Questions for class preparation

Personal Genetics

Papers

- Illumina HumanOmniExpress-24 BeadChips Data Sheet.

Website

- http://www.1000genomes.org/home

Questions

Assigned questions for Rosenberg, et al.:

- What are microsatellite loci?

- What advantages do *autosomal* microsatellite markers provide (versus sex chromosomes)?

- What is the advantage of using only neutral microsatellites?

- What is a “haplotype”. How can this explain 377 microsatellite markers representing 4199 alleles among these populations?

- Describe K clustering.

- What populations might you expect to share similar “membership coefficients” (hint: this refers to K clustering)?

- What physical (geographic) phenomena might you expect to separate the major world divisions of individuals by this method?
• What other characteristics could perhaps explain membership clustering besides geographic proximity?

Questions for in-class discussion:

• Describe how the K clustering in Rosenberg, et al. Fig. 1 explains the origins of the Kalash group despite their geographic location.

• Explain how it might be that the populations sampled from the Americas have such low heterozygosity?

• Given a random sampling of individuals from the U.S., discuss what you might expect our population structure to look like. Why? Given a large enough K, what populations might you expect to emerge?

Assigned questions for the International HapMap Project.: 

• What was the main goal of the International HapMap Project?

• How does a single nucleotide change differ from a microsatellite?

• How can a single nucleotide change at one base confer information about the neighboring region?

• What is linkage disequilibrium (LD)? How does the likelihood of recombination change based on the distance between 2 SNPs (hint: this question and the following question both concern LD)?

• Blocks of genetic sequence that travel together during recombination are called “LD bins”. These are separated by regions where recombination is most likely to occur sometimes called “hot spots”. As crossing over happens from generation to generation, the size of the LD bins shrinks as the ends are slowly chewed back. In what populations might you expect these bins to be the smallest?

• Compare and contrast the primary differences between “direct” and “indirect” association studies.

• How might the data from this study be useful in studying human disease (hint: we sometimes call those types of studies genome-wide association studies, GWAS)?

• What quality control measures were taken in this study to ensure the data presented were accurate (hint: were samples replicated? how were samples assigned for analysis?)?
Assigned questions for the 1000 Genomes Project:

- What was the goal of the 1000 genomes project?

- How is the 1000 genomes dataset different than the HapMap dataset? (hint: what types of data were collected from each?)

- What samples were used for the 1000 genomes project?

- What were the advantages of using the HapMap samples for the 1000 Genomes project?

- The basic pipeline for analysis of the 1000 genomes data was as follows:
  - 1) alignment and variant discovery
  - 2) data filtering
  - 3) sample genotyping
  - 4) validation

  List portions of the genome that might be problematic during the alignment phase of analyzing a genome. How might this be resolved?

- The 1000 genomes project uses multiple mutation calling programs on the data. What might be the advantages and disadvantages of using an intersection (only events identified using ALL programs) versus a union (all events called by all programs) set of events?

- List another technology that has been discussed already in the course (or another) that might be a good source of data for the validation phase and describe why you picked this.

- Some parts of the pilot project got greater sequence coverage (i.e., the trio and exome vs. the “low coverage” genome). What advantages does greater coverage provide? What are the pros and cons of many low-coverage versus a few high-coverage genomes?

- Why is it perhaps not surprising that protein damaging mutations were not found at a high frequency in the population but instead genetically benign common SNPs?

- In what cases might trio sequencing (hint: father, mother, and offspring) be useful?

- What sources of DNA may be more prone to new mutations that are likely not de novo germline events?

- What is the best source of DNA for large-scale studies moving forward? Why (hint: this should be directly relevant to the previous question)?