1.5.1 ESTIMATION OF HAPLOTYPE FREQUENCIES:

• For 2 SNPs, alleles $A_j, B_j$ at locus $j$ there are 4 haplotypes: $A_1A_2, A_1B_2, B_1A_2$ and $B_1B_2$ frequencies $q_1, q_2, q_3, q_4$. Assume HWE at haplotype level.

| Loc-2 | Locus 1 |
|-------|---------|--------|
|       | $A_1A_1$ | $A_1B_1$ | $B_1B_1$ |
| $A_2A_2$ | $q_1^2$ | $2q_1q_3$ | $q_3^2$ | $(q_1 + q_3)^2$ |
| $A_2B_2$ | $2q_1q_2$ | $2(q_1q_4 + q_2q_3)$ | $2q_3q_4$ | $2(q_1 + q_3)(q_2 + q_4)$ |
| $B_2B_2$ | $q_2^2$ | $2q_2q_4$ | $q_4^2$ | $(q_2 + q_4)^2$ |

• Only the double heterozygote $A_1B_1, A_2B_2$ cannot be phased.

• There are 10 two-locus genotypes, but only 9 phenotypes.

• Haplotype frequencies can be estimated from phenotype frequencies via the EM algorithm:

$$P_{t}(A_1A_2/B_1B_2|A_1B_1, A_2B_2) = q_1q_4/(q_1q_4 + q_2q_3)$$

Given a set of current haplotype frequency estimates $q_i, i = 1, \ldots, 4$ and the phenotypic counts, the conditional expected genotypic counts are easily obtained. New haplotype estimates then are the expected multinomial proportions of each haplotype.

• For consistency with PHASE and other software we will temporarily label the two alleles at each SNP 0 and 1.

• At two loci there are 4 haplotypes: 00, 01, 10, 11, freqs $q_{00}, q_{10}, \ldots$.

There are 10 phased two-locus genotypes: eg; 10/00, 11/00, 10/01

There are 9 observable two-locus genotypes: (00,00), (10, 11), ....

• Only the double-heterozygotes (10,10) is ambiguous: 11/00 or 10/01.

For other genotype pairs we just count.

• Suppose there are $H$ double-heterozygotes.

Then $E(\#(11/00) \mid H) = Hq_{11}q_{00}/(q_{11}q_{00} + q_{10}q_{01})$

• So EM is easily implemented.

For two (or a very few) loci it works well.
1.5.2 ESTIMATING PHASE AND HAPLOTYPE FREQUENCIES:

- Consider diallelic loci (e.g. SNPs): Label alleles 0 and 1.
- Then the genotypes are a set of pairs e.g. 00, 10, 11, 10, 10, and haplotypes a string such as 01010. Determining phase is determining which of the 4 possibilities 01111 and 00100, 01110 and 00101, 01101 and 00110, 01100 and 00111 holds.
- For convenience we may write an unphased genotype as (01111,00100) and the phased version as (01111/00100).
- For large samples and/or small numbers of SNPs we may use EM algorithm to estimate haplotype frequencies, and this also provides probabilities of phasings, given the estimated frequencies.
- However, for large numbers of SNPs this does not work well: many sample haplotype freqs are 0, and likelihood surface is high-dimensional and multimodal.

1.5.3 HAPLOTYPING: TWO (OTHER) ALGORITHMS:

- Clark (1990)’s algorithm: note where individuals are homozygous, haplotyping is trivial. Also trivial if heterozygous at just 1 locus.
- Use individuals heterozygous at at most 1 locus to identify haplotypes that must be present. Assuming these, see which other individuals can be explained by one of these haplotypes, plus a new one – add these new ones to the collection, and continue for as long as possible.
- Problems: May not be able to start. May not be able to finish. Final guess may depend on order one adds haplotypes to the pool.
- Stephens’ algorithm (PHASE): use a model that summarizes similarities of haplotypes in a population – the idea is that haplotypes should look like each other "in chunks". Use Monte Carlo to simulate alternative phasings under the model. Produces "probable phasings" with estimates probabilities. (Now also fastPHASE. Scheet and Stephens (2006).)
1.6.1 ALLELIC ASSOCIATION; LINKAGE DISEQUILIBRIUM:

- A measure of allelic association between the two loci is

\[ \Delta = \Pr(A_1 A_2) - \Pr(A_1) \Pr(A_2) \]
\[ = q_1 - (q_1 + q_2)(q_1 + q_3) \]
\[ = (q_1 q_4 - q_2 q_3) \]

since \( q_1 + q_2 + q_3 + q_4 = 1 \). This measure is known as the coefficient of linkage disequilibrium.

- Allelic associations between loci arise from population structure, admixture and history, or from selection. (See 1.6.2 and 1.6.3).
  Example of mixture/subdivision— the “nuisance” case. Vs. case of interest— original mutation on some genetic background.

- Associations are, however, maintained by tight linkage (\( \rho \approx 0 \)).
  LD blocks are the remnants of recombination; they are not caused by linkage, but they survive because of linkage. (See 1.6.5).

- Contrast with HWE: Even for unlinked loci equilibrium (\( \Delta = 0 \)) is not achieved in one generation.

1.6.2 LD ARISES BY MUTATION OR STRUCTURE:

IBD underlies Allelic Association (LD)

- Associations across loci (LD) reflect population level coancestry.
- New variant allele remains in LD with background.
- Population subdivision also results in LD.
1.6.3 LD and POPULATION STRUCTURE:

• Population stratification creates LD, even if there is no LD within subpopulations. If allele frequencies differ so will the frequency of haplotypes.

• Consider populations $i$ in proportions $\alpha_i$ and allele $A$ alleles frequencies $p_{1i}$, $p_{2i}$ at two loci. As before $P_1(A_1) = p_1 = \sum_i \alpha_i p_{1i}$, and $P_1(A_2) = p_2 = \sum_i \alpha_i p_{2i}$.

• Assume there is NO LD within each subpopulation, so the overall frequency of the $A_1A_2$ haplotype is $P_1(A_1A_2) = \sum_i \alpha_i p_{1i}p_{2i}$.

• Then

$$\Delta = P_1(A_1A_2) - P_1(A_1)P_1(A_2) = \sum_i \alpha_i p_{1i}p_{2i} - (\sum_i \alpha_i p_{1i}) (\sum_j \alpha_j p_{2j})$$

$$= \sum_i \alpha_i (p_{1i} - p_1)(p_{2i} - p_2).$$

1.6.4 SEGREGATION OF HAPLOTYPES:

• Homozygous individuals (both loci): for example an $A_1A_1$, $B_2B_2$ individual segregates only $A_1B_2$ haplotypes.

• Homozygote/Heterozygote: for example, an $A_1A_1$, $A_2B_2$ individual passes on $A_1A_2$ or $A_1B_2$ each with probability $1/2$ regardless of recombination probability $\rho$.

• A double-heterozygote individual passes each of the four haplotypes $A_1A_2$, $A_1B_2$, $B_1A_2$ and $B_1B_2$, with probabilities: $(1-\rho)/2$, $\rho/2$, $\rho/2$ and $(1-\rho)/2$ if his genotype is $A_1A_2/B_1B_2$, $\rho/2$, $(1-\rho)/2$, $(1-\rho)/2$, and $\rho/2$ if his genotype is $A_1B_2/B_1A_2$.

• Recombination breaks up chromosomes, but we only see this directly if genotypes are heterozygous at both loci.
1.6.5 DECAY OF LD:

• Suppose current haplotype freqs of $A_1A_2$, $A_1B_2$, $B_1A_2$ and $B_1B_2$ are $q_1$, $q_2$, $q_3$ and $q_4$, and at next generation are $q_1^*$, $q_2^*$, $q_3^*$ and $q_4^*$.

• Now, for example, an offspring $A_1A_2$ haplotype arises from a $A_1A_1$, $A_2A_2$ parent with prob 1.
  - from a $A_1A_1$, $A_2B_2$ or $A_1B_1$, $A_2A_2$, with prob 1/2,
  - from a $A_1B_1$, $A_2B_2$ who is $A_1A_2/B_1B_2$ with prob $(1 - \rho)/2$
  - from a $A_1B_1$, $A_2B_2$ who is $A_1B_2/B_1A_2$ with prob $\rho/2$. Thus

\[
q_1^* = q_1^2 + 2q_1(q_2 + q_3)/2 + 2q_1q_4(1 - \rho)/2 + 2q_2q_3\rho/2
\]
\[
= q_1(q_1 + q_2 + q_3 + q_4) - \rho(q_1q_4 - q_2q_3) = q_1 - \rho\Delta.
\]

Analogously, $q_2^* = q_2 + \rho\Delta$, $q_3^* = q_3 + \rho\Delta$ and $q_4^* = q_4 - \rho\Delta$. Thus, in expectation, allele frequencies are unchanged ($q_1^* + q_2^* = q_1 + q_2$):

\[
\Delta^* = q_1^*q_4^* - q_2^*q_3^*
\]
\[
= (q_1 - \rho\Delta)(q_4 - \rho\Delta) - (q_2 + \rho\Delta)(q_3 + \rho\Delta)
\]
\[
= \Delta - \rho\Delta(q_1 + q_2 + q_3 + q_4) + \rho^2(\Delta^2 - \Delta^2)
\]
\[
= (1 - \rho)\Delta.
\]