Toward a generic feature set defined by consensus peaks as a consistent reference for ATAC-seq data

Qiuchen Meng[1,#], Xinze Wu[1,#], Yubo Zhao[1], Wenchang Chen[1], Chen Li[1], Zheng Wei[1], Jiaqi Li[1,4], Xi Xi[1], Sijie Chen[1], Catherine Zhang[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

The rapid advancement of transposase-accessible chromatin using sequencing (ATAC-seq) technology, particularly with the emergence of single-cell ATAC-seq (scATAC-seq), has accelerated the studies of regulatory element identification, demanding higher precision and uniformity in feature definition. Unlike gene expression data, no consistent feature reference is developed for ATAC-seq data, which hinders single-cell level data analysis and cell atlas creation. Based on a systematic analysis of 1,785 ATAC-seq and 231 scATAC-seq datasets, we found that cells share the same feature set represented by 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 generic feature set 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 establishing cell atlas, and analyzing cells in dynamic or disease states.

Introduction

Chromatin accessibility refers to the availability of chromatinized DNA to directly interact with other macromolecules[1]. The open regions[1,2], also known as accessible regions[2,3], imply positions of many transcriptional regulatory elements such as promoters, enhancers, insulators, and transcription factor (TF) binding sites[4,5]. Studying chromatin accessibility in cells can help to infer gene regulations[2,6–9] in many biological processes such as cell differentiation[10,11], organ development[12,13], and disease occurrence[14–18].

The rapid development of high-throughput sequencing technologies[3], 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[5,17,27–29].

Unlike RNA-seq data with reference transcriptomes[30], ATAC-seq data lack a unified reference for localizing and quantifying chromatin accessibility across the genome. To address this, researchers typically create temporary feature sets for individual analysis. Some 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[5,18,27]. In the case of scATAC-seq data, where peak-calling at the single-cell level is not feasible, multiple cells need to be merged into a pseudo-bulk sample during feature definition[5,18,27,28,33,34]. This results in the loss of single-cell resolution during pseudo-bulk generation. For rare cell types with relatively low ratios, which are often of particular research interest, their specific open chromatin regions are challenging to identify after merging into pseudo-bulk samples. Some studies partitioned the genome into fixed regions and quantified chromatin accessibility in each region[35,36]. This approach loses the biological significance of features. Additionally, it leads to the split of continuous open regions and introduces plenty of redundant features. Several pre-defined sets of open regions were proposed, such as consensus DNase I hypersensitive sites (cDHSs)[37] and the *cis*-element ATLAS (CATLAS)[5]. cDHS shows a common index of DNase I hypersensitive sites generated by DNase I hypersensitive seq (DNase-seq) data[37], and cCREs provided by CATLAS erased the information of open region length[5]. Other open region sets were built to localize and define regulatory elements on the genome[38,39]. Thus, these open region sets are usually used for quality control during ATAC-seq data analysis rather than serve as a reference[3].

The absence of a unified ATAC-seq reference has hindered the comparison and integration of data from different samples. Current practices are to call peaks as features in each individual dataset, and then to integrate features from multiple datasets. In scATAC-seq data analysis, to discover more features, researchers often divide cells into multiple groups followed by feature integration. For feature integration, peak merging[40–43] and iterative removal[5,17] are two common approaches to integrating 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. It is unsuitable for large-scale analysis, especially for single-cell data with multiple clusters. 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, especially for scATAC-seq data[44].

We collected 624 ATAC-seq datasets of multiple tissues to study the distribution of chromatin open regions on the genome across different tissues to define a set of consensus peaks or cPeaks, and collected 231 scATAC-seq datasets to validate the use of cPeaks as a 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 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. Compared with other pre-defined open region sets, cPeaks are more suited for use as a reference, especially for ATAC-seq and scATAC-seq data. 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. It can also be used as a universal feature set to build a large-scale human chromatin accessibility atlas. Additionally, we have established comprehensive resources for cPeak information and provided bioinformatics tools for mapping or converting diverse types of ATAC-seq to cPeaks.

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[38] and ATACdb[45], 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 the beginning, but the growth soon tended to saturate (Fig. 1b). This phenomenon was independent of 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 called it as the shape of the peak (Methods).

102 We further explore the consistency of cPeak shape at the same genomic region under different conditions such as
103 cell types, sequencing technologies, and peak-calling methods. First, we examined whether different tissue types
104 impacted the shapes. The datasets used to generate merged peaks contained two major tissue types: ~51% from
105 blood and ~49% from solid tissues (lung, heart, brain, etc.). We used the same procedure to merge peaks for these
106 two tissue types separately. The number of merged peaks shared by the two groups took up 87% and 76% of the
107 identified merged peaks in the blood data and the solid tissue data, respectively. We found the common merged
108 peaks found from the two groups at the same location tended to have the same shape (an example in Fig. 1c). We
109 measured the shape similarity of each common merged peak by the Pearson correlation coefficient (PCC) of its
110 shapes calculated in the blood or solid tissues (Methods). The common merged peaks showed high PCCs between
111 the two tissue types, with a median of 0.9 (Fig. S1a). Then, we explored whether the sequencing methods influence
112 shapes. Comparing the merged peak shapes with cDHS shapes which generated by aggregating corresponding
113 DNase-seq data[37], we found high similarity in shapes within the same regions (Fig. S1b, Methods). Finally, we
114 conducted experiments testing if shapes correlate with peak-calling methods and parameters. We compared
115 aggregated peak shapes generated by CellRanger[32] and MACS2[31] peak-calling methods with different parameters
116 (Methods). The results showed a high consistency of shapes across varying conditions (Fig. S1c-d). These
117 experiments indicated that the shape of a merged peak is an intrinsic property of the genome, irrelevant to the tissue
118 types, sequencing methods, and peak-calling methods.

119 We found that the shapes of merged peaks were of two major types: unimodal or multimodal (Fig. 1a, Methods). We
120 assumed that a multimodal merged peak was composed of adjacent unimodal peaks, and therefore split multimodal
121 merged peaks into multiple unimodal ones (Fig. 1a, Methods). We took the unimodal merged peaks as the basic unit
122 of the chromatin accessibility reference and named them consensus peaks or cPeaks. In this way, we obtained ~1.4
123 million cPeaks in total (Table S2, an example provided in Fig. S1e) based on the 624 collected ATAC-seq datasets.
124 The lengths of most cPeaks are less than 2,000 bp, with a median length of 525 bp. This is similar to the previous
125 observations that the peak lengths in ATAC-seq data were about 500 bp[46] (Fig. S1f).

126 We assessed the performance of using cPeaks as a chromatin accessibility reference to process scATAC-seq data.
127 We collected 231 scATAC-seq datasets from published studies (Table S3), each with a set of peaks pre-identified
128 with different peak-calling methods[31,32]. For each dataset, we calculated the proportion of the pre-identified peaks
129 that overlapped with cPeaks and defined this proportion as the coverage of cPeaks in the dataset. Most datasets
130 showed coverages greater than 0.9, with an average of 0.98 and a median of 0.99 (Fig. 1d and S1g). We selected 27
131 scATAC-seq datasets from different organs in CATLAS[5] to evaluate the enrichment of sequencing reads on cPeaks.
132 We found that the sequencing reads enriched around centers of cPeaks (Fig. 1e, Methods). On average, 68.8% of
133 sequencing reads in the scATAC-seq data were mapped to the observed cPeaks, higher than the typical mapping rate
134 of ~65% as reported in the technical reports of 10X Genomics[47,48]. This indicates that cPeaks are a reliable set of
135 reference peaks for recovering the signals in scATAC-seq data.

136 We compared the performance of cPeaks against other pre-defined open region sets, including cDHS generated by
137 DNase-seq data and CATLAS cCREs generated via building a human cell atlas. First, we assessed the differences in

coverage of cPeaks and cDHS on ATAC-seq open regions. We collected additional three ATAC-seq and three DNase-seq data on GM12878 cell line[49]. We paired data from two sequencing technologies by keeping their pre-identified peak quantities nearly identical (Methods). We found that relative to cDHS, cPeaks increased the coverage of the ATAC-seq data within each pair, particularly with the open regions uniquely detected in ATAC-seq profiling. Also, the coverage of cPeaks to DNase-seq uniquely detected regions was not reduced (Fig 1f, Methods). It indicates that at peak level, cPeaks cover a broader range of open regions identified by ATAC-seq than cDHS. Then, we compared the coverage of aligned reads on open regions among cPeaks, cDHS, and CATLAS cCREs. We collected scATAC-seq data of hematopoietic differentiation[10] and identified peaks through peak calling as the ground truth of open regions. After mapping sequencing reads to the called peak set and three pre-identified feature sets, we sorted features within each set by the number of mapped reads. Figure 1g illustrates the increase of aggregated mapped reads by adding sorted features. It shows that the curves for cPeaks and called peaks are most similar. Although CATLAS cCREs and cDHS demonstrate a similar growth trend, the capture of mapped reads is not the same due to their trimming procedure of features. Compared to cDHS and CATLAS cCREs, cPeaks are more suitable for ATAC-seq data and maintain the distinct lengths of PORs, thereby serving as a more appropriate reference for both ATAC-seq and scATAC-seq data.

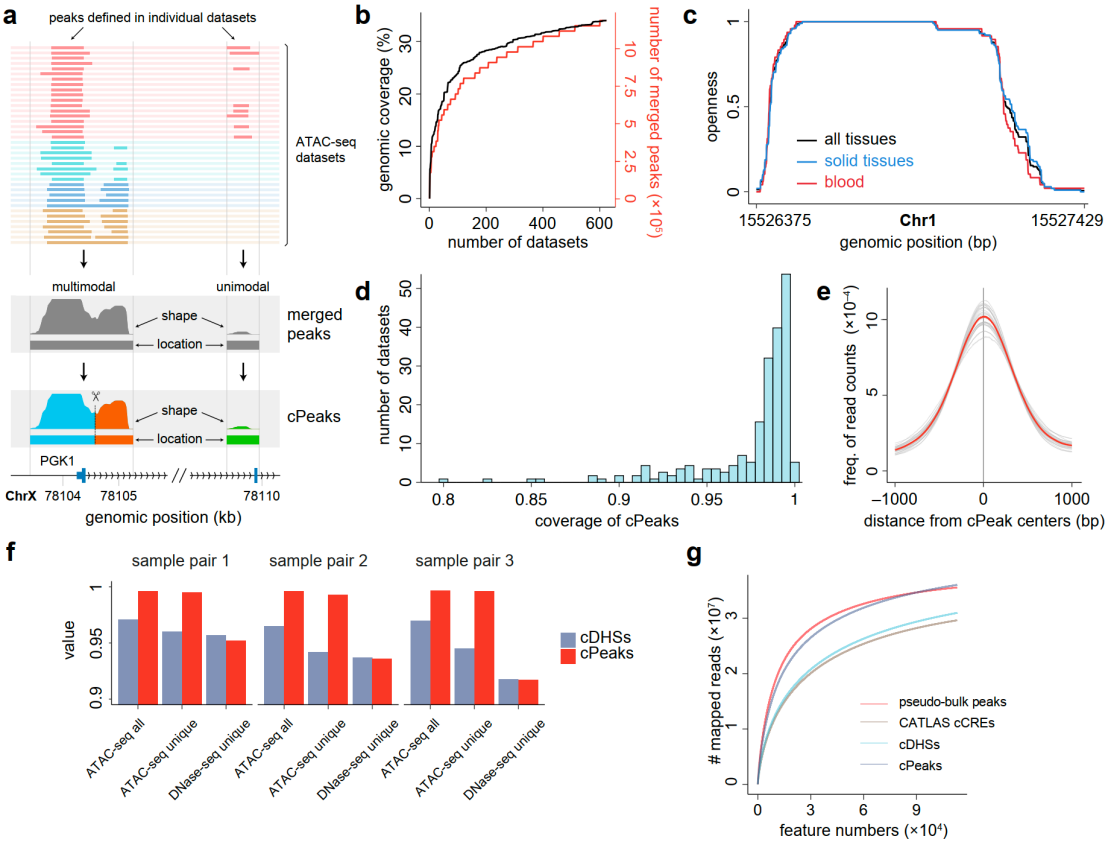


Fig 1. The computation diagram of cPeaks and their basic properties.

(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 represents a pre-identified peak. Different colors

157 indicate different cell types. The middle panel illustrates a multimodal merged peak (left) and a unimodal merged peak (right). In
158 the bottom panel, the multimodal merged peak is clipped as two cPeaks, and the unimodal merged peak forms one peak. The
159 relative positions of the cPeaks with regard to the transcription start site (TSS) and the 2nd exon of PGK1 are shown below the
160 cPeaks. (b) Increase tendency of the genomic coverage (black line) and number of merged peaks (red line) with the addition of
161 more datasets. The x-axis represents the number of aggregated datasets. The left y-axis represents the total genomic coverage of
162 all merged peaks on each step. The right y-axis represents the number of merged peaks on each step. (c) An example of the
163 shapes of merged peaks on chromosome 1, obtained with data of only blood samples (red line), of all solid tissues (blue line), or
164 of all tissues (black line). The shapes obtained with different data are highly correlated along the genome with Pearson
165 correlations of 0.99. (d) Histogram of the coverages of cPeaks in the 231 scATAC-seq datasets. The x-axis represents the
166 coverage of cPeaks, which is the number of overlapped pre-defined peaks with cPeaks divided by the total number of pre-defined
167 peaks in the original data. The y-axis represents the number of datasets. Most datasets get more than 90% coverage and the
168 median coverage is 0.99. (e) Enrichment of sequencing reads around centers of cPeaks. All sequencing reads overlapped with
169 ±1000 bp genomic regions around cPeak centers are aggregated to form the distributions. Each grey line represents a dataset. The
170 red line shows the average distribution of all datasets. (f) The coverage of cPeaks and cDHSs on various open regions. Different
171 blocks represent different pairs of ATAC-seq and DNase-seq data. Where 'ATAC-seq all' represents all open regions detected in
172 the ATAC-seq data, while 'ATAC-seq/DNase-seq unique' represents the unique open regions detected by each sequencing
173 technology. (g) The increase of aggregated mapped reads by adding sorted features. The x-axis represents the number of features
174 added, where the order of these features is sorted from high to low according to the number of mapped reads.

175

176 cPeaks reflect different functional regions on the genome

177 We investigated cPeaks' associations with known regulatory elements or biological functions. By comparing cPeaks
178 with known genomic annotations, we found that cPeaks are enriched around transcriptional start sites (TSSs) and
179 transcription end sites (TESs) (Fig. 2a, Methods). About 13% of cPeaks are localized at promoter regions (±2000 bp
180 around TSS), while the remaining cPeaks were primarily localized at exons or distal intergenic regions (Fig. 2b,
181 Methods). We collected multiple types of epigenetic annotations, including histone modification and CTCF ChIP-
182 seq data (Table S4). Using the genomic and epigenetic annotations, we found that most cPeaks contain or overlap
183 with one or more regulatory elements, such as enhancers, promoters (including known and predicted promoters),
184 untranslated regions (UTRs), or CTCF binding sites (CTCF BSs) (Methods, Fig. 2c).

185 For each cPeak, we defined the maximum value of its shape as its "open ratio" to reflect the ratio of datasets
186 opening in this cPeak (Methods). We defined a cPeak with an open ratio > 90% as a "housekeeping" cPeak (Fig.
187 2d). A total of 24,538 of the observed cPeaks (1.8%) were identified as housekeeping cPeaks. The median length of
188 housekeeping cPeaks was 1441 bp, with most located at promoter regions (Fig. 2e). These promoter regions covered
189 77.4% of the promoters of housekeeping genes[50] (Fisher's exact test p-value $< 1\times10^{-5}$, Methods). The proportion of
190 housekeeping cPeaks and the proportion of housekeeping genes on different chromosomes were highly correlated
191 (Fig. S2a). We performed Gene Ontology (GO) enrichment analysis on genes related to housekeeping cPeaks
192 (Methods) and found that they were enriched in the pathways related to essential biological processes such as
193 ribonucleoprotein complex biogenesis, cell cycle checkpoint signaling, and methylation (Fig. 2f). A recently
194 published article[51] also proposed a similar concept: Housekeeping CREs (HK-CREs). It reached similar conclusions

195 that most HK-CREs were located at core promoters and HK-CREs were associated with biological processes of
196 housekeeping genes. Figure 2g shows that housekeeping cPeaks cover 98.4% of HK-CREs.

197

198 Characterizing genomic regions and transcription factors through cPeak shapes

199 We defined the genomic region corresponding to the maximum value in each cPeak shape as its core open region.
200 The distance from the endpoints of cPeak to the nearest boundary of the core open region is defined as the edge
201 length. Edge length reflects the consistency of the open region boundary across different cells. cPeaks with smaller
202 edge lengths signified more consistent boundaries of open regions in different experiments and suggested the
203 presence of well-positioned flanking nucleosomes[1,52,53]. Conversely, a larger edge length suggested that the flanking
204 nucleosomes around the cPeak boundaries were weakly positioned. We filtered cPeaks with open ratio of more than
205 0.3 to ensure the reliability of shapes. Figure 2h displays the distribution of edge lengths of filtered cPeaks. We
206 observed three peaks with demarcations at 170 bp and 340 bp, which approximately correspond to the length of
207 DNA warps around one and two nucleosomes. We categorized edges less than 170 bp as well-positioned cPeak
208 edges, while others as weakly-positioned edges. Then, cPeaks were classified into three patterns according to the
209 type of two edges. We named them "well-positioned", "one-side well-positioned" and "weakly positioned" cPeaks
210 (Fig. 2i and S2b).

211 We focused on the genomic regions corresponding to well-positioned cPeak edges, which represent more precise
212 regulation. We hypothesized that the occurrence of well-positioned cPeaks may be caused by two biological
213 mechanisms. One is the presence of specific TF binding sites at the inner boundary of cPeaks. The other is the
214 presence of well-positioned flanking nucleosomes outside the boundary of cPeaks. For the first hypothesis, we
215 conducted motif enrichment analysis around well-positioned edges of cPeaks (Methods) and identified some known
216 TFs, such as CTCF, BORIS, Fosl2, Fra2, Fli1, and ETS1 (Fig. S2c). We found that the top 20 enriched TFs shared
217 common DNA binding domains, which include the basic leucine zipper (bZIP) domain, the Zinc finger (Zf) domain,
218 and the erythroblast transformation specific (ETS) domain (Fig. 2j). Other enriched TFs also demonstrated
219 enrichment in domains like the homeobox and basic helix-loop-helix (bHLH) domains (Fig. S2d, Methods). These
220 enriched TFs also showed GO enrichment in essential biological processes, such as the regulation of transcription,
221 cell development, and cell differentiation (Fig. S2e, Methods). We assumed that TFs with such domains tend to have
222 a stronger nucleosome remodeling capability. We found pioneer transcription factors (PTFs)[54–56] were enriched in
223 the enriched TFs with a p-value of 0.016 using the Fisher exact test (Methods), which may facilitate the
224 identification and study of PTFs. For the second hypothesis, we performed *de novo* motif enrichment analysis for the
225 well-positioned edge regions, identifying sequence motifs enriched around the boundaries of cPeaks (Methods). We
226 found several motifs may correlated with well-positioned flanking nucleosome positioning (Fig. 2k & S2f). In
227 summary, the shapes of cPeaks provide new perspectives and insights for genome annotations, categorization of
228 transcription factors, and nucleosome positioning studies.

229

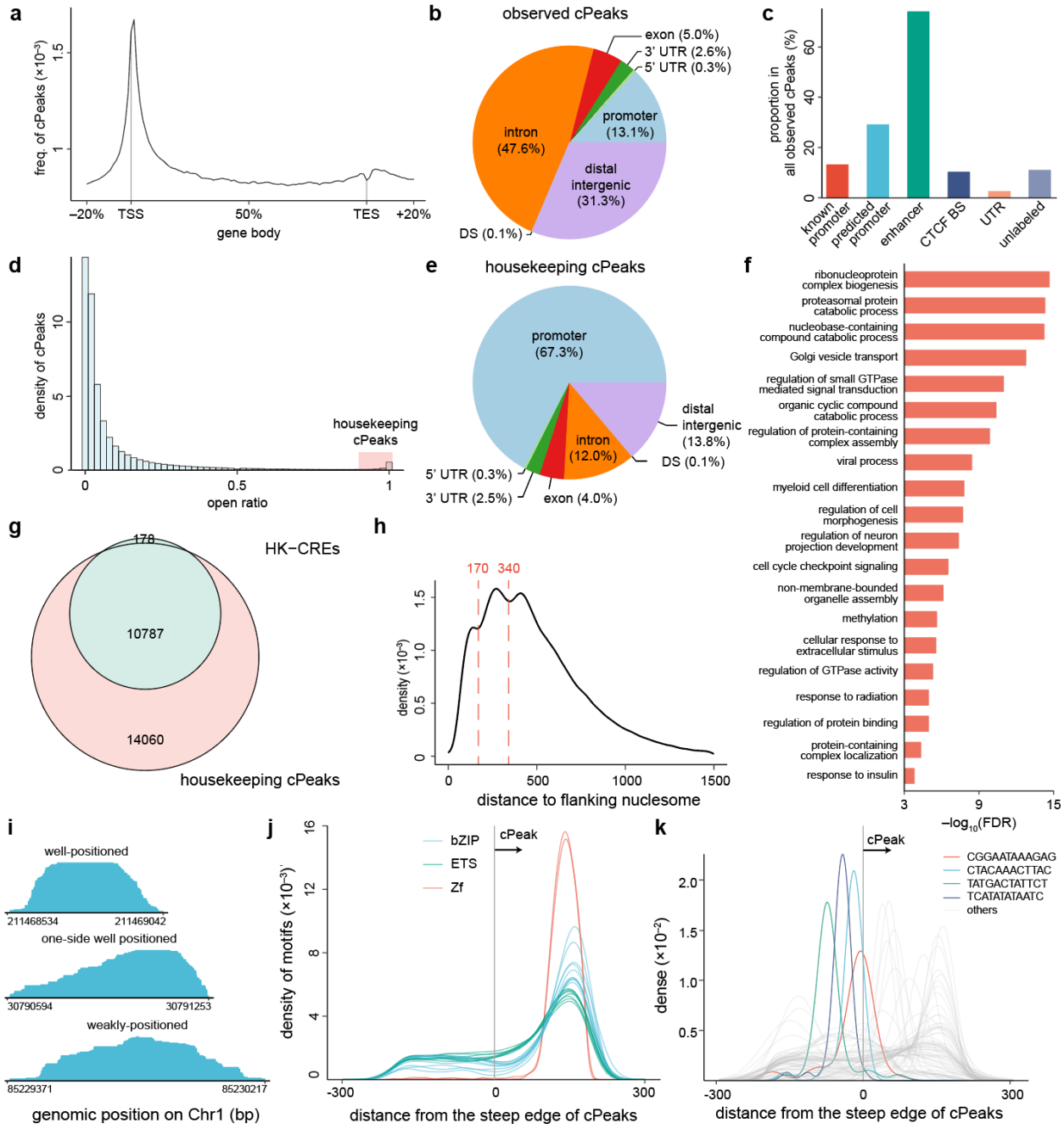


Fig. 2. Properties of observed cPeaks

(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) Venn plot of housekeeping cPeaks and HK-CREs. (h) Length distribution of cPeak edge length. Two dashed lines represented 170 bp and 340 bp. (i) Examples showing the three typical shape patterns of cPeaks. (j) Positions of the top 20 enriched motifs around the steep edges of cPeaks. The x-axis is ±300 bp 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. (k) Positions of the enriched *de novo* motifs around the steep edges of cPeaks. The x-axis is ±300 bp around the edges of cPeaks, from closed regions to open regions. Each line represents a motif. We colored top 4 motif on the closed regions around cPeaks.

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[57] 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. We utilized NeuronMotif[58] to decipher motifs captured by each convolutional kernel, which led to the identification of sequence patterns of positive and negative samples (Methods, Fig. S3a). 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 cPeaks, we applied the trained CNN model to scan the whole human genome, using a sliding window of 500 bp 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. S3b & S3c).

We collected another 1,161 bulk ATAC-seq datasets from ChIP-Atlas[49] 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 mean coverage of observed cPeaks was 91.3%, 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 (the mean 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 form a full set of cPeaks to represent the feature set of ATAC-seq and scATAC-seq data in the human genome for all cell types.

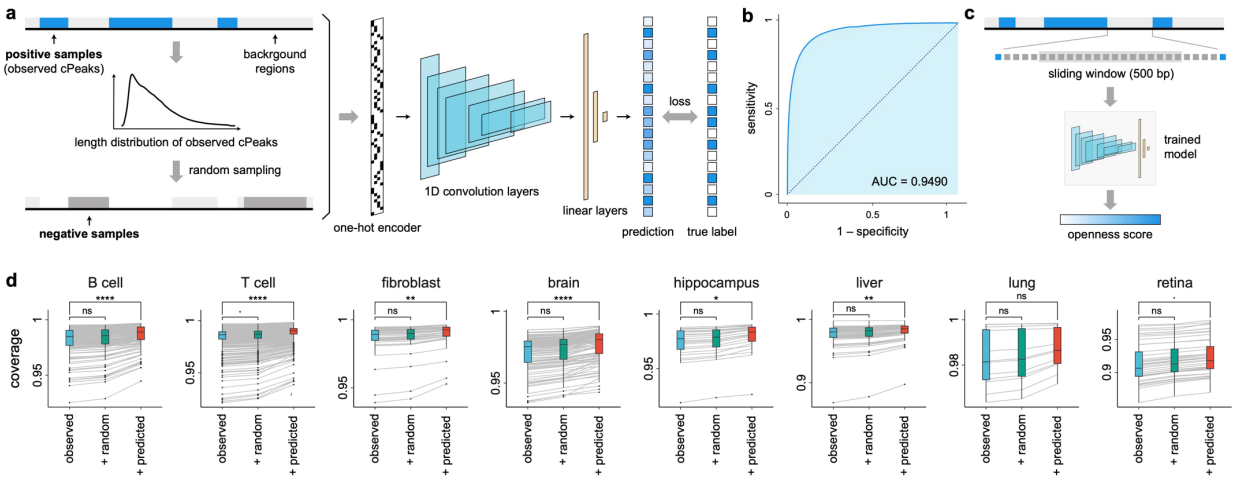


Fig. 3. Computational prediction of cPeaks with DNA sequences using a deep CNN model

(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

301 We compared the performance of using cPeaks as a reference on scATAC-seq data annotation with three commonly
302 used approaches in defining reference features and two pre-identified feature sets. The first approach divided the
303 genome into 2,000 bp bins as the basic units of the chromatin accessibility features[35,36,59]. We name them as
304 "genomic bins" for convenience. The second approach combined all sequenced single cells together as pseudo-bulk
305 samples and called peaks from them as the reference[18,32]. We name the features defined in this way as "pseudo-bulk
306 peaks". The third approach first clusters cells based on features such as pseudo-bulk peaks or labels derived from
307 cell-sorting, then call peaks by combining cells in each cluster, and then unite the peaks from all clusters together as
308 the reference features[5,29]. We name the features calculated in this way as "united peaks". Each approach provided a
309 set of genomic regions as reference features for mapping sequencing reads to generate a cell-by-feature matrix. We
310 also mapped reads to two pre-identified feature sets, cDHS and CATLAS cCREs, to generate matrixes. We
311 compared the three approaches and two feature sets with the cPeak-defined features for their performance in
312 downstream cell annotation tasks (Methods, Fig. 4a). For each approach, we experimented with different selection
313 levels of the features: all features, the top 50,000 highly variable features (HVFs), or the top 5,000 HVFs (Methods,
314 Fig. 4a).

315 We used an scATAC-seq dataset of hematopoietic differentiation[10] for the comparison. The dataset contains 2,034
316 cells with FACS-sorting labels as the ground truth of cell types. We generated the cell-by-feature matrices with each
317 feature-defining approach, used multiple commonly used embedding methods for dimensionality
318 reduction[8,21,32,35,46,60], and conducted the Louvain method[61] to cluster and annotate the cell types (Fig. 4a, Methods).
319 We evaluated the annotation results by comparing them with the known ground truth with the adjusted rand index
320 (ARI), adjusted mutual information (AMI), and homogeneity (H)[62] (Fig. 4a, Methods). Results showed that using
321 cPeaks as the features produced the best or one of the best performances in comparison with other feature-defining
322 approaches for all the annotation and evaluation methods (Fig. 4b and S4). The advantage of the performance
323 achieved with cPeaks over other feature-defining approaches is more obvious when we only used the top 5,000 or
324 top 50,000 features. Considering that the definition of cPeaks was independent of this particular dataset, this
325 performance confirmed that cPeaks captured the inherent PORs in the genome and can be used as a general
326 reference for all future bulk or single-cell ATAC-seq data.

327 There is also a widely used method for iterative dimensionality reduction called ArchR[63], which employs an iterative
328 reduction approach to select appropriate features and refine cell embedding results. We utilized the iterative method
329 of ArchR, conducting 1, 2, and 4 rounds of iteration, to compare the results of cPeak with the five feature sets
330 previously mentioned (Fig. 4a, Methods). The results showed that the performance of one iteration with cPeaks
331 surpassed that of other feature sets subjected to several iterations (Fig. 4c). Also, cPeaks with multiple iterations
332 performed better. The results suggested that cPeak can streamline the ArchR analysis workflow, achieving superior
333 results within one iteration compared to other features' multiple iterations. Additionally, if multiple iterations are
334 performed, cPeak stands out as the optimal set of features for scATAC-seq dataset.

335

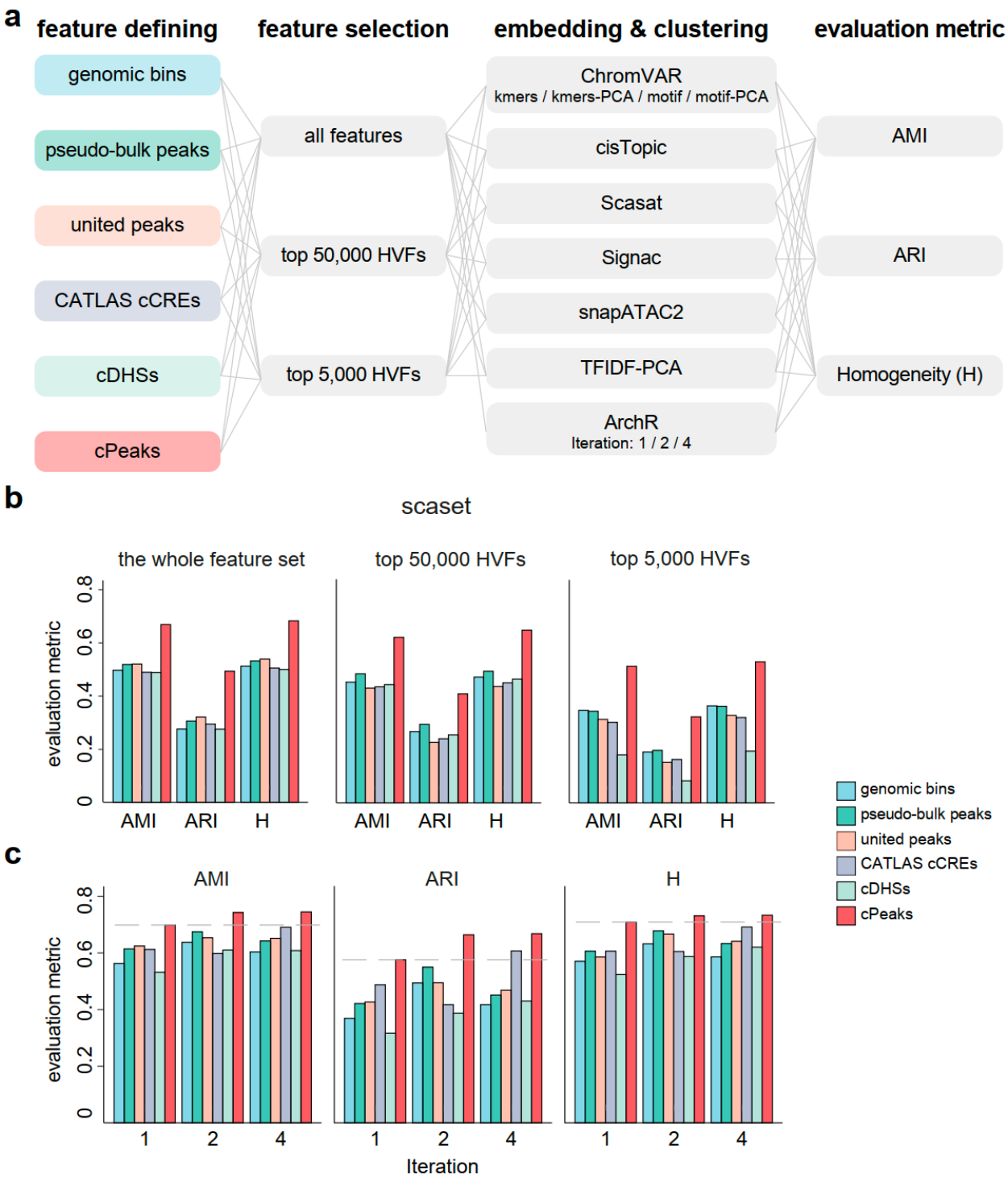


Fig. 4. The performance of cell annotations on the FACS2-sorted hematopoietic differentiation dataset

(a) The analysis workflow of cell annotations. We experimented with 6 feature-defining approaches to obtain references for generating cell-by-feature matrices, 3 ways of feature selection, and 7 strategies to embed the features before clustering analysis, and used three metrics to evaluate the results. (b) Comparison of the performances of the 6 feature-defining approaches at different levels of feature selection, measured by the 3 scores: AMI, ARI, and H. (c) Comparison of the performances of the 6

342 feature-defining approaches at different iterations using ArchR. Each gray dashed line represents the evaluation metric under one
343 iteration using cPeaks as a reference.

344

345 cPeaks enable the discovery and analysis of rare cell types

346 One expected advantage of using a general reference rather than a dataset-dependent reference is the discovery of
347 rare cell types, as cells of rare types may not be sufficient in the dataset to call out the marker peaks that are specific
348 to the rare cell types. We checked the potential of cPeaks with this respect by using them to annotate an scATAC-
349 seq dataset of peripheral blood mononuclear cells (PBMCs) generated by 10X Genomics. It includes 8,728 cells
350 with 87,561 pre-identified pseudo-bulk peaks. We first used cPeaks as the reference to get the cell-by-feature matrix,
351 and applied normalization, feature selection and dimensionality reduction following the common practice
352 (Methods). We then performed graph-based clustering on the reduced dimension and visualized the clustering
353 results using Uniform Manifold Approximation and Projection (UMAP) plots (Fig. S5a). Results showed that the
354 cells were grouped into multiple clusters, with some clusters containing small proportions of cells. We then
355 transferred the cell type labels from an accompanying scRNA-seq PMBC dataset from 10X Genomics by cross-
356 modality integration using the anchor method[64] in Seurat (Fig.5a). For each cell in the scATAC-seq data, we
357 obtained its inferred cell type with a prediction score that indicates the confidence of this cell type prediction.
358 Compared with the results obtained by using the pseudo-bulk peaks to generate the cell-by-feature matrix, cPeak-
359 generated results obtained improved prediction scores (Fig. 5b and S5b).

360 Checking the cPeaks-derived clusters after the cell-type annotation, we found three rare cell types in the dataset:
361 plasmacytoid dendritic cells (pDCs), dendritic cells, and natural killer bright (NK bright) cells. Each of them
362 contains only less than 1% of the cells in the dataset (Fig. S5c). We took pDCs as an example to check their
363 differential open regions in comparison with all other cells (Methods). Results showed that using cPeaks identified
364 1,449 more differential open regions than using the pseudo-bulk peaks (Fig. 5c), most of which were specifically
365 open in pDCs (Fig. 5d and S5d). Due to the low abundance of pDCs, these regions were hard to be detected by
366 pseudo-bulk peaks. These open regions were important for investigating the functions of rare cell types. For
367 example, we found that the promoter region of the PTCRA gene was a pDC-specific open region (Fig. 5e and 5f).
368 According to the scRNA-seq data, PTCRA was only expressed in pDCs (Fig. 5g). Previous studies also confirmed
369 the expression of PTCRA is the feature of human pDCs[65,66].

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

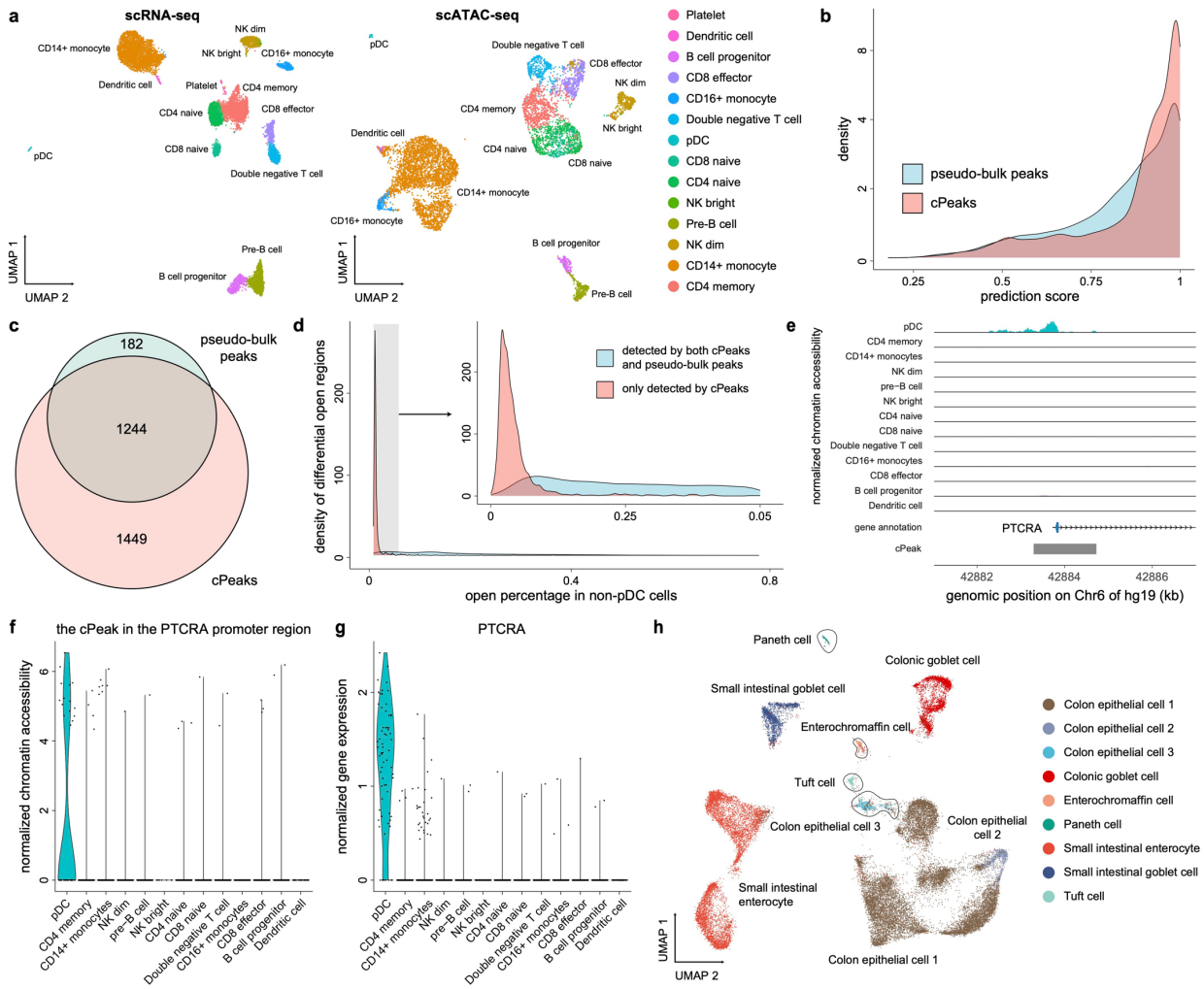


Fig. 5. Rare cell type discoveries using cPeaks as the reference of features

(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 are missed by the analysis based on pseudo-

394 bulk peaks. (e) The chromatin openness in all cell types around the promoter regions of PTCRA, measured by the sum of
395 scATAC-seq reads of all cells of each type. Only pDC cells have a strong openness signal in this region, verifying pDC-specific
396 accessibility of the PTCRA promoter region detected using cPeaks. (f) The distribution of the normalized reads aligned to the
397 cPeak in the PTCRA promoter region across different cell types. The reads are normalized by the frequency-inverse document
398 frequency (TF-IDF) method. Each dot represents a cell. (g) The normalized expression of PTCRA in different cell types in
399 scRNA-seq data. The gene expression is normalized by the total expression in each cell and log-transformed. (h) The UMAP plot
400 of subtypes in GI epithelial cells, clustered using cPeaks as the reference. The colors show the cell types provided by the original
401 paper. The black circles mark the 4 rare cell subtypes: paneth cell, enterochromaffin cell, tuft cell and colon epithelial cell 3.

402

403 Analyze dynamic biological processes and diseases with cPeaks

404 cPeaks were defined using the data of ATAC-seq datasets of adult healthy samples. We investigated their usability
405 and performance in analyzing developmental or disease data. We first collected an scATAC-seq data of human
406 hematopoiesis[32] to study the use of cPeaks in analyzing cell differentiations. We used cPeaks as the reference to
407 construct the differentiation trajectories of human hematopoiesis lineage from these data. The cells displayed
408 continuous state changes in the differentiation process[32] (Fig. 6a). We aligned the reads of each cell in the scATAC-
409 seq data to the cPeaks to get the cell-by-feature matrix, and labeled each cell based on its annotation in the original
410 study (Fig. 6b, Methods). We estimated pseudotime from the cell-by-feature matrix with Monocle 3[67,68] (Fig. 6c,
411 Methods). The estimated pseudotime was consistent with the original study.

412 We then collected the human fetal retinal scATAC-seq dataset covering cells in early and late stages of retina
413 organogenesis[33]. We used cPeaks as the reference to generate the cell-by-feature matrices. We plotted cells with
414 UMAP and labeled them according to the original paper (Methods). For cells in the early stages (8 weeks) of
415 development (Fig. 6d), we observed three trajectories from multipotent progenitor cells to horizontal cells, cones,
416 and retinal ganglion cells (Fig. 6e). We can see that the pseudo-time values of these cells also showed an order in the
417 UMAP (Fig. 6f). For cells in the late stages (13 weeks), we separately analyzed samples collected from the center
418 and the periphery of the retina, and the results showed clear clusters of cell types and their transition states (Fig.
419 S6a-b). It was worth noting that datasets used for generating cPeaks did not include fetal stages or retina organs. The
420 results demonstrated the capability of cPeaks for studying fetal development and unexplored organs, and suggested
421 that the set of PORs constructed from data of adult samples also cover most of the potential open regions in early
422 developmental stages.

423 We then investigated the use of cPeaks in analyzing disease data. Our first experiment on disease data was on an
424 scATAC-seq dataset of both healthy kidneys and kidneys with DKD[18]. DKD or diabetic kidney disease is a chronic
425 inflammatory disease which can cause cellular injury and alter cell types within kidneys. Five of the patient samples
426 in the dataset contain more than 2,000 sequenced cells each. We used these data of 5 patient samples and of one
427 healthy sample to study whether the correct conclusion can be derived using cPeaks as the reference. Fig. 6g showed
428 the compositions of cell types in these samples reported in the original study. We visualized the cells in the healthy
429 samples (Fig. 6h) and patient samples (Fig. 6i, Fig. S6c-f) using the cPeaks-derived cell-by-feature matrices

430 (Methods). We can see that the cell types reported in the original paper are well preserved using cPeaks as the
431 reference. For example, we successfully re-discovered the increased infiltrating leukocytes cells such as B cells, T
432 cells, and mononuclear cells in patient 2 (Fig. 6i), and also verified the increased population of VCAM1+ proximal
433 tubule cells in patient 3 (Fig. S6d), which is a marker of kidney injury[69].

434 The second experiment on diseases was about the clustering of tumor cells. We collected a human gynecologic
435 malignancy scATAC-seq dataset of 74,621 cells in the endometrium or ovary of 11 treatment-naive cancer patients[34]
436 and used cPeaks as the reference to re-analyze the data (Methods). Fig. 6j and S6g showed that cells analyzed with
437 cPeaks could be well clustered by cell types. Cells of the three types of cancers, gastrointestinal stromal tumor
438 (GIST), endometrial cancer, and ovarian cancer, were separated from normal cells. We also labeled the cells in the
439 UMAPs with the copy number variations (CNVs) (Fig. 6k) and expression of KIT, WFDC2, and MUC16 (Fig. S6h)
440 reported by the original study. Besides reproducing the observations in the original study, we found that the ovarian
441 cancer cells in sample 8 were separated into two clusters (Fig. 6j and S6g). Both the clusters expressed the cancer
442 marker gene WFDC2 (Fig. S6h), but only one of them was enriched with CNVs (Fig. 6k, Methods). These new
443 observations suggested that the tumor of sample 8 might contain two subclones with different genomic properties.
444 The results demonstrated that cPeaks can help to identify subclusters within cancer cells, and the use of this general
445 reference may help to identify new clues in a small portion of cells that could be missed in studies using data-
446 derived references.

447

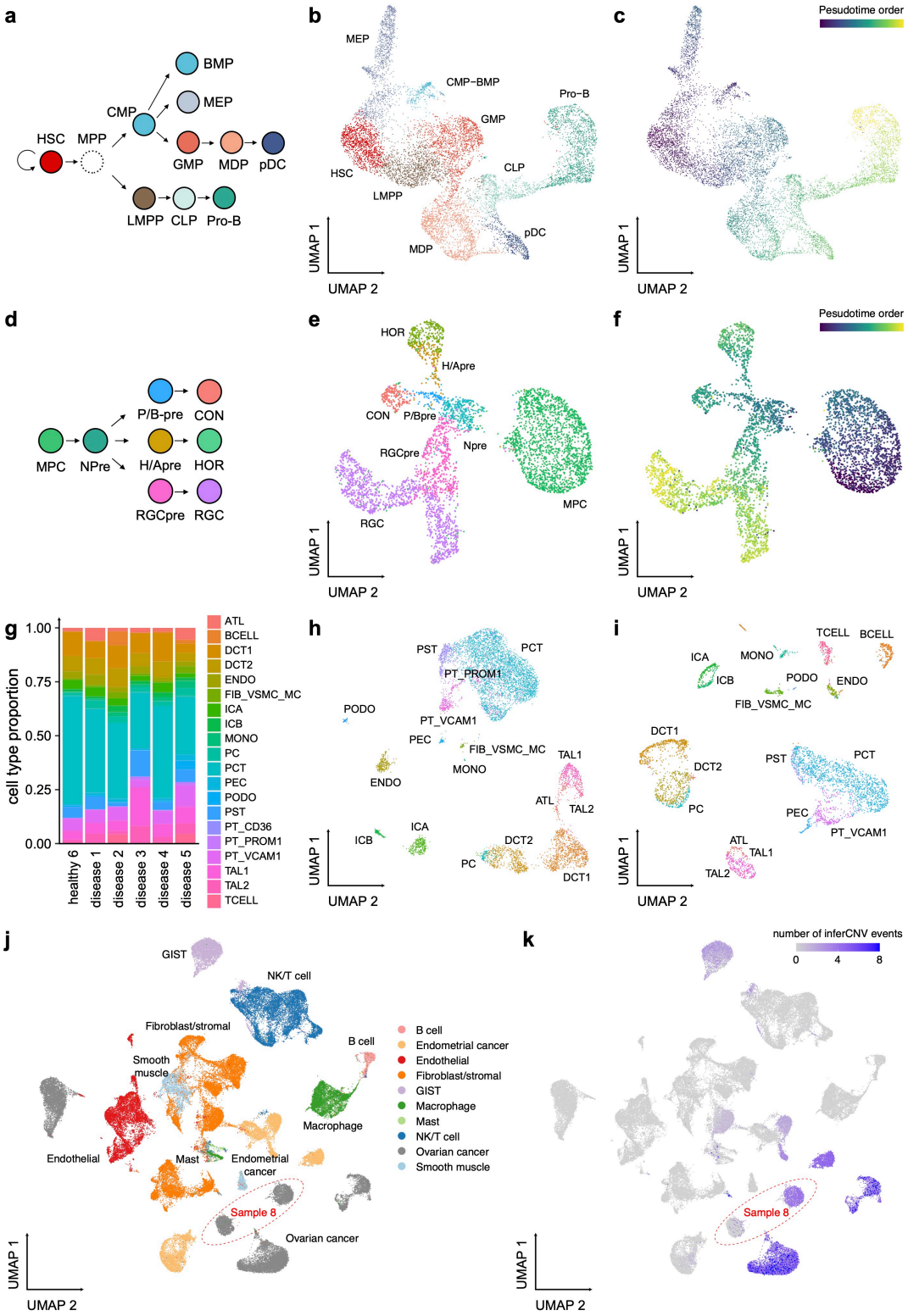


Fig. 6. Analyzing developmental and disease data using cPeaks

(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-

452 erythroid progenitor; CMP, basophils and common myeloid progenitor; BMP, basophil-mast cell progenitor; MEP,
453 megakaryocyte-erythroid progenitor; GMP, granulocyte-macrophage progenitors; MDP, monocyte-dendritic cell progenitor. As
454 MPP is not detected in cells, we use a dashed circle to represent it. (b) The UMAP plot of the human hematopoiesis scATAC-seq
455 data colored by cell types. (c) The UMAP plot of the human hematopoiesis scATAC-seq data colored by pseudotime. (d) The
456 schematic diagram of the cell-fate decision in the early (8 weeks) fetal retina. MPC, multipotent progenitor cell; Npre, neurogenic
457 precursor; P/B-pre, photoreceptor/bipolar precursors; CON, cones; H/Apre, the horizontal/amacrine precursors; HOR, horizontal;
458 RGC, ganglion cells; RGCpre, RGC precursors. (e) The UMAP plot of the 8-week retina scATAC-seq data colored by cell types.
459 (f) The UMAP plot of the 8-week retina scATAC-seq data colored by pseudotime. Combining the pseudotime and the labels of
460 cells in Fig. 6e, the development trajectory of retina cells is consistent with the knowledge in Fig. 6d. (g) The cell type
461 proportions of each sample in the kidney dataset. The x-axis showed the sample name and y-axis showed the cell type proportion.
462 Each color represented one cell type: ATL, ascending thin limb; BCELL, B-cell; DCT1, DCT1-early distal convoluted tubule;
463 DCT2, DCT2-late distal convoluted tubule; ENDO, ENDO-endothelial cells; FIB_VSMC_MC, vascular smooth muscle cells and
464 mesangial cells; ICA, ICA-type A intercalated cells; ICB, ICB-type B intercalated cells; MONO, mononuclear cells; PC,
465 principal cells; PCT, proximal convoluted tubule; PEC, parietal epithelial cells; PODO, podocytes; PST, proximal straight tubule;
466 PT_CD36, CD36(+) proximal tubule; PT_PROM1, PROM1(+) proximal tubule; PT_VCAM1, VCAM1(+) proximal tubule;
467 TAL1, CLDN16(–) thick ascending limb; TAL2-CLDN16(+) thick ascending limb; TCELL, T cell. (h) The UMAP plot of the
468 scATAC-seq data of the healthy sample in the kidney dataset, colored by cell types. (i) The UMAP plot of the scATAC-seq data
469 of the DKD patient 2 sample in the kidney dataset, colored by cell types. (j) The UMAP plot of the human gynecologic
470 malignancy scATAC-seq data. We aggregated all cells in the endometrium and ovary from 11 patients. Colors represent different
471 cell types. Among them, GIST, endometrial cancer, and ovarian cancer are cancer cells. Ovarian cancer cells in sample 8 are
472 circled by the red dashed line. GIST, gastrointestinal stromal tumor; NK/T cell, nature killer cell or T cell. (k) The UMAP plot of
473 the human gynecologic malignancy scATAC-seq data colored by inferred CNVs. The ovarian cancer cells provided by sample 8
474 are circled by red dashed line. The two sub-clusters in these cells have different CNVs. All the results are analyzed with cPeaks
475 as the reference.

476

477 Rebuild human chromatin accessibility atlas with cPeaks

478 We rebuilt a human chromatin accessibility atlas, utilizing scATAC-seq data encompassing 30 adult tissue types[5]. It contains a
479 unified feature set, named CATLAS cCREs, which was generated through pseudobulk peak calling and peak integration. We
480 applied cPeak and CATLAS cCREs as features to do one-round cell clustering following the strategies in the original paper[5]
481 (Methods). We used the 30 major cell groups and 111 cell types annotated by the original paper as the ground truth[5]. We found
482 that both cPeak and CATLAS cCREs showed comparable performance using 30 major cell groups as labels (Fig. 7a), with three
483 evaluation metrics (AMI, ARI, and Homogeneity) scoring around 0.9 (Fig. S7a) and similar silhouette scores (Fig. S7b). When
484 we increased the number of clusters and performed assessments using 111 cell type labels still on a single round of clustering.
485 Results showed that cPeaks were slightly superior to CATLAS cCREs on all evaluations, for example, ARI increased by 0.03,
486 and silhouette score increased by 0.02 (Fig. S7c & S7d).

487 We further classified these 111 cell types into three groups according to their proportions in the atlas. With proportions from low
488 to high, we named three groups as extremely rare cell types, rare cell types, and common cell types (Methods). The number of
489 cell types in each group was similar. The average proportions of a certain cell type in three groups were 0.1%, 0.3%, and 1.2%,
490 respectively. We calculated the clustering indices separately in these three cell groups (Methods, Fig. 7b). For extremely rare cell
491 types, their proportion was 4%. Even when given a complete feature set, these cells cannot form separate clusters due to limited

cell numbers. So, both cPeak and CATLAS cCREs got nearly the same clustering results. To identify these cell types, an extra clustering of the major cell groups is necessary. We have already given an example of the clustering results on the GI epithelial cell group in Figure 5h, which contains four extremely rare cell types. Results showed that these cell types can be distinguished after another round of clustering using cPeaks as features. For common cell types, cPeak achieved slightly higher indices than CATLAS cCREs, with ARI increased by 0.03. As these common cell types constituted 82% of all cells, they governed the overall clustering metric values and the difference between the metrics of cPeak and cCREs (Fig. S7c & S7d). For rare cell types, we observed that in comparison to cCREs, cPeaks demonstrated the greatest increase in clustering metrics for these cells, with ARI increased by 0.07. Although these cell types had low proportions, they were distinguishable after being furnished with sufficient features. The CATLAS cCREs may lose some rare cell type-specific features due to multiple rounds of peak merging and trimming. In contrast, cPeak improved clustering metrics by providing a more comprehensive feature set.

Results showed that cPeak qualifies as a feature set for constructing a human chromatin accessibility cell atlas without the necessity for additional feature generation, as cPeaks can achieve the same performance as the temporary features. Compared to the feature set obtained via clustering and peak integration approaches, cPeaks improves the clustering of cell types, especially the rare cell types with lower cell proportions.

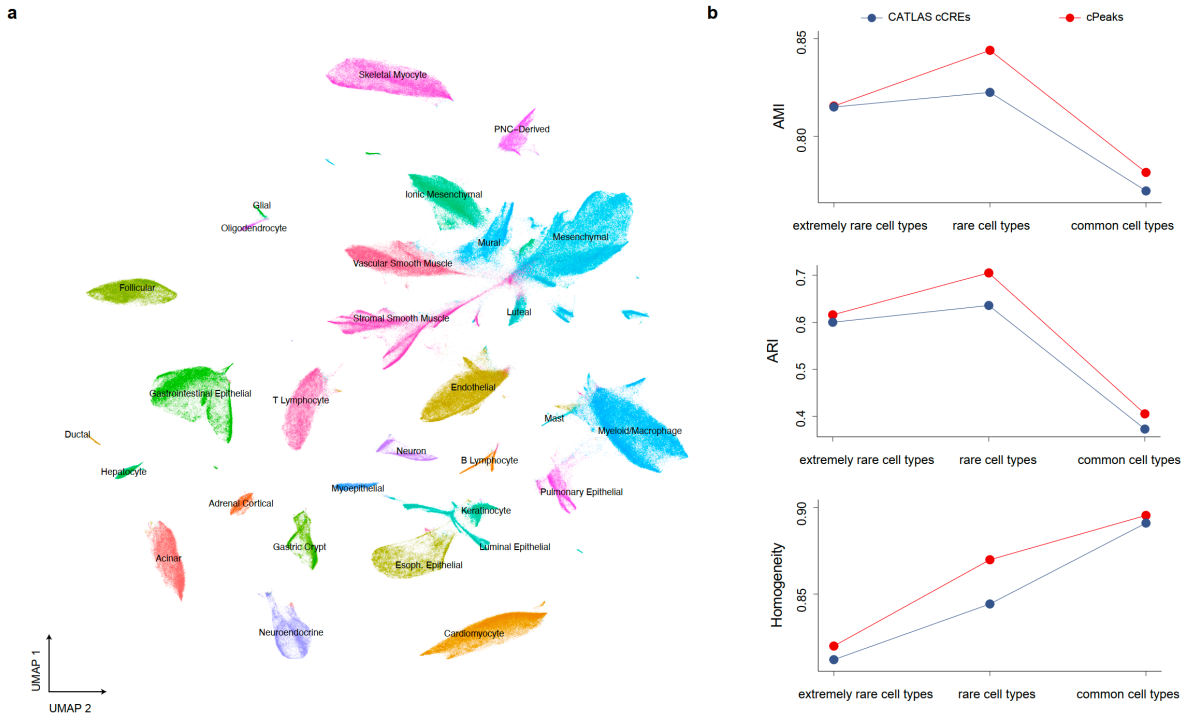


Fig. 7. Visualization and evaluation of building human cell atlas using cPeaks.

(a) The UMAP plot of atlas data colored by major cell groups. (b) The evaluation metrics on different cell groups. Different colors represent different references while mapping sequencing reads.

cPeaks as a universal reference for ATAC-seq data analysis

513 We provided a comprehensive resource for cPeaks, including basic information, cPeak annotation, and integration
514 with biological data. The basic information includes cPeak ID, genomic locations, and length. The cPeak annotation
515 shows the housekeeping status, shape pattern, and inferred regulatory function of each cPeak. Since each cPeak
516 depicts POR on the genome, such as cCREs, we connected cPeaks with other pre-identified feature sets and known
517 regulatory elements. We mapped cPeaks to the genome regions defined by cDHS[37], CATLAS cCREs[5], ReMap[70,71]
518 and rDHS[72], which provided relevant information about regulatory vocabulary, cell-type specificity, and binding
519 TFs associated with each cPeak (Methods, Data Availability). We provided bioinformatics tools for mapping or
520 converting diverse ATAC-seq data types to cPeaks and integrated cPeaks with current widely-used scATAC-seq
521 analysis tools, such as ArchR[63] and SnapATAC2[73] (see Code Availability for details).

522

523 # Discussion

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

536 We conducted a series of experiments to validate the usefulness and benefits of using cPeaks as the general
537 reference for ATAC-seq and scATAC-seq data analyses. The experiments showed that cPeaks can achieve
538 performances equally well with or better than the best existing approaches for building peak-calling references or
539 iteratively clustering in major downstream tasks such as cell type annotation and trajectory analysis. In addition,
540 cPeaks can better identify rare cell types or subtypes from scATAC-seq data. The set of cPeaks derived from healthy
541 adult samples can also work well on data of early developmental stages and data of major inflammatory diseases and
542 cancer. cPeaks also performed well as the feature set for constructing scATAC-seq atlas.

543 The success of cPeaks as a general data-independent reference in all these tasks further supported different cell types
544 shared a common feature set in the human genome. We therefore proposed that all future ATAC-seq or scATAC-
545 seq data can be analyzed by mapping the sequencing reads to the PORs represented by the cPeaks. In this way, we
546 can have cPeaks as the equivalence of the transcriptome in scRNA-seq processing, which will make comparison and
547 integration of data from different studies uniform and standardized. Currently, the emergence of various foundation

548 models based on transcriptions has provided a significant understanding of life science[74–77]. It's also essential to
549 establish foundation models for chromatin accessibility data. One of the key challenges is the inability to standardize
550 features across different datasets. The advent of cPeak provides a groundwork for the development of foundation
551 models on chromatin accessibility data or integrated multi-omics data.

552 We are aware that the set of cPeaks we built in this work is only an initial version based on the current data and
553 computational model. Many extra works can be done to further refine the definition and annotation of cPeaks. Aside
554 from benefiting more convenience and accurate downstream analysis, we hope that cPeak can enable direct
555 comparison of data among different teams, and cPeaks can serve as a bridge to link features to underlying biological
556 content. We provided convenient tools for acquiring and mapping cPeak for various uses on different platforms. We
557 believe that the current data volume is considerable enough to establish a unified feature set, like cPeak. The
558 establishment of this set can promote a more open and convenient epigenomics analysis community.

559

560 Methods

561 Generation of cPeaks

562 Bulk ATAC-seq data for cPeak generation

563 We collected all bulk ATAC-seq data from ENCODE[38] and ATACdb[45]. We selected all adult, healthy datasets with
564 more than 10,000 pre-identified peaks. For each dataset, if the peak number of chromosome Y (ChrY) was less than
565 20, we removed all peaks on ChrY, and the remaining peaks were used for further analysis. We analyzed all the data
566 with the hg38 genome reference. For ATACdb datasets which were analyzed with the hg19 reference in the original
567 study, we used CrossMap[78] (version 0.6.5) to convert positions of all peaks to the hg38 reference. We removed all
568 peaks that altered lengths during conversion.

569 Defining merged peaks and their shapes

570 We defined the merged peaks by joining peaks. Taking advantage of the high quality of ENCODE and the
571 comprehensive cell types of ATACdb, we generated merged peaks for ENCODE and ATACdb separately. The final
572 merged peaks included two parts: (1) all ENCODE merged peaks, and (2) all ATACdb merged peaks that were not
573 overlapped with any ENCODE merged peak. When defining merged peak shapes on ChrY, we only considered the
574 number of datasets containing ChrY after quality control as the denominator.

575 Peak shape consistency calculation

576 First, for peak shape consistency in different cell types, we compared merged peaks generated by 48 blood samples
577 with peaks generated by 101 solid tissue samples in ENCODE. Second, for peak shape consistency in sequencing
578 techniques, we compared merged peaks generated by 149 ENCODE samples with peaks generated by 733 DNase-
579 seq samples. The DNase-seq samples were the peak files for cDHS[37] generation. We downloaded them from
580 https://www.encodeproject.org/annotations/ENCSR857UZV/ and merged peak files. Third, for peak shape

consistency in peak calling methods, we compared merged peaks generated by 149 ENCODE samples with peaks generated by 17 10X samples which were downloaded from https://www.10xgenomics.com/datasets/10k-human-pbmcs-atac-v2-chromium-controller-2-standard. Here, the 17 10x samples were generated by separating sequencing reads following graph clustering results (https://cf.10xgenomics.com/samples/cell-atac/2.1.0/10k_pbmc_ATACv2_nextgem_Chromium_Controller/10k_pbmc_ATACv2_nextgem_Chromium_Controller_analysis.tar.gz). Fourth, for peak shape consistency in peak calling parameters, we compared merged peaks generated by 149 ENCODE samples with peaks generated by all ATACdb samples. Here, we defined the merged peaks generated by all 149 ENCODE files as the ENCODE merged peaks. For each comparison, we selected the ENCODE merged peak region where both merged peak shape maximum values were greater than 0.3. We calculated the Pearson correlation of the merged peak shapes on each selected region. We used permutation to test the significance of each comparison. We randomly shuffled the pair of merged peaks in each comparison on selected regions 100 times. We used spline sampling to equalize the peak length of each mixed pair and calculated the Pearson correlation.

Generation of cPeaks from merged peaks

We used the Hartigan's dip test[79] (the "dip.test" function in the diptest package (version 0.76-0) 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)[80] to cluster the matrix, using the "hdbscan" function in the dbscan package (version v1.1-11) 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 separate 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 and divided it by the total number of aggregated reads.

Comparison of cPeaks and other pre-identified feature sets

For coverage of cPeaks and cDHS on ATAC-seq open region, we downloaded all the DNase-seq and ATAC-seq data of GM12878 from the ChIP-Atlas[49] website. We set the website's parameter "significant" as 50 and filtered the

616 downloaded files by setting the peak number between 30,000 and 100,000. We acquired a total of 65 datasets, which
617 contain 3 DNase-seq datasets. We paired each DNase-seq dataset with the ATAC-seq datasets with the closest
618 number of peaks and got three pairs of datasets, namely SRX5766093 & SRX069089, SRX7124440 & SRX069213,
619 and SRX5766134 & SRX069213. We defined the specific open regions determined by ATAC-seq or DNase-seq in
620 each pair of data sets as ATAC-seq/DNase-seq unique regions.

621

622 Properties of cPeaks

623 Genomic annotations of cPeaks

624 We used the ChIPSeeker[81] package (version 1.35.1) to perform genomic annotations of cPeaks. We defined the
625 promoter region as ±2,000 bp around TSS and annotated cPeaks using the "annotatePeak" function. The annotation
626 database was EnsDb.hg38.v86[30] provided by the ensembldb package. ChIPSeeker annotated cPeaks into seven
627 types: promoter, 5' UTR, 3' UTR, exon, intron, downstream (≤ 300 bp downstream of the gene end), and distal
628 intergenic regions. We used the "plotPeakProf2" function in ChIPSeeker to calculate the enrichment of cPeaks on
629 gene body regions, with a confidence interval of 0.95.

630 Inferring regulatory roles of cPeaks

631 we defined cPeaks overlapped with any gene promoter region as known promoters. We annotated cPeaks
632 overlapped with any 3' UTR as UTRs. We collected 210 CTCF ChIP-seq data in ENCODE[38] and predicted cPeaks
633 that overlapped with any CTCF ChIP-seq signals as CTCF binding sites. We collected histone modification data in
634 ENCODE[38] and used their called peaks as signals. cPeaks overlapping with H3K4Me1 signals were predicted as
635 enhancers. cPeaks overlapping with H3K4Me3 signals and not overlapping with known promoter regions were
636 regarded as predicted promoters.

637 Housekeeping cPeaks

638 We used a pre-defined housekeeping geneset[50] and downloaded it from
639 https://www.tau.ac.il/~elieis/HKG/HK_genes.txt. We obtained a contingency table to describe the amount of
640 housekeeping or non-housekeeping genes related to housekeeping or non-housekeeping cPeaks. We only considered
641 protein coding genes. We used Fisher's exact test to calculate the significance of the contingency table.

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

647 Known motif enrichment and motif scan of cPeaks with different shape patterns

648 The endpoints of cPeaks (ENCODE part) were extended to 200 bp within the cPeak's sides, resulting in well-

649 positioned 200 bp parts and weakly positioned 200 bp parts. We then used Homer[82] "findMotifsGenome.pl" to
650 enrich the motif, with well-positioned parts as the experimental group and weakly positioned parts as the
651 background. To purify the signal, we only chose the endpoints of the unimodal peak before splitting to scan the
652 motif. We used Homer "annotation.pl" to annotate the enriched known motifs. We got the position of each motif
653 around the endpoint and drew the Kernel Density Estimate plot using seaborn.kdeplot[83].

654 **DNA binding domain and GO enrichment of enriched TFs**

655 First, we filtered the enriched TFs with an adjusted p-value of less than 0.1 for domain and GO enrichment. We used
656 the "enrichr" function in the enrichR library for enrichment analysis. For domain enrichment, we use
657 GO_Biological_Process_2023 database and set the adjusted p-value threshold to 1e-5. After obtaining the GO
658 enrichment results, we clustered the enriched GO terms using "GO_similarity" function in the "simplifyEnrichment"
659 library.

660 **The enrichment of pioneer transcription factors (PTF) on enriched TFs**

661 We initially collected 32 known Pioneer Transcriptional Factors (PTFs) from paper[54], which are: 'FOXA1',
662 'FOXA2', 'FOXA3', 'GATA1', 'GATA2', 'GATA3', 'GATA4', 'CEPBA', 'CEBPB', 'ESRRB', 'POU5F1', 'KLF4',
663 'NEUROD1', 'TP53',' FOS', 'FOSL1', 'FOSL2', 'MAFG', 'MAFF', 'MAFK', 'JUN', 'JUNB', 'JUND', 'ATF2', 'ATF3',
664 'ATF4', 'ATF6', 'ATF7', 'SPI1', 'ASCL1', 'HNF1A', and 'PU.1'. We mapped the 440 TFs in the Homer database to
665 these PTFs and identified 25 PTFs in this database, 22 of which were significantly enriched (adjusted p-value <0.1).
666 For the remaining non-PTFs, 269 displayed significant enrichment. We then used Fisher's exact test, resulting in a p-
667 value of 0.016, indicative of a significant association between the PTFs and enriched TFs on well-positioned cPeaks.

668 ***de novo* motif enrichment and motif scan of cPeaks with different shape patterns**

669 We extended to ±200 bp around the endpoints of the well-positioned edges in unimodal cPeaks as an experiment
670 group. We then used Homer "findMotifsGenome.pl" function to find de novo motif on an experiment group, with
671 the random sampled background. We chose the endpoints of the unimodal peak to scan the *de novo* motif with
672 Homer "annotation.pl" function.

673

674 **Extension of cPeaks**

675 **The architecture of the deep CNN model**

676 The architecture of the deep CNN model was modified based on scBasset[57], a one-dimensional (1D) CNN that
677 included six 1D convolutional blocks and two linear blocks. Each convolutional block comprised a 1D convolution
678 layer, a batch normalization layer, a dropout layer, a max pooling layer, and a GELU activation function. Each linear
679 block contained a linear layer, a batch normalization layer, a dropout layer, and a GELU activation function. The
680 numbers of convolutional kernels of the six 1D convolutional layers were 288, 323, 362, 406, 456, and 512,
681 respectively. The kernel sizes were 17, 5, 5, 5, 5, and 5, respectively. After convolution, a linear layer transformed

682 the 1D output of the last convolutional layer from 5120 to 32 dimensions, and another linear layer transformed the
683 output of the previous linear layer from 32 to 2 dimensions. The final activation function is softmax for binary
684 classification.

685 **Generation of negative samples**

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

697 **Model training**

698 We used one-hot encoding to encode the sequences of cPeaks (positive samples) and negative samples. We used
699 [0,0,0,0] to represent base N (unknown nucleotide) in any sequence. We padded all sequences with N at their ends to
700 make their lengths equal to 2,000. We split the positive and negative samples into two groups respectively: 80% for
701 training and 20% for test.

702 We fed these sequences into the model and trained the model to learn the label of these samples (positive samples as
703 1, and negative samples as 0). We regarded the model output as the openness score of this sample. During training,
704 we used the binary cross-entropy loss function and assessed the loss of the test samples after each epoch. We trained
705 the model for 5 epochs and found it had achieved a stable accuracy.

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

710 **Model explanation**

711 We employed NeuronMotif[58] to extract motifs captured in the convolutional kernels of the trained CNN model. We
712 fed the weights of well-trained trained CNN into the NeuronMotif. This software analyzed the weights in each
713 convolutional kernel and explained the weights into motifs. We had 6 layers of the convolutional kernel in the
714 trained CNN model, we chose the third, fourth, and fifth layers of the convolutional kernels to analyze. The output
715 of NeuronMotif was the motif in each convolutional kernel. We randomly selected 20 kernels in each layer to

716 analyze these motifs. Then, we used the meme FIMO algorithm[84] to scan these motifs in both positive and negative
717 samples of CNN model.

718

719 **The resources of cPeaks**

720 It contained the basic information, cPeak annotation, and integration with biological data. It included 15 columns:

721 - **ID**: The ID of this cPeak. We employed a 14-character string as the unique ID for each cPeak. cPeak ID begins
722 with POR for potential open regions, then HS for human, and ends with nine-digital number. For example, the
723 first cPeak was encoded as PORHS000000001.
724 - **chr_hg38**: The chromosome of this cPeak in hg38.
725 - **source**: The source of this cPeak: "observed" or "predicted".
726 - **start_hg38**: start positions of this cPeak in hg38 reference.
727 - **end_hg38**: end positions of this cPeak in hg38 reference.
728 - **housekeeping:** the housekeeping status, whether this cPeak is open in almost all datasets, "TRUE" or
729 "FALSE".
730 - **openMode**: the open mode of cPeaks, "well-positioned", "one-side", "weakly-positioned", or "else".
731 - **inferredElements**: The inferred regulatory elements, "CTCF", "TES", "TSS", "Enhancer" or "Promoter"
732 - **chr_hg19**: The chromosome of this cPeak in hg19 reference.
733 - **start_hg19**: start and end positions of this cPeak in hg19 reference.
734 - **end_hg19**: start and end positions of this cPeak in hg19 reference.
735 - **cDHS_ID**: The overlapped cDHS ID of this cPeak, names with "chr_start_end" in hg38 reference.
736 - **rDHS_ID**: The overlapped rDHS ID of this cPeak, names with "chr_start_end" in hg38 reference.
737 - **ReMap_ID**: The overlapped ReMap ID of this cPeak, names with "chr_start_end" in hg38 reference.
738 - **CATLAS_ID**: The overlapped CATLAS cCREs of this cPeak, names with "chr_start_end" in hg38 reference.

739

740 **Evaluating the cell annotation performance of cPeaks**

741 **Data preprocessing and cell-by-feature matrix generation**

742 We directly downloaded the BAM files of the FACS2-sorting hematological differentiation scATAC-seq data from
743 https://github.com/pinellolab/scATAC-benchmarking/tree/master/Real_Data/Buenrostro_2018/input/sc-
744 bams_nodup, which was provided by a benchmark study[62]. The BAM files contain 2,034 cells with the hg19

745 reference genome.

746 We used six feature sets to generate cell-by-feature matrices: genomic bins, pseudo-bulk peaks, united peaks and
747 cPeaks. Each feature set contained a set of genomic regions on the genome as open regions. For genomic bins, we
748 split the whole genome in the hg19 version into 2,000 bp fix-length bins. Each bin was regarded as a feature. For
749 pseudo-bulk peaks, we used the pre-defined peaks of the original study[10] from the Gene Expression Omnibus (GEO)
750 database under accession code GSE96769. In the original study, all sequenced cells were aggregated as one
751 population for peak calling via MACS2[31]. For united peaks, we used the peak set provided by the benchmark study[62]
752 from https://github.com/pinellolab/scATAC-
753 benchmarking/blob/master/Real_Data/Buenrostro_2018/input/combined.sorted.merged.bed. For cPeaks, we
754 removed the housekeeping cPeaks and extra-long cPeaks (> 2000 bp). In the benchmark study, cells were grouped
755 by their FACS sorting labels. Each group was used to call peaks separately. The called peaks for all groups were
756 merged together as features. For cPeaks, we converted cPeaks from the hg38 version to the hg19 version using
757 CrossMap[78] and used them as features. For cDHSs, we downloaded them from
758 https://zenodo.org/records/3838751/files/DHS_Index_and_Vocabulary_hg19_WM20190703.txt.gz?download=1,
759 this file is hg19 version. For CATALS cCREs, we downloaded them from
760 http://catlas.org/catlas_downloads/humantissues/cCRE_hg38.tsv.gz and converted cCREs from the hg38 version to
761 the hg19 version using CrossMap.

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

765 **Embedding methods for cell representations**

766 **ChromVAR-motif**[46]: We used the ChromVAR package (version 1.16.0) in R to perform embedding. After adding
767 GC bias and resizing features to 500 bp, we used the "getJasparMotifs" function to get motifs from the JASPAR
768 database and matched them to features via the "matchMotifs" function. Then we mapped sequencing reads to motif
769 features and get the cell-by-motif matrix. We transformed this matrix to the cell-by-deviation Z-score matrix using
770 the "computeDevations" function.

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

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

778 **ChromVAR-kmer-PCA**: The generation of the cell-by-deviation Z-score matrix was the same as the ChromVAR-
779 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[8]: We used cisTopic package (version 0.3.0) in R to perform embedding. 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[60]: 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[85]: We first created a Seurat object using Seurat package (v4.3.0) 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.

SnapATAC2[73]: We used cPeaks as the feature set to get a cell-by-cPeaks matrix by "snapatac2.pp.make_peak_matrix" function in snapATAC2 package. We reduced the dimensions to 30 through "snapatac2.tl.spectral" and generated a neighbor graph through "snapatac2.pp.knn". By adjusting the resolution parameter in "snapatac2.tl.leiden", we got 10 types of cells.

TF-IDF-PCA[21,35,86]: 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.

ArchR[63]**:** We used cPeaks as the feature set to get a cell-by-cPeaks matrix by "addFeatureMatrix" function in ArchR. We reduced the dimensions by "addIterativeLSI" function. we got 10 types of cells by "addClusters" function.

We set different iteration times (1,2,4) by "addIterativeLSI" function.

Evaluation of cell representations

After embedding, we performed Louvain clustering[61] 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[85] 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[5]. 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.

846

847 ## Dynamic biological process and diseases data analysis

848 ### Pseudotime calculation of human hematopoiesis lineages

849 We downloaded the fragment files of the human hematopoiesis data[10] from GEO under accession code GSE129785.
850 We generated the cell-by-feature matrix using cPeaks. We followed the same preprocessing, clustering, and low-
851 dimensional visualization steps as the analysis of the PBMC dataset. We used Monocle 3[67,68] to build trajectories,
852 following the steps in the Signac trajectory vignettes (https://stuartlab.org/signac/articles/monocle.html). We built
853 trajectories separately for erythroid and lymphoid lineages using the "learn_graph" function, setting hematopoietic
854 stem cells as the root of the trajectory. We then ordered cells with estimated pseudotimes.

855 ### Processing and embedding of the human fetal development dataset

856 We downloaded the fragment files and metadata of the retinal fetal data[33] from GEO under accession code
857 GSE184386. We followed the preprocessing steps in the original study to filter cells. The generation of the cell-by-
858 feature matrix, clustering and visualizations were the same as the analysis of the PBMC dataset. We labeled each
859 cell using labels from the metadata file. All steps were done separately for each sample.

860 ### Processing and embedding of the kidney dataset

861 We downloaded the fragment files of the kidney dataset[18] from GEO under accession code GSE195460. We
862 download the metadata from https://github.com/p4rkerw/Wilson_Muto_NComm_2022. We selected the healthy
863 sample 6 and all the disease samples with more than 2,000 cells for further analysis. We followed the preprocessing
864 steps in the original study to filter cells. The generation of the cell-by-feature matrix, clustering and visualizations
865 were the same as the analysis of the PBMC dataset. We labeled each cell using labels from the metadata file. All
866 steps were done separately for each sample.

867 ### Processing and embedding of the gynecologic malignancy dataset

868 We downloaded the fragment files and metadata of the gynecologic malignancy dataset[34]from GEO under accession
869 code GSE173682. We combined all cells from 8 patients together to generate the cell-by-feature matrix. We
870 followed the preprocessing steps in the original study to filter cells. The generation of the cell-by-feature matrix,
871 clustering and visualizations were the same as the analysis of the PBMC dataset. We labeled each cell using labels
872 from the metadata file.

873

874 ## Build human chromatin accessibility atlas with cPeaks

875 ### Building human cell atlas

876 We imported 92 fragment files from the CATLAS[5] dataset into snapATAC2. Using the barcodes in the metadata
877 download from http://catlas.org/catlas_downloads/humantissues/fragment/, we got the 615,998 filtered cells. We

878 performed batch correction on these data using "snapatac2.pp.mnc_correct" function. We used the
879 "snapatac2.pp.select_features" function to select 500,000 features. We reduced the dimensions to 30 through
880 "snapatac2.tl.spectral" function and generated a neighbor graph through snapatac2.pp.knn. We use
881 "snapatac2.tl.umap" function to show the UMAP visualization of the cell atlas. By adjusting the resolution
882 parameter in "snapatac2.tl.leiden" function, we got 30 or 111 clusters of cells.

883 **Calculating clustering metrics on different populations of cells**

884 We calculated the proportion of each of the 111 cell types marked in the CATLAS[5] paper in the total number of
885 cells. Depending on the ratio (<0.2%, >0.2% & <0.6%, >0.6%), we categorized cell types into three parts: extremely
886 rare cell types, rare cell types, and common cell types. In the dataset, these three categories comprise 26,238 cells,
887 84,660 cells, and 505,100 cells, accounting for 4.3%, 13.7%, and 82.0% respectively. Using the 111 cell type label
888 as ground truth, we separately calculated AMI, ARI, and Homogeneity (H).

889

890 # Data availability

891 We only used public datasets in this study. The information of bulk ATAC-seq data for cPeak generation can be
892 found in Table S1. The scATAC-seq and ATAC-seq data for cPeaks evaluation can be found in Table S3. The data
893 used for cPeak annotation can be found in Table S4. We downloaded the bam files of the hematological
894 differentiation scATAC-seq dataset from https://github.com/pinellolab/scATAC-
895 benchmarking/tree/master/Real_Data/Buenrostro_2018/input/sc-bams_nodup. We downloaded the PBMC dataset
896 from the 10x Genomics website https://cf.10xgenomics.com/samples/cell-
897 atac/1.0.1/atac_v1_pbmc_10k/atac_v1_pbmc_10k_fragments.tsv.gz. We downloaded the human hematopoiesis
898 dataset from Gene Expression Omnibus (GEO) under accession code GSE129785. We downloaded the fetal retina
899 dataset from GEO under accession code GSE184386. We downloaded the kidney dataset from GEO under accession
900 code GSE195460. We downloaded the gynecologic malignancy dataset from GEO under accession code
901 GSE173682. We downloaded the fragment file and metadata of human cell atlas data from
902 http://catlas.org/catlas_downloads/humantissues/.

903 cPeaks can be downloaded from the link: https://mengqiuchen.github.io/cPeaks/Tutorials. It provides cPeak BED
904 files, containing hg38/hg19 versions, and CSV files which provide detailed information.

905

906 # Code availability

907 All Code is available on GitHub from the link https://github.com/MengQiuchen/cPeaks, including the generation of
908 observed cPeaks, training of the deep CNN model, all applications on scATAC-seq data, and mapping and
909 conversion tools for analyzing diverse data types using cPeaks. We wrote a detailed tutorial on how to use cPeaks at
910 the link https://mengqiuchen.github.io/cPeaks/Tutorials, including integration with existing scATAC-seq analysis

tools, such as SnapATAC2[73] and ArchR[63]. We also developed a Python tool to map fragment files or peak files to cPeaks.

Acknowledgments

This work was partially supported by National Key R&D Program of China (grants 2021YFF1200901 & 2021YFF1200902), National Natural Science Foundation of China (NSFC) (grants 62250005 and 62103227), and Tsinghua-Fuzhou Institute for Data Technology (TFIDT2021005). We acknowledge the helpful discussions of Drs. Wei Xie, Xun Lan and Yinqing Li.

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. Y. Z. contributed to the deep CNN network, cPeak resources construction. W.C. contributed to the mapping tools and tutorial website. Y. Z., W. C. and C.L. tested and improved the mapping tools. Z.W., J.L. (first) and X. Wang assisted in designing the deep CNN network. X.X., Si.C., Catherine.Z., Sh.C., J.L. (second), X.L., W.X. 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. Tsompana, M. & Buck, M. J. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin* **7**, 33 (2014).

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

4. 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.

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

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

7. 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).

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

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

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

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

12. 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.

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

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

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

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

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

18. 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).

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

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

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

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

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

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

978 25. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*
979 **361**, 1380–1385 (2018).

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

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

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

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

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

987 31. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**,
988 1728–1740 (2012).

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

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

993 34. Regner, M. J. *et al.* A multi-omic single-cell landscape of human gynecologic malignancies. *Mol. Cell* **81**, 4924-
994 4941.e10 (2021).

995 35. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309-
996 1324.e18 (2018).

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

999 37. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251
1000 (2020).

1001 38. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

1002 39. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).

1003 40. Nair, V. D. *et al.* Differential analysis of chromatin accessibility and gene expression profiles identifies cis-
1004 regulatory elements in rat adipose and muscle. *Genomics* **113**, 3827–3841 (2021).

1005 41. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data
1006 analysis. *Genome Biol.* **21**, 22 (2020).

1007 42. Daugherty, A. C. *et al.* Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans. *Genome*
1008 *Res.* **27**, 2096–2107 (2017).

1009 43. Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer.
1010 *Nature* **481**, 389–393 (2012).

1011 44. Giansanti, V., Tang, M. & Cittaro, D. Fast analysis of scATAC-seq data using a predefined set of genomic regions.
1012 *F1000Research* **9**, 199 (2020).

1013 45. Wang, F. *et al.* ATACdb: a comprehensive human chromatin accessibility database. *Nucleic Acids Res.* **49**, D55–
1014 D64 (2021).

1015 46. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated
1016 accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).

1017 47. 10x Genomics. 10k_pbmc_ATACv2_nextgem_Chromium_X - Human PBMCs.
1018 https://cf.10xgenomics.com/samples/cell-
1019 atac/2.1.0/10k_pbmc_ATACv2_nextgem_Chromium_X/10k_pbmc_ATACv2_nextgem_Chromium_X_web_su
1020 mmary.html.

1021 48. 10x Genomics. 10k_pbmc_ATACv2_nextgem_Chromium_Controller - Human PBMCs.
1022 https://cf.10xgenomics.com/samples/cell-
1023 atac/2.1.0/10k_pbmc_ATACv2_nextgem_Chromium_Controller/10k_pbmc_ATACv2_nextgem_Chromium_Co
1024 ntroller_web_summary.html.

1025 49. Zou, Z., Ohta, T., Miura, F. & Oki, S. ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic
1026 landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. *Nucleic Acids Res.* **50**, W175–W182
1027 (2022).

1028 50. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).

1029 51. Loza, M., Vandenbon, A. & Nakai, K. Epigenetic characterization of housekeeping core promoters and their
1030 importance in tumor suppression. *Nucleic Acids Res.* **52**, 1107–1119 (2024).

1031 52. Jiang, C. & Pugh, B. F. Nucleosome positioning and gene regulation: advances through genomics. *Nat. Rev. Genet.*
1032 **10**, 161–172 (2009).

1033 53. Struhl, K. & Segal, E. Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.* **20**, 267–273 (2013).

1034 54. Peng, Y. *et al.* Detection of new pioneer transcription factors as cell-type specific nucleosome binders. *eLife* **12**,
1035 (2023).

1036 55. Grossman, S. R. *et al.* Positional specificity of different transcription factor classes within enhancers. *Proc. Natl.*
1037 *Acad. Sci.* **115**, E7222–E7230 (2018).

1038 56. Fernandez Garcia, M. *et al.* Structural Features of Transcription Factors Associating with Nucleosome Binding.
1039 *Mol. Cell* **75**, 921-932.e6 (2019).

1040 57. Yuan, H. & Kelley, D. R. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural
1041 networks. *Nat. Methods* **19**, 1088–1096 (2022).

1042 58. Wei, Z. *et al. NeuronMotif: Deciphering Transcriptional Cis-Regulatory Codes from Deep Neural Networks*.
1043 http://biorxiv.org/lookup/doi/10.1101/2021.02.10.430606 (2021) doi:10.1101/2021.02.10.430606.

1044 59. Fang, R. *et al.* Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* **12**, 1337
1045 (2021).

1046 60. Baker, S. M., Rogerson, C., Hayes, A., Sharrocks, A. D. & Rattray, M. Classifying cells with Scasat, a single-cell
1047 ATAC-seq analysis tool. *Nucleic Acids Res.* **47**, e10–e10 (2019).

1048 61. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks.
1049 *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).

1050 62. Chen, H. *et al.* Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.*
1051 **20**, 241 (2019).

1052 63. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis.
1053 *Nat. Genet.* **53**, 403–411 (2021).

1054 64. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

1055 65. Reizis, B. Regulation of plasmacytoid dendritic cell development. *Curr. Opin. Immunol.* **22**, 206–211 (2010).

1056 66. Ghosh, H. S., Cisse, B., Bunin, A., Lewis, K. L. & Reizis, B. Continuous Expression of the Transcription Factor
1057 E2-2 Maintains the Cell Fate of Mature Plasmacytoid Dendritic Cells. *Immunity* **33**, 905–916 (2010).

1058 67. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of
1059 single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

1060 68. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

1061 69. Muto, Y. *et al.* Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in
1062 the adult human kidney. *Nat. Commun.* **12**, 2190 (2021).

1063 70. Chèneby, J., Gheorghe, M., Artufel, M., Mathelier, A. & Ballester, B. ReMap 2018: an updated atlas of regulatory
1064 regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* **46**, D267–D275
1065 (2018).

1066 71. Hammal, F., de Langen, P., Bergon, A., Lopez, F. & Ballester, B. ReMap 2022: a database of Human, Mouse,
1067 Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing
1068 experiments. *Nucleic Acids Res.* **50**, D316–D325 (2022).

1069 72. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**,
1070 699–710 (2020).

1071 73. Zhang, K., Zemke, N. R., Armand, E. J. & Ren, B. A fast, scalable and versatile tool for analysis of single-cell
1072 omics data. *Nat. Methods* **21**, 217–227 (2024).

1073 74. Bian, H. *et al.* scMulan: a multitask generative pre-trained language model for single-cell analysis.
1074 2024.01.25.577152 Preprint at https://doi.org/10.1101/2024.01.25.577152 (2024).

1075 75. Hao, M. *et al. Large Scale Foundation Model on Single-Cell Transcriptomics*.
1076 http://biorxiv.org/lookup/doi/10.1101/2023.05.29.542705 (2023) doi:10.1101/2023.05.29.542705.

1077 76. Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat.*
1078 *Methods* 1–11 (2024) doi:10.1038/s41592-024-02201-0.

1079 77. Gong, J. *et al.* xTrimoGene: An Efficient and Scalable Representation Learner for Single-Cell RNA-Seq Data. *Adv.*
1080 *Neural Inf. Process. Syst.* **36**, 69391–69403 (2023).

1081 78. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*
1082 **30**, 1006–1007 (2014).

1083 79. Hartigan, J. A. & Hartigan, P. M. The Dip Test of Unimodality. *Ann. Stat.* **13**, (1985).

1084 80. Malzer, C. & Baum, M. A Hybrid Approach To Hierarchical Density-based Cluster Selection. in *2020 IEEE*
1085 *International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)* 223–228 (2020).
1086 doi:10.1109/MFI49285.2020.9235263.

1087 81. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison
1088 and visualization. *Bioinforma. Oxf. Engl.* **31**, 2382–2383 (2015).

1089 82. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory
1090 Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576–589 (2010).

1091 83. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

1092 84. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**,
1093 1017–1018 (2011).

1094 85. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac.
1095 *Nat. Methods* **18**, 1333–1341 (2021).

1096 86. Cusanovich, D. A. *et al.* The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature*
1097 **555**, 538–542 (2018).

1098 87. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and
1099 functional annotation. *Nucleic Acids Res.* **44**, D733-745 (2016).

1100
1101

Supplementary materials

Supplementary Figures

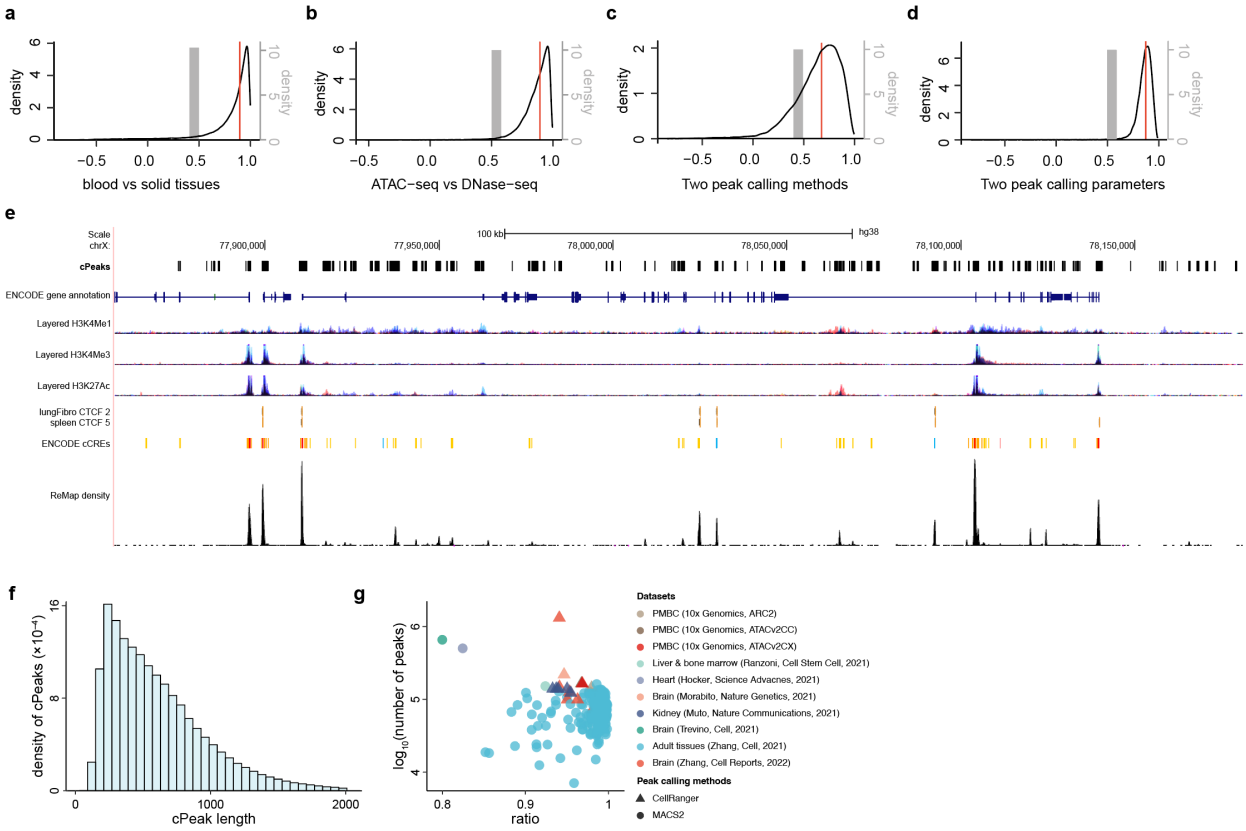


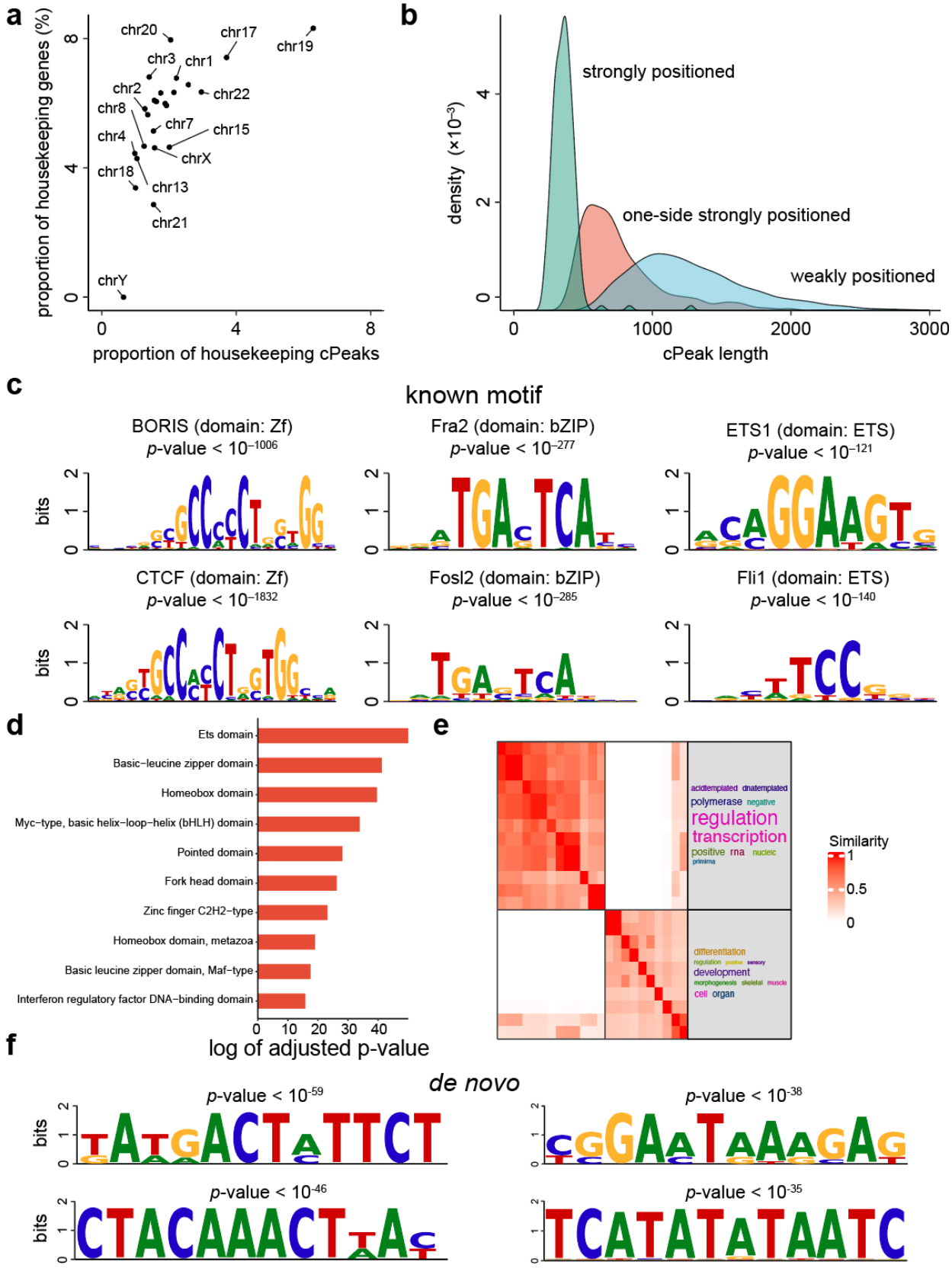
Fig. S1. Details of observed cPeaks

a-d. PCCs of merged peak shapes under different conditions. The red line represents the median of PCC. The gray bars represent the distribution of the median PCCs on 100 permutations.

e. 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[38], layered H3K4Me1 marker[38], layered H3K4Me3 marker[38], layered H3K27Ac marker[38], lung fibroblast CTCF signal[38], spleen CTCF signal[38], ENCODE cCREs[38] and ReMap density[70]. 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[70].

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

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


a
proportion of housekeeping genes (%)
proportion of housekeeping cPeaks
chr20
chr17
chr19
chr3
chr1
chr2
chr22
chr8
chr7
chr15
chr4
chrX
chr18
chr13
chr21
chrY
b
density (×10⁻³)
strongly positioned
one-side strongly positioned
weakly positioned
cPeak length
c
known motif
BORIS (domain: Zf)
p-value < 10⁻¹⁰⁰⁶
Fra2 (domain: bZIP)
p-value < 10⁻²⁷⁷
ETS1 (domain: ETS)
p-value < 10⁻¹²¹
CTCF (domain: Zf)
p-value < 10⁻¹⁸³²
Fosl2 (domain: bZIP)
p-value < 10⁻²⁸⁵
Fli1 (domain: ETS)
p-value < 10⁻¹⁴⁰
bits
d
Ets domain
Basic−leucine zipper domain
Homeobox domain
Myc−type, basic helix−loop−helix (bHLH) domain
Pointed domain
Fork head domain
Zinc finger C2H2−type
Homeobox domain, metazoa
Basic leucine zipper domain, Maf−type
Interferon regulatory factor DNA−binding domain
log of adjusted p-value
e
Similarity
regulation
transcription
polymerase
differentiation
development
cell
organ
f
de novo
p-value < 10⁻⁵⁹
p-value < 10⁻³⁸
p-value < 10⁻⁴⁶
p-value < 10⁻³⁵

1121 **Fig. S2. Properties of observed cPeaks**

1122 a. The proportion of housekeeping cPeaks and housekeeping genes on each chromosome. Each dot
1123 represented a chromosome. For each chromosome, the proportion was calculated by the number of
1124 housekeeping cPeaks/genes divided by the total number of cPeaks/genes on the chromosome. The PCC
1125 between the proportion of housekeeping cPeaks and housekeeping genes is 0.64.

1126 b. The length distribution of cPeaks with different shape patterns. cPeaks with both edges showing fitting
1127 MSEs less than 0.03 were used to illustrate the distribution (see Methods for details).

1128 c. Examples of enriched known motifs found around the well-positioned edges of cPeaks.

1129 d. The top 10 domains enriched around the well-positioned edges of cPeaks. The top 10 domains with the
1130 most significant adjusted P-value were shown.

1131 e. The heatmap of enriched GO terms around well-positioned edges of cPeaks. The row and column of
1132 heatmap represents the enriched GO terms, while the value of heatmap represents term-term similarity.

1133 f. Examples of enriched *de novo* motifs found around the well-positioned edges of cPeaks.

1134

1135

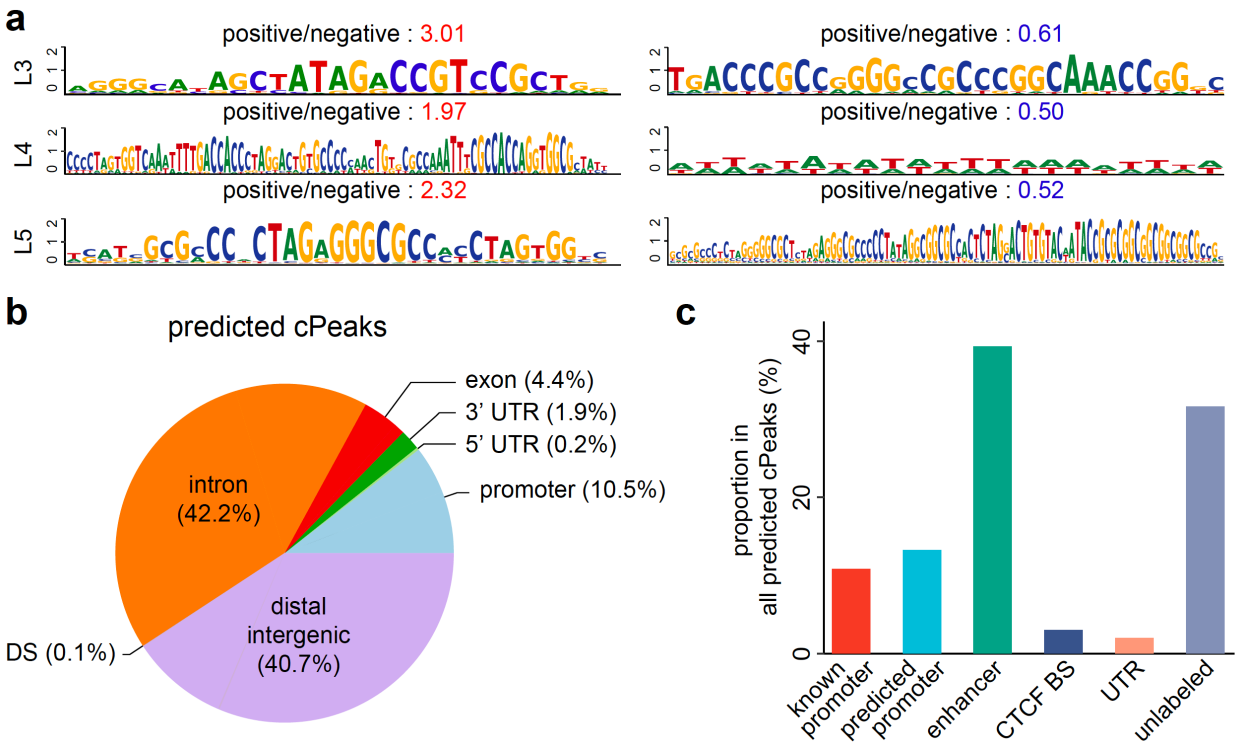


Fig. S3. Properties of predicted cPeaks

a. Examples of motifs captured from the convolution kernel in CNN model. The title represents the ratio of current motif number in positive and negative samples. L3, L4, and L5 represent layer 3, layer 4, and layer 5 respectively.

b. Genomic annotation of predicted 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.

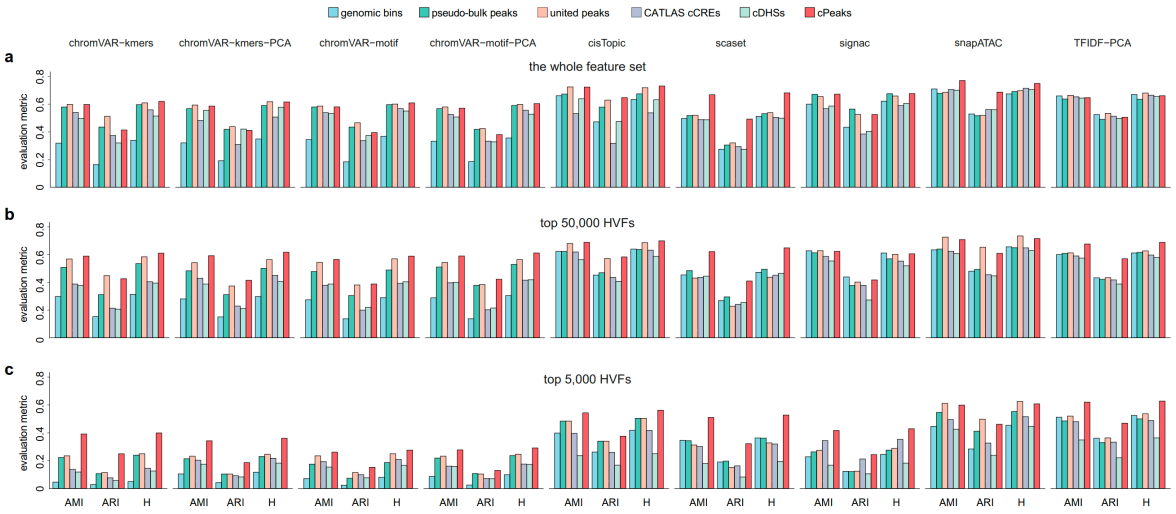


Fig. S4. The performance of cell annotations on FACS2-sorted hematopoietic differentiation data

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.

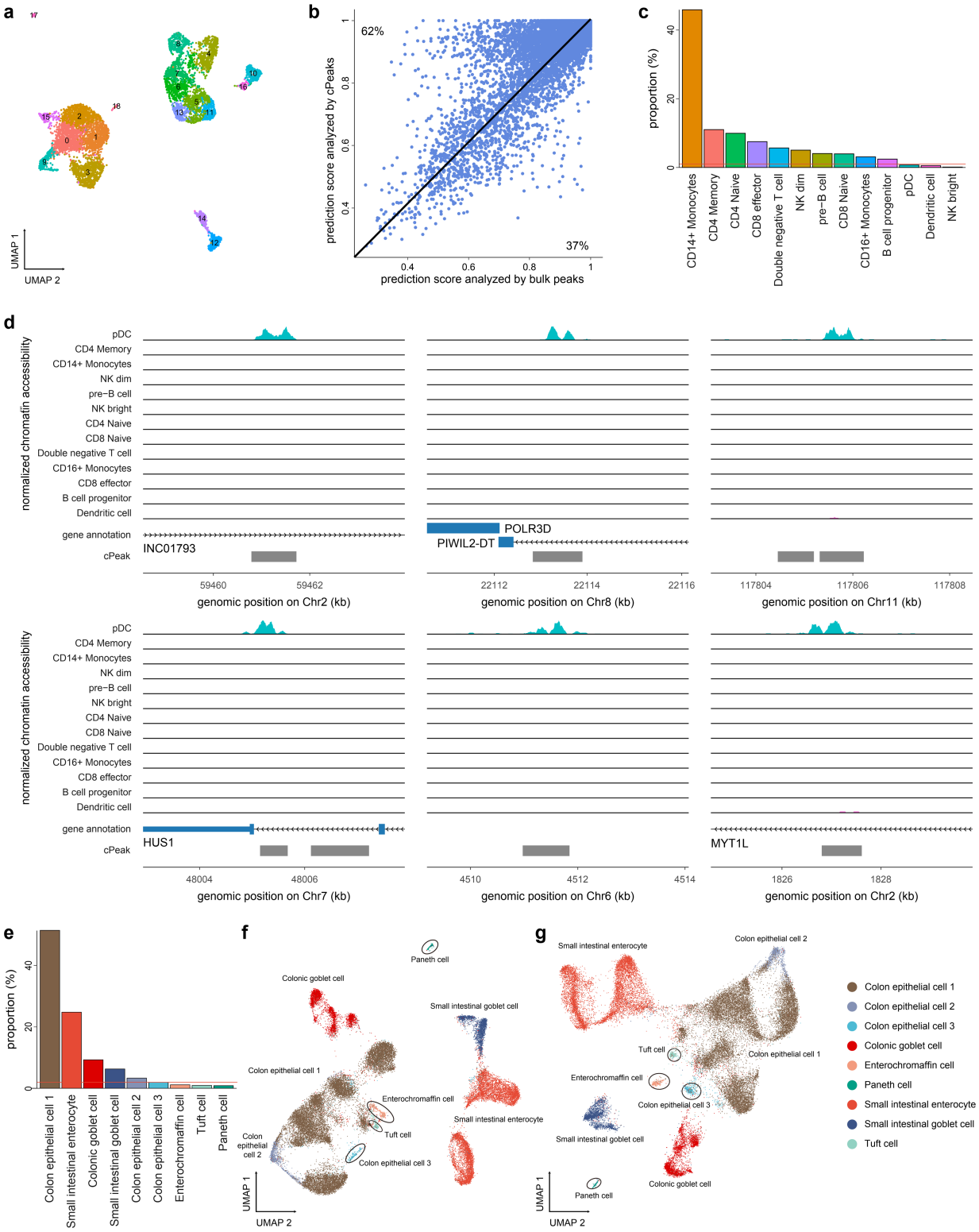


Fig. S5. Rare cell type discovery using cPeaks

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

1158 Different colors represented different clusters.

1159 b. Prediction scores of cells in the PBMC scATAC-seq data calculated by cPeaks and pseudo-bulk peaks.
1160 Each dot represented one cell.

1161 c. The proportion of each cell type in the PBMC scATAC-seq data. Three rare cell types were less than 1%
1162 (the red line): pDCs, dendritic cells and NK bright cells.

1163 d. Genomic tracks of six differentially accessible regions only found by cPeaks, shown in the hg19 reference
1164 genome. We selected six regions with the lowest *p*-values. Genes were annotated by RefSeq[87].

1165 e. The proportion of each subtype in GI epithelial cells. Four rare subtypes were less than 2% (the red line):
1166 paneth cell, enterochromaffin cell, tuft cell and colon epithelial cell 3.

1167 f. The UMAP plot of subtypes in GI epithelial cells using merged peaks in CATLAS. Each dot represented
1168 one cell. Each color represented a cell type provided by the original study[5]. The black lines labeled the four
1169 rare subtypes.

1170 g. The UMAP plot of subtypes in GI epithelial cells using pseudo-bulk peaks called by the population of all
1171 GI epithelial cells. Each dot represented one cell. Each color represented a cell type provided by the
1172 original paper[5]. The black lines labeled the four rare subtypes.
1173

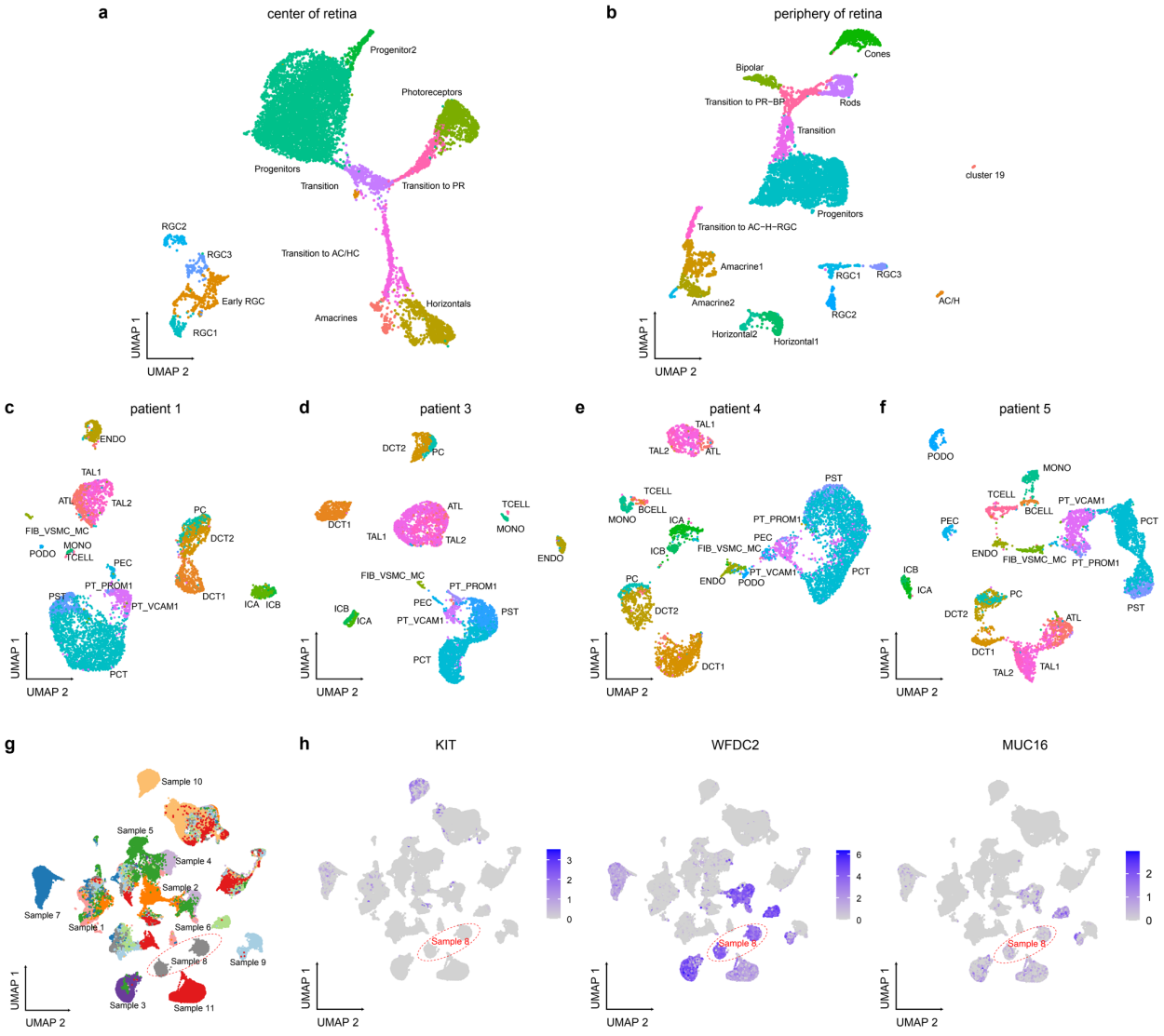


Fig. S6. Analyzing dynamic biological processes and diseases by cPeaks

a. The UMAP plot of the 13-week scATAC-seq data collected from the center of retina colored by cell types. PR, [5]; 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.

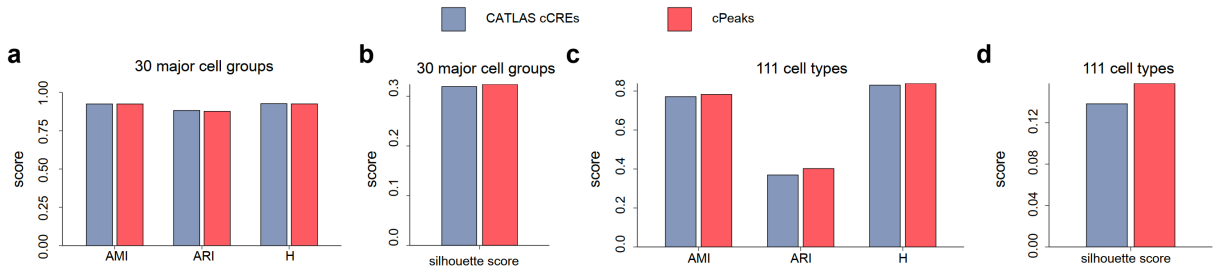


Fig. S7. The performance of cell clustering in human cell atlas

a. The evaluation metric of clustering results using 30 major cell groups as labels.

b. The silhouette score of clustering results using 30 major cell groups as labels.

c. The evaluation metric of clustering results using 111 cell types as labels.

d. The silhouette score of clustering results using 111 cell types as labels.

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.