COS 597D: Data driven statistical genetics

Homework 1 (9/10/14)

Lecturer: Barbara Engelhardt

Read: [Troyanskaya *et al.* 2001] (PDF in Dropbox folder).

In class white-board presentation due 9/15/2014

What scientific problems interest you? What hypotheses are you interested in studying, modeling, and testing? What does the data set that you can test your hypotheses on look like?

Your homework assignment this week is to, individually, identify a set of genomics data (either open source, or accessible to you) and describe three possible hypotheses or explorations that you believe these data will support. You will be asked to give a short presentation at the white board describing the data and also the hypotheses and explorations you are considering.

Open source genomics resources

It may be worthwhile to identify specific studies that have investigated some of these data sets. PubMed and Google Scholar will help you to identify papers that have cited each of these databases and repositories (you can find the "reference" for each resource on their website, and find citations to that reference).

- *dbGaP*: repository for genotype and other genomic data
- *Swiss-Prot/UniProt*: database of gene transcripts, well annotated with functional information
- *Protein Data Bank (PDB)*: database of tertiary structures of tens of thousands of proteins
- *Gene Ontology*: ontology represented as a directed acyclic graph of gene function, cellular location, biological role of many genes
- *Gene Expression Omnibus (GEO)*: gene expression and RNA-sequencing repository
- *The Cancer Genome Atlas (TCGA)*: gene expression and RNA-seq repository for cancer studies
- *Sage Bionetworks Synapse*: repository for many different types of genomic study data
- *Ensembl*: gene transcripts and isoforms, well annotated
- *1000 Genomes/HapMap*: 40M SNPs from > 1000 individuals worldwide, well curated and phased
- *NHGRI GWAS Catalog*: all trait-associated SNPs ever published
- *GenBank (NCBI)*: database of all publicly available annotated genomic sequences
- *ENCODE data*: repository for thousands of regulatory element assays on many different cell types
- *UCSC Genome Browser*: visualization for all of these data sets and more
- Many more; see `http://www.nature.com/scitable/topicpage/genomic-data-resources-challenges-and-pro` for some possibilities (although this is from 2008), or google other options.