The full set of potential open regions (PORs) in the human genome defined by consensus peaks of ATAC-seq data

Qiuchen Meng[1,#], Xinze Wu[1,#], Chen Li[1], Jiaqi Li[1,4], Xi Xi[1], Sijie Chen[1], Shengquan Chen[2], Jiaqi Li[1,5], Xiaowo Wang[1], 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
4. ORCID: 0000-0001-9038-9010
5. ORCID: 0000-0001-8967-5714

[#] These authors contributed equally.

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

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. Based on a systematic analysis of 1,785 ATAC-seq and 231 scATAC-seq datasets, we found that cells share the same set of potential open regions (PORs) on the genome. We proposed a unified reference called consensus peaks (cPeaks) to represent PORs across all observed cell types, and developed a deep-learning model to predict cPeaks unseen in the collected data. The observed and predicted cPeaks defined a full set of PORs in the human genome, which can be used as a reference for all ATAC-seq data to align to. Experiments showed that using this reference to integrate scATAC-seq data can improve cell annotation and facilitate the discovery of rare cell types. cPeaks also performed well in analyzing dynamic biological processes and diseases. The analyses and experiments suggested PORs represent a set of inherent functional regions in the human genome.

Introduction

Chromatin accessibility refers to the availability of chromatinized DNA to be directly interacted with other macromolecules[1]. The accessible regions, also known as open regions, imply positions of many transcriptional regulatory elements such as promoters, enhancers, insulators, and transcription factor (TF) binding sites[2,3]. Studying chromatin accessibility in cells can help to infer gene regulations[4–8] in many 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. The single-cell ATAC-seq or scATAC-seq[21–26] is becoming a major source of information besides single-cell RNA sequencing (scRNA-seq) for investigating the molecular properties of cells of different types or states and 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 the genome. 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]. 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 essential for cross-dataset comparison and cell atlas construction of ATAC-seq data. Current practices are to call peaks in each individual dataset, and then to integrate peaks from multiple datasets. 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.

We collected 624 ATAC-seq datasets of multiple tissues to study the distribution of chromatin open regions on the genome across different tissues and defined a set of consensus peaks or cPeaks, and collected 231 scATAC-seq datasets of 2.4 million cells to validate the use of cPeaks as generic reference of chromatin accessibility. We found that the cPeaks represent the potential open regions or PORs inherent in the genome. We confirmed the association

between the positions of cPeaks and DNA sequence patterns on the genome. We trained a deep learning model with sequence features to predict new cPeaks on the genome that have not been seen in the collected data, and validated the performance of prediction on another 1,161 ATAC-seq datasets. The combination of the ~1.4M observed and ~0.2M predicted cPeaks provides the full set of PORs in the human genome, shared by all tissues and cell types. The cPeaks can be used as the generic reference of ATAC-seq data for read alignments. We provided bioinformatics tools for mapping or converting diverse types of ATAC-seq to cPeaks. We conducted a series of experiments to validate 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 proven to facilitate the analysis of complex biological states, such as cell differentiation, organ development and disease. The experiments further strengthened the hypothesis that PORs are a new set of inherent features on the genome.

Results

Defining cPeaks by aggregating ATAC-seq data

For building a generic reference of chromatin accessibility, we first collected 624 high-quality bulk ATAC-seq datasets from ENCODE[40] and ATACdb[41], covering over 40 human organs and all body systems (Methods, Table S1). Each dataset contained a set of pre-identified peaks, where each peak represented an open region on the genome observed in a dataset (Fig. 1a). We merged the peaks one dataset by one dataset according to their genomic locations to produce a pool of all possible peaks. Non-overlapping peaks were directly added to the pool and overlapping peaks were joined as wider ones. During this process, we observed that the number of total peaks and their genomic coverage grew when new datasets were added at beginning, but the growth soon tended to saturate (Fig. 1b). This phenomenon was independent with the order we added the datasets. It suggested that the number of all possible open regions on the genome is limited and their locations are not random or uniform.

After adding all datasets, we identified ~1.2 million merged peaks covering about 30% of the genome (Methods). A merged peak describes a genomic region that could open in one or more cell types. For each position of a merged peak, we calculated the proportion of datasets that had peaks covering this position, and denoted this proportion as the openness score of the merged peak at this position. We summarized an array of the openness scores at all positions of a merged peak along the 5’-to-3’ direction and call it as the shape of the peak (Methods).

The datasets used to generate merged peaks were of two major tissue types: ~51% from blood and ~49% from solid tissues (lung, heart, brain, etc.). We used the same procedure to merge peaks for these two tissue types separately. The numbers of merged peaks shared by the two groups took up 87% and 76% of the identified merged peaks in the blood data and in the solid tissue data, respectively. We found the common merged peaks found from the two groups at the same location tended to have the same shape (an example in Fig. 1c). We measured the shape similarity of each common merged peak by the Pearson correlation coefficient (PCC) of its shapes calculated in the blood or solid tissues (Methods). The common merged peaks showed high PCCs between the two tissue types, with an average of 0.93 and a median of 0.96 (Fig. S1a). This indicated that the shape of a merged peak might be an intrinsic

property irrelevant to the cell or tissue types.

We found that the shapes of merged peaks were of two major types: unimodal or multimodal (Fig. 1a, Methods). We assumed that a multimodal merged peak was composed of adjacent unimodal peaks, and therefore split multimodal merged peaks into multiple unimodal ones (Fig. 1a, Methods). We took the unimodal merged peaks as the basic unit of the chromatin accessibility reference and named them as consensus peaks or cPeaks. In this way, we obtained ~1.4 million cPeaks in total (Table S2, an example provided in Fig. S1b) based on the 624 collected ATAC-seq datasets. The lengths of most cPeaks are less than 2,000 bp, with median length of 522 bp. This is similar to the previous observations that the peak lengths in ATAC-seq data were about 500 bp[42] (Fig. S1c).

We assessed the performance of using cPeaks as a chromatin accessibility reference to process scATAC-seq data. We collected 231 scATAC-seq datasets from published studies (Table S3), each with a set of peaks pre-identified with different peak-calling methods[31,32]. These data contained about 2.4 million cells. For each dataset, we calculated the proportion of the pre-identified peaks that overlapped with cPeaks and defined this proportion as the coverage of cPeaks in the dataset. Most datasets showed coverages greater than 0.9, with average of 0.98 and median of 0.99 (Fig. 1d and S1d). We selected 27 scATAC-seq datasets 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). On average, 68.8% of sequencing reads in the scATAC-seq data were mapped to the observed cPeaks, higher than the typical mapping rate of ~65% as reported in the technical reports of 10X Genomics[43,44]. This indicates that cPeaks are a reliable set of reference peaks for recovering the signals in scATAC-seq data.

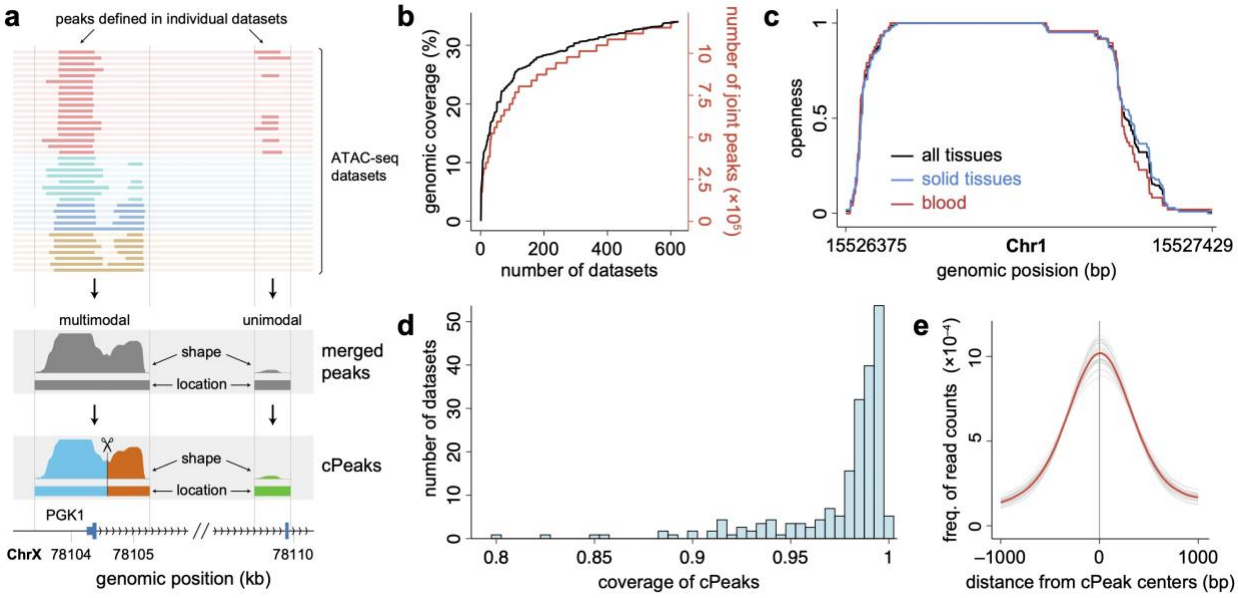

Fig 1. The computation diagram of cPeaks and their basic properties.

(a) The schematic diagram of the computation of cPeaks. A part of the genomic region near PGK1 is shown as an example. In the

upper panel, each row represents an ATAC-seq dataset, and each segment represented a pre-identified peak. Different colors indicate different cell types. The middle panel illustrates a multimodal merged peak (left) and a unimodal merged peak (right). In the bottom panel, the multimodal merged peak is clipped as two cPeaks, and the unimodal merged peak forms one peak. The relative positions of the cPeaks with regard to the transcription start site (TSS) and the 2nd exon of PGK1 are shown below the cPeaks. (b) Increase tendency of the genomic coverage (black line) and number of merged peaks (red line) with the addition of more datasets. The x-axis represents the number of aggregated datasets. The left y-axis represents the total genomic coverage of all merged peaks on each step. The right y-axis represents the number of merged peaks on each step. (c) An example of the shapes of merged peaks on chromosome 1, obtained with data of only blood samples (red line), of all solid tissues (blue line), or of all tissues (black line). The shapes obtained with different data are highly correlated along the genome with Pearson correlations of 0.99. (d) Histogram of the coverages of cPeaks in the 231 scATAC-seq datasets. The x-axis represents the coverage of cPeaks, which is the number of overlapped pre-defined peaks with cPeaks divided by the total number of pre-defined peaks in the original data. The y-axis represents the number of datasets. Most datasets get more than 90% coverage and the median coverage is 0.99. (e) Enrichment of sequencing reads around centers of cPeaks. All sequencing reads overlapped with ±1000 bp genomic regions around cPeak centers are aggregated to form the distributions. Each grey line represents a dataset. The red line shows the average distribution of all datasets.

Properties of observed cPeaks

We investigated cPeaks' associations with known regulatory elements or biological functions. By comparing cPeaks with known genomic annotations, we found that cPeaks are enriched around transcriptional start sites (TSSs) and transcription end sites (TESs) (Fig. 2a, Methods). About 13% of cPeaks are 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 collected multiple types of epigenetic annotations, including histone modification and CTCF ChIP-seq data (Table S4). Using the genomic and epigenetic annotations, we found that most cPeaks contain or overlap with one or more regulatory elements, such as enhancers, promoters (including known and predicted promoters), untranslated regions (UTRs), or CTCF binding sites (CTCF BSs) (Methods, Fig. 2c).

For each cPeak, we defined the maximum value of its shape as its "open ratio" to reflect the ratio of datasets opening in this cPeak (Methods). We defined a cPeak with open ratio > 90% as a "housekeeping" cPeak (Fig. 2d). A total of 24,538 of the observed cPeaks (1.8%) 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[45] (Fisher's exact test p-value $< 1\times10^{-5}$, Methods). The proportion of housekeeping cPeaks and the proportion of housekeeping genes on different chromosomes were highly correlated (Fig. S2a). We 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 of both edges of the shapes (Methods). We named them as "strongly positioned", "one-side strongly positioned", and "weakly positioned" (Fig. 2g and S2b), and hypothesized that these different patterns were related to arrangements of the

nucleosomes around cPeaks. cPeaks with steep edges signified more consistent boundaries of open regions in different experiments and suggested the presence of strongly-positioned nucleosomes around the boundaries. Conversely, a gentle sloop suggested that the nucleosomes around the cPeak boundaries were weakly positioned. We conducted motif enrichment analysis around edges of cPeaks (Methods) and 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 inside the boundaries (Fig. 2h).

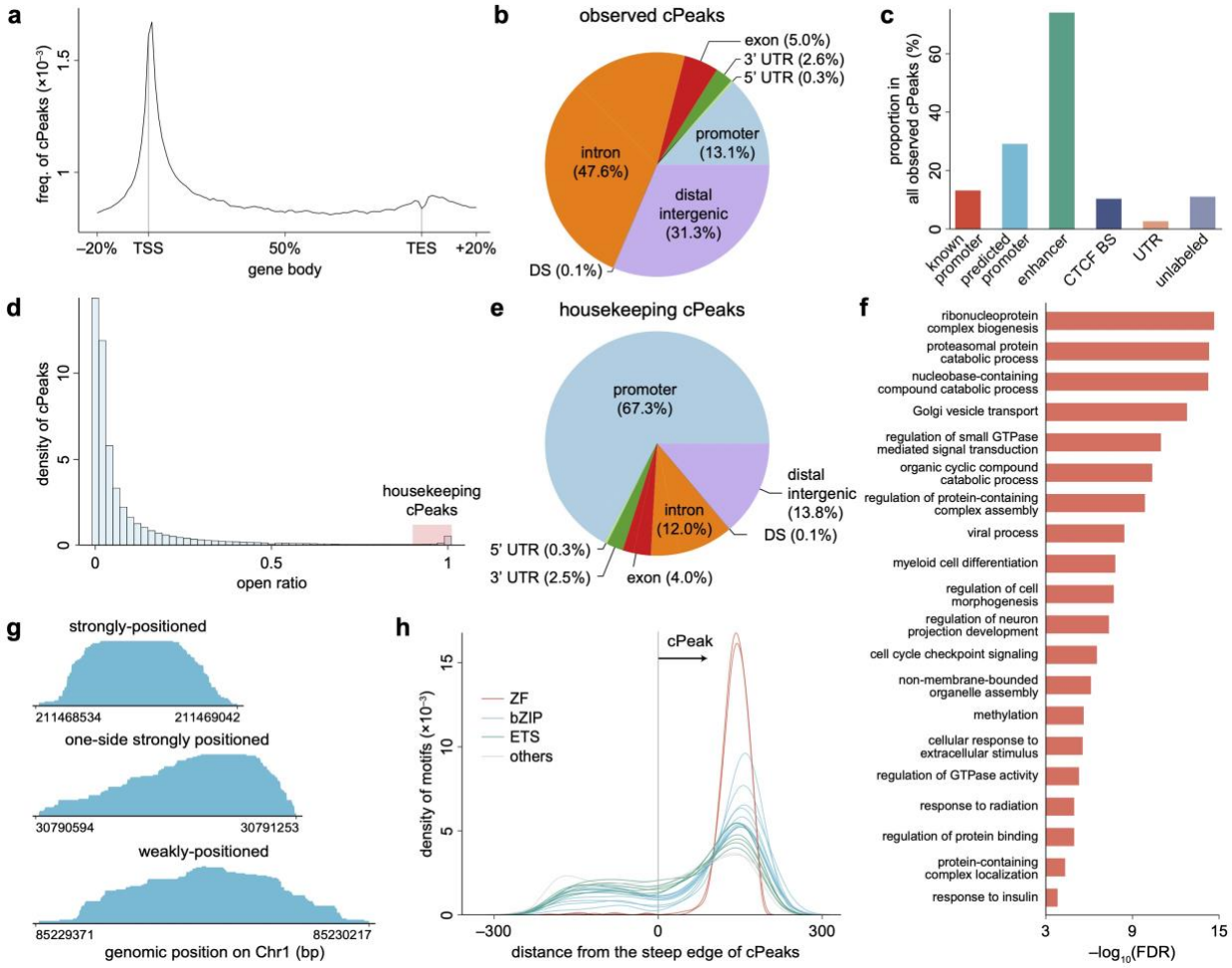

Fig. 2. Properties of observed cPeaks

(a) The enrichment of cPeaks around gene bodies. The x-axis reflects the percentage of total length of gene body regions. The y-axis represents the frequency of observed cPeaks on each site. (b) Pie-chart showing the genomic locations of the observed cPeaks. Genomic annotations of observed cPeaks. DS: downstream (≤ 300 bp downstream of the gene end). (c) The inferred regulatory functions of the observed cPeaks. The x-axis shows the types of inferred functions and the y-axis shows the proportions of observed cPeaks. Note that a cPeak may have multiple inferred functions. CTCF BS: CTCF binding site. (d) The

distribution of open ratios (proportions of tissues open at the cPeak location) of the observed cPeaks. We call the cPeaks with more than 0.9 open ratios (marked by the pink box) as housekeeping cPeaks. (e) Pie-chart showing the genomic locations of the housekeeping cPeaks. We can see that 67.3% of housekeeping cPeaks are in Promoter regions (<=1,000 bp). (f) The top 20 pathways enriched in genes related to housekeeping cPeaks. The 20 pathways with the smallest false discovery rates (FDRs) are shown. (g) Examples showing the three typical shape patterns of cPeaks. (h) Positions of the top 20 enriched motifs around the steep edges of cPeaks. The x-axis is ±300 around the edges of cPeaks, from closed regions to open regions. Each line represents a motif, and each color represents a DNA-binding domain type.

Extend cPeaks through deep learning

We have observed that the locations of cPeaks on the genome were not random (Fig. 1b) and different cell types shared the same set of cPeaks (Fig. 1c). cPeaks also represented various functional regions on the genome, such as promoters, enhancers and CTCF binding sites (Fig. 2c). These findings suggested that cPeaks may be inherent features of the genome. To verify this hypothesis, we investigated whether the positions of cPeaks on the genome can be determined by the corresponding DNA sequences. We adopted the deep convolutional neural network (CNN) model of scBasset[46] to predict an openness score of a candidate region and detect cPeaks on the genome with their DNA sequence patterns (Fig. 3a). We called the regions that were not covered by observed cPeaks as background regions. We trained the model as a binary classifier with observed cPeaks as positive samples and randomly selected background regions as negative samples (Methods). The model achieved high accuracy with an AUC of 94.90% on the test data (Fig. 3b, Methods), demonstrating that the locations of cPeaks were highly correlated with DNA sequences. This result further suggested that cPeaks are a set of inherent features on the genome reflecting all the potential open regions (PORs). The peaks observed in a particular dataset are subsets of PORs that actually open in the studied tissues or cells.

Figure 1b showed that the observed cPeaks tend to saturate but the currently collected data have not reached saturation yet. To get the full set of PORs, we applied the trained CNN model to scan the whole human genome, using a sliding window of 500bp on all the background regions. In this way, we predicted the openness score for each region as its probability as a cPeak (Fig. 3c, Methods). We set the threshold of the openness score as 0.55 to make about 3% of the background regions regarded as predicted cPeaks and obtained ~0.2 million predicted cPeaks in total (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).

We collected another 1,161 bulk ATAC-seq datasets ChIP-Atlas[47] to assess the performance of the predicted cPeaks. They covered 8 cell types or tissues (B cells, T cells, fibroblasts, brain, hippocampus, liver, lung, retina) (Table S3). We calculated the coverages of the observed and predicted cPeaks on the peaks reported in the original publication of these ATAC-seq data. Except for retina, most of the cell types or tissues were also included in the initial datasets we used to generate cPeaks, but none of the new datasets had been included. As a control, 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

the new datasets. We found that observed cPeaks already covered more than 95% of the reported peaks for most new datasets. For retina, the coverage of observed cPeaks was 90.7%, slightly lower than the other tissues. 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). The retina datasets showed the largest improvement in coverage (from 91.3% to 92.3%) after adding predicted cPeaks. These results indicated that predicted cPeaks can elevate the performance of cPeaks to cover open regions in different datasets, especially for cell types or tissues that were not included during cPeak generation.

We combined the predicted cPeaks with the observed cPeaks to for a full set of cPeaks to represent the full set of PORs in the human genome for all cell types. We used these cPeaks as the reference for ATAC-seq data to do read alignment. We provided bioinformatics tools for mapping or converting diverse ATAC-seq data types to cPeaks (see Code Availability for details). We defined an Epigenetics Transfer Format (ETF) to comprehensively describe the information of each cPeak, including its ID, genomic location, length, housekeeping status, shape pattern, and inferred regulatory functions (Methods). The ETF file for all cPeaks was provided in Table S6.

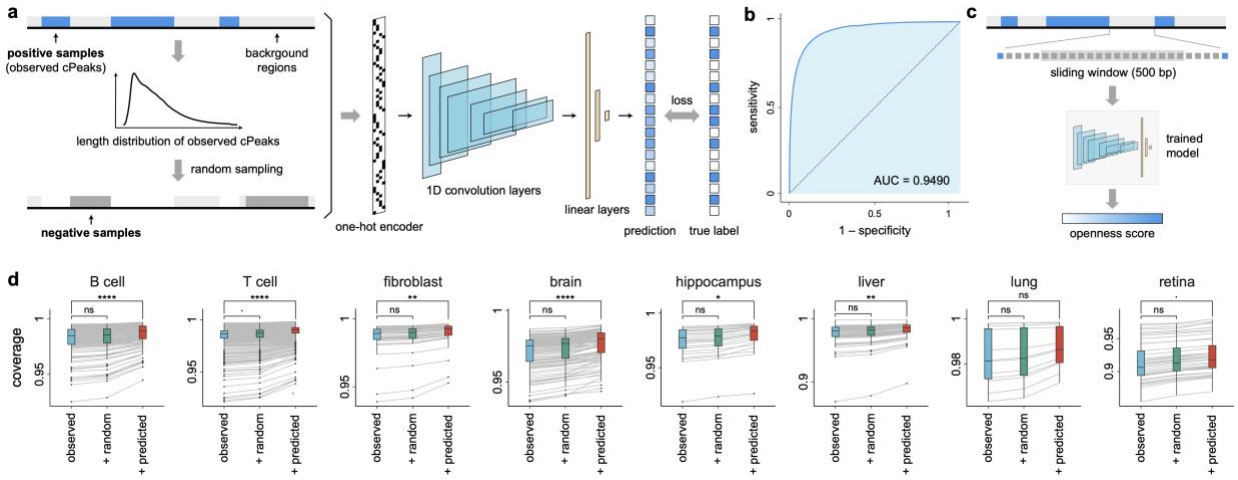

Fig. 3. Computational prediction of cPeaks with DNA sequences using a deep CNN model

(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 extra negative samples from background regions. (b) The ROC curve of the deep CNN model on test data. (c) 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 a predicted openness score. (d) Coverages of the peaks reported in the original papers of the extra ATAC-seq data by three sets of cPeaks: the observed cPeaks based on the initial data, the observed cPeaks plus the predicted cPeaks, and the observed cPeaks plus a random set of regions from the genomic background. Each grey line links the same dataset analyzed by different cPeak sets. The significant levels are 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 using cPeaks as a reference on scATAC-seq data annotation with three commonly used approaches in defining reference features (Methods). The first approach divided the genome into 2,000 bp bins as the basic units of the chromatin accessibility features[33,34,48]. Here we name them as "genomic bins" for convenience. The second approach combined all sequenced single cells together as pseudo-bulk samples and called peaks from them as the reference[17,32]. We name the features defined in this way as "pseudo-bulk peaks". The third approach first cluster cells based on features such as pseudo-bulk peaks or labels derived from cell-sorting, then call peaks by combining cells in each cluster, and then unite the peaks from all clusters together as the reference features[3,29]. We name the features calculated in this way as "united peaks". Each approach provided a set of genomic regions as reference features for mapping sequencing reads to generate a cell-by-feature matrix. We compared the three approaches for defining the features and the cPeak-defined features for their performance in downstream cell annotation tasks. For each approach, we experimented with different selection levels of the features: all features, the top 50,000 highly variable features (HVFs), or the top 5,000 HVFs (Methods, Fig. 4A).

We used an scATAC-seq dataset of hematopoietic differentiation[9] for the comparison. The dataset contains 2,034 cells with FACS-sorting labels as the ground truth of cell types. We generated the cell-by-feature matrices with each feature-defining approach, used multiple commonly-used embedding methods for dimensionality reduction[7,21,32,33,42,49], and conducted the Louvain method[50] to cluster and annotate the cell types (Fig. 4a). We evaluated the annotation results by comparing them with the known ground truth with the adjusted rand index (ARI), adjusted mutual information (AMI), and homogeneity (H)[51] (Fig. 4a, Methods). Results showed that using cPeaks as the features produced the best or one of the best performances in comparison with other feature-defining approaches for all the annotation and evaluation methods (Fig. 4b and S4). The advantage of the performance achieved with cPeaks over other feature-defining approaches is more obvious when we only used the top 5,000 or top 50,000 features. Considering that the definition of cPeaks was independent with this particular dataset, this performance confirmed that cPeaks captured the inherent PORs in the genome and can be used as a general reference for all future bulk or single-cell ATAC-seq data.

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

(a) The analysis workflow of cell annotations. We experimented with 4 feature-defining approaches to obtain references for generating cell-by-feature matrices, 3 ways of feature selection, and 8 strategies to embed the features before clustering analysis, and used three metrics to evaluate the results. (b) Comparison of the performances of the 4 feature-defining approaches at different levels of feature selection, measured by the 3 scores: AMI, ARI and H.

cPeaks enable the discovery of rare cell types

One expected advantage of using a general reference rather than a dataset-dependent reference is the discovery of rare cell types, as cells of rare types may not be sufficient in the dataset to call out the marker peaks that are specific to the rare cell types. We checked the potential of cPeaks with this respect by using them to annotate an scATAC-seq dataset of peripheral blood mononuclear cells (PBMCs) generated by 10X Genomics. It includes 8,728 cells with 87,561 pre-identified pseudo-bulk peaks. We first used cPeaks as the reference to get the cell-by-feature matrix, and applied normalization, feature selection and dimensionality reduction following the common practice (Methods). We then performed graph-based clustering on the reduced dimension and visualized the clustering results using Uniform Manifold Approximation and Projection (UMAP) plots (Fig. S5a). Results showed that the cells were grouped into multiple clusters, with some clusters containing small proportions of cells. We then transferred the cell type labels from an accompanying scRNA-seq PMBC dataset from 10X Genomics by cross-modality integration using the anchor method[52] in Seurat (Fig.5a). For each cell in the scATAC-seq data, we obtained its inferred cell type with a prediction score that indicates the confidence of this cell type prediction. Compared with the results obtained by using the pseudo-bulk peaks to generate the cell-by-feature matrix, cPeak-

generated results obtained improved prediction scores (Fig. 5b and S5b).

Checking the cPeaks-derived clusters after the cell-type annotation, we found three rare cell types in the dataset: plasmacytoid dendritic cells (pDCs), dendritic cells, and natural killer bright (NK bright) cells. Each of them contains only less than 1% of the cells in the dataset (Fig. S5c). We took pDCs as an example to check their differential open regions in comparison with all other cells (Methods). Results showed that using cPeaks identified 1,449 more differential open regions than using the 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 hard 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[53,54].

We also applied cPeaks to help identify rare subtypes of cells within a cell group. We collected the GI epithelial cells in CATLAS[3]. They contain 9 different subtypes including 4 rare subtypes each with less than 2% of the cells (Fig. S5e). Using the pseudo-bulk peaks in the original study cannot separate all subtypes (Fig. S5f), and the original study had to use an additional peak calling step for this group of cells to identify the rare subtypes (Fig. S5g). Using cPeaks to replace the pseudo-bulks peaks as the reference, we were able to distinguish all subtypes using a single step (Fig. 5h, Methods). All results demonstrated that cPeaks can simplify the process of scATAC-seq data processing and facilitate the discovery of rare cell types and their specific features which may otherwise be difficult to be detected using traditional peak-calling approaches.

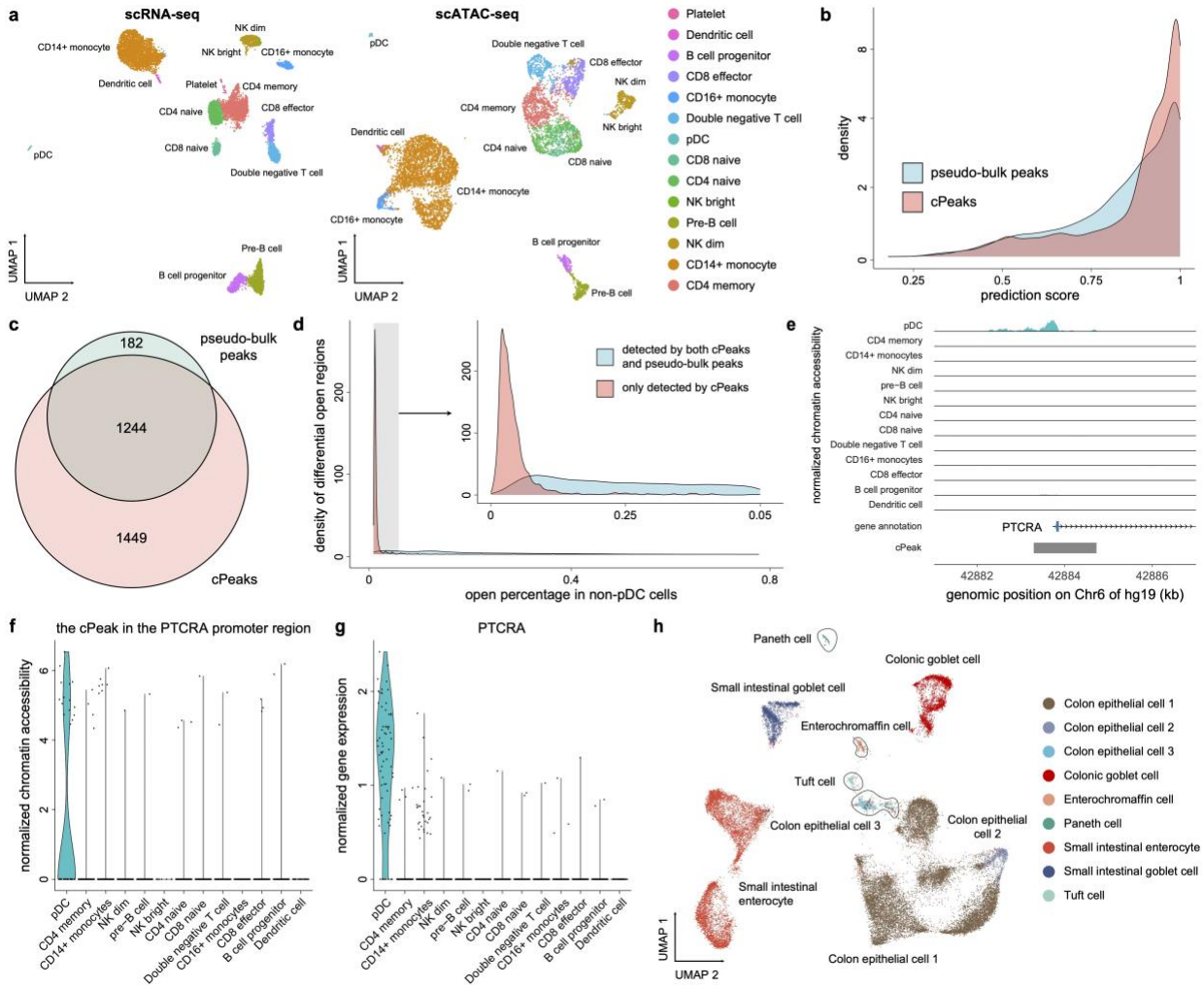

Fig. 5. Rare cell type discoveries using cPeaks as the reference of features

(a) The UMAP plots of scRNA-seq (left panel) and scATAC-seq (right panel) PBMC data, with colors labeling the cell types of the cells. The scRNA-seq cells were assigned to their cell labels according to gene expression, and the labels of scATAC-seq cells were inferred using cross-modality integration. We removed cells in cluster 18 (shown in Fig. S5a) of the scATAC-seq data, as cells in the cluster showed low prediction scores on average (see Methods for details). pDC: Plasmacytoid dendritic cells; NK dim: Nature Killer dim cells; NK bright: Nature Killer bright cells. (b) Histogram of the prediction scores of the cross-modality integration between the scRNA-seq data and scATAC-seq data analyzed with cPeaks (red) or pseudo-bulk peaks (blue). The x-axis represents the prediction score, which reflects the confidence of predicted labels. The y-axis represents the density of cells. (c) The Venn plot of the number of differential open regions calculated by cPeaks (red) or pseudo-bulk peaks (blue) for pDCs. (d) The open percentages of two categories of open regions (as shown in Fig.5c) in the non-pDC cells, showing the specificity of the open regions in pDCs. The red color represents the 1,449 differential open regions only detected with cPeaks. The blue color represents the 1,244 differential open regions detected with both cPeaks and pseudo-bulk peaks. The small panel on the top right is a zoom-in of the density plot of regions with less than 0.05 open percentage in non-pDCs. The open regions that can only be detected using cPeaks as the reference are specific to the rare cell type of pDC and therefore are missed by the analysis based on pseudo-bulk peaks. (e) The chromatin openness in all cell types around the promoter regions of PTCRA, measured by the sum of scATAC-seq reads of all cells of each type. Only pDC cells have a strong openness signal in this region, verifying pDC-specific

accessibility of the PTCRA promoter region detected using cPeaks. (f) The distribution of the normalized reads aligned to the cPeak in the PTCRA promoter region across different cell types. The reads are normalized by the frequency-inverse document frequency (TF-IDF) method. Each dot represents a cell. (g) The normalized expression of PTCRA in different cell types in scRNA-seq data. The gene expression is normalized by the total expression in each cell and log-transformed. (h) The UMAP plot of subtypes in GI epithelial cells, clustered using cPeaks as the reference. The colors show the cell types provided by the original paper. The black circles mark the 4 rare cell subtypes: paneth cell, enterochromaffin cell, tuft cell and colon epithelial cell 3.

Analyze dynamic biological processes and diseases with cPeaks

cPeaks were defined using the data of ATAC-seq datasets of adult healthy samples. We investigated their usability and performance in analyzing developmental or disease data. We first collected an scATAC-seq data of human hematopoiesis[32] to study the use of cPeaks in analyzing cell differentiations. We used cPeaks as the reference to construct the differentiation trajectories of human hematopoiesis lineage from these data. The cells displayed continuous state changes in the differentiation process[32] (Fig. 6a). We aligned the reads of each cell in the scATAC-seq data to the cPeaks to get the cell-by-feature matrix, and labeled each cell based on its annotation in the original study (Fig. 6b, Methods). We estimated pseudotime from the cell-by-feature matrix with Monocle 3[55,56] (Fig. 6c, Methods). The estimated pseudotime was consistent with the original study.

We then collected the human fetal retinal scATAC-seq dataset covering cells in early and late stages of retina organogenesis[57]. We used cPeaks as the reference to generate the cell-by-feature matrices. We plotted cells with UMAP and labeled them according to the original paper (Methods). For cells in the early stages (8 weeks) of development (Fig. 6d), we observed three trajectories from multipotent progenitor cells to horizontal cells, cones, and retinal ganglion cells (Fig. 6e). We can see that the pseudo-time values of these cells also showed an order in the UMAP (Fig. 6f). For cells in the late stages (13 weeks), we separately analyzed samples collected from the center and the periphery of the retina, and the results showed clear clusters of cell types and their transition states (Fig. S6a-b). It was worth noting that datasets 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, and suggested that the set of PORs constructed from data of adult samples also cover most of the potential open regions in early developmental stages.

We then investigated the use of cPeaks in analyzing disease data. Our first experiment on disease data was on an scATAC-seq dataset of both healthy kidneys and kidneys with DKD[17]. DKD or diabetic kidney disease is a chronic inflammatory disease which can cause cellular injury and alter cell types within kidneys. Five of the patient samples in the dataset contain more than 2,000 sequenced cells each. We used these data of 5 samples and of one healthy sample to study whether the correct conclusion can be derived using cPeaks as the reference. Fig. 6g showed the compositions of cell types in these samples reported in the original study. We visualized the cells in the healthy samples (Fig. 6h) and patient samples (Fig. 6i, Fig. S6c-f) using the cPeaks-derived cell-by-feature matrices (Methods). We can see that the cell types reported in the original paper are well preserved using cPeaks as the reference. For example, we successfully re-discovered the increased infiltrating leukocytes cells such as B cells, T

cells, and mononuclear cells in patient 2 (Fig. 6i), and also verified the increased population of VCAM1+ proximal tubule cells in patient 3 (Fig. S6d), which is a marker of kidney injury[58].

The second experiment on diseases was about the clustering of tumor cells. We collected a human gynecologic malignancy scATAC-seq dataset of 74,621 cells in the endometrium or ovary of 11 treatment-naive cancer patients[59] and used cPeaks as the reference to re-analyze the data (Methods). Fig. 6j and S4g showed that cells analyzed with cPeaks could be well clustered by cell types. Cells of the three types of cancers, gastrointestinal stromal tumor (GIST), endometrial cancer, and ovarian cancer, were separated from normal cells. We also labeled the cells in the UMAPs with the copy number variations (CNVs) (Fig. 6k) and expression of KIT, WFDC2, and MUC16 (Fig. S6h) reported by the original study. Besides reproducing the observations in 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). These new observations suggested that the tumor of sample 8 might contain two subclones with different genomic properties. The results demonstrated that cPeaks can help to identify subclusters within cancer cells, and the use of this general reference may help to identify new clues in a small portion of cells that could be missed in studies using data-derived references.

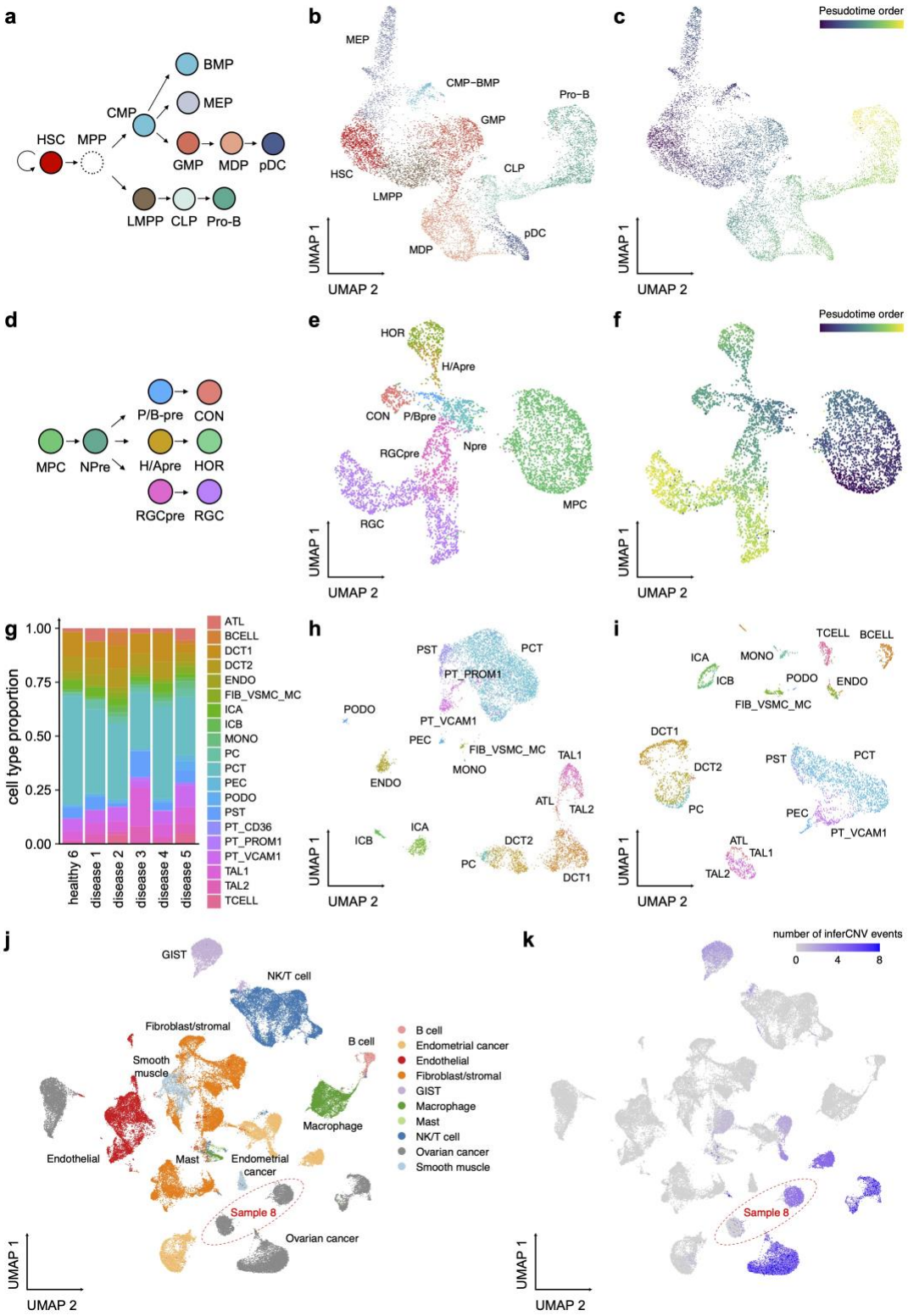

Fig. 6. Analyzing developmental and disease data using cPeaks

(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 is not detected in cells, we use 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. Combining the pseudotime and the labels of cells in Fig. 6e, the development trajectory of retina cells is consistent with the knowledge in Fig. 6d. (g) The cell type proportions of each sample in the kidney dataset. The x-axis showed the sample name and y-axis showed the cell type proportion. Each color represented one cell type: 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. We aggregated all cells in the endometrium and ovary from 11 patients. Colors represent different cell types. Among them, GIST, endometrial cancer, and ovarian cancer are cancer cells. Ovarian cancer cells in sample 8 are 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 are circled by red dashed line. The two sub-clusters in these cells have different CNVs. All the results are analyzed with cPeaks as the reference.

Discussion

A high-quality reference of chromatin accessibility is crucial for the processing and analysis of ATAC-seq data, especially scATAC-seq data. During our effort for building such a general reference, we found that different cell types may share the same set of the potential chromatin open regions in the genome. We called such regions as potential open regions or PORs. They harbor the major regulatory elements in the genome, and may express different openness patterns in different cell types by only opening a subset of the PORs. We observed this phenomenon in 624 ATAC-seq datasets, and confirmed the observation with experiments on additional 1,161 ATAC-seq datasets and 231 scATAC-seq datasets. By merging the peaks from multiple datasets and normalized processing of the shapes of the merged peaks, we defined the concept of consensus peaks or cPeaks, and constructed a set of 1.4M cPeaks based on the 624 ATAC-seq datasets. We built a deep convolution neural network (CNN) model to predict cPeaks with DNA sequences and achieved high accuracy in the prediction. With this model, we further predicted 0.2M cPeaks that have not been seen in the training data. With the observed and predicted cPeaks, we composed the first version of a full set of PORs in the human genome.

We conducted a series of experiments to validate the usefulness and benefits of using cPeaks as the general reference for ATAC-seq and scATAC-seq data analyses. The experiments showed that cPeaks can achieve performances equally well with or better than the best existing approaches for building peak-calling references in major downstream tasks such as cell type annotation and trajectory analysis. In addition, cPeaks can better identify rare cell types or subtypes from scATAC-seq data. The set of cPeaks derived from healthy adult samples can also work well on data of early developmental stages and data of major inflammatory diseases and cancer. The success of cPeaks as a general data-independent reference in all these tasks further supported the hypothesis that PORs are inherent features in the human genome. We therefore propose that all future ATAC-seq or scATAC-seq data can be analyzed by first mapping the sequencing reads to the PORs represented by the cPeaks. In this way, we can have the full set of cPeaks as the equivalence of the transcriptome in scRNA-seq processing, which will make comparison and integration of data from different studies uniform and standardized. The definition and standardization of PORs will also open a new route for the analyses of genomic and epigenomic regulation networks.

We are aware that the concept of PORs as a set of inherent features or elements on the genome is still hypothetical and the set of cPeaks we built in this work is only an initial version based on the current data and computational model. There are many extra works that can be done to further verify the properties and functions of PORs and refine the definition of cPeaks. To facilitate the community to study PORs and use cPeaks in their work, we defined an ETF file format to store all information of cPeaks and also developed a set of basic tools for using cPeaks in various applications.

Methods

Generation of cPeaks

Bulk ATAC-seq data for cPeak generation

We collected all bulk ATAC-seq data from ENCODE[40] and ATACdb[41]. We selected all adult, healthy datasets with more than 10,000 pre-identified peaks. For each dataset, if the peak number of chromosome Y (ChrY) was less than 20, we removed all peaks on ChrY, and the remaining peaks were used for further analysis. We analyzed all the data with the hg38 genome reference. For ATACdb datasets which were analyzed with the hg19 reference in the original study, we used CrossMap[60] (version 0.6.5) to convert positions of all peaks to the hg38 reference. We removed all peaks that altered lengths during conversion.

Defining merged peaks and their shapes

We defined the merged peaks 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 merged peak.

When defining merged peak shapes on ChrY, we only considered the number of datasets containing ChrY after quality control as the denominator.

Calculate shape similarity

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

Generation of cPeaks from merged peaks

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

We split multimodal merged peaks into multiple unimodal ones based on their shapes. For the *i*-th base of a merged peak, we constructed an array x_i the length of which is the same as the number of datasets 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 sampled 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 merged peak length multiplying by 50. We used the cluster ID calculated by HDBSCAN to split the merged peak 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 merged peak.

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[40] and predicted cPeaks that overlapped with any CTCF ChIP-seq signals as CTCF binding sites. We collected histone modification data in ENCODE[40] 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[45] 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 obtained the sequence of ±200 bp around each endpoint with a length of 401 bp. If the sequence was generated from the endpoint of a decreased edge, 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:

- **ID**: The ID of this cPeak. We employed a 14-character string as the unique ID for each cPeak. cPeak ID begins with POR for potential open regions, then HS for human, and ends with nine-digital number. For example, the first cPeak was encoded as PORHS000000001.
- **seqname**: The chromosome of this cPeak.
- **source**: The source of this cPeak: "observed" or "predicted".
- **feature**: "cPeaks"
- **start** & **end**: start and end positions of this cPeak.
- **score**, **strand, frame**: NA
- **attribute**: the additional information about this cPeak. 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[46], 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[51]. 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, united 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 united peaks, we used the peak set provided by the benchmark study[51]

from https://github.com/pinellolab/scATAC-benchmarking/blob/master/Real_Data/Buenrostro_2018/input/combined.sorted.merged.bed. For cPeaks, we removed the housekeeping cPeaks and extra-long cPeaks (> 2000 bp). 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[49]: 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[65]: 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[21,33,66]: 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 Louvain clustering[50] 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[65] 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.

After label transformation, we found cells in cluster 18 had low prediction scores with an average score lower than 0.5. It represented that cells in this small cluster may have low-quality. We removed this cluster from downstream differential accessibility analysis.

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[55,56] 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 fragment files and metadata of the retinal fetal data[57] 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 fragment files of the kidney dataset[17] 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 fragment files and metadata of the gynecologic malignancy dataset[59]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 fetal retina 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 (grants 2021YFF1200901 & 2021YFF1200902) and Tsinghua-Fuzhou Institute for Data Technology (TFIDT2021005).

Author contributions

X.Z., Q.M. and L.W. conceived the study. X.Z. and L.W. supervised the study. Q.M. and X.Wu. collected the data, designed the experiments and performed the analysis. C.L. tested the mapping and conversion tools. J.L. (first) and X. Wang assisted in designing the deep CNN network. X.X., Si.C., Sh.C., J.L. (second) and R.J. contributed to the interpretation of results. Q.M., X.Wu. L.W. and X.Z. wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

1. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).

2. Preissl, S., Gaulton, K. J. & Ren, B. Characterizing cis-regulatory elements using single-cell epigenomics. *Nat. Rev. Genet.* (2022) doi:10.1038/s41576-022-00509-1.

3. Zhang, K. *et al.* A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985-6001.e19 (2021).

4. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell* **71**, 858-871.e8 (2018).

5. Tsompana, M. & Buck, M. J. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin* **7**, 33 (2014).

6. Duren, Z., Chen, X., Jiang, R., Wang, Y. & Wong, W. H. Modeling gene regulation from paired expression and chromatin accessibility data. *Proc. Natl. Acad. Sci.* **114**, E4914–E4923 (2017).

7. Bravo González-Blas, C. *et al.* cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019).

8. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

9. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).

10. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).

11. Preissl, S. *et al. Single nucleus analysis of the chromatin landscape in mouse forebrain development*. http://biorxiv.org/lookup/doi/10.1101/159137 (2017) doi:10.1101/159137.

12. Trevino, A. E. *et al.* Chromatin accessibility dynamics in a model of human forebrain development. *Science* **367**, eaay1645 (2020).

13. Morabito, S. *et al.* Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nat. Genet.* **53**, 1143–1155 (2021).

14. Turner, A. W. *et al.* Single-nucleus chromatin accessibility profiling highlights regulatory mechanisms of coronary artery disease risk. *Nat. Genet.* **54**, 804–816 (2022).

15. Sanghi, A. *et al.* Chromatin accessibility associates with protein-RNA correlation in human cancer. *Nat. Commun.* **12**, 5732 (2021).

16. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).

17. Wilson, P. C. *et al.* Multimodal single cell sequencing implicates chromatin accessibility and genetic background in diabetic kidney disease progression. *Nat. Commun.* **13**, 5253 (2022).

18. Minnoye, L. *et al.* Chromatin accessibility profiling methods. *Nat. Rev. Methods Primer* **1**, 10 (2021).

19. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1-21.29.9 (2015).

20. Grandi, F. C., Modi, H., Kampman, L. & Corces, M. R. Chromatin accessibility profiling by ATAC-seq. *Nat. Protoc.* **17**, 1518–1552 (2022).

21. Cusanovich, D. A. *et al.* Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).

22. Zhu, C. *et al.* An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat. Struct. Mol. Biol.* **26**, 1063–1070 (2019).

23. Lareau, C. A. *et al.* Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916–924 (2019).

24. Mimitou, E. P. *et al.* Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.* **39**, 1246–1258 (2021).

25. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).

26. Chen, X., Miragaia, R. J., Natarajan, K. N. & Teichmann, S. A. A rapid and robust method for single cell chromatin accessibility profiling. *Nat. Commun.* **9**, 5345 (2018).

27. Fullard, J. F. *et al.* An atlas of chromatin accessibility in the adult human brain. *Genome Res.* **28**, 1243–1252 (2018).

28. Domcke, S. *et al.* A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7612 (2020).

29. Markenscoff-Papadimitriou, E. *et al.* A Chromatin Accessibility Atlas of the Developing Human Telencephalon. *Cell* **182**, 754-769.e18 (2020).

30. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).

31. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).

32. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).

33. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309-1324.e18 (2018).

34. Wang, A. *et al.* Single-cell multiomic profiling of human lungs reveals cell-type-specific and age-dynamic control of SARS-CoV2 host genes. *eLife* **9**, e62522 (2020).

35. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

36. Nair, V. D. *et al.* Differential analysis of chromatin accessibility and gene expression profiles identifies cis-regulatory elements in rat adipose and muscle. *Genomics* **113**, 3827–3841 (2021).

37. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.* **21**, 22 (2020).

38. Daugherty, A. C. *et al.* Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans. *Genome Res.* **27**, 2096–2107 (2017).

39. Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).

40. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

41. Wang, F. *et al.* ATACdb: a comprehensive human chromatin accessibility database. *Nucleic Acids Res.* **49**, D55–D64 (2021).

42. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).

43. 10x Genomics. 10k_pbmc_ATACv2_nextgem_Chromium_X - Human PBMCs. https://cf.10xgenomics.com/samples/cell-atac/2.1.0/10k_pbmc_ATACv2_nextgem_Chromium_X/10k_pbmc_ATACv2_nextgem_Chromium_X_web_summary.html.

44. 10x Genomics. 10k_pbmc_ATACv2_nextgem_Chromium_Controller - Human PBMCs. https://cf.10xgenomics.com/samples/cell-atac/2.1.0/10k_pbmc_ATACv2_nextgem_Chromium_Controller/10k_pbmc_ATACv2_nextgem_Chromium_Controller_web_summary.html.

45. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).

46. Yuan, H. & Kelley, D. R. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nat. Methods* **19**, 1088–1096 (2022).

47. 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.* **50**,

W175–W182 (2022).

48. Fang, R. *et al.* Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* **12**, 1337 (2021).

49. Baker, S. M., Rogerson, C., Hayes, A., Sharrocks, A. D. & Rattray, M. Classifying cells with Scasat, a single-cell ATAC-seq analysis tool. *Nucleic Acids Res.* **47**, e10–e10 (2019).

50. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).

51. Chen, H. *et al.* Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.* **20**, 241 (2019).

52. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

53. Reizis, B. Regulation of plasmacytoid dendritic cell development. *Curr. Opin. Immunol.* **22**, 206–211 (2010).

54. Ghosh, H. S., Cisse, B., Bunin, A., Lewis, K. L. & Reizis, B. Continuous Expression of the Transcription Factor E2-2 Maintains the Cell Fate of Mature Plasmacytoid Dendritic Cells. *Immunity* **33**, 905–916 (2010).

55. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

56. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

57. Finkbeiner, C. *et al.* Single-cell ATAC-seq of fetal human retina and stem-cell-derived retinal organoids shows changing chromatin landscapes during cell fate acquisition. *Cell Rep.* **38**, 110294 (2022).

58. Muto, Y. *et al.* Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nat. Commun.* **12**, 2190 (2021).

59. Regner, M. J. *et al.* A multi-omic single-cell landscape of human gynecologic malignancies. *Mol. Cell* **81**, 4924-4941.e10 (2021).

60. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).

61. Hartigan, J. A. & Hartigan, P. M. The Dip Test of Unimodality. *Ann. Stat.* **13**, (1985).

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)* 223–228 (2020). doi:10.1109/MFI49285.2020.9235263.

63. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinforma. Oxf. Engl.* **31**, 2382–2383 (2015).

64. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576–589 (2010).

65. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).

66. Cusanovich, D. A. *et al.* The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **555**, 538–542 (2018).

67. 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.* **46**, D267–D275 (2018).

68. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745 (2016).

Supplementary materials

Supplementary Figures

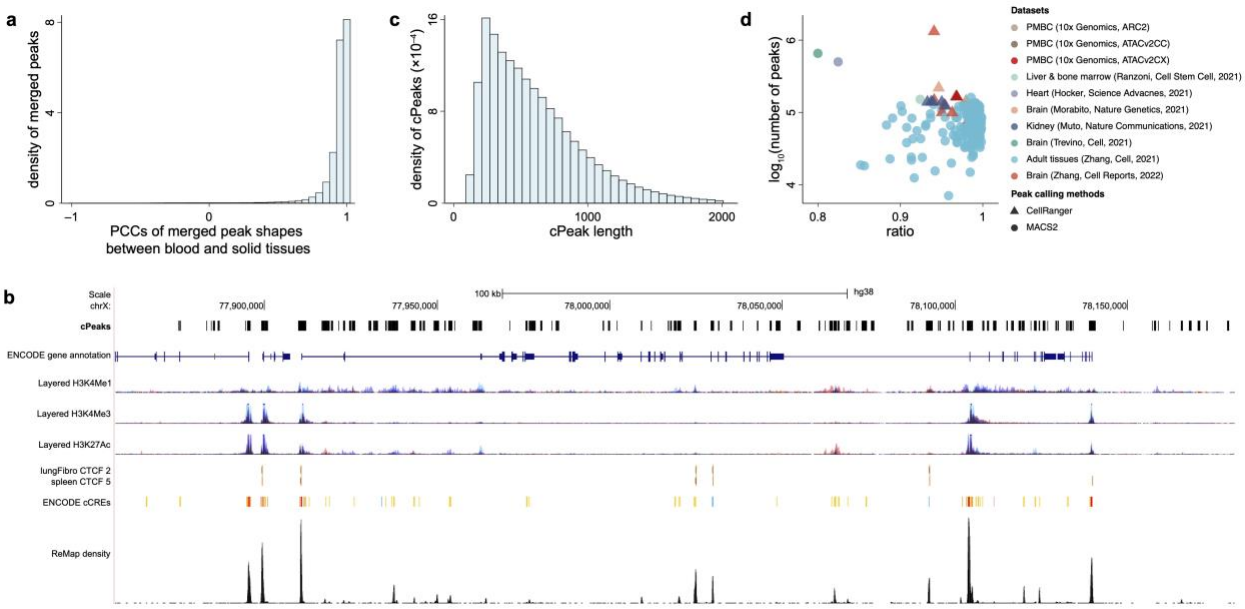

Fig. S1. Details of observed cPeaks

a. PCCs of merged peak 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[40], layered H3K4Me1 marker[40], layered H3K4Me3 marker[40], layered H3K27Ac marker[40], lung fibroblast CTCF signal[40], spleen CTCF signal[40], ENCODE cCREs[40] and ReMap density[67]. 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[67].

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.

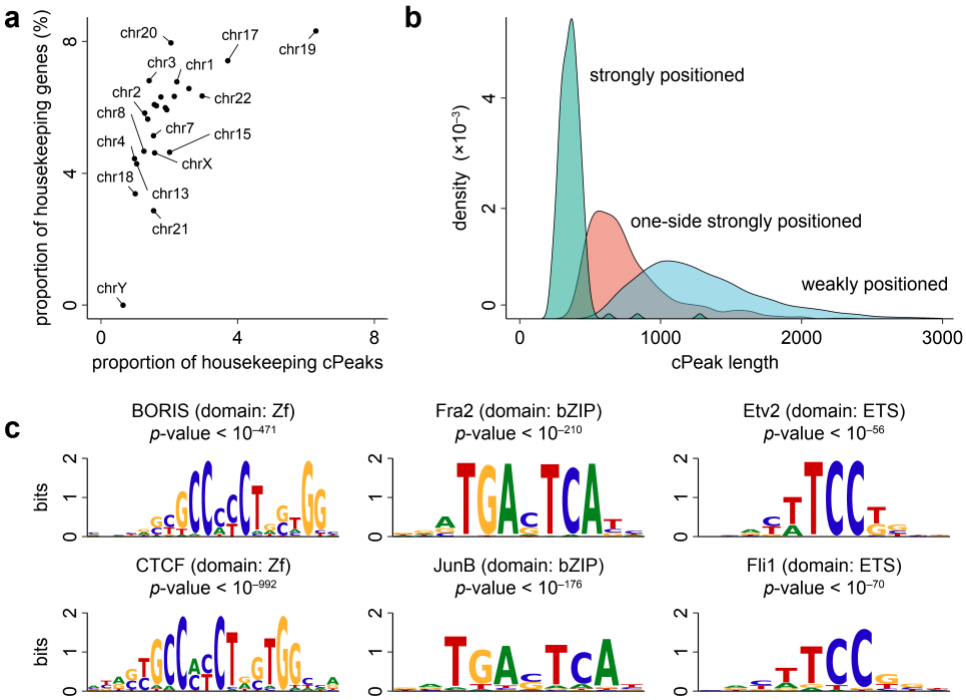

Fig. S2. Properties of observed cPeaks

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.

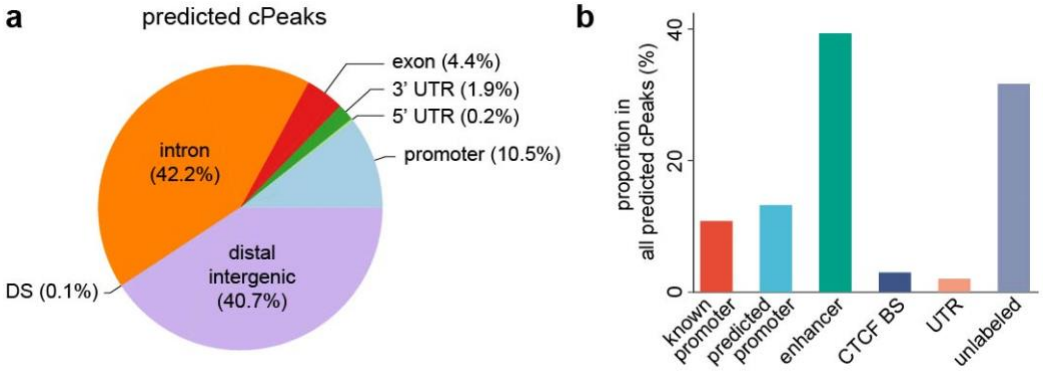

Fig. S3. Properties of predicted cPeaks

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.

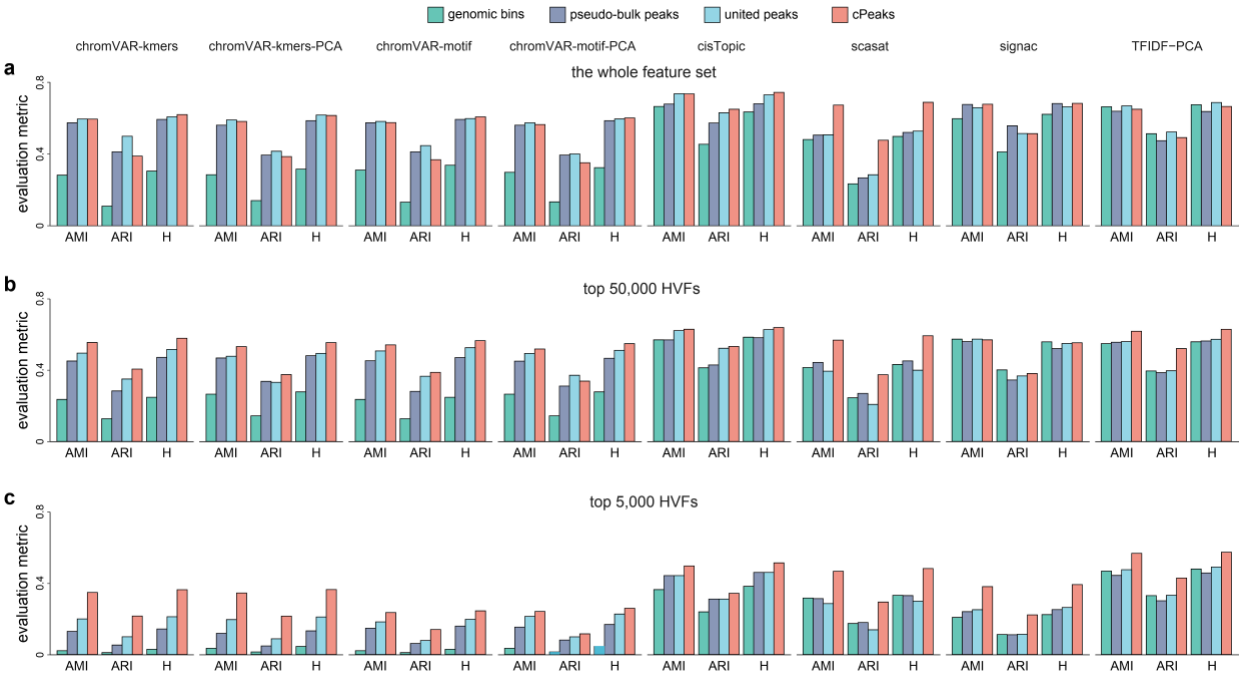

Fig. S4. The performance of cell annotations on FACS2-sorted hematopoietic differentiation data

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.

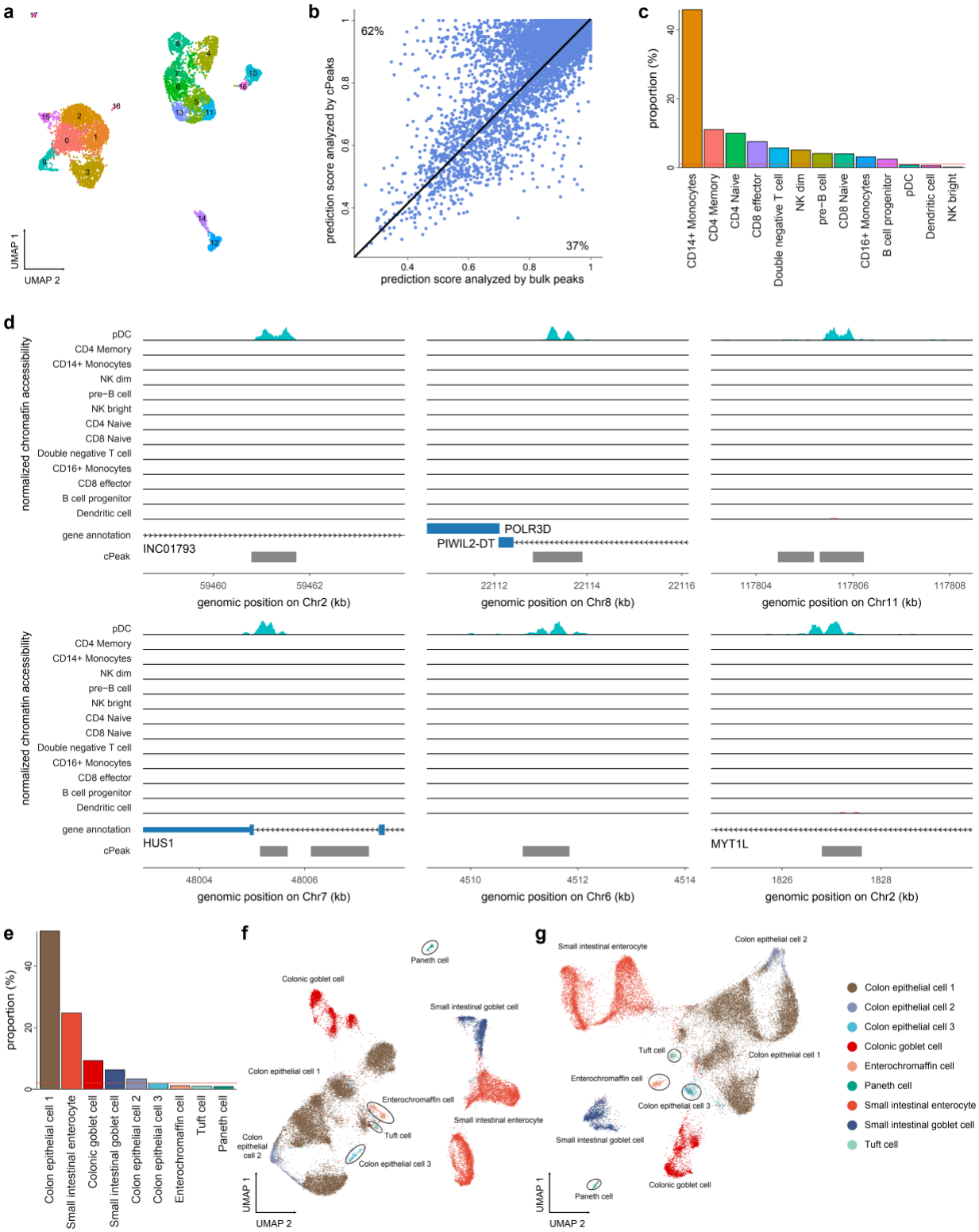

Fig. S5. Rare cell type discovery using cPeaks

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[68].

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.

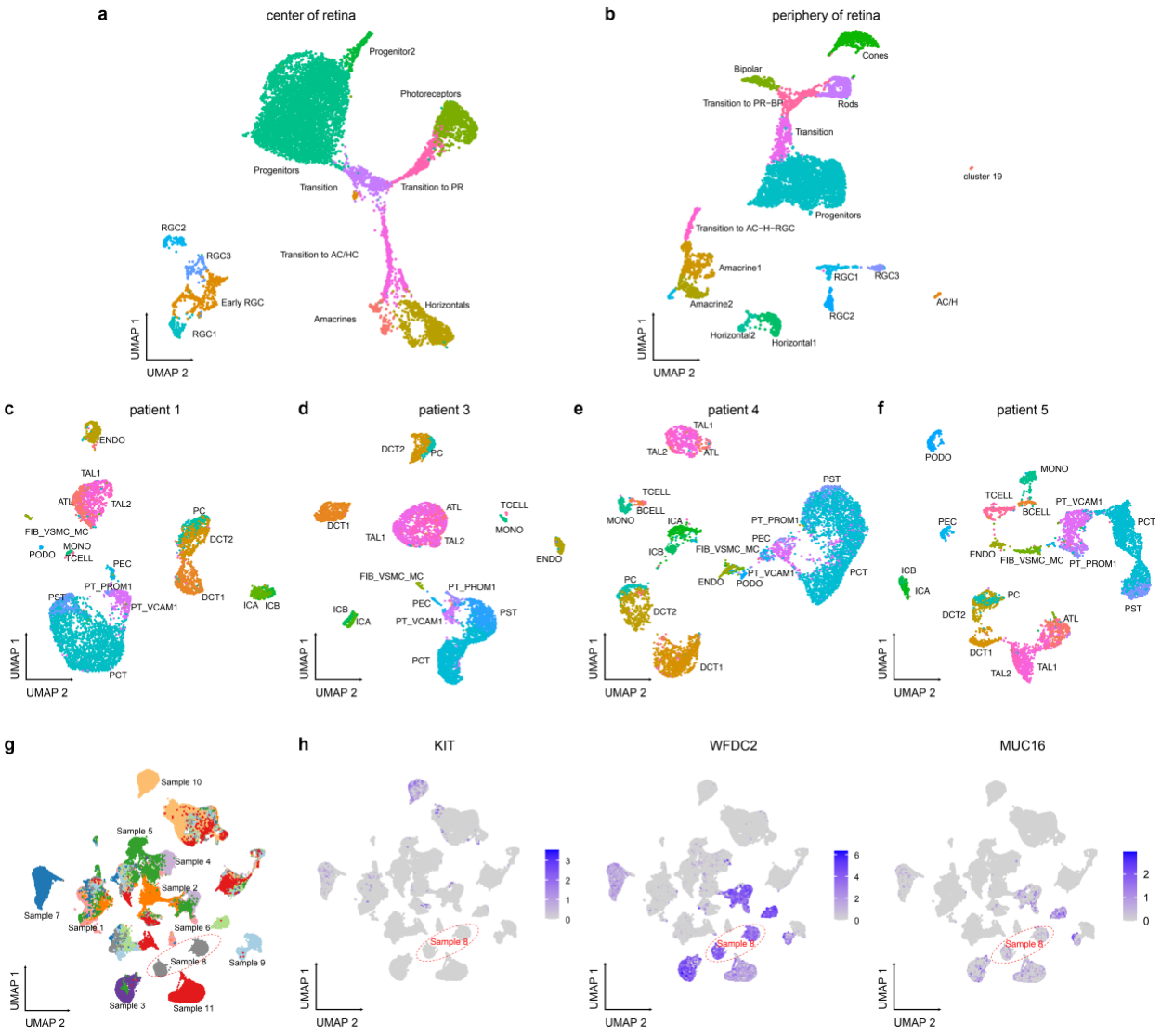

Fig. S6. Analyzing dynamic biological processes and diseases by cPeaks

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.