Project suggestions

- Comparing existing tools
  - Explore risk prediction tools/methods, and the effect of various methods for controlling for ancestry and linkage disequilibrium.
  - Evaluate and compare different methods for annotating non-coding variants.
- Prototype a novel bioinformatics tool
  - Genotyping repeats (e.g. STRs, VNTRs, Alu/Line) from long-read or synthetic long read technologies (10X, Nanopore)
  - Detecting mosaicism at complex variants (e.g. STRs) from next-generation sequencing data
- A visualization tool for making genomic data more accessible to non-experts
  - Visualizing annotations from VCF files (e.g. see http://vcf.iobio.io/ for inspiration)
  - Choose a trait not already presented by 23andMe or DNA.land, and create a web page predicting a person's trait/risk for a complex trait that provides a tutorial behind the genetics of the trait and a visualization of the contribution of different SNPs.
  - A web page for use in forensics, that predicts relevant traits (e.g. ancestry, eye color, skin color, or even height and more complex phenotypes) from a genotype file, along with explanations for lay people.
- Novel analysis of genomic data
  - (idea from Stanford course): tool to simulate "offspring" from a potential "mating" between two genomes. Analyze distribution of traits in siblings.
  - Explore importance of non-coding annotations using variance partitioning.

Public datasets

The following are public datasets that are available for the project:

- The 1000 Genomes Project http://www.internationalgenome.org/ Sequencing data and variant calls for about 3,000 samples worldwide
- the Simons Genome Diversity Project https://www.simonsfoundation.org/life-sciences/simons-genome-diversity-project-dataset/ 300 high coverage whole genomes from diverse population groups.
- The ENCODE Project https://www.genome.gov/10005107/ Histone modifications, transcription factor, expression, and more data from a variety of human cell types.
- The GTEx Project gtexportal.org. Expression data and QTLs from hundreds of human cell types.
- Familinx http://www.familinx.org/. Pedigree info for millions of individuals.