Questions on Menelaou et al (2013) *Bioinformatics* 29:84-91.

Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold.

1. This paper describes methods for deriving accurate genotypes from low pass sequence data. Why is low pass sequencing appealing?

2. The original goal of the 1,000 genomes project was to sequence a large set of individuals at an average depth of ~2-4x.

   Suppose a heterozygous site (with genotype A/B) is sequenced to depth 4 and the sequencing error rate is 1% … What is the probability of observing 0, 1, 2, 3 and 4 copies of allele A? What would happen if you repeated the experiment for a homozygous site (with genotype A/A)?

3. What are genotype likelihoods?

4. The authors use a multivariate normal distribution to describe the relationship between genotyped sites in the haplotype scaffold and sequenced sites where genotypes must be estimated. Can you imagine other situations where a similar model could be used to impute missing data? Provide some examples.

5. The Hidden Markov Model we read about in the Li et al (2010) paper could also be used to estimate genotypes and haplotypes from low pass sequence data. How do the inputs required by the Li et al (2010) method differ from those used here?

6. As the number of individuals and markers being analyzed changes, how do expect computational cost for the standard methods to increase? How does this compare to methods based on Hidden Markov Models similar to the one used by Li et al (2010)?

7. The authors describe an improvement to their method based on *surrogate families*. What are these *surrogate families*? What strategies did the authors consider for defining *surrogate families* and which do they recommend?

8. The authors describe another improvement termed *model averaging*. How does that improvement work?

9. In Table 2, the relative ranking of the methods appears to vary by population. What are some of the possible reasons?

10. What struck you most about the paper?