Joseph Sayad
Submission II

The project I am working on during my time at the Shendure lab is entitled “Evaluating relationships between ChIP-seq experiments and 3D genome structure.” It involves computational analysis of ChIP-seq (chromatin immunoprecipitation followed by deep sequencing) and ChIA-PET (Chromatin interaction analysis by paired-end tag sequencing) datasets. ChIP-seq experiments reveal the genomic locations of transcription factor binding sites; ChIA-PET experiments couple chromosome conformation capture techniques with ChIP-seq to identify spatial arrangements of enhancers and their target genes. In molecular biology, the binding of transcription factors (i.e., proteins involved in gene regulation and expression) can be represented by motifs (specific sequences of nucleotides that represent a binding site). The processing of ChIP-seq data results in a set of peaks which describe the enrichment of transcription factor binding in various regions of the genome. A subset of these peaks possess motifs, while others lack motifs and are difficult to interpret. The objective of my project is to obtain a better interpretation of these peaks. Our theory is that these peaks are more interesting than appreciated: they may be explained by indirect co-localization of transcription factors, and could shed some light on the phenomenon of chromatin looping interactions between regulatory elements and their target genes.

Initial steps of the project required an understanding of the molecular biology at play, and of two programming languages. I spent my two weeks at lab reading scientific papers, and watching video tutorials on molecular biology. My study was greatly supplemented via discussions with my mentor and the lab environment. My third week involved brushing up on Python and R syntax. Both programming languages are useful in the context of genomics data science – extensive data exploration (via computation) can lead to important biological inferences. That being said, the first step of my project involved the development of a data visualization pipeline. The user feeds two datasets to the pipeline, along with a specific genomic location, and the pipeline produces a plot with a chromatin interaction track, a chromosome position track, and a peak-motif annotation track. The R programming language and Gviz (an R package that provides a structured visualization framework to plot various types of data along genomic coordinates) was heavily used in my pipeline. Python was used for dataset preprocessing. My next steps are to visually inspect the tracks generated for apparent patterns in data, and to speak with a post-doc in the Noble lab that has dabbled in this scientific question.

My expectation for my summer research project is to follow where the data leads me, and to arrive at concrete conclusions. It would be great to validate/invalidate our theory with

positive/negative results. I expect to develop into a better programmer, data analyst, and scientist through the remaining duration of my summer. Based on the results of my exploratory data analysis, the next step of the project will include data analysis via statistical methods and machine learning techniques.