Ryan Roper
CSE 527 Final Report
December 15, 2004

APPROPRIATE SELECTION OF TAGGING SNPS IN INDIRECT ASSOCIATION STUDIES

Meiotic Recombination Gives Rise to Genetic Diversity

While the human genome is 99.9% conserved the remaining 0.1% gives rise to a surprising amount of diversity. About 80% of this sequence divergence is in the form of single-nucleotide polymorphisms (SNPs) or single nucleotide substitutions that occur in at least 1% of the human population. The presence of such variants or alleles does not, in itself, provide a high degree of variability in the human population. However, recombination events that occur during gamete formation give rise to an enormous amount of diversity as segments of chromosomes are shuffled and resorted. This occurs before the first division when breaks in aligned chromatids rejoin to their homologous partner as illustrated in Fig.1a. After several generations, many recombination events occurring at different sites can result in a high degree of variability in a population as illustrated in Fig. 2a.

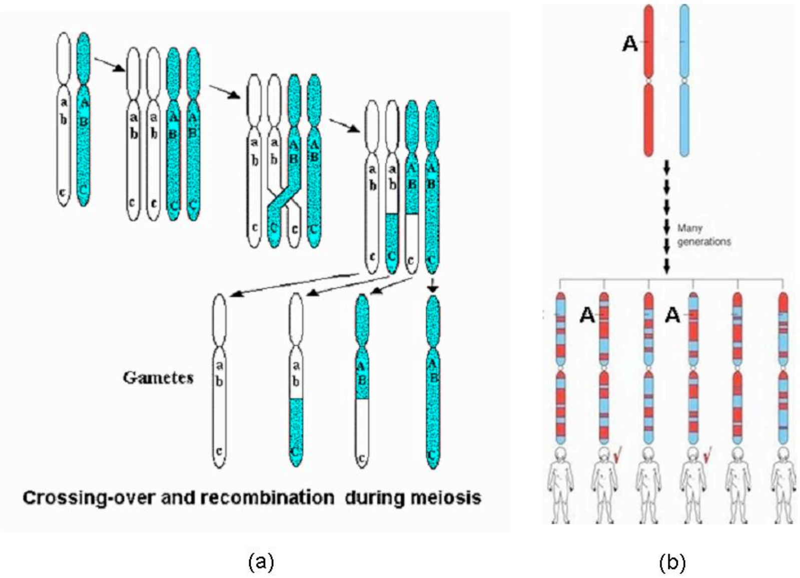

Figure 1. Schematics illustrating (a) recombination by crossing over during meiosis and (b) the formation of haplotypes over many generations

Because the likelihood of a recombination event between two loci is dependent upon the distance separating the loci, alleles are not resorted in an independent manner. This results in chromosomal segments containing SNPs that tend to be transmitted as a group. These are referred to as haplotypes or haplotype blocks.

Exploiting Genetic Diversity to Identify Disease Loci

The diversity generated by meiotic recombination can be exploited to gain insight into genetic determinants of disease. If the occurrence of one variant or the other at a particular locus in a population tends to be correlated with the presence or absence of a particular disease within the population, this provides some evidence that this locus may be a determinant of the disease. With a large enough population, it may be possible to identify a statistically significant correlation between the nucleotide frequency at this locus and the disease. Since many or most diseases are likely to have several genetic determinants, such studies can be elaborated to look at a number of polymorphic loci in addition to this one. This idea forms the basis of association studies in which a large group of individuals is genotyped over a portion or all of their genome to detect statistically significant associations.

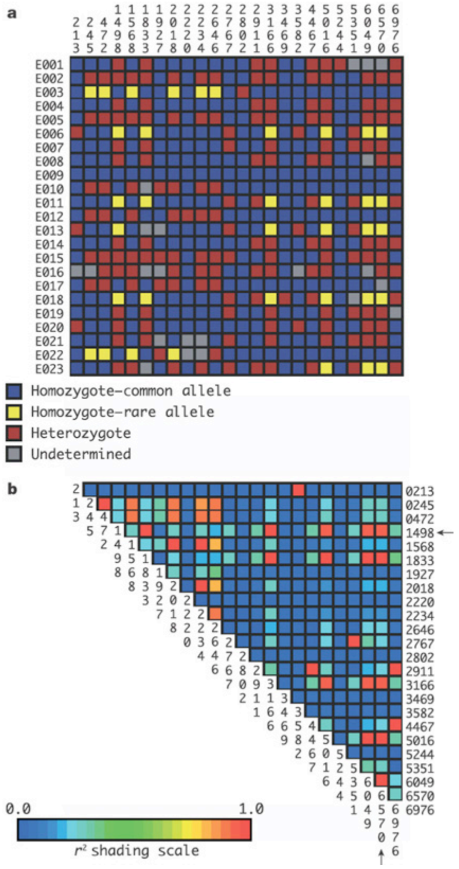

Figure 2. Results of genotyping several SNPs from 23 individuals (top). Linkage disequilibrium (r^2) between loci is indicated in the bottom.

Depending on the amount of information available about a disease, an association study may be either direct or indirect. Direct association studies are hypothesis-driven in that there is some prior knowledge about genes that are suspected to contribute to a particular disease of interest. In such cases, only a relatively small portion of the genome is genotyped and it is often feasible to genotype all SNPs. Thus, all disease loci will be genotyped directly. Indirect association studies, on the other hand, are exploratory or discovery-based in that they are not typically meant to test a hypothesis but rather to identify a set of potential disease-risk sites. These putative disease loci may then be further studied using traditional experimental assays. No prior information about disease-risk loci is necessary, but these studies require genotyping of a much larger portion of the genome. That, combined with the need for a large population size to provide sufficient statistical power, gives rise to practical issues related to the time and cost of genotyping. Indirect studies, therefore, require identification of a set of tagging SNPs (tSNPs) whose frequencies are representative of other nearby SNPs and can serve as surrogate markers for disease loci in the event that the disease-risk SNPs are not directly genotyped. Ideally a minimal set of markers may be selected such that all disease-risk loci are either directly genotyped or are highly correlated with a marker SNP that is genotyped. This task represents a central issue in indirect association studies and is the subject of many studies.

Selection of Tagging SNPs That Uniquely Identify Haplotypes

A tSNP essentially serves as a surrogate marker for the true disease-associated SNP since it exhibits the same properties or occurs in similar frequencies as any disease-risk marker that may be located within the same haplotype. As mentioned previously, a haplotype is a chromosomal segment containing a set of polymorphisms that tend to be transmitted as a group. Such groups of polymorphisms are said to be in linkage disequilibrium (LD) meaning that they are inherited together more often than would be expected by chance. This is usually because their proximity makes the probability of a recombination event within the set of polymorphisms or haplotype relatively low. Two measures of LD are denoted by r^2 and D' and have a value between 0 and 1 where 1 indicates a high degree of LD.

The identification of tSNPs, a procedure known as haplotype tagging, is based on the idea that most of the haplotype structure in a chromosomal region can be captured by genotyping a smaller number of marker SNPs. This is because, the population-wide frequencies or occurrence of the tSNPs are representative of those within the haplotype block. Selection of appropriate tSNPs or markers is essentially an optimization problem where the best markers are those whose frequency of occurrence is most highly correlated (high r^2 or D') to those within the haplotype and least correlated (low r^2 or D') to those contained within the other haplotypes. In other words, the best markers are those that most effectively discriminate between haplotypes. Thus, in indirect association studies, a disease is not actually mapped directly to disease-risk loci unless the genotyped tSNP happens to be a disease-risk SNP. Otherwise the disease is mapped to a haplotype containing the disease-risk locus.

The following sections discuss two approaches to tSNP selection that have been described recently in the literature. The first approach, from Horne and Camp (2004), makes use of a dimension reduction procedure, principle components analysis (PCA), commonly used in analyzing large multivariate data sets. Horne and Camp use a simulated data set to assess the performance of PCA in identifying haplotype structure and selecting appropriate tSNPs. In the second approach, from Carlson et al. (2004a), an algorithm was developed and tested on a set of 100 genes resequenced in a population 47 unrelated individuals. Within the population, 24 are African American and 23 are European American.

PCA in Optimal Selection of tSNPs

Horne and Camp used a two-stage approach in which haplotypes or LD-groups, as they refer to them, are identified in the first stage and tSNPs or group tagging SNPs (gtSNPs), as they refer to them, are identified in the second stage by a separate applications of PCA to each of the LD-groups. In this second stage, SNPs are identified that account for the most variance within the group i.e. those that have the highest loading on the first principle component (PC).

PCA seems well suited for haplotype identification and selection of tSNPs. In PCA, the principle components typically represent orthogonal axes in a high dimensional space that are oriented such that the axis corresponding to the first principle component (that with the highest associated eigenvalue) is aligned along the direction of highest separation of the data (or highest variance). The remaining axes or principle components are aligned along directions of decreasing separation or variance according to their eigenvalues. In this case, data points in this high dimensional space are vectors indicating inheritance patterns of SNPs across genotyped individuals. The factor loading for a given SNP on a given PC provide a measure of the extent to which the SNP contributes to the variance or separation of the data along that PC. In less abstract terms, the loadings provide a measure of how well a particular SNP distinguishes between haplotypes and, therefore, how useful it is as a tSNP. A combination of tSNPs should be selected for each haplotype such that, together, they provide a defining 'signature' for that haplotype that clearly distinguishes it from the rest. To draw a comparison to, say, tumor classification using microarrays, tSNPs would be analogous to genes that show distinct expression profiles in each of the tumors of interest and are, therefore, particularly useful in tumor class discrimination. These hypothetical genes, for example, might be highly up-regulated in one tumor and down-regulated in another.

Haplotype tagging is a concept that is central or fundamental to association analysis. Even though SNPs represent less than 0.1% of the human genome, the enormous number of nucleotides making up our genome results in an estimated 10 million SNPs that would need to be genotyped for a genome-wide association study. Haplotype tagging takes advantage of the fact that alleles within a haplotype are dependent variables. Haplotypes are groups of polymorphisms that tend to be co-inherited because of their proximity to each other and, therefore, because of the low probability of recombination events between them. This effectively reduces the necessary search space because of their redundant nature. In other words, they are not independent variables in much the same way that genes assayed in a microarray experiment are not all independent variables.

It makes sense then that one could use principle components analysis to group SNPs according to haplotype since this is a means of reducing a high dimensional space to a reduced space by taking advantage of the dependence between or the redundant behavior of measured variables. In the case of gene expression this might be co-expressed genes or genes of similar function. In the case of haplotype identification, SNPs within the haplotype tend to be inherited together, thus making them dependent variables. The r^2 value is a measure of this dependence or covariance. The higher this value is for two loci, the higher the likelihood is that the two loci tend to segregate together. As with cluster identification in gene expression analysis, the clusters are not always perfectly clear cut, partly because of the noisy nature of biological data and partly because co-expressed genes or genes of similar function do not necessarily behave in exactly the same way. It makes sense that haplotype identification could be similarly ambiguous. For one reason, close proximity of SNPs, while reducing the probability of their separation by recombination, does not completely eliminate the chance.

Description of an LD-Grouping Algorithm

Carlson et al (2004a) developed an algorithm for assigning SNPs to what are referred to as LD-groups. These groups contain SNPs that tend to have similar inheritance patterns throughout the population that has been genotyped. In this study, the population consists of 47 unrelated individuals, 24 of which are African American (AA) and 23 of which are European American (EA). The algorithm developed by Carlson et al is outlined below:

1) Select from all SNPs exceeding a specified marker allele frequency (MAF)
2) Identify the SNP in linkage disequilibrium (above a specified r^2 threshold) with the most other sites above the specified MAF.
3) Calculate all pairwise r^2 within this group or bin and specify those SNPs that exceed the r^2 threshold with all other sites in the bin as tagSNPs. Only one tagSNP per bin is genotyped.
4) Iterate this process until all SNPs exceeding the MAF threshold are binned. A SNP not exceeding the r^2 threshold with any other SNP is placed alone in a bin.

This algorithm was tested on a set of 100 genes resequenced in these 47 individuals. Results are summarized in Table 1.

LD threshold (r^2)	AA	EA
0.5	5.2 tSNPs per 10 kb (~500,000 genome-wide)	2.6 tSNPs per 10 kb (~250,000 genome-wide)
0.8	8.25 tSNPs per 10 kb (~800,000 genome-wide)	4.2 tSNPs per 10 kb (~400,000 genome-wide)

Table 1. Summary of results obtained by Carlson et al.

From these results it can be seen that LD is sensitive to population stratification. The obvious differences in average tSNP per 10 kb for AA and EA populations is consistent with the idea that there is a greater degree of diversity in the AA population. When selecting an appropriate set of marker loci (tagSNP) it is important to consider stratification in the population. For example, tagSNP obtained by merging the AA and EA populations do not serve as good markers when used for either population individually. In the EA population, LD of 5% of common SNPs did not exceed the threshold r^2 of 0.5 with any tagSNPs. In the AA population, 15% did not exceed this threshold.

REFERENCES

Carlson CS, Eberle MA, Kruglyak L, Nickerson DA. Mapping complex disease loci in whole-genome association studies. *Nature.* 2004 May 27;429(6990):446-52.

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 Jan;74(1):106-20.

Horne BD, Camp NJ. Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. *Genet Epidemiol.* 2004 Jan;26(1):11-21.

Zondervan KT, Cardon LR. The complex interplay among factors that influence allelic association. *Nat Rev Genet.* 2004 Feb;5(2):89-1