CamoTSS: analysis of alternative transcription start sites for cellular phenotypes and regulatory patterns from 5' scRNA-seq data

Ruiyan Hou[1], Chung-Chau Hon[2,3], Yuanhua Huang[1,4,5*]

1 School of Biomedical Sciences, University of Hong Kong, Hong Kong SAR, China
2 RIKEN Center for Integrative Medical Sciences, Yokohama City, Kanagawa 230-0045, Japan
3 Graduate School of Integrated Sciences for Life, Hiroshima University, Higashi-Hiroshima, Japan
4 Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong SAR, China
5 Center for Translational Stem Cell Biology, Hong Kong Science and Technology Park, Hong Kong SAR, China

*** To whom correspondence should be addressed.**

Abstract

Five-prime single-cell RNA-seq (scRNA-seq) has been widely employed to profile cellular transcriptomes, however, its power of analysing transcription start sites (TSS) has not been fully utilised. Here, we present a computational method suite, CamoTSS, to precisely identify TSS and quantify its expression by leveraging the cDNA on read 1, which enables effective detection of alternative TSS usage. With various experimental data sets, we have demonstrated that CamoTSS can accurately identify TSS and the detected alternative TSS usages showed strong specificity in different biological processes, including cell types across human organs, the development of human thymus, and cancer conditions. As evidenced in nasopharyngeal cancer, alternative TSS usage can also reveal regulatory patterns including systematic TSS dysregulations.

Introduction

Alternative usage of different gene architectures enables to differential expression of various mRNA isoforms, including alternative transcription start/end sites and alternative splicing (AS) events, such as exon skipping, intron retention, alternative 5' and 3' splice sites [1–3]. The advances of single-cell transcriptomic technologies have provided a powerful tool to detect cellular heterogeneity in gene-level expression by using the sum of all transcripts originating from the same gene [4]. Several studies have developed computational and statistical approaches to detect and quantify alternative splicing at single-cell resolution [5]. Most of these studies focus on the exon-skipping event, commonly by full-length based platforms like Smart-seq2 [6–8] and also possibly by UMI-based methods like 10X Genomics Chromium [9, 10]. Thanks to the higher throughput, 3' tag-based scRNA-seq in 10x Chromium platform has been broadly adopted to explore gene-level expression. Several groups leveraged the technique characteristic of polyA-biased scRNA-seq and developed computational pipelines to detect alternative 3' end usage, even with potential applicability to detect alternative 5' start site usages [11, 12]. In addition, another recent method, *scraps*, took the advantage of using read 1 (>100bp) to precisely identify polyadenylation sites at a near-nucleotide resolution in scRNA-seq data by 10X Genomics and other TVN-primed libraries [13].

Alternative transcript start site (TSS) usage is another major mechanism to increase transcriptome diversity and its regulation. Cap analysis gene expression (CAGE) has been widely used to capture the 5'-end of transcripts and identify TSS at a single-nucleotide resolution from bulk samples [14], which has been used as the major tool to annotate TSS across mammal genomes in the FANTOM project [15] and to reveal narrow shifts of TSS within a single promoter during zebrafish early embryonic development [16]. The analysis of TSS and its alternative usage has been further fueled by the extensive use of RNA-Seq in multiple international consortium projects, including tissue-specific TSS by the Genotype-Tissue Expression (GTEx) data [1], the cell type specific novel TSS by the RAMPAGE project [17] and cancer type specific promoter regulations from a pan-cancer study [18]. More individual studies also evidenced the importance of TSS regulation for different biological functions, e.g., tumor immune interaction in gastric cancer [19], prognosis in multiple myeloma [20], and synchronized cell-fate transitions in the yeast gametogenesis program [21].

Recently, attention has also been paid to TSS analysis at a single-cell level, for example, a single-cell version of CAGE, *C1 CAGE*, was introduced to identify TSS and enhancer activity with the original sample multiplexing strategy in the C1TM microfluidic system [22]. This was further extended to capture both 5' and 3' by the single-cell RNA Cap And Tail sequencing (scRCAT-seq) method, where UMI became applicable to further reduce the cost [23]. Besides these specialized methods, conventional platforms, e.g., 10x Genomics have commercial kits for constructing a 5' gene expression library (often with V(D)J dual readouts), where the fragmentation does not happen at sequences close to template switch oligo (TSO), suggesting that this part of sequences is an ideal material to detect transcription start site at a (near-) nucleotide resolution (Fig. 1A). Many sequencing centres keep its default setting of equal length paired-end (e.g., 150bp), hence capturing the cDNA in read 1. Indeed, some public 5' 10x Genomics datasets on the GEO repository have such information, bringing an open rich resource to re-explore the TSS usage in various biological contexts at the single-cell level. A software suite called SCAFE has already used this type of sc-end5-seq data to de novo detect TSS at a single-cell resolution, but it mainly paid attention to the cis-regulatory elements (CRE, the proxy of the TSS) rather than the alternative transcription start sites [24]. Therefore, there is an urgent demand for tailored methods to analyse such 5' scRNA-seq both efficiently and accurately, especially on alternative TSS usage.

Here, to identify and quantify potential TSSs and evaluate their differential usages from 5' tag-based scRNA-seq, we fully utilize those "rubbish" sequences in read 1 mentioned above (otherwise trimmed before analysis) and developed CamoTSS (Cap and Motif-based TSS modelling from 5' scRNA-seq data), a computational method suite that calls TSSs by combining clustering of reads distribution and classification with predictive features, followed by window sliding technique to denoise the detection of single-nucleotide-resolution TSS. CamoTSS further focused on the analysis of alternative TSS among different cell populations, tissues, development stages and disease contexts by leveraging our upgraded BRIE2 as a backend engine. The effective application of our method was demonstrated by using public 5' scRNA-seq data containing pair-ends (i.e. read 1 covering information of cDNA) including adult human cell atlas of 15 major organs, primary nasopharyngeal carcinoma and hyperplastic lymphoid tissue, and human thymic cell across development and postnatal life, where alternative usages of TSSs were found with strong specificity on cellular states and showed potential to assess their systematic dysregulation.

Of note, CamoTSS is capable of analysing TSS both at a region (around 100 bp) and a single-nucleotide resolution. For the former, we interchangeably use TSS region, TSS cluster or simply TSS, otherwise, we will specifically call the latter CTSS (CAGE tag-defined transcription start site, a concept borrowed from CAGE [16]). We primarily focus on TSS region/cluster analysis and only introduce the CTSS in the analysis of the thymus development data, considering its biological relevance.

Results

Overview of CamoTSS pipeline

We developed a stepwise computational method CamoTSS to detect alternative transcription start site clusters and quantify their differential usage utilizing 5' tag-based scRNA-seq data alone (Fig.1B). In brief, CamoTSS has three main steps after fetching TSS reads for a certain gene from an aligned bam file: 1) clustering of TSS reads with hierarchical clustering (minimum linkage distance: 100 bp by default), 2) filtering TSS clusters by technical thresholds (minimum UMIs: 50; minimum inter-cluster distance: 300bp) and an embedded classifier to prevent TSS clusters from artefacts by using predictive features (see next paragraph) and 3) annotating these de-novo TSSs (by its summit position) to known annotations (e.g., GENCODE) optimised by a Hungarian algorithm while if a detected TSS cluster does not cover the optimal known TSS position, it remains called as new-TC or novel-TSS (Fig. 1B; Methods).

Due to strand invasion [25] and sequence biases [26], the classification is a critical step to rule out false positives caused by technical artefacts (Fig. 1C). By using ATAC-seq data from a matched sample as ground truth, we labelled the intersecting TSSs captured from 5' scRNA-seq data as true TSS if it overlaps with a high-confidence ATAC peak or false TSS if it overlaps with a low-confidence ATAC peak. First, we found that the four reads-based features (unencoded G percentage, standard deviation, summit count and cluster count) introduced in SCAFE [24] are highly predictive through a logistic regression on both pluripotent stem cells (iPSC) and human dermal fibroblasts (DMFB) lines (10-fold cross-validation; area under the receiver operating characteristic curve, AUC=0.975; Fig. 1D). Further, we introduced a convolutional neural network model (architecture detail in Methods) to examine how well the pure genomic sequence (+/- 100bp) predicts the TSS. Impressively, we found they are also highly predictive (AUC=0.96; same cross-validation split) and can further improve if combined with the four read-based features, no matter with separately-trained or jointly-trained CNN models (AUC=0.983 or 0.986, respectively; Fig. 1D and Supplementary Fig. S1A; Methods). Of note, the utility of the sequence-based component is not only to enhance prediction performance but can also imply regulatory strength (see the NPC section). The high prediction accuracy and improved performance were also observed in each cell line separately (Supplementary Fig. S1B-C) or trained from one sample to predict another (from iPSC to DMFB; AUC=0.985 with the combined features model; Fig. 1D), which proves the reliability of using a pre-trained model across samples. For deployment and user usage, we keep the four reads-based features as the default setting with a pre-trained logistic regression in iPSC and DMFB data, considering its reasonable accuracy, high generalizability and remarkable simplicity.

Furthermore, the quality of detected TSS clusters can be assessed by the consistency with epigenetic signals, including promoter-specific histone modifications (H3K4me3 and H3K27ac), gene-body-specific markers (H3K36me3) [27] and transcription initiation signal (RNA POL2 that collects general transcription factors to form the pre-initiation complex) [28]. In the PBMC dataset, the RNA POL2, H3K4me3 and H3K37ac signals were all highly enriched in the intervals around CamoTSS-identified TSS clusters, while the signal of H3K36me3 was mainly enriched downstream of TSS clusters detected by CamoTSS (Fig. 1E). We also randomly generated the same number of fake-TSS regions with the same sequence length as a negative control, and found no signal surrounding these negative TSS clusters. As an example, Fig. 1F shows one canonical and one novel TSSs of PTPN4 and two canonical TSSs of SCP2 detected by CamoTSS (highlighted by red lines), all with signal support from histone modifications, RNA POL2 and scATAC-seq, which suggests high reliability of CamoTSS. Then we calculated the percentages of annotated and novel TSS clusters detected by CamoTSS in PBMC (Fig. 1G) and it shows the majority of identified TSS clusters (68.7%) are annotated. As we expected, most TSSs (3,205 out of 6148) detected by CamoTSS were mapped to 5'UTR, while a substantial part of TSSs was also mapped to intron (1,804) and exon (1,134) regions, presumably due to the alternative usage of transcription start sites.

CamoTSS facilitates cell identity analysis

Given that CamoTSS has the capability of detecting TSSs at a single-cell level, we wonder how much it can enhance cell identity analysis and to which extent it presents a cell-type specificity. Here, we downloaded 84,363 single-cell transcriptomes (profiled by 10x Genomics, 5' tagged scRNA-seq with reads 1) across 15 organs from one adult donor to detect the variability of TSS usage across cell types and organs [29]. By leveraging CamoTSS, we can obtain matched gene expression and TSS expression for each individual cell, allowing for direct comparison of RNA and TSS clusters without the need for integration methods. We took muscle (5,732 cells) as an example and used the same parameters to preprocess the gene and TSS expression matrices and clustered the cells (Supplementary Fig.S2). Overall, the majority of cells showed consistent clustering between using gene- or TSS-level expressions (Adjusted Rand index: 0.572; Supplementary Fig.S3). Interestingly, we noticed that NK/T cells have different clustering results when using TSS- or gene-level expressions as input (Fig. 2A). The separation at TSS profiling is more consistent with annotation from the original paper in muscle (ARI=0.793 vs 0.218; Fig. 2B). From another perspective, the R7 cluster (identified at the gene level) exhibits two distinct TSS profiles (TSS clusters S4 and part of S9) that are highly concordant with the two subpopulations named GZMK T cell (CD8+ T cell) and IL7R T cell (CD4+ T cell) annotated by the original paper (ARI=0.425; Fig. 2C) [29]. This indicates that TSS data provides complementary information to identify cell clusters at a higher resolution. We speculate the extra information comes from the distinct promoter usage, which links to the regulatory variability between cells that may be evident at the epigenetic level. To validate our hypothesis, we leveraged single-cell regulatory network inference and clustering (SCENIC) analyzing gene expression data to obtain the activities of transcription factors. Surprisingly, R7 displays different regulatory patterns between S4 and S9, for example, IKZF1, a regulator of lymphocyte differentiation (Supplementary Fig. S4A). Otherwise, the clear differential pattern of R7 disappeared when we shuffled the order of cells (Supplementary Fig. S4B), which suggests the TSS-level analysis can capture distinct regulation status that is masked by gene-level analysis. Furthermore, when comparing the transcription factors binding possibility to the top 500 cluster-specific TSS regions by using Homer, we found that S4 and S9 showed enrichment of different transcription factor family motifs, namely nuclear receptors (NR) and basic leucine zipper (bZIP) respectively, as illustrated in Fig. 2D.

Subsequently, we surveyed the TSS profiles in all 15 organs and asked if the TSS profiles are more similar within cell types or organs. In total, 21,125 TSSs were detected by merging all cells from 15 organs and then used to calculate a Pearson's correlation coefficient between each cell type in each organ (at a pseudo-bulk level). By performing the

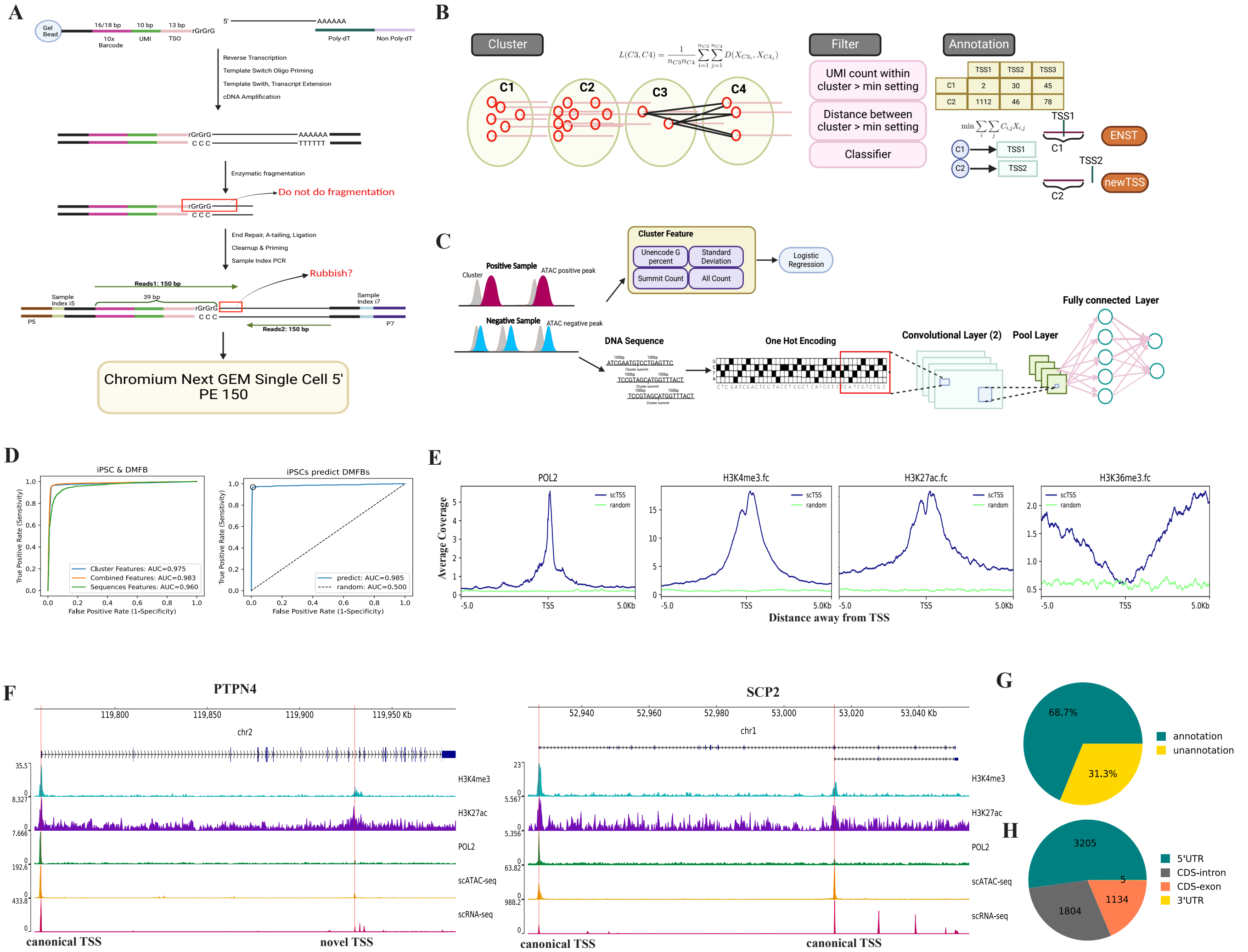


Figure 1. Developing CamoTSS to identify transcription start site (TSS) from 5' tag-based scRNA-seq data. **(A)** A flow chart of the 5' scRNA-seq gene expression library construction (10x Genomics). **(B)** A schematic of CamoTSS which includes clustering, filtering and annotation. **(C)** Classifier embedding in CamoTSS includes a logistic regression model and a convolutional neural network model. Ranked ATAC-seq peaks were used as ground truth labels for the TSS clusters when training classifiers. **(D)** Receiver operating characteristic (ROC) curves for TSS classification with three groups of features by using logistic regressions; the curves are for pooled non-redundant TSSs form iPSC and DMFB datasets (Left panel; Methods; individual sample shown in Supp. Fig. S1). ROC curves showing using iPSC data as train dataset and DMFB as test dataset (Right panel). Ten-fold cross-validation is used for the evaluation. **(E)** The distributions of RNA POL2, H3K4me3, H3K27ac and H3K36me3 signals around the TSSs detected by us and the random regions produced by bedtools. RNA POL2, H3K4me3 and H3K27ac show enrichment around TSSs while H3K36me3 is enriched downstream of our TSSs. **(F)** Tracks plots of two examples (PTPN4 and SCP2) show peaks of scRNA-seq, scATC-seq, POL2, H3K27ac and H3K4me3. Red lines denote the location of our detected TSSs. **(G)** Pie chart of the percentage of our detected TSS regions/clusters as annotated by reference genome or novel TSSs. **(H)** Genomic distribution of the detected TSS regions.

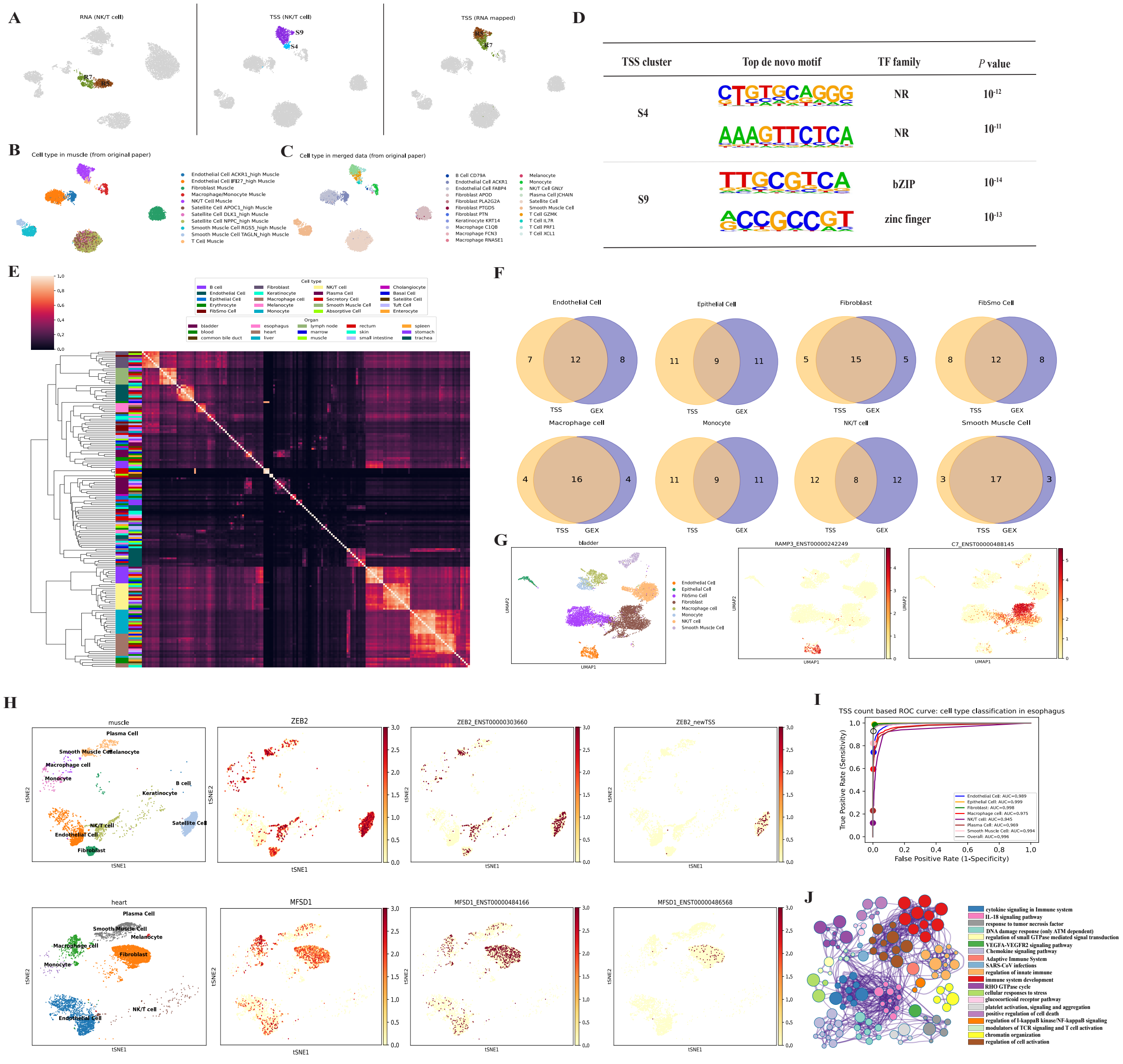
hierarchical clustering on the correlation, the dendrogram shows that samples of the same cell type (across organs) have a higher similarity, with few exceptions (Fig. 2E). This “cell type-dominated clustering” pattern implies that most cell types possess a conserved TSS expression signature. To further illustrate the usage of the TSS profile at cell type identification, we compared the top 20 most significant markers at both TSS and gene levels for each cell type in the bladder (Fig. 2F). Venn diagrams show partial overlap between gene- and TSS-based markers in all cell types, indicating that TSS may serve as additional predictive features for cell type identification, for example, RAMP3_ENST00000242249 as a marker for endothelial cells and C7_ENST00000488145 for fibroblast (Fig. 2G).

Next, we focused on genes which have at least two TSSs. To identify cell-type differential expression at the TSS level, we performed a differential analysis between original annotated cell types and searched for genes with both cell-type-specific TSS shifts (one cell type vs each of others) and non-cell-type-specific TSS (one cell type vs any of others; Methods). We detected 2,301 genes containing such isoform markers in 15 organs (Supplementary Table 1, Supplementary Fig. S5). Fig. 2H top panel shows an example of such TSS from the zinc finger E-box binding homeobox 2 (ZEB2) gene in muscle, which is known to be a transcription factor to regulate epithelial to mesenchymal transition associated with many cancers [30]. While the gene-level expression of ZEB2 is less distinct across cell types, we find that among the 2 TSSs detected by CamoTSS, one TSS without annotation (with minor expression), is almost exclusively expressed in satellite cells. The histone modification, RNA POL2 and scATAC-seq signals all appeared at the same position of TSS (Supplementary Fig. S6), which indicates the reliability of TSS identified by us. As another example, MFSD1 plays an essential role in liver homeostasis as a lysosomal transporter [31]. It undergoes an isoform shift in fibroblast in the heart, where the expression of the ENST00000486568 is significantly higher, suggesting distinct cell-type-specific TSS localization (Fig. 2H bottom panel, Supplementary Fig. S6). To further determine the role of TSS as a cell type marker, we used all of these TSS-level cell-type markers (cell number > 50) to predict cell type annotated by gene expression profile from the original report, and achieved accurate predictions on all cell types in the esophagus (Fig. 2I). We were, then, curious about how many TSS markers of cell type are shared by other organs. To solve this problem, we used the upsetR package [32] to detect the intersection of genes including TSS markers among organs and found there are still some TSS markers overlapping across multiple organs (Supplementary Fig. S7). SH3KBP1 shared the same TSS switch among 7 distinct organs in NK/T cell (Supplementary Fig. S8, Supplementary Fig. S9), indicating the generalizability of TSS as a cell type marker. Next, we explore the biological function of the TSS markers of each cell type (Fig. 2J, Supplementary Fig.S10). Notably, the functions of signature TSS of NK/T cells are mainly enriched in terms related to various immune response processes, including immune system development, cellular responses to stress, regulation of cell activation and regulation of innate immune.

Altered TSS usages in patients with nasopharyngeal carcinoma

Although alternative promoter usage has been found to be cancer type-specific and predictive of patient prognosis via bulk RNA-Seq [18], it remains largely unexplored whether and to which extent the precise TSS usage at a single-cell resolution can further explain the heterogeneity in the cancer microenvironment. To explore this problem, we applied CamoTSS to a nasopharynx dataset (5’ scRNA-seq, 10x Genomics) from patients with either nasopharyngeal carcinoma (NPC; n=7 patients) or nasopharyngeal lymphatic hyperplasia (NLH; n=3 patients), covering 51,001 cells in total (Fig. 3A-B), downloaded from a recent study [33]. Then BRIE2 was run on the 1,784 genes with at least two TSSs across the whole dataset, among which 547 genes were found with significant differential alternative TSS usage between NPC and NLH (FDR$<$ 0.01; Supplementary Table 2). Take T cell as an example (Fig. 3C), among the genes that show alternative-TSS activation in NPC, multiple of them are well-known cancer-related biomarkers such as QSOX1 serving as a prognosis biomarker in breast cancer [34], DTNBP1 relating with memory and executive functions in brain tumor [35], and FAM107B associated with gastric cancer [36]. Particularly, two TSSs of CCND3 have been reported to generate mRNAs with distinct 5’ transcript leaders, resulting in protein isoforms with different N-termini [37] (Fig. 3C).

Additionally, cell type-specific differential TSS usages between cancer status were also identified, for example, SIDT2 is only detected in B cells (Supplementary Fig. S11), which aligns well with a recent report on the cell-type-specific genetic effects to its isoform expression involving TSS changes [38]. Another prominent example is LIMS1, which contains three major cell-type specific TSSs, including TSS1 (LIMS1-215) as a leading TSS for B cells, TSS2 (LIMS1-201) for Myeloid cells and TSS3 (LIMS1-210) for T cells (Fig. 3D). Interestingly if focusing on the proportion of the top two TSSs in T cells (TSS1 and TSS3), we further found that proportion of the minor TSS (TSS1, LIMS1-215) shows a significant up-regulation in cancer condition (NPC) compared to NHL (Fig. 3E), consistent with the trend of expressed cell proportions across patients (Supplementary Fig.12 SA-B). Surprisingly, Myeloid cells show an opposite trend, where the proportion of the



TSS cluster	Top de novo motif	TF family	*P* value
S4	CTGTGCAGGG	NR	10^{-12}
	AAAGTTCTCA	NR	10^{-11}
S9	TTGCGTCA	bZIP	10^{-14}
	ACCGCCGT	zinc finger	10^{-13}

Figure 2. CamoTSS analysis on TSSs between cell types across 15 human organs. **(A)** UMAP projection of RNA profile (left) and TSS profile (middle and right) in muscle. The T cell cluster is highlighted in the colors. All the other cells are colored gray. **(B)** Cell type annotation from the original paper's single organ (muscle). **(C)** Cell type annotation from the original paper's merged annotation (15 organs) with a finer resolution. **(D)** The top de novo motifs enriched in the top 500 cluster-specific peaks of S4 and S9. P values were calculated by using binomial tests. **(E)** Heatmap of Pearson's correlation of expression of common TSS among all cells from all organs. **(F)** Venn diagrams of top 20 significant TSS markers and RNA expression markers in 8 cell clusters of the bladder. **(G)** tSNE plots of TSS data specific marker. **(H)** tSNE plots show alternative TSS marker masked at gene level. **(I)** ROC curves for prediction of cell types from the first 20 PCs of TSS matrix by using a random forest in a multi-label classification. Models were evaluated by using 10-fold cross-validation, whereby the overall average is obtained by merging all cell types at a micro level. **(J)** Enrichment network representing the top 20 enriched terms of significant alternative TSS. The enriched terms that displayed high similarity were grouped together and presented as a network diagram. In this diagram, each node corresponds to an enriched term and is assigned a color based on its cluster. The size of each node reflects the number of enriched genes, while the thickness of the lines connecting nodes represents the similarity score between the enriched terms.

same minor TSS (TSS1) decreases in NPC (Fig. 3F and Supplementary Fig.S12C-D), demonstrating the complexity of TSS regulation and its coupled modulation of cell types and disease conditions. To further understand the potential functions associated with NLH and NPC-specific TSSs at different cell types, we examined GO terms enriched in the TSS-shift gene sets between NLH and NPC for each cell type (Fig. 3G, Supplementary Fig. S13). Specifically, in T cells, the genes with differential TSS usage show enrichment in GO terms related to hemopoiesis, regulation of leukocyte activation, pathway in cancer and negative regulation of apoptotic signalling pathway, suggesting that an abundant TSS-mediated diversity is required for these genes associated with fundamental immune and cancer response properties.

To inform the potential regulatory mechanism leading to the alternative usage of cancer-related TSS, we used FIMO (v4.11.2) to find validated motifs in JASPAR in the NLH- and NPC-elevated TSS sequences. Take CTCF as an example, it can bind three motif patterns in the JASPAR database, including MA0139.1, MA1929.1 and MA1930.1. As Fig. 3H and Supplementary Fig. S14 show, the sequence logo of CTCF binding sequences detected by FIMO is highly similar to that downloaded from the JASPAR database, which confirms the reliability of our method. Then we counted the frequency of the database-curated motifs occurring in the two sets of TSSs (n=528 for T cells and n=556 for B cells) that are elevated in NLH or NPC. Interestingly, more binding sites were consistently observed in the NLH group in both T cells (Fig. 3I) and B cells (Supplementary Fig. S15A), which suggests a global change of transcription factor activities in this cancer. Such systematic bias is not likely caused by random chance, as no obvious difference is observed if the two groups of TSSs (n=528) are randomly selected from the expressed TSSs in the NPC group (Supplementary Fig. S15B).

To explore the underlying reason, the significant differential TFs were counted according to their classes. More than half (55.7%) of the differential TFs belong to C2H2 zinc finger factor class, while this proportion is only one-fifth when counting all TFs (Supplementary Fig. S16). This phenomenon is particularly striking for the top 10 most significant differential TFs, which are all in the C2H2 zinc finger factor class (Supplementary Fig. S17). All of these indicate the C2H2 zinc finger factor class play an essential role in the regulation of alternative TSS in this cancer, which has been reported in a previous study [39]. In addition, we also compared motif patterns bound by the top 10 significant TF with random motifs and discovered top 10 motifs have higher GC percentages, which is consistent with the binding pattern of C2H2 zinc finger factor class (Supplementary Fig. S17).

On top of that, to explain the inclined trend between NLH and NPC, we extracted 200bp (+/- 100bp) sequence around significant TSS start of NLH and NPC separately as two groups test data and exploit our pre-trained convolutional neural network to predict the probability of being a positive sample. As Supplementary Fig. S18 shows, more TSSs in the NLH group were predicted as positive (i.e. the probability is larger than 0.9) compared with the NPC group (n=239 vs 175) and fewer TSSs in the NLH group were predicted as negative sample (i.e. the probability is smaller than 0.1) comparing with NPC group (n=175 vs 229). The difference is significant (P=4.9e-5, Fisher exact test), which indicates the sequence strengths around cancer-specific TSSs become weaker compared to normal samples. To confirm the role of these significant differential TSS in NPC, we explored the function of these TFs and discovered a highly related GO term "transcriptional misregulation in cancer" (Supplementary Fig. S19). We next asked if the expression profile of these TFs aligns with their binding frequency. Taking the critical and well-studied CTCF as an example, though the expression level of expressed cells is similar between these two groups, the proportion of expressed cells is significantly higher in NLH (Fig. 3J, Supplementary Fig. S20), consistent with the report from another group [40]. Also, the whole expression profiles of all significant TFs show a consistent trend with the binding frequency changes, and an unbiased clustering analysis of them well segregated NLH and NPC samples (Fig. 3K).

Transcription start site shifting during human thymic development

Even though scRNA-seq is extensively used to study human development, the actual mechanism of choosing TSS in different development stages in various cell types remains unknown. To explore developmental-stage-specific TSS usage, we analysed a profile of transcription initiation events in three development time points of human thymus, including 11-week and 12-week in prenatal and 30-month in postnatal at distinct cell types (Fig. 4A-B). Next, we focused on genes with multiple TSSs and detected their TSS shift across time points at the single cell level respectively between W11 vs W12, W11 vs M30 and W12 vs M30 (FDR<0.01; Fig. 4C, Supplementary Table3). Pairwise comparisons between time points show that similar genes with differential TSS-shifting patterns appear in 11-week Vs 30-week and 12-week Vs 30-week in T cells, which suggests that the closer the development distance, the more similar alternative TSS usage patterns are. The same trend was also observed in other cell types (Supplementary Fig. S21, Supplementary Table 3). We then summarised genes which are widely significantly differential in all three time points (Fig. 4D, Supplementary Table 4) or just have TSS-shift in a certain development time point (Supplementary Fig. S22, Supplementary Table 4).

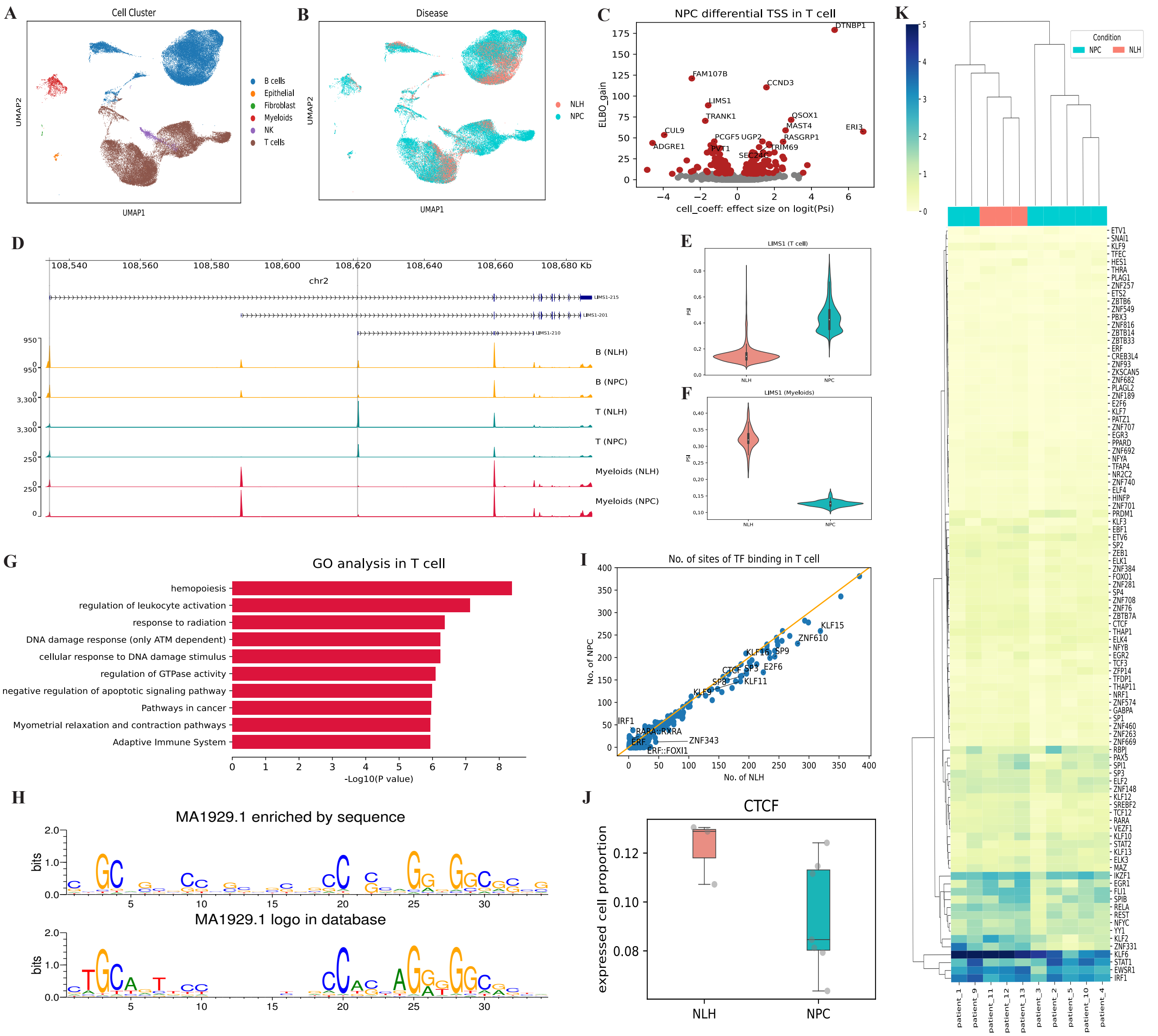


Figure 3. CamoTSS identifies differential alternative TSS usage from nasopharyngeal carcinoma **(A,B)** UMAP plot of gene-level expression, annotated with cell types (A) and disease status (B). **(C)** Volcano plot to show the relationship between ELBO_gain and effect size on logit(PSI) for detecting differential TSS between NPC and NLH patients. **(D)** Genome track plot of LIST1 in different cell types of NLH and NPC patients. **(E,F)** Violin plot on example gene LIST1 for T cell(E) and Myeloids (F) in NLH and NPC patients. The y-axis PSI denotes the proportion of TSS1 (LIST1-215; minor TSS here) among the top two TSSs in each cell type. **(G)** Bar plot showing the enriched terms of genes with differential TSS usage between NLH and NPC patients in the T cell. **(H)** WebLogo of the base frequency of MA1929.1 (i.e. one motif of CTCF) enriched in the sequences detected by FIMO (top) and displayed in the JASPAR database (bottom). **(I)** Scatter plot of the binding frequency of human TFs on 528 TSS regions elevated in NLH and NPC patients (shown is based on T cells). **(J)** Box plot of expressed cell proportion of CTCF between NLH and NPC. **(K)** Heatmap shows the hierarchical clustering of gene expression of significant TFs with differential binding frequency across 10 patients.

For those triple differential TSSs (across all time point pairs), the dynamics patterns can be either monotonic (26 for TSS1 increasing and 44 for TSS1 decreasing) or transient (45 for TSS1 upregulation first and 28 for TSS1 downregulation first; Fig. 4D, Supplementary Table5). Of note, the increase or decrease is relative, depending on the definition of TSS1 and TSS2. Notably, most genes containing TSS shift during three stages are well studied in immune development and promoter research such as RAC2, playing dual roles in neutrophil motility and active retention in zebrafish hematopoietic tissue [41] and SLC3A2, helping branched-chain amino acids (BCAAs) to control Regulatory T cell maintenance [42]. In addition, many genes with differential TSS usage only between two certain stages also play critical roles in development such as FCHO1, involved in T-cell development and function in humans [43] and ST6GAL1 which can enhance B cell development and produce IgG in a CD22-dependent manner in vivo [44] To decipher the function of genes with alternative TSS, we performed GO enrichment analysis and found enriched GO terms are highly relevant to cell cycle, development and stress response (Fig. 4E).

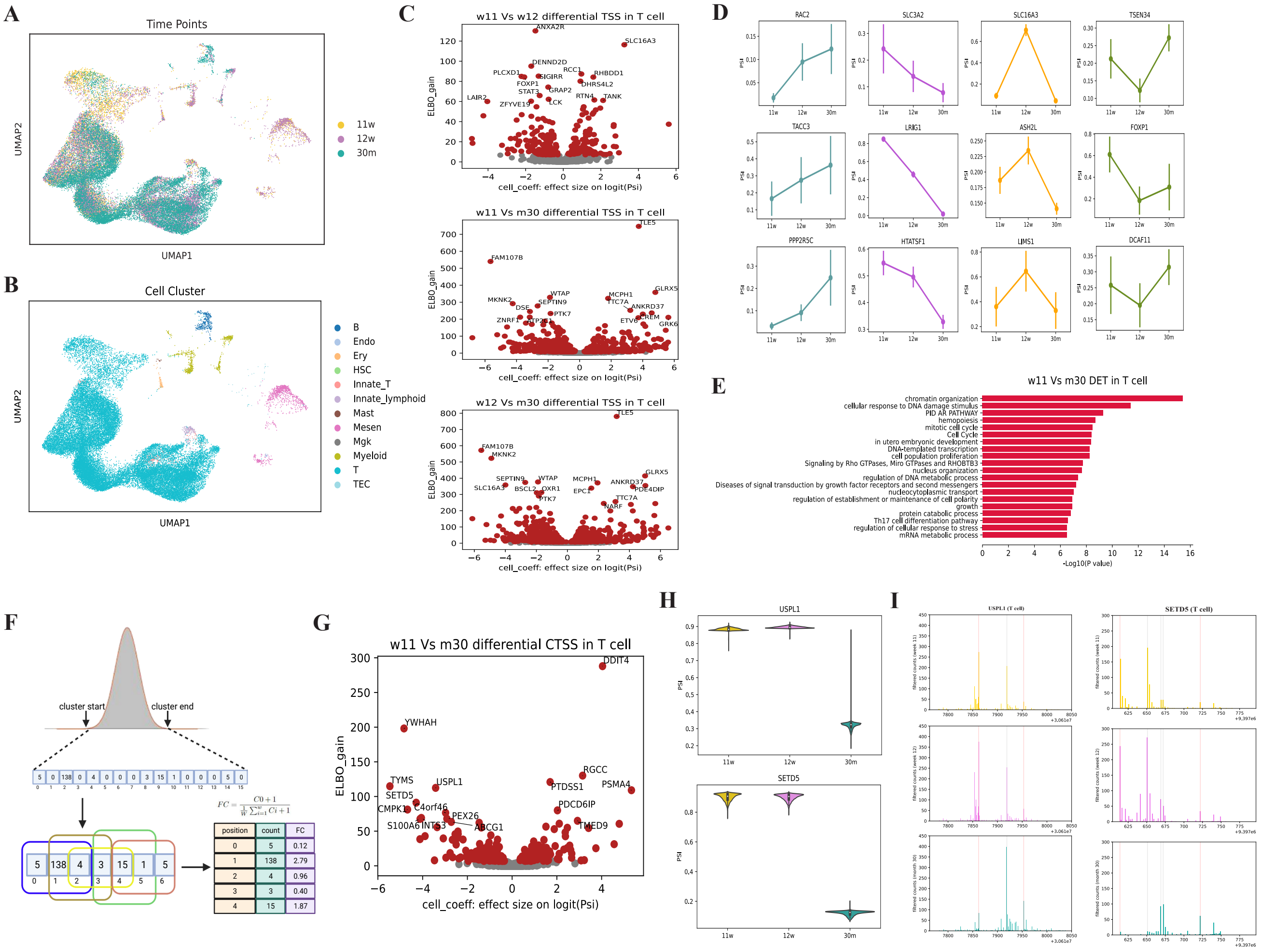
Next, inspired by the work from Haberle and colleagues [16], we also aimed to discover narrow shifts within one TSS region (i.e., cluster) during thymic development. The *bona fide* TSS clusters detected in the first step were selected and then we utilized a sliding-window approach to denoise and get reliable “CTSS” (CAGE-based TSS, i.e., at a single-nucleotide resolution) in one TSS cluster (Fig. 4F; Methods). To detect these narrow shifts within one TSS cluster (usually within 100bp), we first selected two farthest CTSSs within one cluster and then exploited BIRE2 to detect alternative CTSS usage during the three time points in each cell type. In total, we found 1263 genes with significant CTSS shifts between any two development points (163 between W11 and W12, 451 between W11 vs M30, 649 between W12 vs M30; FDR<0.01; Fig. 4G and Supplementary Fig. S24), with higher overlap between the two prenatal stages vs the postnatal stage. In other words, these significant TSS regions contain CTSS shifting from one end to another during thymus development, for example, USPL1 and SETD5 (Fig. 4H-I), both of which were reported as important genes for development in zebrafish [45] or mammals [46]. To unveil the potential functions of these genes with narrow shifting CTSSs between each pair of time points, we performed GO terms enrichment and found that recurrent terms included cell cycle, mitotic and translation (Supplementary Fig. S25), suggesting that the alternative CTSS usage within one TSS region may play an essential role in thymus development.

Discussion

In this work, we present CamoTSS, a computational method for de novo detecting TSS from 5' tag-based scRNA-seq data. This method enjoys a data-driven design and embeds a classifier to accurately detect TSSs, and enables efficient identification of alternative TSS usage between single-cell populations by seamlessly leveraging our BRIE2 model. Specifically, this method first adopts a hierarchical clustering algorithm to determine the potential TSS clusters, followed by filtering the false positive clusters mainly via an embedded classifier from reads and/or sequence-based features. Finally, it annotates those genuine TSS clusters with known transcript annotations (e.g., from a GENCODE GTF file) by a Hungarian algorithm and counts the UMIs of each TSS at a single-cell level. Generally, the four sequence-based features with a logistic regression model provide high accuracy, hence are used in practice by default. On the other hand, our convolutional neural network module, by extracting sequence features from the query TSS, can achieve comparable performance and further improve the accuracy when combined with the reads-based features. Additionally, as shown in the NPC data, this sequence-based model can further reveal the weakened TSS patterns in a disease condition.

While CamoTSS can identify all TSS clusters, we focused on the analysis of alternative TSS usage in this study, covering broad biological scenarios including cell types across 15 human organs, cancer conditions from multiple samples and thymus development with three time points. Differential TSS usage in different cell types was observed in multiple organs, where TSS clusters provide additional molecular signatures otherwise masked by gene expression and help to identify cell clusters with higher purity. Importantly, compared to NLH individuals, differential TSS usage, especially a general preference toward weaker promoter, was detected in all major nasopharynx cell types of NPC patients, which may be regulated by TFs from C2H2 zinc finger factor class. In addition, we also found hundreds of genes with TSS-shift during thymus development stages and many of them have narrow shifts within 100bp in a cell type-specific manner. Taken together, the TSS-level information, especially the alternative TSS usage, can provide more detailed cellular phenotypes and may imply regulatory patterns across various biological contexts. Considering the almost free access to the TSS from the 5' scRNA-seq data, CamoTSS may introduce a new paradigm in analysing such data and resolving the cellular heterogeneity in a finer-grained resolution with better interpretation from the regulatory perspective.

Additionally, there are also open challenges in the TSS analysis. First, when we performed alternative TSS usage, we



$$FC = \frac{C0+1}{\frac{1}{W}\sum_{i=1}^{w} Ci+1}$$

position	count	FC
0	5	0.12
1	138	2.79
2	4	0.96
3	3	0.40
4	15	1.87

Figure 4. CamoTSS identifies differential alternative TSS usage from human thymus development. **(A,B)** UMAP visualization of all cells (35629) in thymus development. Each dot is one cell, with colors coded according to the time points (A) and cell types (B). **(C)** Volcano plot showing the relationship between ELBO_gain and effect size on logit(PSI) for detecting differential TSS between week11 and week12 (Top), week11 and month30 (Middle), week12 and month30 (Bottom). **(D)** Line chart showing four patterns of example genes which are all significant at three development stage pairs. **(E)** Bar plot showing the enriched terms of genes with alternative TSS usage between week 11 and month 30 in the T cell. **(F)** Illustration of the window sliding algorithm for identifying CTSS within one TSS cluster. Count and fold change parameters were used to filter noise. **(G)** Volcano plot between ELBO_gain and effect size on logit(PSI) for detecting differential CTSS between week11 and month30 in T cell. **(H)** Violin plot of PSI value of USPL1 and SETD5 among week11, week12 and month30. The two farthest CTSSs were picked up to calculate PSI for each gene. **(I)** Histogram showing the coverage of reads 1 with unencoded G at the cap obtained from 5' scRNA-seq in USPL1 (Left) and SETD5 (Right). The grey and red lines represent CTSSs identified by CamoTSS, while the red line shows the two farthest CTSS used for differential CTSS analysis with BRIE2.

mainly support the analysis of proportional change of the two major TSSs among cell populations. However, there can be more than two TSSs playing critical roles, especially when ranging from different cell types and diseases in one setting. Therefore, an extended model with support to jointly analyse multiple TSSs would account for such compositional change. Second, a large fraction of public 5' scRNA-seq datasets, e.g., on GEO, may only contain cDNA on read 2 for various reasons. Therefore, extending our CamoTSS to support a read2-only mode with high accuracy can be another timely contribution to data mining from the exponentially increasing single-cell data across the community. For addressing this potential challenge, our sequence-based neural network model would play a more important role in specifying the TSS region. Third, considering that the cell-by-TSS UMI count matrix can serve as a more informative input compared to the conventional cell-by-gene count matrix, it remains to be further examined if the downstream analysis pipeline needs to be adapted, especially when integrating with data from other platforms, e.g., 3' scRNA-seq data.

Materials and Methods

scRNA-seq initial data analysis

The raw fasta files including reads1 (>100bp) and reads2 were downloaded, and then sequences were aligned to the *Homo sapiens* reference genome (hg38) to generate pair-end read alignment bam file by using the cellranger count pipeline (with parameter `--chemistry SC5P-PE`) of 10x Genomics CellRanger (v3.1) software. The possorted_genome_bam.bam file was manually filtered by using `xf:i:25` tag before performing reads counting, same strategy according to the 10x Genomics criteria [47]. In most instances, we aggregated the bam file from each donor sample by using an in-house script (https://github.com/StatBiomed/CamoTSS). In brief, it adds sample ID to the cell barcode by using *pysam* package [48] and then merges all bam files by using *samtools merge* [48].

Construction of CamoTSS method

CamoTSS is a stepwise computational method to identify TSS clusters and it includes three major steps: to cluster reads into TSS clusters/regions, to detect true clusters, and to annotate *bona fide* TSS clusters.

Step1: Cluster for reads start site

Preprocessed bam file was used as input to perform clustering for subsequent promoter evaluation. We fetch all reads 1 for each gene by using BRIE [6]. All obtained reads 1 then filtered according to the cell barcode list specified by users. We remove strand invasion artifacts by aligning the DNA sequence starting from the -14 base and ending at the aligned start sit of reads1 to the TSO sequence (5'-TTTCTTATATGGG-3'). The read is regarded as a strand invader when the edit distance is less than 3. The edit distance was calculated by utilizing *editdistance* python package. The reads were also filtered according to the *SCAFE* criteria to precisely calculate the number of unencoded G [24]. In brief, we require reads 1 that 1) should contain the last 5nt of TS oligo (i.e. ATGGG) and the edit distance is less than 4, 2) start with a softclip region (i.e. "S" in CIGAR string) and the value of "S" is more than 6 and less than 20, 3) the match region following the softclip region is more than 5bp. If the number of fetched reads is more than 10,000, we randomly selected 10,000 reads from all reads in that gene to make sure the efficiency of our software.

The start position of reads was extracted and then input to agglomerative clustering which is a kind of from-bottom-to-up hierarchical clustering method. In brief, it first determines the proximity matrix containing the Euclidean distance between each start position using a distance function. Then, this matrix is updated to display the distance between each cluster. Here, we use the average linkage method to define the distance between two clusters, which can be calculated by 1.

$$L(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} D(X_{A_i}, X_{B_j}). \tag{1}$$

If the linkage distance between two clusters is at or above 100bp, then the two clusters will not be merged.

Step2: Filter false positive cluster

Although we discarded some aberrant reads which are strand invasion artifacts and affect the counting of unencoded G, quite a lot of clusters still can be observed located at the end of the gene body via integrative genomics viewer (IGV). We added another filtering step to efficiently identify high-confidence transcription start site clusters. The total reads counts within one cluster are set to be more than 50 UMIs and the distance between two neighboring clusters is more than 300bp. In addition, We designed and tested three models to distinguish high-confidence TSS from false positive clusters, and compared them with the overlapped peaks from ATAC-seq to select the most suitable model to embed into our pipeline.

1. The first model is "logistic regression" with four reads-based features (next paragraph). We collected the ATAC-seq and scRNA-seq data of human dermal fibroblasts (DMFB) and induced pluripotent stem cells (iPSC) from one batch experiment of a laboratory [24]. For ATAC-seq data, the bigwig file was downloaded and transformed to a bedGraph file by using UCSC Genome Browser's *bigWigToBedGraph* tool. Then *liftOver* was utilized to convert the bedGraph coordinates from hg19 to hg38. Peaks from ATAC-seq were ranked based on the p-value and the top and bottom 5% peaks were defined as ground-truth positive and negative peaks, respectively. Then the bam files of DMFB and iPSC from scRNA-seq were input to our software CamoTSS to detect clusters without filtering with a classifier. These clusters were defined to gold positive (n=5,560) and negative samples (n=5,432) by using bedtools (v2.26.0) [49] to intersect (parameter: `-f 0.1`) with ATAC-based positive and negative peaks, followed by removal of double-detected TSS regions if combining multiple datasets.

Then we extracted four features including cluster count, summit count, the standard deviation, and unencoded G percentage for each cluster. The cluster count refers to the total UMI counts within one cluster and the summit count is the maximum UMI count for a certain position within the cluster. The standard deviation was calculated by *statistics* python package to measure the dispersion of the cluster (treating each UMI as a sample). An intact mRNA with the cap structure can reverse transcribe to cDNA possessing an additional dGMP and cDNA with an extra dGMP cannot be produced by cap-free RNA [50]. This evidence suggests unencoded G percentage within one cluster can be regarded as an essential feature to identify the transcription start site. Because of the number uncertainty of extra added dGMP, reads whose CIGAR string starts with "14S", "15S" and "16S" are all considered as reads with unencoded G. Finally, total samples (n=10,992) with four properties were used to train logistic regression model at 10-fold cross-validation by implementing *Scikit-learn* python package (v1.0.2). We choose 0.5 as the default threshold and classify samples with a probability greater than 0.5 as a true transcription start site cluster.

2. The second model is a "convolutional neural network". The same dataset (n=10,992) used in the logistic regression model was also used to train this deep learning model. We utilized bedtools (v2.26.0) : `getfasta` to extract 200bp sequence shifting around the cluster summit position. These sequences were transformed to numbers between 0 and 3 representing the 4 possible nucleotides and were then one-hot encoded to provide a categorical representation of nucleotide in numerical space to train this neural network (i.e. A: [1,0,0,0], T: [0,1,0,0], C: [0,0,1,0], G: [0,0,0,1]). The convolutional neural network was constructed with PyTorch (v1.12.1) and it consists of two convolution layers connecting to Rectified Linear Unit (ReLU) for activation, followed by batch normalization, one max pooling and a dropout layer (probability of dropout: 0.4). The output from dropout layer was flattened and fed to fully connected layers (32 neurons) with a ReLU activation function. Then the second fully connected layers of 2 neurons were connected with the sigmoid function to calculate the probability of classification. The first convolutional layer has 128 filters with 8-mer width and 4 channels. The second convolutional layer has 64 filters, where the filter size is 4x4. This model can be summarized as follow,

$$\overline{O}_i = f^{Sigmoid} f^{Linear} f^{Linear\_ReLU} f^{Flatten} f^{Dropout} f^{Maxpooling} f^{BatchNorm} f^{Conv\_ReLU} f^{conv\_ReLU} (\overline{X}_i) \tag{2}$$

where $\overline{X}_i$ denotes the one-hot encoded matrix (4, 200). All 10,992 One-Hot-Encoded matrices were split into a training set, test set and validation set according to the 6:2:2 ratio. Then this model was trained with a batch size of 256 and 500 epochs with SGD optimizer with a learning rate of 0.003 and momentum of 0.8. The model with the lowest validation loss during training was kept at last.

3. The third model is the "combination of logistic regression and convolutional neural network". The 32 dimensions features of the first fully connected layer were saved and combined with the four cluster features mentioned above as input features (36 dimensions) to the logistic regression. The same dataset was used to train logistic regression at 10-fold cross-validation. The difference between this model and the first model is the features used to train logistic regression changing from 4 to 36 (i.e., 4+32). Of note, we have also implemented another version of the combined model with jointly training the CNN models, namely concatenating the four reads-based features to the second-last layer before the sigmoid activation (namely it becomes 36 instead of 32). This setting archives minor improvement compared to the separated training, hence is only supported as an option to choose.

We assess the performance of the three methods above by plotting the receiver operating characteristics (ROC) curves and calculating the area under the curve (AUC) values, which can systematically evaluate the sensitivities and specificities of models.

Step3: Annotate clusters

In order to connect detected clusters with existing gene annotation, the start site of transcripts from a comprehensive gene annotation GTF file was assigned to the position with the highest count as the label of clusters by using the Hungarian algorithm. The cost matrix was created by calculating the distance between the start site of each known transcript and the summit position of a query TSS cluster (i.e., with the highest UMI counts within each cluster). Our goal is to find a complete assignment of clusters to transcripts with overall minimal distance, which means to minimize Eq.3,

$$X^* = \text{argmin} \sum_{c=1}^{n_c} \sum_{t=1}^{n_t} C_{c,t} X_{c,t} \tag{3}$$

where X is a boolean matrix (X[c,t]=1 if row c is assigned to column t) and C is the cost matrix, denoting the genomic coordinate distance mentioned above. The `optimize.linear_sum_assignment` function from *scipy* python package was used to solve this assignment problem. After assigning a transcript to one cluster, we named this cluster as the corresponding transcript if the transcript is in the cluster. On the other hand, the cluster is defined as a new TSS.

Detect CTSS within one cluster

Once the TSS clusters (regions) were detected, CamoTSS can further support detection of CTSS at a single-nucleotide resolution. First, reads 1 were filtered to keep those with ‘unencoded G’ (i.e. CIGAR string starts with “14S”, “15S” and “16S”). Then the UMI counts for each position were calculated. To further denoise the signal of CTSS, a sliding window algorithm was used to calculate the fold change, as follows,

$$FC = \frac{C_0 + 1}{\frac{1}{W} \sum_{i=1}^{W} C_i + 1} \tag{4}$$

where W is the window size, and the default value is 15 in CamoTSS, and C_i denotes the UMI counts for the i position within the downstream-oriented window ($i = 0$ means the query position). The *bona fide* CTSSs within each cluster were obtained after filtering according to the fold change and UMI counts values defined by users. In the thymus development dataset, we used the default setting (fold change=6, UMI counts=100).

Evaluation of epigenetic features and RNA POL2 enrichment of detected TSS

The processed histone modification data (i.e. bigWig file) of PBMC were downloaded from the Roadmap project in ENCODE. The target of histone Chip-seq includes H3K27ac (accession: ENCFF067MDM), H3K4me3 (accession: ENCFF074XHZ) and H3K36me3 (accession: ENCFF953FFP). The aligned RNA POL2 data of PBMC obtained from Chip-seq targeting POLR2A was downloaded from ENCODE and the accession is ENCFF595NCO. The fold change (FC) data of each signal compared with the input signal was used for the histone modification data. The scRNA-seq data of PBMC [51] was dealt with CamoTSS to obtain the region of TSS (i.e. bed file). Additionally, we used `bedtools random` to generate a random set of intervals in bed format as the negative control. Then the `computeMatrix` from deepTools was utilized to calculate the histone signal score of TSS and random region [options: `computeMatrix reference-point --referencePoint TSS -b 5000 -a 5000`]. The matrix generated by `computeMatrix` was input to `plotProfile` to create a profile for the score.

Genomic tracks were obtained with pyGenomeTracks (v 3.7). We downloaded aligned bam files of 10k PBMC scATAC-seq from the 10x Genomics website (https://www.10xgenomics.com/resources/datasets/10k-human-pbmcs-atac-v2-chromium-controller-2-standard). For scRNA-seq data and scATAC-seq, we use `bamCoverage` from *deepTools* to convert the alignment file of reads (bam file) to the coverage track (bigWig file). The inputs to *pyGenomeTracks* are bigWig files of scRNA-seq, scATAC-seq, RNA POL2 and other histone markers. The interval of TSS (bed file) detected by CamoTSS was used for highlight and the GENCODE GTF file (hg38) only containing the needed transcript was used for annotation.

Analysis of genomic feature of TSS

We inputted the hg38 annotation file to *gencode_env* package (https://github.com/saketkc/gencode_regions) to obtain the genomic interval of 5’ UTR, 3’ UTR, intron and exon and then selected the start site of the TSS clusters as the symbol of them to count genomic feature distribution of TSS.

Identification of cell-type, disease and development stage-specific TSS and CTSS

Preprocessing of data was done by scanpy (v 1.9.3) [52]. For the raw TSS h5ad files (containing cell-by-TSS) of all three datasets, we filter cells according to the expression h5ad file (containing cell-by-genes). The cell annotation information and UMAP or tSNE visualization coordinates from the expression h5ad file were mapped to cells in the TSS h5ad file. For the 15 organ dataset, we normalized each cell by total counts (target_sum=1e4) over all genes. For NPC and thymic dataset, the TSS UMI reads counts were divided by the total number of reads in the same cell and then multiple with 1e6 to normalize to counts per million (CPM). Then all of these count matrices were transformed with log1p.

For detection of cell-type specific TSS masked at the gene level in the 15 organ dataset, we first filter cell types whose cell number is less than 100. For the remaining cell clusters, a t-test was performed and the difference in expression mean was calculated for each TSS between the cluster and its complement on the log1p count. We picked up TSS clusters that were significantly upregulated in the cluster relative to the complement of the cluster. In addition, the alternative TSS within the same gene cannot display the same pattern. In other words, if the alternative TSS was upregulated compared with the remaining cell clusters, then the degree of upregulation cannot be significant. All t-tests used a significance level of FDR<0.01 (Bonferroni corrected).

For NPC and thymic dataset, BRIE2 (v 2.2.0) [53] was utilized to identify differential disease-associated or development-associated TSS or CTSS at single-cell resolution. We built an h5ad file for each cell type containing two layers for the expression of two alternative TSSs with the highest expression (or two alternative CTSSs with the farthest distance) of the corresponding gene. The file containing cell detection rate and cell state information for each cell type was also created to input to BRIE2 as a design matrix. Detecting differential TSS was performed using the brie-quant module for all pairwise comparisons [options: `--batchSize 1000000 --minCell 10 --interceptMode gene --testBase full --LRTindex 0`] and genes with differential TSS between two diseases or development states were defined as FDR < 0.01.

Hierarchical clustering analysis

To investigate the similarity of TSS profiles across different organs and cell types, we first normalized the combined TSS profile of 15 organs (i.e. combined at bam file level and then run CamoTSS). Specifically, we divided the UMI counts for each TSS by the total UMI counts for all TSS in each cell type. The Pearson correlation coefficient was calculated by using normalized TSS expression of each cell type in each organ and then used to perform a hierarchical clustering analysis.

Functional enrichment and motif enrichment analysis

We utilized the Metascape online web server (v3.5.20230101) [54] to perform GO enrichment analysis [options: `Expression Analysis`] and selected the top 20 enriched terms to do enrichment network visualization. Then we used Cytoscape (v 3.9.1) to modify and visualize the network of enriched terms.

We used `bedtools getfasta` to extract the sequence of NPC- and NLH-elevated TSSs (proportional) and then downloaded 727 human TF motifs from JASPAR CORE 2022 [55]. FIMO (v 4.11.2 MEME-suite) [56] was used to search TF motif occurrences within the TSS sequence of different conditions. Significant occurrences were defined by a q-value threshold of 0.05. Then we counted the occurrence frequency of each TF motif in the TSS sets in each disease state and calculated statistical significance by using Fisher's exact test: p-value < 0.01. Then each patient's average expression of these significant TFs was used to perform hierarchical clustering. Weblogo (CLI) [57] was used to generate sequence logo of FIMO-searched and database-downloaded sequences [options: `-F pdf -A dna --color-scheme classic --fineprint "" --errorbars No`].

Data availability

The 5' scRNA-seq and bulk ATAC-seq of iPSC and Human dermal fibroblasts used to do training dataset were downloaded from ArrayExpress under the accessions E-MTAB-10385 and E-MTAB-10381 [24]. Previously published 5' scRNA-seq data that were reanalyzed here are available in the GEO or ArrayExpress under the primary accession code "GSE111360" (PBMC) [51], "GSE159929" (15 organs) [29], "GSE150825" (nasopharyngeal carcinoma) [33], "E-MTAB-8581" (human thymic development) [58]. The cell type annotation is available within the article and its supplementary files, with a copy in the reproducibility GitHub repository.

Codes availability

CamoTSS is a publicly available Python package at `https://github.com/StatBiomed/CamoTSS`. Detailed documentation and analysis procedures to reproduce results in this paper are also uploaded to this repository.

Author contributions

Y.H. conceived and supervised this study. R.H. implemented the CamoTSS and performed all data analysis. C.C. provided guidance on CTSS and unencoded G analyses. R.H. and Y.H. wrote the manuscript.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Chen Qiao and Weizhong Zheng for technical help on CNN model building and troubleshooting. We thank Xianjie Huang for helping make the package efficient and for troubleshooting. We thank Jiaqi Li for the suggestion on transcription factor detection.

References

1. Alejandro Reyes and Wolfgang Huber. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic acids research*, 46(2):582–592, 2018.
2. Yusuke Shiozawa, Luca Malcovati, Anna Gallì, Aiko Sato-Otsubo, Keisuke Kataoka, Yusuke Sato, Yosaku Watatani, Hiromichi Suzuki, Tetsuichi Yoshizato, Kenichi Yoshida, et al. Aberrant splicing and defective mrna production induced by somatic spliceosome mutations in myelodysplasia. *Nature communications*, 9(1):1–16, 2018.
3. Alicia C Smart, Claire A Margolis, Harold Pimentel, Meng Xiao He, Diana Miao, Dennis Adeegbe, Tim Fugmann, Kwok-Kin Wong, and Eliezer M Van Allen. Intron retention is a source of neoepitopes in cancer. *Nature biotechnology*, 36(11):1056–1058, 2018.
4. Aaron M Horning, Yao Wang, Che-Kuang Lin, Anna D Louie, Rohit R Jadhav, Chia-Nung Hung, Chiou-Miin Wang, Chun-Lin Lin, Nameer B Kirma, Michael A Liss, et al. Single-Cell RNA-seq Reveals a Subpopulation of Prostate Cancer Cells with Enhanced Cell-Cycle–Related Transcription and Attenuated Androgen ResponseHeterogeneous Androgen Responses of Prostate Cancer Cells. *Cancer research*, 78(4):853–864, 2018.
5. Wei Xiong Wen, Adam J Mead, and Supat Thongjuea. Technological advances and computational approaches for alternative splicing analysis in single cells. *Computational and structural biotechnology journal*, 18:332–343, 2020.
6. Yuanhua Huang and Guido Sanguinetti. BRIE: transcriptome-wide splicing quantification in single cells. *Genome biology*, 18(1):1–11, 2017.
7. Yan Song, Olga B Botvinnik, Michael T Lovci, Boyko Kakaradov, Patrick Liu, Jia L Xu, and Gene W Yeo. Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. *Molecular cell*, 67(1):148–161, 2017.
8. Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nature methods*, 7(12):1009–1015, 2010.
9. Julia Eve Olivieri, Roozbeh Dehghannasiri, and Julia Salzman. The SpliZ generalizes 'Percent Spliced In'to reveal regulated splicing at single-cell resolution. *Nature methods*, 19(3):307–310, 2022.
10. Yu Hu, Kai Wang, and Mingyao Li. Detecting differential alternative splicing events in scRNA-seq with or without Unique Molecular Identifiers. *PLoS computational biology*, 16(6):e1007925, 2020.
11. Ralph Patrick, David T Humphreys, Vaibhao Janbandhu, Alicia Oshlack, Joshua WK Ho, Richard P Harvey, and Kitty K Lo. Sierra: discovery of differential transcript usage from polya-captured single-cell rna-seq data. *Genome biology*, 21(1):1–27, 2020.
12. Guo-Wei Li, Fang Nan, Guo-Hua Yuan, Chu-Xiao Liu, Xindong Liu, Ling-Ling Chen, Bin Tian, and Li Yang. SCAPTURE: a deep learning-embedded pipeline that captures polyadenylation information from 3' tag-based RNA-seq of single cells. *Genome biology*, 22(1):1–24, 2021.
13. Rui Fu, Kent A Riemondy, Ryan M Sheridan, Jay R Hesselberth, Craig T Jordan, and Austin E Gillen. scraps: an end-to-end pipeline for measuring alternative polyadenylation at high resolution using single-cell rna-seq. *bioRxiv*, 2022.

14. Toshiyuki Shiraki, Shinji Kondo, Shintaro Katayama, Kazunori Waki, Takeya Kasukawa, Hideya Kawaji, Rimantas Kodzius, Akira Watahiki, Mari Nakamura, Takahiro Arakawa, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences*, 100(26):15776–15781, 2003.

15. The FANTOM Consortium, the RIKEN PMI, and CLST (DGT). A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470, 2014.

16. Vanja Haberle, Nan Li, Yavor Hadzhiev, Charles Plessy, Christopher Previti, Chirag Nepal, Jochen Gehrig, Xianjun Dong, Altuna Akalin, Ana Maria Suzuki, et al. Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature*, 507(7492):381–385, 2014.

17. Jill E Moore, Xiao-Ou Zhang, Shaimae I Elhajjajy, Kaili Fan, Henry E Pratt, Fairlie Reese, Ali Mortazavi, and Zhiping Weng. Integration of high-resolution promoter profiling assays reveals novel, cell type–specific transcription start sites across 115 human cell and tissue types. *Genome Research*, 32(2):389–402, 2022.

18. Deniz Demircioğlu, Engin Cukuroglu, Martin Kindermans, Tannistha Nandi, Claudia Calabrese, Nuno A Fonseca, André Kahles, Kjong-Van Lehmann, Oliver Stegle, Alvis Brazma, et al. A pan-cancer transcriptome analysis reveals pervasive regulation through alternative promoters. *Cell*, 178(6):1465–1477, 2019.

19. Raghav Sundar, Kie-Kyon Huang, Vikrant Kumar, Kalpana Ramnarayanan, Deniz Demircioglu, Zhisheng Her, Xuewen Ong, Zul Fazreen Bin Adam Isa, Manjie Xing, Angie Lay-Keng Tan, et al. Epigenetic promoter alterations in gi tumour immune-editing and resistance to immune checkpoint inhibition. *Gut*, 71(7):1277–1288, 2022.

20. Luis V Valcárcel, Ane Amundarain, Marta Kulis, Stella Charalampopoulou, Ari Melnick, Jesús San Miguel, José I Martín-Subero, Francisco J Planes, Xabier Agirre, and Felipe Prosper. Gene expression derived from alternative promoters improves prognostic stratification in multiple myeloma. *Leukemia*, 35(10):3012–3016, 2021.

21. Minghao Chia, Cai Li, Sueli Marques, Vicente Pelechano, Nicholas M Luscombe, and Folkert J van Werven. High-resolution analysis of cell-state transitions in yeast suggests widespread transcriptional tuning by alternative starts. *Genome biology*, 22(1):1–37, 2021.

22. Tsukasa Kouno, Jonathan Moody, Andrew Tae-Jun Kwon, Youtaro Shibayama, Sachi Kato, Yi Huang, Michael Böttcher, Efthymios Motakis, Mickaël Mendez, Jessica Severin, et al. C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. *Nature communications*, 10(1):1–12, 2019.

23. Youjin Hu, Jiawei Zhong, Yuhua Xiao, Zheng Xing, Katherine Sheu, Shuxin Fan, Qin An, Yuanhui Qiu, Yingfeng Zheng, Xialin Liu, et al. Single-cell RNA cap and tail sequencing (scRCAT-seq) reveals subtype-specific isoforms differing in transcript demarcation. *Nature communications*, 11(1):1–11, 2020.

24. Jonathan Moody, Tsukasa Kouno, Jen-Chien Chang, Yoshinari Ando, Piero Carninci, Jay W Shin, and Chung-Chau Hon. SCAFE: a software suite for analysis of transcribed cis-regulatory elements in single cells. *Bioinformatics*, 38(22):5126–5128, 2022.

25. Xian Adiconis, Adam L Haber, Sean K Simmons, Ami Levy Moonshine, Zhe Ji, Michele A Busby, Xi Shi, Justin Jacques, Madeline A Lancaster, Jen Q Pan, et al. Comprehensive comparative analysis of 5'-end rna-sequencing methods. *Nature methods*, 15(7):505–511, 2018.

26. Nevena Cvetesic, Harry G Leitch, Malgorzata Borkowska, Ferenc Müller, Piero Carninci, Petra Hajkova, and Boris Lenhard. SLIC-CAGE: high-resolution transcription start site mapping using nanogram-levels of total RNA. *Genome Research*, 28(12):1943–1956, 2018.

27. Vu Ngo, Zhao Chen, Kai Zhang, John W Whitaker, Mengchi Wang, and Wei Wang. Epigenomic analysis reveals dna motifs regulating histone modifications in human and mouse. *Proceedings of the National Academy of Sciences*, 116(9):3668–3677, 2019.

28. Sarah Sainsbury, Carrie Bernecky, and Patrick Cramer. Structural basis of transcription initiation by rna polymerase ii. *Nature reviews Molecular cell biology*, 16(3):129–143, 2015.

29. Shuai He, Lin-He Wang, Yang Liu, Yi-Qi Li, Hai-Tian Chen, Jing-Hong Xu, Wan Peng, Guo-Wang Lin, Pan-Pan Wei, Bo Li, et al. Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. *Genome biology*, 21:1–34, 2020.

30. Paul Cheng, Robert C Wirka, Lee Shoa Clarke, Quanyi Zhao, Ramendra Kundu, Trieu Nguyen, Surag Nair, Disha Sharma, Hyun-jung Kim, Huitong Shi, et al. ZEB2 shapes the epigenetic landscape of atherosclerosis. *Circulation*, 145(6):469–485, 2022.

31. David Massa López, Melanie Thelen, Felix Stahl, Christian Thiel, Arne Linhorst, Marc Sylvester, Irm Hermanns-Borgmeyer, Renate Lüllmann-Rauch, Winnie Eskild, Paul Saftig, et al. The lysosomal transporter MFSD1 is essential for liver homeostasis and critically depends on its accessory subunit GLMP. *Elife*, 8:e50025, 2019.

32. Jake R Conway, Alexander Lex, and Nils Gehlenborg. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, 2017.

33. Lanqi Gong, Dora Lai-Wan Kwong, Wei Dai, Pingan Wu, Shanshan Li, Qian Yan, Yu Zhang, Baifeng Zhang, Xiaona Fang, Li Liu, et al. Comprehensive single-cell sequencing reveals the stromal dynamics and tumor-specific characteristics in the microenvironment of nasopharyngeal carcinoma. *Nature Communications*, 12(1):1540, 2021.

34. Nicolas Pernodet, François Hermetet, Pascale Adami, Anne Vejux, Françoise Descotes, Christophe Borg, Marjorie Adams, Jean-René Pallandre, Gabriel Viennet, Frédéric Esnard, et al. High expression of QSOX1 reduces tumorogenesis, and is associated with a better outcome for breast cancer patients. *Breast Cancer Research*, 14(5):1–15, 2012.

35. Denise D Correa, Jaya Satagopan, Kenneth Cheung, Arshi K Arora, Maria Kryza-Lacombe, Youming Xu, Sasan Karimi, John Lyo, Lisa M DeAngelis, and Irene Orlow. COMT, BDNF, and DTNBP1 polymorphisms and cognitive functions in patients with brain tumors. *Neuro-oncology*, 18(10):1425–1433, 2016.

36. Junfu Guo, Yue Bian, Yu Wang, Lisha Chen, Aiwen Yu, and Xiuju Sun. FAM107B is regulated by S100A4 and mediates the effect of S100A4 on the proliferation and migration of MGC803 gastric cancer cells. *Cell Biology International*, 41(10):1103–1109, 2017.

37. Francois-Xavier Dieudonné, Patrick BF O'Connor, Pascale Gubler-Jaquier, Haleh Yasrebi, Beatrice Conne, Sergey Nikolaev, Stylianos Antonarakis, Pavel V Baranov, and Joseph Curran. The effect of heterogeneous Transcription Start Sites (TSS) on the translatome: implications for the mammalian cellular phenotype. *BMC genomics*, 16(1):1–15, 2015.

38. Kensuke Yamaguchi, Kazuyoshi Ishigaki, Akari Suzuki, Yumi Tsuchida, Haruka Tsuchiya, Shuji Sumitomo, Yasuo Nagafuchi, Fuyuki Miya, Tatsuhiko Tsunoda, Hirofumi Shoda, et al. Splicing QTL analysis focusing on coding sequences reveals mechanisms for disease susceptibility loci. *Nature communications*, 13(1):4659, 2022.

39. Jayu Jen and Yi-Ching Wang. Zinc finger proteins in cancer progression. *Journal of biomedical science*, 23(1):1–9, 2016.

40. Larry Ka-Yue Chow, Dittman Lai-Shun Chung, Lihua Tao, Kui Fat Chan, Stewart Yuk Tung, Roger Kai Cheong Ngan, Wai Tong Ng, Anne Wing-Mui Lee, Chun Chung Yau, Dora Lai-Wan Kwong, et al. Epigenomic landscape study reveals molecular subtypes and EBV-associated regulatory epigenome reprogramming in nasopharyngeal carcinoma. *EBioMedicine*, 86:104357, 2022.

41. Qing Deng, Sa Kan Yoo, Peter J. Cavnar, Julie M. Green, and Anna Huttenlocher. Dual Roles for Rac2 in Neutrophil Motility and Active Retention in Zebrafish Hematopoietic Tissue. *Developmental Cell*, 21(4):735–745, 2011.

42. Kayo Ikeda, Makoto Kinoshita, Hisako Kayama, Shushi Nagamori, Pornparn Kongpracha, Eiji Umemoto, Ryu Okumura, Takashi Kurakawa, Mari Murakami, Norihisa Mikami, et al. Slc3a2 mediates branched-chain amino-acid-dependent maintenance of regulatory T cells. *Cell reports*, 21(7):1824–1838, 2017.

43. Marcin Lyszkiewicz, Natalia Zietara, Laura Frey, Ulrich Pannicke, Marcel Stern, Yanshan Liu, Yanxin Fan, Jacek Puchalka, Sebastian Hollizeck, Ido Somekh, et al. Human FCHO1 deficiency reveals role for clathrin-mediated endocytosis in development and function of T cells. *Nature communications*, 11(1):1031, 2020.

44. Eric E Irons, Patrick R Punch, and Joseph TY Lau. Blood-borne ST6GAL1 regulates immunoglobulin production in B cells. *Frontiers in Immunology*, 11:617, 2020.

45. Sarah Schulz, Georgia Chachami, Lukasz Kozaczkiewicz, Ulrike Winter, Nicolas Stankovic-Valentin, Petra Haas, Kay Hofmann, Henning Urlaub, Huib Ovaa, Joachim Wittbrodt, et al. Ubiquitin-specific protease-like 1 (USPL1) is a SUMO isopeptidase with essential, non-catalytic functions. *EMBO reports*, 13(10):930–938, 2012.

46. Anna B Osipovich, Rama Gangula, Pedro G Vianna, and Mark A Magnuson. Setd5 is essential for mammalian development and the co-transcriptional regulation of histone acetylation. *Development*, 143(24):4595–4607, 2016.

47. Navigating 10x genomics barcoded bam files. `https://www.10xgenomics.com/resources/analysis-guides/tutorial-navigating-10x-barcoded-bam-files`. Accessed: 2021-10-14.

48. Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *bioinformatics*, 25(16):2078–2079, 2009.

49. Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

50. Hideki Ohtake, Kuniyo Ohtoko, Yoshihiro Ishimaru, and Seishi Kato. Determination of the capped site sequence of mrna based on the detection of cap-dependent nucleotide addition using an anchor ligation method. *DNA research*, 11(4):305–309, 2004.

51. James T Neal, Xingnan Li, Junjie Zhu, Valeria Giangarra, Caitlin L Grzeskowiak, Jihang Ju, Iris H Liu, Shin-Heng Chiou, Ameen A Salahudeen, Amber R Smith, et al. Organoid modeling of the tumor immune microenvironment. *Cell*, 175(7):1972–1988, 2018.

52. F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

53. Yuanhua Huang and Guido Sanguinetti. BRIE2: computational identification of splicing phenotypes from single-cell transcriptomic experiments. *Genome biology*, 22(1):1–15, 2021.

54. Yingyao Zhou, Bin Zhou, Lars Pache, Max Chang, Alireza Hadj Khodabakhshi, Olga Tanaseichuk, Christopher Benner, and Sumit K Chanda. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature communications*, 10(1):1523, 2019.

55. Jaime A Castro-Mondragon, Rafael Riudavets-Puig, Ieva Rauluseviciute, Roza Berhanu Lemma, Laura Turchi, Romain Blanc-Mathieu, Jeremy Lucas, Paul Boddie, Aziz Khan, Nicolás Manosalva Pérez, et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic acids research*, 50(D1):D165–D173, 2022.

56. Charles E Grant, Timothy L Bailey, and William Stafford Noble. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

57. Gavin E Crooks, Gary Hon, John-Marc Chandonia, and Steven E Brenner. Weblogo: a sequence logo generator. *Genome research*, 14(6):1188–1190, 2004.

58. Jong-Eun Park, Rachel A Botting, Cecilia Domínguez Conde, Dorin-Mirel Popescu, Marieke Lavaert, Daniel J Kunz, Issac Goh, Emily Stephenson, Roberta Ragazzini, Elizabeth Tuck, et al. A cell atlas of human thymic development defines T cell repertoire formation. *Science*, 367(6480):eaay3224, 2020.