

Tools for Visualization of Geographic Structure in Population Genomic Data

UNIVERSITY OF CHICAGO

PI: NOVEMBRE, JOHN

Grant Number: 1 U01 CA198933-01

Large sample sizes are increasingly common in genetics/genomics, particularly in human genetics where sample sizes must be large (>1,000s of individuals) to detect variant associations with complex disease traits. A common feature of data from large samples is that the individuals within the study have varying levels of similarity with one another that can become problematic for downstream analyses (e.g. causing spurious associations) if not understood. Thus uncovering population structure and dissecting it to understand its source is a common and important practice in large-scale studies. Here, we aim to solve challenges for visualizing population structure that regularly arise when researchers interact with large-scale population genomic data sets. In Aim 1 we will develop a software tool for visualizing population structure using principal components analysis (PCA). This tool will make straightforward several steps that are commonly reinvented by data scientists as they analyze PCA outputs from genetic data. It will also make more clear whether PCA analyses may be returning anomalous results. In Aim 2 we will develop a tool for producing geographic allele frequency maps of publicly available or user-generated allele frequency data. In Aim 3 we will develop a visualization approach for displaying geographic regions where populations show unexpectedly high or low levels of differentiation. In Aim 4 we will integrate these pieces of software into a single suite and link them to externally generated data sources and existing genome browsers. By developing these sets of tools we help to remove the need for unnecessary script generation by independent researchers and increase the pace of genomics research. Throughout the project we will pay special attention to developing user-friendly interactive data displays such as those generated by the Data Driven Documents (d3) JavaScript visualization libraries. Where possible we will use simple, yet flexible python backends and provide complementary R libraries to facilitate customizations and integration with existing analysis pipelines. While population genetic applications will motivate our work, the tools we are generating will be generally applicable to other forms of structured biomedical data.

PUBLIC HEALTH RELEVANCE PUBLIC HEALTH RELEVANCE: This project will provide tools for visualizing large-scale genetic data with population structure. While numerous advanced algorithms for summarizing population structure exist, the human interface to the outputs of these methods is lacking and has become a time sink during the analysis of large samples. In this project we will provide user-friendly tools that lower the barrier to understanding genetic variation datasets. In particular we will develop tools for visualizing compressed representations of genetic variation (i.e. PCA results) and how genetic diversity is distributed across geographic space in a sample.