Part 3 - Novel Analysis Plan
Wanjun Gu

The proposed project aims to investigate the enrichment of transcription factor footprints (referred to as footprints henceforth) within genomic loci demonstrating robust evidence of positive natural selection. Specifically, I intend to focus on adaptation to high-altitude environments characterized by hypoxic conditions. Through this study, I will examine the presence of TF footprints in genomic regions that exhibit indications of positive natural selection in response to hypoxia-induced environmental stress. This approach will shed light on the role of transcription factors in the adaptive evolution of organisms under challenging environmental conditions, contributing valuable insights to our understanding of genomic responses to selective pressures. For this project, two primary data sources will be utilized. The first data source involves genomic loci under positive natural selection in Tibetans, who are native highlanders. This information is extracted from Simonson et al. 2010 (DOI: 10.1126/science.1189406). The second data source encompasses footprinting data, which is available on the website https://www.vierstra.org/resources/dgf. Specifically, BED files based on different cell types can be downloaded from the "directory listing" in the "Footprints identified in individual datasets" section. Additionally, consensus footprints can be obtained from the "Consensus Footprints" section on the same website. The hypothesis for this study posits a null hypothesis suggesting no enrichment of footprints in regions under positive natural selection due to exposure to low-oxygen conditions. Conversely, the alternative hypothesis suggests that there is indeed an enrichment of footprints in regions undergoing positive natural selection as a consequence of adapting to the low-oxygen conditions associated with high-altitude living.

The statistical methods employed in this project will leverage the supplementary material of Simonson et al. 2010, allowing us to download genomic regions exhibiting strong positive selection. For each of these identified regions, I will select the leading marker, defined as the top marker from the population genetics selection scan, and extend the analysis by 200 kb both upstream and downstream to encompass the entire region. Subsequently, I will calculate the total number of transcription factor (TF) footprints within these expanded regions. To establish a baseline for comparison, I will repeat this analysis on randomly generated genomic regions that match in size and distribution across chromosomes. Through a bootstrap analysis, I will assess whether the TF footprints more frequently occur in regions under natural selection compared to randomly defined genomic regions. Additionally, I will also examine if the fold enrichment increases in relation to both the strength of footprinting and the evidence for natural selection. The software utilized for data cleaning and pre-processing will include BEDTools (https://bedtools.readthedocs.io/en/latest/), liftover (https://genome.ucsc.edu/cgi-bin/hgLiftOver), and R (https://cran.r-project.org/bin/windows/base/). R will further be employed for enrichment analysis and visualization, ensuring a comprehensive and rigorous approach to analyzing the relationship between TF footprints and genomic regions under positive natural selection.