype sample for the voting process.

E. Selection, crossover and mutation

In this work we used the well-known method of roulette wheel selection [28], which is one of the most common and easy to implement selection mechanisms. Basically, roulette wheel selection works as follows: each chromosome in the population is associated with a sector on a virtual wheel. A sector will cover a larger area on this wheel when the corresponding chromosome has a higher fitness value, while a lower fitness value is represented by a smaller sector.

After the selection process crossover is implemented, a crucial operation of the genetic algorithm. In this paper, we utilize an intelligent crossover operation in order to avoid generating forbidden offsprings. At first, a group of crossover mask was created. A crossover mask is simply a string of binary bits randomly produced with a chance of 0.5. Figure 3 shows here to positions on the mask change the corresponding positions on the parent chromosomes. We randomly generated a crossover mask {0, 1, 0, 0, 1, 0}, meaning that the two genes matched on the parent chromosomes are exchanged. This produces the two offspring chromosomes {0, 0, 0, 1, 1, 0} and {1, 1, 0, 1, 0, 1} after crossover.

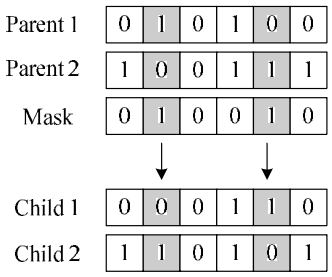

Fig. 3 Crossover procedure

Mutation procedures can be used in GA to prevent evolution from slowing down too much, and to generate a large variety of chromosomes, which avoids local optima solutions. With a mutation rate M_{rate} given, a random number between 0 and 1 is generated for each gene (i.e. each dimension). If the random number is smaller than M_{rate}, then the corresponding gene will be mutated (i.e. 0→1 or 1→0). We used a mutation rate M_{rate} of 0.01 in this study. If this operator generates a chromosome which does not satisfy the capacity constraint, it will be ignored and the original chromosome is retained for further calculations.

F. Amendment procedure

After implementation of the crossover and mutation procedures, the number of selected tag SNPs has to be adapted in an amendment process. If the number of selected tag SNPs is greater than the requested number, some randomly selected tag SNPs from the group of selected SNPs will be un-selected. On the other hand, if the number of selected tag SNPs is smaller than the requested number, some originally not selected SNPs will randomly be selected.

IV. EXPERIMENTAL DATA SETS

Four published experimental SNP data sets were downloaded from HAPMAP (http://www.hapmap.org/) for evaluation.

A. ENCODE Regions from HapMap

Regions ENm013, ENr112 and ENr113 from 30 CEPH family trios obtained from HapMap (2008). These data were collected from chromosome 7q21.13, 2p16.3 and 4q26. The number of SNPs genotyped in each region was 376, 439 and 523 (note that the SNP numbers of our data were greater than in He *et al.*). We also used 90 genotypes corresponding to the parents from each data set.

B. Chromosome 5q31

The data set from the study of Daly *et al.* (2001) [7] was derived from the 616 kilobase region of human chromosome 5q31 from 129 family trios.

C. Other Gene Region from HapMap

We used two sets of SNPs spanning the two genes STEAP and TRPM8 collected from 30 CEPH family trios. The number of SNPs in each region was 37 and 107.

TABLE I
THE NUMBER OF TAG SNPS NEEDED TO REACH PREDICTION ACCURACIES OF BETWEEN 80% AND 99% FOR MLR-TAGGING, SVM/STSA, AND GA METHODS.

TABLE I.A. MLR-TAGGING METHOD

Group	Data set (num of SNPs)	Prediction accuracy (%)											
		80	85	90	91	92	93	94	95	96	97	98	99
1	STEAP (22)	1	1	1	2	2	2	2	2	3	3	4	4
	TRPM8 (101)	1	2	4	5	5	6	7	8	10	15	15	24
	5q31 (103)	1	2	5	7	7	9	13	16	21	31	41	55
2	ENm013 (360)	2	3	6	6	7	8	9	9	11	15	22	254
	ENr112 (411)	6	9	14	16	18	20	24	33	63	95	126	187
	ENr113 (514)	4	5	10	11	13	15	18	40	55	80	104	200

TABLE I.B. SVM/STSA METHOD

Group	Data set (num of SNPs)	Prediction accuracy (%)											
		80	85	90	91	92	93	94	95	96	97	98	99
1	STEAP (22)	1	1	1	1	1	1	1	2	2	2	2	2
	TRPM8 (101)	1	1	2	5	5	6	7	8	10	15	15	24
	5q31 (103)	1	1	3	3	4	5	6	8	10	22	42	51

SVM/STSA method didn't available on group2 data sets [22].

TABLE I.C. GA METHOD

Group	Data set (num of SNPs)	Prediction accuracy (%)											
		80	85	90	91	92	93	94	95	96	97	98	99
1	STEAP (22)	1	1	1	1	1	1	1	1	1	1	1	2
	TRPM8 (101)	1	1	1	1	1	1	1	1	2	2	4	10
	5q31 (103)	1	1	1	1	1	1	1	1	2	2	5	12
2	ENm013 (360)	1	1	1	1	1	1	2	4	5	7	16	116
	ENr112 (411)	1	1	1	1	1	1	1	3	4	4	8	26
	ENr113 (514)	1	1	1	1	1	1	1	2	8	14	51	124

V. RESULT AND DISCUSSION

The termination condition of the GA in this study was reached at a pre-specified number of iterations (in our case, the number of iterations was 50). Parameters of the genetic algorithms used here were: population size of 50, number of iteration set to 50, crossover rate of 0.9, and mutation rate of 0.01. The chromosome length is the SNP number of the data set.

In this study, we introduce GA based on KNN for the tag SNP selection problem. Table I show the number of tag SNPs needed to reach prediction accuracies of 80% to 99% for MLR-tagging, SVM/STSA, and GA methods. The test data sets used were the same as the ones in He *et al.* [22]. These 6 data sets can be divided into two groups. The group1 contains three smaller data sets (STEAP, TRPM8 and 5q31). Three methods were applied to test these data sets. Under the same prediction accuracy, the number of tag SNPs selected by the proposed method is much smaller than the number of tag SNPs selected by the SVM based method, which in turn generally has a smaller number of tag SNPs selected by the SVM based method, which in turn generally has a smaller number of tag SNPs selected than MLR-tagging. The prediction accuracy of the proposed method when only one SNP was selected reached 97% for data set STEAP and 95% for the TRPM and 5q31data sets, respectively.

Obviously, the prediction accuracy of the proposed method is superior to the one of the SVM/STSA method under these circumstances. In our experiment, the prediction accuracy of the proposed method was better than the one of SVM/STSA for an equal number of tag SNPs. Group2 consists three large data sets ENm013, ENr112 and ENr113. Only the GA and MLR-tagging method were applied to these test data sets since no data was provided for the SVM/STSA method in the original literature [22]. The number of tag SNP selected by the proposed method is much smaller than the number of selected SNPs for MLR-tagging. The prediction accuracy of the proposed method when only one SNP is selected was 93% for the ENm013 data set, and 94% for the ENr112 and ENr113 data sets, respectively.

The prediction accuracy of the GA can easily reach 90% when just a single tag SNP is used. In fact, there are two reasons for this phenomenon. First, like He *et al.* said that too more major allele in data. Therefore, in the design of LOOCV procedure, we implement KNN to select k tags and voting. The predicted allele of untagged SNP was major allele naturally. In other words, if one predicted each SNP as 0 (homozygous with major allele) then the prediction accuracy were higher. For example of STEAP, it directly predicts each SNP as major allele. The accuracy of this prediction is 89.28%, so that it always can reach 90% by selecting any tag SNP. Second, the data type of our study was genotype data. Usually, a haplotype is represented by two strings 0 and 1. 0 represents the homozygous with major allele, and 1 represents the homozygous with minor allele. Respectively, each genotype is represented by 0, 1 and 2. 0 represents the homozygous with

TABLE II. TAG SNP NUMBERS BY GA, MLR-TAGGING, STAMPA AND RLRP TO ACHIEVE THE ACCURACY OF 80% AND 90% IN LOOCV TEST.

Accuracy	Algorithm	ENm013 (376)	ENr112 (439)	ENr113 (523)	STEAP (37)	TRPM8 (101)	5q31 (103)
80%	GA	**1**	**1**	**1**	**1**	**1**	**1**
	MLR-tagging	2	6	4	**1**	**1**	**1**
	STAMPA	5	9	11	2	3	2
	RLRP	11	17	35	4	9	10
90%	GA	**1**	**1**	**1**	**1**	**1**	**1**
	MLR-tagging	6	14	10	**1**	4	5
	STAMPA	12	17	18	2	6	6
	RLRP	48	52	58	8	22	35

Because of the data set were updated, so these datasets we used were greater than He *et al.*.

TABLE III. COMPARISON OF ACCURACY AND RUNTIME BY GA□SVM/STSA AND MLR METHODS.

Datasets (num of SNPs)	methods	Number of tag SNPs					
		1	2	4	6	8	10
5q31 (103)	GA	95.88	97.55	98.44	98.81	98.91	98.98
	SVM/STSA	86.81	89.32	92.24	94.09	95.28	96.09
	MLR-tagging	81.15	83.84	88.15	90.91	92.66	93.49
TRPM8 (107)	GA	95.90	97.08	98.09	98.31	98.88	99.15
	SVM/STSA	88.89	90.50	90.67	93.67	95.56	96.74
	MLR-tagging	80.68	85.32	90.75	93.74	95.16	96.38
STEAP (37)	GA	98.31	99.11	99.70	99.91	99.97	100
	SVM/STSA	94.02	98.18	99.68	99.73	99.79	99.80
	MLR-tagging	90.79	96.16	99.13	99.71	99.78	99.78

All experiments were performed on a computer with Intel(R) Core (TM) 2, 1.86GHz processor, and 1.5GB RAM.

major allele, 1 represents the homozygous with minor allele, and 2 represents the heterozygous site. Heterozygous site can represent both 0 and 1.

For the same part of tag SNP number, we followed the comparison model of He *et al.* [22]. We compared GA with MLR-tagging [22], STAMPA [16], and RLRP [30] (see Table II). It schemed two thresholds of 80% and 90%, and then compared the tag SNP number of differ methods. GA and MLR-tagging were selected one tag to achieve 90% in some datasets. It shows these data were suited to the prediction of major allele (GA and MLR-tagging). With regard to STAMPA, it predicted by inspecting the two closest tag SNPs from both left and right SNPs. If there has no neighbor at both sides, it would not able to predict the SNPs. That's why STAMPA requires at least two tags for prediction and the effect was not very well.

The advantage of SVM/STSA was the smallest number (i.e., tag SNP number = 1) to achieve the higher prediction accuracy [22]. We consider that runtime of SVM/STSA was too much. As the study of He *et al.* [22], this method for 5q31 dataset needs 3 hour to select one tag SNP. On the contrary, MLR-tagging only needs 0.77 second and GA needs 5 second. The time spent would increase by adding selected tag SNP number. In another way, time spent would also increase with the data updating phenomenon. Although the prediction quality of SVM/STSA was better than MLR-tagging, but this method was time-consuming to process. As shown in Table III, we used the same condition with He *et al.* and comparing the result by the smallest data STEAP (include 22 SNPs). Proposed GA selected 2 tag SNPs to achieve the accuracy of 99%, and 100% with 8 tags. The SVM/STSA method used two tags to predict at 98%, but it couldn't achieve 100% at the same tags with GA by 10 tag SNPs. It would presume that both of MLR-tagging or SVM/STSA would weakness in the increasing SNP number. They were only congruence with certain dataset which had less SNP.

VI. CONCLUSION

In this paper, we present a novel approach to tag SNPs prediction based on genetic algorithm (GA) and a K-nearest neighbor (K-NN) serves as an evaluator of the GA. The experimental data we used is genotype information taken from HapMap project. We compared the proposed method with state-of-the-art tag SNP selection algorithms from these literatures. The prediction accuracy of proposed method consistently identified tag SNPs with considerably better than the method support vector machines method (SVM), multiple linear regressions (MLR), and STAMPA from the test problems. Furthermore, the number of tag SNPs selecting by the proposed method was smaller than the number of tag SNPs in any other methods (include SVM/STSA, MLR-tagging, and STAMPA methods). The run time of our proposed method was much shorter than the run time of the SVM/STSA method when the same accuracy was reached.

ACKNOWLEDGMENT

This work is partly supported by the National Science Council in Taiwan under grant NSC96-2622-E-151-019-CC3, NSC96-2221-E-214-050-MY3, NSC95-2221-E-214-087.

REFERENCES

[1] D. Brinza and A. Zelikovsky, "2SNP: scalable phasing based on 2-SNP haplotypes," *Bioinformatics,* vol. 22, pp. 371-3, Feb 1 2006.

[2] S. Buch, C. Schafmayer, H. Volzke, C. Becker, A. Franke, H. von Eller-Eberstein, C. Kluck, I. Bassmann, M. Brosch, F. Lammert, J. F. Miquel, F. Nervi, M. Wittig, D. Rosskopf, B. Timm, C. Holl, M. Seeger, A. ElSharawy, T. Lu, J. Egberts, F. Fandrich, U. R. Folsch, M. Krawczak, S. Schreiber, P. Nurnberg, J. Tepel, and J. Hampe, "A genome-wide association scan identifies the hepatic cholesterol transporter ABCG8 as a susceptibility factor for human gallstone disease," *Nat Genet,* vol. 39, pp. 995-9, Aug 2007.

[3] B. W. Zanke, C. M. Greenwood, J. Rangrej, R. Kustra, A. Tenesa, S. M. Farrington, J. Prendergast, S. Olschwang, T. Chiang, E. Crowdy, V. Ferretti, P. Laflamme, S. Sundararajan, S. Roumy, J. F. Olivier, F. Robidoux, R. Sladek, A. Montpetit, P. Campbell, S. Bezieau, A. M. O'Shea, G. Zogopoulos, M. Cotterchio, P. Newcomb, J. McLaughlin, B. Younghusband, R. Green, J. Green, M. E. Porteous, H. Campbell, H. Blanche, M. Sahbatou, E. Tubacher, C. Bonaiti-Pellie, B. Buecher, E. Riboli, S. Kury, S. J. Chanock, J. Potter, G. Thomas, S. Gallinger, T. J. Hudson, and M. G. Dunlop, "Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24," *Nat Genet,* vol. 39, pp. 989-94, Aug 2007.

[4] Y. J. Yoo, J. Tang, R. A. Kaslow, and K. Zhang, "Haplotype inference for present absent genotype data using previously identified haplotypes and haplotype patterns." vol. 23: Oxford Univ Press, 2007, p. 2399.

[5] O. Delaneau, C. Coulonges, P. Y. Boelle, G. Nelson, J. L. Spadoni, and J. F. Zagury, "ISHAPE: new rapid and accurate software for haplotyping," *BMC Bioinformatics,* vol. 8, p. 205, 2007.

[6] S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, and D. Altshuler, "The structure of haplotype blocks in the human genome," *Science,* vol. 296, pp. 2225-9, Jun 21 2002.

[7] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander, "High-resolution haplotype structure in the human genome," *Nat Genet,* vol. 29, pp. 229-32, Oct 2001.

[8] N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. Nguyen, M. C. Norris, J. B. Sheehan, N. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O. Trulson, K. R. Vyas, K. A. Frazer, S. P. Fodor, and D. R. Cox, "Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21," *Science,* vol. 294, pp. 1719-23, Nov 23 2001.

[9] X. Ke and L. R. Cardon, "Efficient selective screening of haplotype tag SNPs," *Bioinformatics,* vol. 19, pp. 287-8, Jan 22 2003.

[10] K. Zhang, M. Deng, T. Chen, M. S. Waterman, and F. Sun, "A dynamic programming algorithm for haplotype block partitioning," *Proc Natl Acad Sci U S A,* vol. 99, pp. 7335-9, May 28 2002.

[11] K. Zhang and L. Jin, "HaploBlockFinder: haplotype block analyses," *Bioinformatics,* vol. 19, pp. 1300-1, Jul 1 2003.

[12] G. C. Johnson, L. Esposito, B. J. Barratt, A. N. Smith, J. Heward, G. Di Genova, H. Ueda, H. J. Cordell, I. A. Eaves, F. Dudbridge, R. C. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S. C. Gough, D. G. Clayton, and J. A. Todd, "Haplotype tagging for the identification of common disease genes," *Nat Genet,* vol. 29, pp. 233-7, Oct 2001.

[13] T. U. M. Phuong, Z. Lin, and R. B. Altman, "CHOOSING SNPs USING FEATURE SELECTION." vol. 4, 2006, pp. 241-257.

[14] V. Bafna, B. V. Halldorsson, R. Schwartz, A. G. Clark, and S. Istrail, "Haplotypes and informative SNP selection algorithms: don't block out information," ACM New York, NY, USA, 2003, pp. 19-27.

[15] B. V. Halldorsson, V. Bafna, R. Lippert, R. Schwartz, F. M. De La Vega, A. G. Clark, and S. Istrail, "Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies," *Genome Res,* vol. 14, pp. 1633-40, Aug 2004.

[16] E. Halperin, G. Kimmel, and R. Shamir, "Tag SNP selection in genotype data for maximizing SNP prediction accuracy," *Bioinformatics,* vol. 21 Suppl 1, pp. i195-203, Jun 2005.

[17] P. H. Lee and H. Shatkay, "BNTagger: improved tagging SNP selection using Bayesian networks," *Bioinformatics,* vol. 22, pp. e211-9, Jul 15 2006.

[18] Z. Liu, S. Lin, and M. Tan, "Genome-wide tagging SNPs with entropy-based Monte Carlo method," *J Comput Biol,* vol. 13, pp. 1606-14, Nov 2006.

[19] C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson, "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium," *Am J Hum Genet,* vol. 74, pp. 106-20, Jan 2004.

[20] K. Zhang, Z. Qin, T. Chen, J. S. Liu, M. S. Waterman, and F. Sun, "HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms," *Bioinformatics,* vol. 21, pp. 131-4, Jan 1 2005.

[21] J. He and A. Zelikovsky, "MLR-tagging: informative SNP selection for unphased genotypes based on multiple linear regression," *Bioinformatics,* vol. 22, pp. 2558-61, Oct 15 2006.

[22] J. He and A. Zelikovsky, "Informative SNP selection methods based on SNP prediction," *IEEE Trans Nanobioscience,* vol. 6, pp. 60-7, Mar 2007.

[23] K. Zhang, T. Chen, M. S. Waterman, and F. Sun, "A Set of Dynamic Programming Algorithms for Haplotype Block Partitioning and Tag SNP Selection via Haplotype Data or Genotype Data," pp. 1–26.

[24] B. Devlin and N. Risch, "A comparison of linkage disequilibrium measures for fine-scale mapping," *Genomics,* vol. 29, pp. 311-22, Sep 20 1995.

[25] H. I. Avi-Itzhak, X. Su, and F. M. De La Vega, "Selection of minimum subsets of single nucleotide polymorphisms to capture haplotype block diversity," *Pac Symp Biocomput,* pp. 466-77, 2003.

[26] J. H. Holland, *Adaptation in natural and artificial systems*: MIT Press Cambridge, MA, USA, 1992.

[27] E. Fix and J. Hodges, "Discriminatory Analysis-Nonparametric Discrimination: Consistency Properties," Storming Media, 1951.

[28] D. E. Goldberg and K. Deb, "A comparative analysis of selection schemes used in genetic algorithms." vol. 1, 1991, pp. 69-93.

[29] G. A. Thorisson, A. V. Smith, L. Krishnan, and L. D. Stein, "The International HapMap Project Web site," *Genome Res,* vol. 15, pp. 1592-3, Nov 2005.

[30] J. He, K. Westbrooks, and A. Zelikovsky, "Linear reduction method for predictive and informative tag SNP selection," *Int J Bioinform Res Appl,* vol. 1, pp. 249-60, 2005.