Questions on Li et al (2010) *Genet Epidemiol*. 34:816-34.

MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.

1. What prompted the development of methods for genotype imputation?

2. The paper makes a connection between methods for genotype imputation and methods for estimating haplotypes… Why?

3. What is the basic idea behind a Hidden Markov Model? In what other settings have you encountered HMM?

4. What is the set of possible states? What are the transition probabilities?

5. A series of summaries for imputation quality are proposed. What are they? Do you expect these measures would apply equally well to common and rare variants?

6. The paper describes an algorithm to reduce memory requirements for Markov algorithms from linear (proportional to the number of markers, for example) to something much smaller (proportional to the square root of the number of markers). How do these work? What is their downside?

7. The paper describes an evaluation of the approach in diverse populations. Consider applying the approach to admixed individuals. In these individuals, different stretches of chromosome may have different ancestries. Can you imagine how to refine the HMM to account for this?

8. The paper describes an approach to limit computational cost, by limiting the number of template haplotypes. How do you think that approach might be refined?

9. In addition to describing a new method, the paper includes quite a number of additional experiments, using both real and simulated data. Did you find that unusual? What do you think prompted that?

10. What struck you most about the paper?