Report: Data Wrangling, Analysis, and Visualization

Project: Genetic Diversity in Human Populations
Author: Greg Gunterson (ggunters@ucsc.edu)
Data Sources:

- http://www.hagsc.org/hgdp/files.html
- http://www.cephb.fr/common/HGDPid_populations.xls

Visualization Source:

- https://bost.ocks.org/mike/miserables/
- https://d3js.org/d3.v2.min.js

GitHub Repository:

- https://github.com/ggunters/genetic_diversity

Scripts:

- `_1generate_pops_tsv.py`
- `_2generate_rand_sample.py`
- `_3parse_data.py`
- `_4generate_diffs_map.py`
- `_5compute_fixation_indices.py`
- `_6generate_json.py`

Summary of Data Wrangling:

Data from `HGDPid_populations.xls` was parsed and output into `populations.tsv` using the script 1. A random sample of the genotype data contained in the file `HGDP_FinalReport_Forward.txt` was generated and output into `rand_sample.tsv` using script 2. Data from these `.tsv` files were then parsed in script 3 and stored in data structures `genotypes_matrix` and `pops_map`, which were saved using Python's 'pickle' module.

Summary of Data Analysis:

Analysis was done in scripts 4 and 5. Script 4 counts the number of nucleotide differences per nucleotide site for each pair of individuals and stores these values in the `diffs_map` data structure, which is saved. Script 5 loads this data structure, as well as `pops_map` and uses these to compute fixation indices for each pair of populations. Each fixation index was computed using $(\pi_{\text{between}} - \pi_{\text{within}}) / \pi_{\text{between}}$, where each π value is computed as the average number of pairwise nucleotide differences per site, per individual[1]. Fixation indices were stored in the `f_sets` data structure, which was saved.

Summary of Visualization:

Data from the `f_sets` data structure was loaded and output as JSON for compatibility with d3 and JavaScript in script 6. This JSON was then copied and hard-coded into the visualization file `genetic_diversity.htm`. JavaScript source code contained in the visualization file was adapted from the page linked above, with modifications for a color legend and tooltip.

[1] see https://en.wikipedia.org/wiki/Fixation_index#Estimation