Consensus peaks of chromatin accessibility in the human genome

Qiuchen Meng[1,#], Xinze Wu[1,#], Chen Li[1], Jiaqi Li[1], Xi Xi[1], Shengquan Chen[2], Rui Jiang[1], Lei Wei[1,*], Xuegong Zhang[1,3,*]

1. MOE Key Laboratory of Bioinformatics and Bioinformatics Division, BNRIST, Department of Automation, Tsinghua University, Beijing 100084, China
2. School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, China
3. School of Life Sciences and School of Medicine, Center for Synthetic and Systems Biology, Tsinghua University, Beijing 100084, China

[#] These authors contributed equally.

[*] Corresponding author: weilei92@tsinghua.edu.cn (L.W.); zhangxg@tsinghua.edu.cn (X.Z.)

Abstract

Chromatin accessibility profiling methods such as assay for transposase-accessible chromatin using sequencing (ATAC-seq) have been promoting the identification of gene regulatory elements and the characterization of epigenetic landscapes. Unlike gene expression data, there is no consistent reference for chromatin accessibility data, which hinders large-scale integration analysis. By analyzing many more than 1000 ATAC-seq samples and 100 scATAC-seq samples, we found that cells share the same set of potential open regions. We thus proposed a reference called consensus peaks (cPeaks) to represent open regions across different cell types, and developed a deep-learning model to predict all potential open regions in the human genome. We showed that cPeaks can be regarded as a new set of epigenomic elements in the human genome, and using cPeaks can increase the performance of cell annotations and facilitate the discovery of rare cell types. cPeaks also performed well in analyzing dynamic biological processes and diseases. cPeaks can serve as a general reference for epigenetic studies, much like the reference genome for genomic studies, making the research faster, more accurate, and more scalable.

Introduction

Chromatin accessibility refers to the ability of nuclear macromolecules to physically interact with chromatinized DNA[1]. The accessible regions, also known as open regions, imply the positions of many transcriptional regulatory elements, including promoters, enhancers, insulators, and transcription factor (TF) binding sites[2,3]. Studying chromatin accessibility helps to infer gene regulatory networks[4–8] and understand various biological processes such as cell differentiation[9,10], organ development[11,12], and disease occurance[13–17].

The rapid development of high-throughput sequencing technologies[18], especially Assay for Transposase-Accessible

Chromatin using sequencing (ATAC-seq)[19,20], provides fast and easy genome-wide measurement of chromatin accessibility. scATAC-seq[21–26], the single-cell version of ATAC-seq, is becoming the second major resource of information besides scRNA-seq for building the biomolecular atlases of cells[3,16,27–29].

Unlike RNA-seq data with reference transcriptomes[30], ATAC-seq data lack a unified reference for localizing and quantifying chromatin accessibility across genomes. Until now, some studies used peak-calling algorithms such as MACS2[31] or Cell Ranger ATAC[32] to define genomic regions with significantly enriched read signals as open regions and named them as peaks[3,17,27]. Some studies partitioned the genome into fixed regions and quantified chromatin accessibility in each region[33,34]. Several pre-defined sets of open regions were proposed, such as candidate cis-regulatory elements (cCREs) provided by the ENCODE project[35] or the *cis*-element ATLAS (CATLAS)[3]. These sets aimed to define the functional elements across the genome from a biological perspective instead of building a reference for ATAC-seq. cCREs provided by ENCODE were calculated from DNase I hypersensitive seq (DNase-seq) data[35], and cCREs provided by CATLAS erased the information of open region length[3]. Thus, these open region sets are usually used for quality control during ATAC-seq data analysis rather than serve as a reference.

The absence of a unified ATAC-seq reference has hindered the integration of data from different studies. Data integration is an essential part of cross-dataset comparison and cell atlas construction of ATAC-seq data. Peak merging[36–39] and iterative removal[3] are two common approaches to integrate ATAC-seq data. The merging approach is to join overlapped peaks and expand them to a wider one. It may get extra-long peaks after applying many merging steps, and thus is unsuitable for large-scale analysis, especially for single-cell atlas-level analysis. Besides, the peaks need to be re-unified each time when new data are added. The iterative removal approach is to select the most statistically significant peak out of all overlapped peaks. This approach ignores the information contained in the removed peaks and is impractical in building cell atlases with peaks that are obtained through different statistical methods. A generic reference of chromatin accessibility designed from a data-processing perspective is needed for integrating existing and future ATAC-seq data.

To build such a chromatin accessibility reference, we collected 624 high-quality human ATAC-seq datasets[40,41] covering ~40 organs. We found different cell types shared a limited set of potential open regions (PORs), which showed confidence in establishing a generic reference for all cell types. We obtained ~1.2M observed consensus Peaks (cPeaks) form these PORs. We discovered the sequence patterns of the observed cPeaks by deep learning and predicted more cPeaks as predicted ones to enhance the performance of cPeaks in more cell types. We combined the observed and predicted cPeaks as a whole reference for chromatin accessibility. We conducted a series of experiments to validate the rationality of cPeaks and the performance of cPeaks in downstream analysis. cPeaks enhanced the annotation of scATAC-seq data and facilitated the identification of rare cell types. cPeaks were proved to enable simple and clear analysis of many intricate biological states, such as cell differentiation, organ development and disease. Together with the cPeak reference set, we provided bioinformatics tools for mapping or converting diverse ATAC-seq data types to cPeaks.

Results

Define cPeaks by aggregating ATAC-seq data

We collected 624 high-quality bulk ATAC-seq samples from ENCODE[41] and ATACdb[40], covering over 40 human organs and all body systems (Methods, Table S1). Each sample contained a set of pre-identified peaks, where each peak represented an open region on the genome (Fig. 1a). We randomly aggregated these samples one by one by merging the overlapped peaks as a wider one at each step. We found that the number and genomic coverage of merged peaks grew gradually with the addition of samples, but the growth rate became slower (Fig. 1b). The results suggested that open regions are not randomly located on the genome.

After adding all samples, we identified about 1.2 million merged peaks with about 30% genomic coverage. We defined each joint peak as a potential open region (POR) (Methods). A POR describes a genomic region that could open in one or more cell types. For each position of a POR, we calculated the proportion of samples that contained peaks in this position, and denoted this proportion as the openness at this position. For each POR, we constructed an array that contained the openness at all positions of this POR along the 5'-to-3' direction and defined the array as the shape of this POR (Methods).

The samples used to generate PORs contained two major tissue types: blood (such as T cells, B cells, and NK cells) and solid tissues (such as lung, heart, and brain), which constituted 51% and 49% of the samples, respectively. We regenerated PORs for these two tissue types separately to investigate whether POR shapes were consistent (Fig. 1c). We defined PORs which overlapped between the two tissue types as a set of shared PORs. The number of shared POR sets corresponded to 87% and 76% of total PORs in liquid and solid tissue, respectively. We then investigated the shape similarity of each shared PORs by calculating the Pearson's correlation coefficient (PCC) of the shapes of shared PORs between blood and solid tissues (Methods). Shared PORs showed high PCCs between the two tissue types, with an average of 0.93 and a median of 0.96 (Fig. S1a). It indicated that the POR shape might be an intrinsic property, irrelevant to cell or tissue types.

We found that there were two types of PORs according to their shapes: unimodal PORs and multimodal PORs (Fig. 1a, Methods). We assumed that multimodal PORs were composed of adjacent unimodal PORs and split each multimodal POR into several unimodal PORs (Fig. 1a, Methods). We considered each unimodal POR as a basic element of the chromatin accessibility reference and named it as an observed consensus peak (cPeak). We obtained ~1.4 million observed cPeaks in total (Table S2, an example provided in Fig. S1b). We found that lengths of most cPeaks were less than 2,000 bp. The median length of cPeaks was 522 bp, similar to the previous observation that the peak length was about 500 bp[42] (Fig. S1c).

We then assessed the performance of cPeaks as a chromatin accessibility reference when dealing with scATAC-seq data. We collected 231 scATAC-seq samples from published studies (Table S3), each with a set of peaks pre-identified by different peak-calling methods[31,32]. For each sample, we calculated the proportion of the pre-identified peaks that overlapped with cPeaks and defined the proportion as the coverage of cPeaks. Most samples showed coverages greater than 90%, with an average of 0.98 and a median of 0.99 (Fig. 1d and S1d). We selected 27

scATAC-seq samples from different organs in CATLAS[3] to evaluate the enrichment of sequencing reads on cPeaks. We found that the sequencing reads enriched around centers of cPeaks (Fig. 1e, Methods), and on average, 70.3% of sequencing reads were overlapped with cPeaks, indicating that cPeaks could cover a significant proportion of scATAC-seq signals.

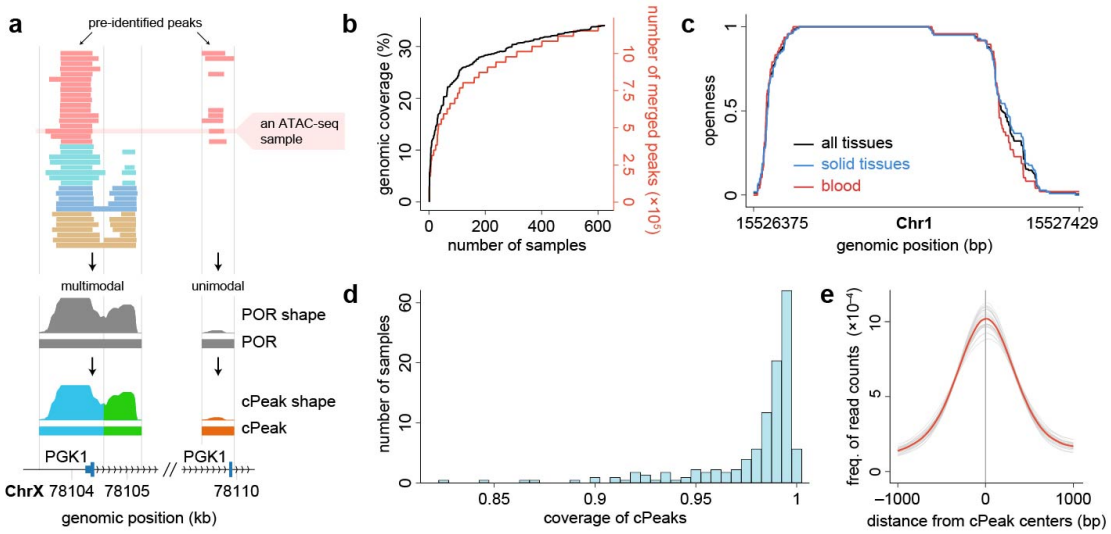

Fig 1. The generation of cPeaks

a. A schematic diagram about cPeak generation. Here we showed a part of the genomic region near PGK1 as an example. In the upper part, each horizontal row represented an ATAC-seq sample, and each segment represented a pre-identified peak. Different colors represented different cell types.

b. Change of the genomic coverage (black line) and number (red line) of merged peaks during the addition of samples.

c. An example of POR shapes generated by blood, solid tissues, or all tissues.

d. Coverages of cPeaks in 231 scATAC-seq samples.

e. Enrichment of sequencing reads around centers of cPeaks. All sequencing reads overlapped with ±1000 bp genomic regions around cPeak centers were aggregated to form the distributions. Each grey line represented a sample. The red line showed the average distribution of all samples.

Properties of observed cPeaks

We investigated whether cPeaks are associated with known regulatory elements or biological functions. By comparing cPeaks with known genomic annotations, we found that cPeaks were enriched around transcriptional start sites (TSSs) and transcription end sites (TESs) (Fig. 2a, Methods). About 13% of cPeaks were localized at promoter regions (±2000 bp around TSS), while the remaining cPeaks were primarily localized at exons or distal intergenic regions (Fig. 2b, Methods). We further collected multiple types of epigenetic annotations, including histone

modification and CTCF ChIP-seq data (Table S4). Using the genomic and epigenetic annotations, we inferred each cPeak as one or more regulatory elements, such as promoters (including known and predicted promoters), untranslated regions (UTRs), enhancers, or CTCF binding sites (CTCF BSs) (Methods). We found that most cPeaks were inferred as enhancers or promoters (Fig. 2c).

For each cPeak, we defined the maximum value of its shape as its open ratio to reflect the ratio of samples opening in this cPeak (Methods). We defined a cPeak with an open ratio > 90% as a housekeeping cPeak (Fig. 2d). 24,538 cPeaks (1.8% of all cPeaks) were identified as housekeeping cPeaks. The median length of housekeeping cPeaks was 1441 bp, with most locating at promoter regions (Fig. 2e). These promoter regions covered 77.4% of the promoters of housekeeping genes[43] (Fisher's exact test p-value $< 1\times10^{-5}$, Methods). The proportion of housekeeping cPeaks and the proportion of housekeeping genes on difference chromosomes were highly correlated (Fig. S2a). We also performed Gene Ontology (GO) enrichment analysis on genes related to housekeeping cPeaks (Methods) and found that they were enriched in the pathways related to essential biological processes such as ribonucleoprotein complex biogenesis, cell cycle checkpoint signaling, and methylation (Fig. 2f).

We found that there were three different patterns of cPeak shapes according to the steepness on both edges of cPeak shapes (Methods). We hypothesized that these different shape patterns were related to different arrangements of the nucleosomes around cPeaks and named them as strongly positioned, one-side strongly positioned, and weakly positioned cPeaks (Fig. 2g and S2b). cPeaks with steep edges signified more consistent boundaries of open regions detected in different experiments and suggested the presence of strongly-positioned nucleosomes around the boundaries. Conversely, a gentle edge suggested that the nucleosomes around cPeak boundaries were weakly positioned.

We conducted motif enrichment analysis around edges of cPeaks (Methods). We identified some known TFs associated with the steep edges of cPeaks, such as CTCF, BORIS, Fosl2, Fra2, Fra1, Fos and Atf3. Most of these TFs shared the same DNA binding domains: the basic leucine zipper (bZIP) domain, the Zinc finger (Zf) domain, and the erythroblast transformation specific (ETS) domain (Fig. S2c). Most motifs were located on cPeaks about 150 bp away from the boundaries (Fig. 2h).

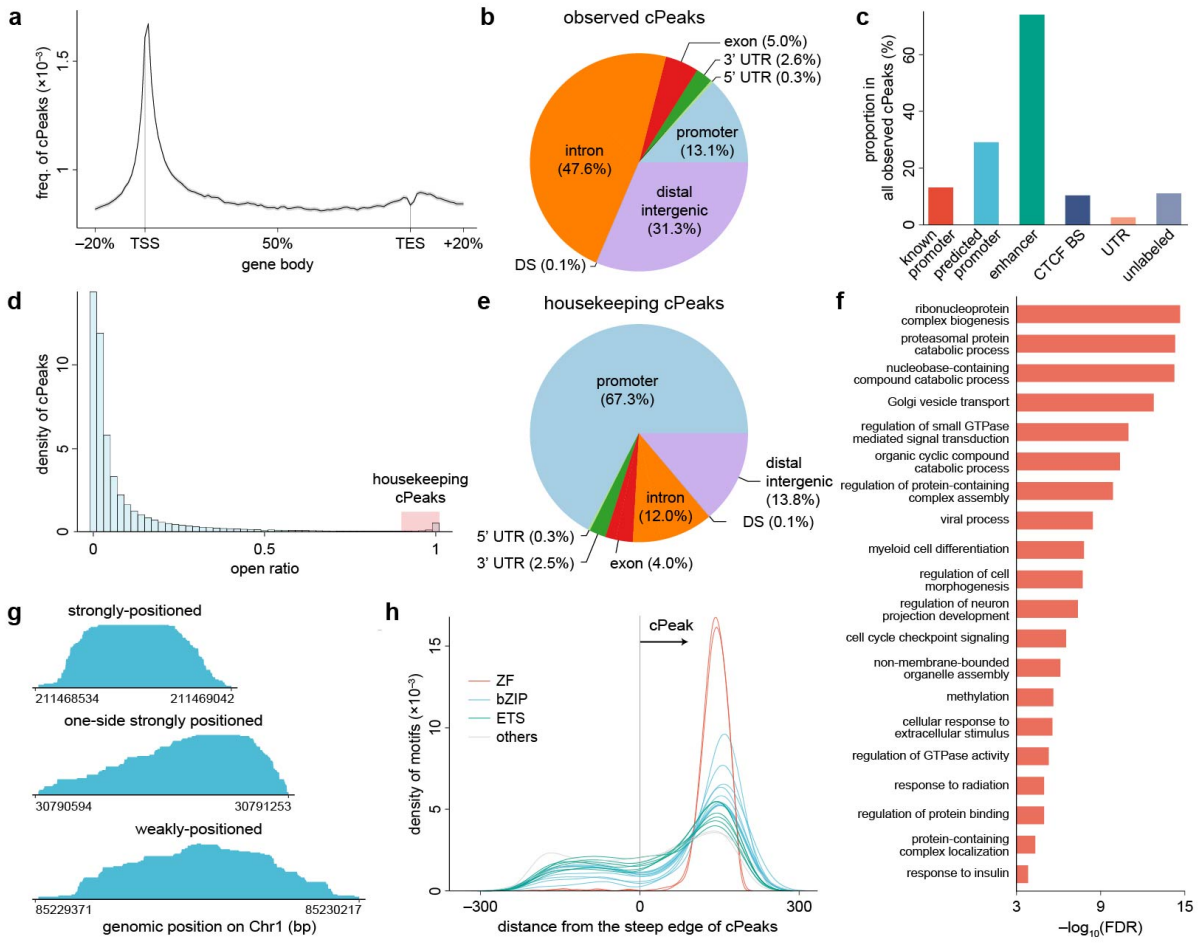

Fig. 2. Properties of observed cPeaks

a. The enrichment of cPeaks around gene bodies. The x-axis reflected the percentage of total length of gene body regions. The black line represented the average value and the grey intervals represented the 95% confidence intervals.

b. Genomic annotations of observed cPeaks. DS: downstream (≤ 300 bp downstream of the gene end).

c. The inferred regulatory roles of observed cPeaks. Each cPeak may have multiple inferred functions. CTCF BS: CTCF binding site.

d. The distribution of open ratios of observed cPeaks.

e. Genomic annotations of housekeeping cPeaks.

f. Top pathways enriched in genes related to housekeeping cPeaks. The 20 pathways with the smallest false discovery rates (FDRs) were shown.

g. Examples of cPeaks with different shape patterns.

h. Positions of the top 20 enriched motifs around the steep edges of cPeaks. Each line represented a motif, and

each color represented a DNA-binding domain type.

Extend cPeaks through deep learning

We obtained observed cPeaks based on limited cell types. As shown in Fig. 1b, there are likely to show some open regions not contained by these observed cPeaks when considering a new sample, especially for cell types or tissues not contained in the datasets used to generate observed cPeaks. We developed a machine learning method that learned the patterns of observed cPeaks to predict more cPeaks to deal with this problem. We observed that different cell types share the same location and shape of PORs (Fig. 1c), suggesting that the occurrence of cPeaks are associated with gene sequences. A previous study also showed the correlation between DNA sequence and chromatin accessibility[44]. Thus, we adopted the deep convolutional neural network (CNN) model of scBasset[44] to predict more cPeaks from DNA sequences (Fig. 3a, Methods). We defined the regions that were not covered by observed cPeaks as background regions. We trained the model using a binary classification approach with observed cPeaks as positive samples and randomly selected background regions as negative samples (Methods). The model achieved high accuracy with an AUC of 87.93% (Methods). The results demonstrated that cPeaks can be predicted from DNA sequences.

We then applied the trained model to predict more cPeaks. We used a sliding window with a length of 500bp to scan the background regions and got an openness score for each region to represent its probability as a cPeak (Fig. 3b, Methods). We selected regions with openness scores higher than 0.55 as predicted cPeaks (Fig. 3c) and obtained ~0.2 million predicted cPeaks (Table S5). We performed genomic annotations and inferred regulatory roles of these predicted cPeaks and found that their properties were similar to observed cPeaks (Fig. S3).

To assess the performance of predicted cPeaks, we calculated their coverages using new ATAC-seq samples collected from ChIP-Atlas[45], including 1,161 bulk ATAC-seq samples covering 8 cell types or tissues (B cells, T cells, fibroblasts, brain, hippocampus, liver, lung, retina) (Table S3). Most of these cell types or tissues existed in the data used to generate cPeaks except for retina, but none of these samples were included during cPeak generation. We randomly generated a set of background regions (random regions) with the same length and number as predicted cPeaks. We evaluated the coverages of observed cPeaks, observed cPeaks plus random regions, and observed cPeaks plus predicted cPeaks on these new-collected samples. We found that observed cPeaks already covered more than 95% of the pre-identified peaks for most samples. For retina, the coverage of observed cPeaks was 90.7%, slightly lower than others. The addition of random regions to observed cPeaks showed no significant improvement on coverages compared to observed cPeaks, but the addition of predicted cPeaks significantly increased coverages for most cell types and tissues (Fig. 3d). Notably, the retina samples showed the largest improvement in coverage after adding predicted cPeaks, from 91.3% to 92.3%. These results indicated that predicted cPeaks can elevate the performance of cPeaks to cover open regions in different samples, especially for cell types or tissues that were not included during cPeak generation. Thus, we combined predicted cPeaks with observed cPeaks to create an extended set of cPeaks, named integrated cPeaks. We provided bioinformatics tools for mapping or converting diverse

ATAC-seq data types to integrated cPeaks (see Code Availability for details). We defined an Epigenetics Transfer Format (ETF) to comprehensively describe the information of each cPeak, including its genomic location, length, housekeeping status, shape pattern, and inferred regulatory functions (Methods). The ETF file for all cPeaks was provided in Table S6.

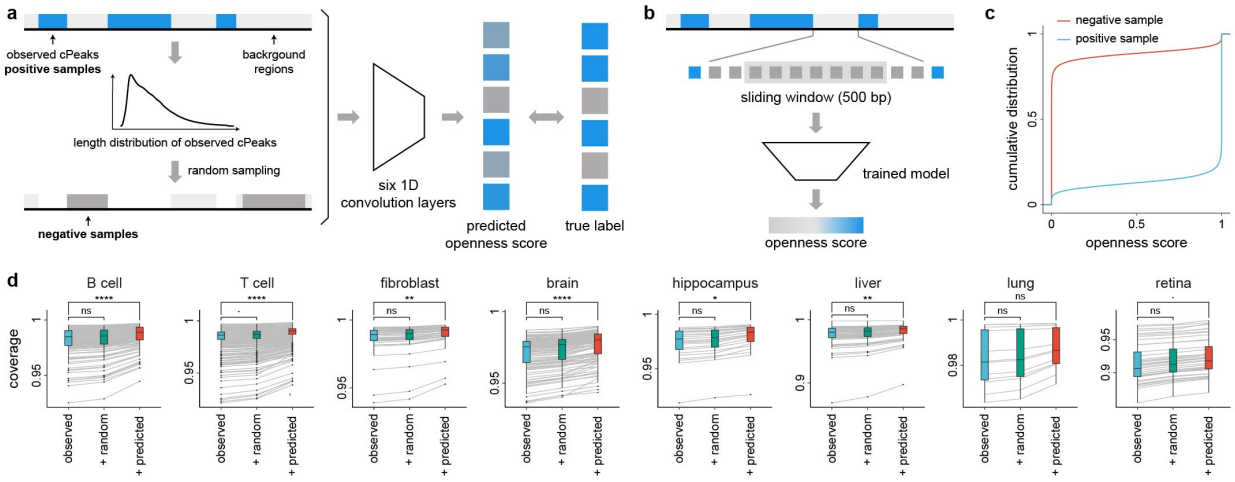

Fig. 3. Extend cPeaks by a deep CNN model

a. The one-dimensional CNN model for predicting whether a genomic region is a cPeak. We considered cPeaks as positive samples and followed their length distribution to generate negative samples from background regions.

b. Predicting cPeaks by the CNN model. We scanned all background regions by a window of 500 bp with a stride of 250 bp. Each region was fed into the trained model to get an openness score.

c. An empirical cumulative distribution of openness scores of all training samples of the CNN model. The red line represented the negative samples randomly generated from background regions, and the blue line represented the positive samples (observed cPeaks).

d. Coverages of different cPeak sets on ATAC-seq data. Each grey line linked the same sample analyzed by different cPeak sets. The significant levels were calculated by the paired Wilcoxon signed-rank test (*p*-values denoted as ****, < 0.0001; ***, < 0.001; **, < 0.01; *, < 0.05; ·, < 0.1; ns, ≥ 0.1).

cPeaks benefit cell annotation of scATAC-seq data

We compared the performance of cPeaks as a reference on scATAC-seq data annotation with three commonly used approaches in defining open regions (Methods). The first approach divided the genome into 2,000 bp bins as the basic elements[33,34,46], and here we named them as genomic bins. The second approach combined all sequencing single cells together as pseudo-bulk samples and called peaks from this pseudo-bulk sample[17,47]. We named the

called peaks as pseudo-bulk peaks. The third approach first clustered cells based on features such as pseudo-bulk peaks or labels derived from cell-sorting, called pseudo-bulk peaks for cells in each cluster, and then merged the called peaks together[3,29]. We named the final peaks as merged peaks. Each approach provided a genomic region set as a reference for mapping sequencing reads to generate a cell-by-chromatin accessibility feature matrix. We used all features, top 50,000 highly variable features (HVFs), and top 5,000 HVFs to perform cell annotation, respectively (Methods).

We used a hematopoietic differentiation scATAC-seq dataset[9], which contains 2,034 cells with FACS-sorting labels as the ground truth of cell type labels. We generated cell-by-feature matrices by different feature-defining approaches, applied different feature selection methods, and then performed different commonly-used embedding methods[7,21,32,33,33,42,48] (Fig. 4a). We evaluated the annotation results by the adjusted rand index (ARI), adjusted mutual information (AMI), and homogeneity (H)[49] (Fig. 4a, Methods). The results showed that cPeaks got similar or better performance in comparison with other feature-defining approaches under all annotation and evaluation methods (Fig. 4b and S4). Remarkably, when considering only 50,000 or 5,000 features, cPeaks still exhibited superior performance in cell type annotation, outperforming all other methods.

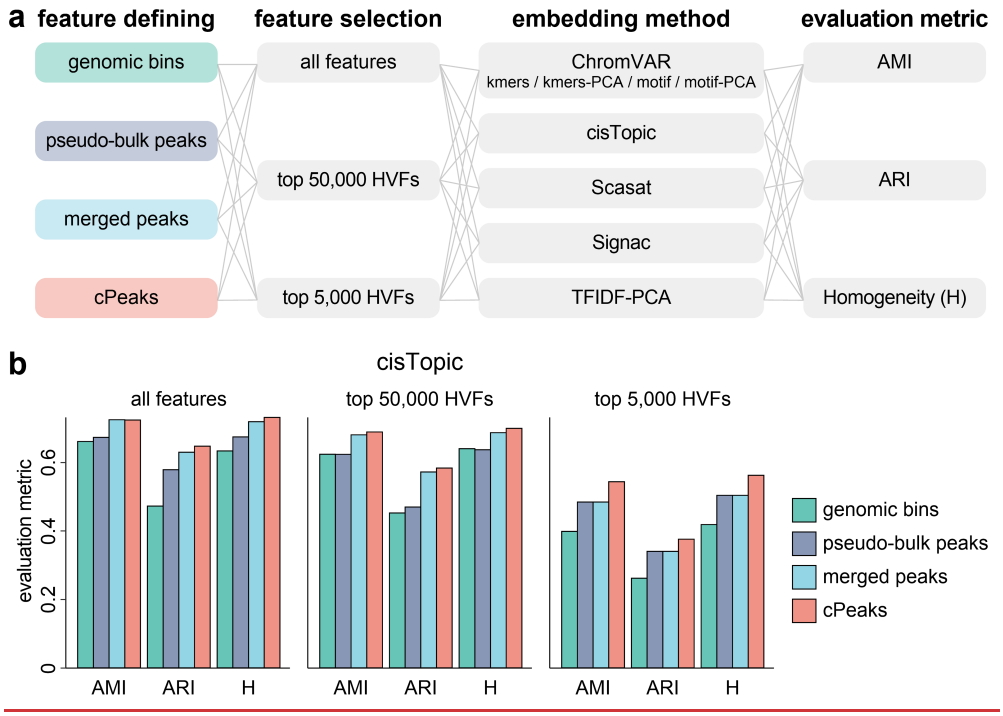

Fig. 4. The performance of cell annotations on the FACS2-sorted hematopoietic differentiation dataset

a. The analysis workflow of cell annotations. We used four different feature-defining approaches to obtain references for generating cell-by-feature matrices. We used three ways to select features, applied eight different strategies to embed the features, and used three metrics to evaluate the results.

b. Performance of different feature-defining approaches with different feature selection ways and evaluation metrics via the cisTopic embedding method.

cPeaks enable the discovery of rare cell types

We utilized cPeaks to annotate a scATAC-seq dataset of peripheral blood mononuclear cells (PBMCs) generated by 10X Genomics, including 8,728 cells with 87,561 pre-identified pseudo-bulk peaks. We used cPeaks as the reference to generate a cell-by-feature matrix for this dataset (Fig. S5a, Methods). We conducted cross-modality integration of this dataset and a scRNA-seq PBMC dataset provided by 10x Genomics via the anchor method[50] in Seurat to assign a pre-defined cell-type label in the scRNA-seq data to each cell in the scATAC-seq data (Methods). The assigned labels of the scATAC-seq data were shown to be highly consistent on the Uniform Manifold Approximation and Projection (UMAP) plot (Fig. 5a). We also calculated the prediction score which indicated the confidence of cell type assignment in cross-modality integration for each cell[50] (Methods), and the results showed that cPeaks could improve the prediction score compared to the original peak calling approach (Fig. 5b and S5b).

After cell type annotation, we identified three rare cell types, named as plasmacytoid dendritic cells (pDCs), dendritic cells, and natural killer bright (NK bright) cells, each with a proportion of less than 1% (Fig. S5c). We focused on pDCs and identified their differential open regions in comparison with all other cells (Methods). The results showed that analysis using cPeaks identified 1,400 more differential open regions than using pseudo-bulk peaks (Fig. 5c), most of which were specifically open in pDCs (Fig. 5d and S5d). Due to the low abundance of pDCs, these regions were unlikely to be detected by pseudo-bulk peaks. These open regions were important for investigating the functions of rare cell types. For example, we found that the promoter region of the PTCRA gene was a pDC-specific open region (Fig. 5e and 5f). According to the scRNA-seq data, PTCRA was only expressed in pDCs (Fig. 5g). Previous studies also confirmed the expression of PTCRA is the feature of human pDCs[51,52].

cPeaks also have the potential to identify rare cell subtypes that are difficult to be detected using pseudo-bulk peaks. We applied cPeaks as references to cluster GI epithelial cells in CATLAS[3] (Methods). In the original study, iterative clustering was used to identify nine different cell subtypes, including four rare subtypes each with a ratio less than 2% (Fig. S5e). However, the peaks generated by the original study failed to clearly visualize all the rare subtypes (Fig. S5f). An additional peak calling step for GI epithelial cell types was essential to separate these subtypes on the UMAP plot (Fig. S5g). Using cPeaks as references, we directly generated the cell-by-feature matrix and separated all rare subtypes clearly (Fig. 5h). All the results demonstrated that cPeaks simplified the process of scATAC-seq data processing and facilitated the discovery of rare cell types and their specific open regions which would otherwise be difficult to be detected using traditional peak calling approaches.

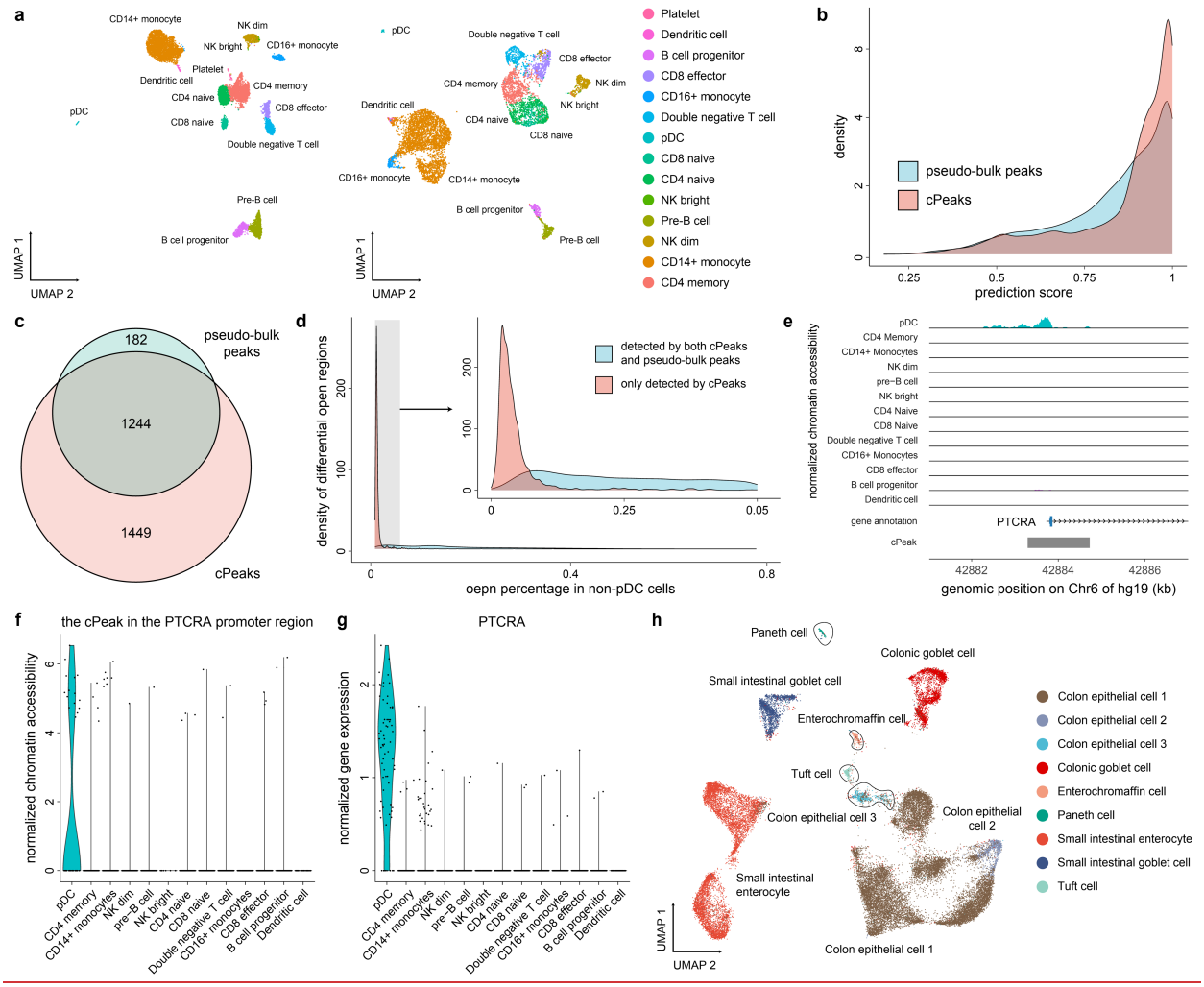

Fig. 5. Rare cell type discoveries using cPeaks

a. The UMAP plots of scRNA-seq and scATAC-seq PBMC data. The colors in scRNA-seq were cell labels defined by gene expression. The corresponding colors in scATAC-seq were the inferred cell type labels by cross-modality integration.

b. Prediction scores of the cross-modality integration between the scRNA-seq data and scATAC-seq data analyzed by cPeaks (red) or pseudo-bulk peaks (blue).

c. The Venn plot of the number of differential open regions calculated by cPeaks (red) or pseudo-bulk peaks (blue).

d. The open percentage of two categories of regions in non-pDC cells. The open percentage of a region meant the percentage of cells open on this region. Red: differential open regions of pDC cells only detected by cPeaks; blue: differential open regions of pDC cells detected by both cPeaks and pseudo-bulk peaks.

e. The pseudo-bulk chromatin accessibility in different cell types around the promoter region of PTCRA, shown in the hg19 reference genome.

f. The normalized chromatin accessibility of the cPeak located in the PTCRA promoter region. Each dot represented one cell. The chromatin accessibility was normalized by the frequency-inverse document frequency (TF-IDF) method.

g. The normalized expression of PTCRA in different cell types in scRNA-seq data. The gene expression was normalized by the total expression in each cell and log-transformed. Each dot represented one cell.

h. The UMAP plot of subtypes in GI epithelial cells, clustered using cPeaks as a reference. Each dot represented one cell. Each color represented a cell type provided by the original paper[3]. The black lines labeled four rare subtypes: paneth cell, enterochromaffin cell, tuft cell and colon epithelial cell 3.

Analyze dynamic biological processes and diseases by cPeaks

Although cPeaks were generated by the data of adult healthy samples, they can also handle the analysis of intricate biological states, such as cell differentiation, organ development and disease. We first used cPeaks to construct the differentiation trajectories of human hematopoiesis lineages[47]. These cells displayed continuous state changes in the differential process (Fig. 6a). We aligned the fragments of each cell in the scATAC-seq data to cPeaks and labeled each cell based on its annotation in the original study (Fig. 6b, Methods). We estimated pseudotimes for the erythroid and lymphoid lineage separately using the cell-by-feature matrix generated by cPeaks via a pseudotime construction tool[53] (Fig. 6c, Method). The estimated pseudotime was consistent with the original study[47].

We applied cPeaks to a human fetal retinal dataset[54], which included scATAC-seq covering the early and late stages of retina organogenesis. We used cPeaks as the reference to generate the cell-by-feature matrices. We plotted cells with UMAP and labeled them based on the original paper (Methods). For cells in the early stages (8 weeks) of development (Fig. 6d), we identified three trajectories from multipotent progenitor cells to horizontal cells, cones, and retinal ganglion cells (Fig. 6e). The pseudotime values of these cells were also well arranged (Fig. 6f). For cells in the late stages (13 weeks), we separately analyzed samples collected from the center and the periphery of the retina with cPeaks, and the results showed clear clusters of cell types and their transition states (Fig. S6a-b). It was worth noting that samples used for generating cPeaks did not include fetal stages or retina organs. The results demonstrated the capability of cPeaks for studying fetal development and unexplored organs.

We applied cPeaks to study kidneys with diabetic kidney disease (DKD). DKD is a chronic inflammatory disease which can cause cellular injury and alter cell types within kidneys. We used a scATAC-seq dataset with both healthy kidneys and kidneys with DKD[55]. We selected one healthy and five patient samples for analysis, each with more than 2,000 sequenced cells. Fig. 6g showed the differences of cell types among these samples. We visualized the cells in the healthy (Fig. 6h) and patient samples (Fig. 6i, Fig. S6c-f) separately (Methods). Using cPeaks, we successfully annotated the increased infiltrating leukocytes cells such as B cells, T cells, and mononuclear cells in patient 2 (Fig. 6i). We also found the increased population of VCAM1+ proximal tubule cells in patient 3 (Fig. S6d), which is a marker of kidney injury[56].

We then evaluated the performance of cPeaks on the clustering of tumor cells. We applied cPeaks to a human gynecologic malignancy dataset[57] (Methods). This dataset contains 74,621 cells found in the endometrium and ovary of 11 treatment-naive patients. Fig. 6j and S4g showed that cells analyzed with cPeaks could be well clustered by cell type. Three types of cancer cells (gastrointestinal stromal tumor (GIST), endometrial cancer, and ovarian cancer) were separated from normal cells. We also annotated the cells with copy number variations (CNVs) (Fig. 6k) and expression of KIT, WFDC2, and MUC16 (Fig. S6h) inferred by the original study. We found that the ovarian cancer cells in sample 8 were separated into two clusters (Fig. 6j and S6g). Both the clusters expressed the cancer marker gene WFDC2 (Fig. S6h), but only one of them was enriched with CNVs (Fig. 6k, Methods). The results demonstrated that cPeaks can help to identify subclusters within cancer cells.

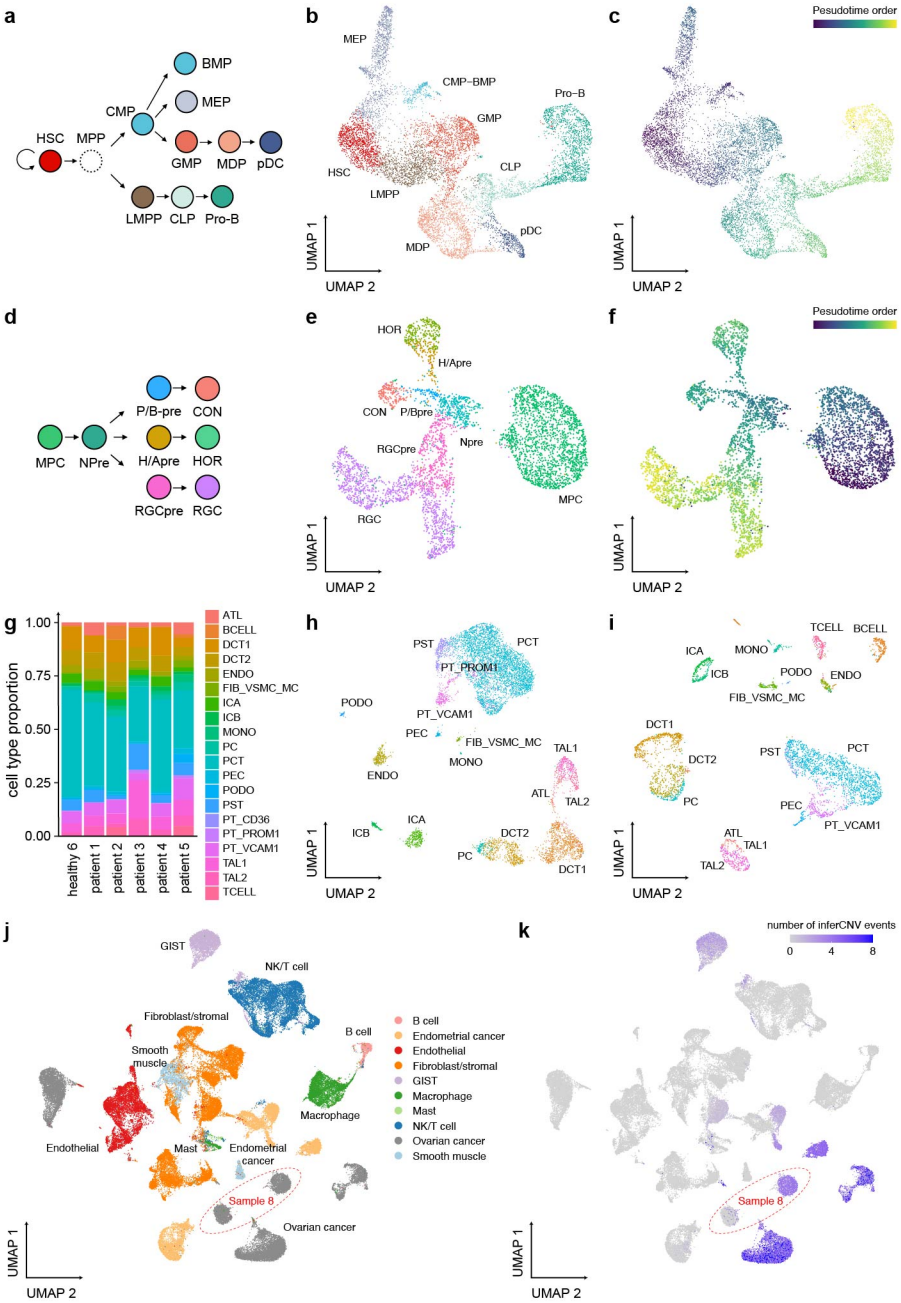

Fig. 6. Analyzing dynamic biological processes and diseases by cPeaks

a. The schematic diagram of cell types in human hematopoiesis lineages. HSC, hematopoietic stem cell; MPP, multipotent progenitor; LMPP, lymphoid-primed multipotent progenitor; CLP, common lymphoid progenitor; Pro-B, megakaryocyte-erythroid progenitor; CMP, basophils and common myeloid progenitor; BMP, basophil-mast cell progenitor; MEP, megakaryocyte-erythroid progenitor; GMP, granulocyte-macrophage progenitors; MDP, monocyte-dendritic cell progenitor. As MPP was not detected in cells, we used a dashed circle to represent it.

b. The UMAP plot of the human hematopoiesis scATAC-seq data colored by cell types.

c. The UMAP plot of the human hematopoiesis scATAC-seq data colored by pseudotime.

d. The schematic diagram of the cell-fate decision in the early (8 weeks) fetal retina. MPC, multipotent progenitor cell; Npre, neurogenic precursor; P/B-pre, photoreceptor/bipolar precursors; CON, cones; H/Apre, the horizontal/amacrine precursors; HOR, horizontal; RGC, ganglion cells; RGCpre, RGC precursors.

e. The UMAP plot of the 8-week retina scATAC-seq data colored by cell types.

f. The UMAP plot of the 8-week retina scATAC-seq data colored by pseudotime.

g. The cell type proportions of each sample in the kidney dataset. ATL, ascending thin limb; BCELL, B-cell; DCT1, DCT1-early distal convoluted tubule; DCT2, DCT2-late distal convoluted tubule; ENDO, ENDO-endothelial cells; FIB_VSMC_MC, vascular smooth muscle cells and mesangial cells; ICA, ICA-type A intercalated cells; ICB, ICB-type B intercalated cells; MONO, mononuclear cells; PC, principal cells; PCT, proximal convoluted tubule; PEC, parietal epithelial cells; PODO, podocytes; PST, proximal straight tubule; PT_CD36, CD36(+) proximal tubule; PT_PROM1, PROM1(+) proximal tubule; PT_VCAM1, VCAM1(+) proximal tubule; TAL1, CLDN16(–) thick ascending limb; TAL2-CLDN16(+) thick ascending limb; TCELL, T cell.

h. The UMAP plot of the scATAC-seq data of the healthy sample in the kidney dataset, colored by cell types.

i. The UMAP plot of the scATAC-seq data of the DKD patient 2 sample in the kidney dataset, colored by cell types.

j. The UMAP plot of the human gynecologic malignancy scATAC-seq data colored by cell types. We aggregated all cells in the endometrium and ovary from 11 patients together. Ovarian cancer cells in sample 8 were circled by the red dashed line. GIST, gastrointestinal stromal tumor; NK/T cell, nature killer cell or T cell.

k. The UMAP plot of the human gynecologic malignancy scATAC-seq data colored by inferred CNVs. The ovarian cancer cells provided by sample 8 were circled by red dashed line. All the results are analyzed with cPeaks as the reference. For (b-c), (e-f), and (h-k), each dot represented one cell.

Discussion

A high-quality reference of chromatin accessibility is crucial for the storage, processing, and analysis of ATAC-seq data, especially rapidly-accumulating scATAC-seq data. We hypothesized that chromatin accessibility profiles across different datasets and cell types share a common open region set and thus proposed cPeaks as a possible reference. We integrated a large number of ATAC-seq datasets to get observed cPeaks and trained a deep CNN network to generate predicted cPeaks. The efficiency of the CNN network suggested that chromatin accessibility was, at least partly, encoded in the DNA sequence, which is also supported by previous studies[44]. The convergent

trend of genomic coverage during sample aggregation and the high correlation of cPeak shapes further indicated that the potential open regions may be consistent across different cell types and datasets. All these evidences demonstrated the rationalization of cPeak design.

cPeaks, as a chromatin accessibility reference, will benefit the analysis of scATAC-seq data to achieve a real single-cell level. Unlike traditional approaches, cPeaks only require read alignments and do not rely on pseudo-bulk peak-calling or iteratively clustering. This simplifies the analysis of scATAC-seq data and facilitates the integration of different sources of scATAC-seq data with high resolution. Moreover, cPeaks can enhance the performance of cell annotation and rare cell type discovery by precisely quantifying ATAC-seq features. As cPeaks do not rely on any pseudo-bulk approach, they can capture regulatory elements that are specific to rare cell types, which helps to discover new cell types. Although cPeaks were defined using healthy adult ATAC-seq data, they also performed well in dynamic and disease scATAC-seq data, such as fetal development and tumor. These findings demonstrated the potential of cPeaks in improving the analysis of scATAC-seq data in diverse contexts.

Though our approach has shown promising results, there are some issues to be refined in future studies. For instance, the generation of predicted cPeaks can be further improved by introducing a probability model to estimate cPeak shapes and thus to refine the locations of cPeaks. Additionally, the CNN model can be improved to generate predicted cPeaks with variable lengths. Furthermore, cPeaks were generated based on the standard genome reference, which may not include disease-specific open regions caused by genetic alterations. Hence, cPeaks can be expanded by integrating genotype data to discover mutation-induced open regions. This will also benefit the studies of disease-associated non-coding genomic variations to reveal the casual paths between genotypes and phenotypes.

We classified cPeaks based on their shapes and annotated their candidate regulatory roles, and more properties can be further studied such as the co-accessibilities of cPeaks across different cell types or organs. Besides, cPeaks in different species can be identified to investigate their properties and regulatory roles in evolution. To facilitate the usage of cPeaks, we have provided an ETF file to store all information of cPeaks as well as mapping and conversion tools for better analyzing scATAC-seq data with cPeaks. In the future, more scATAC-seq analysis tools can be further developed to better utilize cPeaks on different applications, such as data imputation, batch effect removal, and cross-modality integration. we will build a resource to provide integrative, user-friendly information for all cPeaks. Moreover, we can develop a portraiture for each cPeak, similar to gene portraitures in ECA[58], to further illustrate its information, such as the marker cPeaks in each cell type.

Methods

Generation of cPeaks

Bulk ATAC-seq data for cPeak generation

We collected all bulk ATAC-seq data from ENCODE[41] and ATACdb[59]. We selected all adult, healthy samples with more than 10,000 pre-identified peaks. For each sample, if the peak number of chromosome Y (ChrY) was less than

20, we removed all peaks in ChrY, and the remaining peaks were used for further analysis. We analyzed all the data with the hg38 genome reference. For ATACdb samples which were analyzed with the hg19 reference in the original study, we used CrossMap[60] (version 0.6.5) to converted positions of all peaks to the hg38 reference. We removed all peaks that altered lengths during conversion.

Defining PORs and POR shapes

We defined the PORs by joining peaks. Taking advantage of the high quality of ENCODE and the comprehensive cell types of ATACdb, we generated merged peaks for ENCODE and ATACdb separately. The final merged peaks included two parts: (1) all ENCODE merged peaks, and (2) all ATACdb merged peaks that were not overlapped with any ENCODE joint peak. We defined each joint peak as a POR.

When defining POR shapes on ChrY, we only considered the number of samples containing ChrY after quality control as the denominator.

Calculate shape similarity

We generated PORs of blood and solid tissues separately. We only considered PORs with open ratios larger than 0.1. If PORs from the two tissue types respectively were overlapped, we defined them as a set of shared PORs. For each set of shared PORs, we calculated the outermost border among these PORs, and padded each POR's shape to the border with zeros. We selected a padded POR for each tissue type, and calculated PCC between the shapes of these two PORs, regarding as the shape similarity.

Generation of cPeaks from PORs

We used the Hartigan's dip test[61] (the "dip.test" function in the diptest package in R) to measure the multimodality of POR shapes. We defined PORs with p-values less than 0.1 as multimodal PORs.

We split multimodal PORs into unimodal PORs based on their shapes. For the i-th base of a POR, we constructed an array x_i the length of which is the same as the number of samples that showed a peak on this position. All elements in x_i were set as i. We concatenated all x_i together to get an array $\boldsymbol{x}$. We then generated another array y with the same length of $\boldsymbol{x}$, and each element in $\boldsymbol{y}$ was randomly sample from a Gaussian distribution $N(0, 0.1)$. We paired $\boldsymbol{x}$ and $\boldsymbol{y}$ to constructed a two-dimensional matrix. Then, we performed Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)[62] to cluster the matrix, using the "hdbscan" function in the dbscan package in R with variable minimum size of clusters (minPts) set as the length of $\boldsymbol{x}$ dividing by the POR length multiplying by 50. We used the cluster ID calculated by HDBSCAN to split the POR into several separated regions. For each separated region, we decreased the value of minPts by 5 and repeatedly performed clustering and splitting until all separated regions were unimodal or achieved the minimum length of 800 bp. After, all the separated regions were regarded as cPeaks generated from the multimodal POR.

The distribution of sequencing reads around cPeak centers

We aligned the centers of all cPeaks. The reads mapped ±1000 bp around cPeak centers were collected to form an aggregated distribution. We calculated the frequency of read counts on each position, divided it by the total number

of aggregated reads.

Properties of cPeaks

Genomic annotations of cPeaks

We used the ChIPSeeker[63] package to perform genomic annotations of cPeaks. We defined the promoter region as ±2,000 bp around TSS and annotated cPeaks using the "annotatePeak" function. The annotation database was EnsDb.hg38.v86[30] provided by the ensembldb package. ChIPSeeker annotated cPeaks into seven types: promoter, 5' UTR, 3' UTR, exon, intron, downstream (≤ 300 bp downstream of the gene end), and distal intergenic regions. We used the "plotPeakProf2" function in ChIPSeeker to calculate the enrichment of cPeaks on gene body regions, with a confidence interval of 0.95.

Inferring regulatory roles of cPeaks

we defined cPeaks overlapped with any gene promoter region as known promoters. We annotated cPeaks overlapped with any 3' UTR as UTRs. We collected 210 CTCF ChIP-seq data in ENCODE[41] and predicted cPeaks that overlapped with any CTCF ChIP-seq signals as CTCF binding sites. We collected histone modification data in ENCODE[41] and used their called peaks as signals. cPeaks overlapping with H3K4Me1 signals were predicted as enhancers. cPeaks overlapping with H3K4Me3 signals and not overlapping with known promoter regions were regarded as predicted promoters.

Housekeeping cPeaks

We used a pre-defined housekeeping geneset[43] and downloaded it from https://www.tau.ac.il/~elieis/HKG/HK_genes.txt. We obtained a contingency table to describe the amount of housekeeping or non-housekeeping genes related to housekeeping or non-housekeeping cPeaks. We only considered protein coding genes. We used Fisher's exact test to calculate the significance of the contingency table.

We investigated housekeeping cPeaks that were annotated as known promoters, and defined the genes driven by these promoters as genes related to housekeeping cPeaks. We performed functional enrichment analysis of GO biological process terms on all genes related to housekeeping cPeaks. We used the "enrichGO" function in the ChIPSeeker package in R to perform the enrichment analysis and summarized the enriched pathways by the "simplify" function in the clusterProfiler package in R.

Shape patterns of cPeaks

For each cPeak, we defined all the positions with the maximum openness as the plateau. We segmented each cPeak region into three parts according to its shape: an increased edge, a plateau, and a decreased edge. We reversed the direction of the decreased edge. We used linear regression to fit the shape of edges and used the mean squared error (MSE) to measure the performance of fitting. We selected edges with fitting MSEs less than 0.03 and obtained the distribution of the length of these edges. We found the distribution was bimodal and used 170 bp as the threshold to

divide the distribution into two unimodal ones. Edges with lengths less than the threshold were regarded as strongly positioned edges or steep edges, and Edges with lengths greater than the threshold were regarded as weakly positioned edges. For each cPeak, if both edges were strongly positioned, we defined it as a strongly positioned cPeak. If only one side was strongly positioned, we defined it as a one-side strongly positioned cPeak. If both edges were weakly positioned, we defined it as a weakly positioned cPeak.

Motif enrichment analysis around cPeaks

For each cPeak, we extended 200 bp on both endpoints to get an endpoint-around sequence with a length of 401 bp. If the sequence was overlapped with the decreased edge of the cPeak, we converted the sequence into its reverse complementary counterpart. We conducted motif enrichment analysis on the endpoint-around sequences of all cPeaks using the "findMotifsGenome.pl" function in Homer[64] which examined the enrichment of known motifs. After finding the motif, we used the "annotatePeaks.pl function" in Homer to annotate the motif location.

The Epigenetics Transfer Format (ETF)

ETF is a tab-delimited text format based on the General Feature Format (GFF). It contained the basic information and properties of cPeaks. It included:

- **seqname**: The chromosome of the cPeak.
- **source**: The source of the cPeak: "observed" or "predicted".
- **feature**: "cPeaks"
- **start** & **end**: start and end positions of the cPeak.
- **score**, **strand, frame**: NA
- **attribute**: the additional information about cPeaks. It contains "housekeeping" (the housekeeping status), "shapePattern" (the shape pattern) and "inferredElements" (the inferred regulatory elements).

Extension of cPeaks

The architecture of the deep CNN model

The architecture of the deep CNN model was modified based on scBasset[44], a one-dimensional (1D) CNN that included six 1D convolutional blocks and two linear blocks. Each convolutional block comprised a 1D convolution layer, a batch normalization layer, a dropout layer, a max pooling layer, and a GELU activation function. Each linear block contained a linear layer, a batch normalization layer, a dropout layer, and a GELU activation function. The numbers of convolutional kernels of the six 1D convolutional layers were 288, 323, 362, 406, 456, and 512, respectively. The kernel sizes were 17, 5, 5, 5, 5, and 5, respectively. After convolution, a linear layer transformed the 1D output of the last convolutional layer from 5120 to 32 dimensions, and another linear layer transformed the output of the previous linear layer from 32 to 2 dimensions. The final activation function is softmax for binary

classification.

Generation of negative samples

We defined the regions that were not covered by cPeaks as background regions. We generated negative samples from them. We first selected cPeaks and background regions with lengths between 100 bp and 2,000 bp. To make the distribution of negative samples the same as the cPeaks, we divided the range of 100 to 2,000 bp into 100 equal-length bins. We quantified the number of cPeaks and background regions respectively within each bin according to their lengths. For each bin, if the number of background regions is more than the number of cPeaks, we randomly sampled from the background regions to make the sample number the same as the number of cPeaks. If the number of background regions was less than the number of cPeaks, we used all the background regions, and supplemented some regions to make the sample number the same as the number of cPeaks. The supplemented regions were generated by randomly sampling the background regions with lengths greater than 2,000 bp and truncating them into the same length as the cPeaks in the bin. Finally, all these produced background regions collected from each bin were used as the negative samples.

Model training

We used one-hot encoding to encode the sequences of cPeaks (positive samples) and negative samples. We used [0,0,0,0] to represent base N (unknown nucleotide) in any sequence. We padded all sequences with N at their ends to make their lengths equal to 2,000. We split the positive and negative samples into two groups respectively: 80% for training and 20% for test.

We fed these sequences into the model and trained the model to learn the label of these samples (positive samples as 1, and negative samples as 0). We regarded the model output as the openness score of this sample. During training, we used the binary cross-entropy loss function and assessed the loss of the test samples after each epoch. We trained the model for 5 epochs and found it had achieved a stable accuracy.

We used the trained model to predict more cPeaks. We generated candidate sequences by sliding a 500 bp window with a stride of 250 bp in background regions. These candidate sequences were inputted into the trained model to compute the openness score. We selected candidate sequences whose openness scores were larger than 0.55 as predicted cPeaks.

Evaluating the cell annotation performance of cPeaks

Data preprocessing and cell-by-feature matrix generation

We directly downloaded the BAM files of the FACS2-sorting hematological differentiation scATAC-seq data from https://github.com/pinellolab/scATAC-benchmarking/tree/master/Real_Data/Buenrostro_2018/input/sc-bams_nodup, which was provided by a benchmark study[65]. The BAM files contain 2,034 cells with the hg19 reference genome.

We used four features to generate cell-by-feature matrices: genomic bins, pseudo-bulk peaks, merged peaks and cPeaks. Each feature set contained a set of genomic regions on the genome as open regions. For genomic bins, we split the whole genome in the hg19 version into 2,000 bp fix-length bins. Each bin was regarded as a feature. For pseudo-bulk peaks, we used the pre-defined peaks of the original study[9] from the Gene Expression Omnibus (GEO) database under accession code GSE96769. In the original study, all sequenced cells were aggregated as one population for peak calling via MACS2[31]. For merged peaks, we used the peak set provided by the benchmark study[65] from https://github.com/pinellolab/scATAC-benchmarking/blob/master/Real_Data/Buenrostro_2018/input/combined.sorted.merged.bed. In the benchmark study, cells were grouped by their FACS sorting labels. Each group was used to call peaks separately. The called peaks for all groups were merged together as features. For cPeaks, we converted cPeaks from the hg38 version to the hg19 version using CrossMap[60] and used them as features.

For each cell, we mapped the sequencing reads in the BAM file to a feature set and counted the number of reads that overlapped with each feature as its chromatin accessibility. Combining the chromatin accessibility on all features of all cells, we got the cell-by-feature matrix.

Embedding methods for cell representations

ChromVAR-motif[42]: We used the ChromVAR package in R to perform embedding. After adding GC bias and resizing features to 500 bp, we used the "getJasparMotifs" function to get motifs from the JASPAR database and matched them to features via the "matchMotifs" function. Then we mapped sequencing reads to motif features and get the cell-by-motif matrix. We transformed this matrix to the cell-by-deviation Z-score matrix using the "computeDevations" function.

ChromVAR-motif-PCA: The generation of the cell-by-deviation Z-score matrix was the same as the ChromVAR-motif method. We added a principal component analysis (PCA) to this matrix and selected the top 10 principal components (PCs) to generate final cell-by-PC matrix.

ChromVAR-kmer: We used the ChromVAR package in R to perform embedding. After adding GC bias and resizing features to 500 bp, instead of matching features to motifs, we used k-mers ($k = 6$) as features via the "matchKmers" function to generate a cell-by-kmer matrix. Then we transformed this matrix to the cell-by-deviation Z-score matrix.

ChromVAR-kmer-PCA: The generation of the cell-by-deviation Z-score matrix was the same as the ChromVAR-kmer method. We added a Principal component analysis (PCA) to this matrix and selected the top 10 PCs to generate final cell-by-PC matrix.

cisTopic[7]: We applied latent Dirichlet allocation (LDA) to the cell-by-feature matrix to generate the cell-by-topic probability matrix as embedding results. We used the cisTopic package in R and used the "runCGSModels" function to build the model with the number of topics as 10, 20, 25, 30, 35, and 40. We selected the best model using the "selectModel" function. The final normalized cell-by-topic assignment matrix was generated by the "modelMatSelection" function with "method = probability".

Scasat[48]: We calculated cell-cell Jaccard distance based on the cell-by-feature matrix using the "jaccard" function. Then, we performed multidimensional scaling (MDS) to reduce the dimension to 15.

Signac[66]: We first created a Seurat object by the cell-by-feature matrix. Then we used the "RunTFIDF" function in Seurat package to perform frequency-inverse document frequency (TF-IDF) to get a cell-by-normalized feature matrix. Then we run singular value decomposition (SVD) on the cell-by-normalized feature matrix. These two steps were called latent semantic indexing (LSI). We removed the first LSI which was highly correlated with the sequencing depth and selected the 2~50 LSI as the reduced dimension.

TF-IDF-PCA[21,33,67]: We first filtered the features on the sex chromosomes, then performed TF-IDF transformation on the remaining features. Then SVD was performed to generate a cell-by-PC matrix. We selected the top 150 PCs as the embedding features.

Evaluation of cell representations

After embedding, we performed *k*-means clustering with the same number of clusters as the number of FACS-sorted labels. After clustering, we evaluated the clustering results based on three indexes: Adjusted Mutual Information (AMI), Adjusted Rand Index (ARI), and Homogeneity (H). All these indexes were calculated using the sklearn.metrics.cluster library in Python.

Selection of highly variable features (HVFs)

We first applied TF-IDF transformation to normalize the cell-by-feature matrix. We then calculated the variance of each feature across all cells after normalization. We sorted the features by their variance and selected the top features as HVFs.

Rare cell type discovery

Cell type annotation of the PBMC dataset

We download the fragment file of the PBMC dataset from the 10x Genomic website (https://cf.10xgenomics.com/samples/cell-atac/1.0.1/atac_v1_pbmc_10k/atac_v1_pbmc_10k_fragments.tsv.gz). We mapped all fragments to cPeaks and counted the number of fragments that overlapped with each cPeak as its accessibility. We then generated the cell-by-feature matrix for cPeaks. We directly downloaded the matrix using pseudo-bulk peaks as features from the same website (https://cf.10xgenomics.com/samples/cell-atac/1.0.1/atac_v1_pbmc_10k/atac_v1_pbmc_10k_filtered_peak_bc_matrix.h5). Then we followed the steps in the Signac[66] PBMC vignette (https://stuartlab.org/signac/articles/pbmc_vignette.html) to preprocess these two cell-by-feature matrices separately. We then performed clustering in the low-dimensional space to create the UMAP plot. We used graph-based clustering and UMAP visualization on dimensions 2 ~30.

We annotated cells in the scATAC-seq data by integrating them with a labeled PBMC scRNA-seq dataset provided by 10x Genomics. The raw scRNA-seq data was downloaded from https://support.10xgenomics.com/single-cell-

gene-expression/datasets/3.0.0/pbmc_10k_v3, and the pre-processed scRNA-seq Seurat object was downloaded from https://signac-objects.s3.amazonaws.com/pbmc_10k_v3.rds. We utilized the "FindTransferAnchors" function in Seurat for cross-modality integration and used the "TransferData" function in Seurat to infer the labels on scATAC-seq data and their corresponding prediction scores.

Differential accessibility analysis between pDCs and non-pDCs

We performed differential accessibility analysis to find differentially accessible regions in pDCs compared to other cell types. We used "FindMarkers" in the Signac package with test.use = "LR" and latent.vars = "peak_region_fragments" to perform differential analysis.

Cell subtype annotation of GI epithelial cells

We selected cells labeled as GI epithelial cells from CATLAS[3]. We generated cell-by-feature matrices using cPeaks, cCREs defined by CATLAS, and MACS2[31] separately. We selected non-housekeeping cPeaks that were open in more than 50 cells as features to generate the cell-by-feature matrix. We called the peaks in GI epithelial cells using MACS2 with parameters --shift 0 --nomodel --call-summits --nolambda --keep-dup all. We followed the same preprocessing steps in the analysis of PBMCs to deal with each cell-by-feature matrix. After preprocessing, we performed graph-based clustering and UMAP visualizations on dimensions 2:30.

Dynamic biological process and diseases data analysis

Pseudotime calculation of human hematopoiesis lineages

We downloaded the fragment files of the human hematopoiesis data[9] from GEO under accession code GSE129785. We generated the cell-by-feature matrix using cPeaks. We followed the same preprocessing, clustering, and low-dimensional visualization steps as the analysis of the PBMC dataset. We used Monocle 3[53] to build trajectories, following the steps in the Signac trajectory vignettes (https://stuartlab.org/signac/articles/monocle.html). We built trajectories separately for erythroid and lymphoid lineages using the "learn_graph" function, setting hematopoietic stem cells as the root of the trajectory. We then ordered cells with estimated pseudotimes.

Processing and embedding of the human fetal development dataset

We downloaded the fragments files and metadata of the retinal fetal data[54] from GEO under accession code GSE184386. We followed the preprocessing steps in the original study to filter cells. The generation of the cell-by-feature matrix, clustering and visualizations were the same as the analysis of the PBMC dataset. We labeled each cell using labels from the metadata file. All steps were done separately for each sample.

Processing and embedding of the kidney dataset

We downloaded the fragments files of the kidney dataset[55] from GEO under accession code GSE195460. We download the metadata from https://github.com/p4rkerw/Wilson_Muto_NComm_2022. We selected the healthy sample 6 and all the disease samples with more than 2,000 cells for further analysis. We followed the preprocessing

steps in the original study to filter cells. The generation of the cell-by-feature matrix, clustering and visualizations were the same as the analysis of the PBMC dataset. We labeled each cell using labels from the metadata file. All steps were done separately for each sample.

Processing and embedding of the gynecologic malignancy dataset

We downloaded the fragments files and metadata of the gynecologic malignancy dataset[57]from GEO under accession code GSE173682. We combined all cells from 8 patients together to generate the cell-by-feature matrix. We followed the preprocessing steps in the original study to filter cells. The generation of the cell-by-feature matrix, clustering and visualizations were the same as the analysis of the PBMC dataset. We labeled each cell using labels from the metadata file.

Data availability

We only used public datasets in this study. The information of bulk ATAC-seq data for cPeak generation can be found in Table S1. The scATAC-seq and ATAC-seq data for cPeaks evaluation can be found in Table S3. The data used for cPeak annotation can be found in Table S4. We downloaded the bam files of the hematological differentiation scATAC-seq dataset from https://github.com/pinellolab/scATAC-benchmarking/tree/master/Real_Data/Buenrostro_2018/input/sc-bams_nodup. We downloaded the PBMC dataset from the 10x Genomics website https://cf.10xgenomics.com/samples/cell-atac/1.0.1/atac_v1_pbmc_10k/atac_v1_pbmc_10k_fragments.tsv.gz. We downloaded the human hematopoiesis dataset from Gene Expression Omnibus (GEO) under accession code GSE129785. We downloaded the retinal fetal dataset from GEO under accession code GSE184386. We downloaded the kidney dataset from GEO under accession code GSE195460. We downloaded the gynecologic malignancy dataset from GEO under accession code GSE173682.

Code availability

All Code is available on GitHub from the link https://github.com/MengQiuchen/cPeaks, including the generation of observed cPeaks, training of the deep CNN model, all applications on scATAC-seq data, and mapping and conversion tools for analyzing diverse data types using cPeaks.

Acknowledgements

This work was partially supported by National Natural Science Foundation of China (NSFC) (grants 62250005, 61721003 and 62103227), National Key R&D Program of China (grant 2021YFF1200901) and Tsinghua-Fuzhou Institute for Data Technology (TFIDT2021005).

Author contributions

X.Z., L.W., Q.M. and X.Z. conceived the study. Q.M. and X.Z. wrote the manuscript. X.Z. and L.W. supervised the study. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Supplementary materials

Supplementary Figures

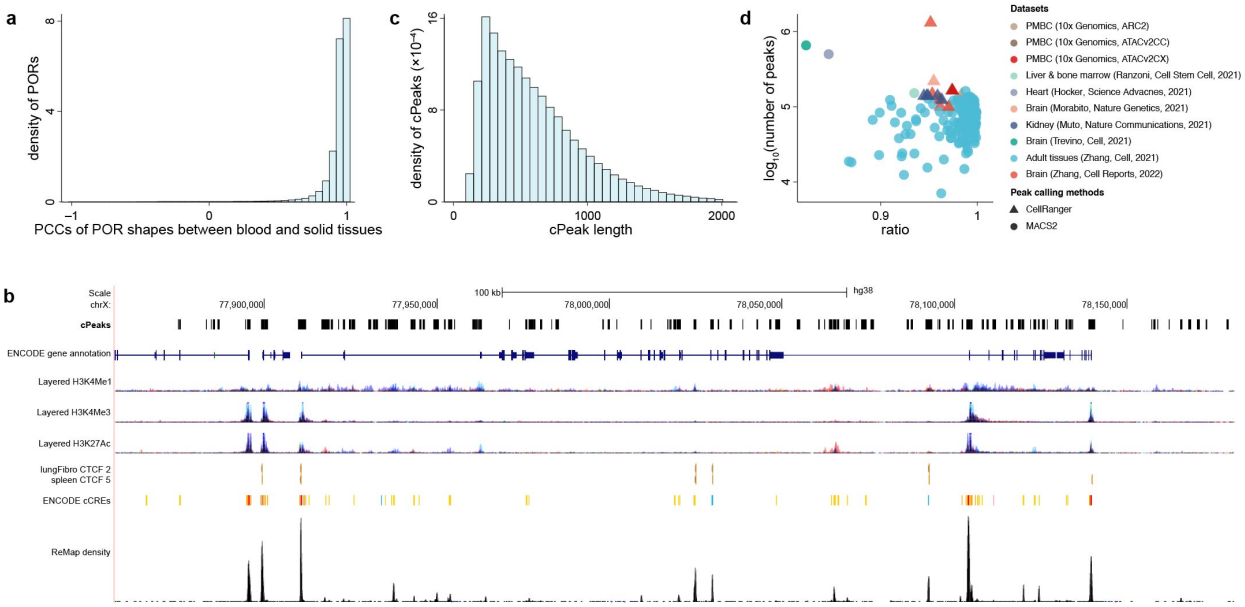

Fig. S1. Details of observed cPeaks

a. PCCs of POR shapes between blood and solid tissues.

b. An example of cPeaks in the UCSC genome browser on the regions around the PGK1 gene (chrX:77,856,686-78,181,002). The selected tracks from top to bottom were cPeaks, ENCODE gene annotation[41], layered H3K4Me1 marker[41], layered H3K4Me3 marker[41], layered H3K27Ac marker[41], lung fibroblast CTCF signal[41], spleen CTCF signal[41], ENCODE cCREs[41] and ReMap density[68]. The layered H3K4Me1/H3K3Me3/H3K27Ac marker combined the corresponding histone modification signals of 7 cell lines (GM12878, H1-hESC, HSMM, HUVEC, K562, NHEK, NHLF) in ENCODE together. Different colors and layers represented different cell lines. The ReMap density track represented the density of

overlapped transcriptional regulators provided by ReMap Atlas[68].

c. The length distribution of observed cPeaks. cPeaks with lengths less than 2000 bp were shown.

d. The coverage of cPeaks in 231 scATAC-seq data. The colors represented different dataset sources, details of which can be found in Table S3.

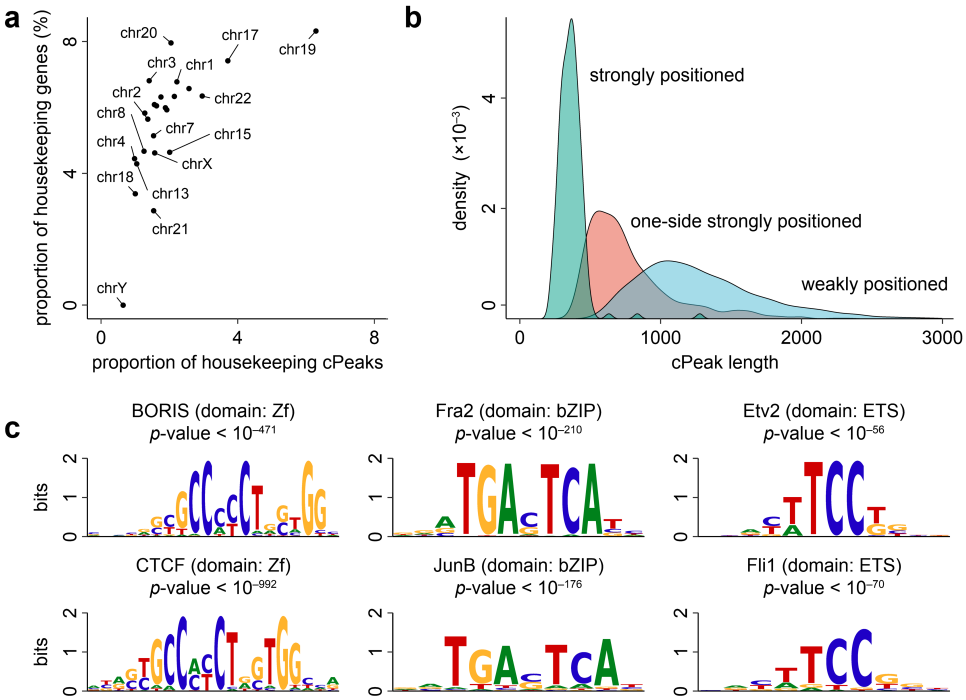

Fig. S2. Properties of observed cPeaks

a. The proportion of housekeeping cPeaks and housekeeping genes on each chromosome. Each dot represented a chromosome. For each chromosome, the proportion was calculated by the number of housekeeping cPeaks/genes divided by the total number of cPeaks/genes on the chromosome. The PCC between the proportion of housekeeping cPeaks and housekeeping genes is 0.64.

b. The length distribution of cPeaks with different shape patterns. cPeaks with both edges showing fitting MSEs less than 0.03 were used to illustrate the distribution (see Methods for details).

c. Examples of enriched motifs found around the steep edges of cPeaks.

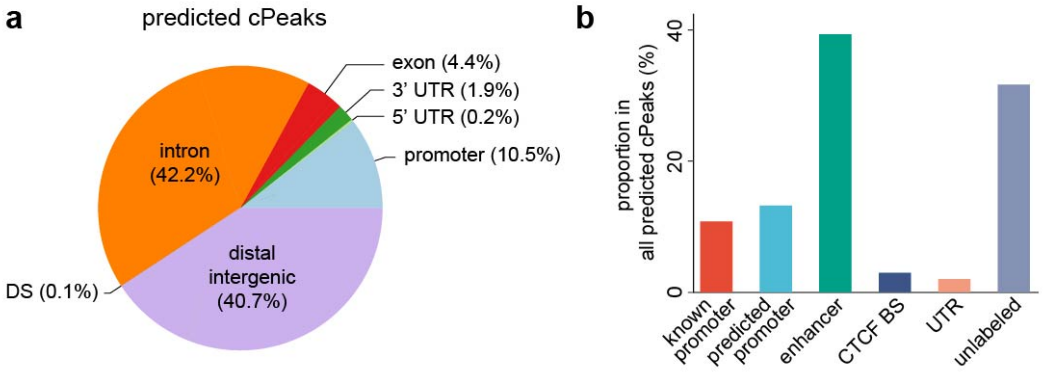

Fig. S3. Properties of predicted cPeaks

a. Genomic annotation of predicted cPeaks. DS: downstream (≤ 300 bp downstream of the gene end).

b. The inferred regulatory roles of observed cPeaks. Each cPeak may have multiple inferred functions. CTCF BS: CTCF binding site.

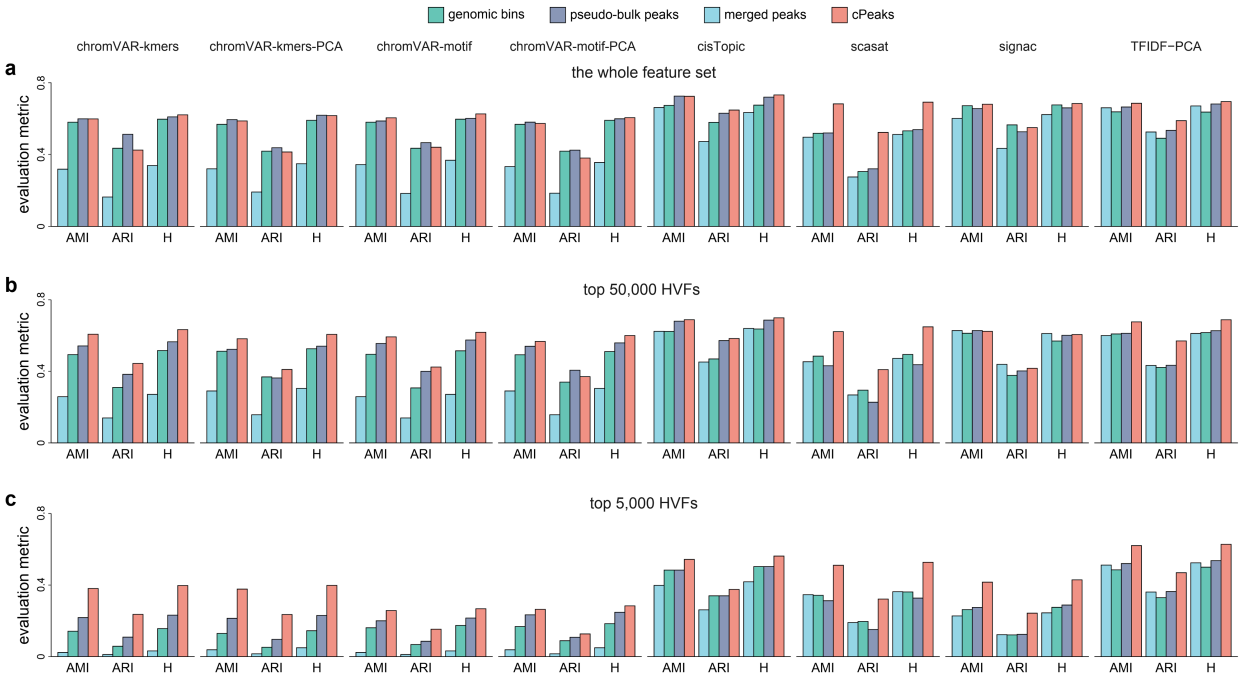

Fig. S4. The performance of cell annotations on FACS2-sorted hematopoietic differentiation data

a. Performance of different feature-defining approaches with different evaluation metrics via different embedding methods. All features were used for analysis.

b. Performance of different feature-defining approaches with different evaluation metrics via different embedding methods. The top 50,000 HVFs were used for analysis.

c. Performance of different feature-defining approaches with different evaluation metrics via different embedding methods. The top 5,000 HVFs were used for analysis.

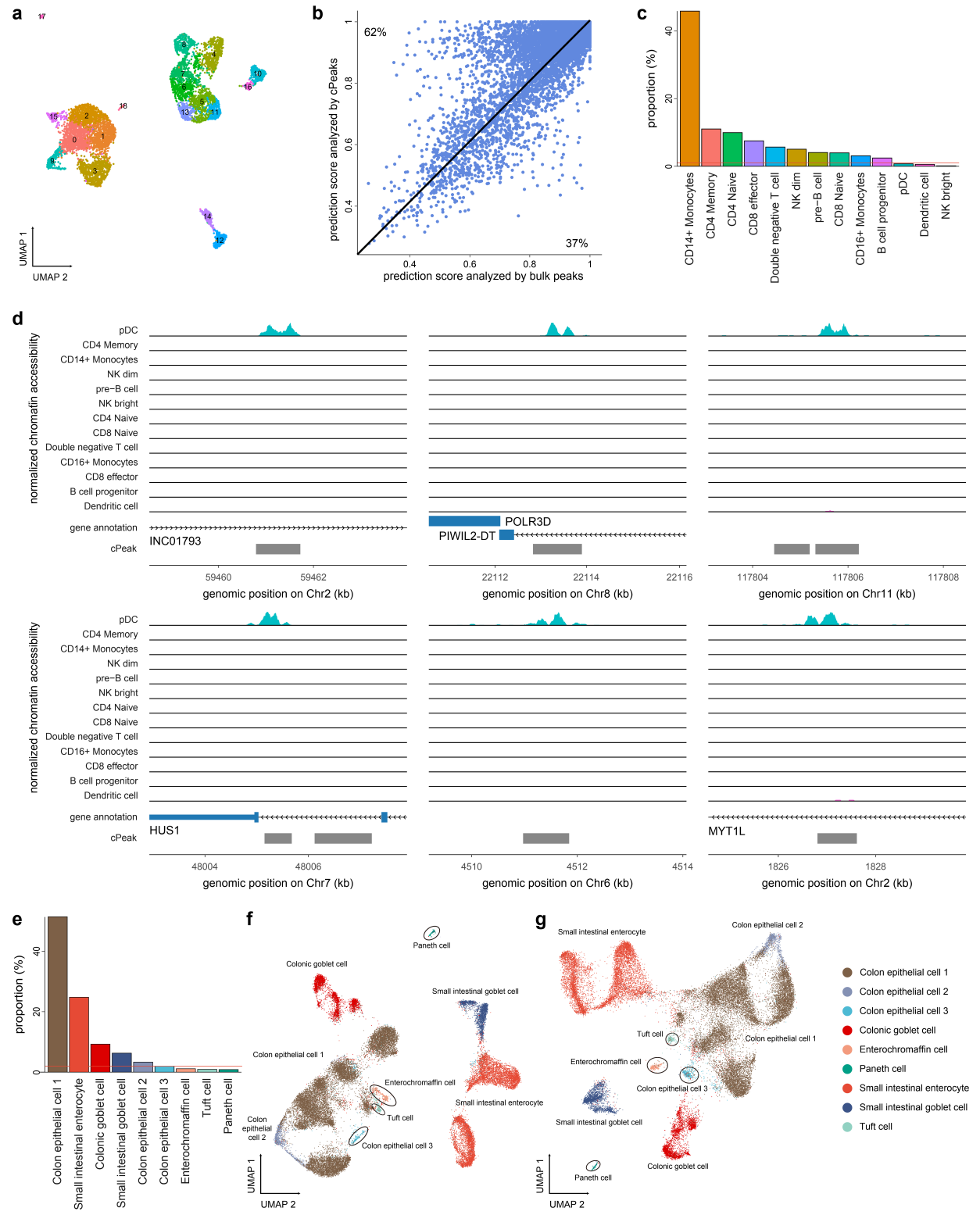

Fig. S5. Rare cell type discovery using cPeaks

a. The UMAP plot of the PBMC scATAC-seq data analyzed by cPeaks. Each dot represented one cell.

Different colors represented different clusters.

b. Prediction scores of cells in the PBMC scATAC-seq data calculated by cPeaks and pseudo-bulk peaks. Each dot represented one cell.

c. The proportion of each cell type in the PBMC scATAC-seq data. Three rare cell types were less than 1% (the red line): pDCs, dendritic cells and NK bright cells.

d. Genomic tracks of six differentially accessible regions only found by cPeaks, shown in the hg19 reference genome. We selected six regions with the lowest *p*-values. Genes were annotated by RefSeq[69].

e. The proportion of each subtype in GI epithelial cells. Four rare subtypes were less than 2% (the red line): paneth cell, enterochromaffin cell, tuft cell and colon epithelial cell 3.

f. The UMAP plot of subtypes in GI epithelial cells using merged peaks in CATLAS. Each dot represented one cell. Each color represented a cell type provided by the original study[3]. The black lines labeled the four rare subtypes.

g. The UMAP plot of subtypes in GI epithelial cells using pseudo-bulk peaks called by the population of all GI epithelial cells. Each dot represented one cell. Each color represented a cell type provided by the original paper[3]. The black lines labeled the four rare subtypes.

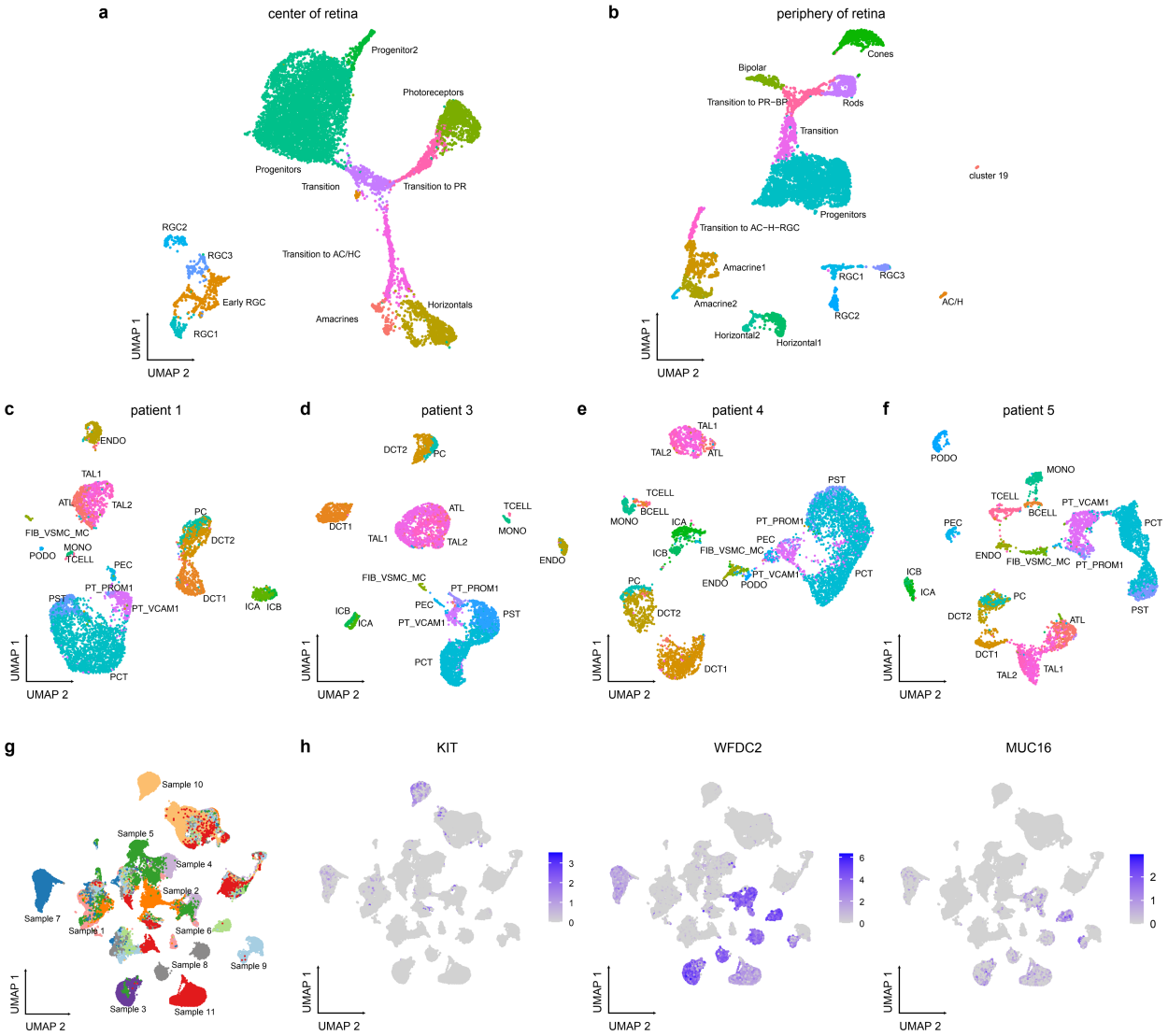

Fig. S6. Analyzing dynamic biological processes and diseases by cPeaks

a. The UMAP plot of the 13-week scATAC-seq data collected from the center of retina colored by cell types. PR, photoreceptors; AC, amacrine cell; HC, Horizontal cell; RGC, ganglion cell.

b. The UMAP plot of the 13-week scATAC-seq data collected from the periphery of retina colored by cell types. PR, Photoreceptors; BP, Bipolar and Rods; A-H-RGC, amacrine cell, horizontal cell or ganglion cell.

c-f. The UMAP plot of the scATAC-seq data of the DKD patient 1, 3, 4, and 5 sample in the kidney dataset, colored by cell types.

g. The UMAP of the human gynecologic malignancy scATAC-seq data colored by sample IDs.

h. The UMAP of the human gynecologic malignancy scATAC-seq data colored by inferred gene expression. For all figures, each dot represented one cell and the abbreviations of label were same as Fig. 6.

Supplementary Tables

Table S1. Description of bulk ATAC-seq data for cPeak generation.

Table S2. Genomic locations of observed cPeaks.

Table S3. Description of scATAC-seq and bulk ATAC-seq data for cPeak evaluation.

Table S4. List of epigenetic and CTCF ChIP-seq data for cPeak annotations.

Table S5. Genomic locations and openness scores of predicted cPeaks.

Table S6. Information of all cPeaks in the ETF format.

References

1. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet*. 2019;20(4):207-220. doi:10.1038/s41576-018-0089-8

2. Preissl S, Gaulton KJ, Ren B. Characterizing cis-regulatory elements using single-cell epigenomics. *Nat Rev Genet*. Published online July 15, 2022. doi:10.1038/s41576-022-00509-1

3. Zhang K, Hocker JD, Miller M, et al. A single-cell atlas of chromatin accessibility in the human genome. *Cell*. 2021;184(24):5985-6001.e19. doi:10.1016/j.cell.2021.10.024

4. Pliner HA, Packer JS, McFaline-Figueroa JL, et al. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell*. 2018;71(5):858-871.e8. doi:10.1016/j.molcel.2018.06.044

5. Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin*. 2014;7(1):33. doi:10.1186/1756-8935-7-33

6. Duren Z, Chen X, Jiang R, Wang Y, Wong WH. Modeling gene regulation from paired expression and chromatin accessibility data. *Proc Natl Acad Sci*. 2017;114(25):E4914-E4923. doi:10.1073/pnas.1704553114

7. Bravo González-Blas C, Minnoye L, Papasokrati D, et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat Methods*. 2019;16(5):397-400. doi:10.1038/s41592-019-0367-1

8. Buenrostro JD, Wu B, Litzenburger UM, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*. 2015;523(7561):486-490. doi:10.1038/nature14590

9. Buenrostro JD, Corces MR, Lareau CA, et al. Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell*. 2018;173(6):1535-1548.e16. doi:10.1016/j.cell.2018.03.074

10. Corces MR, Buenrostro JD, Wu B, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet*. 2016;48(10):1193-1203. doi:10.1038/ng.3646

11. Preissl S, Fang R, Zhao Y, et al. *Single Nucleus Analysis of the Chromatin Landscape in Mouse Forebrain Development*. Genomics; 2017. doi:10.1101/159137

12. Trevino AE, Sinnott-Armstrong N, Andersen J, et al. Chromatin accessibility dynamics in a model of human forebrain development. *Science*. 2020;367(6476):eaay1645. doi:10.1126/science.aay1645

13. Morabito S, Miyoshi E, Michael N, et al. Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nat Genet*. 2021;53(8):1143-1155. doi:10.1038/s41588-021-00894-z

14. Turner AW, Hu SS, Mosquera JV, et al. Single-nucleus chromatin accessibility profiling highlights regulatory mechanisms of coronary artery disease risk. *Nat Genet*. 2022;54(6):804-816. doi:10.1038/s41588-022-01069-0

15. Sanghi A, Gruber JJ, Metwally A, et al. Chromatin accessibility associates with protein-RNA correlation in human cancer. *Nat Commun*. 2021;12(1):5732. doi:10.1038/s41467-021-25872-1

16. Corces MR, Granja JM, Shams S, et al. The chromatin accessibility landscape of primary human cancers. *Science*. 2018;362(6413):eaav1898. doi:10.1126/science.aav1898

17. Wilson PC, Muto Y, Wu H, Karihaloo A, Waikar SS, Humphreys BD. Multimodal single cell sequencing implicates chromatin accessibility and genetic background in diabetic kidney disease progression. *Nat Commun*. 2022;13(1):5253. doi:10.1038/s41467-022-32972-z

18. Minnoye L, Marinov GK, Krausgruber T, et al. Chromatin accessibility profiling methods. *Nat Rev Methods Primer*. 2021;1(1):10. doi:10.1038/s43586-020-00008-9

19. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol*. 2015;109(1):21.29.1-21.29.9. doi:https://doi.org/10.1002/0471142727.mb2129s109

20. Grandi FC, Modi H, Kampman L, Corces MR. Chromatin accessibility profiling by ATAC-seq. *Nat Protoc*. 2022;17(6):1518-1552. doi:10.1038/s41596-022-00692-9

21. Cusanovich DA, Daza R, Adey A, et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*. 2015;348(6237):910-914. doi:10.1126/science.aab1601

22. Zhu C, Yu M, Huang H, et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat Struct Mol Biol*. 2019;26(11):1063-1070. doi:10.1038/s41594-019-0323-x

23. Lareau CA, Duarte FM, Chew JG, et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat Biotechnol*. 2019;37(8):916-924. doi:10.1038/s41587-019-0147-6

24. Mimitou EP, Lareau CA, Chen KY, et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat Biotechnol*. 2021;39(10):1246-1258. doi:10.1038/s41587-021-00927-2

25. Cao J, Cusanovich DA, Ramani V, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*. 2018;361(6409):1380-1385. doi:10.1126/science.aau0730

26. Chen X, Miragaia RJ, Natarajan KN, Teichmann SA. A rapid and robust method for single cell chromatin accessibility profiling. *Nat Commun*. 2018;9(1):5345. doi:10.1038/s41467-018-07771-0

27. Fullard JF, Hauberg ME, Bendl J, et al. An atlas of chromatin accessibility in the adult human brain. *Genome Res*. 2018;28(8):1243-1252. doi:10.1101/gr.232488.117

28. Domcke S, Hill AJ, Daza RM, et al. A human cell atlas of fetal chromatin accessibility. *Science*. 2020;370(6518):eaba7612. doi:10.1126/science.aba7612

29. Markenscoff-Papadimitriou E, Whalen S, Przytycki P, et al. A Chromatin Accessibility Atlas of the Developing Human Telencephalon. *Cell*. 2020;182(3):754-769.e18. doi:10.1016/j.cell.2020.06.002

30. Cunningham F, Allen JE, Allen J, et al. Ensembl 2022. *Nucleic Acids Res*. 2022;50(D1):D988-D995. doi:10.1093/nar/gkab1049

31. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. *Nat Protoc*. 2012;7(9):1728-1740. doi:10.1038/nprot.2012.101

32. Satpathy AT, Granja JM, Yost KE, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol*. 2019;37(8):925-936. doi:10.1038/s41587-019-0206-z

33. Cusanovich DA, Hill AJ, Aghamirzaie D, et al. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell*. 2018;174(5):1309-1324.e18. doi:10.1016/j.cell.2018.06.052

34. Wang A, Chiou J, Poirion OB, et al. Single-cell multiomic profiling of human lungs reveals cell-type-specific and age-dynamic control of SARS-CoV2 host genes. Morrisey EE, De Langhe S, eds. *eLife*. 2020;9:e62522. doi:10.7554/eLife.62522

35. Moore JE, Purcaro MJ, Pratt HE, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*. 2020;583(7818):699-710. doi:10.1038/s41586-020-2493-4

36. Nair VD, Vasoya M, Nair V, et al. Differential analysis of chromatin accessibility and gene expression profiles identifies cis-regulatory elements in rat adipose and muscle. *Genomics*. 2021;113(6):3827-3841. doi:10.1016/j.ygeno.2021.09.013

37. Yan F, Powell DR, Curtis DJ, Wong NC. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol*. 2020;21(1):22. doi:10.1186/s13059-020-1929-3

38. Daugherty AC, Yeo RW, Buenrostro JD, Greenleaf WJ, Kundaje A, Brunet A. Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans. *Genome Res*. 2017;27(12):2096-2107. doi:10.1101/gr.226233.117

39. Ross-Innes CS, Stark R, Teschendorff AE, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*. 2012;481(7381):389-393. doi:10.1038/nature10730

40. Wang F, Bai X, Wang Y, et al. ATACdb: a comprehensive human chromatin accessibility database. *Nucleic Acids Res*. 2021;49(D1):D55-D64. doi:10.1093/nar/gkaa943

41. Dunham I, Kundaje A, Aldred SF, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74. doi:10.1038/nature11247

42. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods*. 2017;14(10):975-978. doi:10.1038/nmeth.4401

43. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. *Trends Genet*. 2013;29(10):569-574. doi:10.1016/j.tig.2013.05.010

44. Yuan H, Kelley DR. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nat Methods*. 2022;19(9):1088-1096. doi:10.1038/s41592-022-01562-8

45. Zou Z, Ohta T, Miura F, Oki S. ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. *Nucleic Acids Res*. 2022;50(W1):W175-W182. doi:10.1093/nar/gkac199

46. Fang R, Preissl S, Li Y, et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat Commun*. 2021;12(1):1337. doi:10.1038/s41467-021-21583-9

47. Satpathy AT, Granja JM, Yost KE, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol*. 2019;37(8):925-936. doi:10.1038/s41587-019-0206-z

48. Baker SM, Rogerson C, Hayes A, Sharrocks AD, Rattray M. Classifying cells with Scasat, a single-cell ATAC-seq analysis tool. *Nucleic Acids Res*. 2019;47(2):e10-e10. doi:10.1093/nar/gky950

49. Chen H, Lareau C, Andreani T, et al. Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol*. 2019;20(1):241. doi:10.1186/s13059-019-1854-5

50. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. *Cell*. 2019;177(7):1888-1902.e21. doi:10.1016/j.cell.2019.05.031

51. Reizis B. Regulation of plasmacytoid dendritic cell development. *Curr Opin Immunol*. 2010;22(2):206-211. doi:10.1016/j.coi.2010.01.005

52. Ghosh HS, Cisse B, Bunin A, Lewis KL, Reizis B. Continuous Expression of the Transcription Factor E2-2 Maintains the Cell Fate of Mature Plasmacytoid Dendritic Cells. *Immunity*. 2010;33(6):905-916. doi:10.1016/j.immuni.2010.11.023

53. Cao J, Spielmann M, Qiu X, et al. The single cell transcriptional landscape of mammalian organogenesis. *Nature*. 2019;566(7745):496-502. doi:10.1038/s41586-019-0969-x

54. Finkbeiner C, Ortuño-Lizarán I, Sridhar A, Hooper M, Petter S, Reh TA. Single-cell ATAC-seq of fetal human retina and stem-cell-derived retinal organoids shows changing chromatin landscapes during cell fate acquisition. *Cell Rep*. 2022;38(4):110294. doi:10.1016/j.celrep.2021.110294

55. Wilson PC, Muto Y, Wu H, Karihaloo A, Waikar SS, Humphreys BD. Multimodal single cell sequencing implicates chromatin accessibility and genetic background in diabetic kidney disease progression. *Nat Commun*. 2022;13(1):5253. doi:10.1038/s41467-022-32972-z

56. Muto Y, Wilson PC, Ledru N, et al. Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nat Commun*. 2021;12:2190. doi:10.1038/s41467-021-22368-w

57. Regner MJ, Wisniewska K, Garcia-Recio S, et al. A multi-omic single-cell landscape of human gynecologic malignancies. *Mol Cell*. 2021;81(23):4924-4941.e10. doi:10.1016/j.molcel.2021.10.013

58. Chen S, Luo Y, Gao H, et al. hECA: The cell-centric assembly of a cell atlas. *iScience*. 2022;25(5). doi:10.1016/j.isci.2022.104318

59. Wang F, Bai X, Wang Y, et al. ATACdb: a comprehensive human chromatin accessibility database. *Nucleic Acids Res*. 2021;49(D1):D55-D64. doi:10.1093/nar/gkaa943

60. Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*. 2014;30(7):1006-1007. doi:10.1093/bioinformatics/btt730

61. Hartigan JA, Hartigan PM. The Dip Test of Unimodality. *Ann Stat*. 1985;13(1). doi:10.1214/aos/1176346577

62. Malzer C, Baum M. A Hybrid Approach To Hierarchical Density-based Cluster Selection. In: *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. ; 2020:223-228. doi:10.1109/MFI49285.2020.9235263

63. Yu G, Wang LG, He QY. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinforma Oxf Engl*. 2015;31(14):2382-2383. doi:10.1093/bioinformatics/btv145

64. Heinz S, Benner C, Spann N, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell*. 2010;38(4):576-589. doi:10.1016/j.molcel.2010.05.004

65. Chen H, Lareau C, Andreani T, et al. Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol*. 2019;20(1):241. doi:10.1186/s13059-019-1854-5

66. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Single-cell chromatin state analysis with Signac. *Nat Methods*. 2021;18(11):1333-1341. doi:10.1038/s41592-021-01282-5

67. Cusanovich DA, Reddington JP, Garfield DA, et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature*. 2018;555(7697):538-542. doi:10.1038/nature25981

68. Chèneby J, Gheorghe M, Artufel M, Mathelier A, Ballester B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res*. 2018;46(D1):D267-D275. doi:10.1093/nar/gkx1092

69. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44(D1):D733-745. doi:10.1093/nar/gkv1189