**Introduction to Algorithms and Data Structures**

BIO514 Spring 2016

**Final Projects**

For your final project, you should form small groups of 1-3 people each. The dataset is provided on the course web site:

* Human sequencing data from the 1000 Genomes Project (one chromosome)

Your tasks:

1. Develop an efficient storage algorithm and provide a Python implementation. It should achieve at least comparable compression/space efficiency as ZIP. Describe the algorithm briefly in prose and qualitatively assess its runtime efficiency.
2. Using the variance/covariance matrix between individuals, or other similarity matrix, develop an algorithm that determines the degree to which there exist distinct subpopulations within the sample. Possibilities are principal component plots, clustering algorithms, or community detection algorithm. Choose one, apply it to your data of choice, and discuss the results, their statistical significance, and the algorithm's numerical feature (e.g. complexity and stability).
3. In Task 2, you may decide to use an existing approach to detect the sub-populations in your dataset. If you decide to do this, then Task 3 is mandatory. Otherwise, it is optional.

Develop a matrix norm for your genetic similarity matrix. Develop an algorithm that, for each sub-population, identifies a subset of genetic loci (min 1000 loci) whose genetic similarity matrix (just computed based on the selected subset of loci) has maximal distance to overall genetic similarity matrix (computed based on all genetic loci), with respect to your matrix norm

1. As a group, present the project and its results in class, and submit its implementation/code.
2. Individually, subject a ~two page writeup (plus citations) including one (or at most two) figure(s) addressing the two items above. Citations won't count toward the two pages, which are a soft limit (although don't expect us to grade a 30-page Cell paper!) See the next few pages for a brief example.

So again, presentations, write-up and code/data file submissions are one per group. Each team will give one 15-minute presentation (plus ~5 min. questions) describing your problem, your approach, and your results, and the team will submit the entire packet of code and data (as appropriate) with which you tackled the problem. **\*\*\*Final project submissions (both one per group for results, and one per individual for writeups), will be due by the end of the term\*\*\***.