GenBio AI
Sequence Data Engineering Homework

Tyler Katz

March 30, 2025

1 Summary

1.1 Human Reference Genomes

I began this task by first exploring the ATAC-seq and CAGE data available in the ENCODE Functional Genomics database. According to the ATAC-seq Data Standards and Processing Pipeline, alignment files for these experiments are mapped to the GRCh38 reference genome. Given that around 49% of CAGE data mapped to the human genome in ENCODE uses the GRCh38 reference as well, I chose to focus on this subset of data. Therefore, I first implemented a Downloader class to download the GRCh38 reference genome, v24 annotation, and chromosome sizes using the ENCODE API.

1.2 ATAC-seq and CAGE Data

To extract regions of the genome with ATAC-seq and CAGE signal, I referenced the processing pipeline documentation for each assay and decided to focus on merged Irreproducible Discovery Rate (IDR) peak sets. These are peaks merged across replicates or pseudoreplicates (subsampling of reads) to which a statistical procedure is applied to produce a high confidence and reproducible set of peaks. I also chose to filter experiments that were labeled as having "extremely low read depth" or a "low FRiP score" to filter low quality or noisy data. Another approach to subsample the data I considered was using only experiments labeled "Reference Epigenome", however this was only availble for ATAC-seq and not CAGE. I then utilized the API to download the relevant BED files of peaks, as well as experimental metadata. Because experimental conditions vary and have an affect on the peaks identified, I chose to analyze the data at the experimental level and at the assay level. For the experimental level, I extended peaks 10 kb in each direction using pybedtools and merged them for each experiment individually. I then used the genome annotation to find Ensembl transcript IDs that intersect these regions and used the UniProt API to map these IDs to UniProt proteins. To analyze the data at the assay level, for each assay I applied the same procedure, except I concatenated peaks from all experiments before extending, merging, and mapping.

1.3 Code

All code can be found at the Github repository linked here. `https://github.com/katztyler01/genbio_hw`

2 Summary Statistics & Visualizations

2.1 Experimental Level

In the following figures, the data used was at the experimental level, which allowed me to capture the experimental metadata.

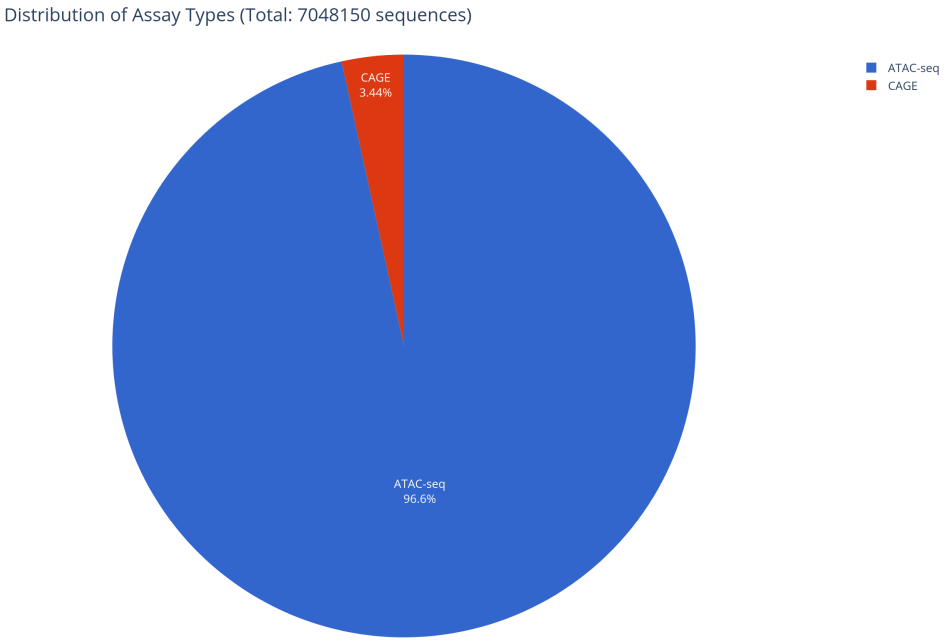

Figure 1: Percentage of Sequences Belonging to Each Assay Type

In Figure 1, we can see at the experimental level there is a total of 7048150 sequences for which a large majority come from ATAC-seq experiments.

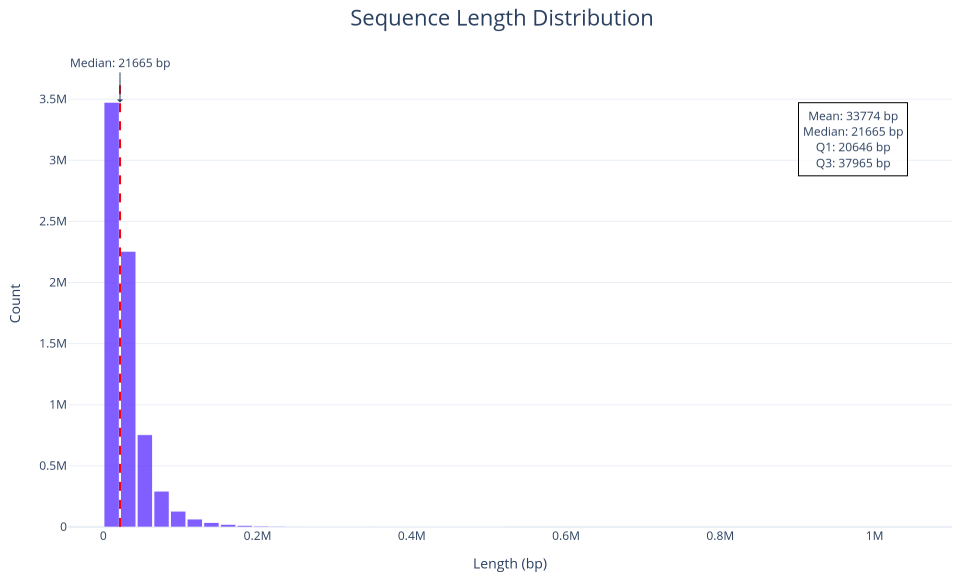

Figure 2: Distribution of Sequence Lengths

In Figure 2, we can see that the median length of sequences at the experimental level is 21665 nucleotides.

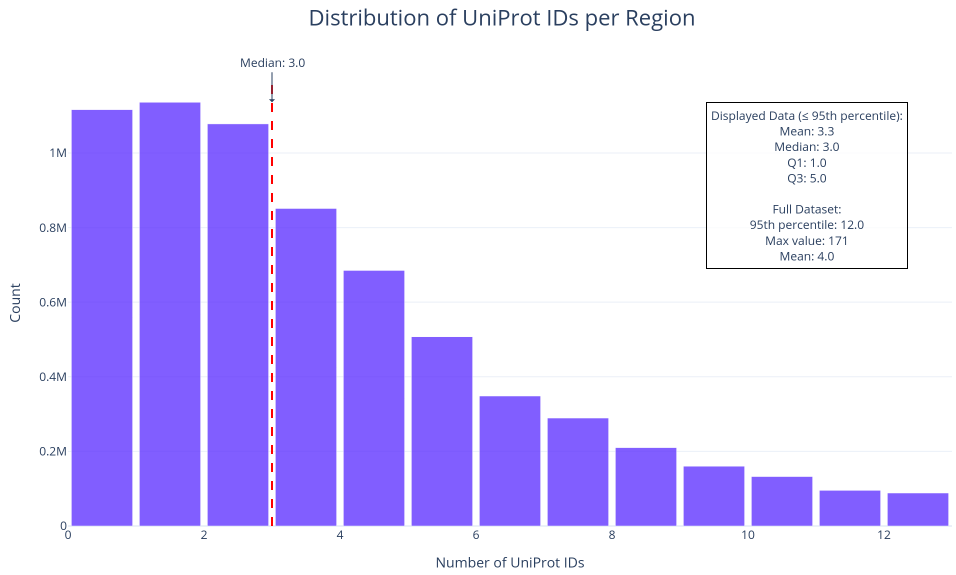

Figure 3: Distribution of Counts of Uniprot Ids Mapped

Figure 3 shows that a median of 3 Uniprot IDs were mapped to each region. Due to a very skewed distribution, some outliers were excluded from the plot.

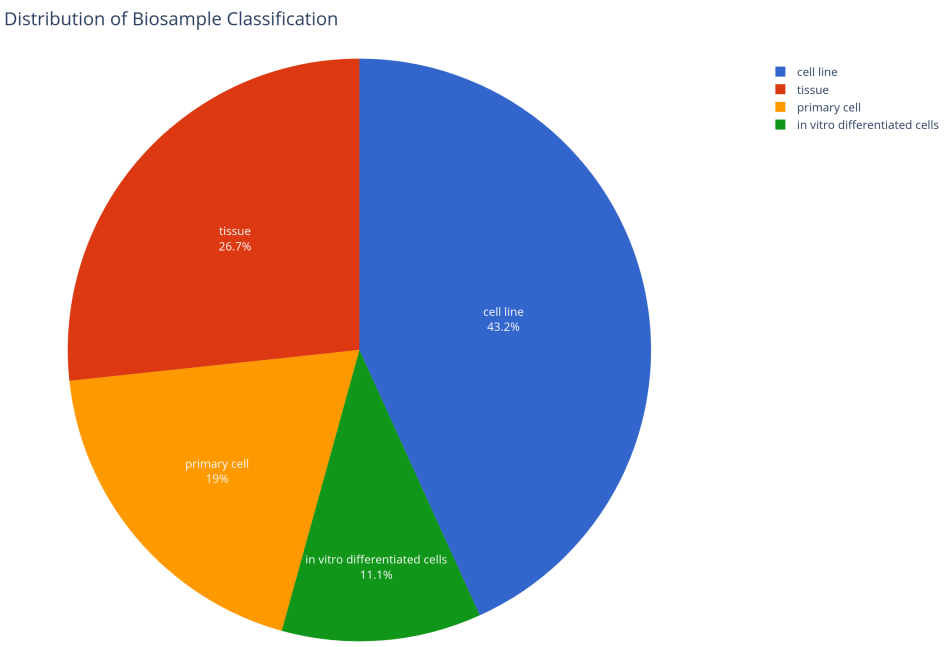

Figure 4: Proportion of Sequences per Biosample Classification

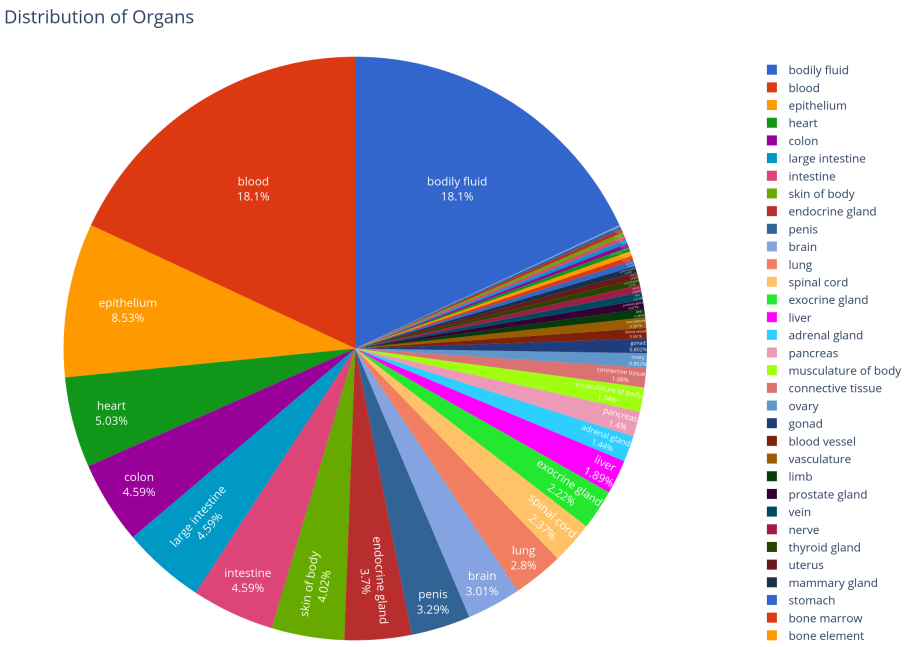

Figure 5: Proportion of Sequences per Organ SLIM

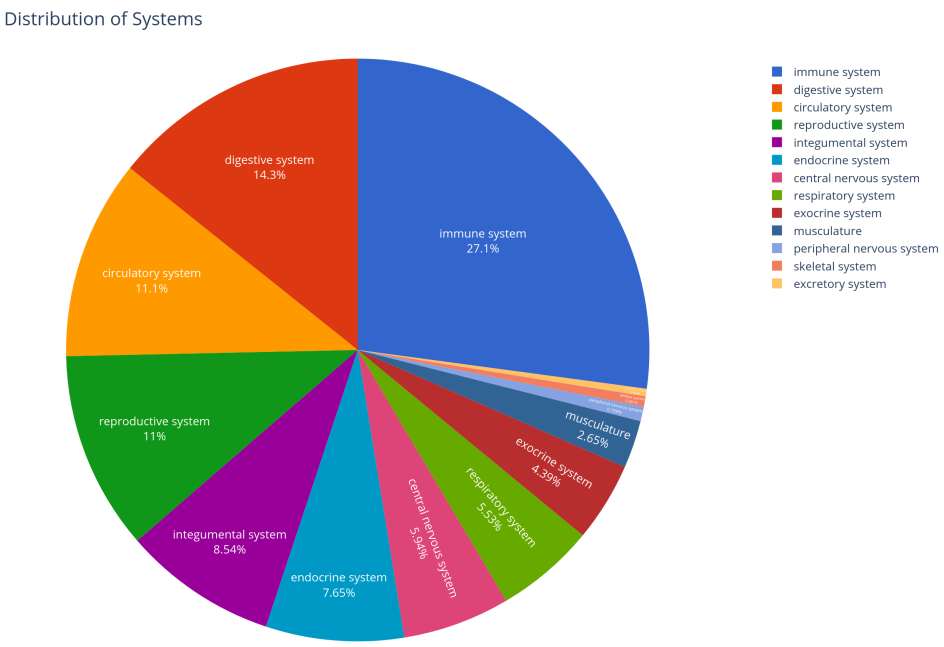

Figure 6: Proportion of Sequences per System SLIM

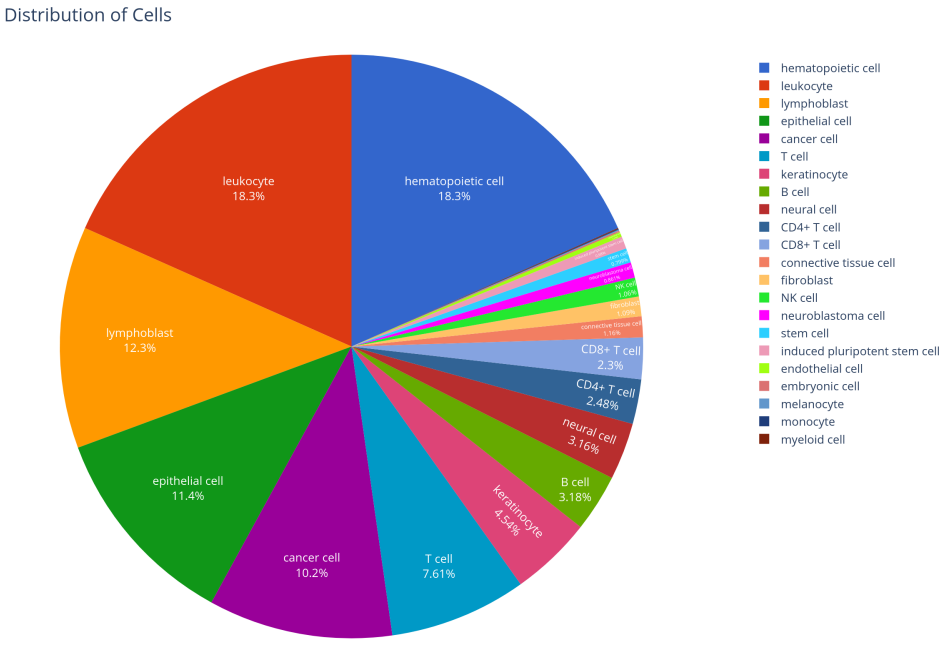

Figure 7: Proportion of Sequences per Cell SLIM

In Figures 4, 5, 6, 7, we can see the proportions of sequences coming from experiments with certain condition annotations. Each SLIM type corresponds to an ontology employed by the ENCODE database to tag experiments with their experimental conditions.

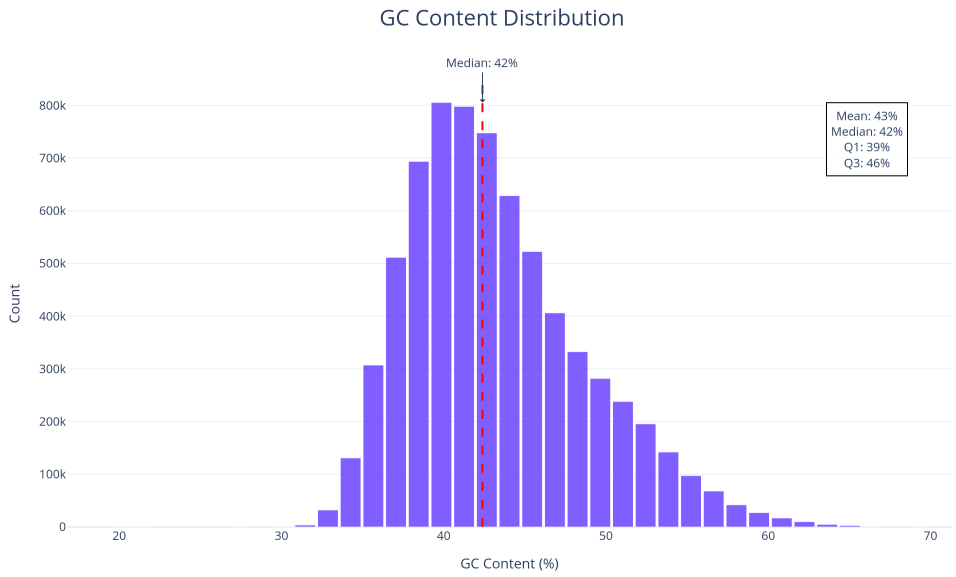

Figure 8: Distribution of GC Content

Lastly, in Figure 8, I've shown the distribution of GC content across sequences given that promoter regions are typically rich in GC nucleotides, unmethylated, and therefore may correspond with open chromatin.

2.2 Assay Level

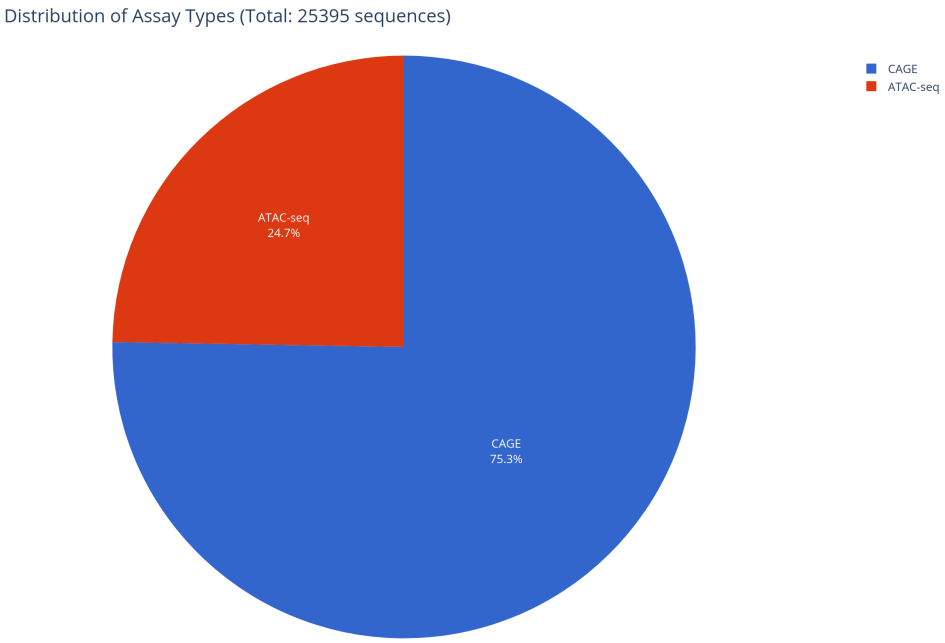

Figure 9: Proportion of Sequences Belonging to Assay

In Figure 9, we can see that now at the assay level, the majority of sequences come from CAGE experiments. However, in Figure 10, we can see the the majority of nucleotides come from ATAC-seq experiments. This indicates redundancy in the regions identified by ATAC-seq across the experimental data in ENCODE and indicates that the merging process joined together larger regions of the genome for the ATAC-seq data.

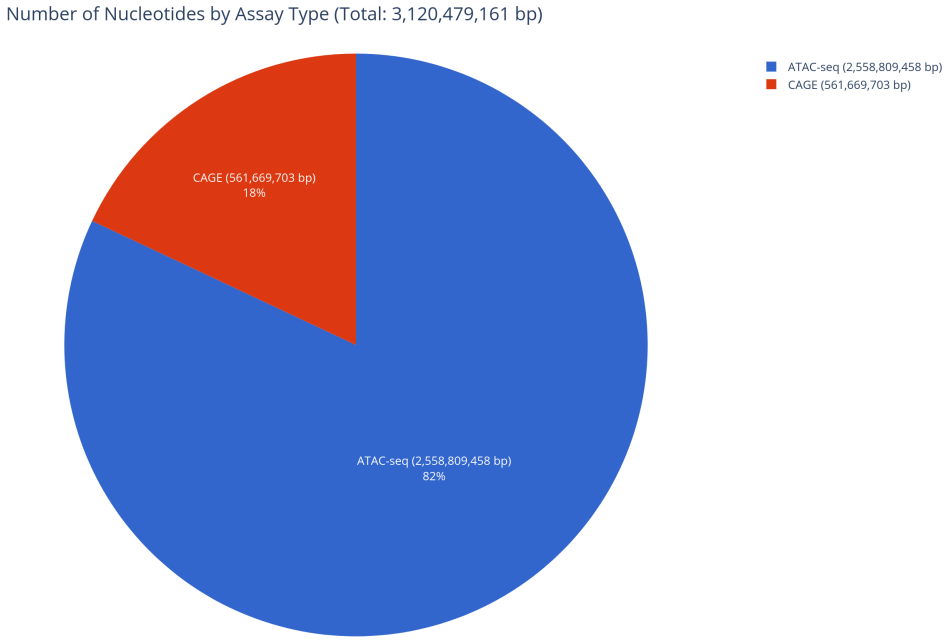

Figure 10: Proportion of Nucleotides Belonging to Assay

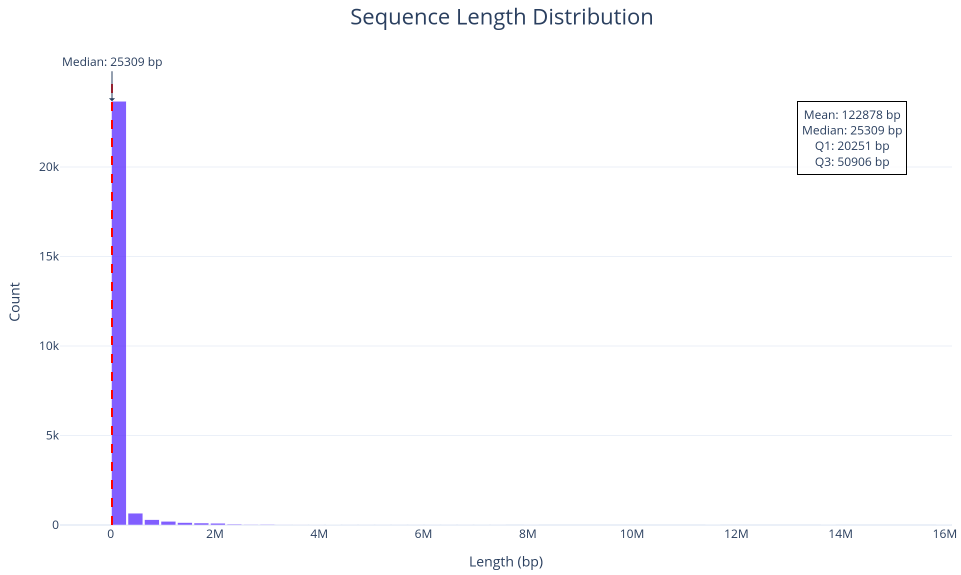

Figure 11: Distribution of Sequence Length

In Figure 11, we see the distribution of sequence lengths is again very skewed and confirms that the merging approach employed results in some larger regions.

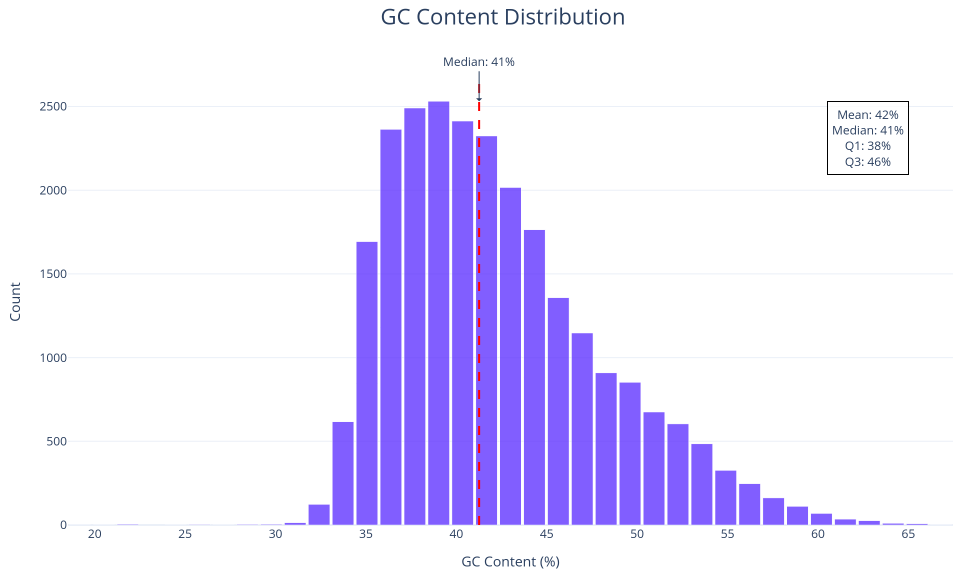

Figure 12: Distribution of GC Content

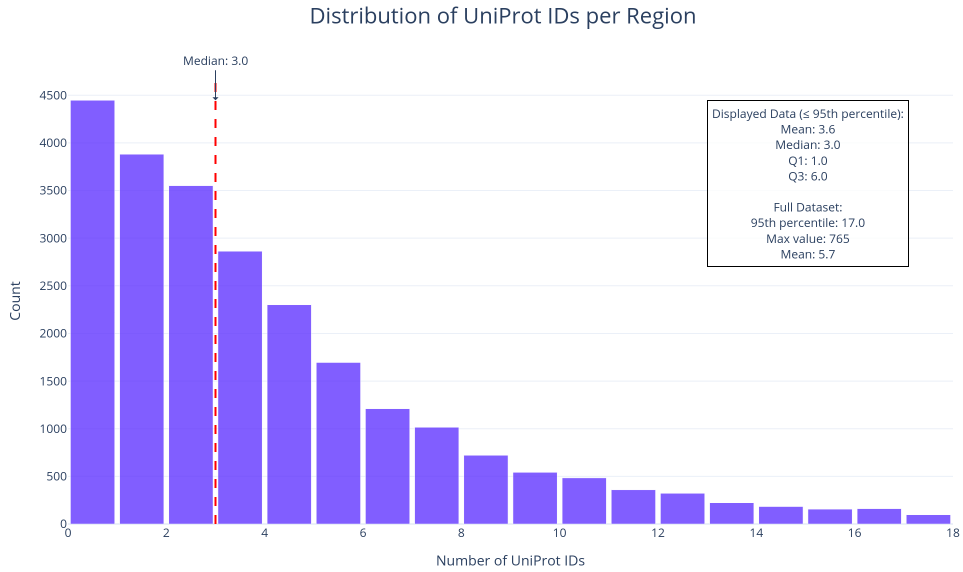

Figure 13: Distribution of Counts of UniProt IDs Mapped

Figures 12, 13, similar to above show the distributions of GC content and UniProt ID counts for the assay level data. While largely similar, we can see that the accumulation of larger regions also resulted in some regions with higher numbers of UniProt IDs mapped with a max count of 765 IDs.

2.3 Discussion

Overall, the two levels of analysis correspond to very different datasets (7048150 sequences vs 25395 sequences). This highlights the need for important considerations and trade-offs in creating a mixed-modality pretraining dataset. Firstly, how to treat experimental conditions must be considered as the utility in many functional genomics assays is to perform some differential analysis between experimental conditions. To maintain this granularity, a more robust data processing pipeline could be leveraged to first cluster experimental data according to some ontology like system or organ SLIMs for which one might expect to see more similarity across experimental profiles. Additionally, a less inclusive approach could be taken in which the intersection of regions across experiments is captured. This approach would capture regions of ubiquitously open chromatin (ATAC-seq) or ubiquitously expressed transcripts (CAGE) across the various experimental conditions. Lastly, a more robust filtering protocol should be employed and could ameliorate some of these issues. Further discussion of the goals of the model would be necessary to make decisions regarding these considerations.

3 Data Splits

After compiling a pretraining dataset for a DNA foundation model, a protocol to split data into train/validation/test sets must be employed to rigorously evaluate a model's ability to generalize. To start, sequences could be split according to chromosome in which certain chromosomes are held out for validation and testing and the rest are used for training. However this approach may result in similar sequences across splits given the data is extracted from ATAC-seq and CAGE experiments and a more robust protocol may need to be employed to capture similarity of the regulatory elements found in these regions across distinct chromosomes. A potential approach to address this might involve representing the sequences of the dataset as nodes in a graph. Edges could be added to fully connect sequences from the same chromosome and inter-chromosomal edges could be added between nodes for which a similarity metric falls above some threshold. The similarity metric chosen could be an efficient sketching algorithm like Mash/MinHash to be able to compute distances quickly. Then, a community detection algorithm could be applied to the graph to find train/validation/test communities that limits inter-split similarity.