DECODE the History of Cancer Evolution from Bulk DNA-Sequencing Data

Elliott Seo (Mentor: Dr. Khanh N. Dinh, 2024 IICD SRP)

As DNA-sequencing becomes a common diagnostic tool for examining tumors, one of the most important goals has been the inference of tumor development history, toward better patient classification and treatment selection. A common summary statistic for bulk DNA data is the Site Frequency Spectrum (SFS). Recently, MOBSTER [1] was introduced as a method to decompose the mutational SFS into separate clusters of high frequencies, which indicate past selective sweeps, and the neutral tail, which is formed as cells mutate during divisions. However, MOBSTER assumes binomial coverage, which does not align with real DNA data [2]. We have recently developed DECODE, a SFS deconvolution method that accounts for sample-specific sequencing coverage and loss of rare mutations due to mutation calling algorithms.

In this work, we will first test the performance of DECODE against MOBSTER. The mutational data will be simulated using CINner [3] under different assumptions of tumor evolution patterns. The two methods will then be compared on the basis of neutral tail detection and power, number of selective clusters, and the number of mutations and frequencies of each cluster, against ground truth from CINner.

We will then apply DECODE and MOBSTER to 1,900 tumor samples from the International Cancer Genome Consortium (ICGC). To analyze the accuracy of the results, the predicted truncal cluster frequencies will be compared against sample purity. We will then estimate the age of the tumor's most recent common ancestor (MRCA) from purity-corrected neutral tail power and number of mutations in the truncal cluster. Finally, the MRCA age will be compared against clinical information, including patient survival and relapse status, to find whether it can serve as a biomarker for patient outcome. The approach promises to provide a novel method to aid personalized cancer treatment.

1. Caravagna, G. *et al.* Subclonal Reconstruction of Tumors by Using Machine Learning and Population Genetics. *Nature Genetics* **52**, 898-907 (2020).

2. Dinh, K. N. *et al.* Statistical Inference for the Evolutionary History of Cancer Genomics. *Statistical Science* **35**, 129-144 (2020).

3. Dinh, K. N. *et al.* CINner: Modeling and Simulation of Chromosomal Instability in Cancer at Single-Cell Resolution. *bioRxiv* (2024).