The Pennsylvania State University

The Graduate School

Eberly College of Science

AN EM BASED TAGGING SNP SELECTION ALGORITHM INCORPORATING GENOTYPING ERRORS

A Thesis in

Statistics

by

Tao Yang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2014

The thesis of Tao Yang was reviewed and approved* by the following:

Vernon M. Chinchilli
Professor of Statistics
Thesis Advisor

Bruce G. Lindsay
Professor of Statistics

Aleksandra Slavkovic
Associate Professor of Statistics
Chair of Graduate Program

*Signatures are on file in the Graduate School

ABSTRACT

Many tagging SNP selection methods depend heavily on the estimated haplotype frequencies. One limitation of the existing tagging SNP selection algorithms is that they assume the reported genotypes are error-free. However, genotyping errors are often unavoidable in practice. Recent studies have demonstrated that even slight genotyping errors can lead to serious consequences with regard to haplotype reconstruction and frequency estimation. In this thesis, we present a tagging SNP selection method that allows for genotyping errors. Our method is based on the pair-wise r^2 tagging SNP selection algorithm proposed by Carlson et al. [4]. We modified the standard EM algorithm in Carlson's method to incorporate genotyping errors, in an attempt to obtain better estimates of the haplotype frequencies and r^2 measure. Through extensive simulation studies we compared the performance of our algorithm with that of the original algorithm. We found that the number of tags selected by both methods increased with increasing genotyping errors, though our method led to smaller increase. The power of haplotype association tests using the selected tags decreased dramatically with increasing genotyping errors. The power of single marker tests also decreased, but the reduction was not as much as the reduction in power of haplotype tests. When restricting the mean number of tags selected by both methods to be similar to the baseline number, Carlson's method and our method led to similar power for the subsequent haplotype and single marker tests. Our results showed that, by incorporating random genotyping errors, our method can select tagging SNPs more efficiently than Carlson's method. The computer program that implements our tagging SNP selection algorithm is available at our web site: http://www.personal.psu.edu/tuy104/.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Single nucleotide polymorphisms (SNPs) serve as effective markers in disease gene mapping, especially in association studies. SNPs represent the most frequent form of polymorphism in the human genome (Figure 1). With the recent development of sequencing technology, the availability of SNP markers is expanding quickly. A SNP occurs once about 290 base pairs, which implies the existence of 11 million SNPs among the 3.2 billion base pairs of the human genome [1]. The high abundance of the SNPs requires efficient selection of the SNPs for genotyping since many SNPs are redundant in high linkage disequilibrium (LD) region. If chosen carefully, only a few SNPs in a high LD region will provide enough information to predict much of the information regarding common SNPs in that region. Optimally selecting the minimal informative subset of SNPs will not only reduce genotyping costs, but also reduce the risk of losing the true association in a sea of noise. As a result, many strategies to optimize the choice of tagging SNPs for association analyses have been developed, such as the haplotype diversity based method proposed by Johnson et al. [2], the coefficient of determination method proposed by Stram et al. [3] and the pair wise r^2 method proposed by Carlson et al. [4].

SNPs serve as effective markers for localizing disease susceptibility genes, but current genotyping technologies are inadequate for genotyping all available SNP markers

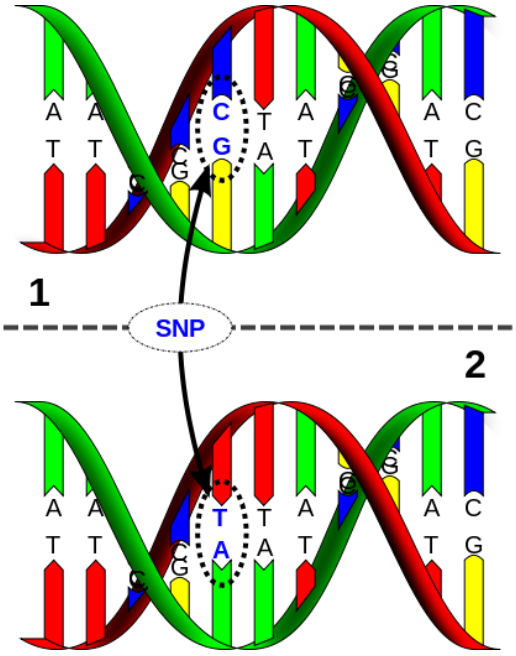

Figure 1: Single nucleotide polymorphisms. Note that DNA molecule 1 differs from DNA molecule 2 at a single base-pair location (a C/T polymorphism).

in a typical linkage/association study. Much attention has recently been paid to methods for selecting the minimal informative subset of SNPs in identifying haplotypes, but there has been little investigation of the effect of missing or erroneous genotypes on the performance of these SNP selection algorithms and subsequent association tests using the selected tagging SNPs. One implicit assumption of current tagging SNP selection algorithms is that there are no genotyping errors. However, even with modern automated sequencing technologies, the problem of erroneous genotypes is still not uncommon. Genotyping errors occur when the genotypes identified by molecular analysis do not correspond to the latent true genotypes of the individuals under study. Virtually every genetic data set includes some erroneous genotypes. Low quality or quantity of the DNA sample, human oversights, shortcomings in genotype scoring software, and biochemical

anomalies all could lead to mistyping. Genotyping methods vary from lab to lab around the world, and can differ dramatically. Sobel et al. [5] estimated the error rates for genotypes generated on an ABI PRISM 377 DNA sequencer and found that the error rate for genotypes directly assigned by the genotyping software could be as high as 13%. In an application of high-density oligonucleotide array based analysis, to determine the distant history of SNPs in current human populations, Hacia et al. [18] found an average error rate of 7% using multiple replicates. Wang et al. [19] studied two SNP sequencing techniques, gel-based sequencing and high-density variation-detection DNA chips, and found that both techniques yielded over 10% error rates. Liu et al. [6] have previously studied the impact of missing and erroneous genotypes on tagging SNPs selected by three algorithms and showed that genotyping errors could have a severe impact on tagging SNP selection. The number of tagging SNPs selected could increase quickly with increasing genotyping errors and the power of subsequent haplotype association test using the selected tagging SNPs could decrease dramatically in the presence of erroneous genotypes. Therefore, genotyping errors need to be taken into account properly in tagging SNP selection.

Since most genotyping technologies only lead to individual marker genotypes, statistical methods are needed to estimate the haplotype frequencies (Figure 2). For unrelated individuals, the Expectation Maximization (EM) algorithms [7,8] are often used to estimate the haplotype frequencies. The haplotype frequency estimates play important roles in many tagging SNP selection algorithms since they are often functions of the haplotype frequencies. One limitation of the classical EM algorithm is that it assumes that

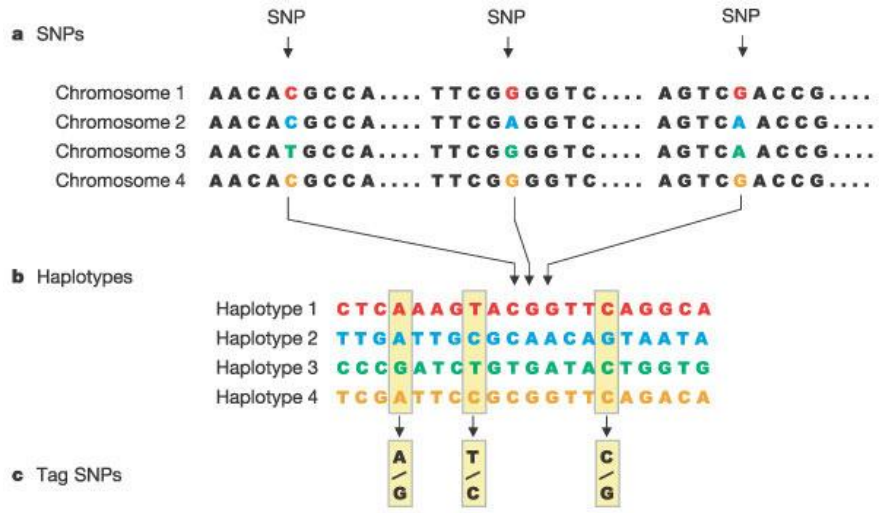

Figure 2: SNPs, Haplotypes, and Tagging SNPs. Note that genotyping just the three tag SNPs out of the 20 SNPs is sufficient to identify these four haplotypes uniquely.

the genotypes are error-free. Kirk and Cardon [9] have shown that even slight amounts of genotyping errors could significantly decrease the accuracy of haplotype reconstruction and frequency estimation. To cope with this problem, Zou and Zhao [10] developed an EM algorithm that accounts for random genotyping errors. Their simulation studies showed that their EM algorithm could give satisfactorily precise haplotype frequency estimates.

In this thesis, we develop a tagging SNP selection algorithm that incorporates genotyping errors. Through two sets of simulations, we compared the performance of our algorithm with that of original Carlson's algorithm in the presence of erroneous genotypes.

Chapter 2

Related Work

Tagging SNP Selection Methods

In this section, we briefly review three tagging SNP selection algorithms which involve haplotype frequencies estimated using the EM algorithm: Clayton's diversity based method htstep [2] , Stram's coefficient of determination based method tagsnp [3], and Carlson's linkage disequilibrium (LD) based method ldSelect [4].

1. Clayton's diversity based algorithm [2].

   Diversity at locus j is defined as:

$$D_j = \sum_{i=1}^{N}\sum_{k=1}^{N}\left(Z_{ij} - Z_{kj}\right)^2 = 2\left[N\sum_{i=1}^{N} z_{ij}^2 - \left(\sum_{i=1}^{N} Z_{ij}\right)^2\right]$$

where N is the total number of chromosomes in the data set. For locus j, if haplotypes i and k have same alleles, then $z_{ij} - z_{kj} = 0$; otherwise, $z_{ij} - z_{kj} = \pm 1$. The haplotype diversity over all loci is then

$$D = \sum_{j=1}^{L} D_j = \sum_{j=1}^{L} 2\left(N_{n_{1j}} - n_{1j}^2\right) = \sum_{j=1}^{L} 2n_{1j}n_{0j} = N^2\sum_{j=1}^{L} 2p_{0j}p_{1j}$$

where n_{0j} and n_{1j} are the number of 0's and 1's at locus j. Suppose the tagging SNPs classify the N haplotypes into G groups, then the residual diversity is defined as within group diversity:

$$R = \sum_{j=1}^{L}\sum_{g=1}^{G} 2n_{0jg}n_{1jg} = N^2 \sum_{j=1}^{L}\sum_{g=1}^{G} 2p_{0jg}p_{1jg}$$

where n_{0jg} and n_{1jg} are the number of 0's and 1's at locus j within group g.

We can define the proportion of diversity explained by the tagging SNPs as

$$P = 1 - \frac{R}{D} = 1 - \frac{\sum_{j=1}^{L}\sum_{g=1}^{G} 2p_{0jg}p_{1jg}}{\sum_{j=1}^{L} 2p_{0j}p_{1j}} > \alpha$$

Tagging SNPs can be selected by defining the cutoff value α.

2. Stram's coefficient of determination (R_h^2) algorithm [3] .

Stram et al. defined haplotype dosage, $\delta_h(H)$ as the count of the number of copies of haplotype h contained in the true haplotype pair H carried by that individual (i.e. $\delta_h(H)$ = 0, 1 or 2). Under HWE, it estimates $E\{\delta_h(H)|G_i\}$ from subject i with genotype G_i can be expressed as

$$E\{\delta_h(H_i)\,|\,G_i\} = \frac{\sum_{H\sim G_i} \delta_h(H) p_{h_1} p_{h_2}}{\sum_{H\sim G_i} p_{h_1} p_{h_2}}$$

where $H\sim G_i$ indicates haplotype pairs (h_1, h_2) that are compatible with the observed genotype G_i . p_{hj} refers to the frequency of haplotype j.

The squared correlation R_h^2 between the true and predicted haplotype dosage ($E\{\delta_h(H)|G_i\}$) can be expressed as the proportion of the total variance of $\delta_h(H_i)$ explained by the genotype data

$$R_h^2 = \frac{\mathrm{var}[E\{\delta_h(H_i)\,|\,G_i\}]}{2p_h(1-p_h)} = \frac{\sum_G \{E(\delta_h\,|\,G)^2 P(G)\} - (2p_h)^2}{2p_h(1-p_h)}$$

For the reduced set of tagging SNPs, R_h^2 can be computed as above. The algorithm looks for the best set of tagging SNPs that maximize the minimum value of R_h^2 calculated for each common haplotype.

3. Carlson's r^2 linkage disequilibrium based algorithm [4].

Compute r^2 statistic for all SNP pairs based on the estimated two-locus haplotype frequencies. Select the single locus exceeding the r^2 threshold with the maximum number of other loci. Group this locus and all associated loci into one bin. Within one bin, any SNP locus exceeding the r^2 threshold with all other loci can be specified as a tagging SNP for the bin. Thus, one or more SNPs within one bin could be specified as tagging

SNPs, and only one tagging SNP would need to be genotyped per bin. The binning process is iterated until all loci are binned. If an SNP does not exceed the r^2 threshold with any other SNP, then it is treated as a singleton bin.

Chapter 3

Methods

EM algorithm

The EM algorithm developed by Zou and Zhao [10] takes random genotyping errors into account. They assumed the error rate from one allele to another allele is the same at the same marker locus and used ε_l to denote the error rate at marker locus l, where l = 1, …, k. Suppose the total number of haplotypes is H. If the haplotype frequency estimates at the j^{th} iteration are $p^{(j)} = (p_1^{(j)}, \ldots, p_H^{(j)})$, then

1. The E-step:

$$Q(p \mid p^{(j)}) = \sum_{i=1}^{n} E_{d_i \mid (g_i, p^{(j)})} \left[\log P(g_i, d_i)\right]$$

$$= \sum_{i=1}^{n} \sum_{d} \log P(g_i, d) \cdot P(d \mid g_i, p^{(j)})$$

$$= \sum_{i=1}^{n} \sum_{d} \left[\log P(g_i \mid d) + \log P(d)\right] \cdot P(d \mid g_i, p^{(j)})$$

and

$$P(d \mid g_i, p^{(j)}) = \frac{P(g_i \mid d) \cdot P(d \mid p^{(j)})}{P(g_i \mid p^{(j)})}$$

$$= \frac{P(g_i \mid d) \cdot P(d \mid p^{(j)})}{\sum_d P(g_i \mid d) \cdot P(d \mid p^{(j)})}$$

where g_i denote the observed genotype of individual i (i = 1, …, n) and d_i stands for the true diplotype of individual i, and $\sum_d$ means sum over all possible diplotypes.

Let $d_{i(u)}$ denote the possible diplotype of individual i consistent with the observed genotype g_i, then

$$P(g_i \mid d) = \sum_{u=1}^{U_i} P(d_{i(u)} \mid d) = \sum_{u=1}^{U_i} \prod_{l=1}^{k} P(d_{i(u)}^{l} \mid d^{l})$$

$$= \sum_{u=1}^{U_i} \prod_{l=1}^{k} [1-(I_l - 1)\varepsilon_l]^{z_{iuld}} \, \varepsilon_l^{2-z_{iuld}}$$

where $d_{i(u)}^{l}$ is the diplotype of individual i consistent with the observed genotype at locus l and d^l is the true diplotype of individual i at locus l. I_l is the total number of alleles at locus l and Z_{iuld} represents the number of the same alleles on the same chromosome for $d_{i(u)}^{l}$ and d^l.

For $d = h_s h_t$, $\quad P(d \mid p^{(j)}) = \begin{cases} p_s^{(j)\,2} & s = t \\ 2 p_s^{(j)} p_t^{(j)} & s \neq t \end{cases}$

2. The M-step:

The maximization step then provides the new haplotype frequency estimates, $p^{(j+1)}$. For haplotype t, the new frequency estimate is

$$p_t^{(j+1)} = \frac{1}{2n}\sum_{i=1}^{n}\sum_{d}\delta_{dt}\cdot P(d \mid g_i, p^{(j)})$$

where δ_{dt} denotes the number of copies of haplotype t within diplotype d. The new haplotype frequency estimates then serve as the initial values for another iteration. The process continues until successive values are sufficiently close.

Discussion

Suppose the allele to allele error rate is ε, which is the same across all allele changes at any specified marker locus, then

$$\Pr(T_{AA}\ \&\ O_{Aa\_or\_aa}) = \Pr(O_{Aa} \mid T_{AA})\Pr(T_{AA}) + \Pr(O_{aa} \mid T_{AA})\Pr(T_{AA}) = 2\varepsilon(1-\varepsilon)P_A^2 + \varepsilon^2 P_A^2$$

where A and a are the two alleles at this locus, P_A is the allele frequency of allele A, T denotes the true genotype and O stands for the observed genotype. Similarly, we have

$$\Pr(T_{aa} \,\&\, O_{Aa\_or\_AA}) = 2\varepsilon(1-\varepsilon)P_a^2 + \varepsilon^2 P_a^2$$
$$\Pr(T_{Aa} \,\&\, O_{aa\_or\_AA}) = 4\varepsilon(1-\varepsilon)P_A P_a$$

The overall error rate at this locus is the sum of the above three probabilities:

$$P(error) = 2\varepsilon(1-\varepsilon) + \varepsilon^2(P_A^2 + P_a^2)$$

Therefore, the overall genotyping error rate at a given locus is a function of the allele to allele error rate ε and marker allele frequencies. Since ε is usually low, $\varepsilon^2(P_A^2 + P_a^2)$ is likely to be a very small number. Thus, the overall genotype error rate approximates 2ε(1-ε) and is about the same across loci with different allele frequencies.

Tagging SNP selection algorithm that accounts for genotyping error

Carlson's r^2 linkage disequilibrium based algorithm [4] computes r^2 statistic for all SNP pairs based on the two-locus haplotype frequencies estimated using the standard EM algorithm. It first selects the single locus exceeding the r^2 threshold with the maximum number of other loci. Then, it groups this locus and all associated loci into one bin. Within one bin, any SNP locus exceeding the r^2 threshold with all other loci can be specified as a tagging SNP for the bin. Thus, one or more SNPs within one bin could be specified as tagging SNPs, and only one tagging SNP would need to be genotyped per bin. The binning process is iterated until all loci are binned. If an SNP does not exceed

the r^2 threshold with any other SNP, it is treated as a singleton bin. We modified Carlson's algorithm to allow for random genotyping errors by replacing the standard EM with the above EM algorithm.

Chapter 4

Experiments

Simulation

We conducted two sets of simulations. The first set was simulated according to the haplotype frequencies in the CYP19 region analyzed by Stram et al. [3]. The CYP19 region had a total of 19 SNPs. These 19 SNPs are within one haplotype block as judged by the methods of Gabriel et al [11]. In each simulated data set, the reported haplotype frequencies in the CYP19 region were used as weights in the sampling of individual's chromosomes. For each individual, a pair of random numbers was generated, each corresponding to a particular haplotype. One individual's genotype was formed by randomly pairing two simulated haplotypes and treating the phase as unknown. One locus was chosen to be the causal locus (locus 12 with minor allele frequency of 0.38) and disease phenotypes were simulated according to the genotypes at this locus. We simulated three disease inheritance models: dominant, additive and recessive models. The penetrance, causal allele frequency and relative risk are listed in Table 1. Genotypes in the second set of simulations were simulated using a coalescent algorithm that assumes a constant-sized population with a constant recombination rate over the entire region as implemented in computer program CoaSim [12]. 30 evenly spaced SNPs within a roughly 0.01 cM interval were simulated. (rho = 4Nr were set to 4, where N represents the effective population size (10,000) and r corresponds to the recombination rate). Each

of the 30 markers had a minor allele frequency greater than 0.05. The bi-allelic disease causal locus was placed between marker 15 and 16 and had a mutant allele frequency between 0.1 and 0.2. Case-control status was generated under an additive disease inheritance model (Table 1).

Table 1: Penetrance and relative risk of the simulated data sets

Data sets	Disease model	Causal allele frequency	Penetrance			Relative Risk		
			DD	Dd	dd	DD	Dd	dd
CYP19 region	Dominant	0.3783	0.3	0.3	0.1	3	3	1
	Additive	0.3783	0.3	0.2	0.1	3	2	1
	Recessive	0.3783	0.3	0.1	0.1	3	1	1
Coalescent model	Additive	0.1 - 0.2	0.5	0.3	0.1	5	3	1

We simulated 1,000 data sets for each scenario. Each simulated data set composed of two sub data sets: a training data set for tagging SNP selection and a test data set with genotypes only at the selected tagging SNP loci for association studies. The CYP19 data sets had 25 cases and 25 controls in each training data set and 100 cases and 100 controls in each test data set. The coalescent data sets had 30 cases and 30 controls in each training data set and 150 cases and 150 controls in each test data set. After simulating the disease phenotypes, we removed the genotypes at the causal locus and pretended the causal locus was unmeasured. In both sets of simulations, the causal locus was withheld from both the training and test data sets.

Genotyping errors were introduced randomly at each locus within each individual according to the random error model (e.g. $\Pr(O_{aa} \mid T_{AA}) = \varepsilon^2$, etc.). The allele to allele error rate, ε, was the same across all loci.

Tagging SNPs were selected using the training data set. Since recent work [13,14] has suggested that cases and controls should be pooled together for haplotype association studies, all 100 chromosomes in the training data set were used to identify the tagging SNPs. For the CYP19 data sets, r^2 threshold for binning SNPs was set to 0.8. For the coalescent data sets, r^2 threshold for binning SNPs was set to 0.5. For single marker analysis, the association between the disease phenotype and marker allele or genotype frequencies was tested at each tagging SNP locus. Since many cells had expected counts less than 5, Fisher's exact tests were used instead of the usual Chi-square tests. Fisher's exact test is based on exact probabilities from the hypergeometric distribution. The Chi-square test relies on a large sample approximation. Therefore, Fisher's exact test is preferred in situations where a large sample approximation is inappropriate. Within each simulated data set multiple tagging SNP loci were tested for association with the disease phenotype, and a data set was considered to have significant association if at least one SNP had a p-value less than the cutoff value. Therefore, to adjust for multiple comparisons, the p-values obtained from the single marker analysis were corrected by Bonferroni correction. Usually, Bonferroni correction would be too conservative to adjust for multiple testing on SNPs because of the correlation among SNPs. However, we do not expect Bonferroni correction to be too conservative here because the SNPs tested were

selected as tagging SNPs and thus were unlikely to be highly correlated with each other. The power of the single marker tests was estimated as the proportion of data sets with at least one SNP with adjusted p-value less than 0.05. Haplotype association tests were also performed in the test data sets based on genotypes at the tagging SNP loci using the score test implemented in the haplo.stats program [13]. The power of the haplotype tests were estimated as the proportion of data sets with a global p-value less than 0.05.

Results

We compared the performance of Carlson's method with that of our modified method in the presence of random genotyping errors through simulation studies. For the data set simulated according to the CYP19 haplotype frequencies, the mean number of selected tagging SNPs and power of haplotype and single marker tests using the selected tagging SNPs are shown in Table 2. The allele to allele error rate ε was set to be 0, 0.01, 0.02, 0.03, 0.04 and 0.05. The overall error rate per locus was approximately 0, 0.02, 0.039, 0.058, 0.077 and 0.095 respectively. We noticed that at baseline with no errors (ε equaled 0), Carlson's method and our modified method selected exactly the same number of tags as expected. In addition, we found several trends. First, the tagging SNPs selected by both methods increased with increasing genotyping errors though the number of tags selected by our method increased at a lower rate. The discrepancy between the numbers of tags selected by the two methods increased with increasing erroneous genotypes. When ε equaled 0.05, Carlson's method selected almost all the SNPs as tagging SNPs (17.3 tags out of a total of 18 SNPs) while our method selected about 7 tags less. Second,

the power of haplotype association tests using tags selected by both methods reduced dramatically with increasing genotyping errors. Our method led to higher power, especially for cases with high genotyping errors. The disease phenotype was caused by a single causal locus and the alleles at adjacent loci were likely to be associated with the causal allele. In the absence of erroneous genotypes, the data sets were likely to be dominated by several common haplotypes. With increasing genotyping errors, more and more low frequency haplotypes were probably introduced and the effect of the causal allele was probably being split up over many more erroneous haplotypes. Therefore, the power of the haplotype tests decreased dramatically. Carlson's method picked more tags than our methods in the presence of genotyping errors, thus more low frequency haplotypes were probably introduced which resulted in lower power. Third, the power of both allele-wise and genotype-wise single marker tests decreased with increasing errors, but the reduction is not as much as the reduction in power of haplotype tests. Carlson's method led to higher power, especially for cases with high genotyping errors. This is probably because within a certain simulated data set, random genotyping error may cause one or more associated loci to become non-significant. However, it was unlikely for the random genotyping errors to simultaneously cause all associated loci to become non-significant. Thus, there was not much decrease in the power of the single marker tests. Carlson's method probably resulted in more associated loci being tested thus it was more difficult for random genotyping error to cause all associated loci to become non-significant. Therefore, Carlson's method led to higher power. We saw similar patterns in all three disease inheritance models.

Table 2: Number of tags selected and power of subsequent single marker and haplotype association tests based on the 1,000 data sets simulated according to the haplotype frequencies in the CYP19 region.

Disease model	Allele wise error rate (ε)	Algorithm	Mean number of tags [(a)]	Agree-ment% [(b)]	Power of Haplotype test	Power of allele test	Power of genotype test
Additive model	0	Carlson	6.566 (0.727)	100%	0.640	0.733	0.671
		Modified	6.566 (0.727)	100%	0.640	0.733	0.671
	0.01	Carlson	8.429 (1.199)	100%	0.481	0.690	0.646
		Modified	7.82(0.981)	84.24%	0.483	0.698	0.645
	0.02	Carlson	11.131 (1.76)	100%	0.351	0.670	0.625
		Modified	8.595 (1.165)	68.17%	0.379	0.664	0.612
	0.03	Carlson	14.11 (1.875)	100%	0.251	0.662	0.622
		Modified	9.284 (1.392)	61.38%	0.333	0.634	0.598
	0.04	Carlson	16.192 (1.44)	100%	0.205	0.659	0.637
		Modified	9.953 (1.497)	59.75%	0.284	0.625	0.558
	0.05	Carlson	17.306 (0.95)	100%	0.164	0.687	0.649
		Modified	10.621 (1.54)	60.82%	0.233	0.601	0.546
Dominant model	0	Carlson	6.543 (0.724)	100%	0.725	0.773	0.890
		Modified	6.543 (0.724)	100%	0.725	0.773	0.890
	0.01	Carlson	8.386 (1.222)	100%	0.551	0.754	0.871
		Modified	7.707 (0.985)	82.17%	0.571	0.745	0.861
	0.02	Carlson	11.185 (1.74)	100%	0.434	0.736	0.859
		Modified	8.509 (1.169)	67.42%	0.466	0.715	0.831
	0.03	Carlson	14.254 (1.81)	100%	0.318	0.726	0.860
		Modified	9.243 (1.357)	60.87%	0.392	0.697	0.810
	0.04	Carlson	16.338 (1.36)	100%	0.256	0.733	0.856
		Modified	9.95 (1.473)	59.36%	0.312	0.684	0.787
	0.05	Carlson	17.39 (0.876)	100%	0.211	0.736	0.861
		Modified	10.683 (1.43)	60.97%	0.266	0.657	0.766
Recessive model	0	Carlson	6.491 (0.739)	100%	0.553	0.666	0.757
		Modified	6.491 (0.739)	100%	0.553	0.666	0.757
	0.01	Carlson	8.394 (1.259)	100%	0.375	0.633	0.719
		Modified	7.679 (1.034)	83.37%	0.379	0.632	0.726
	0.02	Carlson	11.182 (1.78)	100%	0.255	0.622	0.715
		Modified	8.49 (1.211)	67.42%	0.293	0.595	0.696
	0.03	Carlson	14.205 (1.82)	100%	0.210	0.609	0.713
		Modified	9.251 (1.396)	61.13%	0.256	0.588	0.699
	0.04	Carlson	16.308 (1.43)	100%	0.154	0.621	0.715
		Modified	9.945 (1.450)	59.54%	0.220	0.559	0.645
	0.05	Carlson	17.345 (0.92)	100%	0.140	0.624	0.714
		Modified	10.57 (1.466)	60.35%	0.172	0.543	0.621

(a) Mean and standard deviation of the number of tagging SNPs selected.

(b) Average percent overlap of the tagging SNPs selected by the two methods. Percent overlap was obtained by dividing the number of common SNPs selected by the number of tags selected by Carlson's method.

To make fair power comparisons, we restricted the mean number of tags selected in all scenarios to be similar by adjusting the r^2 thresholds in both algorithms. Table 3 lists the results. Again, the power of the haplotype test decreased quickly with increasing errors and the power of single marker tests also decreased, but not as fast as that of the haplotype test. Despite the fact that some SNPs selected by Carlson's method and our method were different as can be seem from the percent agreement column, the two methods led to similar power for both the haplotype test and single marker tests. For the additive disease inheritance model, the power of haplotype tests using tags selected by both methods decreased from 64% at baseline to 28% with ε equaled 0.05. For single marker analyses, when ε was set to 0.05, both methods led to about 15% reduction in power compared to baseline. Again, we saw similar patterns across all three disease inheritance models though the three disease models had different power at baseline.

Similar trends showed up in results obtained from data sets simulated under the coalescent model. Table 4 lists the mean number of tags selected and power of haplotype and single marker tests using the selected tagging SNPs. Both Carlson's method and our modified method selected more tagging SNPs with increasing genotyping errors though our method picked less number of tags especially with more genotyping errors. At baseline, both methods selected 9.6 tags. When ε was 0.05, Carlson's method selected over 18 tags, a one fold increase in tag number compared to baseline while our method picked about 13.8 tags, a less than one half fold increase in tag number compared to baseline. Both methods led to a big reduction in power of haplotype tests though our method led to slightly higher power with high genotyping errors. Interestingly, the power

Table 3: Number of tags selected and power of subsequent single marker and haplotype association tests based on the 1,000 data sets simulated according to the haplotype frequencies in the CYP19 region when the average numbers of tags selected were restricted to be close to the baseline number.

Disease model	Allele wise error rate(ε)	Algorithm [(a)]	Mean number of tags [(b)]	Agree-ment% [(c)]	Power of Haplotype test	Power of allele test	Power of genotype test
Additive model	0	Carlson (0.8)	6.566 (0.727)	100%	0.640	0.733	0.671
		Modified (0.8)	6.566 (0.727)	100%	0.640	0.733	0.671
	0.01	Carlson (0.65)	6.581 (1.057)	100%	0.505	0.708	0.648
		Modified (0.68)	6.619 (1.022)	94.87%	0.511	0.711	0.649
	0.02	Carlson (0.57)	6.555 (1.000)	100%	0.432	0.673	0.615
		Modified (0.64)	6.583 (0.943)	90.77%	0.438	0.668	0.622
	0.03	Carlson (0.51)	6.598 (1.085)	100%	0.380	0.649	0.603
		Modified (0.61)	6.592 (0.989)	87.32%	0.383	0.653	0.598
	0.04	Carlson (0.45)	6.578 (1.149)	100%	0.365	0.606	0.550
		Modified (0.58)	6.584 (1.027)	85.05%	0.346	0.607	0.546
	0.05	Carlson (0.4)	6.585 (1.215)	100%	0.282	0.578	0.524
		Modified (0.55)	6.57 (1.140)	82.62%	0.282	0.590	0.507
Dominant model	0	Carlson (0.8)	6.543 (0.724)	100%	0.725	0.773	0.890
		Modified (0.8)	6.543 (0.724)	100%	0.725	0.773	0.890
	0.01	Carlson (0.65)	6.529 (1.034)	100%	0.607	0.763	0.877
		Modified (0.68)	6.58 (1.029)	94.95%	0.604	0.761	0.877
	0.02	Carlson (0.57)	6.57 (1.017)	100%	0.491	0.713	0.837
		Modified (0.64)	6.613 (0.971)	91.14%	0.509	0.718	0.840
	0.03	Carlson (0.5)	6.545 (1.065)	100%	0.431	0.688	0.804
		Modified (0.6)	6.559 (0.991)	88.30%	0.424	0.677	0.792
	0.04	Carlson (0.44)	6.491 (1.103)	100%	0.386	0.650	0.758
		Modified (0.57)	6.51 (1.016)	85.65%	0.383	0.665	0.769
	0.05	Carlson (0.39)	6.5 (1.159)	100%	0.363	0.623	0.731
		Modified (0.54)	6.497 (1.068)	83.12%	0.346	0.613	0.733
Recessive model	0	Carlson (0.8)	6.491 (0.739)	100%	0.553	0.666	0.757
		Modified (0.8)	6.491 (0.739)	100%	0.553	0.666	0.757
	0.01	Carlson (0.65)	6.481 (1.057)	100%	0.413	0.639	0.727
		Modified (0.68)	6.501 (1.050)	94.44%	0.411	0.638	0.728
	0.02	Carlson (0.57)	6.493 (1.052)	100%	0.353	0.601	0.706
		Modified (0.64)	6.509 (0.986)	90.65%	0.339	0.604	0.701
	0.03	Carlson (0.5)	6.447 (1.045)	100%	0.315	0.579	0.661
		Modified (0.6)	6.45 (0.985)	87.70%	0.305	0.591	0.668
	0.04	Carlson (0.44)	6.44 (1.158)	100%	0.271	0.558	0.627
		Modified (0.57)	6.465 (1.047)	84.31%	0.274	0.548	0.616
	0.05	Carlson (0.39)	6.478 (1.224)	100%	0.247	0.545	0.617
		Modified (0.54)	6.45 (1.123)	82.06%	0.237	0.530	0.579

(a) The numbers inside the parentheses are the r^2 threshold used with different ε.
(b) Mean and standard deviation of the number of tagging SNPs selected.
(c) Average percent overlap of the tagging SNPs selected by the two methods. Percent overlap was obtained by dividing the number of common SNPs selected by the number of tags selected by Carlson's method.

Table 4: Number of tags selected and power of subsequent single marker and haplotype association tests based on the 1,000 data sets simulated under the additive coalescent model.

Allele wise error rate (ε)	Algorithm	Mean number of tags [(a)]	Agree-ment% [(b)]	Power of Haplotype test	Power of allele test	Power of genotype test
0	Carlson	9.635 (2.118)	100%	0.930	0.873	0.848
	Modified	9.635 (2.118)	100%	0.930	0.873	0.848
0.01	Carlson	10.684 (2.382)	100%	0.889	0.889	0.868
	Modified	10.264 (2.332)	88.83%	0.884	0.881	0.862
0.02	Carlson	12.238 (2.866)	100%	0.838	0.886	0.861
	Modified	11.04 (2.598)	78.74%	0.845	0.878	0.860
0.03	Carlson	14.272 (3.394)	100%	0.759	0.893	0.873
	Modified	12.039 (2.950)	72.80%	0.757	0.879	0.861
0.04	Carlson	16.362 (3.838)	100%	0.659	0.894	0.884
	Modified	12.978 (3.251)	69.56%	0.670	0.880	0.855
0.05	Carlson	18.587 (4.102)	100%	0.596	0.907	0.887
	Modified	13.861 (3.623)	66.16%	0.638	0.878	0.855

(a) Mean and standard deviation of the number of tagging SNPs selected.
(b) Average percent overlap of the tagging SNPs selected by the two methods. Percent overlap was obtained by dividing the number of common SNPs selected by the number of tags selected by Carlson's method..

of single marker tests using tags selected by both methods slightly increased with increasing erroneous genotypes. Carlson's method led to slightly higher power in the presence of error. For instance, both methods had 87.3% power in the allele-wise test at baseline. When ε increased to 0.05 the power of Carlson's method increased to 90.7%, while our modified method had 87.8% power. The slight increase in power may be caused by the fact that more tags were selected with increasing errors. As a result, single marker tests were performed probably at more associated loci and thus power increased a little bit.

Similar to what we did for the CYP19 data sets, we restricted the mean number of tags selected under different genotyping error rates to be similar to the means at baseline by adjusting the r^2 thresholds. The results are shown in Table 5. A similar pattern can be seen as in the CYP19 data sets. The power of haplotype tests decreased much faster than that of the allele-wise and genotype wise single marker tests. Although some tags selected by the two methods were different, both methods led to similar power for both the haplotype test and single marker tests.

The tagging SNP selection algorithms we considered here utilize r^2 measure to pick tags. Devlin and Risch [15] studied the properties of several LD measures and found that r^2 was not invariant to the sampling strategies. Their study found that r^2 was affected by case-control sampling strategies since disease haplotypes were sampled at a rate higher than their population frequencies in a typical case-control study. To study the impact of sampling strategies on tagging SNP selection and power of subsequent single

marker and haplotype association tests, we considered three designs for the CYP19 training data sets: half cases and half controls (25 affected individuals and 25 unaffected individuals), cases only (50 affected individuals) and controls only (50 unaffected individuals). The data sets were simulated under the additive disease inheritance model (Table 6), and ε was set to be 0, 0.01, 0.03, and 0.05. We noticed that for each method, with a certain level of errors, similar numbers of tags were selected using training data sets generated by all three sampling strategies. The subsequent power of haplotype and single marker tests using tags selected by each method was similar across the three sampling strategies under each genotyping error rate.

Table 5: Number of tags selected and power of subsequent single marker and haplotype association tests based on the 1,000 data sets simulated under the additive coalescent model when the average numbers of tags selected were restricted to be close to the baseline number.

Allele wise error rate (ε)	Algorithm [(a)]	Mean number of tags [(b)]	Agree-ment% [(c)]	Power of Haplotype test	Power of allele test	Power of genotype test
0	Carlson (0.5)	9.635 (2.118)	100%	0.930	0.873	0.848
	Modified (0.5)	9.635 (2.118)	100%	0.930	0.873	0.848
0.01	Carlson (0.43)	9.598 (2.171)	100%	0.883	0.863	0.839
	Modified (0.46)	9.65 (2.180)	94.59%	0.875	0.869	0.847
0.02	Carlson (0.37)	9.683 (2.292)	100%	0.827	0.858	0.836
	Modified (0.41)	9.575 (2.267)	88.52%	0.839	0.849	0.830
0.03	Carlson (0.31)	9.622 (2.327)	100%	0.771	0.847	0.824
	Modified (0.37)	9.599 (2.340)	84.83%	0.769	0.850	0.826
0.04	Carlson (0.27)	9.709 (2.458)	100%	0.703	0.829	0.816
	Modified (0.34)	9.661 (2.442)	82.31%	0.712	0.832	0.807
0.05	Carlson (0.23)	9.665 (2.494)	100%	0.667	0.831	0.809
	Modified (0.31)	9.678 (2.518)	81.05%	0.679	0.831	0.807

(a) The numbers inside the parentheses are the r^2 threshold used with different ε.
(b) Mean and standard deviation of the number of tagging SNPs selected.
(c) Average percent overlap of the tagging SNPs selected by the two methods. Percent overlap was obtained by dividing the number of common SNPs selected by the number of tags selected by Carlson's method.

Table 6: Number of tags selected and power of subsequent single marker and haplotype association tests based on the 1,000 data sets simulated according to the haplotype frequencies in the CYP19 region when different sampling strategies were applied.

Allele wise error rate (ε)	Sampling strategy	Algorithm	Mean number of tags [(a)]	Agree-ment% [(b)]	Power of Haplotype test	Power of allele test	Power of genotype test
0	25 case/ 25 control	Calson	6.566 (0.727)	100%	0.640	0.733	0.671
		Modified	6.566 (0.727)	100%	0.640	0.733	0.671
	50 case	Calson	6.351 (0.833)	100%	0.655	0.710	0.662
		Modified	6.351 (0.833)	100%	0.655	0.710	0.662
	50 control	Calson	6.685 (0.647)	100%	0.638	0.698	0.663
		Modified	6.685 (0.647)	100%	0.638	0.698	0.663
0.01	25 case/ 25 control	Calson	8.429 (1.199)	100%	0.481	0.690	0.646
		Modified	7.820 (0.981)	84.24%	0.483	0.698	0.645
	50 case	Calson	8.453 (1.320)	100%	0.481	0.658	0.632
		Modified	7.726 (1.137)	81.95%	0.482	0.664	0.628
	50 control	Calson	8.424 (1.164)	100%	0.494	0.666	0.623
		Modified	7.77 (0.851)	82.68%	0.488	0.675	0.631
0.03	25 case/ 25 control	Calson	14.11 (1.875)	100%	0.251	0.662	0.622
		Modified	9.284 (1.392)	61.38%	0.333	0.634	0.598
	50 case	Calson	14.392 (1.844)	100%	0.237	0.639	0.600
		Modified	9.776 (1.577)	64.03%	0.303	0.618	0.579
	50 control	Calson	14.416 (1.907)	100%	0.271	0.649	0.618
		Modified	9.046 (1.277)	59.00%	0.313	0.639	0.566
0.05	25 case/ 25 control	Calson	17.306 (0.948)	100%	0.164	0.687	0.649
		Modified	10.621 (1.536)	60.82%	0.233	0.601	0.546
	50 case	Calson	17.389 (0.908)	100%	0.165	0.661	0.629
		Modified	11.34 (1.514)	64.70%	0.205	0.573	0.524
	50 control	Calson	17.409 (0.859)	100%	0.157	0.665	0.620
		Modified	10.3 (1.539)	58.73%	0.248	0.587	0.546

(a) Mean and standard deviation of the number of tagging SNPs selected.

(b) Average percent overlap of the tagging SNPs selected by the two methods. Percent overlap was obtained by dividing the number of common SNPs selected by the number of tags selected by Carlson's method.

Discussion

Liu et al. [6] showed that erroneous genotypes could have a severe impact on tagging SNP selection and power of subsequent association tests by increasing the number of tags selected and decreasing the power of haplotype association test using the selected tags. Therefore, genotype errors should not be neglected in tagging SNP selection. In this study, we modified Carlson's pair wise r^2 based tagging SNP selection algorithm to allow for random genotyping errors by replacing the standard EM algorithm in haplotype frequency estimation by the EM algorithm that takes random genotyping errors into account. Our simulation results showed that our modified algorithm selected less number of tags compared to those selected by Carlson's algorithm when genotyping errors were present. When restricting the mean number of tags selected by both methods to be similar to the baseline number, Carlson's method and our method led to similar power for the subsequent haplotype and single maker tests. These findings imply that our modified method can select the tagging SNPs more efficiently than Carlson's method in the presence of erroneous genotypes and should be employed when one suspects to have high genotyping errors in the data set.

In our simulation studies, we have assumed that the allele to allele error rate, ε is known. In reality, the error rate can be estimated by genotyping a set of samples multiple times assuming that the true genotype can be inferred correctly if a specimen is genotyped enough times. Taberlet et al. [16] conducted simulation studies and developed a set of rules to assign individual's genotype based on genotyping the specimen multiple

times. For instance, if a sample was genotyped three times and the observed genotypes at a certain locus was AA, AB, BB, according to their rules, the correct genotype was almost certainly the heterozygous, AB. After inferring the correct genotypes, genotyping error rates can be estimated directly based on the counts of the erroneous genotypes and the total number of genotypes. The estimated error rate can be used for the same genotyping technique for other experiments. Alternatively, we can simultaneously estimate the allele frequencies and genotyping error rates by constructing a likelihood function of the observed genotype combinations. More recently, Kalinowski et al. [17] have developed a DNA census algorithm that estimates dropout and misprint rates for every specimen before clustering specimens into sets to evaluate the evidence of identity.

Chapter 5

Conclusion

In this thesis, we have developed a tagging SNP selection algorithm to allow for random genotyping errors. In addition to random genotyping errors, we can also incorporate other types of errors. The only thing needs to be changed would be $P(d_{i(u)}^{l} \mid d^{l})$ in calculating $P(g_i \mid d)$ in the estimation step of the EM algorithm. For instance, if the homozygous to homozygous error rate is ε_1 and heterozygous to homozygous error rate is ε_2, then $P(d_{i(u)}^{l} \mid d^{l})$ would be ε_2 for $d_{i(u)}^{l}$ being AA and d^{l} being AB.

Our study has direct implications for optimal selection of tagging SNPs in disease gene mapping. Our method can select fewer number of tags relative to that selected by the original Carlson's method and keep the power of subsequent association tests comparable to the power resulted from the original method. The resources saved on genotyping more SNPs can be used to genotype more individuals to increase the statistical power of the association tests. The computer program that implements our tagging SNP selection algorithm is available at our web site: http://www.personal.psu.edu/tuy104/.

References

1. Kruglyak L, Nickerson DA: Variation is the spice of life. Nat Genet 2001;27:234-236.
2. Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA: Haplotype tagging for the identification of common disease genes. Nat Genet 2001;29:233-237.
3. Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Pike MC: Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. Hum Hered 2003;55:27-36.
4. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 2004;74:106-120.
5. Sobel E, Papp JC, Lange K: Detection and integration of genotyping errors in statistical genetics. Am J Hum Genet 2002;70:496-508.
6. Liu W, Zhao W, Chase GA: The impact of missing and erroneous genotypes on tagging SNP selection and power of subsequent association tests. Hum Hered 2006;61:31-44.
7. Hill WG: Estimation of linkage disequilibrium in randomly mating populations. Heredity 1974;33:229-239.

8. Excoffier L, Slatkin M: Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 1995;12:921-927.

9. Kirk KM, Cardon LR: The impact of genotyping error on haplotype reconstruction and frequency estimation. Eur J Hum Genet 2002;10:616-622.

10. Zou G, Zhao H: Haplotype frequency estimation in the presence of genotyping errors. Hum Hered 2003;56:131-138.

11. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D: The structure of haplotype blocks in the human genome. Science 2002;296:2225-2229.

12. Mailund T, Schierup MH, Pedersen CN, Mechlenborg PJ, Madsen JN, Schauser L: CoaSim: a flexible environment for simulating genetic data under coalescent models. BMC Bioinformatics 2005;6:252.

13. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA: Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 2002;70:425-434.

14. Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG: Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 2002;53:79-91.

15. Devlin B, Risch N: A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 1995;29:311-322.

16. Taberlet P, Griffin S, Goossens B, Questiau S, Manceau V, Escaravage N, Waits LP, Bouvet J: Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic Acids Res 1996;24:3189-3194.

17. Kalinowski ST, Taper, ML, Creel, S: Using DNA from non-invasive samples to identify individuals and census populations: an evidential approach tolerant of genotyping errors. Conservation Genetics 2006;7:319 - 329.

18. Hacia, J. G., Fan, J. B., Ryder, O., Jin, L., Edgemon, K., Ghandour, G., Mayer, R. A., Sun, B., Hsie, L., Robbins, C. M., Brody, L. C.,Wang, D., Lander, E. S., Lipshutz, R., Fodor, S. P., Collins, F. S. : Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. Nat Genet 1999; 22, 164–167.

19. Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lipshutz, R., Chee, M., Lander, E. S. : Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 1998; 280, 1077–1082.