Measures of LD

Assume we have two binary loci, one denoted $X$ with genotypes $a, A$ and $Y$ with $b, B$. Assume we are either considering haploid organisms, or more likely, looking at each copy of the genome (so one diploid organism is two samples). We can describe the joint distribution of the two loci via a $2 \times 2$ table:

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>b</th>
<th>B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td></td>
<td>$p_{ab}$</td>
<td>$p_{aB}$</td>
<td>$p_a$</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td>$p_{Ab}$</td>
<td>$p_{AB}$</td>
<td>$p_A$</td>
</tr>
<tr>
<td>Total</td>
<td>$p_b$</td>
<td>$p_B$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(We can add hats and write $\hat{p}_{ab}, \hat{p}_{aB}, \ldots$ to differentiate observed data distributions from theoretical distributions).

We are interested in understanding whether the sites $X, Y$ are “associated” by LD and how much. Intuitively this means that by knowing $X$ we have information on $Y$.

A simple measure: **Lewontin’s $D$:** $D = p_{ab} - p_ap_b = -(p_{aB} - p_ap_B) = \ldots = \text{Cov}(X, Y)$.

**Example:** MRCA is AB, mutation $A \rightarrow a$, followed by $B \rightarrow b$ giving:

<table>
<thead>
<tr>
<th>$X$ \ $Y$</th>
<th>b</th>
<th>B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.3</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Total</td>
<td>0.3</td>
<td>0.7</td>
<td></td>
</tr>
</tbody>
</table>

For this table $D = 0.3 - 0.15 = 0.15$. However this tree has gone through no recombination!

An alternative measure which respects the phylogenetic order is $D'$ which is $D$, normalized to the range $-1 \leq D \leq 1$ given then marginal distributions of $X, Y$:

$$D' = \frac{D}{m(p_a, p_b, \text{sign}(D))}, \quad m = \begin{cases} 
\min(p_a, p_b) - p_ap_b & \text{if } D > 0 \\
 p_ap_b - \max(p_a + p_b - 1, 0) & \text{if } D < 0
\end{cases}.$$
For the example above we would get \( m = 0.3 - 0.15 = 0.15 \), so not surprisingly \( D' = 1 \).

Claim: For a pair of loci with no recombinations, \( D' = 1 \).

The problem with \( D' \) (to some extent also \( D \)): Not really clear how the values relate directly to the “amount of information \( X \) carries on \( Y \).”

Squared correlation / variance explained \( r^2 \):

\[
r^2 = \text{cor}^2(X,Y) = \frac{D^2}{P_a P_a P_b P_B}.
\]

Recall the interpretation from regression as the “variance explained” by regressing \( Y \) on \( X \) or \( X \) on \( Y \).

For the example above: \( r^2 = \frac{0.15^2}{0.21 \times 0.25} = 0.42 \).

\( r^2 \) and \( D \) combine information on:

1. Whether there is recombinations breaking the correlation
2. The “phylogenetic context”, i.e., whether the mutations happened in a similar place in the tree

\( r^2 \approx 1 \) means that both conditions hold – few or no violations of the tree, and similar phylogenetic context.

Important Note: \( r^2 \) and \( D \) are not monotone decreasing as \( X, Y \) move further away along the genome — recombinations are increasing for sure, but far away mutations can still have similar phylogenetic context!

Conclusion: If \( X \) is causative for some disease, and \( r^2(X,Y) \) is big, then \( Y \) is likely to also be associated with the disease only due to this correlation. This should be taken into account:

- What happens if we did not measure \( X \) at all, only \( Y \)?
- What should we conclude if we see many associated loci close together: are there independent associations, or is it all due to one association and LD? How can we use \( r^2 \) values to distinguish?

Expectation-Maximization (EM) to estimate stratification by ancestry

(This section is primarily based on the paper [Estimation of Individual Admixture by Tang et al., Genetic Epidemiology, (2005)].)

For this section we will assume that we have:

- \( I \) individuals from \( K \) different ethnic origins. For simplicity we assume \( K = 2 \), mixture of European (Eu) and African (Af) ancestry, as in African-Americans. We assume in the \( I \) we have:
- $I_0$ of mixed ancestry (unknown mixture proportions)
- $I_1 = I - I_0$ of known ancestry (typically 100% from one ancestry), $I_1 = 0$ is possible

- On each individual we observe $M$ genetic markers ($\times 2$ for two chromosomes), which may have a different distribution in Eu and Af, and therefore carry information on ancestry
- Each marker $m$ has $L_m$ possible values. For SNPs usually $L_m = 2$, but the markers can also be other elements like STRs with $L_m > 2$.

Notations:
- $G = \{G_{ima}\}$ – Value of the $m$ marker in the $i$ individual, copy $a \in \{1, 2\}$. This is a random variable.
- $P = \{P_{mlk}\}$ – Proportion of value $l$ for marker $m$ in population $k$. For example, if SNP $j$ is always $A$ in African and has 50% $A$ in Europe, then $P_{j,A,Af} = 1$, $P_{j,A,Eu} = 0.5$. These are unknown parameters.
- $Q = \{Q_{ik}\}$ – Proportion of ancestry $k$ in individual $i$. For $i > I_0$ this is a known binary vector $Q_{ik} \in \{0, 1\}$, while for $i \leq I_0$ this is an unknown parameter vector on the simplex.

Assumptions:
- $G_{im1a1}, G_{im2a2}$ are independent $\forall m_1, m_2, a_1, a_2$. This entails two assumptions:
  1. No LD between the markers $m_1, m_2$. This may not be very problematic if the $M$ markers were samples for the sole purpose of estimating ancestry, so there are not too many of them and they are far apart.
  2. The two chromosomes of the same individual are independent, so for $m_1 = m_2$ the two copies are still independent. This is known as the Hardy-Weinberg Equilibrium (HWE) assumption, and is violated for example by marriages between relatives.

The resulting log-likelihood function:

$$\ell(P, Q; G) = \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{a=1}^{2} \sum_{l=1}^{L_m} \mathbb{1}\{G_{ima} = l\} \log \left( \sum_{k=1}^{K} P_{mlk} Q_{ik} \right),$$

where the last term is the probability of the value $l$ in the $m$ for person $i$, summarized over her ancestry distribution.

The paper describes several interesting solutions for this estimation problem, we will focus on one that uses a well known approach we can refresh and use: Expectation-Maximization (EM).

**Reminder: EM algorithm**

Assume we have a parameter vector $\Theta$, some observed data $X$ and some unobserved data $Y$. We want to calculate the MLE of $\Theta$ given the observed data $X$, however the calculations are much easier if we had known $Y$ as well, that is calculating $\ell(\Theta; X, Y)$ is easier than directly $\ell(\Theta; X)$. 

3
Then the EM algorithm is an iterative algorithm. At stage \(r\), we have a “current guess” \(\Theta^{(r)}\), and we use it to calculate:

\[
E-\text{Step: } \ell_r^E(\Theta) = \mathbb{E}_{\Theta^{(r)}}(\ell(\Theta; X, Y)|X),
\]

that is, the expected value of the log-likelihood, integrated over the unknown \(Y\), and using the current vector \(\Theta^{(r)}\) in the distribution of \(Y|X\). Note that \(\Theta\) plays two roles here – one, where \(\Theta^{(r)}\) is used to calculate conditional expectation, and two, where \(\Theta\) is used symbolically in the likelihood. For example, if \(Y\) appears only linearly in the log-likelihood, then we simply plug \(\mathbb{E}_{\Theta^{(r)}}Y|X\) into this to obtain \(\ell_r^E\).

The next step is the M-step, which finds the best value of \(\Theta\) given the current integrated likelihood \(\ell_r^E\):

\[
M-\text{Step: } \Theta_{r+1} = \arg \max_{\Theta} \ell_r^E(\Theta).
\]

The theoretical guarantee we get is that \(\ell(\Theta^{(r)}; X)\) is an increasing function of \(r\), which converges to a local maximum (not necessarily the MLE, which is the global maximum). For convex problems, it will eventually converge to the MLE.

**EM for our problem**

Define as unobserved data: \(Z = \{Z_{ima}\} \in \{1, \ldots, K\}\) the ethnic origin (e.g., Eu or Af) of the \(a\)th copy of the \(m\)th marker in the \(i\)th individual.

For \(i > I_0\), \(Z_{ima} = Q_{ia}\) is in fact known, since \(Q_{ia} \in \{0, 1\}\). For \(i \leq I_0\), under our assumptions \(Z_{ima} \sim \text{multinom}(Q_{ia})\).

The log-likelihood of the complete data:

\[
\ell(P, Q; G, Z) = \sum_i \sum_m \sum_a \sum_l \sum_k \mathbb{I}\{G_{ima} = l, Z_{ima} = k\} \log(P_{mlk}Q_{ik}).
\]

From this it is easy to see the form of the E-step:

\[
\ell_r^E(P, Q) = \mathbb{E}_{Q^{(r)}, P^{(r)}}(\ell(P, Q; X, Y)|X) = \sum_i \sum_m \sum_a \sum_l \sum_k \mathbb{I}\{G_{ima} = l\} \mathbb{P}_{Q^{(r)}, P^{(r)}}(Z_{ima} = k|G) \log(P_{mlk}Q_{ik}).
\]

We have to calculate the probability / expectation, denote

\[
E_{imak}^{(r)} = \mathbb{P}_{Q^{(r)}, P^{(r)}}(Z_{ima} = k|G),
\]

and assume \(G_{ima} = l\) is given, then:

\[
E_{imak}^{(r)} = \mathbb{P}(Z_{ima} = k|G_{ima} = l; P_{mlk}^{(r)}, Q_{ik}^{(r)}) = \frac{\mathbb{P}(Z_{ima} = k, G_{ima} = l)}{\mathbb{P}(G_{ima} = l)} = \frac{P_{mlk}^{(r)}Q_{ik}^{(r)}}{\sum_{u=1}^K P_{mlu}^{(r)}Q_{iu}^{(r)}},
\]

\(4\)
where the equality (*) is due to the independence assumptions we made above (each \(Z_{ima}\) is drawn independently than all other \(Z\)'s, and \(G_{ima}\) depends only on \(Z_{ima}\).)

Now we can write the explicit integrated likelihood to move to the M-step:

\[
\ell^E_r(P, Q) = \sum_{i,m,a,l,k} \mathbb{1}\{G_{ima} = l\} E^{(r)}_{imak} \log (P_{mlk}Q_{ik}) = \\
= \sum_{m,k,l} \left[ \log (P_{mlk}) \sum_{i,a} \mathbb{1}\{G_{ima} = l\} E^{(r)}_{imak} \right] + \sum_{i,k} \left[ \log (Q_{ik}) \sum_{m,l,a} \mathbb{1}\{G_{ima} = l\} E^{(r)}_{imak} \right],
\]

and maximizing this to find \(P^{(r+1)}\), \(Q^{(r+1)}\) is easy:

\[
P^{(r+1)}_{mlk} = \frac{\sum_{i=1}^{I} \sum_{a=1}^{2} \mathbb{1}\{G_{ima} = l\} E^{(r)}_{imak}}{\sum_{i=1}^{I} \sum_{a=1}^{2} E^{(r)}_{imak}} \\
Q^{(r+1)}_{ik} = \frac{\sum_{m=1}^{M} \sum_{a=1}^{2} E^{(r)}_{imak}}{2M} \text{ for } i \leq I_0 \\
Q^{(r+1)}_{ik} = Q_{ik} \text{ known for } i > I_0.
\]