Stats for comp bio midterm 2018
Due April 3

Pr. 1 Consider the two mixture of multinomial models stated in "Inference of Population Structure Using Multilocus Genotype Data": Write out the EM algorithm to estimate the model parameters for the model with and without admixture.

Pr. 2 For the same eQTL data that was used in Homework 2.

1. State how you would use LASSO to implement the regression problem. State the model and optimization problem that needs to be solved to estimate the β’s;
2. Estimate the β parameters using Least Angle Regression.
3. Explain your results and contrast then to a standard shrinkage model.
4. Explain the difference between cis or proximal qtls and trans or distal eqtls.

Pr. 3 Read the following paper: "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.”

1. State what the multiple hypothesis testing problem is for this paper. How are FDR and FWER rates controlled?
2. How would you adapt this problem to the case where one is given genotype data instead of expression data. You are still interested in gene set enrichment and you have a list that assigns SNPs to genes.
3. How would you adapt the statistics in this paper to the case where the expression data was not given as array data but as high throughput sequencing data.

Pr. 4 Read “A New Statistical Method for Haplotype Reconstruction from Population Data”
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1275651/
by Matthew Stephens, Nicholas J. Smith, and Peter Donnelly.
State the model and explain the inference procedure.

Pr. 5 Write out the likelihood model for a mixture of regressions
https://www.ceremade.dauphine.fr/ xian/jcgsrobert.pdf
write out an EM algorithm to infer the parameters.