Human SNP haplotypes

Statistics 246, Spring 2002
Week 15, Lecture 1
Human single nucleotide polymorphisms

The majority of human sequence variation is due to **substitutions** that have occurred **once** in the history of mankind at **individual base pairs**, SNPs (Patil et al, 2001 listed at the end, and refs therein).

It has been estimated that > **5 million common SNPs**, each with a frequency of 10% - 50% account for the bulk of human DNA sequence difference.

Such SNPs are present in the human genome about **1 in every 600** base pairs.

Alleles making up **blocks** of such SNPs in close physical proximity are often correlated, and define a limited number of **SNP haplotypes**, each of which reflects descent from a single, ancient ancestral chromosome.
The Daly et al (2001) data set

This consists of 103 common SNPs (>5% minor allele frequency) in a 500 kb region implicated in Crohn disease, genotyped in 129 trios (mom, pop, kid) from a European derived population, giving 258 transmitted and 258 untransmitted chromosomes.

Studies to date have revealed great variability in local haplotype structure: the relative contributions of mutation, recombination, selection, population history, and stochastic events seems to vary unpredictably. Some haplotypes extend only a few kb, while others extend for > 100 kb.

Here is some evidence from Figure 1 of Daly et al, 2001. Linkage disequilibrium (LD) between an arbitrary marker (#26 in a, #61 in c, see *) and every other marker in the data set is indicated, using the normalized association measure \( D' = \frac{(ad-bc)}{(a+c)(c+d)} \) of LD. Note the noisiness of the plot.
Daly et al (2001), Figure 1
Measures of association in 2×2 tables

Given positive observed frequencies from a 2×2 table, say a, b, c and d for the cells 11, 10, 01 and 00 respectively, how do we measure association between the two classifications? Put a+b+c+d=n. Geneticists like to use

\[ D = p_{11} - p_{1+}p_{+1} \]

where \( p_{11} = a/n, \) \( p_{1+} = (a+b)/n \) and \( p_{+1} = (a+c)/n. \) One long recognised trouble with this measure is that its values can be greater or smaller, depending on the marginal proportions \( p_{1+} \) and \( p_{+1}. \) Ideally, one would like a measure of association which captured just association, and was parametrically independent of the marginal frequencies. One exists, namely the odds ratio \( \phi = ad/bc, \) equivalently, \( \lambda = \log \phi = \log(ad/bc). \)

This has the nice property that for any specified marginal probabilities \( p_{+1} \) and \( p_{+1} \) between 0 and 1 and any value of \( \lambda, \) there is a unique 2×2 table with these marginals and log odds ratio. Despite this wonderful result, geneticists continue to use a normalized \( D, \) namely, \( D' = D/D_{\text{max}} \) where \( D_{\text{max}} \) is the largest value of \( D \) with the given marginals. If \( D > 0, \) we can show (Exercise!)

\[ D_{\text{max}} = \min \{ p_{1+}(1-p_{+1}), (1-p_{1+})p_{+1} \}. \]

Check that this leads to the formula quoted in the previous slide but one.
Human SNP haplotypes, cont.

If we identify the underlying haplotypes, the LD picture becomes clearer. In Figure 1b, a multi-allelic form of D’ is used to plot LD between the maximum likelihood haplotype group assignment at the location of the 26th marker and that assignment at the location of every other marker in the set. Here the haplotypes have been blocked (details later), and each block treated as an allele. Figure 1d repeats 1b, but with the 61st marker.

Note that when haplotypes rather than single SNPs are used, there is much less noise.

There is a $r \times c$ table analogue of the result cited earlier, involving $(r-1) \times (c-1)$ log odds ratios and $r+s-1$ marginal frequencies, but what geneticists want here is a single number summarizing the association in an $r \times c$ table where $\max(r,c) > 2$. No entirely satisfactory single number exists, though many have been tried and many are in use.

For the multi-allelic form of D’ used above, see Hedrick, *Genetics* 117, 331-341, 1987, “Gametic disequilibrium measures: proceed with caution”.
The block structure of haplotypes

Daly et al (2001) were able to infer offspring haplotypes largely from parents, with a little help from the EM when parents and children were both heterozygous, see last week. They say that “it became evident that the region could be largely decomposed into discrete haplotype blocks, each with a striking lack of diversity (Fig. 2)”.

The haplotype blocks span up to 100kb and contain 5 or more common SNPs. For example, one 84 kb block of 8 SNPs shows just two distinct haplotypes accounting for 95% of the observed chromosomes (Table 1).
A long haplotype block

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGACAACCC</td>
<td>283 (83.2%) haplotype A</td>
</tr>
<tr>
<td>AATTCGGGG</td>
<td>40 (11.8%) haplotype B</td>
</tr>
<tr>
<td>GATTAGGCC</td>
<td>2 (0.6%)</td>
</tr>
<tr>
<td>GGTCAGCC</td>
<td>2 (0.6%)</td>
</tr>
</tbody>
</table>

*Another 13 chromosomes (3.8%) were observed that matched haplotype A or B at all alleles except one, and might represent gene conversion or an undetected genotyping error.
Construction of the haplotype blocks

If I have time I’ll describe Daly’s method of determining haplotype blocks. Basically they define an HMM rather like the one used to map markers on mouse chromosomes (MapMaker) and estimate what they term the “historical recombination frequency $\theta$” between each pair of consecutive SNPs. Their “data” is an assignment of each chromosome to one of four ancestral haplotypes. Consecutive SNPs are then in the same block if $\theta < 1\% \ (73/103)$, with 14 having $1\% < \theta < 4\%$ and 9 with $\theta > 4\%$.

The approach is justified by the observation that the visually defined haplotype blocks have only a few (2-4) haplotypes which show no evidence of being derived from one another by recombination, and which account for nearly all chromosomes ($>90\%$) in the sample. Further, the discrete blocks are separated by intervals in which several independent recombination events seem to have occurred, giving rise to greater haplotype diversity in regions spanning the blocks, see Figure 2.

Finally, we see that the haplotypes at the various blocks can be readily assigned to one of just four ancestral long-range haplotypes.
Daly et al (2001) Figure 2
The data in this paper derives from a publicly available panel of 24 ethnically diverse individuals, and concerns chromosome 21 SNPs. The two chromosomes of each individual were separated using rodent-human somatic cell hybrid technology, and so were able to be typed separately, leading directly to haplotypes. Overall, 20 independent copies of chr 21 were analyzed for SNP discovery and haplotype structure.

The typing was done on specially constructed high-density oligonucleotide arrays (Affymetrix), and in total, they identified 35,989 SNPs in their sample of 20 chromosomes.

The allele frequency distribution is depicted in Figure 1A, see next page. The 32 Mbp of chr 21 DNA was then divided into 200 kb segments, and the observed heterozygosity was used to calculate an average nucleotide diversity for each segment, and these are plotted in Figure 1B. Finally, Fig 1C shows the distribution of distances between consecutive SNPs.
Figure 1 of Patil et al (2001)
SNP block structure in chromosome 21

What do we mean in this context by a haplotype block? Informally, a block is a set of $s$ consecutive SNPs, which, although in theory could generate as many as $2^s$ different haplotypes, in fact shows markedly fewer in our sample of $n$, perhaps as few as $s+1$. In this case, there will be a subset of SNPs in the block whose alleles in our sample essentially determine those of the remaining SNPs in the block. These have been called haplotype tags. Finally, we’d like the set of SNPs constituting a block to be maximal with respect to this property, i.e., if we enlarge it, lose some of its economy.

Formally defining blocks is a mathematical exercise. How many there are, and where their boundaries should go, is a question whose answer largely depends on the criterion to be optimized, that is, by how and to what extent do we wish to trade off the diversity permitted in a block’s haplotypes against the number of haplotype tags, both locally and globally.

Before turning to mathematics, let’s look at part of the blocking defined by Patil et al, 2001.
The **haplotype patterns** for 20 independent globally diverse chromosomes defined by 147 common human chr 21 SNPs spanning 106 kb of genomic sequence.

Each **row** represents an **SNP**. **Blue box** = major, **yellow** = minor allele. Each **column** represents a single **chromosome**.

The 147 SNPs are divided into 18 **blocks** defined by **black** lines.

The **expanded box on the right** is an SNP block of 26 SNPs over 19kb of genomic DNA. The 4 most common of 7 different haplotypes include 80% of the chromosomes, and can be distinguished with 2 SNPs.

Figure 2 of Patil et al
SNP block structure in chromosome 21

How do we define contiguous blocks of SNPs spanning the 32.4 Mb of chr 21, while minimizing the number of SNPs required to define a haplotype? Here is the greedy algorithm of Patil et al.

Begin by considering all possible blocks of ≥1 consecutive SNPs. Next, exclude all blocks in which < 80% of the chromosomes in the data are defined by haplotypes represented more than once in the block (80% coverage). [Ambiguous haplotypes are treated as missing data and not included when calculating % coverage.]

Considering the remaining overlapping blocks simultaneously, select the one which maximizes the ratio of total SNPs in the block to the number required to uniquely discriminate haplotypes represented more than once in the block. Any of the remaining blocks that physically overlap with the selected block are discarded, and the process repeated until we have selected a set of contiguous, non-overlapping blocks that cover the 32.4 Mb of chr 21 with no gaps and with every SNP assigned to a block.
Results of the Patil et al’ algorithm

Using the algorithm just described, their data set of 24,047 common SNPs on a sample of 20 chromosomes, yielded 4,135 blocks of SNPs.

A total of 589 blocks (14% of the total) contain >10 SNPs/block, and comprise 44% of the total 32.4 Mb. In contrast, 2,138 blocks (52% of the total) contain <3 SNPs/block, and make up only 20% of the physical length of the chromosome.

The largest block contains 114 common SNPs and spans 115 kb of DNA. Average block length is 7.8 kb. Also, on average there are 2.7 common haplotypes per block, common here meaning represented on multiple chromosomes.
Patil et al (2001), completed

One extra thing these authors did was determine subsets of the 24,047 common SNPs to capture any desired fraction of the common haplotype information. Common haplotype information is defined as complete information for haplotypes that are present more than once, and include more than 80% of the sample across the entire 32.4 Mb.

Example result: a minimum of 4,563 SNPs are required to capture all the common haplotype information, but only 2,793 SNPs are required to capture the common haplotype information in blocks containing 3 or more SNPs, which cover 81% of the 32.4Mb. A total of 1,794 SNPs are required to capture all the common haplotype information in genic DNA, representing 220 distinct genes.
Number of SNPs required to capture common haplotype information
Haplotype tagging for the identification of common disease genes

This is the paper Johnson et al (2001).
References


