A Critical Review of Strategies for Selecting Haplotype Tag SNPs

Ruby Lee
June 4, 2011

Introduction

Single nucleotide polymorphisms (SNPs) are sequence variations observed across populations that are found at single points in the genome. Typically, either a major or minor allele (i.e. one of two possible nucleotide bases) is observed at each SNP position. In genomics, SNPs are used in a wide variety of applications, including the prediction of specific traits, the classification of patients with varying drug reactions during clinical trials, and the genome wide association studies of complex diseases [1]. There are currently 3-4 million known SNPs in the human genome, or approximately one in every 1200 base pairs [2].

Haplotypes are regions in the genome containing a series of contiguous SNPs that are co-inherited, because they are close together and recombination does not occur between them during meosis. Thus, contiguous SNPs on a given chromosome can be inherited in blocks of haplotypes, and when there exists a high degree of linkage disequilibrium (i.e. correlation between the SNPs), a subset of SNPs can be selected that captures the full haplotype information. The SNPs in this subset are referred to as haplotype tag SNPs (htSNPs) [3]. The selection of htSNPs eliminates the need to genotype all of the SNPs in a particular region, which is both cost-effective and efficient.

In the past decade, many computational strategies have been developed for the efficient and accurate finding of htSNPs for particular chromosomal regions [3]. Here,

we review three strategies that emerged during the early stages of htSNP research. Many of the current strategies have arisen based on modifications of these models.

Greedy Algorithm

Patil, et. al. [4] reported the implementation of a greedy optimization algorithm in order to define a set of haplotype blocks, spanning chromosome 21. The procedure proceeds as follows:

1. Define a "block" as a set of adjacent base pairs on the same chromosome, consisting of one or more SNP.
2. Include only the blocks with >80% coverage, meaning that some haplotype represented more than once in the block must define at least 80% of the chromosomes in the block.
3. Select the block with the maximum *m*:*n* ratio, where *m* is the total number of SNPs in the block, and *n* is the number of htSNPs in the block, i.e. the minimal number of SNPs needed to differentiate the haplotypes that are represented more than once in the block.
4. Discard all of the blocks that overlap with the selected block.
5. From the remaining blocks, repeat steps 3 and 4 until a contiguous set of non-overlapping blocks has been selected.

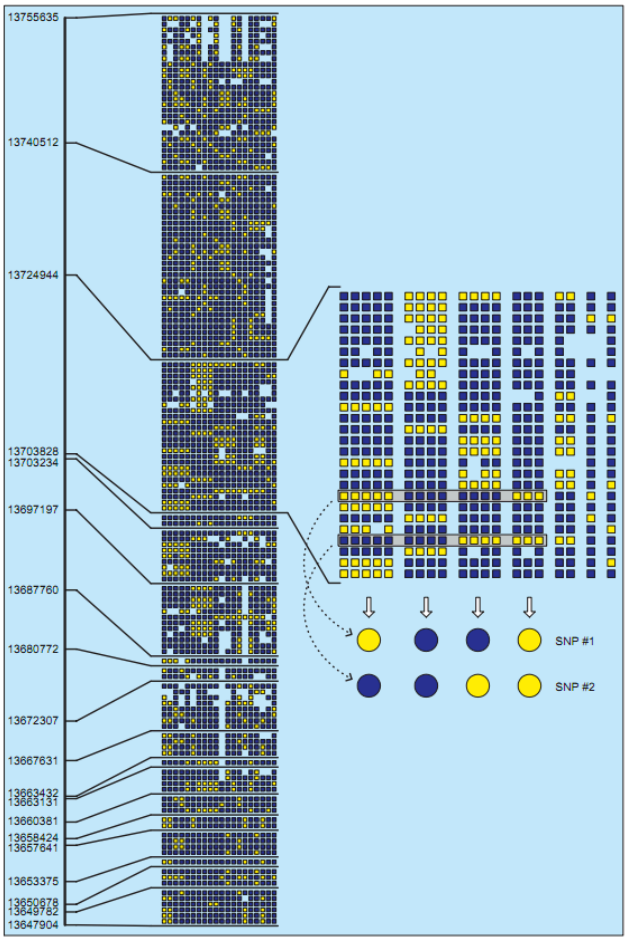

Fig 1 (from Patil, et. al.). A schematic of the greedy algorithm for htSNP selection. Each column represents a human chromosome, and each box represents a SNP (blue boxes represent the major allele, yellow boxes represent the minor allele). The blocks have been demarcated by the black lines pointing to numerical labels. The htSNP selection process for block 13724944 is depicted. The chromosomes are organized by haplotype, and the four most common haplotypes define at least 80% of the chromosomes in the block. The selected htSNPs are represented by circles, and their pairings uniquely differentiate the haplotypes in the block.

By analyzing 20 independent copies of chromosome 21, the authors identified 35,989 SNPs, which was narrowed to 24,047 common SNPs which had a minor allele present two or more times in the set of samples. Using this method, 4135 blocks with an average size of 7.8 kb were identified on chromosome 21. 14% of the blocks had more than 10 SNPs, which accounted for 44% of the length of the chromosome; 52% of the blocks had less than 3 SNPs, which accounted for 20% of the length of the chromosome. When the coverage percentage was increased from 80% to 90%, the method yielded larger numbers of shorter blocks, as expected.

While this procedure is relatively simple and straightforward to implement, it has several disadvantages as a greedy algorithm. It gives an approximate solution, but it cannot guarantee that its solution is optimal [5]. Stage 3 of the algorithm makes a decision in block selection given only the information at hand; it does not consider the effects of its decision on future iterations. In general, greedy algorithms make locally optimal choices, which may or may not always lead to globally optimal choices [6].

Dynamic Programming Algorithm

Zhang, et. al. [5] implemented a dynamic programming algorithm to find a set of representative htSNPs using the same chromosomal sequence data as Patil, et. al. This algorithm recursively finds the minimal number of SNPs required to distinguish the set coverage percentage of haplotypes in each block, for the smallest number of blocks. The algorithm proceeds as follows:

Given K haplotypes and n consecutive SNPs, let r_i be a K-dimensional vector, where i = 1, 2, …, n and $r_i(k)$ = 0,1, or 2 is the allele of the kth haplotype at the ith SNP site (0 indicates missing data). A block is then defined by r_i … r_j.

Haplotypes are *compatible* if the alleles at each SNP site (sites without missing data) are identical. A haplotype is *ambiguous* if it is compatible with two haplotypes that are imcompatible with each other. For example, for h_1 = (1, 1, 0, 2), h_2 = (1, 1, 2, 0), and h_3 = (1, 1, 1, 2), h_1 is compatible with h_2 and h_3, because they share alleles at every point that does not have missing data. However, h_2 is clearly not compatible with h_3, so h_1 is ambiguous. This algorithm treats two unambiguous, compatible haplotypes as identical.

Let S_j be the number of htSNPs used in the optimal blocking of first j SNPs, r_1 … r_j, and set S_0 = 0. Let f(r_i … r_j) be the minimal number of SNPs needed to differentiate the coverage percentage of haplotypes that are represented more than once in the defined block. Let block(r_i … r_j) be a Boolean function that equals 1 if at least the coverage percentage of haplotypes in the block is represented more than once. Then,

$$S_j = \min\{S_{i-1} + f(r_i \ldots r_j), \text{ if } 1 \le i \le j \text{ and block}(r_i \ldots r_j) = 1\}$$

computes the minimum number of htSNPs for the blocking of n SNPs.

Because several blocking schemes may use the same number of htSNPs, another equation can also be used to compute the minimum number of required blocks, C_j. Setting C_0 = 0,

$$C_j = \min\{C_{i-1} + 1, \text{if } 1 \le i \le j \text{ and block}(r_i \ldots r_j) = 1 \text{ and } S_j = S_{i-1} + f(r_i \ldots r_j)\}.$$

By running this algorithm on the data from Patil, et. al. and comparing to their analysis, the number of htSNPs in chromosome 21 was reduced by 21.5% (4563 to 3582) and the number of blocks was reduced by 37.7% (4135 to 2575). When the coverage percentage was increased from 80% to 90%, the method yielded larger numbers of shorter blocks, as expected. This method produces optimal solutions, but due to its recursive nature, may require an overwhelming amount of computational resources to run on long sequences. In contrast to the greedy algorithm, this dynamic programming algorithm solves the problem of assignment by first tackling smaller problems involving subsets of blocking, which guarantees that the solution is optimal [7].

Statistical Method

The previous two methods have focused on minimizing the number of representative SNPs to account for most of the haplotypes in each block. However, Stram, et. al. [8] proposed that the minimum set of htSNPs is not always the optimal one, and instead uses a statistic similar to the coefficient of determination to choose the optimal set of htSNPs. The process proceeds as follows:

1. Identify blocks of high linkage disequilibrium, using the definitions as outlined in Gabriel, et. al. [9]:
    a. A haplotype block is defined as a region over which <5% of comparisons among SNP pairs show strong evidence of historical recombination.

Strong evidence of historical recombination is defined as when the upper confidence bound on allelic association, D', is less than 0.9.

2. For each common haplotype (i.e. those with an estimated frequency of >5%), calculate an estimate of the haplotype dosage, $\delta_h(H)$, or the expected number of copies a haplotype *h* will be contained in a haplotype pair $H_i=(h_1,h_2)$. Note that $\delta_h(H_i) = 0, 1, \text{or } 2$.

$$E\{\delta_h(H_i)|G_i\} = (\Sigma_{H \sim Gi}\, \delta_h(H_i)p_{h1}p_{h2})/(\Sigma_{H \sim Gi}\, p_{h1}p_{h2}),$$

where $E\{\delta_h(H_i)|G_i\}$ is the haplotype dosage for a subject *i* with genotype G_i, $\Sigma_{H \sim Gi}$ is a summation over the haplotype pairs, and p_{h1} is the frequency of the first haplotype.

3. Find the set of htSNPs that maximizes the minimum value of R_h^2 for the common haplotypes. R_h^2 is the squared correlation between the estimate $E\{\delta_h(H_i)|G_i\}$ and the true value $\delta_h(H_i)$.

    a. In other words, R_h^2 is the ratio between the variance of δ_h explained by the genotype data to the total variance of $\delta_h(H_i)$ (see Fig. 2 for details).

$$R_h^2 = (\text{Var}[E\{\delta_h(H_i)|G_i\}])/(2p_h(1-p_h))$$

    Note that the variance is calculated by averaging $E\{\delta_h(H_i)|G_i\}$ over all of the genotypes, weighted by the probability of each genotype.

Genotype, G	(0,0)	(0,1)	(0,2)	(1,0)	(1,1)	(1,2)	(2,0)	(2,1)	(2,2)
Haplotype pair, H	{(0,0),(0,0)}	{(0,0),(0,1)}	{(0,1),(0,1)}	{(1,0),(0,0)}	{(1,0),(0,1)} {(1,1),(0,0)}	{(1,1),(0,1)}	{(1,0),(1,0)}	{(1,1),(1,0)}	{(1,1),(1,1)}
$P(G)$	p_0^2	$2p_0p_1$	p_1^2	$2p_2p_0$	$2(p_2p_1 + p_3p_0)$	$2p_3p_1$	p_2^2	$2p_3p_2$	p_3^2
$E\{\delta_{h0}(H)\|G\}$	2	1	0	1	$\frac{p_3p_0}{p_2p_1 + p_3p_0}$	0	0	0	0
$E\{\delta_{h1}(H)\|G\}$	0	1	2	0	$\frac{p_2p_1}{p_2p_1 + p_3p_0}$	1	0	0	0
$E\{\delta_{h2}(H)\|G\}$	0	0	0	1	$\frac{p_2p_1}{p_2p_1 + p_3p_0}$	0	2	1	0
$E\{\delta_{h3}(H)\|G\}$	0	0	0	0	$\frac{p_3p_0}{p_2p_1 + p_3p_0}$	1	0	1	2

Fig 2 (from Stram, et. al.). Details on the calculation of Var[E{$\delta_h(H_i)|G_i$}] and R_h^2 for two SNPs.

It seems like a potential disadvantage of this method would be the number of calculations necessary to find E{$\delta_h(H_i)|G_i$} when many SNPs are involved. However, this can be mediated by a "divide-and-conquer" process, in which the calculations are broken up into pseudo-blocks of approximately five contiguous SNPs, and only the haplotypes that have non-zero frequency after each set of pseudo-block calculations continue to be considered.

Finding the set of htSNPs that maximizes the minimum value of R_h^2 for the common haplotypes also requires a tedious number of calculations. This process can also be shortened by implementing a modified stepwise procedure. Instead of exhaustively checking every possible set of SNPs, the algorithm first finds the single best SNP (i.e. the SNP that produces the greatest increase in the maximal minimum value of R_h^2, which is calculated from the remaining SNPs). It then looks backward to find the next SNP that will further increase R_h^2, until the desired number of htSNPs is found.

Although using this procedure no longer guarantees that the best set of htSNPs will be found, the authors report that the results remain "very favorable."

Unlike the two previously described algorithms, this method does not compute a value for the number of htSNPs needed to represent the haplotypes in each block. The stepwise procedure can be repeated until the desired number of htSNPs is found. It therefore may be possible to combine the merits of both the dynamic programming algorithm and the statistical method, using the first to compute the minimum number of htSNPs needed for optimal blocking and the second to determine, based on R_h^2, the set of htSNPs.

Conclusion

The papers that initially proposed the greedy algorithm, the dynamic programming algorithm, and the statistical method have been highly cited, by 1042, 312, and 372 sources, respectively (Google Scholar). Many other strategies for determining htSNPs have also emerged following these basic methods, including entropy-based selection [10], usage of a hidden Markov model to define block structure [11], and selection by principle components analysis [3].

Overall, of those analyzed, the greedy algorithm is the easiest to implement. It, however, does not guarantee the output of a globally optimal decision. The dynamic programming algorithm does produce optimal solutions, but because it relies on recursion, it may be much more computationally intensive, especially when running on larger regions. In contrast, instead of focusing purely on minimizing the number htSNPs,

the statistical method uses R_h^2 to minimize the uncertainty in the prediction of common haplotypes based on SNP genotypes. There are a number of measures that can be taken to reduce the number of calculations required by this method, while still preserving the quality of its results. Thus, the statistical method seems favorable in terms of its approach and efficiency. Ultimately, there exist a host of strategies for htSNP searching, and each should be evaluated on the basis of performance, speed, and optimization for the desired application of the algorithm.

References

[1] Brutlag, Douglas. "Simple Nucleotide Polymorphisms (SNPs)." BIOCHEM 218. 19 May 2011.

[2] Snyder, Mike. "Structural Variation in the Human Genome." 24 May 2011.

[3] Lin, Zhen, et. al. "Finding Haplotype Tagging SNPs by Use of Principal Components Analysis." *Am. J. Hum. Genet.* 75: 850-861, 2004.

[4] Patil, Nila, et. al. "Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21." *Science* 294: 1719-1723 (2001).

[5] Zhang, Kui, et. al. "A dynamic programming algorithm for haplotype block partitioning." *PNAS* 99: 7335-7339 (2002).

[6] Cormen, Thomas, et. al. *Introduction to Algorithms*. Cambridge: The MIT Press, 2000.

[7] "Glossary." Nature. 4 June 2011 <http://www.nature.com/nrg/journal/v6/n4/glossary/nrg1576_glossary.html>.

[8] Stram, Daniel O., et. al. “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*. 55 (2003): 27-36.

[9] Gabriel, Stacey B., et. al. “The Structure of Haplotype Blocks in the Human Genome.” *Science* 296 (2002): 2225-2229.

[10] Hampe, Jochen, et. al. “Entropy-based SNP selection for genetic association studies.” *Hum. Genet.* 114 (2003): 36-43.

[11] Daly, Mark J. , et. al. “High-resolution haplotype structure in the human genome.” *Nature Genetics* 29 (2001): 229-232.