Development of a custom ADMIXTURE to assess disparities in cancer outcomes in TCGA

Timothy J Sears, PID: A14109446, Varun Surapaneni, PID: A59019115

Option 1: Implement a method we talked about in class and apply it to real data

We will build a version of the ADMIXTURE tool that allows us to assess the genetic ancestry of a given dataset. Similar to the original tool, the user will have to pre specify a number of populations that the subject is supposed to have been descended from. This tool will be implemented as a command line utility, where the source code is written in python. All analysis will be run within a jupyter notebook and visualizations of data will be preserved. The tool will be published on a public github repository. The original ADMIXTURE tool will be run on the same data as a baseline comparison and ground truth.

Disparities in cancer outcomes by race is a pernicious issue in the United States. While it is the hope of all cancer researchers and medical providers that each patient receives adequate care regardless of their race, cancer types such as prostate, bladder, and breast have varying levels of detection and outcome by race.

We aim to improve our understanding of these findings by applying a version of the ADMIXTURE tool on germline data from TCGA to cluster patients into approximately three ancestry groups. The resulting data will be used to stratify patients into several populations, and their cancer outcome (measured by overall survival, adjusted for tumor type) will be assessed. We also aim to see if admixed populations have different outcomes than patients belonging entirely to specific ethnic groups.

We expect that certain cancer types will have larger racial disparities than others, and that our ADMIXTURE-like tool will sufficiently replicate what is already known by the medical community.

Our team already has access to the necessary datasets and all appropriate measures to protect patient identities will be adhered to.