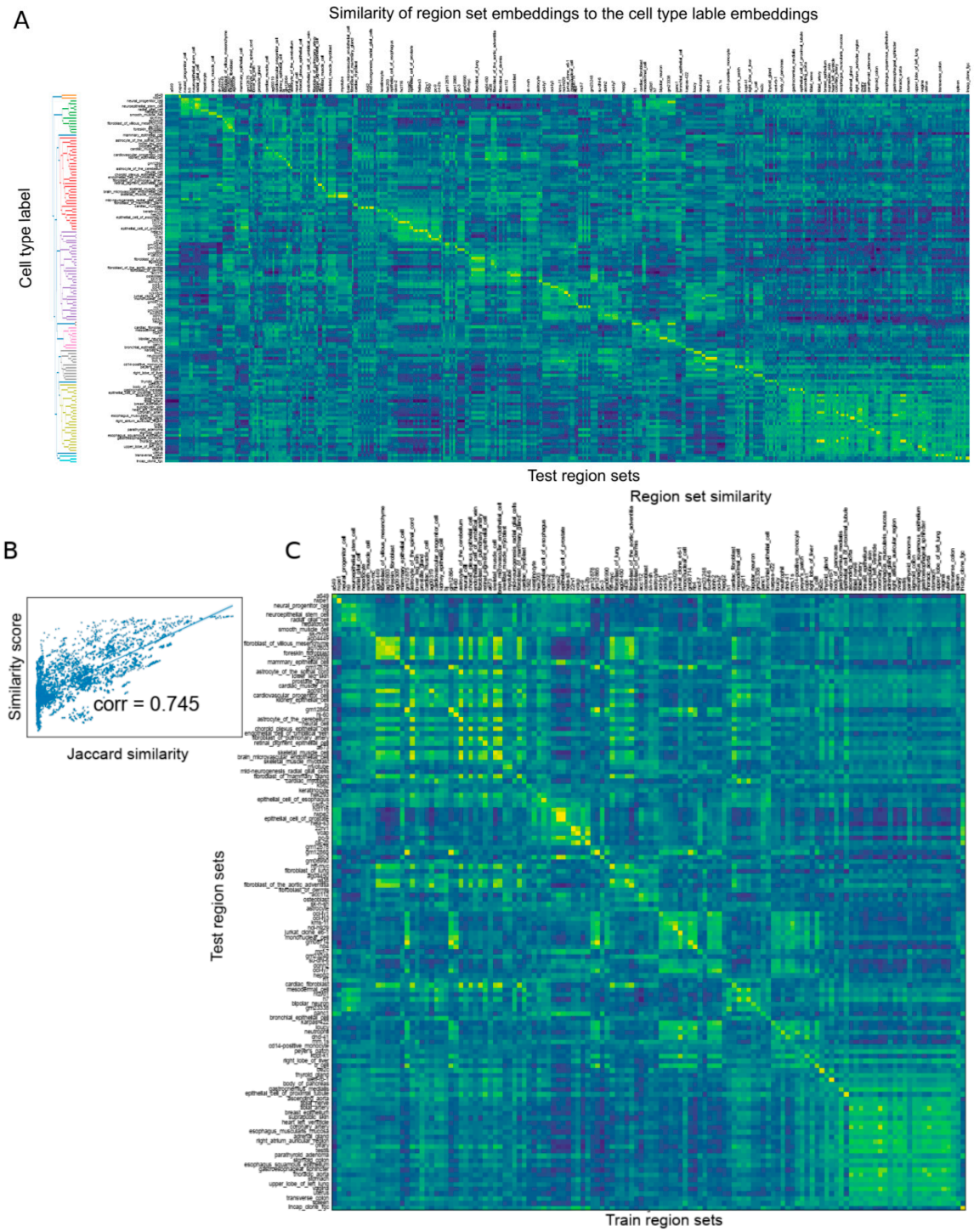
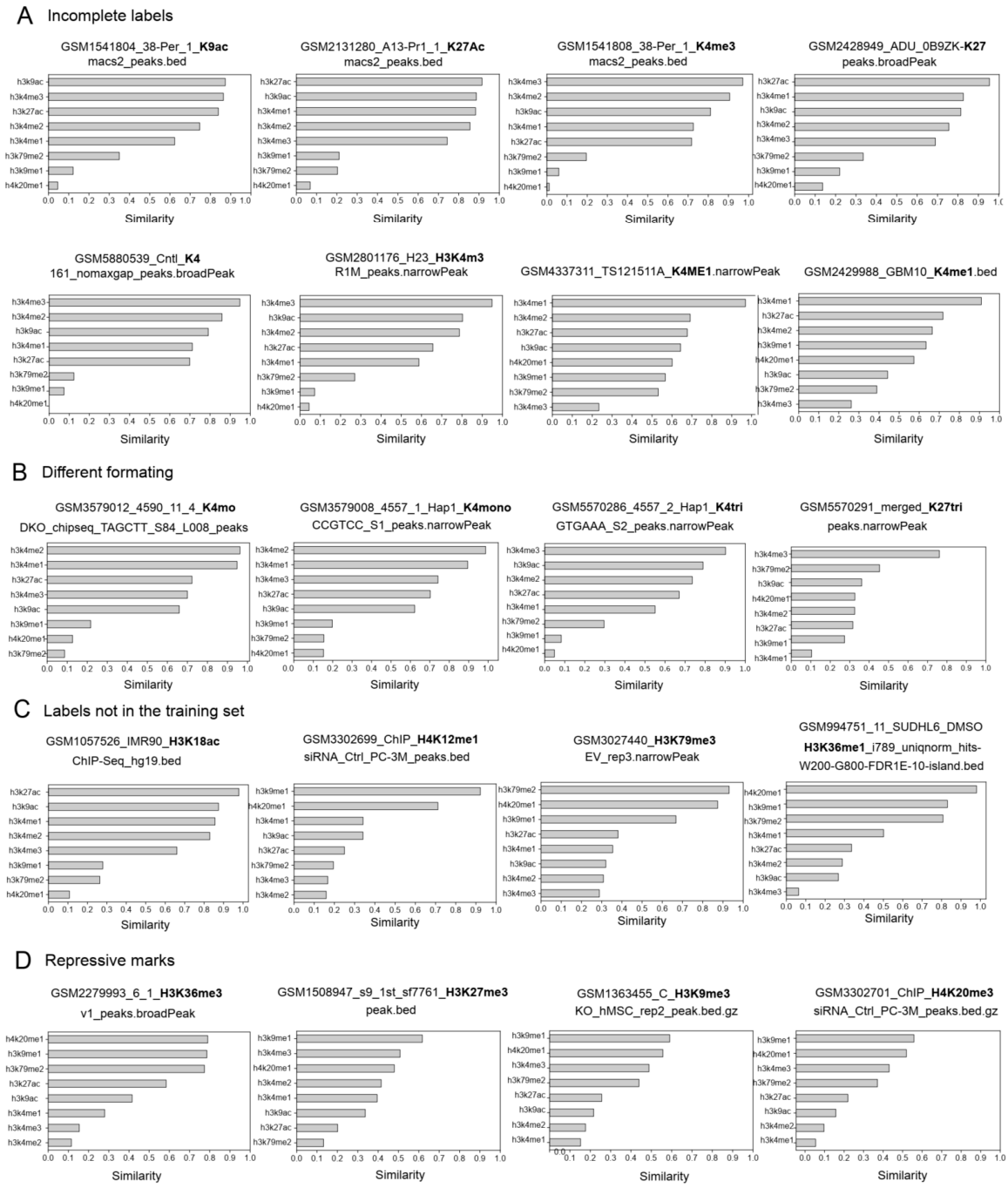
Supplemental figures

Joint representation learning for retrieval and annotation of genomic interval sets



Supplementary Figure S1: Global performance trends. A) Heatmap showing the similarity between the cell type label embeddings and the embeddings of the region sets in the test dataset. The dendrogram depicts relationships among the label embeddings. B) The pair-wise similarity of a subset of BED file embeddings and their Jaccard similarity is correlated.C) The heatmap plot of the similarity between the test region set embeddings and the embeddings of the region sets in the training dataset. The model is trained on the cell type labels.



Supplementary Figure S2: Examples of corrected labels from public data sources. These examples show that the model can be used to A) complete the labels for files with incomplete labels; B) standardize metadata from different sources; C) identify the most related label when the label does not exist in the training model; and D) differentiate activating and repressive marks (for repressive marks, no high-confidence label is found).