TagSNP Selection Based on Pairwise LD Criteria and Power Analysis in Association Studies

Shyam Gopalakrishnan and Zhaohui S. Qin

Pacific Symposium on Biocomputing 11:511-522(2006)

TAGSNP SELECTION BASED ON PAIRWISE LD CRITERIA AND POWER ANALYSIS IN ASSOCIATION STUDIES

SHYAM GOPALAKRISHNAN, ZHAOHUI S QIN

Center for Statistical Genetics, Department of Biostatistics, School of Public Health, University of Michigan,
Ann Arbor, MI 48109-2029, USA
E-mail: gopalakr@umich.edu

TagSNP selection is an important step in designing case control association studies. Among selection methods that have proliferated, the ones based on pairwise LD measurement are attractive for the purpose of designing association studies. The goal is to minimize the number of markers selected for genotyping in a particular platform and therefore reduce genotyping cost while simultaneously representing information provided by all other markers. Depending on the platform, it is also important to select sets that are robust against occasional genotyping failure. An array of methods has been proposed to effectively select these tagSNPs using various criteria. In this study, we extend the algorithms used in FESTA, a computer program we previously developed for picking tagSNPs using r^2 criteria. We applied FESTA to the HapMap whole chromosome data in two different populations, and we also performed a power analysis for case-control association studies using simulated data. FESTA chooses 294322 tagSNPs in the autosomes in the CEPH samples. The YORUBA samples require 61.5% more tagSNPs than the CEPH samples. The power study showed that limiting ourselves to only tagSNPs, instead of choosing all SNPs in the interval for an association study, results in a power loss of only about 5-10%.

1. Introduction

Rapid advancement in genotyping technologies together with the successful deployment of the International HapMap project [1,2] further popularized the genome-wide association studies, where a dense set of SNP markers across the whole genome is assayed to locate a susceptible chromosomal region that potentially harbors disease predisposing genetic variants. An important initial step in designing an association study is to choose a set of SNPs to represent all variants in the genomic regions of interest. Various algorithms have been proposed for selecting these so called tagSNPs[4,5,6,7,8,9,10,11,12,13]. Most of these strategies aim at choosing "haplotype tagging" SNPs, which are able to capture most of the haplotype diversity, and therefore, could

potentially capture most of the information for association between a trait and the marker loci[14]. Recently, Zhang and Jin[15] and Carlson et al.[4] introduced a simpler criterion for choosing tagSNPs which is based on the pairwise LD measure[16]. These methods search for a small set of SNPs that are in strong LD (measured through pairwise r^2) with all the other SNPs that are not selected for genotyping. Pairwise r^2 is an attractive criterion for tagSNP selection since it is closely related to statistical power for case control association studies, where a directly associated SNP is replaced with an indirectly associated tagSNP[17].

In studies conducted in this manuscript, we adopted the newly developed pairwise LD-based algorithm named **F**ragmented **E**xhaustive **S**earch for **TA**gSNPs[18]. FESTA implements a novel partition step to allow comprehensive search to be carried out. Therefore, it produces fewer tagSNPs than the greedy approach. FESTA also incorporates alternative solution picking according to additional criteria; it can force certain markers in or out of the tagSNP set; and find double coverage tagSNPs. FESTA readily identifies equivalent tagSNP sets, so that additional selection criteria can be incorporated.

We extended the FESTA algorithms by adding a new user-defined criterion, which can be used to pick among the alternative tagSNP sets identified by FESTA. This added flexibility can be quite useful under some situations. We also applied FESTA to whole chromosome HapMap data to identify tagSNPs genome wide. Next, we conducted a simulation study for power analysis; comparing power of detecting association to the disease causing variant using tagSNPs chosen by FESTA. We use two benchmarks to compare the performance of the tagSNPs, (a) all the SNPs in the interval and (b) the same number of random SNPs in the interval.

2. Methods

2.1. *FESTA: Algorithms*

In this section, for the sake of completeness, we briefly review the algorithm implemented in FESTA. The basic idea is to replace a greedy search, where the most connected markers are added sequentially to the tagSNP set, with an exhaustive search where all marker combinations are evaluated. In most settings, our method is guaranteed to find the optimal tagSNP set(s) defined by the r^2 criterion. The details of the FESTA program and the results of comparison with the greedy approach can be found in Qin et al[18].

Define $\mathbb{S}$ to be the set of all SNPs in the precinct under consideration.

Our aim is to find a tagSNP set, denoted by T, a subset of $\mathbb{S}$ such that for all a_i not in T, there exists a_j in T such that $r^2(a_i, a_j) \geq r_0$. In our explanation of the algorithm, we introduce two intermediate SNP sets, P and Q. The candidate set P contains all the markers that are eligible to be chosen as tagSNPs and the target set Q contains all the markers that are yet to be tagged, i.e. no marker in Q is in LD with any tagSNP in T. Typically, the candidate set P is the complement of the tagSNP set T, and $P = Q$. We describe several different algorithms for updating P, Q and T starting with a greedy approach[4]. We then outline successive refinements of a partition and exhaustive search algorithm, designed to allow processing of very large number of markers. Finally, we discuss enhancements to our algorithm.

2.1.1. *Greedy Approach*

The greedy algorithm[4] constructs a tagSNP set by adding the most connected marker to the tagSNP set. It then removes the chosen marker and all connected markers from consideration. This is repeated till there are no markers to be considered. Though the greedy approach is efficient, it does not always find the optimal solution[18].

2.1.2. *Exhaustive Search and Partitioning*

An exhaustive search guarantees the minimum tagSNP set. Genome-wide tagSNP selection requires considering thousands of SNP markers. In these cases, exhaustive searches can not be directly applied due to prohibitive computation costs. Here we use the spatial locality property of LD, i.e. high LD can only be maintained over short distances; therefore we can decompose the set of markers into disjoint precincts such that no marker in a precinct is in high LD with any marker outside the precinct.

After the partitioning step, we perform the tagSNP selection within each precinct using exhaustive search. The result of the greedy algorithm can be used as an upper bound on the number of tagSNPs required in the precinct. The detailed algorithm follows;

(1) Apply BFS[20] to decompose the entire set of markers into precincts $\mathbb{S}_i$ such that strong LD can only be observed within precincts. $\mathbb{S} = \bigcup_{i=1}^{n} \mathbb{S}_i$, and $\mathbb{S}_i \bigcap \mathbb{S}_j = \varnothing \forall i \neq j$;
(2) Within each precinct $\mathbb{S}_i$, set $k_i = 1$,
    a Enumerate all possible k_i-marker combinations. $P_i = Q_i =$

$\mathbb{S}_i$. If no such combination can cover the entire precinct, set $k_i = k_i + 1$ and repeat this step;

b Record all tagSNP sets that can cover the precinct. These form the complete minimum tagSNP sets $T_i^j : j = 1, ..., J_i$, where J_i is the number of such minimum tagSNP sets.

(3) Any combination of tagSNP sets identified from all disjoint subsets forms a tagSNP set for the whole set $\mathbb{S}$, the overall size of such minimum tagSNP sets is $\sum_{i=1}^{n} k_i$, and the total number of minimum tagSNP sets is $\prod_{i=1}^{n} J_i$.

FESTA uses a hybrid of greedy and exhaustive algorithms to solve precincts that are not computationally feasible.

In addition to the basic tagSNP selection, we have implemented the following additional features to assist in tagSNP selection.

(1) Include/Exclude tagSNP markers: As discussed earlier, it may be important to include/exclude some SNPs in the tagSNP set to reduce genotyping cost or ensure genotyping success. Specific SNPs may be included/excluded from the tagSNP set using the mandatory/exclude option in FESTA respectively.
(2) Choosing between alternate solutions based on LD: Exhaustive search may return more than one tagSNP solution for a given precinct. All these sets contain the same number of tagSNPs. Three additional criteria were implemented in FESTA to select one set, (a) Maximize the average r^2 between tagSNPs and the untagged SNPs they represent; (b) Maximize the lowest r^2 between tagSNPs and the untagged SNPs they represent; (c) Minimize the average r^2 among all pairs of tagSNPs;
(3) Double coverage: Current pairwise LD based tagSNP picking algorithms aim to find a tagSNP set such that each SNP marker is either a tagSNP itself or is in LD with at least one of the tagSNPs. Random genotyping failure or error on these tagSNPs can result in loss of power. To be more robust, FESTA implemented a more stringent criterion requiring that, if possible, every untyped marker should be in LD with at least two tagSNPs.

2.1.3. *Extensions to FESTA*

How to choose an optimal set of tagSNPs for genotyping is a practical problem. Specific issues may arise in various scenarios, therefore it is im-

perative that the tagSNP selection tool is flexible enough to let the user impose different optimization rules or apply certain restrictions by themselves. One idea is to introduce an additional criteria to constrain all the available results.

It is common to obtain a large number of tagSNP sets of the same size using the pairwise LD criterion based tagging tools such as FESTA. To select a particular tagSNP set for a particular study, additional criteria need to be introduced to narrow down to the ultimate optimal tagSNP set according to the study requirement. In addition to the additional criteria described above which are already implemented in FESTA, we added a new feature to the FESTA program to allow optimization based on user-specified *ad hoc* variables. An example of such a variable is the quality or design scores of some genotyping platforms such as Illumina. The design score is a continuous variable, which ranges from 0 to 100, where high scores indicate higher genotyping success rate. By entering such scores for all the candidate SNPs, FESTA will to identify the tagSNP set that is optimized in the sense of high design scores among all the tagSNP sets that have the same size according to the pairwise r^2 criteria alone. By adding this constraint, genotyping failure rate will be reduced. Another variable that can be assigned to each SNP is the minor allele frequency (MAF). In this case FESTA can produce a tagSNP set that maximizes the average MAF of all tagSNPs.

This additional variable can also be discrete. For example, whether or not this SNP is in the coding region (cSNP), missense SNP or double hit SNP can be indicated using this variable. FESTA can then report which of the tagSNP sets contained the largest number of such desired SNPs. Another example, some of the markers are "preferred" because they may have already been typed in earlier rounds of studies. To minimize the cost of retyping them, an additional indicator variable can be added showing whether this marker is preferred. Then the tagSNP set containing the most number of preferred tagSNPs can be selected among all tagSNP sets picked by pairwise LD criterion alone. In association studies, these practical constraints can be quite valuable.

2.2. *Power Analysis*

Similar to the power comparison study[21] by Zhang et al., we conducted a simulation study to assess the power of performing case control association studies using tagSNPs identified by FESTA.

2.2.1. *Simulation scheme*

We first simulated a large number of chromosomes consisting of many consecutive SNPs across a 500 kb genomic region using the 'ms' program[22]. It assumes the standard coalescent approximation to the Wright-Fisher model. We assume a constant population size, without subpopulation or gene conversion. We further assume a constant mutation rate throughout this region. The mutation parameter, $\theta = 4N_0\mu$, was chosen to be 200, where N_0 is the effective diploid population size, and $\mu = 10^{-8}$ is the neutral mutation rate per site for this segment. The recombination parameter, $\rho = 4N_0r$ was set to 20. Here, $r = 10^{-9}$ is the probability of recombination in this interval. A hundred populations each containing 2200 chromosomes were generated.

After generating the haplotypes, we randomly chose a marker locus as the disease locus as long as its minor allele frequency was greater than 0.05. The remaining marker loci that had minor allele frequency greater than 0.05 were also retained.

The tagSNP sets were selected using FESTA. The pairwise LD measurement, r^2, was calculated from two marker haplotype frequencies and allele frequencies calculated from the first 200 chromosomes in each population. The case control samples were generated using the remaining 2000 chromosomes, where a hypothetical individual was formed by randomly picking two chromosomes from the pool. The disease status for each hypothetical individual was determined by the penetrance of the individual's disease locus genotype. We assumed the four common disease models: additive, multiplicative, dominant and recessive. In order to mimic a common disease, common variant situation, we specified population disease prevalence to be $P = 0.05$ and $P = 0.1$, and the sibling recurrence risk ratio λ_s[23] was fixed at 1.02.

The association test is based on the difference in allele frequency between case individuals and control individuals[24]. Suppose N is the number of case/control individuals, n_i and m_i are the number of allele A_i in case and control individuals, and p_i and q_i are the frequency of allele A_i in case and control individuals respectively. The test statistic is

$$\chi^2 = \sum_{i=1}^{2} \frac{(n_i - m_i)^2}{n_i + m_i} = 2n \sum_{i=1}^{2} \frac{(p_i - q_i)^2}{p_i + q_i} \tag{1}$$

The above test statistic approximately has a χ^2 distribution with 1 degree of freedom under the null hypothesis of no association. The use of this test

statistic assumes Hardy-Weinberg Equilibrium[25], as shown by Sasieni.

Since there are a large number of marker loci involved, multiple testing is a critical issue for the performance of association test. Simple adjustment approaches such as Bonferroni correction do not perform well. An alternative approach is the permutation test using a Monte Carlo strategy[26]. The maximum value of the test statistic from all markers, denoted as χ^2_{max}, was taken as the test statistic for the association test of the interval. The same test statistic is also calculated for each of the permuted case control samples (generated by switching case control labels for randomly picked individuals). The overall p-value is calculated as the proportion of permuted case control samples that have higher χ^2_{max} value than the one observed from the original case control sample.

The following procedure illustrates our simulation scheme:

(1) Generate 100 populations of 2200 chromosomes using ms program.
(2) In each population, use the first 200 chromosomes to calculate the pairwise r^2, and select tagSNPs using the FESTA program. Subsequently, generate 500 case and 500 controls by sampling from the remaining 2000 chromosomes.
(3) Calculate the test statistic χ^2_{max} under three different cases:

    a. all SNPs in this segment,
    b. only the tagSNPs,
    c. the same number of randomly chosen SNPs.

(4) Perform random permutation 100 times within each case control sample, and calculate the same test statistics χ^2_{max}.
(5) Calculate the overall p-value by determining the proportion of the permuted samples that have higher test statistics than the ones observed from the original case control sample.

Since FESTA provides multiple tagSNP sets for each precinct, and thus for the entire set of markers as well, we chose 3 tagSNP sets, which we used to calculate the power of the tagSNPs to associate the interval to a disease. We used the 3 in-built criteria based on LD, described in 2.1.2, to choose from the alternative solutions. The power of the tagSNPs was reported as the average power of the 3 chosen solutions.

3. Results

3.1. *Genome wide tagSNP selection*

We applied FESTA to pick tagSNPs in the entire human genome using the HapMap data. Two different populations with African and European ancestry were used. A minor allele frequency (maf) threshold of 0.05 was used to prune the SNP map. The total number of SNPs is 742180 in the CEPH samples (European ancestry) and it is 775420 in the YORUBA samples (African ancestry). The total number of tagSNPs in autosomes, using a threshold of $r^2 = 0.8$, identified in the CEPH samples is 294322 and it is 475307 in the YORUBA samples. The YORUBA samples contain almost 61.5% more tagSNPs compared to the CEPH samples. The summary of percentage of tagSNPs in each chromosome is summarized in table 1. It is interesting to note that chromosome 19, the most gene-rich of all human chromosomes, has the largest proportion of tagSNPs in three out of four cases.

Table 1. Proportions of tagSNPs in 22 human autosomes in 2 populations.

	$r^2 > 0.5$			$r^2 > 0.8$		
	Mean	Min	Max	Mean	Min	Max
CEPH	0.251	0.178(8)	0.352(19)	0.410	0.312(8)	0.517(19)
YORUBA	0.434	0.335(8)	0.534(19)	0.624	0.514(9)	0.705(16)

The average computation time, in seconds, in the CEPH samples is 603.54 with a minimum of 169.7 (chr 20) and a maximum of 1313.25 (chr 3), whereas in the YORUBA samples, the average computation time is 628.05, with a minimum of 183.87 (chr 19) and a maximum of 1421.15 (chr 8).

Figure 1 shows the proportion of tagSNps selected in the 22 autosomes in the two populations using the thresholds of $r^2 = 0.5$ and $r^2 = 0.8$.

3.2. *Power results*

We analyzed the power of the tagSNPs to detect association of a disease to the interval as mentioned in the previous section. We simulated a 100 populations of 2200 haplotypes each. The number of selected SNPs (MAF > 0.05) in a population ranges from 236 to 902 with an average of about 584 SNPs per population. The number of tagSNP in a population ranges from 66 to 346, and the average number of tagSNPs selected in a population was about 162 markers. On average, 28% of SNPs were selected as tagSNPs,

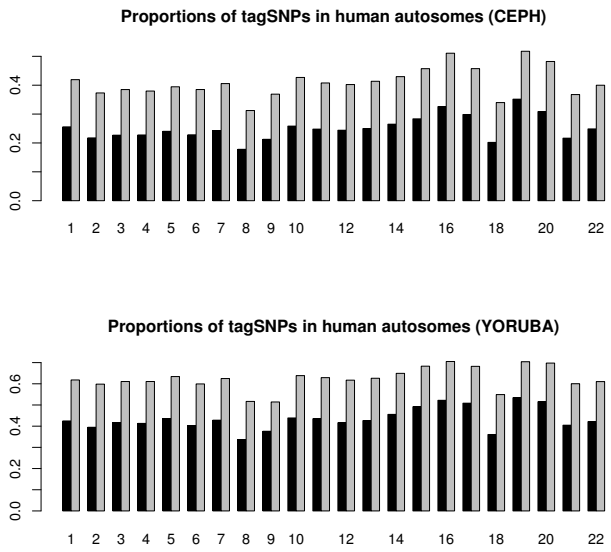

Figure 1. Proportion of tagSNPs in human autosomes using $r^2 = 0.5$ (black) and $r^2 = 0.8$ (gray) thresholds.

across the 100 populations. The disease marker had an average MAF of about 0.216.

Table 2. Power analysis results with disease marker included.

Disease Model	$Prevalence = 0.05$			$Prevalence = 0.1$		
	All SNPs	TagSNPs	Random	All SNPs	TagSNPs	Random
Additive	0.49	0.45	0.407	0.63	0.603	0.563
Multiplicative	0.11	0.11	0.087	0.145	0.1367	0.13
Dominant	0.45	0.42	0.417	0.54	0.547	0.513
Recessive	0.42	0.427	0.407	0.47	0.47	0.437

Table 3. Power analysis results with disease marker excluded.

Disease Model	$Prevalence = 0.05$			$Prevalence = 0.1$		
	All SNPs	TagSNPs	Random	All SNPs	TagSNPs	Random
Additive	0.61	0.603	0.586	0.6	0.6	0.6
Multiplicative	0.1	0.115	0.087	0.15	0.13	0.123
Dominant	0.41	0.39	0.37	0.49	0.4833	0.453
Recessive	0.51	0.49	0.487	0.59	0.58	0.58

We also conducted the power analysis in two ways, by: (i) including the

disease marker in the set of simulated SNPs and (ii) excluding the disease marker from the set of simulated SNPs. Exclusion of the disease marker ensures that none of the sets being analyzed contain the disease marker. We used a threshold of $r^2 = 0.8$ for FESTA to identify the tagSNPs in both cases. The results of the simulation study are summarized in tables 2 and 3 given above.

4. Discussion

As can be observed from the results in the above tables, the loss of power is minimal in the case of tagSNPs selected by the FESTA program. However, it is higher when using a random set of SNPs to represent the information in the interval. There is about a 5-10% loss of power when we choose only tagSNPs instead of all the SNPs in the interval, whereas choosing the same number of random SNPs results in a power loss of about 20%. With higher prevalence of the disease, we get better power to associate the interval to the disease.

We also compared the performance of the algorithm under two other situations, viz., when the disease marker is central to the interval, i.e. if we represent the interval as (0, 1), the disease marker lies in the region [0.4 0.6], and when the disease marker is not central to the interval. We find that a central location of the disease marker favors tagSNPs more heavily than it does random SNPs. If, however the marker is not centrally located, random SNPs perform almost as well as tagSNPs; e.g., in the multiplicative model with $P = 0.1$, the power of the tagSNPs is about 0.17 whereas the random SNPs show a power of 0.13 when the disease marker is central; when the disease marker is not central both the tagSNPs and the random SNPs exhibit a power of about 0.14.

Our current simulation study is still very limited. More comprehensive comparison is needed for us to better understand the effect of tagSNPs under different scenarios.

Pairwise LD is just one criterion for choosing tagSNPs. An interesting alternative is to consider multipoint LD instead. Since a marker may not be in high LD with any single marker, but may be correlated well with haplotypes consisting of multiple linked markers. Therefore, typically multipoint LD[12,27] based tagSNP selection algorithms such as Tagger[28] produce fewer tagSNPs compared to pairwise LD based approaches. However, when conducting association studies using single markers, tagSNPs picked based on pairwise LD criterion are likely to show better power.

The extended FESTA program is freely available at http://www.sph.umich.edu/csg/qin/FESTA.

Acknowledgements

This work is partially supported by NIH RO1-HG002651-01. We would like to thank Dr. Gonçalo Abecasis for insightful discussion about this project, and the four anonymous reviewers for their constructive comments and suggestions.

References

1. The International HapMap Consortium, The International HapMap Project. *Nature* **426**, 789-796, (2003).
2. Sachidanandam R, International SNP Map Working Group A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-933, (2001).
3. Avi-Itzhak HI, Su X, De La Vega FM Selection of minimum subsets of single nucleotide polymorphisms to capture haplotype block diversity. *Pac Symp Biocomputing* 466-477, (2003).
4. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L and Nickerson DA, Selecting a maximally informative set of single-nucleotide polymorphisms for association analysis using linkage disequilibrium. *Am. J. Hum. Genet.* **74**, 106-120, (2004).
5. Hampe J, Schreiber S, Krawczak M Entropy-based SNP selection for genetic association studies. *Hum Genet.* **114**, 36-43, (2003).
6. Halldórsson BV, Bafna V, Lippert R, Schwartz R, De La Vega FM, Clark AG, Istrail S. Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. *Genome Res.* **14**, 1633-1640, (2004).
7. Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29**, 233-237, (2001).
8. Ke X, Cardon LR Efficient selective screening of haplotype tag SNPs. *Bioinformatics* **19**, 287-288, (2003).
9. Lin Z, Altman RB Finding haplotype tagging SNPs by use of principal components analysis. *Am. J. Hum. Genet.* **75**, 850-861, (2004).
10. Meng Z, Zaykin DV, Xu CF, Wagner M, Ehm MG. Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am. J. Hum. Genet.* **73**, 115-130, (2004).
11. Sebastiani P, Lazarus R, Weiss ST, Lunkel LM, Kohane IS and Romani MF, Minimal haplotype tagging *Proc. Natl. Acad. Sci. USA* **100**, 9900-9905, (2003).
12. Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE and Pike MC Choosing haplotype-tagging SNPs based on unphased

genotype data using preliminary sample of unrelated subjects with an example from the multiethnic cohort study. *Hum. Hered.* **55**, 27-36 (2003).
13. Zhang K, Deng M, Chen T, Waterman MS and Sun F A dynamic programming algorithm for haplotype partitioning. *Proc. Natl. Acad. Sci. USA* **99**, 7335-7339, (2002).
14. Chapman JM, Cooper JD, Todd JA, Clayton DG, Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* **56**, 1831, (2003).
15. Zhang, K. and Jin, L. HaploBlockFinder: Haplotype block analysis. *Bioinformatics* **19**, 1300-1301, (2003).
16. Delvin B, Risch N, A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311-322, (1995).
17. Pritchard JK, Przeworski M Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1-14, (2001).
18. Qin ZS, Gopalakrishnan S, Abecasis G, An efficient comprehensive search algorithm for tagSNP selection using linkage disequilibrium criteria. *Unpublished manuscript*. http://www.sph.umich.edu/csg/qin/FESTA. (2005).
19. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee D H, Marjoribanks C, McDonough DP, et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719-1723, (2001).
20. Cormen TH, Leiserson CE, Rivest RL, Introduction to algorithms.*McGraw-Hill Publications*, (2001).
21. Zhang K, Calabrese P, Nordborg M, Sun F, Haplotype Block Structure and Its Applications to Association Studies: Power and Study Designs *Am J Hum Genet.* **71(6)**. 13861394, 2002.
22. Hudson RR, Generating samples under a Wright-Fisher neutral model. *Bioinformatics*, **18**, 337-338, (2002).
23. Risch N, Linkage strategies for genetically complex traits II: The power of affected relative pairs. *Am. J. Hum. Genet.* **46**, 229-41, (1990).
24. Olson JM, Wijsman EM, Design and sample size considerations in the detection of linkage disequilibrium with a disease locus. *Am J Hum Genet* **55**, 574-580, (1994).
25. Sasieni PD, From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253-1261, (1997).
26. McIntyre LM, Martin ER, Simonsen KL, Kaplan NL, Circumventing multiple testing: a multilocus Monte Carlo approach to testing for association. *Genet Epidemiol* **19**, 18-29, (2000).
27. Stram DO, Tag SNP selection for association studies. *Genet Epidemiol.* **27**, 365-374 (2005).
28. Paul de Bakker, Tagger http://www.broad.mit.edu/mpg/tagger