1.7 Haplotypes and allelic association
1.7.1 Estimating Phase: 2 loci

- Consider diallelic loci (e.g. SNPs). Label alleles 0 and 1.
- At two loci there are 4 haplotypes: 00, 01, 10, 11.
- There are 10 phased two-locus genotypes e.g; 10/00.
- Observable (unphased) genotype is a pair of pairs. There are 9 (3 x 3) observable two-locus genotypes. e.g. (00,00), (10, 11), ...
- Only the double-heterozygotes (10,10) is ambiguous. For other genotype pairs we just count.
- Suppose the current estimates of haplotype frequencies are q00, q01, q10, q11, and there are H double-hets.
- Then E(#(11/00) | H) = H q11 q00/(q11 q00 +q10 q01).
- So EM is easily implemented.
- For two (or a very few) loci it works well.

1.7.2 Estimating phase: multiple loci

- Consider diallelic loci. Label alleles 0 and 1.
- A haplotype is a string such as 01010.
- Genotypes are a set of pairs e.g. 00, 10, 11, 10, 10.
- Determining phase is determining which of the 4 possibilities 01111 and 00100, 01110 and 00101, 01101 and 00110, 01100 and 00111holds.
- Notation: 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 give probabilities of phasings.
- 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.

1.7.3 Two other haplotyping algorithms

- Clarke'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 estimated probabilities. (Now also FASTPHASE.)

1.8 Maintaining variation
1.8.1 Mutation and selection

- Mutation provides new variation.
- Directional selection removes variation.
- In equilibrium, ``loss" = ``gain". Hence, indirect estimates of mutation rates.
- For example, recessive with selection coefficient s. We lose 2 A alleles, with prob, s, for each AA individual. We gain μ A alleles, in each of 2N meioses (approx.). So Ns 2 q^2 = 2N μ or μ = sq^2, and q=√(μ/s).
- For dominant with selection s. We lose 1 A allele, with prob. s, for each AB individual. Gain as before. So Ns 2q(1-q) = 2N μ, or μ = sq, and q= μ/s (approx.).

1.8.2 Random genetic drift

- Real populations are finite (and have structure, and history, …).
- Let X(t) be number of A alleles at time t in popn size 2N genes.
- Wright-Fisher model: (X(t) | X(t-1)) is Bin(2N, X(t-1)/2N).
- Then E(X(t)) = E(E(X(t)|X(t-1))) = E(2N (X(t-1)/2N)) = E(X(t-1)) = …. = X(0).
- Hence E(X(∞))=X(0) so P(X(∞) = 2N) = X(0)/2N.
- Note E(X^2) = var(X) + (E(X))^2, so E(X(t)^2) = E(E(X(t)^2 | X(t-1))) = E(var(X(t)|X(t-1))) + E(X(t-1)^2)
- Homozygosity increases relative to time 0, because the allele frequency has increasing chance of being closer to 0 or 1, but population is still in HWE.

Over time, populations diverge

- Let V(t) = var(X(t)) and X(t) and Y(t) be time-t counts of alleles type A in two independent populations with same X(0)=Y(0) at time 0.
- E((X(t) - Y(t))^2) = E(X^2) - 2 E(XY) + E(Y^2) = (V(t) + X(0)^2) -2 X(0)^2 + (V(t) + X(0)^2) = 2 V(t).
- At first generation V(1) = 2N (X(0)/2N) (1 -X(0)/2N).
- So if allele freqs do not change much (N large or t small), 2 V(t) is approx (4 N t) (X(0)/2N) (1 -X(0)/2N).