1 Introduction

This lectures concerns the phenomenon of nonuniform "recombination structure" in the human genome. The phenomenon stems from the observation that LD (over relatively short genomic regions) behaves differently in different populations (e.g., Caucasian and African origin) and in different locations of the genome. The notion of a *haplotype block* provides a convenient technical vechile to investigate the recombination structure by computational means based on SNP genotype data. Within each haplotype block, the haplotype diversity is limited and can often be captured by a relatively small subset of the SNPs within the block, called *tag SNPs*. First, we will study a few ways to formulate mathematically the notion of a haplotype block. Then, a couple of ways to define tag SNPs are considered. Finally, we will formulate the haplotype discovery problem with respect to these definitions and give a dynamic programming algorithm that finds an optimal haplotype block decomposition.

Generic methodology to be learned in this lecture (by examples): concept formulation; dynamic programming.

2 The phenomenon of nonuniform recombination structure

In general, the linkage disequilibrium (LD) between two loci decreases with the number of generations. In terms of the linkage disequilibrium coefficient, $D = p_{AB} - p_A p_B$, we have

$$D_t = (1 - c)^t D_0 \,,$$

where the subscript D_t refers to D in the population after t generations, and $c \in [0, 0.5]$ is the genetic distance between the two loci, that is, the probability that in one meiosis an odd number of cross-overs will occur between the the loci. (See, e.g., Wikipedia on LD for a proof of the equation.)

Thus, if a haplotype appears with some frequency in the population at time $t = 0$, then after some generations, the haplotype is likely to have a lower frequency, since parts of it have been recombined with parts of some other haplotypes. In this way the haplotype diversity grows as t grows.

However, there are also forces that affect in the opposite way. *Random drift* makes some haplotypes (as well as individual alleles) become common just by chance. *Population bottlenecks* do effectively the same thing: a relatively few haplotypes (out of many) are chosen as the founders of a new population.

These biological processes are sufficient for explaining the observation that LD decays much more rapidly in African populations that in Caucasian populations (due to the out-of-Africa bottleneck). In other words, conserved haplotypes (i.e., haplotypes with a substantial frequency, say, > 5 %) are on average considerably shorter in African populations than in Caucasian populations.

But, they are not quite sufficient for explaining another obervation: the genetic distance as a function of the physical distance varies a lot in different regions of the genome. Indeed,

it has been observed that there are so-called *recombination hotspots*, that is, short regions that are particularly prone to cross-over events; in between the hotspots there are "colder" regions within which recombinations are much rarer.

3 The motivation of haplotype blocks and tag SNPs

The nonuniform recombination structure has implications to genetic mapping of diseases and other traits of interest. Due to cold regions and conserved haplotypes, it may suffice to genotype only a fraction of the known SNPs to capture the haplotype(s) of an individual (with ot without the trait of interest). In African populations a larger number SNPs is probably needed than in Caucasian populations. Because genotyping a SNP is expensive, it is desirable to find a small informative set of SNPs for a given population. This calls for a proper mathematical formulation of the task and computational means for finding such SNPs.

A haplotype block is an idealization of the fact that the haplotype diversity is relatively low within a cold region between two recombination hotspots. Generally, a haplotype block is defined as a region of the genome where the haplotype diversity is low. Accordingly, a haplotype block has fixed endpoints and is relative to a fixed population. Various definitions for low haplotype diversity have been proposed, some of which we consider in the next section.

Because within a haplotype the haplotype diversity is low, there may be a relatively small set of *tag SNPs* that are sufficient for "predicting" or "capturing" the haplotypes. Then it is sufficient to genotype only these tag SNPs, as the other SNPs within the block do not provide any (or, much) further information. Again, various definitions for predicting/capturing have been proposed, some of which we consider in a later section below.

4 Definitions of haplotype blocks

We consider three different definitions of haplotype blocks.

4.1 The coverage fo common haplotypes (Patil et al., Science 2001)

A sample of haplotypes over a region is called a haplotype block if a fraction α or more of the not-necessarily distinct haplotypes occur at least two times in the sample. In other words, the fraction of haplotypes that occur just once is at most $1 - \alpha$. Patil et al. (2001) used the value $\alpha = 0.80$.

The definition is extended to haplotype data that contain missing alleles, as follows. First, we say that two haplotypes are *compatible* with each other if at every locus their alleles *match*, that is, the two alleles are the same or at least one of the alleles is missing. Second, a haplotype is called *ambiguous* (w.r.t. the sample) if it is compatible with two other haplotypes that are themselves incompatible. For example, 1??0 is compatible with both 110? and 1110, which are, however, mutually incompatible; here "?" denotes missing data, that is, an unobserved allele. Finally, the definition in the previous paragraph is applied to the haplotype sample that is obtained by removing all ambiguous haplotypes and by

treating mutually compatible haplotypes identical.

4.2 No historical recombination (Wang et al., AJHG 2002)

This is one of the simplest possible definitions of haplotype blocks. A sample of haplotypes over a region is called a haplotype block if for every pair of SNPs the sample contains at most three of the four possible haplotypes, AB, Ab, aB, ab; this test is often referred to as the *four gamete test*. The rationale behind this definition stems from the fact that, in the absence of recurrent and/or backward mutation, the only explanation for observing all four gametes between a pair of loci is the occurrence of at least one historical recombination event.

4.3 LD-based blocks (Gabriel et al., Science 2002)

This amazingly complex definition is concerned with pairwise LD between the markers. For a pair of markers with p and q the frequencies of the dominating alleles, A and B, at the two loci, respectively, LD is measured by

$$D' = \begin{cases} \frac{D}{\min\{p(1-q),(1-p)q\}} & \text{if } D \geq 0, \\ \frac{-D}{\min\{pq,(1-p)(1-q)\}} & \text{if } D \leq 0; \end{cases}$$

recall that D denotes the disequilibrium coefficient. Because D' is sensitive to small sample sizes, LD is actually measured not by the point estimate of D' but from the estimates confidence interval, obtained (approximatively) by a common method known as bootstrapping. A quote from Gabriel et al. (2002):

> We define pairs to be in ??strong LD?? if the one-sided upper 95% confidence bound on D' is 0.98 (that is, consistent with no historical recombination) and the lower bound is above 0.7. Conversely, we term ??strong evidence for historical recombination?? pairs for which the upper confidence bound on D' is less than 0.9. On average, 87% of all pairs of markers with minor allele frequency 0.2 fell into one of these two categories (and were thus termed ??informative?? marker pairs). This method should be robust to study-specific differences in the frequencies of SNPs and sample sizes examined, because it relies on those pairs for which narrow confidence intervals (that is, precise estimates) have been obtained.

We call their "strong evidence for historical recombination" *weak LD* to contrast with *strong LD*.

Now, a sample of haplotypes over a region is called a haplotype block if the outermost pair of markers is in strong LD and if, for all pairwise comparisons in the block, the number of pairs in strong LD is at least 19-fold greater than the number of pairs in weak LD.

5 Definitions of tag SNPs

We consider two different definitions of tag SNPs.

5.1 Common haplotypes (Patil et al., Science 2001)

For a sample of haplotypes over a region, a set of SNPs is called a tagging set if the SNPs uniquely distinguish a faction β of all the (not necessarily dintinct) haplotypes in the sample, and the size of the set is the smallest possible. A suitable value for β could be, for instance, 0.80.

Here, we say a set $S \subseteq \{1, 2, \ldots, m\}$ uniquely distinguishes a haplotype h in H if

$$h' \in H \text{ and } h_S = h'_S \text{ imply } h = h',$$

where h_S denotes the sequence $h_{s_1} h_{s_2} \cdots h_{s_k}$ for $S = \{s_1, s_2, \ldots, s_k\}$ and $s_1 < s_2 < \cdots < s_k$. Thus, a tagging set is a minimum-size set S for which there exist at least $\beta|H|$ haplotypes H that are uniquely distinguished by S; here H is understood as a multiset and $|H|$ the total number of haplotypes, not necessarily dintinct.

5.2 Best worst-case pairwise correlation (Carlson et al., AJHG 2003)

This definition is concerned with pairwise correlations of SNPs between tag SNPs and other SNPs in the region, motivated by association studies for gene mapping. The pairwise correlation of two SNPs, s and t, is measured by the square of the correlation coefficient, Δ^2_{st}. For a sample of haplotypes over a region, a set of SNPs S is called a tagging set if its "prediction power", defined as

$$\min_{s \in \bar{S}} \max_{t \in S} \Delta^2_{st},$$

exceeds a given threshold θ, and the size $|S|$ is as small as possible; here $\bar{S}$ denotes the complement set of S with respect to the markers in the region. In words, the prediction power for an individual SNP s absent in the set S is taken as the maximum of the Δ^2_{st} over the SNPs t present in the set S, and the total prediction power is taken as the minimum of these individual prediction powers over the SNPs absent in the set S.

6 Generic problem formulations: few tag SNPs or large coverage?

Consider a sample of n haplotypes over m markers (SNPs). For any marker interval $[s, t]$ with $1 \leq s \leq t \leq m$, we assume the following functions have been specified:

Block indicator: the function $B(s, t)$ takes value 1 if the haplotypes projected onto the marker interval form a haplotype block, and value 0 otherwise. If $B(s, t) = 1$, then the interval $[s, t]$ is called a *block*.

Number of tag SNPs: the function $K(s, t)$ is the size of a (minimum-size) tagging set for the haplotypes projected onto the marker interval.

Length of interval: the function $L(s, t)$ is the length of the marker interval, measured, e.g., by the number of markers $t - s + 1$ or by the physical length of the genome region spanned by the interval.

Any of the definitions given in the previous sections for haplotype blocks and tagging sets can be used here to settle the definition of the functions B and K.

For a set of marker intervals $I = \{I_1, I_2, \ldots, I_\ell\}$ we extend the functions K and L by

$$K(I) = K(I_1) + K(I_2) + \cdots + K(I_\ell)\,, \quad L(I) = L(I_1) + L(I_2) + \cdots + L(I_\ell)\,.$$

In what follows, we will mainly consider sets I that consist of disjoint blocks.

Armed with these definitions, we are ready to two formulations of the problem of finding in some sense optimal haplotype blocks (and tag SNPs) for a given sample of long haplotypes over a genome region. In the first formulation the total number of tag SNPs is bounded and the task is to cover the maximum length of the region.

Definition 1. *The* blocks with a fixed number fo tag SNPs (BFNT) *problem is as follows. Given n haplotypes over m markers, and an integer k, find a set of disjoint blocks I with $K(I) \leq k$ such that $L(I)$ is maximized.*

In the second, a sort of dual formulation, the length of the region to be covered by blocks is fixed and the task is to find a block decomposition with a minimum number of tag SNPs

Definition 2. *The* blocks with a fixed genome coverage (BFGC) *problem is as follows. Given n haplotypes over m markers, and a number $\beta \leq 1$, find a set of disjoint blocks I with $L(I) \geq \beta L(1, m)$ such that $K(I)$ is minimized.*

In the next section we present a dynamic programming algorithm for the BFNT problem. For the BFGC problem an interesting, parametric dynamic programming algorithm has been given by Zhang et al. (AJHG 2003); however, that algorithm is more involved, and its description is beyond the scope of this course.

7 A dynamic programming algorithm for BFNT

Let $f(t, j)$ be the maximum length of the genome that is covered by some disjoint set of blocks I on the first t SNPs with at most j tag SNPs, that is, each block of I is contained in $[1, t]$, $K(I) \leq j$, and $f(t, j) = L(I)$ is the largest possible; we let $f(0, j) = 0$ for any $j \geq 0$ and $f(0, j) = -\infty$ for any $j < 0$. Now, we observe the recurrence

$$f(t, j) = \max\Big\{ f(t-1, j), \max\big\{ f\big(s-1, j-K(s,t)\big) + L(s,t) : 1 \leq s \leq t \text{ and } B(s,t) = 1 \big\} \Big\}\,.$$

Indeed, suppose I is a set of disjoint block on $[1, t]$ such that $L(I)$ is maximized subject to the constraint $K(I) \leq j$. Then either the last block I_ℓ in I ends before t, implying $f(t, j) = f(t-1, j)$, or I_ℓ ends exactly at t and starts at some s, with $1 \leq s \leq t$, implying $f(t, j) = f\big(s-1, j-K(I_\ell)\big) + L(I_\ell) = f\big(t-1, j-K(s,t)\big) + L(s,t)$.

Using the recurrence the values of $f(t, j)$ can be evaluated for all $1 \leq t \leq m$ and $1 \leq j \leq k$ by dynamic programming in $O(mkM)$, assuming that M is number of SNPs in the largest block and that the functions B, K, and L have been precomputed. Note that while M can in theory be as large as m, in practice, m is usually much smaller than m; e.g., $m = 1000$ and $M = 40$. Once the entire table of values $f(t, j)$ has been computed, an optimal set of disjoint blocks can be constructed by backtracking the entries that contribute to $f(t, j)$ (see the exercises).

It remains to discuss the computation of the functions B, K, and L. Of these, L is expected to be fast to evaluate, particularly if the length is measured simply in the number of markers. Evaluating B amounts to determining whether the haplotypes on a given interval form a haplotype block according to one of the many possible definitions; any of the three definitions presented above admits an algorithm that runs in time polynomial in the number of haplotypes n and the number of markers in the interval. The most demanding case is the function K, for it basically requires a search through all possible subsets of the markers in the interval. For example, Zhang et al. (PNAS 2002) have shown that finding a tagging set according to the Patil et al. definition is an NP-complete problem, meaning it is unlikely that a polynomial time algorithm exists for the problem. In practice, however, usually a small number, say five, tag SNPs suffice for a haplotype block, and thus a brute force enumeration of marker subsets upt to size is computationally feasible.

Exercises

IV:1 Does the following sample form a haplotype block under (a) the Patil et al. definition (b) the Wang et al. definition (four gamete test)?

```
11100
01011
00?11
1?100
01011
10101
1010?
010?1
?1001
11001
01011
```

IV:2 Suppose the missing data in the above exercise (IV:1) are replaced by 1s. (a) Which haplotypes are uniquely distinguished by the first two SNPs? (b) Find a tagging set for the haplotypes according to the Patil et al. definition with $\beta = 0.80$.

IV:3 Suppose the values $f(t, j)$ have been computed and stored into a respective array for all $1 \leq t \leq m$ and $1 \leq j \leq k$. Present a pseudo code for an algorithm that, given such an array as input, constructs and outputs an optimal disjoint set of blocks, that is, an I such that $K(I) \leq k$ and $L(I) = f(m, k)$. What is the asymptotic running time of your algorithm in terms of m, k, and the maximum block length M?