Towards Unifying Tagging SNP Selection Algorithms

Andy Ly

Thesis Advisor: Sorin Istrail
Second Reader: Charles Lawrence

May 2016

1 Abstract

Single nucleotide polymorphism (SNP) can be described as the variation in a single nucleotide of the genome of a species. These variations may occur between different members of the same species, and are currently studied to determine correlation to diseases. A tagging SNP, on the other hand, is a representative single nucleotide polymorphism that can represent a haplotype, a set of SNPs on a chromatid of a chromosome. As there exist many SNPs in an organism, identifying tagging SNPs reduces the complexity and computation needed. There are algorithms that have been developed to select tagging SNPs from a set of SNPs, but each of their approaches are different.

The algorithms of interest are the LD-Select algorithm, the Informativeness algorithm, and principal component analysis (PCA). While there are similarities between these algorithms, there are also differences. For example, both the application of PCA and the LD-Select algorithms rely on haplotype blocks, while the Informativeness algorithm is "block-free". Another aspect is while the LD-Select and the Informativeness algorithms use linkage disequilibrium (LD), a statistical measure for determining non-random association of SNPs, the algorithm that uses PCA relies on a different statistical measure more inherit to PCA, but is still related to LD.

In this study, the above algorithms will be discussed. These algorithms have different results due to each other's approaches and assumptions. With knowledge from these algorithms we will define a metric for comparing two loci, and a method of using that metric for selecting tagging SNPs.

2 Tagging SNPs

Single nucleotide polymorphisms, or SNPs, are the genetic variations in DNA at a specific site on a genome. For example, let us say there is a site that has the base "A" most of the time. At the same time, there is a small population that has a different base, "T", at the same site. This constitutes a SNP at that site. On average, a SNP occurs every 300 nucleotides, resulting in around 10 million SNPs in the human genome, which is 3.2 billion bases long. SNPs may be unique to an individual or to a population. SNPs may also be "harmless", as multiple sequences of three long nucleotides may code for the same amino acid (degeneracy), resulting in the same protein. In other situations, this may not hold.

SNPs are of interest because an individual may be more susceptible to certain diseases if they have a SNP at a specific site in their genome. This is heavily studied in Genome-Wide Association Studies (GWAS), where SNPs are analyzed for variations among individuals carrying a certain disease of interest and individuals that do not carry such disease, in clinical settings, in hope of determining genetic factors associated to the disease of interest. In GWAS, individuals exhibiting certain phenotypes of interest are analyzed and compared to those that do not, creating a case and control test. While there are pairwise relations between SNPs, it is not proper to run tests for each and every possible pair of SNPs. Multiple comparison is used instead, as the problem is not represented pairwise but as a whole (all SNPs in set) instead. As more and more features (SNPs) are added, the likelihood of a difference occurring among individuals through random change increases.

A tagging SNP is a SNP that is representative of a set of SNPs of a haplotype (sequence of SNPs for an individual). The tagging SNP is in high linkage disequilibrium with the set of SNPs it was selected from. A reason behind this is SNPs may be inherited together, specifically the genes the SNPs are from. Genetic linkage occurs during the phase when gametes are formed from cell separation, and this separation may not be completely random. When conducting a test, one may have many SNPs of interest, so reducing the number of features, or SNPs, is beneficial. Mapping individuals with a disease may be costly, both through time and resources. Because GWAS uses SNP arrays to identify a subset of genetic material and not the whole genome, tagging SNPs are essential.

3 Linkage Disequilibrium

Linkage disequilibrium can be described as the non-random association between two alleles at two different loci. The opposite, where there is random association, is called linkage equilibrium. For linkage disequilibrium, a few measures are used to described this association: D, D', and r^2.

First, let us say A and a correspond to the major allele and minor allele for locus 1, and B and b respectively for locus 2. To define the frequencies (normalized), we can define them with their corresponding haplotype as:

Haplotype	Frequency
AB	p_{AB}
Ab	p_{Ab}
aB	p_{aB}
ab	p_{ab}

In addition, we can define the individual allele frequencies as:

Allele	Frequency
A	p_A
a	p_a
B	p_B
b	p_b

If there is no linkage disequilibrium, we can assume frequencies as independent, specifically $p_{AB} = p_A \cdot p_B$, and all other haplotype and allele frequencies. When there is linkage disequilibrium, these values are not equal, and instead, the difference can be defined as D, which can be defined as the following relation:

$$D = p_{AB} - p_A \cdot p_B = p_{AB} \cdot p_{ab} - p_{Ab} \cdot p_{aB}$$

D can vary, and to normalize this value, D' was created, which can be calculated as:

$$D' = \begin{cases} \dfrac{D}{min(p_A \cdot p_b, p_a \cdot p_B)}, & \text{if } D \geq 0, \\ \dfrac{D}{min(p_A \cdot p_B, p_a \cdot p_b)}, & \text{otherwise } (D < 0) \end{cases}$$

Finally, there is a correlation coefficient that can be calculated and used to describe linkage disequilibrium, which is defined as:

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}$$

4 Power of Association

Pritchard and Przeworski [8] described the relation between linkage disequilibrium (r^2) and the power of association studies. Here we will describe this relation in more detail. The variables used in this description are defined as:

N_1 : number of individuals tested at locus 1 (disease susceptibility locus)

N_2 : number of individuals tested at locus 2 (nearby marker locus)

A, a : alleles at locus 1

B, b : alleles at locus 2

π_{DA} : frequency of allele A in individuals with disease

π_{CA} : frequency of allele A in individuals with control

π_{DB} : frequency of allele B in individuals with disease

π_{CB} : frequency of allele B in individuals with control

π_A : frequency of allele A

π_B : frequency of allele B

q_{AB} : probability with allele A at locus 1 and allele B at locus 2

q_{aB} : probability with allele a at locus 1 and allele B at locus 2

ϕ : fraction of sample that are case (disease)

$1 - \phi$: fraction of sample that are control

*Note, variables defined later with a $\hat{}$ are sample frequencies, while variables with a $\bar{}$ are approximations of the population frequencies.

Let us first consider the Wald (test) statistic:

$$\frac{(\hat{\theta} - \theta_0)^2}{Var(\hat{\theta})}$$

Here, we are interested in seeing if there is a difference between the case and control, so we will rewrite the above equation as:

$$\frac{(\pi_{\hat{DA}} - \pi_{\hat{CA}})^2}{Var(\pi_{\hat{DA}} - \pi_{\hat{CA}})}$$

where $\hat{\theta}$ is $\pi_{\hat{DA}} - \pi_{\hat{CA}}$ and θ_0 is 0. We can determine the variance as:

$$Var(\pi_{\hat{DA}} - \pi_{\hat{CA}}) = \frac{\pi_{\hat{DA}}(1 - \pi_{\hat{DA}})}{2N_{1D}} + \frac{\pi_{\hat{CA}}(1 - \pi_{\hat{CA}})}{2N_{1C}}$$

where N_{1D} and N_{1C} are the number of individuals that contain allele A at locus 1 and are diseased (case) and control respectively, and add up to N_1. Here, we are assuming independence between the case and control, and the variances are determined by using the rule (assuming independence again):

$$Var(X \pm Y) = Var(X) + Var(Y)$$

Also, we are modeling with n number of Bernoulli trials for both case and control.

If Hardy-Weinberg equilibrium holds (null), we can simplify the variance as:

$$\hat{\pi_A}(1-\hat{\pi_A})\Big[\frac{1}{2N_{1D}}+\frac{1}{2N_{1C}}\Big]$$

with $\pi_{\hat{D}A} = \pi_{\hat{C}A}$. A 2 is added here to account for humans having 2 alleles at each loci.

We can rewrite the the N's with $N_{1D} = \phi \times N_1$ and $N_{1C} = (1-\phi) \times N_1$, and merge the two fractions together.

$$\begin{aligned}
\frac{1}{2N_{1D}}+\frac{1}{2N_{1C}} &= \frac{N_{1C}}{2N_{1D}N_{1C}}+\frac{N_{1D}}{2N_{1D}N_{1C}} \\
&= \frac{N_{1D}+N_{1C}}{2N_{1D}N_{1C}} \\
&= \frac{N_1}{2N_{1D}N_{1C}} \\
&= \frac{N_1}{2N_1\phi N_1(1-\phi)} \\
&= \frac{1}{2N_1\phi(1-\phi)}
\end{aligned}$$

Substituting this back in, we can see our test statistic is:

$$\frac{(\pi_{\hat{D}A}-\pi_{\hat{C}A})^2 2N_1\phi(1-\phi)}{\hat{\pi_A}(1-\hat{\pi_A})}$$

This can be compared back to a normal distribution, where the test statistic will simply be the square root of the above. Given k independent, standard normal variables:

$$Z_1, Z_2, ..., Z_k$$

The sum of the squares:

$$\sum_{i=1}^{k}(Z_i)^2$$

follows a chi-squared distribution with a degrees of freedom of k. Using this definition, our test statistic defined earlier follows a chi-squared distribution with a degrees of freedom of 1:

$$\chi_1^2 = \frac{(\pi_{\hat{D}A}-\pi_{\hat{C}A})^2 2N_1\phi(1-\phi)}{\hat{\pi_A}(1-\hat{\pi_A})}$$

The same can be done with allele B at locus 2, where we will define the test statistic as:

$$\chi_2^2 = \frac{(\hat{\pi_{DB}} - \hat{\pi_{CB}})^2 2N_2\phi(1-\phi)}{\hat{\pi_B}(1-\hat{\pi_B})}$$

As both χ_1^2 and χ_2^2 are approximately the squares of normal random variables, we can take the square root of them.

$$(\pi_{DA} - \pi_{CA})\Big[\frac{2N_1\phi(1-\phi)}{\bar{\pi_A}(1-\bar{\pi_A})}\Big]^{\frac{1}{2}}$$

Here, an approximation is used for determining the population with allele A at locus 1:

$$\bar{\pi_A} = \phi\pi_{DA} + (1-\phi)\pi_{CA} \approx \pi_A$$

which can be derived from the Law of Total Probability, considering π_{DA} and π_{CA} as conditional probabilities and summing up the marginal probabilities.

The authors defined a relation between frequencies of one locus and another:

$$\pi_{DB} - \pi_{CB} = (\pi_{DA} - \pi_{CA})(q_{AB} - q_{aB})$$

The same process for χ_1^2 can be used for χ_2^2, but we will substitute the relation just defined:

$$(\pi_{DA} - \pi_{CA})(q_{AB} - q_{aB})\Big[\frac{2N_2\phi(1-\phi)}{\bar{\pi_B}(1-\bar{\pi_B})}\Big]^{\frac{1}{2}}$$

The author also defined a r^2 measure:

$$r^2 = (q_{AB} - q_{aB})^2\pi_A(1-\pi_A)\pi_B^{-1}(1-\pi_B^{-1})$$

One can observe that this r^2 measure follows the distribution of χ_1^2 and χ_2^2 are approximately the same if $N_2 = \frac{N_1}{r^2}$. The relation tells us for one to have same power between both markers, the sample size must be increased by $\frac{1}{r^2}$.

5 Data Representation

Here we will briefly cover how data will be represented and used. With a given set of haplotypes:

Haplotype 1	A	A	C	G	T
Haplotype 2	A	T	C	G	T
Haplotype 3	A	A	C	G	T
Haplotype 4	A	A	G	G	T

we replace all major alleles (common) with a 0 and minor alleles (rare) with a 1:

Haplotype 1	1	1	1	1	1
Haplotype 2	1	0	1	1	1
Haplotype 3	1	1	1	1	1
Haplotype 4	1	1	0	1	1

We apply this binary representation to haplotypes because of the infinite alleles model, which is the assumption that because there is a large number of alleles, any new mutation will occur in a new location. The algorithms discussed in the following sections will use data in this format.

6 Algorithms for Determining Tagging SNPs

Selecting tagging SNPs is important to GWAS, as the number of features (SNPs) are reduced before they are analyzed. Here, we will be describing a few algorithms that select tagging SNPs from a set of haplotypes.

6.1 LD-Select [3]

First, starting with a set of haplotypes, loci where the minor allele frequency (MAF) exceeds a certain threshold are chosen. r^2 is calculated for each loci that have a MAF exceeding the threshold and all other loci. The one locus with a MAF exceeding the threshold that has the maximum number of r^2 connections that exceed another set threshold is chosen, and all loci connected are binned together. This process is repeated for all other loci selected earlier based on their MAF. If a locus does not have any connection to another loci with r^2 exceeding a defined threshold, they are put into a bin by themself. For each locus in a bin, r^2 is calculated with all other locus, and for loci that have r^2 with all other loci exceed a threshold, it will be chosen as the tagging SNP of that bin. Because of this condition, multiple loci in a bin can be defined as a tagging SNP. This is a greedy algorithm that relies on linkage disequilibrium, to create a "Dominating Set".

6.2 Informativeness [4]

This algorithm is centered around information theory, and modeled as a "Set Cover" problem. With a set of haplotypes, at each loci, a sub-graph is created connecting alleles among haplotypes that differ. Then a set of loci are selected so that all possible connections between haplotypes are covered. The optimal set in this case would be the smallest subset of loci.

An algorithm for selecting this optimal set was described, with a complexity of $O(nk2^w)$, and will provide us with the k optimal SNPs.

k-MIS algorithm 1 $O(nk2^w)$

for $s := 1$ to n **do**
  for $l := 1$ to k **do**
    for all assignments A_s **do**
      $A_s^0 \leftarrow 0A_s[1...w-1]$;
      $A_s^1 \leftarrow 1A_s[1...w-1]$;
      $I(s,l,A_s) \leftarrow I(S(A_s),s) + max(I(s-1,l-A_s[w],A_s^0), I(s-1,l-A_s[w],A_s^1))$;

Let us describe this algorithm from above. An informative measure is used to compare two loci, s and t. It is defined as:

$$I(s,t) \simeq \frac{|E(s) \cap E(t)|}{|E(t)|}$$

Where $E(s)$ and $E(t)$ are number of connections at respective loci. This measure is also used for comparing a set of SNPs S' and another SNP t:

$$I(S',t) \simeq \frac{|E(S') \cap E(t)|}{|E(t)|}$$

An assignment is defined as, for SNP s, all SNPs neighboring it that are w SNPs away. This w window is user defined. A_s can be constructed by:

$$A_s[i] = \begin{cases} 1, & \text{if SNP } s - \lfloor \frac{w}{2} \rfloor + i \in S', \\ 0, & \text{otherwise} \end{cases}$$

$I(s,l,A_s)$ is the most informative subset of l SNPs chosen from SNPs 1 to s. This is determined by finding the informativeness measure of A_s and s, and by using dynamic programming, using previous informativeness measures calculated, and maximizing on which potential allele precedes the A_s set.

6.3 Principal Component Analysis [7]

Principal component analysis, or PCA, is a technique that determines linear combinations (principal components) that compresses data via dimension reduction. In general, the steps of PCA can be described as: normalization of data points, covariance matrix calculation, eigenvectors and eigenvalues of such matrix, component selection, and data reduction/mapping. For tagging SNPs, a few methods were created relying on PCA. Here, we will be describing one of them. This process consists of two steps. First, PCA is used to determine principal components, which is referred as eigenSNPs. Haplotypes are placed together in a matrix, where at each loci the major allele is labeled as 0 and the minor allele is labeled as 1, when observed. From here, instead of a covariance matrix, a correlation matrix is generated. This new matrix is defined as R, and linkage disequilibrium is used as the measure, between each SNP. Eigenvectors ($E = e_1, e_2, ...$) and eigenvalues ($\Lambda = \lambda_1, \lambda_2, ...$) are calculated by

solving $Re_j = \lambda_j e_j$. k eigenvectors with the largest eigenvalues are chosen of the set of eigenvectors, and each correspond to an eigenSNP. Each eigenSNP s is a weighted sum of SNPs, and can be defined as:

$$s_i = \sum_{j=1}^{p} e_{ij} x_j, i = 1, ..., k$$

To determine the variance captured by selected SNPs, the following relation can be used:

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

which can be described as the sum of eigenvalues of eigenSNPs selected over the sum of eigenvalues of all SNPs.

Next, these eigenSNPs are mapped, to determine tagging SNPs. Two methods for this step were described. The first is called the "Greedy-discard method", which assumes eigenSNPs with a small eigenvalue are not as important, along with those that have a high correlation to them. Starting with eigenvectors with the smallest eigenvalue to the $(p-k)$th smallest eigenvalue (the ones that were not selected as the k eigenSNPs), the SNP that has the largest coefficient with the $(p-k)$th eigenSNP and has not been previously rejected is rejected. With the remaining k SNPs, they are mapped to k eigenSNPs in reverse order, resulting in k tagging SNPs. The second method described was the "Varimax-rotation method", whose goal is to maximize the "the sum of the variances of the squared coefficients within each eigenvector". The orthogonal transformation T is determined by the relation $E^r = ET = e_1^r, e_2^r, ..., e_p^r$. From here, for all k eigenSNPs, the average coefficient are:

$$\Gamma_i = \frac{1}{k} \sum_{j=1}^{k} |e_{ij}^r|, i = 1, ..., n$$

and for the remaining $(p-k)$ eigenSNPs, the average cofficient are:

$$\gamma_i = \frac{1}{p-k} \sum_{j=k+1}^{p} |e_{ij}^r|, i = 1, ..., n$$

For SNP i, if $\Gamma_i > \gamma_i$, it is chosen.

7 Graphical Representations

In the previous section, two algorithms, LD-Select and Informativeness, were described. Mentioned was what graph theory problems they were modeled after. In this section, the graph theory problems along with their mutual reductions will be described.

7.1 Dominating Set

For a given graph $G(V, E)$ where V are the vertices and E are the edges, the dominating set is defined as a set of vertices D, which is a subset of V, where every vertex not in D is adjacent to at least one vertex in D.

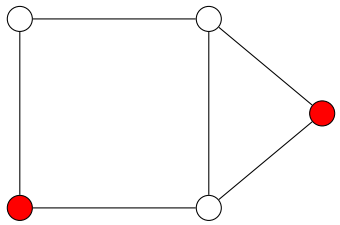


Figure 1: An example dominating set

The LD-Select algorithm is an algorithm that determines a dominating set, with vertices as SNPs and edges as r^2 between SNPs. The solution it returns may not be minimal dominating set, due to the use of a greedy fashion maximizing r^2 and binning with a set r^2 threshold and grouping of SNPs connected to one another with that measure. Through its use of binning, blocks are created as result.

7.2 Set Cover

For a given universe U containing n elements $(1, 2, ..., n)$, and a collection C where each element $S_i$in C is a subset of U, a set cover is defined as a subset of C where all elements in U are captured. The Informativeness algorithm determines a set cover, with the optimal set cover being the minimum subset of C that covers U. We can model this as a bipartite graph, where a clique (for a given set of nodes, there exist an edge between each pair) of loci are on one side and on the other is an independent set (for a given set of nodes, there are no edges connecting them) of connections between haplotypes:

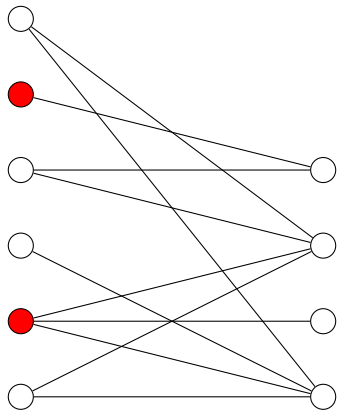


Figure 2: An example set cover on a bipartite graph

7.3 Reductions

Kann [6] describes proofs for the reduction from a minimum set cover to a minimum dominating set, and vice versa. Here, we will briefly cover them.

7.3.1 Minimum Set Cover to Minimum Dominating Set

Let us look at a split graph. On one side, we have an clique, where the set of nodes are defined as X. On the other side we have an independent that is defined as (U, C), where U is the universe of elements and C as the subsets of the elements in U. Let us assume the nodes of U and X are disjoint. We can construct a graph $G(V, E)$ where $V = U \cup I$ and E consist of edges between nodes of the clique, and between the clique and independent set.

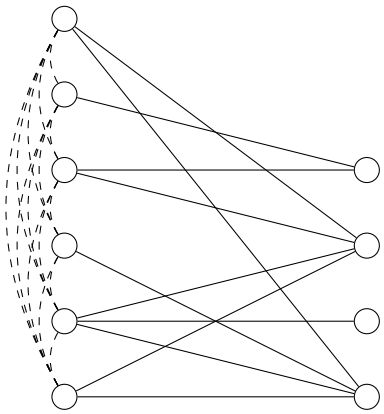


Figure 3: An example split graph

With a dominating set, we can derive a set cover. Given a dominating set $V' \subseteq V$ of G, for every node in $x \in V'$, we can substitute it with any node in the independent set where this node is connected to. Once this is done with all nodes in the dominating set, a set cover is created of the same size.
With a set cover $C' \subseteq C$, we can derive a dominating set $V' \subseteq V$ of G of the same size by including the nodes corresponding to the set cover $i : S_i \in C'$.

7.3.2 Minimum Dominating Set to Minimum Set Cover

Here, we will be using a connected graph where (U, C) is defined with the universe U being all vertices and C as subsets for each vertex and its neighbors. In addition, this graph is defined as $G(V, E)$ where V consists of all vertices (similar to U) and E consist of all edges connecting the vertices of this graph.

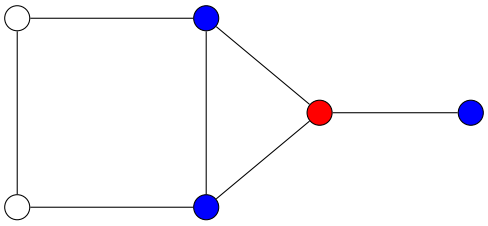


Figure 4: An example graph

Graphically, we can describe it as, with the figure above, for the vertex colored red, a set in C will consist of both the red vertex and the blue vertices.
Say we are given a dominating set D. We can create a set cover from each vertex $v \in D$, and because each subset constructed from vertex v covers all

neighbors, it is a set cover of the same size as its corresponding dominating set. It is possible to generate a dominating set with nodes from both the clique and the independent set. In that case, we can simply map a node from the clique to the independent set node in the dominating set. This will yield a set cover of a size equal to or less than the original dominating set.

7.4 Comments

Both determining a minimum dominating set and a minimum set cover are NP-complete problems. Approximate methods are used instead, and while reductions exist between both problems, finding the optimal set from one to another is also complex.

8 Metric

Linkage disequilibrium relies on allele frequencies of a give sample between two loci of interest. Depending on the samples, this metric may be skewed, as it heavily relies on sample size and allele frequencies being representative of the population. Linkage disequilibrium is also effected by many factors, such as recombination, mutation, and population structure. Here, we will be describing a potential metric that is based on information, similar to the Informativeness algorithm.

Let us define $e_{ki} \in E_i$ as connections between haplotypes at one locus, locus i. These connections are based on differences in alleles, specifically a connection is formed between haplotypes if at the same locus the allele differ. Similar to linkage disequilibrium, we will be considering pairwise relations only. When comparing two loci, we are interested in how much information we "gained" when selecting them pairwise. To define this, we can consider how many unique connections are obtained. If two loci do not yield anything (haplotypes have same allele) or have the same connections, we are not as interested, and those loci should be pooled together. A way to determine how many unique connections we obtain all together, we can consider the number of unique connections over the total number of connections, for two loci, i and j:

$$\frac{|E_i \cup E_j|}{|E_i| + |E_j|}$$

This way, we also take into account the overlap of the information between the two loci. This value ranges from 0.5, where two loci have the same information, and 1, where the two loci contain different information. To normalize this into $[0, 1]$, we can rewrite the metric with a normalization factor:

$$\varsigma_{ij} = 2 \times \Big(\frac{|E_i \cup E_j|}{|E_i| + |E_j|} - 0.5 \Big)$$

Next, we compare this metric with linkage disequilibrium r^2 for equivalent loci pairs.

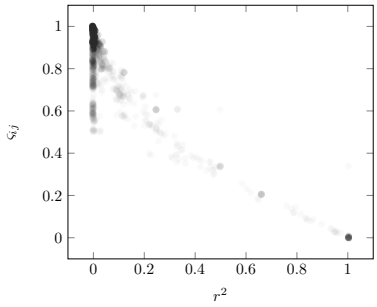


Figure 5: BRCA2 MKK

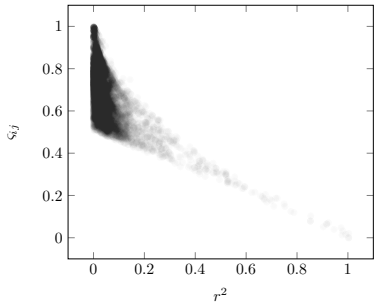
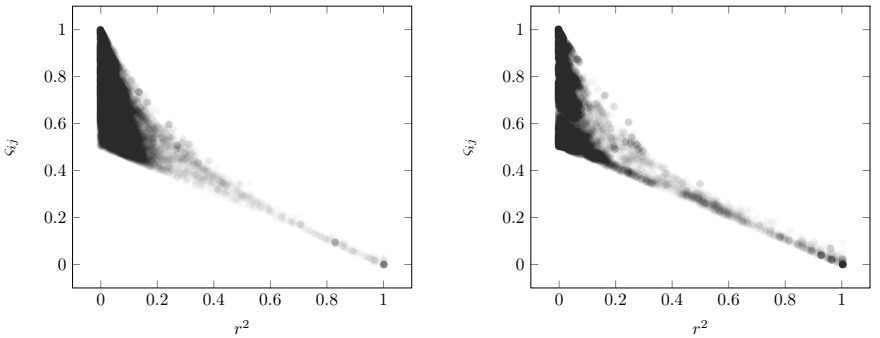
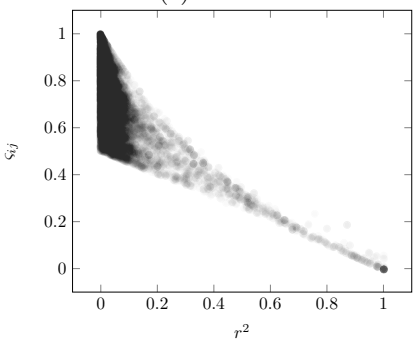


Figure 6: chr9 660k-760k MKK

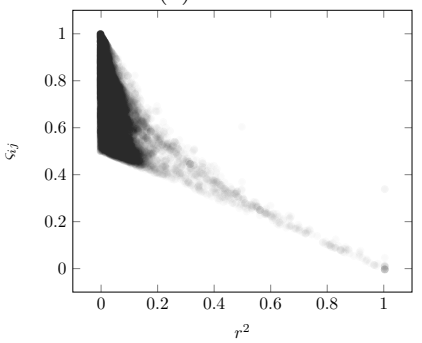


(a) AWS

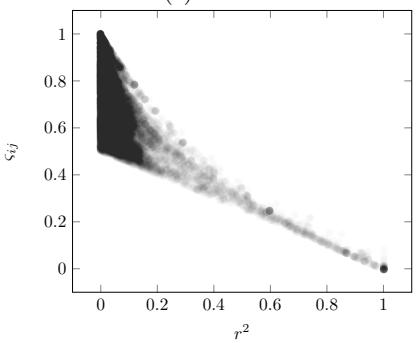
(b) CHD



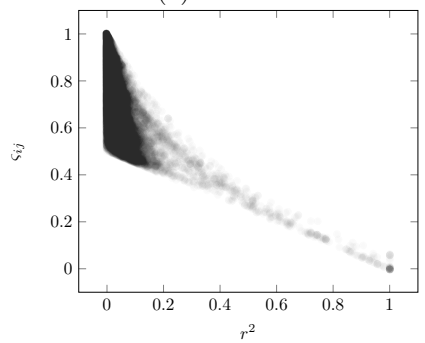
(c) GIH



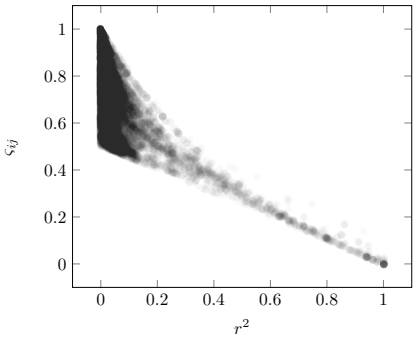
(d) LWK



(e) MEX



(f) MKK



(g) TSI

Figure 7: ENm010

Looking at a few sites and comparing r^2 and this metric, we can see there is some relation. For high r^2, the informativeness is low, while when r^2 is low, the informativeness is high. Also, as r^2 decreases, the informativeness metric varies much more, but ends at 0.5, which tells us there is approximately half overlap in connections between haplotypes among the two loci.

9 Method

Let us now propose a method that relies on the metric just defined. Looking at sample preliminary results of this metric and tagging SNPs selected from a set cover implementation, a certain pattern was found. If one looks at the average value of this metric for one SNP and all other SNPs, if the SNP was selected as a tagging SNP, it tends to have a number of this pairwise metric close to its average.

To describe this process, assume we are working with a split graph. With a split graph, we have a clique and an independent set. As all nodes in the clique are connected to one another, we cannot use the clique as is for selecting a set of nodes, and instead, will use the metric we defined earlier. This process can be group into two parts: selecting nodes based on number of connections with a constraint on the metric, and mapping remaining nodes in the independent set not covered by the set of nodes selected in the first step to other nodes in the clique.

We first start off with x haplotypes of n SNPs long. A preliminary step is we remove SNPs that are "redundant", or those that contain the same information. Now, we calculate the metric between all pairs of SNPs, excluding with itself. With each SNP, we calculate the average of its pairwise metrics with other SNPs, and count the number of pairwise metrics that fall within a range from this average (0.1 was used here). Next, a greedy approach is used to select the SNP with the highest count, or pairwise metrics close to the average for that SNP. If there are multiple, we choose a random one and proceed. This SNP will now be selected as a tagging SNP and removed from the graph. For all other SNPs remaining, connections of the tagging SNP selected are removed. After a certain number of iterations, we will not be able to select a tagging SNP by this manner, due to having less connections remaining, and variability in the metric increasing as a result. We now have some connections we want to cover, and here, we again will greedily select SNPs that will encompass them. The reason behind this approach goes back to the reduction of dominating set to set cover. With a given dominating set of a graph (in this case, we have a split graph, and nodes selected for this dominating set may be of the clique or independent set), we can generate a set cover of the same size, by simply selecting nodes from the clique that correspond with the node in the independent set. Through this selection, there is a possibility of the number of nodes in the set decreasing.

10 Comparison

Using data from ENm010 for MKK population (265 SNPs), LD-Select, Informativeness, and this process were ran. The results are as follows (for binary sequences, at each position, 1 is denoted as SNP being selected as tagging SNP):

LD-select: 150 tagging SNPs
00100000101010000000000011010100001110100101101111111100100111110011
01000101011101110110111110111000101111001110101101011110001100111011
01000110100001110011101010011111111011111110101100101101100101101010
01111011101010101110011101011111100100010111110101011000001 11
Coverage: 0.999985190012

Informativeness: 172 tagging SNPs (with $w = 21$) 01110101111010000001101 0
10000101101111100110100110111111100111010100101001111111111111011011
00111000111011111110101111111010111101111001010101101101000110111110
10001111101001111111111110111101011010111100011011111100111111001001 11
11001111110101111111000000111001001 11
Coverage: 1.0

Informativeness: 134 tagging SNPs (with $w = 41$) 01110101111010001001011 0
00011110000000000010100111101010100010111111101111000000100111000011
10110000010111001111111110000100001000111011110101100001000100111110
01111000000110001110011111001101110101101100001011111000011000110110
11011010100111101111100010011000000 01
Coverage: 0.999985190012

Process (using 0.1 for window): 41 tagging SNPs
00000100010100001001000100000010000010000000110000000000100000010000
00111000000000000000010001010111000000001010000001100000000000000100
00011000001000000001000000000001000000000010000000100000001000000000
00000000100000000000000000000100000100001001000000000001001000
Coverage: 0.430022807381 (21 SNPs with first step)
Coverage: 1.0 (both steps)

For LD-Select, r^2 increases, the number of SNPs decrease, but at a slight cost of coverage. For Informativeness, as w is increased, coverage also decreases.

11 Discussion

Looking back at the algorithms covered here, they all approach determining tagging SNPs in a different manner. One relies on determining SNPs by maximizing a measure and binning. Another relies on dynamic programming and maximizing information captured. Comparing LD-Select, Informativeness, and

the process described here, we noticed some tagging SNPs overlap, telling us the measures used among these algorithms, they tend to favor certain SNPs consistently. With the currently implementation of the process described, as certain optimizations were not made, some computations are intensive, specifically comparing sets. In addition, while there is a way of selecting SNPs in the clique that maximizes connectedness, there is not really a measure for mapping nodes from the independent set back to the clique.

References

[1] What are single nucleotide polymorphisms (snps)? `https://ghr.nlm.nih.gov/primer/genomicresearch/snp`.

[2] Alan A. Bertossi. Dominating sets for split and bipartite graphs. *Information Processing Letters*, 19(1):37 – 40, 1984.

[3] 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.*, 74(1):106–120, Jan 2004.

[4] 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.*, 14(8):1633–1640, Aug 2004.

[5] Randall C. Johnson, George W. Nelson, Jennifer L. Troyer, James A. Lautenberger, Bailey D. Kessing, Cheryl A. Winkler, and Stephen J. O'Brien. Accounting for multiple comparisons in a genome-wide association study (gwas). *BMC Genomics*, 11(1):1–6, 2010.

[6] Viggo Kann. *On the Approximability of NP-complete Optimization Problems*. PhD thesis, Royal Institute of Technology, May 1992.

[7] Z. Lin and R. B. Altman. Finding haplotype tagging SNPs by use of principal components analysis. *Am. J. Hum. Genet.*, 75(5):850–861, Nov 2004.

[8] J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.*, 69(1):1–14, Jul 2001.

[9] D. J. Schaid and S. J. Jacobsen. Biased tests of association: comparisons of allele frequencies when departing from Hardy-Weinberg proportions. *Am. J. Epidemiol.*, 149(8):706–711, Apr 1999.