Briefings in Bioinformatics, 2025, **26(6)**, bbaf646
https://doi.org/10.1093/bib/bbaf646
Problem Solving Protocol

scAED: a framework for mapping the enhancer state at single-cell resolution

Avinash Veerappa[1], Jai Chand Patel[1], Sushil Shakyawar[1], Sankarasubramanian Jagadesan[1], Chittibabu Guda[1,2,*]

[1]Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, S 45th St, Omaha, NE 68198, United States
[2]Center for Biomedical Informatics Research and Innovation, S 45th St, University of Nebraska Medical Center, Omaha, NE 68198, United States

*Corresponding author. Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, S 45th St, Omaha, NE 68198 United States. Fax: 402-559-3400; E-mail: babu.guda@unmc.edu

Abstract

Cells within the same tissue exhibit heterogeneity, influenced by several factors including enhancers. Enhancers regulate precise spatial and temporal patterns of gene expression by switching between active or inactive states. Therefore, it is imperative to capture this dynamic dimension of enhancers that regulate genes by initiating transcription across large distances in each cell. Existing enhancer databases rely on data from pooled-cell sequencing or single-cell clustering. In both approaches, the dynamics of cell-specific enhancer states are concealed due to aggregation of data across multiple cells, which limits our understanding of cellular heterogeneity. To address this, we developed a novel computational framework to extract the chromatin activation state of each enhancer in each cell using sc-Multiome and matched snATACseq-snRNAseq datasets that resulted in the development of single-cell Active Enhancer Database (scAED). scAED is a perpetual project that currently contains active enhancers in one brain region (putamen) and two physiological states (control and healthy) of the pancreas comprising of 21 cell types. We have catalogued 2 291 987 unique active enhancer regions from a total of 34 124 988 active enhancer regions. Besides the characterization of active enhancers at single-cell resolution, scAED also introduces several formative advancements including the characterization of bidirectional enhancers, capture of trans-acting elements and their precise binding coordinates, incorporation of strand-specific information, and establishment of a unique enhancer ID system. These innovations collectively represent significant novel advancements in the field of enhancer biology. We will continue to grow scAED as new datasets become available on other organs, tissues, and cell types. We anticipate that the widespread adoption of this platform would accelerate generation of testable hypotheses on novel regulatory mechanisms to understand the molecular underpinnings of health and disease.
Data accessibility: Data can be queried and downloaded from the scAED website at https://www.gudalab-rtools.net/scAED/. Additionally, the custom scripts for the scAED framework are available in the GitHub repository at https://github.com/GudaLab/scAED.

Keywords: enhancer; single-cell; database; scAED; multiome; snATACseq; snRNAseq; transcription factor

Introduction

Enhancers are distal DNA elements located in distinct chromosomal regions that participate in cis-regulation of their target gene transcription to enhance gene expression [1]. Enhancers, including super-enhancers and shadow-enhancers, collaborate in complex networks to regulate target gene expression [2]. Active enhancers exhibit specific chromatin organization patterns, such as an open chromatin conformation and histone modifications, while showing unique cofactor interactions that reveal unique motif sequence characteristics and regulatory mechanisms [3, 4]. Enhancers play a crucial role in cell and organ function by precisely enhancing their target genes, from development to senescence. Recent studies have underscored the significant impact of enhancer dysfunction on various diseases, with disruptions arising from genetic variations, chromosomal rearrangements, or epigenetic modulations contributing to congenital disorders, cancers, and complex diseases [5, 6] with integrative systems-omics approaches highlighting substance use disorders and neuroadaptations [7].

Known enhancers listed in the existing databases pose significant background noise, imprecision, and other inconsistencies [8–15] because most of them were identified through pooled-cell sequencing technologies, resulting in a loss of sub-anatomical region-specific and cell-type-specific information. Even the enhancers derived from the recent single-cell technologies average out their active status because they were profiled at the cluster level. Moreover, enhancer databases are often disorganized, with no universal IDs assigned to them, leading to lack of a standardized naming conventions. Finally, none of the existing databases provide information on the transcription factors (TFs) bound (motifs) within the enhancers and their precise TF binding site (TFBS) coordinates.

Profiling genome-wide active enhancers at single-cell resolution has been challenging due to the difficulty in capturing their

Received: May 29, 2025. **Revised:** September 29, 2025. **Accepted:** November 14, 2025

unpredictable positions and the lack of high-throughput capture technologies to profile them [4]. In contrast, the rise of multiome and single-nucleus Assay for Transposase-Accessible Chromatin with high-throughput sequencing (sn-ATACseq) offers an excellent opportunity for their identification in the open chromatin regions (OCRs) at single-cell resolution. Sequenced reads from either Multiome or snATAC-seq experiments can detect regions of accessible chromatin, including active enhancers. Additionally, regions of accessible chromatin that bear active enhancers can be used to perform TF foot printing to identify TFBS, their motifs, and their precise TFBS coordinates.

We have developed the first active enhancer database at single-cell resolution, named the Single Cell Active Enhancer Database (scAED) using sequence reads from snMultiome or snATAC-seq data. This was achieved by developing custom scripts, optimizing parameters along with utilizing specialized third-party tools and scalable cloud-based workflows such as CloudATAC to streamline large-scale ATAC-seq processing [16]. scAED also provides active enhancers in healthy and diseased cell states, potentially serving as a platform for identifying candidate enhancers contributing to cell and organ dysfunction. In this study, active enhancers (AEs) are defined as enhancer regions characterized by open and accessible chromatin, along with the binding of trans-acting elements, such as TFs. We achieved several firsts by developing scAED that include: identifying enhancers at single-cell resolution, characterizing bidirectional enhancers, capturing trans-acting elements in active enhancers, establishing strand specificity for both TFs and target genes, and developing and maintaining unique enhancer ID nomenclature. We will continue to develop and update scAED database as relevant data become available on other organs, tissues, and cell types.

Materials and Methods

Datasets

In this study, we utilized a combination of tools and in-house scripts to effectively capture, profile, annotate, and categorize AEs in OCRs of the brain and pancreas. Figure 1 outlines the steps involved in each processing stage during the development of this framework. The datasets are from the putamen region (MM_566) and primary motor cortex (MM_410) of the human brain, sourced from the BICCN project. They contain snRNA-seq and snATAC-seq data derived from the same tissue slide blocks of the donor. Another dataset includes samples from Pancreatic Ductal Adenocarcinoma (PDAC) and Normal Adjacent (NA) samples from a PDAC donor, generated using 10x Genomics multiome library (GSE241896). We used 10x Genomics' Cell Ranger Count and Cell Ranger ATAC pipelines to process FASTQ files with the GRCh38p14 reference, performing alignment, de-duplication, filtering, feature-barcode matrix generation, and gene expression and chromatin accessibility analysis. For Multiome samples, we used Cell Ranger-ARC Count to align with GRCh38p14, filtering, barcode counting, peak calling, and counting of both ATAC and GEX molecules. We used Cytograph qc [17] for analyzing gene expression (GEX) data from the brain's putamen and primary motor cortex region of the BICCN project [18, 19], while Seurat [20] was used for analyzing GEX data from pancreatic head region to examine PDAC sample, and NA tissue samples from the study by Pratt, Ma [21]. SnapATAC2 [22] was employed to analyze chromatin accessibility (ATAC) data in both cases and to characterize and annotate cell clusters by integrating GEX and ATAC. Figure 1 shows the graphical representation of each major step adopted in this framework.

Data preprocessing and quality control of snATACseq

We generated fragment files from the processed bam files obtained from the cell ranger output to indicate chromatin accessibility, which was imported using 'snap.pp.import_data()'. This function computed basic quality control (QC) metrics, such as counts of fragment size (Supplementary Fig. 1a and b), TSS enrichment and unique fragments per cell (Fig. 2a), while storing the data in an AnnData object. We analyzed the fragment size distribution to identify key features: nucleosome-free regions, mononucleosomes, and di-nucleosomes using 'snap.pl.frag_size_distr()' (Supplementary Fig. 1a). TSS enrichment scores were computed for each cell using 'snap.metrics.tsse()', enabling the identification of high-quality cells. Based on TSS enrichment and fragment counts (Fig. 2a), we filtered cells that did not meet the predefined quality thresholds using 'snap.pp.filter_cells()', specifying minimum and maximum counts of fragments and a threshold for TSS enrichment.

Feature selection and dimensionality reduction

The cell-by-bin matrix was constructed to record insertion counts using 'snap.pp.add_tile_matrix()'. Feature selection was performed to identify the most accessible genomic features using 'snap.pp.select_features()', where parameters such as the number of features and potential blacklist or whitelist files were specified. We then removed potential doublets using a scrublet algorithm ('snap.pp.scrublet()'). We applied spectral embedding for dimensionality reduction ('snap.tl.spectral()'), followed by UMAP for two-dimensional visualization purposes ('snap.tl.umap()').

Clustering, imputation, and annotation

We conducted graph-based clustering using the Leiden algorithm by building a k-nearest neighbor graph ('snap.pp.knn()') and then identifying densely connected subgraphs or clusters ('snap.tl.leiden()') (Fig. 2b). A cell-by-gene activity matrix was generated using 'snap.pp.make_gene_matrix()' for cell cluster annotation. We utilized the MAGIC algorithm [23] ('sc.external.pp.magic()') for data imputation and smoothing, facilitating the visualization of marker genes and enhancing data interpretation.

snRNAseq analysis—data preprocessing and quality control

For the snRNAseq analysis before integration with snATAC-seq data, we employed a detailed workflow encompassing data QC, clustering, and annotation processes, utilizing the Cytograph qc [17]. We processed snRNAseq samples and their replicates using the 'cytograph qc' command, leveraging DoubletFinder to calculate a doublet score for each cell. We further filtered cells based on unique molecular identifiers (UMIs), percentages of unspliced RNA, and doublet scores. Cells were excluded if they had fewer than 800 UMIs, <40% unspliced RNA, or a doublet score below 0.3.

Dimension reduction and clustering analysis

For clustering, we utilized the Cytograph package [17]. We executed the 'cytograph process' command with parameters configured to optimize data handling without removing low-quality or doublet cells. This setup included 'min_umis: 0', 'doublets_action: None,' and 'features: variance.' We used principal components analysis (PCA) on the most variable genes to construct a k-nearest neighbors' graph, which was input for Louvain clustering via the

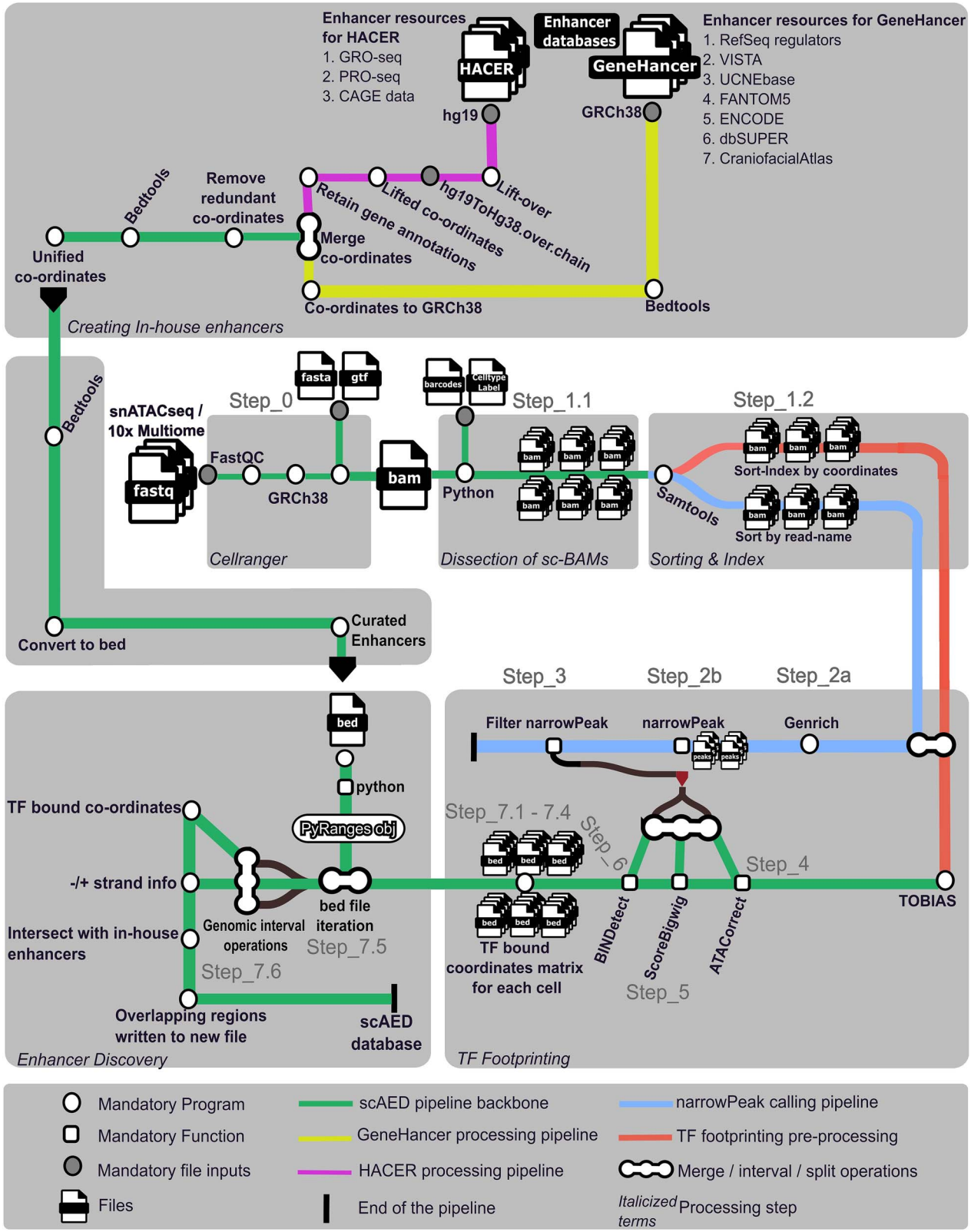

Figure 1. Overview of the workflow for developing the scAED framework. This figure illustrates the sequential steps involved in resolving enhancer data to the single-cell level. It includes performing TF footprinting, intersecting OCRs with in-house curated enhancers, identifying active enhancers, and compiling a cell-specific enhancer database. This comprehensive pipeline facilitates the systematic identification and analysis of active enhancers at single-cell resolution. The step numbers listed above each processing step correspond to the script numbers provided in the GitHub repository.

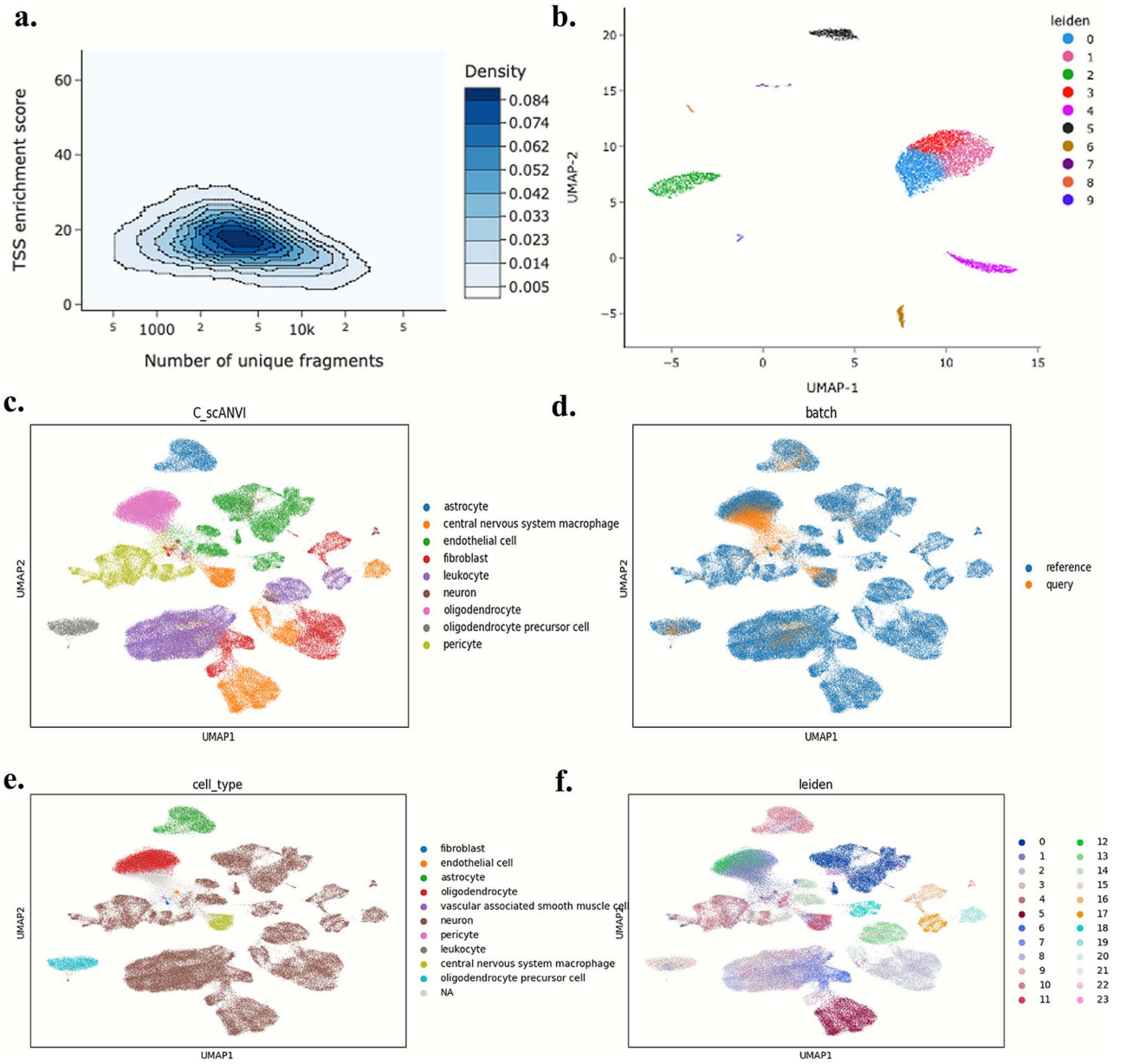

Figure 2. Quality control, model training, and integration of scATAC-seq and scRNA-seq data. a. Displays a density plot of TSS enrichment scores versus the number of unique fragments. b. Presents a UMAP visualization of clustered cells based on scATAC-seq data. c and d illustrate the integrated UMAP embedding of scATAC-seq and scRNA-seq data, with cells colored by cell type and batch. e and f show the cell type label transfer, and UMAP plots colored by cell type and Leiden clusters, respectively.

scikit-network. All cells passing QC were pooled and clustered together. Additional QC at the cluster level involved removing clusters with high mitochondrial UMI percentages or clusters flagged with multiple specific auto-annotations.

Clusters and subclusters that did not express at least one statistically enriched gene were merged with the nearest cluster based on the t-SNE, using false discovery rate-corrected p-values calculated against a null distribution. The final t-SNE and dendrogram were generated based on Euclidean distances between cluster means of the 2000 most variable genes, normalized and standardized using sci-kit-learn's StandardScaler.

Cell annotation strategies

We employed two distinct cell annotation approaches, trinarization [17] for the brain and GPTCelltype [24] for the pancreas. The trinarization approach was chosen for brain tissue due to its effective utilization of a well-curated three-state reference atlas, enabling accurate annotation of cell populations. In contrast, the pancreas lacks a definitive reference atlas, particularly under physiological conditions such as PDAC. Therefore, we applied the generative AI-based GPTCelltype approach, which allowed for robust prediction and annotation of diverse cell types in the pancreatic dataset.

Trinarization approach

For the putamen brain region, cell annotation was performed using the trinarization method as described by the Linnarsson Lab [17]. This approach involved transforming gene expression data into a three-state system (trinary values: −1, 0, +1) to reflect gene repression, no change, or activation, respectively. The trinarization was applied to the gene expression matrix, where gene expression values were first normalized and then categorized based

on predefined thresholds. Subsequently, trinarized profiles were matched to a reference atlas, allowing for robust and automated cell type annotation. The method's application resulted in high-confidence assignments of cell identities across the dataset, leveraging the contrast between gene activation and repression states to delineate distinct cell populations.

GPTCelltype approach

For the cells of the pancreatic head region, cell annotation was conducted using the GPTCelltype framework as described by Hou *et al.* (https://github.com/Winnie09/GPTCelltype) [24]. For each cell cluster, the top 10 differentially expressed genes were selected as input features to guide the annotation process. Both GPT-4o (13 June 2023 version) and GPT-3.5-turbo-0125 (13 June 2023 version) models were employed for initial testing. Due to the higher reproducibility observed with the GPT-3.5-turbo-0125 version, this model was chosen for final annotations. The annotation process was iterated 20 times to ensure high-confidence predictions, allowing for robust identification of cell types across the dataset. This iterative approach provided consistent and reliable cell type assignments, leveraging the power of large language models in biological data interpretation.

To assess GPTCelltype's accuracy, we first benchmarked it on the putamen snRNA-seq dataset, using the BICCN trinarization labels as ground truth. For each cluster we provided the top 10 differentially expressed genes and ran GPT-4o and GPT-3.5 20 times each; the consensus label from GPT-3.5, chosen for its higher run-to-run reproducibility was then compared with the reference. Across all putamen clusters the model achieved 83.3%–90.1% agreement. Misclassifications were confined to fine-grained interneuron subtypes (ex. VIP+ and LAMP5+ cells). Although GPTCelltype sometimes failed to assign these specific subtype labels, it nevertheless classified the cells correctly as neurons, and the affected clusters were flagged for manual curation. Because the pancreas lacks such deep subtype hierarchy, we next applied the same GPT-3.5 pipeline to the pancreatic dataset, again iterating 20 times per cluster to obtain high-confidence consensus annotations.

snATACseq-snRNAseq integration

We integrated snRNAseq data with snATACseq data obtained from the anatomically dissected regions of the same donor. Using scanpy and scvi-tools, we transferred cell type annotations from the snRNAseq cell clusters to the corresponding snATACseq data. We adjusted the model to accommodate batch effects and account for highly variable genes ('sc.pp.highly_variable_genes()') and trained a variational autoencoder (VAE) model using 'scvi.model.SCVI()' to perform the integration (Supplementary Fig. 1c and d). The integration process was validated by comparing predicted cell type labels with known annotations, ensuring consistency and accuracy in data interpretation (Fig. 2c and d) (Supplementary Fig. 1c and d). Finally, the cell type labels were transferred from GEX to ATAC clusters (Fig. 2e and f).

Setting up active enhancer database framework

Single-cell BAM processing and TF footprinting

We developed a custom script to utilize an integrated single-cell dataset from a '.h5ad' file into an AnnData object. A predefined list of cell types and cell labels was specified, and the script iterated over each cell type in the list, filtering the AnnData object to retain only the cells in the specified list. It extracted the barcodes of these filtered cells, which were unique identifiers for each cell. Through this process, the script effectively segregated and saved the barcodes of cells belonging to different cell types. The barcode file corresponding to each cell type was generated and used as an input in the next step. For each barcode file, the script initialized a dictionary to hold file handles for each barcode. It read the barcodes from the file and prepared output BAM file paths for each barcode. It then opened a new BAM file for each barcode, setting up the file handles. The script proceeded to read the original BAM file and iterated over each read, extracting the barcode from the CB (Cell Barcode) tag. If a read's barcode matched one in the dictionary, the read was written to the respective BAM file. Processing all reads resulted in the creation of filtered BAM files, each containing only the reads associated with a specific cell barcode, thus enabling extraction of single-cell BAM files (scBAM) for each cell of all cell types. These scBAM files were then processed with SAMTools [25] to sort and index to optimize them for efficient data retrieval and downstream processing. Peak calling was conducted using Genrich [26] with parameters tailored to identify narrowPeak OCRs. The resulting narrowPeaks were further filtered for high confidence, which applied quality thresholds to remove artifacts and low-confidence peaks.

Next, TOBIAS ATACorrect [27] function was utilized to normalize the ATAC-seq by performing bias correction of ATAC-seq reads in OCRs. ScoreBigwig [27] was further used to calculate footprint scores from previously corrected cut sites. Further, BINDetect [27] was used to identify bound TFs in the OCRs based on scores, sequence and motifs in each cell. The data consolidation was then automated by organizing TF bound genomic coordinate sites into respective bed files and the multiple TF BED files consisting of their binding coordinates were merged into single files named after their cell IDs. The merged BED files were intersected with the in-house developed enhancer resource via BEDTools [28].

Creating an in-house enhancer resource

We collated the vast repositories of enhancer regions from the HACER [13] and GeneHancer [29] databases, which provided a rich collection of enhancer data from more than 300 cell lines. Regarding the HACER database, enhancers were retrieved based on the catalog of active enhancers derived from GRO-seq, PRO-seq, and CAGE experiments from the 265 cell lines in hg19 coordinates. List of 265 cell lines can be accessed here https://bioinfo.vanderbilt.edu/AE/HACER/download.html. Regarding the GeneHancer database, access to enhancer coordinates was granted after making a written request to GeneCards. Enhancers were provided to us in Hg38 format containing experimental data from ChIP-seq, DNase-seq, ATAC-seq, P300, POLR2A, CAGE, ChIA-PET, GRO-seq, STARR-seq and MPRA from the following eight data sources. Motivated by the hypothesis that if an enhancer validated in one cell line is also located within an OCR in another cell type, it suggests active involvement in regulating gene expression across both cell types, potentially influencing similar biological functions or processes, we created an in-house enhancer resource. This extensive dataset served as the foundational resource for constructing a robust in-house enhancer resource. Each enhancer region was meticulously converted to the GRCh38 human genome build using LiftOver [30], ensuring that the genomic coordinates were updated to align with the latest genome annotation standards. Crucially, during this conversion process, the original enhancer-gene associations present in HACER [13] database were preserved. This retention was vital as it maintained the functional context of each enhancer, linking them to potential gene targets based on experimental data. With the GRCh38-aligned enhancers and their associated gene links, an in-house database was created,

tailored to support detailed analyses of enhancer activity across various cell types. The final step in this workflow involved intersecting these enhancer sequences with TF footprints for each cell type being studied. This intersection identified enhancers that were not just theoretically active but were also empirically engaged in transcriptional regulation via TF binding, providing a powerful tool for probing the functional dynamics of enhancers and understanding their roles in gene regulatory networks across different cellular contexts.

Further, to compare the proportion of contribution of enhancers from GeneHancer and HACER databases individually, we quantified, for each cell type, the number of overlapping enhancers in GeneHancer versus HACER databases and summarized their relationship using the ratio HACER/GeneHancer and log2(HACER/GeneHancer) (0 = equal representation; negative values indicate GeneHancer-enriched). In the Supplementary Table 1, we provide a comparison of enhancer results obtained from different reference databases (HACER and GeneHancer). Across 64 cell types, results are highly consistent in relative trends with median ratio ~ 0.257 and median log2 ratio ~ −1.96, indicating GeneHancer generally contains more elements per cell type, as expected for a larger, more integrative catalog (Supplementary Table 1). The range of ratios spans 0.135–0.493 (e.g. OGC_2 = 0.135; SMC = 0.493), showing that while absolute coverage differs between databases, the directionality is coherent across diverse neuronal (VIP/PVALB/SST/LAMP5) and glial (oligodendrocyte/microglia/astrocyte) classes (Supplementary Table 1). Aggregating two independent enhancer resources yielded concordant relative cell-type patterns, with predictable differences in absolute counts due to catalog scope. This supports the robustness of our annotation strategy to database choice.

Creating scAED web application

The active enhancers intersected between the in-house enhancer resource and OCR of each single-cell was made as a database. We developed the database using the Shiny application [31] and used several R packages for the implementation to enhance functionality and user experience. The 'shiny' package provides the core interactive web application structure, while 'DT' was utilized to incorporate interactive tables that allow users to explore and manipulate the displayed data efficiently. Customization and theming were achieved through 'shinythemes', ensuring a user-friendly interface. Data management and file system access were handled by 'shinyFiles', which facilitated efficient navigation and data handling within the application. Visualization capabilities were significantly enhanced by 'ggplot2' [32] and 'ggpubr' [32], which enabled the creation of graphs and plots. Lastly, the 'data.table' package was used for high-performance data processing, ensuring rapid handling of large datasets. The database is hosted on a web server having 256 GB RAM, 12 CPUs, and a 10 Gb bandwidth internet connection.

Pairwise cell–cell similarity analysis based on enhancer overlap

We compared active enhancers across single cells using a custom Python pipeline to identify common and unique chromosomal coordinates ('Chromosome,' 'Start,' 'End') from the enhancer data. Pairwise enhancer similarity was quantified with a custom Python (3.11) workflow that operated directly on the single-cell files containing enhancer identifiers. Within each cell, duplicate enhancer names were removed so that every enhancer contributed exactly once to the analysis. All unordered cell pairs $(\mathbf{n}(\mathbf{n}-1)/2)$ were generated with *itertools.combinations*, and their similarity was computed as the Jaccard Index

$$\text{Jaccard}\ (i,j) = \frac{|\ \text{Ei} \cap \text{Ej}\ |}{|\ \text{Ei} \cap \text{Ej}\ |}$$

where E_i *and* E_j are the sets of enhancers in cells i and j, respectively. Computations were parallelized with the multiprocessing module, and the resulting Jaccard scores were multiplied by 100 to yield percent overlap values. These values were written to an $n \times n$ similarity matrix whose diagonal was fixed at 100% (self-comparison) and whose off-diagonal entries captured inter-cell overlap. The matrix was visualized as a heatmap (Fig. 3) with diagonal elements highlighted in red for clarity and exported alongside the raw data and a cumulative distribution plot of pairwise overlaps. Descriptive statistics for the overlap distribution (mean, median, quartiles, min–max, σ) are provided in Table 2. Before performing multivariate and clustering analyses, the matrix was z-score standardized with scikit-learn's *StandardScaler* to place all similarity features on a common scale (Supplementary Fig. 2a–e).

Clustering enhancers based on common and unique chromosomal coordinates

To determine the optimal number of clusters, we applied the Elbow Method ('KMeans().fit()' with 'inertia_' parameter) (Supplementary Fig. 3a) and Silhouette Method ('silhouette_score()') (Supplementary Fig. 3b), followed by K-means clustering ('KMeans(n_clusters = optimal_clusters)'). We visualized the clusters using t-SNE ('TSNE(n_components = 2)') (Fig. 3a) and performed hierarchical clustering ('linkage()' with Ward's method) (Fig. 3b) and provide insights into cellular heterogeneity and regulatory dynamics at single-cell resolution.

Mitigating sparsity to reliably perform single-cell enhancer calling

We controlled scATAC sparsity at four checkpoints, (i) we retained only high-quality nuclei (≥5 k unique fragments, TSS-enrichment ≥ 10), verified by the expected mono- and di-nucleosome periodicity in the fragment-length distribution (Supplementary Figs 1a and 4a). (ii) To rule out depth-driven artefacts we re-sampled every library; this produced comparable curves across single cells. (Supplementary Fig. 4b) (iii) Genrich peaks were kept only if supported by ≥ 5 fragments; a correlation heat-map shows total reads and peak count remain linearly coupled (r = 0.99), confirming reproducible discovery (Supplementary Fig. 4c). (iv) Within those peaks, a TF footprint was accepted only when its TOBIAS BINDetect score lay above the cell-specific 95th-percentile inflection in the FLD score-density plot (bimodal separation) (Supplementary Fig. 4d) and the site overlapped a HACER or GeneHancer enhancer. (v) correlation between sequencing depth and TF occupancy in enhancer regions for several brain regions (Supplementary Fig. 7). We tested whether enhancer annotation (mean % TF occupancy within annotated enhancers) is driven by sequencing depth (read count per BAM). Across eight cortical regions (a) primary motor cortex (Pearson r = 0.051), (b) primary auditory cortex (Pearson r = 0.069), (c) inferior temporal gyrus (Pearson r = 0.086), (d) temporal polar cortex (Pearson r = 0.100), (e) frontal agranular insular cortex, (f) parastriate cortex, (Pearson r = 0.119) (g) rostroventral DFC (Pearson r = 0.108), and (h) primary visual cortex (Pearson r = 0.091), scatter plots with linear fits show no meaningful correlation between depth and TF occupancy (Supplementary Fig. 7). Samples spanning low, moderate, and high sequencing depths exhibit a similarly broad range of TF

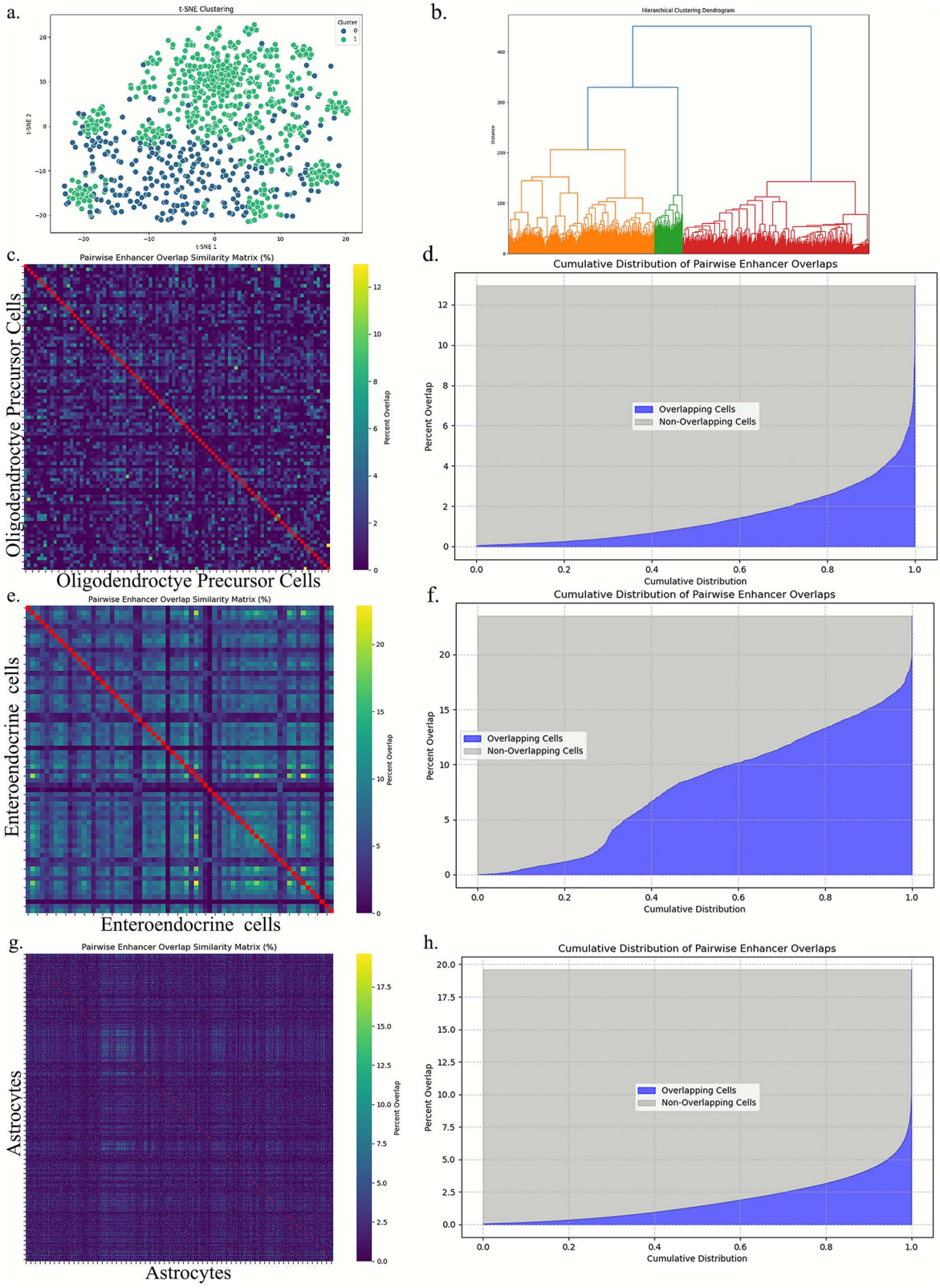

Figure 3. Clustering and pairwise cell heterogeneity matrix of enhancers. a. Visualizes t-SNE clustering, displaying two well-separated clusters. b. Displays a hierarchical clustering dendrogram, identifying three distinct clusters of astrocytes based on common and unique chromosomal coordinates ('chromosome,' 'start,' 'end'). c, e, and g display the enhancer overlap similarity matrix as a percentage for OPCs from the putamen brain region, enteroendocrine cells from the NA pancreatic head, and astrocytes from the putamen brain region. Panels d, f, and h illustrate the distribution of pairwise enhancer overlaps in 5% bin sizes for OPCs, enteroendocrine cells and astrocytes, respectively.

Table 1. Shows the total cell count, cell types, total enhancers, singleton enhancers, and singleton bi-directional enhancers for the putamen of brain and pancreatic head of pancreas

Organ	Condition	Region	Total cell count	Cell types	Total enhancers	Singleton enhancer counts	Singleton bi-directional enhancers	Total non-redundant	% of bi-directional Enhancers
Brain	Healthy	Putamen	210	Pericytes	120 382	42 753	19 667	62 420	31.51
			93	Oligodendrocyte precursor cells	67 323	29 794	14 645	44 439	32.96
			3216	Neurons	1 726 055	100 144	44 445	144 589	30.74
			1224	Astrocytes	925 154	87 269	39 487	126 756	31.15
			1388	CNS macrophages	915 172	106 690	47 320	154 010	30.73
Total			**6131**	***n* = 5**	**3 754 086**	**366 650**	**165 564**	**532 214**	**-**
Pancreas	Normal adjacent	Pancreatic head	66	Endothelial cells	271 702	74 244	35 058	109 302	32.07
			107	Other cells	226 866	60 137	28 591	88 728	32.22
			146	Satellite cells	3980	3682	2105	5787	36.37
			1229	Goblet cells	4 326 588	141 403	63 759	205 162	31.08
			173	T cells	464 070	78 437	38 007	116 444	32.64
			1589	Acinar cells	5 628 411	149 449	67 187	216 636	31.01
			735	Neurons	2 165 283	125 141	56 480	181 621	31.10
			882	Epithelial cells	3 873 248	130 270	59 336	189 606	31.29
Total			**4927**	***n* = 8**	**16 840 986**	**999 143**	**456 751**	**1 113 286**	**-**
Pancreas	PDAC	Pancreatic head	600	Macrophages/Microglia	3 019 527	136 505	63 179	199 684	31.64
			257	Other cells	662 767	94 132	44 006	138 138	31.86
			69	Enteroendocrine cells	421 991	77 631	37 594	115 225	32.63
			279	Fibroblasts	1 767 993	138 830	63 785	202 615	31.48
			402	T Cells	1 017 150	103 123	48 787	151 910	32.12
			501	Acinar cells	2 809 187	127 240	58 647	185 887	31.55
			306	Neurons/Neuroendocrine	1 096 383	108 944	50 321	159 265	31.60
			324	Cancer cells	2 734 918	139 789	64 327	204 116	31.51
Total			**2738**	***n* = 8**	**13 529 916**	**926 194**	**430 646**	**1 356 840**	-
Grand total			**13 796**	***n* = 21**	**34 124 988**	**2 291 987**	**1 052 961**	**3 002 340**	**-**

The rows in BOLD signify the respective totals for the sub-groups.

occupancy values, indicating that depth within our study range does not constrain detection (Supplementary Fig. 7). Thus, enhancer TF-occupancy estimates are not depth-dependent in a way that would undermine annotation validity. Even after applying the depth-related quality filters, we still capture the expected enhancer heterogeneity across cell types, supporting that scAED inferences reflect biology rather than sampling noise.

Results

Active enhancers from a single cell

In this study, AEs are defined as enhancer regions characterized by open and accessible chromatin, allowing binding of trans-acting elements, such as TFs. We selected two organs—brain (putamen) [18, 19] and pancreas (pancreatic head) Pratt, Ma [21] and three physiological conditions (Healthy, Tumor, and Normal Adjacent (NA) to Tumor) for this project, based on the availability of scMultiome and scATAC-seq datasets. Our comprehensive analysis revealed a total of 34 124 988 AEs from 13 796 cells across 21 distinct cell types (Table 1). Specifically, in the putamen region of a healthy donor brain, we identified 3 754 086 AEs distributed among five cell types. In contrast, the pancreatic head region yielded 16 840 986 and 13 529 916 AEs from eight cell types in NA and tumor tissue samples, respectively. Upon removing the redundant enhancer coordinates (coordinates that occur more than once), we found a total of 366 650 unique active enhancers in all celltypes of putamen, 999 143 from normal adjacent pancreatic head, and 926 194 from PDAC tumors (Table 1). Further, we wanted to compare active enhancer counts between the pancreatic head of NA tissue (n = 16 840 986) and the putamen (n = 3 754 086) to identify fold difference. To account for differences in cell numbers between these two tissues, we normalized active enhancer counts by dividing the total enhancers by the number of cells in each tissue. The putamen had 613 enhancers per cell, while the NA pancreatic head tissue had 3418 enhancers per cell, resulting in a 5.58-fold difference per cell in the NA pancreatic head compared to the putamen. Our framework, in this way, demonstrated the capability of extracting the enhancers from diverse platforms, such as multiome and snATACseq. Besides this tractability, our framework can also capture the active enhancers across organs.

Enhancer clustering analysis

To further group the cells based on common and unique enhancer coordinates, a clustering analysis of all profiled enhancers across each cell was conducted. Supplementary Fig. 3 shows the results of the clustering analysis for astrocytes of putamen. The Elbow Method identified an inflection point around three clusters (Supplementary Fig. 3a), suggesting an optimal cluster count in this range. However, the Silhouette Method, with a maximum score at two clusters, indicated a clearer separation at this level (Supplementary Fig. 3b). This two-cluster separation was further supported by t-SNE visualization, which revealed two well-defined clusters (Fig. 3a). Hierarchical clustering, as depicted by the dendrogram, identified three distinct clusters, marked by orange, green, and red branches, suggesting a more complex

Table 2. Pair-wise comparison of cells within each cell type, highlighting the minimum and maximum active enhancers coordinate overlap. This analysis also includes metrics such as mean, median, standard deviation, 25th percentile and 75th percentile

Organ	Condition	Region	Cell types	Overlap of active enhancers within cell types (in %)						
				Min	Max	Mean	Median	Standard deviation	25th Percentile	75th Percentile
Brain	Healthy	Putamen	Pericytes	0.02	21.11	1.34	0.69	1.54	0.26	1.96
			Oligodendrocyte precursor cells	0.04	12.95	1.49	1.01	1.48	0.34	2.25
			Neurons	0.03	36.57	1.95	1.29	1.91	0.44	2.97
			Astrocytes	0.02	19.6	1.86	1.40	1.68	0.48	2.81
			CNS_macrophages	0.02	21.38	1.15	0.66	1.25	0.27	1.64
Pancreas	Normal adjacent	Pancreatic head	Endothelial cells	0.01	22.86	5.33	5.18	3.17	2.77	7.42
			Non cells	0.02	14.26	3.94	3.72	2.42	2.15	5.48
			Satellite cells	0.04	24.17	6.05	5.66	3.43	3.62	7.99
			Goblet cells	0.01	33.68	7.19	7.60	4.04	3.82	10.31
			T cells	0.02	26.21	5.63	5.51	2.71	3.72	7.39
			Acinar cells	0.01	37.14	7.48	8.03	4.15	4.15	10.59
			Neurons	0.01	26.27	6.59	7.05	3.69	3.70	9.36
			Epithelial cells	0.01	43.86	8.32	9.12	4.47	4.80	11.71
Pancreas	PDAC	Pancreatic Head	Macrophages/Microglia	0.01	29.81	7.02	7.24	4.36	3.20	10.50
			Unknown cells	0.01	17.29	4.48	4.16	2.89	2.24	6.34
			Enteroendocrine cells	0.01	25.81	6.64	7.14	4.04	3.08	9.77
			Fibroblasts	0.01	28.50	6.50	6.98	4.38	2.16	9.83
			T cells	0.01	39.19	5.53	5.39	3.29	2.86	7.92
			Acinar cells	0.01	28.41	8.58	9.59	5.23	3.48	12.91
			Neurons/Neuroendocrine	0.01	25.81	6.64	7.14	4.04	3.08	9.77
			Cancer cells	0.01	31.92	8.10	7.70	6.20	1.91	13.13
Total			***n* = 21**	-	-	-	-	-	-	-

structure than indicated by the Silhouette and t-SNE analyses (Fig. 3b). The Elbow Method's identification of three clusters aligns with this, indicating the presence of a third, distinct subset of cells with unique enhancer activity. These findings underscore the necessity of profiling the enhancers and cataloguing them at single-cell resolution to fully capture the intricate enhancer heterogeneity within the tissue.

Active enhancers display high levels of cellular heterogeneity across cell types

We computed pairwise similarity between cells to estimate the level of heterogeneity based on shared enhancer regions (enhancer overlaps), ensuring that redundant enhancers were removed so that each enhancer was counted only once within each cell. Pairwise comparisons of enhancer sets between cells were conducted using the Jaccard Index [33], which quantifies similarity as the proportion of shared enhancers relative to the total number of unique enhancers across both cells. The resulting similarity scores were stored in a matrix and visualized as a heatmap (Fig. 3c, e, g) and corresponding Supplementary data has been provided. The analysis revealed significant variability in enhancer overlap across different cell types, as shown in Table 2. In healthy brain tissue, neurons exhibited the highest enhancer overlap, reaching a maximum of 36.57%. In contrast, OPCs (Fig. 3c) showed the lowest enhancer overlap, followed by pericytes, CNS macrophages, and astrocytes (Fig. 3g), which clustered around 21%. In the pancreatic head, epithelial cells from NA tissue demonstrated the highest overlap at 43.86% (Table 2). Endothelial cells, satellite cells, and neurons clustered around overlaps above 24%, while goblet cells and acinar cells exhibited overlaps greater than 34%. Furthermore, examination of T cells from PDAC tumor samples revealed a maximum overlap of 39.19%. Additionally, enteroendocrine cells (Fig. 3e), neuroendocrine cells, fibroblasts, macrophages/microglia, acinar cells, and cancer cells were found to have overlaps exceeding 25% (Table 2). The distribution of enhancer overlap computed based on pairwise similarity illustrates the cumulative distribution of pairwise overlaps for OPCs (Fig. 3d), enteroendocrine cells (from NA pancreatic head) (Fig. 3f), and astrocytes (Fig. 3h). The pairwise similarity matrix for enhancers and their cumulative distribution across all cell types is provided in the supplemental information (Supplementary Information). Pages 1–3 describe the remaining cell types from the putamen region, while pages 4–11 detail the remaining cell types from the NA pancreatic head, and pages 12–18 cover the cell types from tumor samples of the pancreatic head. Besides the pairwise similarity matrix metrics, we also include summary statistics, including mean, median, standard deviation, and percentiles (Table 2). We extended scAED to an independent scATAC-seq blood dataset containing T cells from three conditions: active HIV infection, HIV + ART therapy, and uninfected controls. The enhancer calling in these T cells mirrored the quantitative and qualitative patterns observed in the 21 brain and pancreas cell types. Pairwise similarity overlap heat maps for each condition are shown in Supplementary Fig. 5a–c, with maximum enhancer overlap reaching 25% among HIV-infected cells, 20% in ART-treated cells, and 16% in uninfected controls, values consistent with those observed in putamen and pancreatic cell types. The distributions of the number of cell pairs based on varying bin sizes for these conditions can be found in Supplementary Fig. 5d–f. The distribution of cell pairs across varying overlap bins was comparable between the datasets. These observations demonstrate that scAED can identify enhancers even in simpler cell types, such as CD4+ T cells, and in datasets

with multiple confounding variables, including disease status, treatment, and uninfected conditions.

Robustness of scAED framework and database in capturing heterogeneity

While cell heterogeneity is inherent to genetics of a cell, however, there are no studies that has demonstrated the exact levels of heterogeneity or homogeneity at single cell level resolution as of writing this manuscript. By assessing the OCRs and TF Footprints within them with known enhancer regions at the granular level of true single-cell resolution (not cluster-level) for the first time, we show that enhancer activity is actually heterogeneous even within the cell subtypes and not just primary celltypes within an organ. To demonstrate this, we take the example of cells obtained from human primary motor cortex where we profiled 8391 cells (from sample MM_410 from the BICCN project) and footprinted the OCRs. Utilizing the scAED framework, we show varied percent of homo- and heterogeneity across all neuronal and non-neuronal celltypes and their functional subtypes (Supplementary Table 2). The celltype that showed the maximum homogeneity was OGC_3 (Oligodendrocyte subtype 3) with a min-max proportion of enhancer homogeneity value of 0.00%–17.35% between the cells of that subtype, meaning some cells had no same enhancers between them while some cells showed up to 17.35% similarity. Interestingly, celltype ITL5_4 (Intratelencephalic Layer 5 neuron subtype 4) showed min-max percent of enhancer homogeneity value of 0.00%–0.00% indicating no shared enhancers even within the same functional subtypes and within the same sampled brain region (Supplementary Table 2). Some subtypes showed minimum homogeneity of 2.12% with ITL23_2 and ~2.83% with ITL6_1_2 subtypes (Supplementary Table 2) proving that scAED framework has the capabilities to truly detect the homogeneity-heterogeneity levels at the granular single-cell subtype levels.

Similarly, we also cross-compared enhancers by cell type within organ/condition and intersected them across all organs (Supplementary Table 3). We analyzed ~80 cell types spanning brain (primary adult cortex; multiple neuronal and glial subtypes), PDAC (tumor), NA (normal adjacent), and blood (CD4+ T cells) (Supplementary Table 3). This cross-comparison demonstrates that brain cell types showed strong organ exclusivity. Many neuronal and glial subtypes have ~60%–70% organ-exclusive enhancers (e.g. OGC_2: 66.8%, ITL23_5: 64.0%, PVALB_1: 64.2%), indicating organ-specific regulatory programs dominate in cortex (Supplementary Table 3). Contrastingly, pancreas (PDAC and NA) shows lower exclusivity (~5%–11%), e.g. celltypes such as, acinar (10.7%), neuroendocrine (10.6%), and epithelial (8.9%), reflecting greater sharing with other organs and/or broader/common regulatory activity, despite large absolute enhancer counts. CD4+ T cells of blood shows minimal exclusivity (0.11%) in this cross-organ intersection, suggesting that the specific enhancer set we detect in CD4+ T is largely shared with other contexts captured here since it is under-sampled relative to brain/pancreas breadth). This demonstrates the robustness of the scAED framework and database in capturing comparable enhancers across various organs and conditions, including brain primary tissue, pancreatic tumors versus normal tissue, and blood. These results indicate that the scAED framework and database capture biologically meaningful, cell type- and tissue-dependent regulatory differences.

scAED database

The enhancers have been compiled into a database using the Shiny application framework [31]. This application encompasses enhancers for all currently profiled cell types, displaying them in a tabular format with several columns: Chromosome, Start, End (coordinates), Genes, Cell Line, TF, and Motifs (Fig. 4) (Supplementary Fig. 6). Users can apply various filters, such as organ and region, to visually inspect a summary table describing the available cell types (Fig. 4a–d). Upon selecting a cell type of interest, users can explore detailed enhancer tables and graphs specific to that cell type (Fig. 4a and b). These data are available for local download for further analyses. Additionally, the 'Overlaps' tab provides in-depth information on TF binding for each enhancer, including specific binding coordinates and strand specificity within the enhancer region (Fig. 4c and d).

Search function and development of enhancer identifiers

scAED can be queried for enhancers using five methods: chromosomal genomic coordinates, target genes, TFs targeting enhancers, cell lines, and enhancer identifiers (EIDs). Besides this, users can also visualize all the enhancers that are active in a cell type by selecting appropriate filters. In contrast to the established identifiers for genomic features such as rsIDs for single nucleotide polymorphisms (SNPs), Promoter IDs, Exon IDs, Intron IDs, and Gene IDs, a universal identifier for enhancer regions has not been systematically defined. Recognizing the absence of a standardized system for enhancer identification and given that our single-cell methodology proved adept at comprehensively capturing active enhancer regions, we developed a new EID system. The EIDs are unique identifiers that are a set of 12 characters long, with the first two characters, 'En' denoting 'Enhancer,' followed by two characters to depict 'chromosome number,' and the next eight characters were numeric assigned to each enhancer region to facilitate a structured and consistent reference system. We recommend the adoption of EID as a universal standard to aid researchers in the precise and standardized referencing of enhancer regions.

Bidirectional enhancers

Our novel pipeline framework has demonstrated the ability to identify bidirectional active enhancers along both sense and antisense strands. They typically bind different sets of TFs on each strand to regulate gene expression in both directions (Supplementary File 1). Our methodology is the first of its kind capable of effectively capturing and cataloging these regulatory elements and providing database users with a robust tool for bidirectional enhancer detection. Analysis of bidirectional enhancer distribution revealed a consistent presence of ~30% across cell types and organs (Table 1).

Capturing trans elements in active enhancers

Another key innovative feature of our pipeline is its capacity to concurrently capture the predicted binding of trans-acting elements within active enhancer regions, providing comprehensive data on their specific binding coordinates, motifs, and strand orientation. This multidimensional exploration reveals a complex landscape of TF occupancy across enhancers. The number of TFs bound to individual active enhancers exhibited substantial variability, ranging from single TF occupancy to as many as >30 TFs co-occupying a single enhancer region (Supplementary File 1). The DNA sequence in the file name of Supplementary File 1 indicates the cell ID. The trans elements in AEs for each cell exist in three distinct states: (i) AE chromatin bound by transcription activators and regulators at precise coordinates, (ii) AE chromatin bound by transcription repressors at precise coordinates, and (iii)

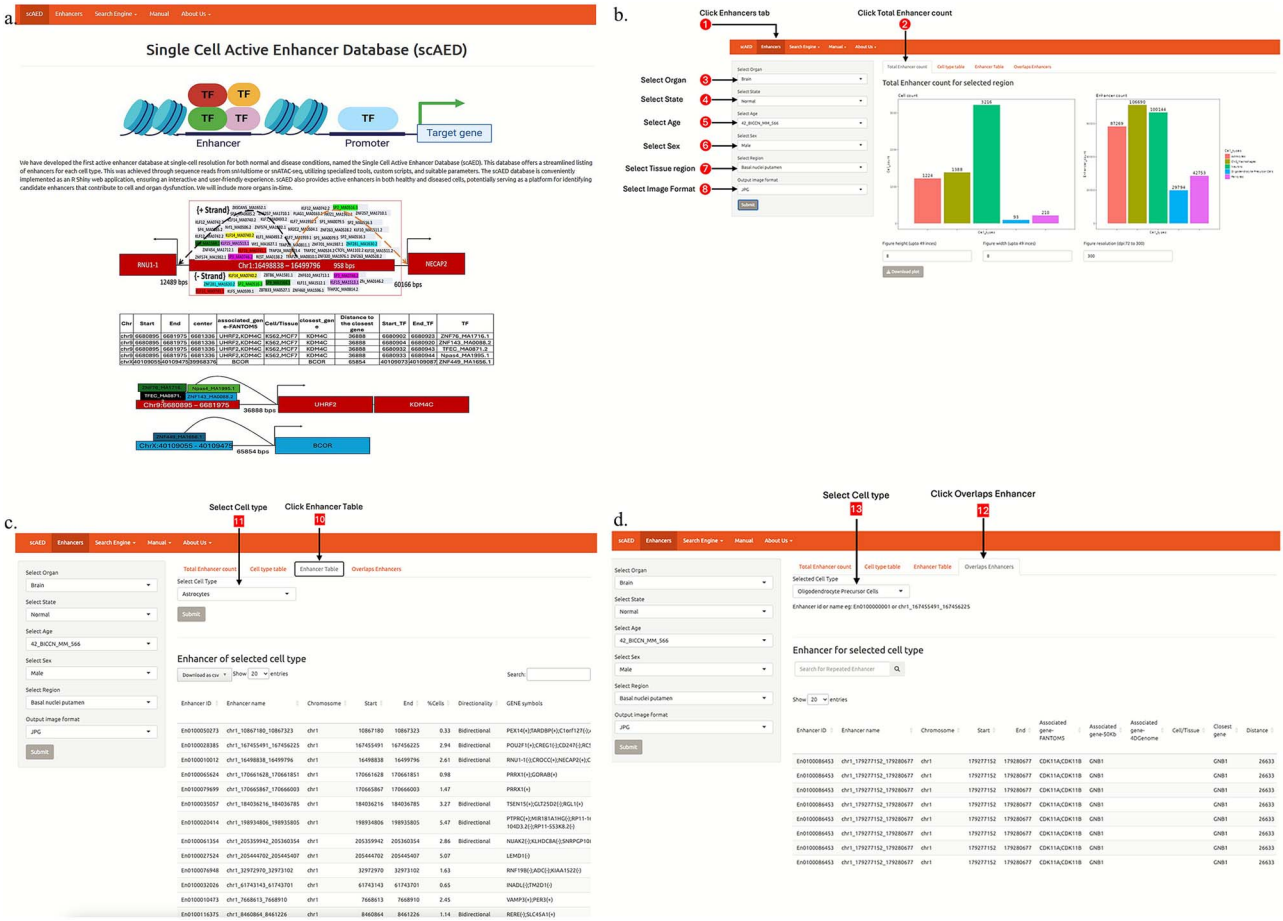

Figure 4. Annotation of scAED Data Access Pages. a. Shows the homepage of scAED website. b. Visualizes the total enhancer counts based on organ, physiological condition, age, sex, and tissue types. c. Presents the enhancer table, which includes the enhancer ID, enhancer name, chromosome, start and end positions, percentage of cells, directionality, target genes, and their strand locations, among other features. d. Illustrates the overlaps for the enhancers. In this context, 'overlaps' refers to each enhancer coordinate being repeated in rows to accurately display the TFs and their coordinates for each respective enhancer.

AE chromatin bound by both transcription repressors and regulators at precise coordinates. Each row of Supplementary File 2 provides these distinct states for the example cell ID with barcode AAACTACCAGACATAGTGGCGT.

Strand specificity and distance

scAED also includes strand information for each TF bound within the enhancer region, allowing users to understand strand specific TF binding preferences. An example file is provided as (Supplementary File 2). Another important feature of this database is the 'Distance' feature, which indicates the distance in base pairs (bp) from the enhancer region to its target genes (Fig. 4) (Supplementary Fig. 6).

Discussion

Our novel framework for the analysis of OCRs at single-cell resolution in multiple cell types from two distinct organs under three physiological conditions has generated a comprehensive resource of AEs (Table 1). scAED, as of writing this manuscript carries over 34.12 million AEs, providing a unique opportunity to elucidate the complex regulatory landscape governing gene expression. The ability to extract AEs from diverse platforms, including sc-multiome and snATAC-seq and leveraging the snRNAseq data for expression correlation, underscores the versatility and broad applicability of our methodological framework. In this framework, we avoided using unsupervised or supervised learning approaches for extracting enhancer information to prevent data loss and overfitting [34]. Instead, we took a more direct and novel approach in retrieving the regulatory information, ensuring complete control over preserving biological information. Some studies have used unsupervised machine learning approaches to extract enhancer information from bulk clusters of single cells; however, this method can obscure important cell-to-cell variability, leading to inaccurate enhancer identification due to averaging effects, and introduce biases that compromise the biological relevance of the findings [8–15]. Since the computational tools and frameworks previously developed for profiling enhancers in single-cell datasets [8–12] are currently inaccessible or unavailable in their repositories, and the remaining references represent bulk [13, 14] and experimental approaches [15], we were unable to perform direct comparisons of data preservation capabilities. This limitation in accessibility reinforces the need for our open-sourced and reproducible framework to generate and access active enhancer data in different types and states of single cells. Table 3 provides a concise overview comparing key features, such as accessibility, support for single-cell data, and enhancer resolution across several existing databases alongside scAED, highlighting the major limitations of current databases that scAED successfully addresses.

Table 3. Comparing key features, such as accessibility, support for single-cell data, enhancer resolution, among others, across several existing databases alongside our own resource, scAED

Database	Accessibility	Enhancer discovery		Enhancer resolution			Ref
		Single-cell data	Bulk data	Tissue level	Single cell-type (cluster) level	Single-cell level	
scEnhancer (http://enhanceratlas.net/scenhancer/)	NO	Yes	NO	NO	Yes	NO	[8]
EnhFFL (http://lcbb.swjtu.edu.cn/EnhFFL)	Yes	NO	Yes	Yes	NO	NO	[9]
SEdb (http://www.licpathway.net/sedb)	NO	NO	Yes	Yes	NO	NO	[10]
HEDD (http://zdzlab.einstein.yu.edu/1/hedd.php)	NO	NO	Yes	Yes	NO	NO	[11]
CenhANCER (http://cenhancer.chenzxlab.cn/)	NO	NO	Yes	Yes	NO	NO	[12]
HACER (https://bioinfo.vanderbilt.edu/AE/HACER/)	Yes	NO	Yes	Yes	NO	NO	[13]
EnhancerAtlas 2.0 (http://enhanceratlas.org/)	Yes	NO	Yes	Yes	NO	NO	[14]
VISTA Enhancer Browser (https://enhancer.lbl.gov)	Yes	NO	Yes	Yes	NO	NO	[15]
scAED (https://www.gudalab-rtools.net/scAED/)	Yes	Yes	NO	Yes	Yes	Yes	NA

Spectrum of enhancer heterogeneity

Our analysis of pairwise similarity in enhancer usage across all cells and cell types revealed a higher degree of heterogeneity than previously recognized (Fig. 3) (Table 2). The maximum enhancer overlaps between two cells varied significantly, ranging from as low as 12.95% in OPCs of the putamen to as high as 43.86% in epithelial cells from NA pancreatic head tissue (Table 2). This broad spectrum of enhancer overlap indicates that, even within the same cell type, individual cells share only less than half of their active enhancers, suggesting substantial variability in enhancer regulation across the profiled sections. scAED provides the ability to identify enhancers that are active on a cell-by-cell resolution. Furthermore, the minimum enhancer overlaps consistently fell between 0.01% and 0.04%, highlighting instances where cells shared only a few enhancers or, in some cases, nearly none (Table 2). This underscores the highly heterogeneous nature of enhancer regulation at the single-cell level. To our knowledge, this is the first study to systematically quantify enhancer relatedness at such a granular resolution, providing novel insights into the limited conservation of enhancer landscapes among individual cells within a given cell type. These findings offer a new perspective on the dynamic and cell-specific nature of enhancer usage, emphasizing the complexity of gene regulatory mechanisms at the single-cell level.

scAED offers several features for the first-time

To enable seamless exploration and dissemination of this rich dataset, we developed a user-friendly Shiny application [31], scAED. This web-based scAED platform combines various tools for exploring and visualizing active enhancers, providing a streamlined approach for users to focus on the biological interpretation and hypothesis generation. scAED has the ability to query enhancers using various parameters, including chromosomal coordinates, target genes, TFs, cell types, and EIDs, underscoring the versatility and user-friendliness of the scAED platform. This multifaceted search functionality empowers researchers to efficiently navigate the enhancer landscape and rapidly extract relevant information to aid in their investigations. scAED platform represents a significant advancement in this regard, providing a structured and universally applicable system for enhancer referencing through the introduction of EIDs.

Notably, the identification of bidirectional enhancers represents a crucial innovation enabled by our computational framework. Most cells, with only a few exceptions, consistently show the prevalence of bidirectional enhancers to be between 31% and 36% (Table 1), highlighting the highly prevalent nature of this regulatory mechanism. This finding provides evidence at the most granular level showing the prevalence and precise proportion of bidirectional enhancer-mediated regulation and opens new avenues for exploring the complexity of gene expression control. The comprehensive capture of trans-acting elements within active enhancer regions is another key feature of our platform. Supplementary File 2 shows the active enhancers identified in a cell with each row showing the precise binding coordinates of bound TFs (Supplementary File 2) (See column headers 'Start TF Binding', 'End TF Binding' and 'Transcription factors'). The detailed information on TFs, TF binding coordinates, motifs, and strand orientation provides valuable insights into the intricate regulatory networks on TF usage governing enhancer function. The observed variability in TF binding density across cell types highlights the cell-specific nature of enhancer-mediated regulation, underscoring the importance of single-cell resolution in enhancer mapping. The status of AEs was found to vary within a cell; while they are open and accessible, the nature of the bound TFs differed. Some AEs were exclusively bound with repressor TFs, while others were bound with activators, and some were found binding both activators and regulators (Supplementary File 2). In this way, scAED offers the versatility of providing the distinct states of AEs within a cell.

Enhancer distribution shows higher degree of cell–cell variability

Clustering analysis was performed on all cell types in the database (data not shown). However, the clustering analysis shown in Fig. 3a and b specifically highlights the distinct

enhancer profiles across astrocyte cell populations, based on common and unique enhancer coordinates. The identification of three clusters, particularly the third subset highlighted by hierarchical clustering (Fig. 3b), suggests the existence of previously unrecognized cell states or subpopulations with unique enhancer dynamics. This complexity in enhancer distribution, especially the observed heterogeneity within and between clusters, may reflect differential gene regulatory mechanisms that are key in maintaining cell subtype specific functions. The substantial inter-cluster separation, as indicated by t-SNE analysis, reinforces the notion that these enhancer profiles are not merely reflective of broad cell-type classifications but likely represent fine-scale regulatory differences with significant biological implications. These findings emphasize the importance of high-resolution; single-cell approaches to unravel the nuanced regulatory networks that underlie cellular heterogeneity, with potential implications for understanding tissue-specific functions and disease mechanisms. This analysis of enhancer overlaps within cell types revealed a striking degree of cell-to-cell variability, challenging the notion of a universal enhancer landscape and underscoring the dynamic and context-dependent nature of enhancer activation. Furthermore, we examined the relationship between active enhancer counts and gene density across chromosomes (Supplementary Fig. 2f). The strong positive linear relationship observed between gene density and enhancer counts across chromosomes suggest that chromosomes with higher gene density tend to have more active enhancers.

The integration of strand-specific information for TF binding and the inclusion of the 'distance' feature, which indicates the proximity of enhancers to their target genes, further enhances the utility of the scAED platform. These functionalities provide researchers with valuable insights into the directionality and spatial organization of enhancer-promoter interactions, vital for deciphering the intricate regulatory networks underlying gene expression programs.

The differential enhancer activity observed among the cells of intra- and inter-organ specimens underlines the distinct regulatory landscapes of these organs. The higher and broader enhancer frequency ranges in pancreatic cells, particularly in normal adjacent tissues, may reflect the organ's complex regulatory demands and pronounced cellular heterogeneity. This heterogeneity is further emphasized by the broad enhancer activity observed across different pancreatic cell types, demonstrating the complex regulatory dynamics at play. Further, scAED was found to be reliable in calling enhancers even in tissues with simple hierarchies or noisier ATAC-seq data (Supplementary Figs 4a–d and 5) confirming that scAED performs equally well in simple cell types such as T cells and in datasets prone to noisy signals. Further, across tissues, enhancer usage showed clear tissue dependence: cortical neuronal and glial subtypes display high organ exclusivity (~60%–70% unique enhancers), consistent with specialized regulatory programs in the brain, whereas pancreatic cell types exhibit lower exclusivity (~5%–11%) and CD4+ T cells show minimal exclusivity, reflecting broader sharing of regulatory elements. These cross-organ patterns, reproduced across conditions, indicate that our annotations capture biologically meaningful, tissue-specific regulation rather than artifacts of a single reference, and position scAED to dissect context-dependent enhancer landscapes at cell-type resolution.

Our single-cell active enhancer framework proved highly sensitive, capturing this cellular diversity with precision and computationally validating its capability to detect subtle variations in enhancer activity across individual cells. Importantly, these findings highlight how the use of averaging effects from databases hosting enhancers from pooled-cell sequencing technologies or cluster-level resolution could have obfuscated the identification of precise active enhancers, particularly those varying across different degrees within the cell population. By contrast, our single-cell approach ensures that these subtle, yet remarkable regulatory differences are not masked, thereby underscoring the critical importance of acquiring regulatory information at single-cell resolution to fully understand the intricate cellular diversity and its implications in both healthy and diseased states. Because scAED operates on standard fragment or BAM files with barcode tags, it should be compatible with other snATAC-seq chemistries that also generate unique barcodes, including chemistries of Bio-Rad ddSEQ scATAC-Seq, BD Rhapsody ATAC-Seq Assay, Takara Bio ICELL8 cx scATAC-Seq, Parse Biosciences Evercode ATAC (split-pool), ScaleBio QuantumScale™/scATAC Pre-Indexing, and Fluidigm (Standard BioTools) C1 scATAC-Seq, and while we have not yet benchmarked all of these platforms, we plan to incorporate such datasets in future extensions of scAED.

Conclusion

In conclusion, the scAED platform represents a significant advancement in the field of enhancer biology, providing a reliable framework and user-friendly resource for the exploration of active enhancers across diverse cell types and physiological conditions. We will continue to update scAED with newly identified enhancers from other organ, tissue, and cell types as relevant datasets become available. The innovative features, including the EID system, bidirectional enhancer identification, and the capture of trans-acting elements, position scAED as a powerful tool for unraveling the complexities of gene regulation in different cell types and states. We anticipate that the widespread adoption of this platform will accelerate the discovery of novel regulatory mechanisms and drive meaningful progress in understanding the molecular underpinnings of health and disease.

Key Points

- We developed **scAED**, the first single-cell active enhancer database, which identifies and catalogs enhancer activity at true single-cell resolution across diverse organs and physiological conditions, overcoming limitations of pooled-cell and cluster-level datasets.
- Our analysis uncovered **34.12 million active enhancers** across 13 796 single cells from 21 cell types and revealed a **high degree of enhancer heterogeneity** both within and between cell types, with overlap between cells often below 40%.
- scAED introduces **Enhancer IDs (EIDs)** —a structured naming system enabling universal reference of enhancer regions—and supports detailed querying by genomic coordinates, TFs, gene targets, or cell types via a Shiny-based user interface.
- The framework captures **bidirectional enhancer activity**, transcription factor binding coordinates, motif usage, and **strand specificity**, providing an unprecedented view of the trans-regulatory landscape at single-cell resolution.
- Enhancer clustering and similarity analyses demonstrated **fine-scale cell subpopulations** and revealed enhancer activity variability tied to gene density and

cell identity, highlighting the importance of single-cell resolution in regulatory genomics.

Acknowledgements

The authors would like to thank the Bioinformatics and Systems Biology Core (BSBC) at UNMC and the Holland Computing Center at the University of Nebraska for providing the computational and storage infrastructure.

Conflict of interest: The authors declare no competing interests.

Funding

The corresponding author is partly supported by multiple NIH awards—5P20GM103427, 5P30CA036727.

Data availability

The dataset used in the current study are available from the original authors on reasonable request.

Ethics approval and consent to participate

The original authors of the dataset had received appropriate ethics approval and consent.

Consent for publication

All authors listed in this manuscript consent towards the publication.

References

1. Karnuta JM, Scacheri PC. Enhancers: bridging the gap between gene control and human disease. *Hum Mol Genet* 2018;**27**:R219–27.
2. Vitale E, Gugnoni M, Ciarrocchi A. No need to stick together to Be connected: multiple types of enhancers' networking. *Cancers (Basel)* 2021;**13**:1–16.
3. Barral A, Dejardin J. The chromatin signatures of enhancers and their dynamic regulation. *Nucleus* 2023;**14**:2160551.
4. Marand AP, Zhang T, Zhu B. *et al.* Towards genome-wide prediction and characterization of enhancers in plants. *Biochim Biophys Acta Gene Regul Mech* 2017;**1860**:131–9. https://doi.org/10.1016/j.bbagrm.2016.06.006
5. Zaugg JB, Sahlén P, Andersson R. *et al.* Current challenges in understanding the role of enhancers in disease. *Nat Struct Mol Biol* 2022;**29**:1148–58. https://doi.org/10.1038/s41594-022-00896-3
6. Claringbould A, Zaugg JB. Enhancers in disease: molecular basis and emerging treatment strategies. *Trends Mol Med* 2021;**27**:1060–73.
7. Veerappa A, Pendyala G, Guda C. A systems omics-based approach to decode substance use disorders and neuroadaptations. *Neurosci Biobehav Rev* 2021;**130**:61–80.
8. Gao T, Zheng Z, Pan Y. *et al.* scEnhancer: a single-cell enhancer resource with annotation across hundreds of tissue/cell types in three species. *Nucleic Acids Res* 2022;**50**:D371–9.
9. Kang R, Tan Z, Lang M. *et al.* EnhFFL: a database of enhancer mediated feed-forward loops for human and mouse. *Precis Clin Med* 2021;**4**:129–35. https://doi.org/10.1093/pcmedi/pbab006
10. Jiang Y, Qian F, Bai X. *et al.* SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Res* 2019;**47**:D235–43. https://doi.org/10.1093/nar/gky1025
11. Wang Z, Zhang Q, Zhang W. *et al.* HEDD: Human enhancer disease database. *Nucleic Acids Res* 2018;**46**:D113–20. https://doi.org/10.1093/nar/gkx988
12. Luo ZH, Shi MW, Zhang Y. *et al.* CenhANCER: a comprehensive cancer enhancer database for primary tissues and cell lines. *Database (Oxford)* 2023;**2023**:1–7. https://doi.org/10.1093/database/baad022
13. Wang J, Dai X, Berry LD. *et al.* HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res* 2019;**47**:D106–12. https://doi.org/10.1093/nar/gky864
14. Gao T, Qian J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res* 2020;**48**:D58–64. https://doi.org/10.1093/nar/gkz980
15. Visel A, Minovitsky S, Dubchak I. *et al.* VISTA enhancer browser–a database of tissue-specific human enhancers. *Nucleic Acids Res* 2007;**35**:D88–92. https://doi.org/10.1093/nar/gkl822
16. Veerappa AM, Rowley MJ, Maggio A. *et al.* CloudATAC: a cloud-based framework for ATAC-Seq data analysis. *Brief Bioinform* 2024;**25**:1–12. https://doi.org/10.1093/bib/bbae090
17. https://github.com/linnarsson-lab/adult-human-brain, *Cytograph qc*. 2024.
18. Li YE, Preissl S, Miller M. *et al.* A comparative atlas of single-cell chromatin accessibility in the human brain. *Science* 2023;**382**:eadf7044. https://doi.org/10.1126/science.adf7044
19. Siletti K, Hodge R, Mossi Albiach A. *et al.* Transcriptomic diversity of cell types across the adult human brain. *Science* 2023;**382**:eadd7046. https://doi.org/10.1126/science.add7046
20. Butler A, Hoffman P, Smibert P. *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;**36**:411–20. https://doi.org/10.1038/nbt.4096
21. Pratt HG, Ma L, Dziadowicz SA. *et al.* Analysis of single nuclear chromatin accessibility reveals unique myeloid populations in human pancreatic ductal adenocarcinoma. *Clin Transl Med* 2024;**14**:1–21.e1595. https://doi.org/10.1002/ctm2.1595
22. Zhang K, Zemke NR, Armand EJ. *et al.* A fast, scalable and versatile tool for analysis of single-cell omics data. *Nat Methods* 2024;**21**:217–27. https://doi.org/10.1038/s41592-023-02139-9
23. van Dijk D, Sharma R, Nainys J. *et al.* Recovering gene interactions from single-cell data using data diffusion. *Cell* 2018;**174**:716–729.e27. https://doi.org/10.1016/j.cell.2018.05.061
24. https://github.com/Winnie09/GPTCelltype, *GPTCelltype*. 2024.
25. Danecek P, Bonfield JK, Liddle J. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* 2021;**10**:1–4. https://doi.org/10.1093/gigascience/giab008
26. Gaspar JM. John M Gaspar (ed.), *Genrich: Detecting Sites of Genomic Enrichment, in Github*. 2018. University of New Hampshire.
27. Bentsen M, Goymann P, Schultheis H. *et al.* ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat Commun* 2020;**11**:4267. https://doi.org/10.1038/s41467-020-18035-1
28. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2.
29. Fishilevich S, Nudel R, Rappaport N. *et al.* GeneHancer: genome-wide integration of enhancers and target genes

in GeneCards. *Database (Oxford)* 2017;**2017**:1–17. https://doi.org/10.1093/database/bax028
30. Hinrichs AS, Karolchik D, Baertsch R. *et al.* The UCSC genome browser database: update 2006. *Nucleic Acids Res* 2006;**34**:D590–8. https://doi.org/10.1093/nar/gkj144
31. Winston Chang JC, Allaire JJ, Sievert C. *et al. Shiny: Web Application Framework for R*2024. USA: Posit Software, PBC Publisher. https://shiny.posit.co/.
32. Wickham H. Hadley Wickham (ed.), *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag, 2016.
33. Raimundo Real JMV. The probabilistic basis of Jaccard's index of similarity. *Syst Biol* 1996;**45**:380–5.
34. Johansson U, Sönströd C, Norinder U. *et al.* Trade-off between accuracy and interpretability for predictive in silico modeling. *Future Med Chem* 2011;**3**:647–63. https://doi.org/10.4155/fmc.11.23

Briefings in Bioinformatics, 2025, **26(6)**, bbaf646
https://doi.org/10.1093/bib/bbaf646
Problem Solving Protocol