Optimization of cell clustering, cluster annotation, and peak calling procedures

Cell Clustering

To fully dissect the heterogeneity within each tissue, we utilized an iterative clustering strategy. The first round of clustering analysis was performed on individual samples to identify initial clusters and candidate peaks. Using the peaks called in the first round of clustering analysis, we then generated a single binary cell-by-peak matrix using cells from all samples and again performed the dimension reduction followed by graph-based clustering to obtain the major cell groups across the entire dataset. We next performed sub-clustering analysis for each of the identified major cell group to identify subclusters.

The resolution parameter in the leiden algorithm was chosen according to the joint consideration of cluster separation (measured by the average silhouette width) and the stability of clustering results. Silhouette width is designed to assess the quality of clusters with a convex shape. Although single-cell clusters are generally non-convex, the spectral clustering technique employed by SnapATAC is a non-linear dimensionality reduction method and is able to transform the clusters into a convex shape suitable for applying silhouette width to measure cluster separation. Indeed, using benchmarking datasets we showed that the silhouette width is extremely useful for selecting appropriate clustering parameters: silhouette width was highly correlated with the adjusted rand index (ARI), which measures the consistency between the clustering result and the ground truth. We also noticed that parameters that produced the optimal number of clusters were generally associated with high clustering stability. We therefore used a criterion of stability greater than 0.85 to filter clustering parameters. The parameters selected according to the criterion above were further tuned with biological considerations. Rationales of parameter determination in sub-clustering analysis are listed in the Table below.

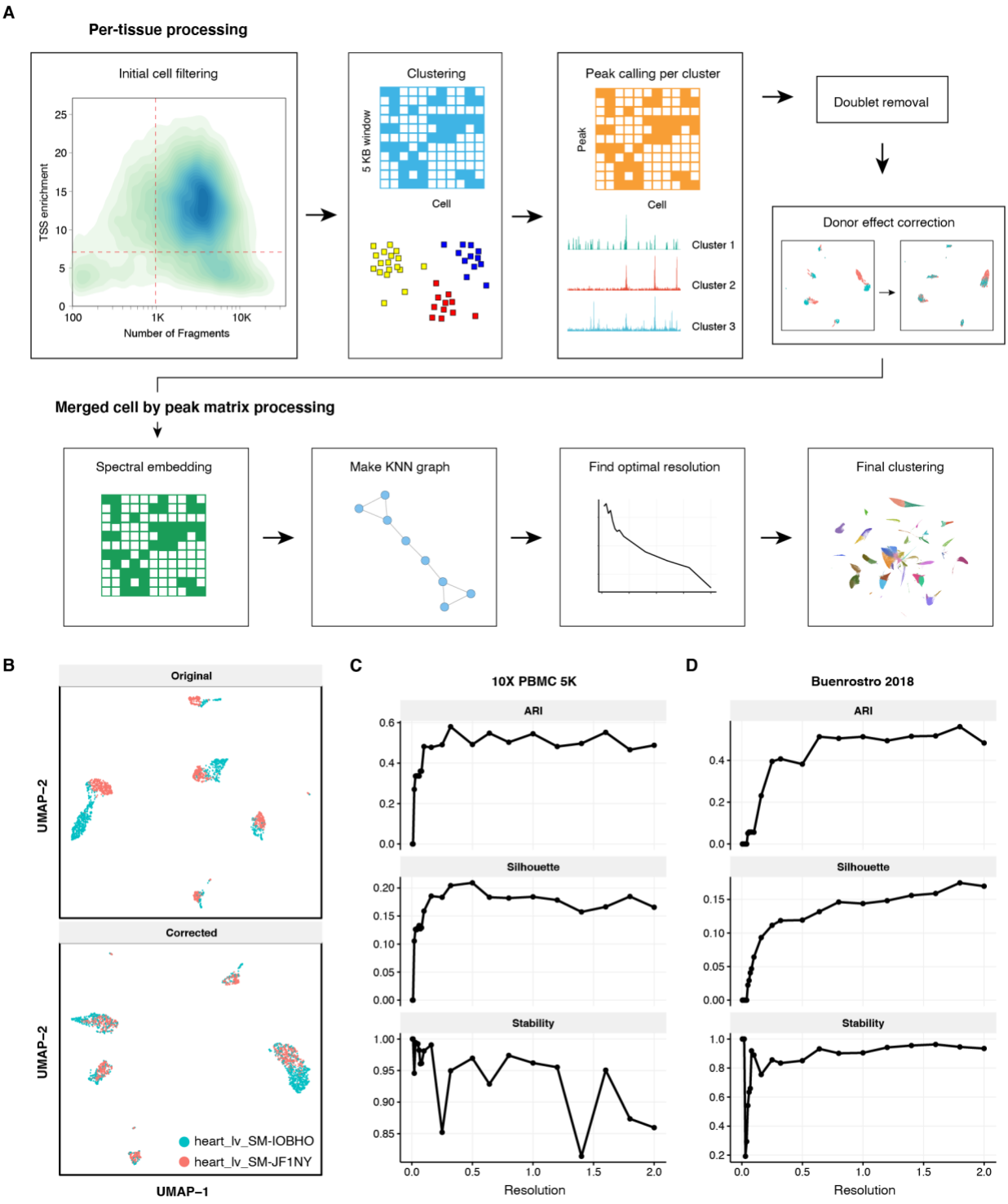
In total, we identified 30 highly stable major cell groups, and then subclustered each to arrive at a total of 111 adult cell types.

Annotation of cell clusters

To annotate the cell clusters, we first curated a set of marker genes from the PanglaoDB (Franzén et al., 2019) corresponding to expected cell types. We aggregated open chromatin fragments from each cluster and utilized the promoter accessibility, defined as RPM of +/- 1kb around TSS, as the proxy for gene activity. We then computed the raw cell type enrichment score as the logarithm of the geometric mean of marker genes' activity. The final enrichment scores were obtained by applying two rounds of z-score transformation, first across cell types and then across cell clusters, on raw enrichment scores. For each cluster, we picked the cell type that showed strongest enrichment to make initial assignments. Finally, we manually reviewed these assignments and made adjustments based on focused consideration of marker gene accessibility in conjunction with information about tissue(s) of origin. For example, the figures below show the marker gene accessibility for neuroendocrine cell types and non-neuroendocrine pancreatic cell types as controls at the genes encoding *GCG*, *INS-IGF2*, *SST*, and *GHRL*.

Optimizing peak calling procedure

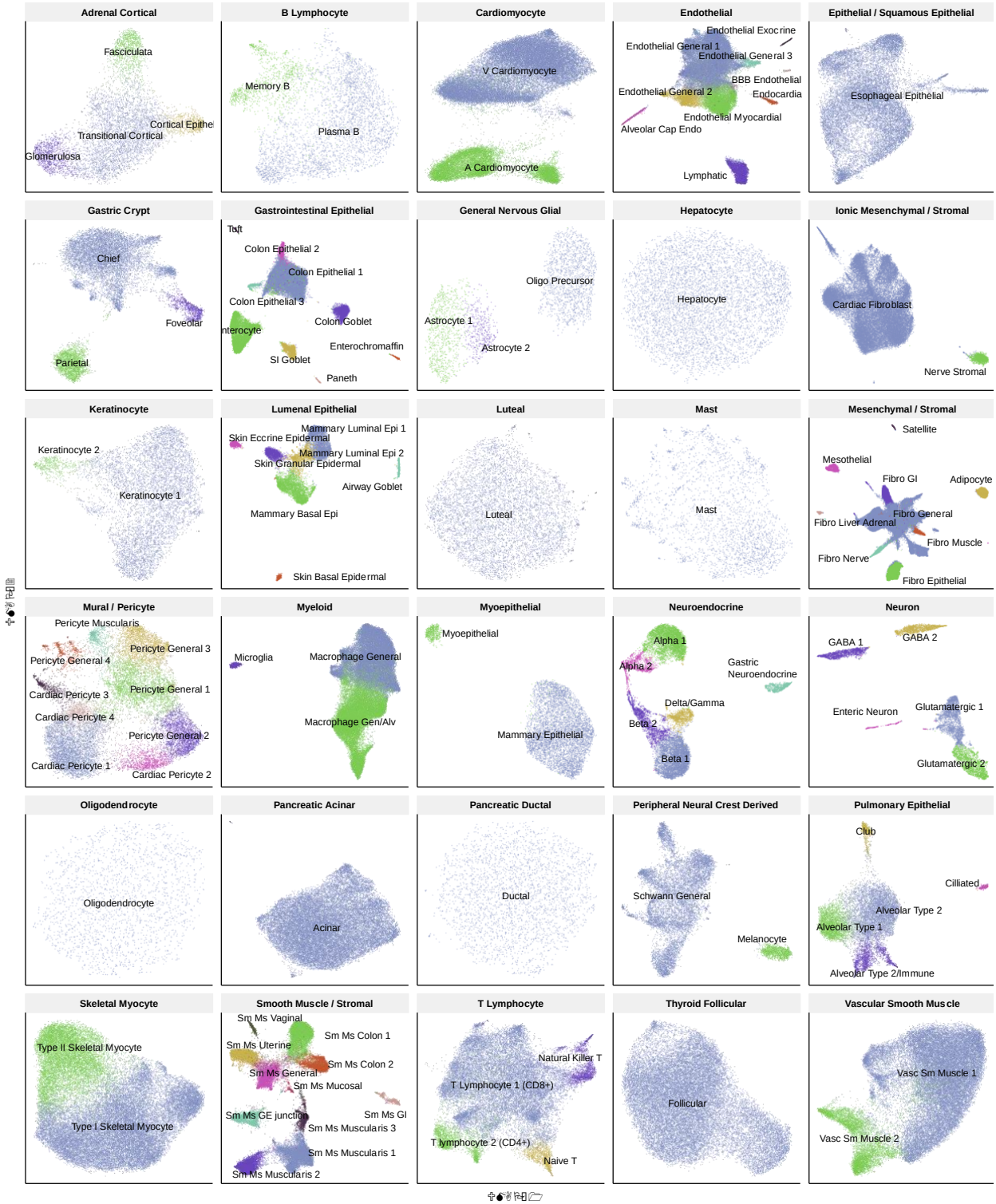
Single-cell assays typically profile different cell types at different sequencing depths due to the varying abundance of cell types in each tissue. Using simulated datasets, we found that adapting peak calling cutoffs to each cell type's sequencing depth increased the sensitivity to detect legitimate peaks in rare cell types and decreased false discovery rate for calling peaks in cell types with high relative abundance. To further improve our overall confidence in identified peaks, we next adopted a peak-filtering protocol (Li et al., 2021) by removing peaks that were not significantly more accessible than background at the single-cell level.



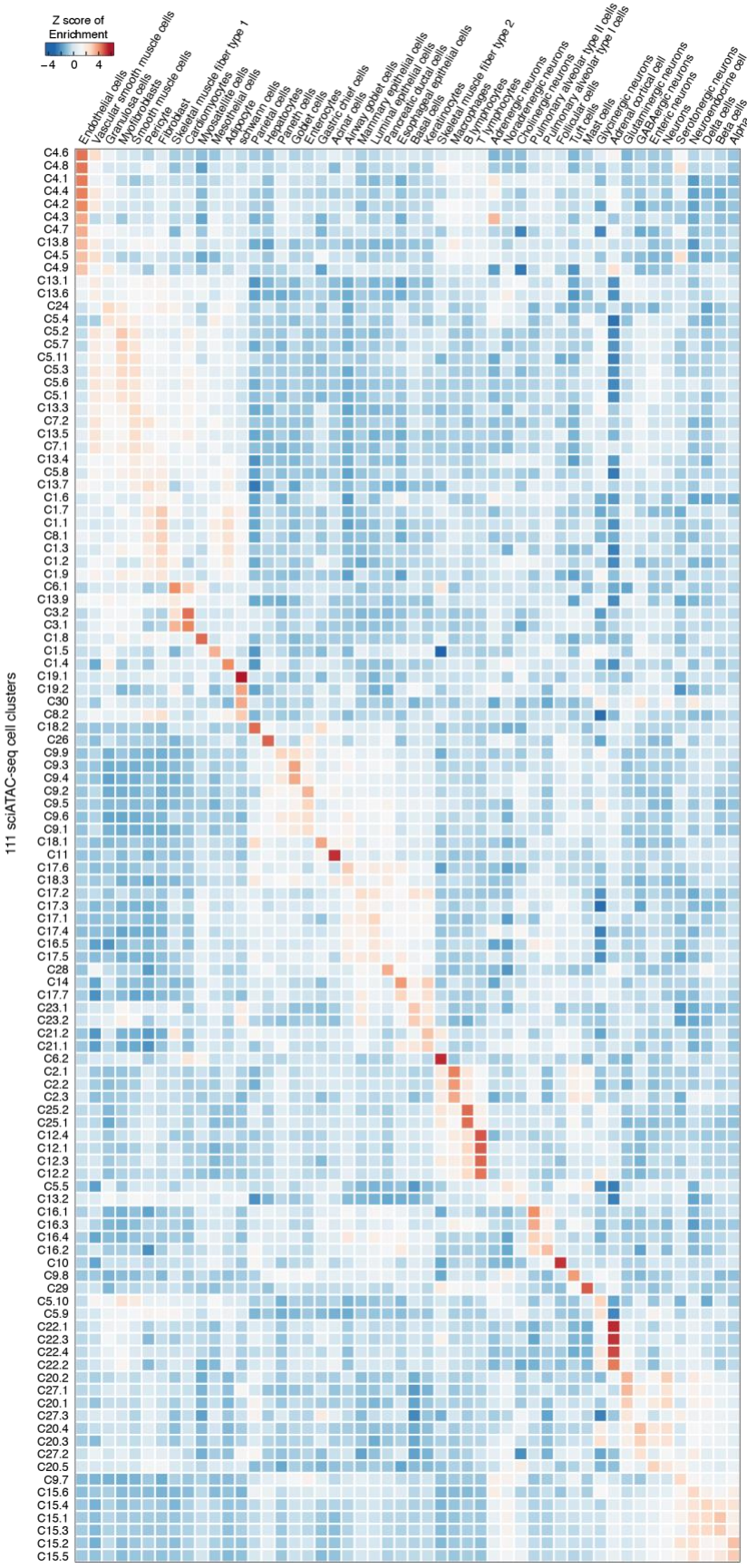
Computational framework for analyzing sci-ATAC-seq data. A) Schematic illustrating the workflow of the analysis pipeline. **B)** Example scatter plots showing the UMAP embedding of nuclei before and after batch correction. Dots with the same color are derived from the same donor or batch. **C,D)** Line plots showing the Adjusted Rand Index (ARI), average silhouette width, and stability of clustering results as a function of resolution parameter in the Leiden algorithm. ARI was computed based the cell annotations from the previous study (Chen et al., 2019). To compute the stability under a particular resolution, five perturbations were conducted on the kNN graph. During each perturbation 2% of the edges were randomly selected and subjected to removal. The clustering was performed on the perturbed graph and the average ARI between different runs were taken as the stability.

Major cluster ID	Resolution with highest silhouette (# of clusters)	Selected Resolution (# of clusters)	Comments
C1	0.005 (5)	0.01 (9)	The highest and second highest silhouette scores were very close, and UMAP embedding inidcated that the second highest score yielded more clusters. The resolution with the second highest silhouette was chosen.
C2	0.003 (2)	0.03 (3)	Original resolution missed Alveolar Macrophage population.
C3	0.04 (2)	0.04 (2)	
C4	0.0075 (2)	0.05 (9)	UMAP embedding inidcated that the resolution with the second highest silhouette was revealed more clusters.
C5	0.01 (2)	0.06 (11)	UMAP embedding inidcated that resolutions with lower silhouette revealed more clusters, so the resolution was adjusted to reveal these clusters.
C6	0.04 (1)	0.05 (2)	Original resolution failed to separate the type 1 and type 2 skeletal myocytes.
C7	0.03 (2)	0.03 (2)	
C8	0.01 (2)	0.01 (2)	
C9	0.07 (9)	0.07 (9)	
C12	0.04 (2)	0.06 (4)	Original resolution failed to reveal T cell subtypes. The resolution with the second highest sihouette was selected.
C13	0.32 (9)	0.32 (9)	
C14	N/A	N/A	
C15	0.01 (2)	0.32 (6)	Original resolution failed to identify Delta/Gamma cells.
C16	0.07 (2)	0.23 (5)	Original resolution failed to identify Alveolar Type 1 cells.
C17	0.003 (2)	0.16 (7)	UMAP embedding indicated that more clusters were present than revealed be highest silhouette value are. The resolution with the second highest silhouette was chosen.
C18	1 (8)	0.07 (3)	The original resolution appeared to lead to overclustering as some of the resulting clusters shared highly similar marker genes.
C19	0.05 (2)	0.05 (2)	
C20	0.25 (5)	0.25 (5)	
C21	0.1 (2)	0.1 (2)	
C22	0.64 (5)	0.16 (4)	The original resolution appeared to lead to overclustering as some of the resulting clusters shared highly similar marker genes.
C23	0.1 (2)	0.1 (2)	
C25	0.25 (2)	0.25 (2)	
C27	0.5 (3)	0.5 (3)	

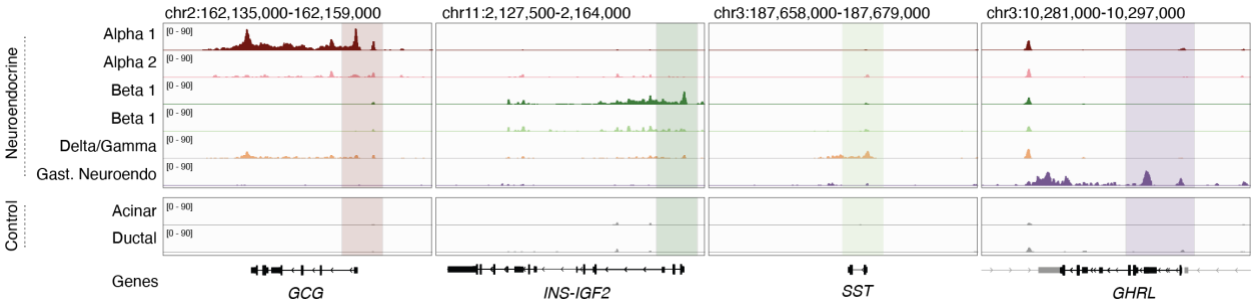
Resolution parameters used in the sub-clustering analysis. Major cell groups that do not have subclusters were omitted from the table, including C10, C11, C14, C24, C26, C28, C29, and C30.



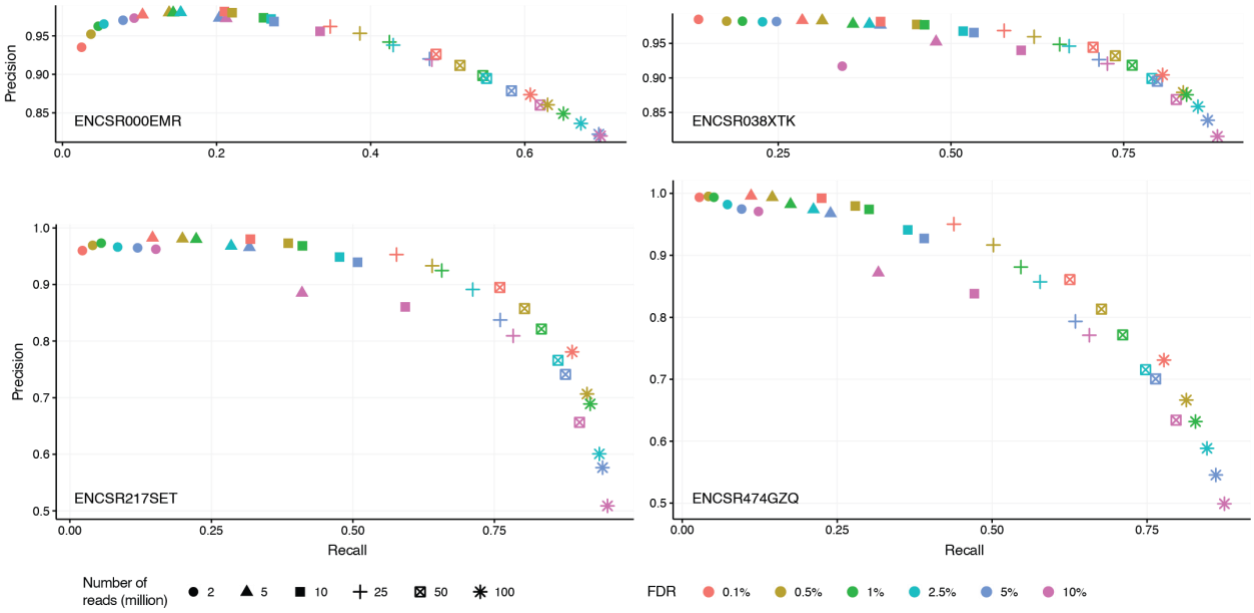
Iterative clustering analysis of the 30 major cell groups. Each scatter plot shows the UMAP embedding of nuclei from one of 30 major cell groups. Subclusters are indicated by different colors. 22 out of 30 major cell groups had more than one subcluster.



Evidence supporting the annotation of 111 cell clusters. Heatmap representation showing the marker gene enrichment of cell types. The marker genes were downloaded from the PanglaoDB (Franzén et al., 2019).



Example of focused lineage marker gene consideration for cell type annotation. Genome browser tracks show chromatin accessibility for six neuroendocrine cell types subclustered from the neuroendocrine major cell group. Non-neuroendocrine pancreatic cell types are included as controls. Neuroendocrine cell marker genes encoding *GCG, INS-IGF2, SST,* and *GHRL* are indicated in black, neighboring genes are indicated in gray. Transcription start site(s) of the indicated genes are highlighted. Gast. Neuroendo = Gastric Neuroendocrine.



Peak call benchmarking using different FDR cutoff and down-sampling rate. Each plot shows the precision (y-axis) and recall (x-axis) of peaks called by MACS2 (Zhang et al., 2008) under different combinations of FDR (color) and number of reads (shape). In all cases, the ground truth was taken as the peaks produced by the ENCODE consortium using all reads.