Profiling of transcribed *cis*-regulatory elements in single cells

Jonathan Moody[*,1], Tsukasa Kouno[*,1], Akari Suzuki[1], Youtaro Shibayama[1], Chikashi Terao[1], Jen-Chien Chang[1], Fernando López-Redondo[1], Chi Wai Yip[1], Yoshinari Ando[1], Kazuhiko Yamamoto[1], Piero Carninci[1,2], Jay W. Shin[†,1], Chung-Chau Hon[†,1]

[1] RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan 230-0045.
[2] Human Technopole, Via Cristina Belgioioso 171, Milano 20157, Italy.
[*] These authors contributed equally
[†] Correspondence should be addressed to C.C.H (chungchau.hon@riken.jp) or J.W.S (jay.shin@riken.jp)

Abstract

Profiling of *cis*-regulatory elements (CREs, mostly promoters and enhancers) in single cells allows the interrogation of the cell-type and -state specific contexts of gene regulation and genetic predisposition to diseases. Here we demonstrate single-cell RNA-5′end-sequencing (sc-end5-seq) methods can detect transcribed CREs (tCREs), enabling simultaneous quantification of gene expression and enhancer activities in a single assay with no extra cost. We show enhancer RNAs can be effectively detected using sc-end5-seq methods with either random or oligo(dT) priming. To analyze tCREs in single cells, we developed *SCAFE* (Single Cell Analysis of Five-prime Ends) to identify genuine tCREs and analyze their activities (https://github.com/chung-lab/scafe). As compared to accessible CRE (aCRE, based on chromatin accessibility), tCREs are more accurate in predicting CRE interactions by co-activity, more sensitive in detecting shifts in alternative promoter usage and more enriched in diseases heritability. Our results highlight additional dimensions within sc-end5-seq data which can be used for interrogating gene regulation and disease heritability.

Main text

Introduction

Expression of genes specifying cell identity (i.e. cell-types and -states) is primarily controlled by the activities of their cognate CREs, mostly promoters[1] and enhancers[2]. These CREs are highly enriched in disease associated variants[3], reflecting the importance of gene regulation in diseases. Therefore, understanding the cell-identity specific CRE activities not only helps to decipher the principles of gene regulation[4,5], but also the cellular contexts of genetic predisposition to diseases[6]. While gene expression can be quantified with single-cell RNA-sequencing methods (sc-RNA-seq)[7,8], profiling of CREs primarily relies on single-cell Assay for Transposase Accessible Chromatin using sequencing (sc-ATAC-seq)[9,10], which measures the accessibility of chromatin regions in a binary manner (i.e. accessible or non-accessible)[11]. Several methods were developed for joint profiling of gene expression and chromatin accessibility within the same cell[12–15], allowing the prediction of CREs interactions to their target genes (i.e. enhancer-to-promoter (EP) interactions), through cell-to-cell co-variations of their activities[13,16]. However, the close-to-binary nature and excess sparsity of chromatin accessibility data render the analyses of individual CREs in single cells challenging[17]. Also, a substantial fraction of accessible CREs that are distant from annotated promoters (i.e. distal aCREs) do not show the epigenomic features of active enhancers[18]. While an unknown fraction of these non-enhancer distal aCREs could be regulatory, e.g. insulators[19] or silencers[20], their overall relevance in gene regulation remains elusive.

Alternatively, measuring the transcription at CRE (or tCRE) can be used as a proxy for their activity[2], which can be achieved by sequencing the 5′ends of RNA[21] representing the transcription start sites (TSS) within the CREs[1]. Such measurement is highly quantitative and is ranked as the top feature for predicting active EP interactions in a machine-learning approach, compared to other epigenomic features[22]. In fact, the co-variation of transcription signals between CREs were shown to accurately predict individual cell-type-specific EP interactions[23]. In addition, transgenic enhancer assays showed endogenous transcription at a distal CRE is highly correlated with its ability to function as an enhancer[24]. These observations suggest distal CRE identified and quantified by transcription evidence, compared to solely chromatin accessibility evidence, could be more relevant to enhancer activation of gene expression.

Previously, we demonstrated the application of sc-end5-seq in the integrated fluidic circuit-based C1™ platform (Fluidigm®) for detection of known TSS in hundreds of single cells[25]. In this study, we evaluated the sc-end5-seq methods on the droplet-based Chromium™ platform (10x Genomics®), with random and oligo(dT) priming, for *de novo* discovery and quantification of tCREs in thousands of single cells. Unexpectedly, both random and oligo(dT) priming methods effectively detected enhancer RNAs, which are supposed to be mostly non-polyadenylated (non-polyA)[2]. A major challenge in *de novo* discovery of tCREs from sc-end5-seq data is artifactual template switching (TS) reactions producing false TSS[26,27]. Therefore, we have devised a multiple logistic regression classifier to identify genuine TSS and effectively minimize false positives. Applying both sc-end5-seq and sc-ATAC-seq to peripheral blood mononuclear cells (PBMC) with immuno-stimulation, we compared the performance of tCREs and aCREs in: 1) identification of cell-type specific CREs, 2) detection of stimulation-induced transcription factor (TF) activities and 3) shifts in alternative promoter usage, 4) prediction of CRE interactions by co-activity, 5) enrichment in diseases heritability and 6) functional interpretations of disease associated variants. Finally, we developed *SCAFE*, a command-line tool to annotate genuine tCREs and predict their interactions from RNA-5′end-sequencing data.

Results

Assessing the performance of 3′end and 5′end sc-RNA-seq methods

While the sc-end5-seq method on the Chromium™ platform is primed with oligo(dT) (sc-end5-dT), we modified the protocol with random hexamer priming (sc-end5-rand), aiming for enhanced detection of non-polyA RNAs (see Methods)[28,29]. We then performed both sc-end5-dT and sc-end5-rand methods, along with the oligo(dT) primed 3′end sc-RNA-seq method (sc-end3-dT), on human dermal fibroblasts (DMFB) and induced pluripotent stem cells (iPSC) (Supplementary Fig. 1a). For comparison, CAGE, RNA-seq, and ATAC-seq were also performed in bulk on both cell lines (Fig. 1). In the following section, we focus on iPSC for the sake of clarity (Fig. 2; see Supplementary Fig. 2 for DMFB).

First, we assessed the global distributions of reads. As expected, in sc-end5-rand more reads are mapped to ribosomal RNA (rRNA) (~15%) than in sc-end5-dT (~2%) (Fig. 2a, *upper panel*, *whole genome*). When considering reads mapped within genes (i.e., genic reads), the percentage of reads mapped to TSS is lower in both sc-end5-seq (~60%) than in bulk-CAGE (~85%) (Fig. 2a, *upper panel*, *within gene*), reflecting the greater extent of non-specific artefacts in sc-end5-seq, as discussed in the next section. We also note the genic reads of both sc-end5-seq methods and bulk-CAGE are strongly enriched at the 5′end of genes (Fig. 2a, *middle panel*) and peaked precisely at the annotated TSS (Fig. 2a, *lower panel*), suggesting both sc-end5-seq methods can precisely pinpoint TSS.

Next, we assessed the sensitivity of gene detection by sc-RNA-seq methods. When considering pooled single cells (i.e., pseudo-bulk), all three sc-RNA-seq methods showed similar sensitivities (Fig. 2b, *left panel*). However, when considering per-cell, both oligo(dT)-primed methods (i.e., sc-end3-dT and sc-end5-dT) detected ~30% more genes than the random-primed method (i.e., sc-end5-rand) at matched sequencing depths (Fig. 2b, *right panel*). This might be explained by lower complexity of the sc-end5-rand per-cell libraries, attributed to its higher rRNA read percentage and higher reads per unique molecular identifier (UMI) (Supplementary Fig. 1). Overall, the pseudo-bulk expression level of genes among the three sc-

RNA-seq methods are highly correlated (Fig. 2c), allowing datasets from these three sc-RNA-seq methods to be robustly integrated (Supplementary Fig. 3), and opening the possibility of joint-analyses of sc-end5-seq datasets with the many available sc-end3-dT public datasets.

To further examine the differences between the two priming methods, we tested for the enrichment of subcellular compartment-specific RNAs (see Methods), non-polyA histone RNAs[30], and long or short RNAs (see Methods). In the genes expressed higher in sc-end5-rand compared to sc-end5-dT, we observed strong enrichment of non-polyA histone RNAs (FDR <0.005, Fig. 2d). This is supported by the enrichment of chromatin-bound RNAs (FDR <0.005, Fig. 2d), which contain many nascent RNAs and non-polyA RNAs[31]. The significant enrichment of long RNAs (FDR <0.005, Fig. 2d) might be attributed to the higher reverse transcription efficiency of random priming within the body of the longer transcripts, in contrast to oligo(dT) priming which mainly from the 3′end of transcripts. Unexpectedly, sc-end5-dT also detected non-polyA histone RNAs with moderate expression (Fig. 2d), suggesting potential internal priming at A-rich sequences in sc-end5-dT, which has been also observed extensively in sc-end3-dT method[32–34]. In summary, these observations suggest a comparable performance in gene detection for the three sc-RNA-seq methods, with sc-end5-rand showing slightly lower per-cell sensitivity and sc-end5-dT showing unexpected detection of non-polyA RNAs.

TSS identification using sc-end5-seq methods

Previous reports suggested a fraction of TSS identified based on read 5′ends from TS reactions may not be genuine[26,27], attributed to various artefacts including strand invasion[27] and other sources[35]. This results in excessive artifactual TSS, especially along the gene body known as “exon painting”[36]. While a fraction of these “exon painting” reads could be attributed to cleavage and recapping[37], their exact molecular origins remain elusive. To this end, we developed a novel method in *SCAFE* to identify genuine TSS (Supplementary Fig. 4).

First, we filter strand invasion artefacts based on the complementarity to TS oligo sequence[35] and found more strand invasion artefacts in sc-end5-rand (~5% reads) than in sc-end5-dT (~3% reads) (Supplementary Fig. 5). The filtered reads were then clustered. We found the proportion of TSS clusters along the gene body in both sc-end5-seq methods were still substantially higher than bulk-CAGE (Supplementary Fig. 6), consistent with the fact that “exon painting” is more prevalent in TS-based methods[26]. We benchmarked the properties of TSS clusters (Fig. 3a) and devised a classifier for genuine TSS using multiple logistic regression (see Methods) (Fig. 3b). Here we focus on the sc-end5-dT iPSC dataset for simplicity. First, the UMI counts within the TSS cluster (i.e. cluster counts) performed the worst (Area Under Receiver Operating Characteristic (ROC) Curve (AUC)=0.641) (Fig. 3a), and its performance decreases with sequencing depth (Fig. 3c). Two other common metrics, UMI count at TSS summit (i.e. summit count, AUC=0.725) and within ±75nt flanking its summit (i.e. flanking count, AUC=0.737) performed only marginally better than the cluster count (Fig. 3a,c), suggesting these commonly used metrics are at best mediocre classifiers for TSS. Since “exon painting” artefacts should be positively correlated with transcript abundance, we examined other metrics that are independent of RNA expression level, including UMI counts corrected for background expression (i.e. corrected expression, see Methods) and percentage of reads with 5′ mismatched G[26] (i.e. unencoded-G percentage, see Methods). We found both metrics performed well across sequencing depths with AUC >0.9 (Fig. 3c).

To devise a TSS classifier, we combined metrics using multiple logistic regression. We found the combination of flanking count, unencoded-G percentage and corrected expression is sufficient to achieve the best performance, with AUC >0.98 across sequencing depths (Fig. 3b,c). Its accuracy is high and robust for TSS clusters located in various genomic regions and across a wide range of cutoffs (Supplementary Fig. 7a), which is well-validated by chromatin accessibility, promoter motifs, CpG island, sequence conservation (Supplementary Fig. 7b,c,d,e,f) and histone marks (Fig. 3d). At the default cutoff of 0.5, ~98% of sense exonic TSS clusters were removed (Fig. 3d, *3rd row*). These removed TSS clusters are void of marks for active CREs (e.g., H3K27ac, H3K4me1 and H3K4me3) but overlap marks for transcription elongation (e.g., H3K36me3), suggesting our TSS classifier effectively removed “exon painting” artifacts.

In addition, the TSS clusters located at gene TSS are marked with a bimodal H3K4me1 pattern which indicates active promoters, in contrast to the others that are marked with relatively unimodal H3K4me1 pattern which indicates active enhancers[38,39]. In summary, the *SCAFE* TSS classifier robustly distinguishes genuine TSS from artifacts.

Defining tCRE using sc-end5-seq methods

tCREs are defined in *SCAFE* by merging closely located TSS clusters and classified as either proximal or distal based on their distance to annotated gene TSS (Fig. 4a). Proximal tCRE can be interpreted as promoters of genes and promoter upstream transcripts (PROMPTs)[40]. Distal tCRE can be interpreted as mostly enhancers[41], with an unknown, but likely minor, fraction of them as unannotated promoters (e.g. alternative promoters). To benchmark the sensitivity of tCRE detection, we also performed bulk-CAGE on chromatin-bound RNA (bulk-Chrom-CAGE), which captures the 5′ends of nascent transcripts for sensitive detection of short-lived RNAs (e.g. enhancer RNAs)[31] and can thus be viewed as a permissive baseline for their detection. First, we found similar proportions of tCREs defined as distal in sc-end5-dT (~10%) and sc-end5-rand (~12%) (Fig. 4b, *all tCRE*), suggesting a similar sensitivity of enhancer RNA detection in both methods. In addition, amongst distal tCREs the proportions of exonic, intergenic and intronic were similar across the bulk and single-cell 5′end methods (Fig. 4b, *distal tCRE*). Considering the excessive exonic TSS cluster in sc-end5-seq before filtering (Supplementary Fig. 6), it suggests the filtering step effectively minimized the "exon painting" artefacts in sc-end5-seq.

Next we assessed the sensitivity of tCRE detection in various methods (Fig. 4c,d,e). As expected, bulk-Chrom-CAGE showed the highest sensitivity (Fig. 4c). Both sc-end5-seq methods detected ~50% to ~80% of those detected by bulk methods at matched sequencing depths (Fig. 4c). In pseudo-bulk, although sc-end5-rand seemed slightly more sensitive at lower depths (Fig. 4d, at ~50M), the sensitivity of both methods are similar at higher depths (Fig. 4d, at ~150M). When considered per-cell, however, sc-end5-dT is substantially more sensitive than sc-end5-rand (Fig. 4e). The tCREs identified in both methods are largely overlapping (Fig. 4f) and their expression levels are highly correlated (Fig. 4g). The high concordance of distal tCREs between sc-end5-dT and sc-end5-rand is unexpected, assuming a considerable fraction of these distal tCREs are enhancers, which produce mostly non-polyA RNAs[2]. To further investigate this, we examined the balanced bidirectionally transcribed enhancer loci in DMFB and iPSC (defined by bulk-CAGE as previously described[2]). Both sc-end5-dT and sc-end5-rand recapitulated these bulk-defined cell-type specific bidirectional transcription pattern at comparable number of enhancer loci (Fig. 4h), confirming that both sc-end5-seq methods detected enhancer RNAs with similar sensitivity. The unexpected detection of enhancer RNAs by sc-end5-dT could be attributed to the potential internal priming[32–34], as discussed. In view of their similar pseudo-bulk performances (Fig. 4d,f,g,h) and the superior per-cell performance of sc-end5-dT (Fig. 4e), we performed sc-end5-dT and sc-ATAC-seq in PBMC for the comparison of tCRE and aCRE.

Comparing tCRE and aCRE in PBMC

We next defined tCREs (n =30,180) and aCREs (n =157,055) in PBMCs treated with PMA/ionomycin (i.e., stimulated cell state) or DMSO (i.e., resting cell state) (Fig. 1, Supplementary Fig. 1). Gene-based cell-type annotations were transferred from the tCRE cells to aCRE cells using CCA[42] (Supplementary Fig. 8). Either UMAPs based on tCRE or aCRE show similar separation of cell-types and excellent integration of cell-states (Fig. 5a). Examining a subset of aCREs with cell-type specific chromatin accessibility (see Methods, Fig. 5b, *top row*), we found concordant patterns of cell-type specific RNA expression at the overlapping tCREs (Fig.5b, *bottom row*). To examine cell-type specific TF activity, we applied *ChromVAR*[43] to both aCRE and tCRE to estimate TF motif activities and defined cell-type specific motifs (see Methods). These cell-type specific motifs based on aCRE and tCRE are significantly concordant in most cell-types (Supplementary Fig. 9a, Fisher's exact test, P <0.05). Clustering of cell-types using TF motif activities appears to be consistent within broad categories with co-clustering of monocytes, lymphocytes and cytotoxic T-cells between aCRE and tCRE (Supplementary Fig. 9b). We further examined the activation of TF upon stimulation (see Methods) and observed a generally consistent upregulation of TF motif activities between aCREs and tCREs (Fig. 5c, mean Pearson's r=0.84), which is mostly driven by *JUN/FOS* related motifs that are components of the early immune responses. These results suggest both tCRE and aCRE can recover cell-type and -state specific contexts of gene regulation (i.e. CRE and TF activities).

Co-activity of a pair of CREs can be used to predict their physical interactions[16]. Here we compared the accuracy of tCREs and aCREs in prediction of interacting CREs, using the co-activity estimated in *Cicero*[16], benchmarked against promoter-capture Hi-C (PCHi-C)[44] (see Methods). Co-activity scores were estimated separately using cells within individual cell-types (i.e. cell-type sets) or all cells (i.e. pooled set). Here, we focus on a subset of CREs that is overlapping between tCRE and aCRE. First, we observed significantly higher co-activity scores for tCRE-pairs than aCRE-pairs (Fig. 5d, $P < 2.2 \times 10^{-16}$ in K-S test for the pooled set, *solid line*). At co-activity scores ≥0.2, we found the linked tCRE-pairs are significantly more likely to be validated by PCHi-C (~40%) than the linked aCRE-pairs (~10%) (Fig. 5e, $P < 7 \times 10^{-6}$, paired *t*-test for the cell-type sets). These results suggest tCREs are more accurate in predicting CRE interactions by co-activity.

Alternative promoter usage is an important mechanism to increase transcriptome diversity for generation functionally distinct isoforms[45]. Here we examined the power of tCRE and aCRE to detect shifts in alternative promoter usage upon stimulation. First, we found 123 genes with significant shifts in tCRE (i.e. alternative promoter) usage upon stimulation in at least one cell-type (FDR <0.05 in *t*-test). We then examined the chromatin accessibility signals at the corresponding tCREs and observed only minimal extent of shifts in accessibility (Fig. 5f, *horizontal box plot, top*). Highlighting the *DHX30* locus (Fig. 5g), in T-cell:CD8:naïve, its expression shifts from Promoter#1 to Promoter#2 upon stimulation, whereas in Monocyte:CD14, no shift in expression occurs (Fig. 5f,h; Supplementary Fig. 10). In contrast, the chromatin accessibility at the two promoters remains mostly constant between the two states in all cell-types (Fig. 5h). These results suggest tCREs are generally more sensitive in detecting shifts in alternative promoter usage upon cell-state changes.

Enrichment of trait associated variants in tCRE

For interpretation of genetic predisposition, we examined the enrichment of trait heritability[46] in CREs from PBMCs. For comparison, we used tCRE defined with default and lenient logistic probability cutoffs (see Methods). As expected, we found both tCREs and aCREs are enriched in hematologic and immunologic traits, but generally not in psychiatric and metabolic traits (Fig. 6a, *top row*). The pattern is similar when considering proximal and distal CREs separately (Fig. 6a, *middle and bottom row*), implying that distal tCREs are biologically relevant. In addition, the enrichment in default tCREs is generally higher than that of lenient tCREs, particularly for distal tCREs (Fig. 6a), suggesting a higher proportion of default tCRE is biologically relevant. Nonetheless, we also noticed the default tCRE are less sensitive in terms of reaching statistical significance, which can be attributed to the smaller number of SNPs in default tCRE leading to larger estimates of standard error as reported[47]. For the sake of statistical power, we thus used lenient tCREs in the rest of the heritability enrichment analyses.

As we observed a generally higher level of enrichment in default distal tCREs than in aCREs (Fig. 6a, *bottom row*), thus we reasoned transcription at CRE could be indicative to its activity and thus biological relevance. To this end, we investigated the heritability enrichment in aCREs with various levels of transcription evidence (Fig. 6b). About 45% of all aCREs showed evidence of transcription (i.e. transcribed aCRE, Fig. 6b, *top row, right panel*). This percentage is comparable to our estimate that ~47% of aCREs are transcribed in DMFB based on bulk-CAGE with an unprecedented sequencing depth of 12,000M reads (Supplementary Fig. 11, based on FANTOM6 CAGE datasets[48], see Methods), suggesting this percentage of transcribed aCRE in PBMC is a reasonable estimate despite limited sequencing depth at ~1,000M reads (Supplementary Fig. 1a). Untranscribed aCREs may be poised promoters, untranscribed enhancers, silencers, insulators or technical artifacts of sc-ATAC-seq[18–20]. These untranscribed aCREs are not enriched in heritability for most traits, in contrast to the transcribed aCREs which showed significant heritability enrichment (Fig. 6b, *top row, left panel*, FDR <0.05). The enrichment levels are dependent on the level of transcription, particularly in distal aCREs, where only ~15% of which showed high evidence of transcription and are highly enriched in trait heritability (Fig. 6b, *bottom row*). These observations are consistent with the previous reports[2,24] and highlight the importance of considering the evidence of transcription to identify active enhancers.

We next examined the enrichment of heritability in cell-type specific CREs, which may be used to identify trait relevant cell-types (Fig. 6c,d; see Methods). As expected, immune cell-type specific CREs are not enriched in heritability of psychiatric and metabolic traits. Also, monocyte count heritability is enriched in monocyte specific CREs and leukocyte count heritability is enriched in CREs specific to most cell-types (Fig. 6c, *hematologic panel, solid dots*, FDR <0.05). Investigating the heritability of immunologic disorders, we found consistent and significant enrichment of T-cell, B-cell and NK cell-specific CREs in most disorders (FDR<0.05), recapitulating the general relevance of lymphoid cells in these disorders[49]. While the sensitivities of tCRE and aCRE in detection of heritability enrichments are generally comparable in most diseases (Fig.6c, *solid dots*), we observed a slightly higher sensitivity in aCRE in some diseases, such as SLE and rheumatoid arthritis. Next we compared the extent of cell-type specific enrichment of heritability[46] in tCRE and aCRE as a metric to prioritize cell-type relevance for each trait (Fig. 6d). We found an overall consistent cell-type ranking between tCRE and aCRE (mean Pearson's r=0.61). Particularly, in Eczema with Pearson's r of 0.90, both tCRE and aCRE consistently ranked CD4+ T-cells as the most relevant cell-type, recapitulating the pivotal roles of Type-1 and -2 immune responses in skin inflammation[50]. We have also performed the same analyses for stimulation-responsive CREs in various cell-types, with similar conclusions (Supplementary Fig. 12). In summary, our data demonstrates the usability of tCRE in the identification and prioritization of trait relevant cell-types, which is comparable to that of aCREs.

Functional annotation disease-associated variants using tCRE

Lastly, we compared the use of tCRE and aCRE in functional annotation of disease-associated variants by linking to their target genes in relevant cell-types (see Methods). Using tCREs, on average ~41% of the trait-associated loci could be connected to a relevant cell-type specific CRE, compared to ~68% by aCRE (Fig. 6e). In addition, we found the number of genes associated by distal CRE is on average ~4.5 times lower in tCRE than aCRE. Since the total number of distal aCRE (n=129,679) is much larger than distal tCRE (n=26,266), the higher number of genes associated by distal aCREs is not surprising. However, given the lack of heritability enrichment in distal aCRE with no (62%) or low (23%) transcription evidence (Fig. 6b), as well as the generally lower PCHi-C validation rate of aCRE co-activity links (Fig. 5c), the relevance of the genes associated by these untranscribed distal aCREs remains elusive, despite the high number. To this end, we highlighted an example gene, Prostaglandin E2 receptor 4 (*PTERG4*), located in proximity to the linkage disequilibrium (LD) block associated with multiple sclerosis, allergy, asthma, Crohn's disease and ulcerative colitis (Fig. 6f). We found a cluster of distal tCREs within these LD blocks (Supplementary Fig. 13), overlapping with multiple trait-associated variants and are linked by co-activity to the proximal tCREs of *PTERG4* (Fig. 6f). Finally, both distal and proximal tCREs of *PTERG4* are highly enriched in T-cells, agreeing with the pivotal roles of T-cells in autoimmune disorders[51] (Fig. 6g). These findings are consistent with a previous report demonstrating that this distal CREs found in Crohn's disease risk locus might regulate the expression of *PTGER4*[52]. In summary, these observations demonstrate the usability of single-cell tCRE activities in functional annotation of trait-associated variants with epigenomic and cellular contexts.

Discussion

Here we outlined an analysis framework using sc-end5-seq data to define tCRE in single cells, for interrogating gene regulation and disease heritability with cell-type specific contexts. Compared to accessibility data which is close-to-binary in nature[17], transcription data is quantitative[23] and has a wider dynamic range. This might explain the higher accuracy in prediction of CRE interactions by co-activity in tCRE (Fig. 5e). In addition, the dynamic nature of transcriptome might better capture the fine granularities of gene regulation during rapid cell-state changes, which is reflected in the detection of shifts in alternative promoter usage by transcription data, but not by accessibility data (Fig. 5h). The lack of heritability enrichment in untranscribed aCREs (Fig. 6b), as well as the higher levels of heritability enrichment in distal tCRE (Fig. 6a), also highlight the importance of considering the evidence of transcription to identify active and biologically relevant CREs. Although we demonstrated that sc-end5-seq methods can detect enhancer RNAs (Fig. 4h), the high level of dropouts (due to their low abundance) renders the analyses of enhancer RNAs in single cells challenging. One might partially alleviate the problem by pooling data from multiple cells (as meta-cells) for downstream analyses. Alternatively, constructing the sc-end5-seq libraries with nuclei instead of whole cells[53] or targeted capturing of a subset of enhancer RNAs[54], should enrich enhancer RNAs in the library to improve dropouts. Currently, most datasets generated on the Chromium™ platform are from sc-end3-dT, while the sc-end5-dT method is used only when T- or B-cell receptor repertoire is a matter of concern. Although it is well-known that sc-end5-seq data can theoretically detect CRE activity with no extra cost, the lack of dedicated tools for data analyses, in particular *de novo* CRE discovery, prevented the wider adoption of this analysis framework. Here we developed *SCAFE* for dedicated analyses of tCREs (Supplementary Fig. 4) and we anticipate wide applications of sc-end5-seq methods along with this tool in the future for interrogating CREs in single cells.

Data availability

Data from this study have been submitted to ENA (Accession: ######). This data may be viewed on the Zenbu genome browser at http://fantom.gsc.riken.jp/zenbu/gLyphs/#config=sc_tCRE_methods

Code availability

The *SCAFE* tool for processing 5′end RNA-seq data is available at https://github.com/chung-lab/scafe

Methods

Human ethics

All human samples examined in this study were either exempted material or were obtained with informed consent and covered under the research protocol (no. H30-9) approved by the ethics committees of the RIKEN Yokohama Campus.

Genome version and gene models

Human genome assembly version hg19 and gene models from GENCODE[55] version v32lift37 were used in all analyses of this study, unless otherwise stated.

Preparing DMFB and iPSC samples

DMFB from neonatal foreskin were purchased (Lonza®). Cells were cultured in Gibco Dulbecco's Modified Eagle Medium (DMEM, high glucose with L-glutamine) supplemented with 10% Fetal bovine serum (FBS) and penicillin/streptomycin. Cells were dissociated with trypsin 0.25% Ethylenediaminetetraacetic acid (EDTA) for 5 minutes (mins) at 37°C and washed twice in 0.04% Bovine serum albumin (BSA) in Phosphate-buffered saline (PBS). iPSC[56] were cultured in StemFit™ medium (Reprocell®) under feeder-free conditions at 37°C in a 5% CO2 incubator. The cells were plated on a culture dish pre-coated with iMatrix-511™ (Nippi®). Rock inhibitor (FUJIFILM Wako®) was added to the cells at a final concentration of 10µM during the first day of culturing. StemFit™ medium is refreshed daily until harvesting. The cells were dissociated and detached by incubating with TrypLE™ Select (Thermo Fisher®) followed by scrapping in StemFit™ medium. The cells were spin down and washed with 0.04% BSA in PBS twice.

Preparing PBMC samples

Human PBMCs were prepared from whole blood of a male healthy donor with Leucosep™ (Greiner®). Isolated 2×10^6 PBMC cells were incubated with PMA/ionomycin (i.e. stimulated) (Cell Activation Cocktail with Brefeldin A, Biolegend®), or DMSO as control (i.e. resting), for six hours.

Isolating cytoplasmic, nucleoplasmic, and chromatin-bound RNAs

Cell fractionation was carried out according to a previous study[57]. Briefly, cells grown to ~90% confluency in 10cm dishes were collected by trypsinization and washed once in PBS. The cells were lysed in lysis buffer, followed by separation of the nucleus from the cytoplasmic material by centrifugation in a sucrose cushion. The isolated nucleus was rinsed once in PBS-EDTA and lysed by adding glycerol buffer and urea buffer in equal volumes. The precipitate, which contained the chromatin-RNA complex, was isolated by centrifugation and washed once in PBS-EDTA. RNA from each of the three subcellular compartments was isolated by Trizol^{TM} (Thermo Fisher$^{®}$).

Bulk CAGE, RNA-seq and ATAC-seq library construction and sequencing for DMFB and iPSC

Bulk CAGE libraries were generated by the nAnT-iCAGE[58] method as previously described and sequenced on HiSeq^{TM} 2500 (Illumina$^{®}$) as 50bp single-end reads. Bulk RNA-seq libraries was generated as previously described[2] and sequenced on HiSeq^{TM} 2500 (Illumina$^{®}$) as 100bp paired-end reads. Bulk ATAC-seq was performed as previously described[59] with slight modifications. Briefly, 2.5×10^4 cells/ml were used for library preparation. Due to the more resistant membrane properties of DMFB, 0.25% IGEPAL^{TM} CA-630 (Sigma-Aldrich$^{®}$) were used for cell lysis. Transposase reaction was carried out as described in the protocol followed by 10 to 12 cycles of PCR amplification. Amplified DNA fragments were purified with MinElute^{TM} PCR Purification Kit (QIAGEN$^{®}$) and size-selected with AMPure^{TM} XP (Beckman Coulter$^{®}$). All libraries were examined in Bioanalyzer^{TM} (Agilent$^{®}$) for size profiles and quantified by KAPA^{TM} Library Quantification Kits (Kapa Biosystems$^{®}$). Bulk ATAC-seq libraries were sequenced on HiSeq^{TM} 2500 (Illumina$^{®}$) as 50bp paired-end reads.

sc-RNA-seq library construction and sequencing for DMFB and iPSC

Freshly prepared iPSC and DMFB cells were loaded onto the Chromium^{TM} Controller (10x Genomics$^{®}$) on different days. Cell number and viability were measured by Countess^{TM} II Automated Cell Counter (Thermo Fisher$^{®}$). Final cell density was adjusted to 1.0×10^6cells/ml with >95% viability. Both cells were targeting ~5,000 cells per reaction. For sc-end3-dT libraries, we used Chromium^{TM} Single Cell 3′ Library kit v2 (10x Genomics$^{®}$). Briefly, single cell suspensions were mixed with the Single cell Master Mix using Reverse transcription (RT) Primer (AAGCAGTGGTATCAACGCAGAGTACATr–GrGrG) and loaded together with 3′ gel beads and partitioning oil into a Single Cell A Chips according to the manufacturer's instructions (10x Genomics$^{®}$). For sc-end5-dT and sc-end5-rand libraries, used Single Cell 5′ Library kit v1.1 (10x Genomics$^{®}$). Single cell suspension was mixed with Single cell Master Mix using oligo(dT) RT primer (AAGCAGTGGTATCAACGCAGAGTACGAGAC–T(30)–VN) or random hexamer RT primer (AAGCAGTGGTATCAACGCAGAGTACNNNNNN) and loaded together with 5′ gel beads and partitioning oil into a Single Cell A Chips according to the manufacturer's instructions. RNAs within single cells were uniquely barcoded and reverse transcribed within droplets. Both methods used Veriti^{TM} Thermal Cycler (Applied Biosystems$^{®}$) for RT reaction. After collecting cDNAs prepared from each method, they were amplified using cDNA primer mix from the kit, followed by the standard steps according to manufacturer's instructions. For iPSC and DMFB, six libraries (i.e. 3 methods × 2 cell lines) were barcoded by different indexes from i7 sample index plate (10x Genomics$^{®}$). The libraries were examined in Bioanalyzer^{TM} (Agilent$^{®}$) for size profiles and quantified by KAPA^{TM} Library Quantification Kits (Kapa Biosystems$^{®}$). All libraries were sequenced on HiSeq^{TM} 2500 (Illumina$^{®}$) as 75 bp paired-end reads.

sc-end5-dT and sc-ATAC-seq library construction and sequencing for PBMC

Freshly prepared resting and stimulated PBMCs were subjected to sc-end5-dT (Single Cell 5′ Library kit v1.1) and sc-ATAC-seq (Single Cell ATAC kit v1.1) library construction on the same day using the Chromium^{TM} platform according to manufacturer’s instructions (10x Genomics$^{®}$). About 5,000 cells/nuclei were targeted per reaction. sc-end5-dT and sc-ATAC-seq libraries were sequenced on HiSeq^{TM} 2500 (Illumina$^{®}$) as 75bp and 100bp paired-end reads respectively.

Processing cell line bulk RNA-seq and CAGE data

Reads were aligned to hg19 with *hisat2 v2.0.4*[60]. For each sample, the first aligned base at the 5’end of read 1 was piled up to a ctss (capped TSS) bed file using custom *perl* scripts. The ctss bed files were used for down sampling, feature intersection and counting.

Processing of FANTOM6 bulk-CAGE data for DMFB

Publicly available bulk-CAGE dataset on DMFB (n=1,163) were obtained[48]. Alignment bam files (on hg38) were converted to ctss files as described above and lifted over to hg19 using *liftover* (http://genome.ucsc.edu). All ctss files were pooled and subsampled to various depths. These subsampled ctss files were processed in the *SCAFE* workflows for *de novo* definition of TSS clusters and calculation of their logistic probabilities as described below.

Processing of bulk ATAC-seq data

The bulk ATAC-seq data for DMFB and iPSC were processed using pipelines developed by the ENCODE consortium (https://github.com/kundajelab/atac_dnase_pipelines). The –log(P) signal tracks for pooled replicates were used to defined gold-standards for training of the TSS classifiers.

Processing of cell line sc-RNA-seq data

Reads were aligned to hg19 with *Cellranger*, and bam files were processed with *SCAFE* to generate filtered ctss files and *de novo* define tCRE. Annotation counts were produced by intersecting ctss files with GENCODE gene models. Metagene plots from overlapping ctss files with exons binned with Bioconductor *equisplit* using *foverlaps*. Enrichment of genesets in sc-end5-dT versus sc-end5-rand was tested using *fgsea v1.16.0*[61] with nperm = 1000. Genesets were defined as: 1) cytoplasmic, nucleoplasmic, and chromatin-bound RNAs: $\log_2$ fold change ≥2 in fractionated CAGE compared to total CAGE, 2) long and short RNAs: maximum transcript length per gene ≥25,000nt and <1,000nt, 3) Non-polyA histone RNAs: histone RNAs with $\log_2$ fold-change ≥2 in non-polyA fraction in a previous study[30].

Processing of PBMC sc-end5-dT data

Reads were aligned to hg19 with *Cellranger* and then processed with *Seurat v3*[62]. Cells were excluded with ≥4 median absolute deviation from the mean for number of features, UMI count, and percentage of mitochondrial UMI. Top 2,000 variable features were selected. Resting and stimulated PBMC samples were integrated with *Suerat CCA* using principal component (PC) 1 to 20 based on gene-based expression matirx. Bam files were processed with *SCAFE* to generate filtered ctss files and *de novo* define tCRE. tCRE matrices from *SCAFE* were added to the *Seurat* object for downstream analysis. Cell annotation was performed combining annotation from *scMatch (version GitHub master at 2020-10-10)*[63] and known marker genes. sc-end5-dT cell-type specific markers and stimulation specific markers were defined with modified *Seurat FindMarkers* to return all results (min.pct = 0, return.thresh = Inf,logfc.threshold = 0,min.cells.group = 0).

Processing of PBMC sc-ATAC-seq

PBMC sc-ATAC resting and stimulated cells were processed with *SnapATAC v1.0.0*[64] with default parameters, selecting cells with ≥40% reads in peaks. Integrated with *Harmony v1.0*[65] using PC 1 to 20. sc-ATAC and sc-end5-dT were integrated using *SnapATAC FindTransferAnchors* and *TransferData* functions to transfer cell cluster annotations to the sc-ATAC-seq cells. sc-ATAC-seq peaks were defined per cell-type using *SnapATAC runMACS*, then merged. Cell-type specific markers and stimulation-specific markers were defined with *SnapATAC findDAR*.

Estimating TF Motif activity

ChromVAR v1.12.0[43] was used to calculate per cell TF motif activities for the JASPAR2018[66] core motif set for tCRE or aCRE excluding chrM. The tCRE matrix was binarized prior to running. Fisher's exact tests and correlations of the top 80 motifs by *ChromVAR* deviation score per cell-type were used in Supplementary Fig 9.

Predicting CRE interaction by co-activity

Cicero v1.3.4.11[16] was used for tCRE and aCRE present in 3 or more cells (all cells included, and separately subset to each cell-type) following default parameters. For comparisons between tCREs and aCREs, only a subset of CRE that are overlapped between tCREs and aCREs were used. We also excluded CREs pairs located within 10kb. A pair of CRE with co-activity score ≥0.2 is defined as "linked". PCHi-C connections[44] (without cutoffs) from all cell-types were pooled and used for validation of co-activity linked CREs pairs.

Detecting shifts in alternative promoter usage

For each cell type cluster (excluding dendritic cells due to low cell count), knn clustering of the *Seurat* SNN matrix (k=50) was used to generate metacells. The proportion of each genes UMI arising from proximal tCREs was calculated for each metacell. Cell type specific tCRE switching events were identified using a t-test for differences in the proportion of gene UMI contributed from each tCRE between metacells of selected clusters and a background of all other clusters. ATAC-seq signal at a tCRE was defined as the maximum signal in cluster specific bigwig files generated with *SnapATAC runMACS*.

Removal of strand invasion artifacts

Strand invasion artifacts, i.e. strand invaders, can be identified based on complementarity of genomic sequence upstream of the mapped reads to TS oligo sequence, according to a study[35]. Briefly, we extracted a 13nt genomic sequence immediately upstream of the 5′end of mapped reads, then globally aligned to the TS oligo sequence (TTTCTTATATGGG) and calculated the edit distance. A read is considered as an artifact of strand invasion when 1) the edit distance ≤ 5 and two of the three nucleotides immediately upstream were guanosines (Supplementary Fig.5), based on the previously proposed thresholds[35].

Identifying unencoded G

Previous studies suggest most reads derived from capped RNAs begin with an unencoded "G", which can be used to distinguish genuine TSS from artifacts[26,67]. To precisely calculate the number of unencoded G for each mapped read, we first identify the junction between TS oligo and cDNA sequence and then examine the cDNA 5′end. Specifically, to precisely locate the TS oligo-cDNA junction, we considered only the reads 1) containing the last 5nt (i.e. 3′end) of TS oligo sequence (i.e. ATGGG) with maximum one mismatch, 2) starting with a softclip region (i.e. "S" in CIGAR string[68]) of ± 50% of the TS oligo sequence length (i.e. 6 to 20nt), 3) with a match region ≥ 5nt (i.e. "M" in CIGAR string) following the softclip region. The 5′end of cDNA was defined as the first nucleotide immediately following the last nucleotide of the TS oligo sequence. The first 3nt of cDNA sequence was compared to the genomic sequence at their corresponding aligned position, and the number of Gs that are mismatched was defined as the number of unencoded G for the examined read.

Defining TSS clusters and their properties

The 5′ positions of reads (i.e. TSS) in *Cellranger* alignment *bam* files were extracted, piled-up by UMI, and clustered using *Paraclu* [69] using default parameters. Only TSS clusters with total UMI ≥5 and summit UMI ≥ 3 were retained. The following properties were extracted for each TSS cluster: 1) cluster count, 2) summit count, 3) flank count, 4) corrected expression and 5) unencoded G percentage. Cluster, summit and flank count refers to UMI counts within the cluster, at its summit, and within a region flanking its summit (±75nt). Corrected expression refers to an expression value relative to its local background, based on the assumption that the level of exon painting artefact is positively correlated with the transcript abundance. Specifically, if the summit of a TSS cluster is located within genic regions, it will be assigned to either exon or intron, in either sense or antisense strand of the corresponding gene, or otherwise assigned to intergenic, as its local background. All annotated TSS regions (±250nt) were masked from these local backgrounds. The density of UMI per nucleotide within each local background is calculated (i.e. local background density). The corrected expression of a TSS cluster is calculated as the ratio of the density of UMI within the region flanking its summit (±75nt) to the density of its local background. Unencoded G percentage refers to the percentage of UMI within the cluster that has ≥1 unencoded G.

Building a TSS classifier

To combine the five properties into a single classifier, we used multiple logistic regression implemented in the *caret*[70] R package. For training of this classifier, we defined a set of "gold standard" TSS clusters based on their ATAC-seq signal (as averaged –log*P* within TSS cluster). Specifically, the top and bottom 5% of TSS clusters, ranked by their ATAC-seq signal, were defined as positive and negative gold standards, and used for training of the logistic models at 5-fold cross-validation. The resulting logistic probability was used as the TSS classifier. The performance of this TSS classifier, as well as its constituent metrics, is measured as AUC, using the top and bottom 10% of TSS clusters as positive and negative gold standards for testing. The default cutoff of logistic probability at 0.5 is defined as the default threshold. All the TSS clusters in this study are filtered with this default cutoff. In the PBMC datasets, corresponding sc-ATAC-seq datasets were used for training and an additional lenient logistic probability cutoff of 0.028 was also used, which corresponds to a specificity of 0.5.

Defining tCRE and aCRE

tCREs are defined by merging closely located TSS clusters. Briefly, TSS clusters located within ±500nt of annotated gene TSS were classified as proximal, or as distal otherwise. All TSS clusters were then extended 400nt upstream and 100nt downstream. These extended ranges were merged using *bedtools*[71], in a strand-specific manner for proximal TSS clusters and non-strand-specific manner for distal TSS clusters, as proximal and distal tCRE respectively. Distal tCRE were then assigned to either exonic, intronic or intergenic, in this order. aCREs are defined by the ATAC peak ranges output from *SnapATAC*. aCREs located within ±500nt of annotated gene TSS were classified as proximal, or as distal otherwise.

Developing *SCAFE* tools

SCAFE (Single Cell Analysis of Five-prime Ends) consists of a set of command-line tools written in *perl* and *R* programming languages, providing an end-to-end solution for processing of sc-end5-seq data. Briefly, it takes the read alignment file (bam), maps the cDNA 5'ends, identifies genuine TSS clusters, defines tCREs, annotated tCREs to gene models, quantify their expression and predicts tCRE interactions by co-activity. The tools in *SCAFE* can be ran individually as independent tools or ran serially as predefined workflows. For details please visit: https://github.com/chung-lab/scafe

Processing of GWAS data

For heritability enrichment, GWAS summary statistics were obtained from (1) UK biobank heritability browser (https://nealelab.github.io/UKBB_ldsc/index.html), (2) Dr. Alkes Price group site (https://alkesgroup.broadinstitute.org/) and (3) Japanese encyclopedia of genetic associations (JENGER, http://jenger.riken.jp/). Summary statistics obtained from (1) and (2) were directly used for heritability enrichment analyses, while the summary statistics obtained from (3) were pre-processed using *"munge_sumstats.py"* scripts in *LDSC* software[72]. For linking trait associated variants to candidate genes, lead variants ($P < 5\times10^{-8}$) were obtained from (1) GWASdb[73] (as of 19th August 2015, http://jjwanglab.org/gwasdb) and (2) NHGRI-EBI GWAS Catalog[74] (release r2020-07-15). The variants within the LD block of these lead variants (i.e. proxy variants) were searched for using *PLINK v1.9*[75] with an r2 ≥0.5 within ±500kb in matched population panels of Phase 3 1000 Genomes Project downloaded from MAGMA website[76] (http://ctg.cncr.nl/software/MAGMA/ref_data/). The final set trait-associated variants contain 158,745 variants for 10 immune disorders and 2 blood traits.

Estimating enrichment of trait heritability

Enrichment of trait heritability in tCRE (or aCRE) was assessed by stratified LD score regression (S-LDSC) implemented in *LDSC* software. Briefly, sets of tCRE (or aCRE) were defined based on their proximity to annotated TSS (i.e. all, proximal or distal). Additional sets of tCREs were generated based on a more lenient logistic probability cutoff as mentioned above. Additional sets of aCREs were generated based on evidence of transcription (i.e. number of UMI from RNA reads). Annotation files and LD score files were generated for each set of tCRE (or aCRE) using the "*make_annot.py*" and "*ldsc.py*" scripts using default parameters. Each set of tCRE (or aCRE) was added onto the 97 annotations of the baseline-LD model v2.2 and heritability enrichment (i.e. ratio of proportion of heritability to proportion of SNP) for each trait was estimated using the "*ldsc.py*" script with *"--h2"* flag in default parameters.

Evaluating cell-type specificity of trait heritability

Cell-type specificity of trait heritability was assessed by LD score regression for specifically expressed genes (LDSC-SEG) implemented in *LDSC* software[77]. Briefly, enrichment of each tCRE (or aCRE) in each cell type were calculated using *findDAR* implemented in *SnapATAC* and *FindMarkers* in *Seurat*, respectively. Sets of "cell-type specific" tCRE (or aCRE) were defined as the top 20% of tCRE (or aCRE) ranked by the enrichment P for each cell type. A set of "core" tCRE (or aCRE) was defined as all tCREs (or aCREs) that are not "cell-type specific" to any of the cell types. Annotation files and LD score files were generated for each set of "cell-type specific" and "core" tCREs (or aCREs) using the "*make_annot.py*" and "*ldsc.py*" scripts using default parameters. For each cell type, sets of "cell-type specific" and "core" tCRE (or aCRE) were added onto the 53 annotations of baseline-LD model v1.2 and the contribution of "cell-type specific" tCRE (or aCRE) to trait heritability (i.e. regression coefficient) for each trait was estimated using the "*ldsc.py*" script with *"--h2-cts"* flag in default parameters.

Connecting trait-associated variants to candidate genes

Trait associated variants were defined as mentioned above. A tCRE (or aCRE) is associated with a trait if it overlaps at least one trait-associated variant. A gene is associated with a trait when its proximal tCRE (or aCRE) is associated with a trait, or a distal tCRE (or aCRE) is associated with a trait and connected to its proximal tCRE by co-activity score ≥0.2.

Zenbu genome browser

Most datasets in this study can be visualized in Zenbu genome browser. The Zenbu genome browser features on-the-fly demultiplexing single-cell or cell-type signals. Thus, single-nucleotide resolution signal within each single cell could be convenient interrogated. For details please visit: https://fantom.gsc.riken.jp/zenbu/gLyphs/#config=sc_tCRE_methods

Data visualization and statistics

We used R (https://www.r-project.org/) and the *ggplot2* R package[78] unless otherwise noted for visualizations.

Acknowledgements

This research was supported by research grant to RIKEN Center for Integrative Medical Sciences from Ministry of Education, Culture, Sports, Science and Technology (MEXT). We would like to extend our thanks to GeNAS sequencing platform in IMS.

Conflict of Interest

None

Author contributions

CCH, JWS, PC conceived the project and supervised the research. TK optimized experiments and constructed single cell libraries. JM and CCH analyzed most of the data. JM, TK, JWS, CCH wrote the manuscript. JCC processed the bulk-ATAC data. CWY processed the DMFB bulk CAGE data. CT assisted the heritability enrichment analysis. AS, KY performed the PBMC stimulation experiments. YS performed cell fractionated bulk RNA experiments. FLR performed bulk-ATAC-seq experiments. YA supported the logistics of sample collection.

References

1. Forrest, A. R. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
2. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
3. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190–1195 (2012).
4. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.* **16**, 144–154 (2015).
5. Catarino, R. R. & Stark, A. Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes Dev.* **32**, 202–223 (2018).

6. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300–307 (2021).

7. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).

8. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160 (2016).

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

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

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

12. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).

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

14. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116.e20 (2020).

15. Granja, J. M. *et al.* Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat. Biotechnol.* **37**, 1458–1465 (2019).

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

17. Xiong, L. *et al.* SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat. Commun.* **10**, 4576 (2019).

18. Thibodeau, A., Uyar, A., Khetan, S., Stitzel, M. L. & Ucar, D. A neural network based model effectively predicts enhancers from clinical ATAC-seq samples. *Sci. Rep.* **8**, 16048 (2018).

19. Kim, T. H. *et al.* Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231–1245 (2007).

20. Pang, B. & Snyder, M. P. Systematic identification of silencers in human cells. *Nat. Genet.* **52**, 254–263 (2020).

21. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 15776–15781 (2003).

22. Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* **48**, 488–496 (2016).

23. Rennie, S., Dalby, M., van Duin, L. & Andersson, R. Transcriptional decomposition reveals active chromatin architectures and cell specific regulatory interactions. *Nat. Commun.* **9**, 487 (2018).

24. Mikhaylichenko, O. *et al.* The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes Dev.* **32**, 42–57 (2018).

25. Kouno, T. *et al.* C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. *Nat. Commun.* **10**, 360 (2019).

26. Cumbie, J. S., Ivanchenko, M. G. & Megraw, M. NanoCAGE-XL and CapFilter: an approach to genome wide identification of high confidence transcription start sites. *BMC Genomics* **16**, 597 (2015).

27. Adiconis, X. *et al.* Comprehensive comparative analysis of 5'-end RNA-sequencing methods. *Nat. Methods* **15**, 505–511 (2018).

28. Cheng, J. *et al.* Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution. *Science* **308**, 1149–1154 (2005).

29. Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nat. Methods* **3**, 211–222 (2006).

30. Yang, L., Duff, M. O., Graveley, B. R., Carmichael, G. G. & Chen, L.-L. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol.* **12**, R16 (2011).

31. Hirabayashi, S. *et al.* NET-CAGE characterizes the dynamics and topology of human transcribed cis-regulatory elements. *Nat. Genet.* **51**, 1369–1379 (2019).

32. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

33. Gaidatzis, D., Burger, L., Florescu, M. & Stadler, M. B. Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat. Biotechnol.* **33**, 722–729 (2015).

34. 10x GENOMICS. Technical Note – Interpreting Intronic and Antisense Reads in Single Cell Gene Expression Data. support.10xgenomics.com/permalink/3ItKYUsoESnDpnFNnfgvNT.

35. Cvetesic, N. *et al.* SLIC-CAGE: high-resolution transcription start site mapping using nanogram-levels of total RNA. *Genome Res.* **28**, 1943–1956 (2018).

36. Kanamori-Katayama, M. *et al.* Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res.* **21**, 1150–1159 (2011).

37. Affymetrix ENCODE Transcriptome Project & Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* **457**, 1028–1032 (2009).

38. Cheng, J. *et al.* A role for H3K4 monomethylation in gene repression and partitioning of chromatin readers. *Mol. Cell* **53**, 979–992 (2014).

39. Bae, S. & Lesch, B. J. H3K4me1 Distribution Predicts Transcription State and Poising at Promoters. *Front. Cell Dev. Biol.* **8**, 289 (2020).

40. Preker, P. *et al.* PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. *Nucleic Acids Res.* **39**, 7179–7193 (2011).

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

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

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

44. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384.e19 (2016).

45. Zhang, P. *et al.* Relatively frequent switching of transcription start sites during cerebellar development. *BMC Genomics* **18**, 461 (2017).

46. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

47. Ni, G., Moser, G., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Wray, N. R. & Lee, S. H. Estimation of Genetic Correlation via Linkage Disequilibrium Score Regression and Genomic Restricted Maximum Likelihood. *Am. J. Hum. Genet.* **102**, 1185–1194 (2018).

48. Ramilowski, J. A. *et al.* Functional annotation of human long noncoding RNAs via molecular phenotyping. *Genome Res.* **30**, 1060–1072 (2020).

49. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).

50. Akdis, M. *et al.* T helper (Th) 2 predominance in atopic diseases is due to preferential apoptosis of circulating memory/effector Th1 cells. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **17**, 1026–1035 (2003).

51. Skapenko, A., Leipe, J., Lipsky, P. E. & Schulze-Koops, H. The role of the T cell in autoimmune inflammation. *Arthritis Res. Ther.* **7**, S4 (2005).

52. Libioulle, C. *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.* **3**, e58 (2007).

53. Grindberg, R. V. *et al.* RNA-sequencing from single nuclei. *Proc. Natl. Acad. Sci.* **110**, 19802–19807 (2013).

54. Mercer, T. R. *et al.* Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.* **30**, 99–104 (2012).

55. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

56. Fort, A. *et al.* Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat. Genet.* **46**, 558–566 (2014).

57. Conrad, T. & Ørom, U. A. Cellular Fractionation and Isolation of Chromatin-Associated RNA. *Methods Mol. Biol. Clifton NJ* **1468**, 1–9 (2017).

58. Murata, M. *et al.* Detecting expressed genes using CAGE. *Methods Mol. Biol. Clifton NJ* **1164**, 67–85 (2014).

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

60. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

61. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* 060012 (2021) doi:10.1101/060012.

62. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

63. Hou, R., Denisenko, E. & Forrest, A. R. R. scMatch: a single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics* **35**, 4688–4695 (2019).

64. Fang, R. *et al.* SnapATAC: A Comprehensive Analysis Package for Single Cell ATAC-seq. *bioRxiv* 615179 (2020) doi:10.1101/615179.

65. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

66. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D260–D266 (2018).

67. Kawaji, H. *et al.* Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Res.* **24**, 708–717 (2014).

68. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).

69. Frith, M. C. *et al.* A code for transcription initiation in mammalian genomes. *Genome Res.* **18**, 1–12 (2008).

70. Kuhn, M. Building Predictive Models in *R* Using the **caret** Package. *J. Stat. Softw.* **28**, (2008).

71. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma. Oxf. Engl.* **26**, 841–842 (2010).

72. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

73. Li, M. J. *et al.* GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* **44**, D869-876 (2016).

74. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

75. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

76. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).

77. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).

78. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).

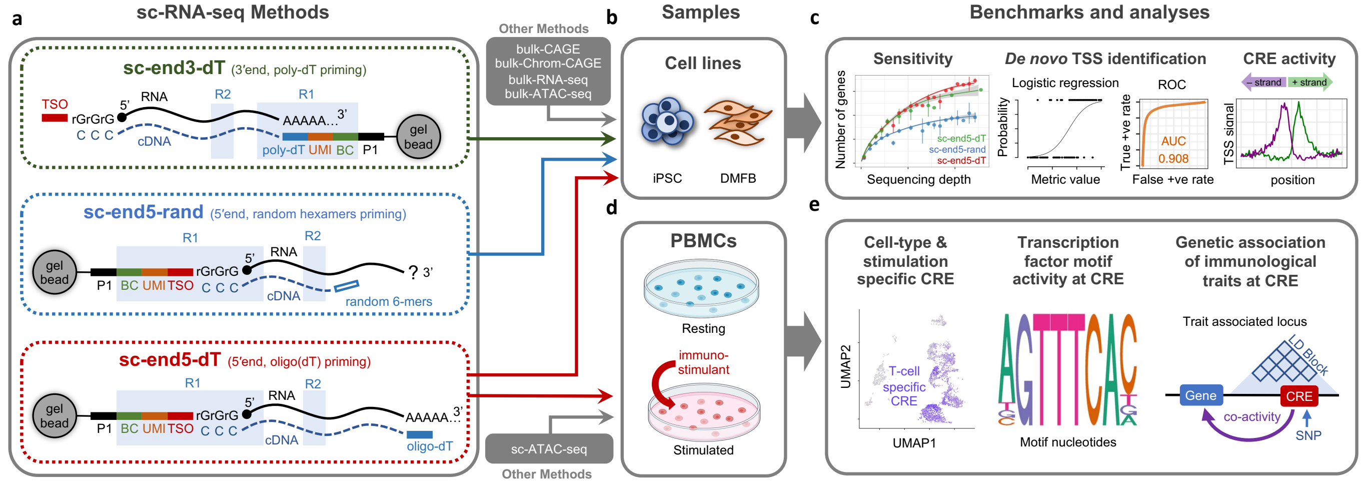


Fig. 1: Overview of the experimental designs and benchmark analysis. **a**, sc-RNA-seq methods used in this study. sc-end5-rand method is a custom method, and the other two methods are original methods on Chromium$^{\mathrm{TM}}$ platform (10x Genomics$^{®}$). (*BC*: cell barcode, *UMI*: unique molecular identifier, *TSO*: template switching oligonucleotide, *R1*: read 1, *R2*: read 2) **b**, two cell lines are used to compare the performance of the three sc-RNA-seq methods, with matched bulk transcriptome and epigenome datasets. **c**, the datasets from **(b)** were used for sensitivity assessment, *de novo* identifying TSS, detecting CRE activity. **d**, PMBCs, at resting and stimulated states, were profiled using sc-ATAC-seq and sc-end5-dT methods. **e**, the datasets from the two methods in **(d)** were compared in terms detection of cell-type/stimulation specific CRE, transcription factor motif activity and genetic association of traits.

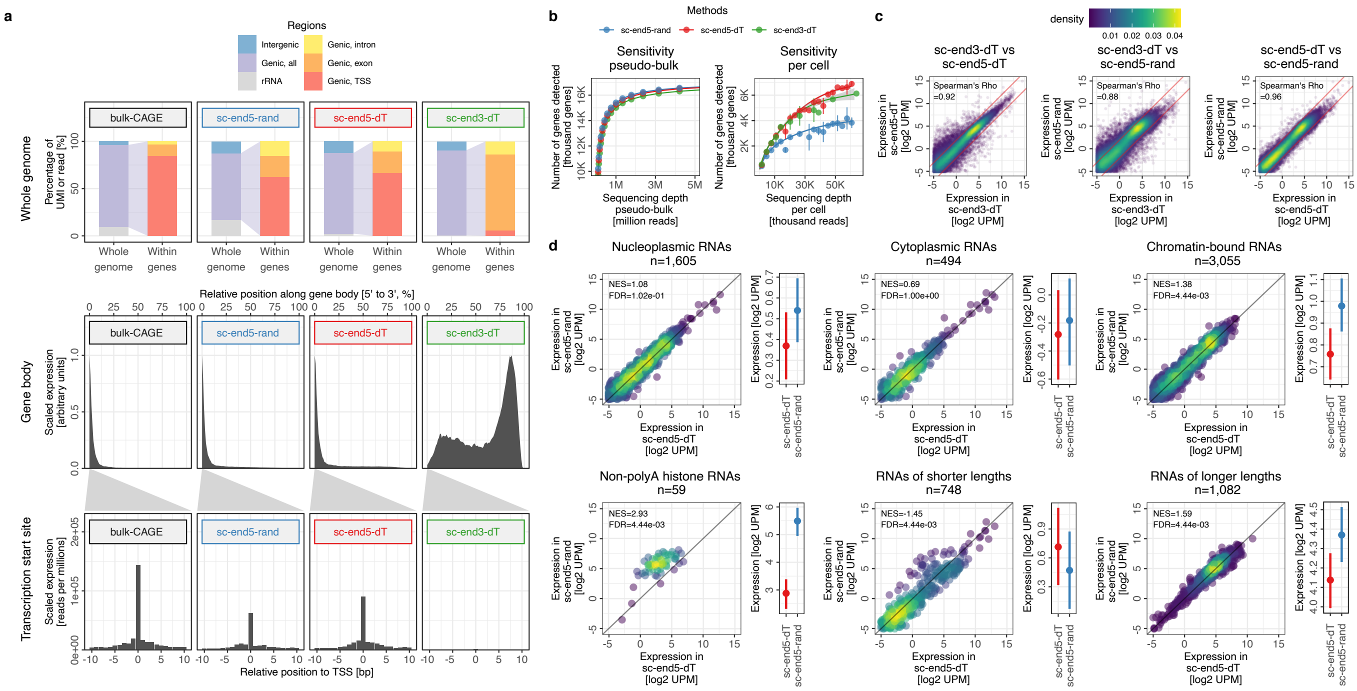


Fig. 2: Performance of sc-RNA-seq methods. a, distribution of reads from bulk-CAGE and sc-RNA-seq methods. *top*, distribution of reads in the whole genome; *middle*, distribution of reads along the gene body; *bottom*, distribution of reads in around annotated TSS. **b**, Sensitivity of gene detection in pseudo-bulk (*left*) and in single cells (*right*) across sequencing depth. Error bars represent standard deviation. The genes that are detected in bulk-RNA-seq were used as the scope. **c**, Correlation of gene expression levels between the pseudo-bulk data of the three sc-RNA-seq methods. *red line*, ±2-fold differences. UPM, UMI per million. Color represents the density of points. **d**, differences in the expression levels of RNAs with various properties between sc-end5-rand and sc-end5-dT. Gene Set Enrichment Analysis (GSEA) was performed on each RNA set; NES and FDR, normalized enrichment score and false discovery rate of GSEA. Color represents the density of points. A positive NES value with FDR <0.05 refers to a significantly higher abundance of an RNA set in sc-end5-rand. (*right*) mean and standard errors.

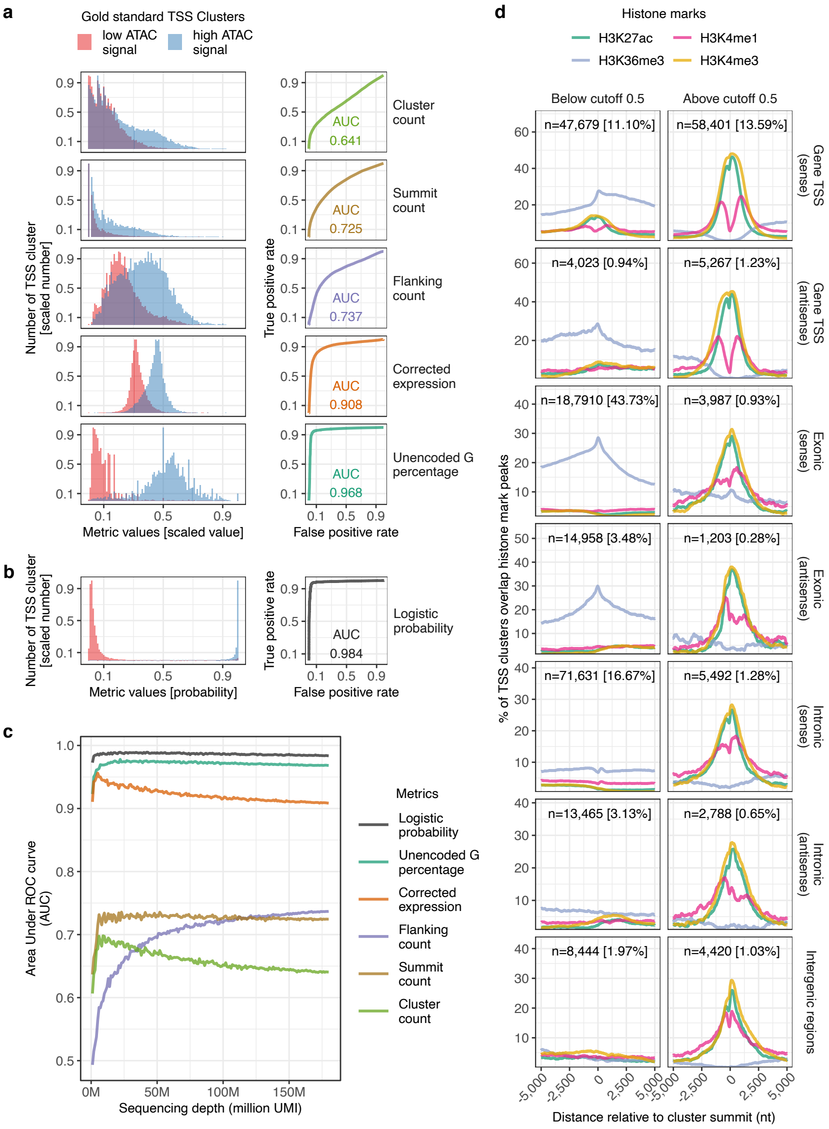


Fig. 3: *De novo* identification of genuine TSS. a, properties of gold-standard TSS clusters (*left*) and their performance as a TSS classifier measured as Area Under Receiver Operating Characteristic (ROC) Curve (AUC) (*right*). **b**, logistic probability of gold-standard TSS clusters (*left*) and its performance as a TSS classifier measured as AUC (*right*). **c**, performance of various metrics as a TSS classifier in (**a**) and (**b**) across various sequencing depth. **d**, histone marks at TSS clusters with logistic probability below (*left*) or above (*right*) 0.5 cutoff, at annotated gene TSS, exonic or intronic regions in sense or antisense orientations, or otherwise intergenic regions. n, number of TSS clusters. %, percentage of TSS clusters in all genomic locations regardless of logistic probability thresholds.

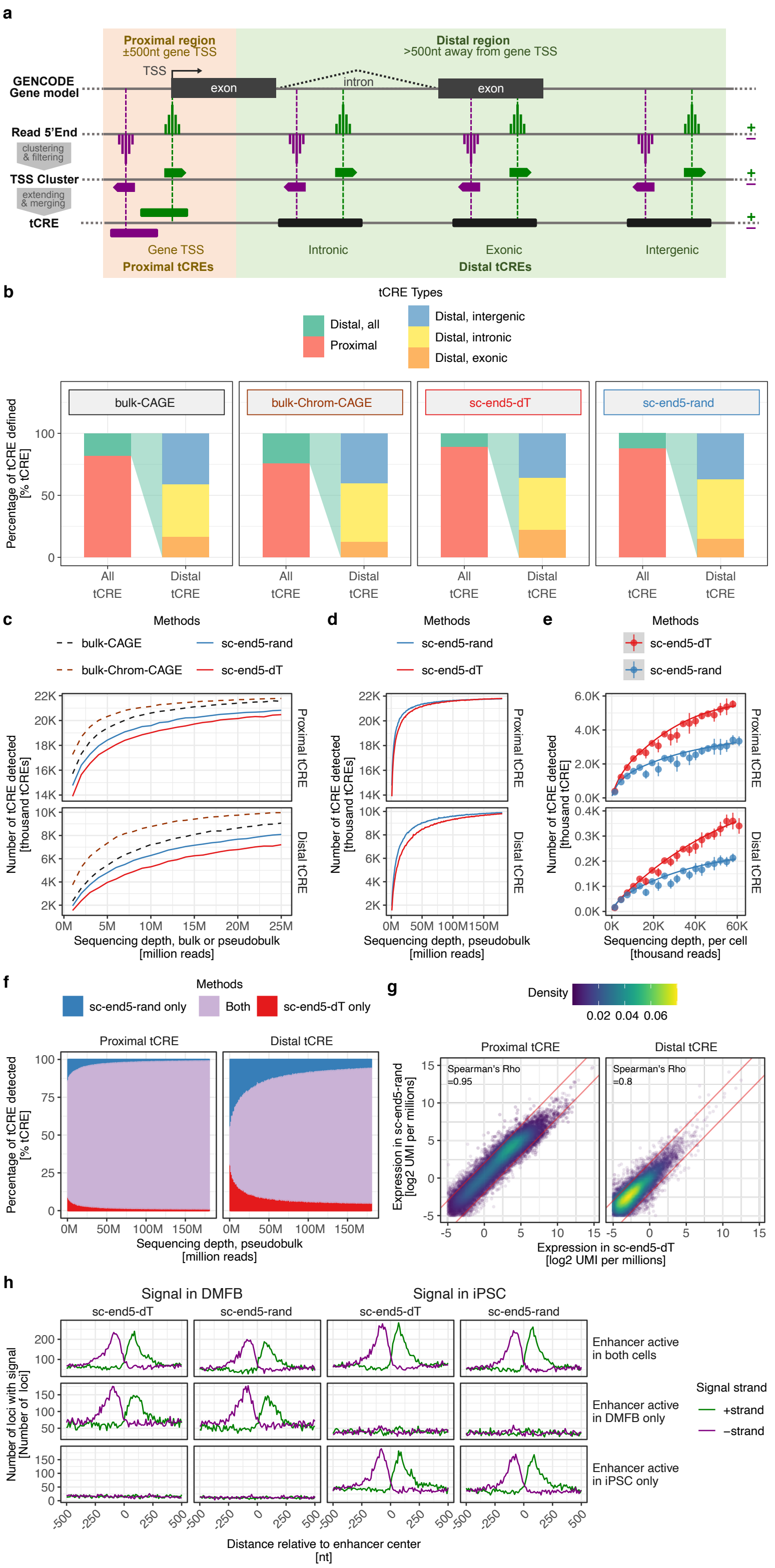


Fig. 4: Definition and properties of tCRE. a, defining tCRE by merging closely located TSS clusters. Distance to gene TSS was used as the criteria to define proximal or distal tCRE. Proximal and distal TSS clusters were merged in stranded and strandless manner, respectively. Distal tCREs are further classified as intronic, exonic, or otherwise intergenic. **b**, proportion of tCREs types defined from sc-end5-dT and sc-end5-rand pseudo-bulk, compared to bulk-CAGE and bulk-Chrom-CAGE. All four libraries were subsampled to 25 million reads. **c**, Sensitivity of tCRE detection in sc-end5-dT and sc-end5-rand pseudo-bulk, compared to bulk-CAGE and bulk-Chrom-CAGE, from 1 to 25 million reads. **d**, sensitivity of tCRE detection in sc-end5-dT and sc-end5-rand pseudo-bulk, from 1 to 150 million reads. **e**, Sensitivity of tCRE detection in sc-end5-dT and sc-end5-rand in single cells, from 1,000 to 60,000 reads per cell. Error bars represent standard deviation. **f**, Proportion of overlap in tCRE detected in sc-end5-seq pseudo-bulk from 1 to 150 million reads. **g**, correlation of tCRE levels between the pseudo-bulk data of the two sc-end5-seq methods. *red line*, ±2-fold differences. UPM, UMI per million. **h**, count of overlapping enhancer loci in pseudo-bulk sc-end5-dT and sc-end5-rand at bidirectional enhancer loci defined in bulk-CAGE, separated into cell-type specificity by overlap with bulk-ATAC-seq peaks.

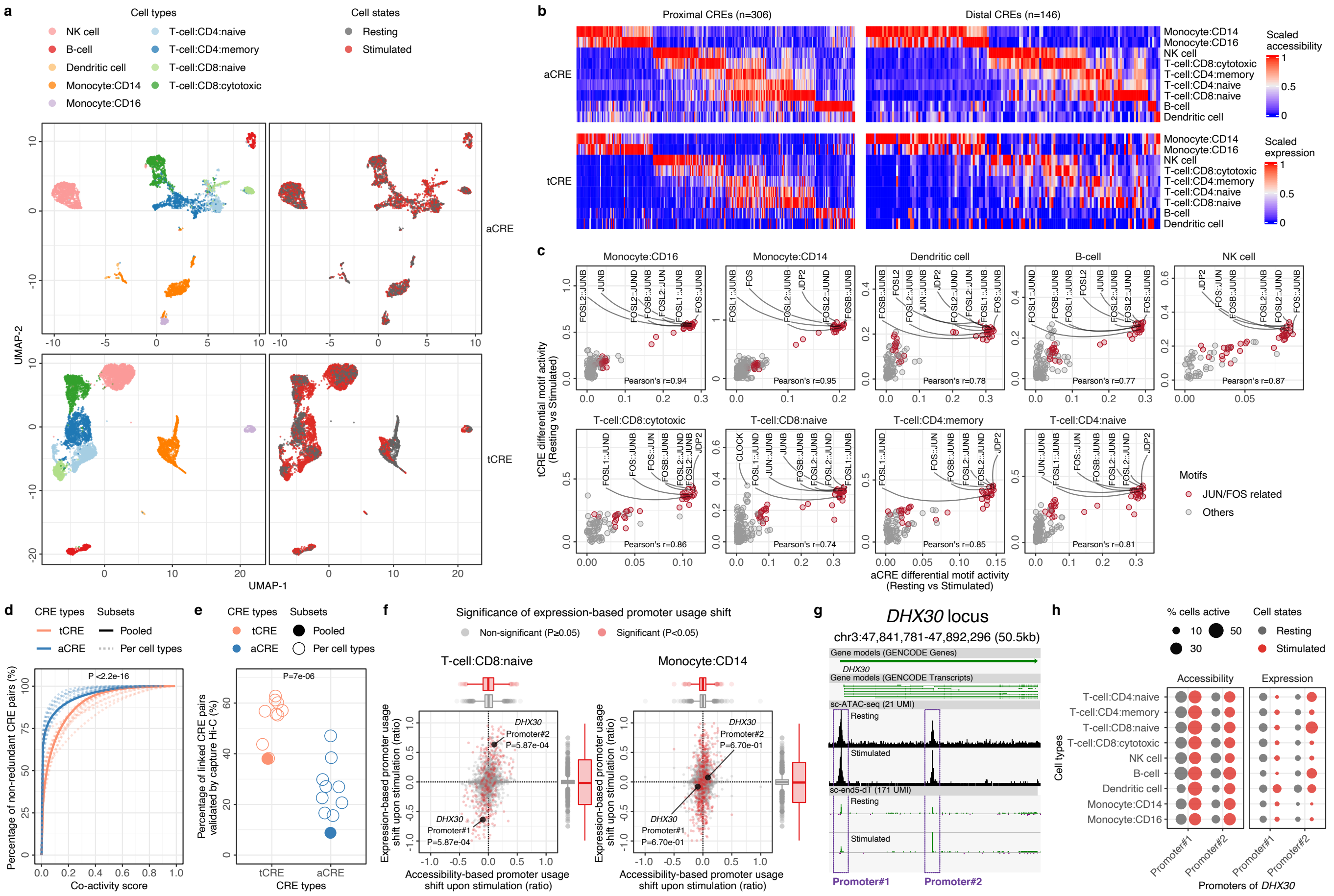


Fig. 5: Comparison of tCRE and aCRE in PBMCs. a, UMAP of cells based on aCRE features *(upper*) and tCRE features (*lower*). Colored by cell cluster (*left*) or stimulation status (*right*). **b**, Heatmap of cell type specific aCRE features (*upper* - color indicates ratio of cells with open aCRE, normalized to the maximum value per loci) and the transcriptional expression present at those loci (*lower* - color indicates the mean expression per cluster normalized to the maximum value per loci). **c,** Motif activity difference between resting and stimulated cells in aCRE (*x-axis*) and tCRE (*y-axis*) per cell cluster. *FOS/JUN* family motifs are highlighted. **d**, Distribution of *Cicero* coactivity scores for tCRE (*orange*) and aCRE (*blue*) within each cluster (*dashed lines*) and all cells pooled (*solid line*). K-S test statistic for difference of distribution in all cell pooled shown. **e**, Number of identified *Cicer*o connections per cluster using co-activity threshold of 0.2 for tCRE (*orange*) and aCRE (*blue*), per cell type (*hollow circles*) and for all cells pooled (*solid circles*). T-test for difference of tCRE and aCRE means shown. **f,** Per cell type alternative promoter usage change upon stimulation for genes with multiple proximal tCRE. (*x-axis*) change in ATAC-seq peak height within tCRE upon stimulation, (*y-axis*) mean change in proportion of gene expression from tCRE across metacells (k=50) upon stimulation. Mann-Whitney U test for change in tCRE usage between metacells shown. **g,** Zenbu genome browser view of highlighted *DHX30* alternative promoters. ATAC-seq signal in resting and stimulated (*upper*), 5′ UMI count in resting and stimulated (*lower*). **h,** Per cell type dot plots of *DHX30* alternative promoters. Proportion of cells with accessible aCRE (*left*) and transcribing tCRE (*right*) colored by stimulation state.

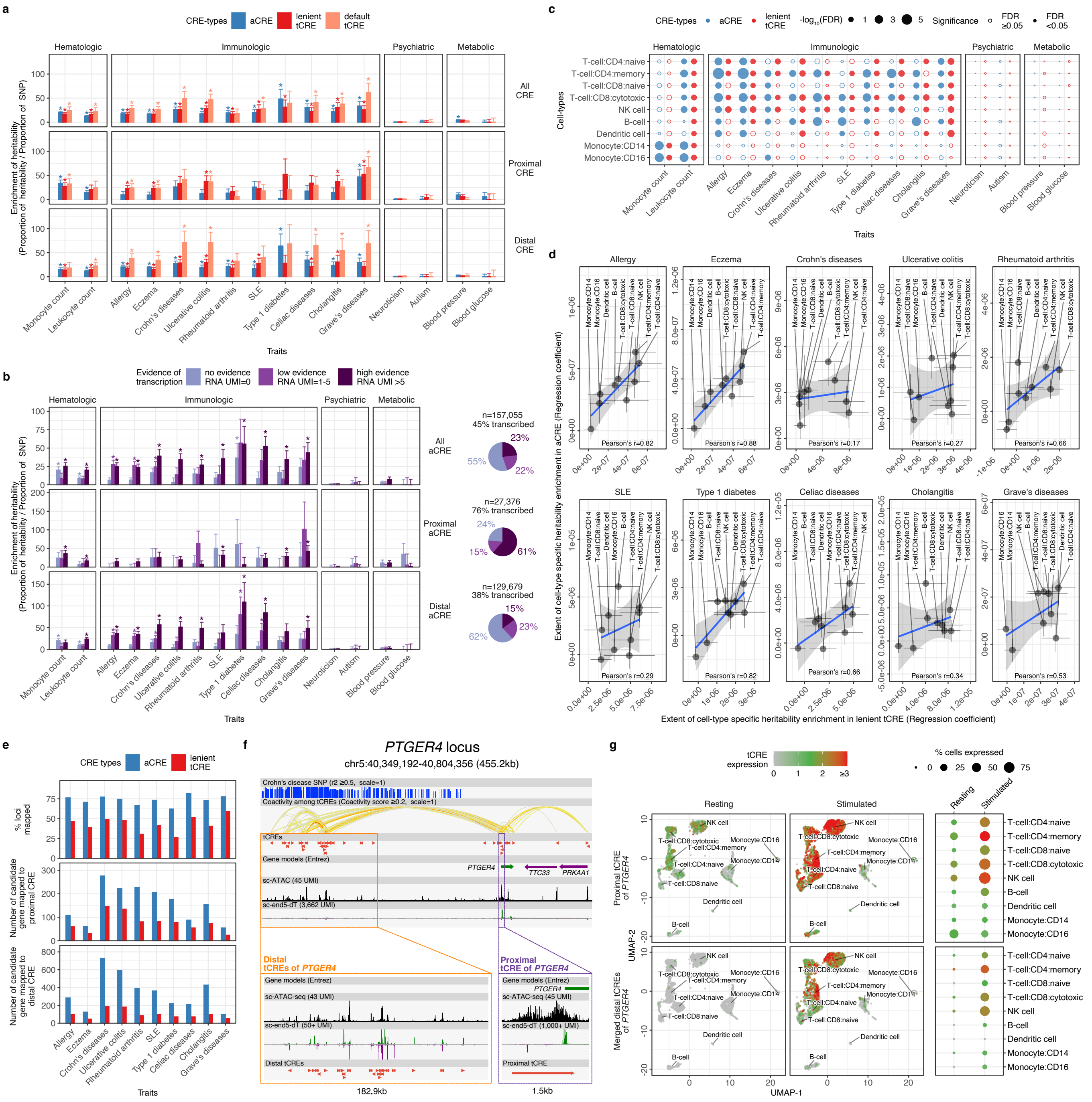
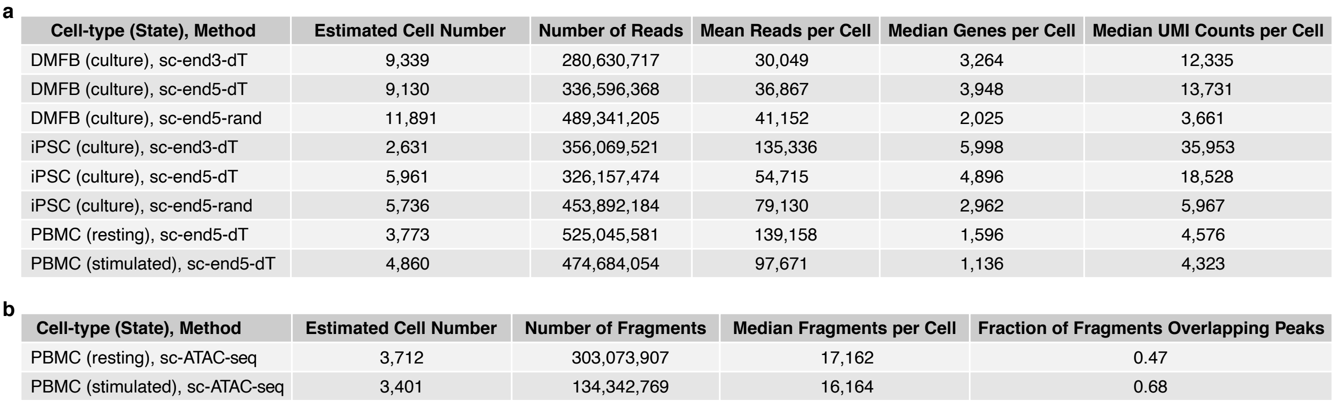


Fig. 6: Disease-associated variants at tCRE and aCRE in PBMCs. a, Enrichment of heritability in various CRE types. *Y-axis*, enrichment of heritability is measured as the ratio of proportion of heritability to proportion of SNP, in LDSC. *Error bars*, standard error of the estimate. Asterisks, significant enrichments with FDR <0.05. **b,** Enrichment of heritability in aCREs with various levels of evidence of transcription. *Y-axis*, *error bars*, and *asterisks* are the same as in **(a)**. **c,** Enrichment of heritability in cell-type specific CREs. *Solid circles*, significant enrichments with FDR <0.05. **d,** Ranking of cell-type relevance to diseases based on heritability enrichment. Regression coefficient, from the analysis in **(c)**, can be interpreted as the extent of heritability enrichment, and thus cell-type relevance. *Error bars*, standard error of the estimate. *Blue line* and *grey shade*, linear regression mean and 95% confidence intervals. **e,** Mapping disease-associated variants to candidate genes using CREs with cell-type/state contexts. Top, percentage of loci with at least 1 candidate gene mapped. *Middle* and *bottom*, number of candidate genes mapped using proximal and distal CREs, respectively, with cell-type/state contexts. **f,** Genetic signals and tCREs at a Crohn's disease risk locus in close to *PTGER4*. Crohn's disease SNP, in LD with $r^2 \geq 0.5$, represented by the height of the bars. Co-activity among tCREs, with score ≥0.2 in *Cicero*, represented by the color of the arcs. Resting and stimulated PBMC data were pooled in the sc-ATAC-seq and sc-end5-dT tracks. *Green* and *blue bars* in the sc-end5-dT track represent the forward and reverse strand signal. The view was generated in the Zenbu genome browser with modifications. **g,** Cell-types/states specific activity of proximal and distal tCREs of *PTGER4*. Merged distal tCREs refers to the sum of expression values of 20 closely located distal tCREs, as detailed in Supplementary Fig.13.



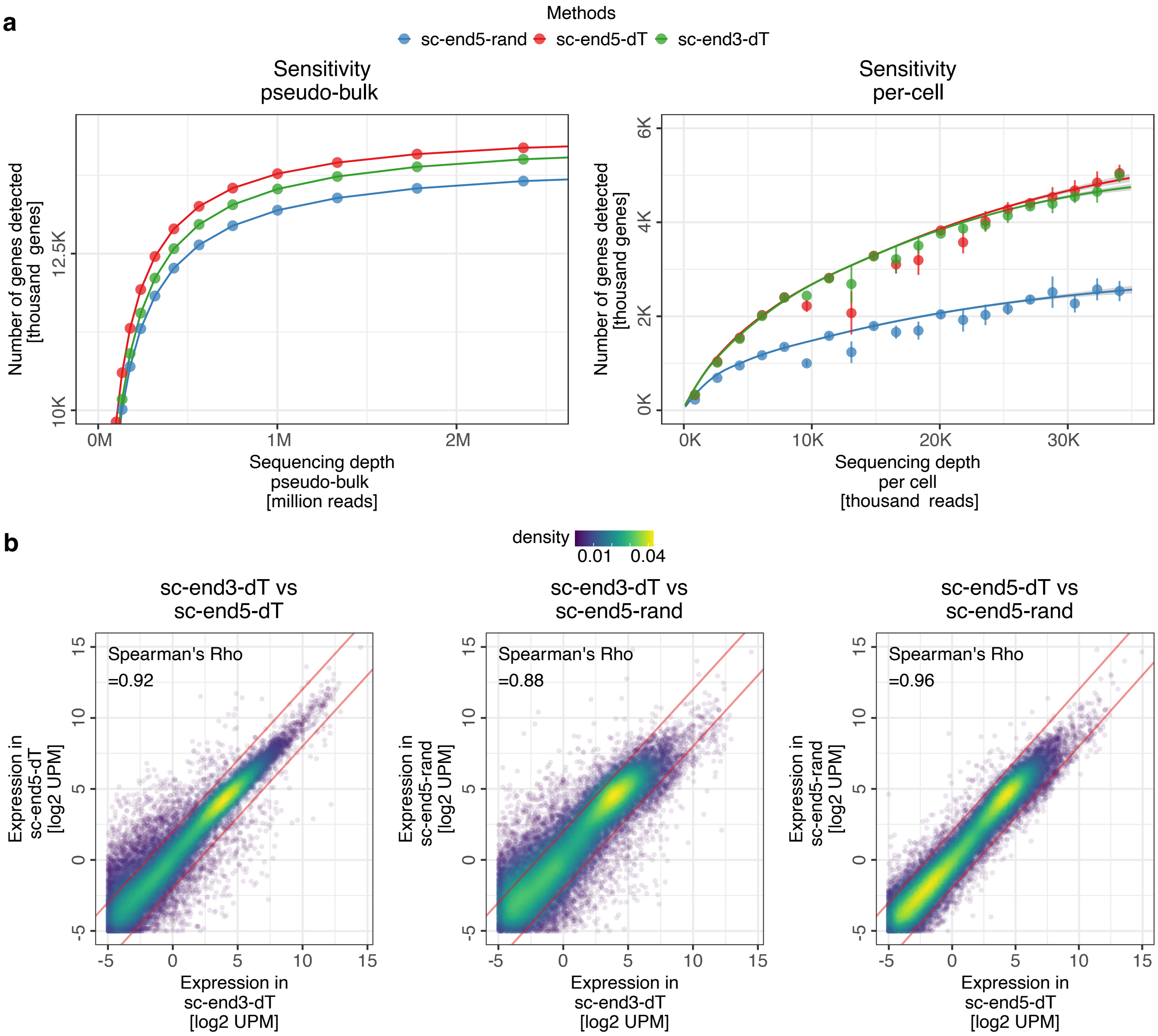
a

Cell-type (State), Method	Estimated Cell Number	Number of Reads	Mean Reads per Cell	Median Genes per Cell	Median UMI Counts per Cell
DMFB (culture), sc-end3-dT	9,339	280,630,717	30,049	3,264	12,335
DMFB (culture), sc-end5-dT	9,130	336,596,368	36,867	3,948	13,731
DMFB (culture), sc-end5-rand	11,891	489,341,205	41,152	2,025	3,661
iPSC (culture), sc-end3-dT	2,631	356,069,521	135,336	5,998	35,953
iPSC (culture), sc-end5-dT	5,961	326,157,474	54,715	4,896	18,528
iPSC (culture), sc-end5-rand	5,736	453,892,184	79,130	2,962	5,967
PBMC (resting), sc-end5-dT	3,773	525,045,581	139,158	1,596	4,576
PBMC (stimulated), sc-end5-dT	4,860	474,684,054	97,671	1,136	4,323

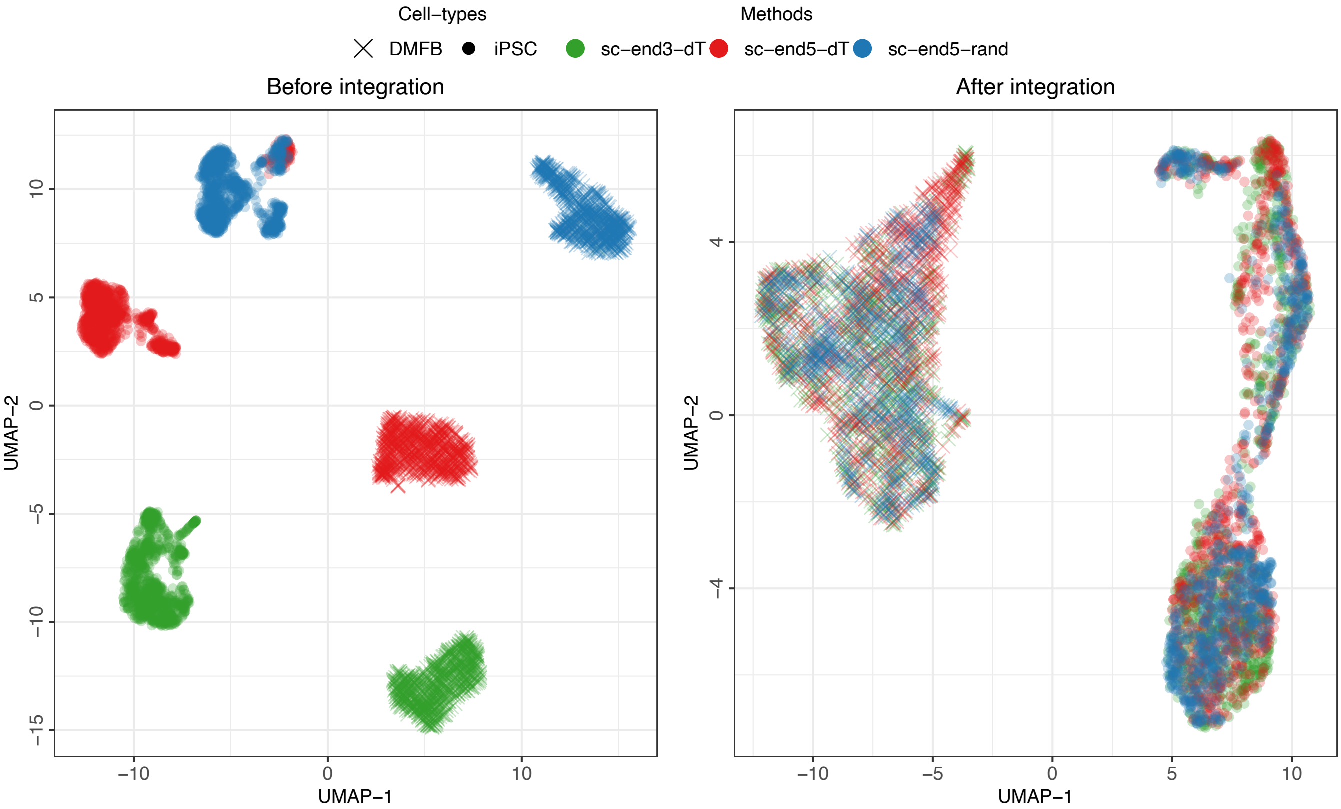
b

Cell-type (State), Method	Estimated Cell Number	Number of Fragments	Median Fragments per Cell	Fraction of Fragments Overlapping Peaks
PBMC (resting), sc-ATAC-seq	3,712	303,073,907	17,162	0.47
PBMC (stimulated), sc-ATAC-seq	3,401	134,342,769	16,164	0.68

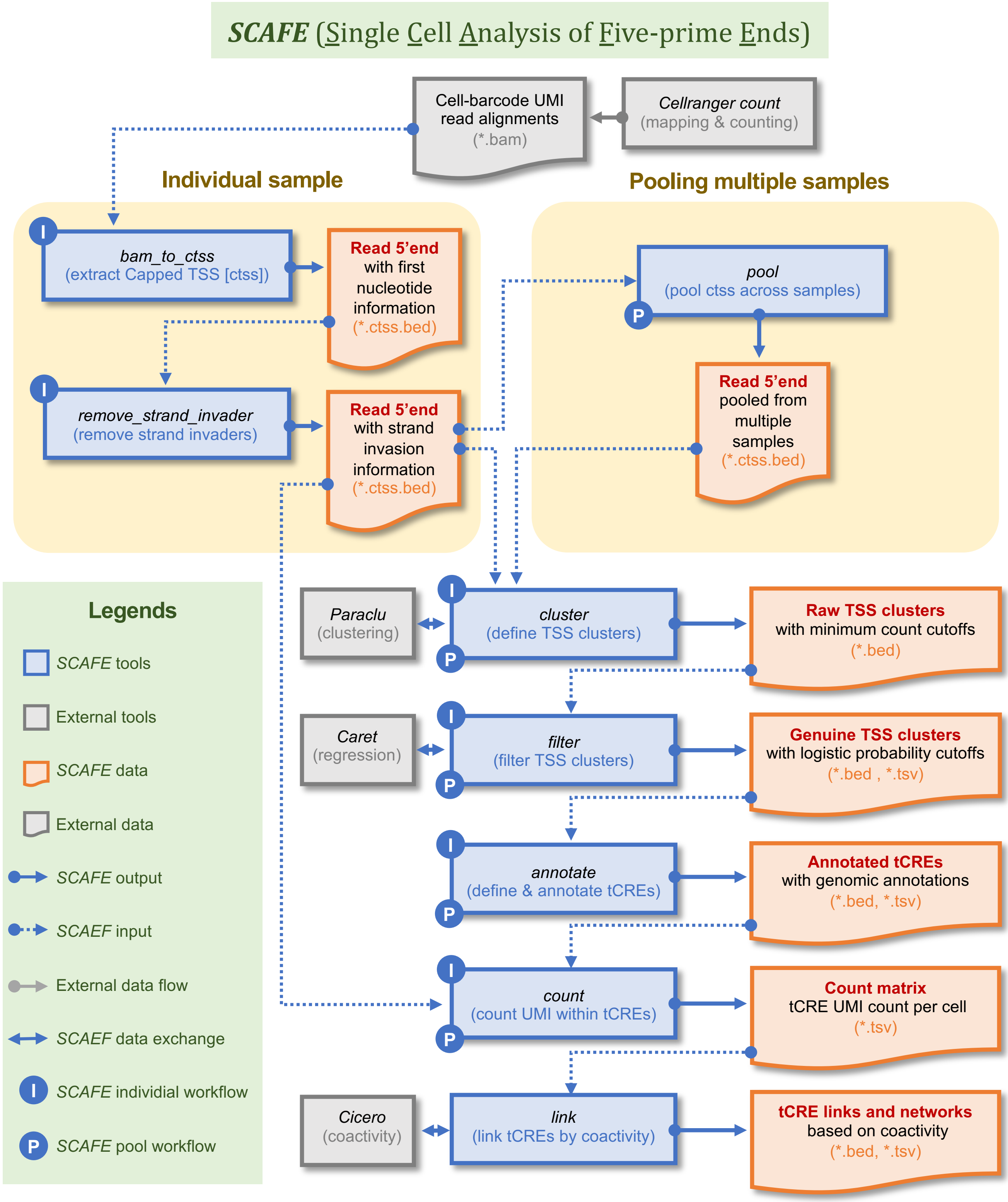
Supplementary Fig. 1: Statistics of sc-RNA-seq and sc-ATAC-seq libraries in this study. a, Statistics of sc-RNA-seq libraries. **b,** Statistics of sc-ATAC libraries. All numbers were extracted from the reports generated from standard 10x Genomics^{TM} tool *Cellranger*.



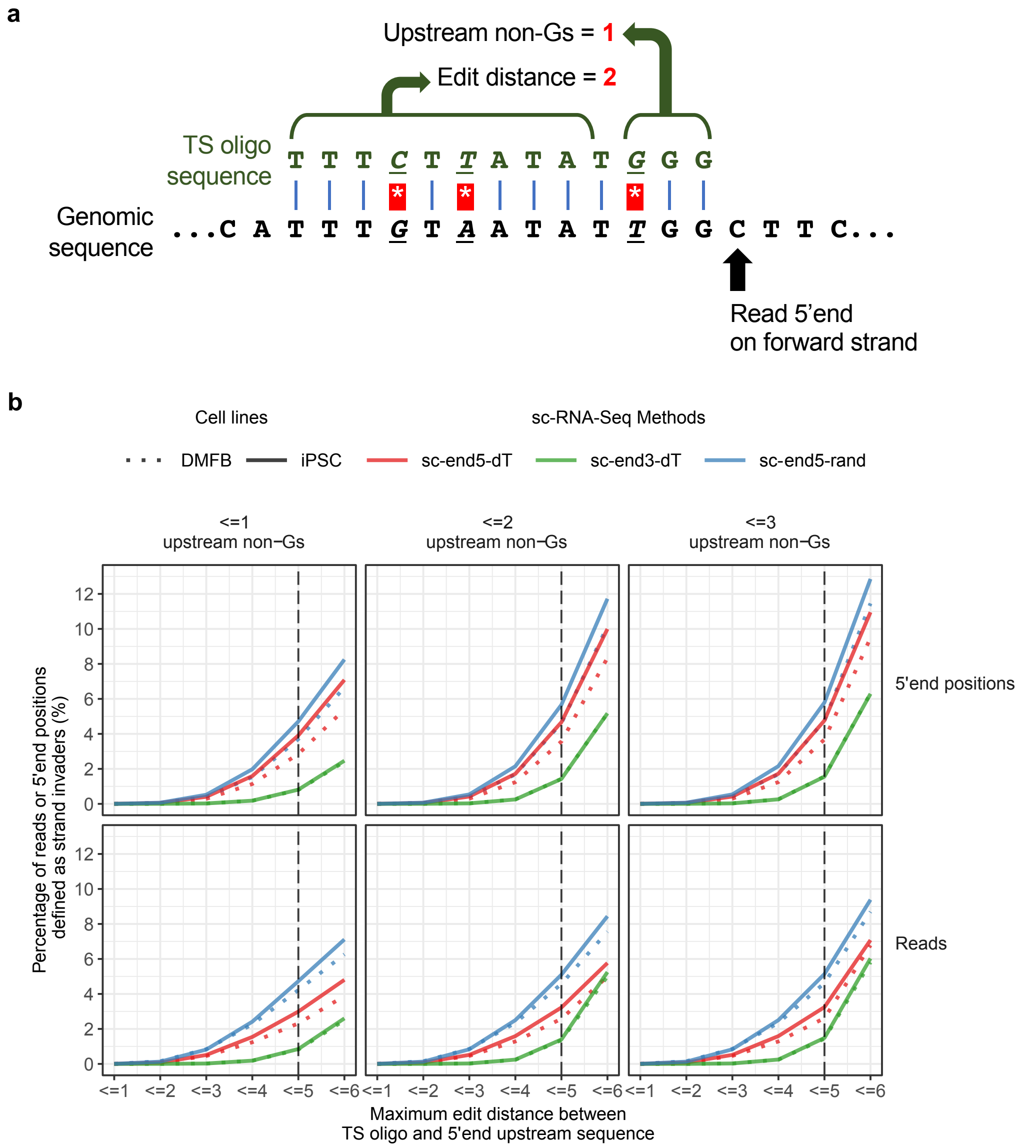
Supplementary Fig. 2: Performance of sc-RNA-seq methods in DMFB. **a**, Sensitivity of detection of genes in pseudo-bulk (*left*) and in single cells (*right*) across sequencing depth. Error bars represent standard deviation. The genes that are detected in bulk-RNA-seq were used as the scope. **b**, Correlation of gene expression levels between the pseudo-bulk data of the three sc-RNA-seq methods. *red line*, ±2-fold differences. UPM, UMI per million. Color represents the density of points.



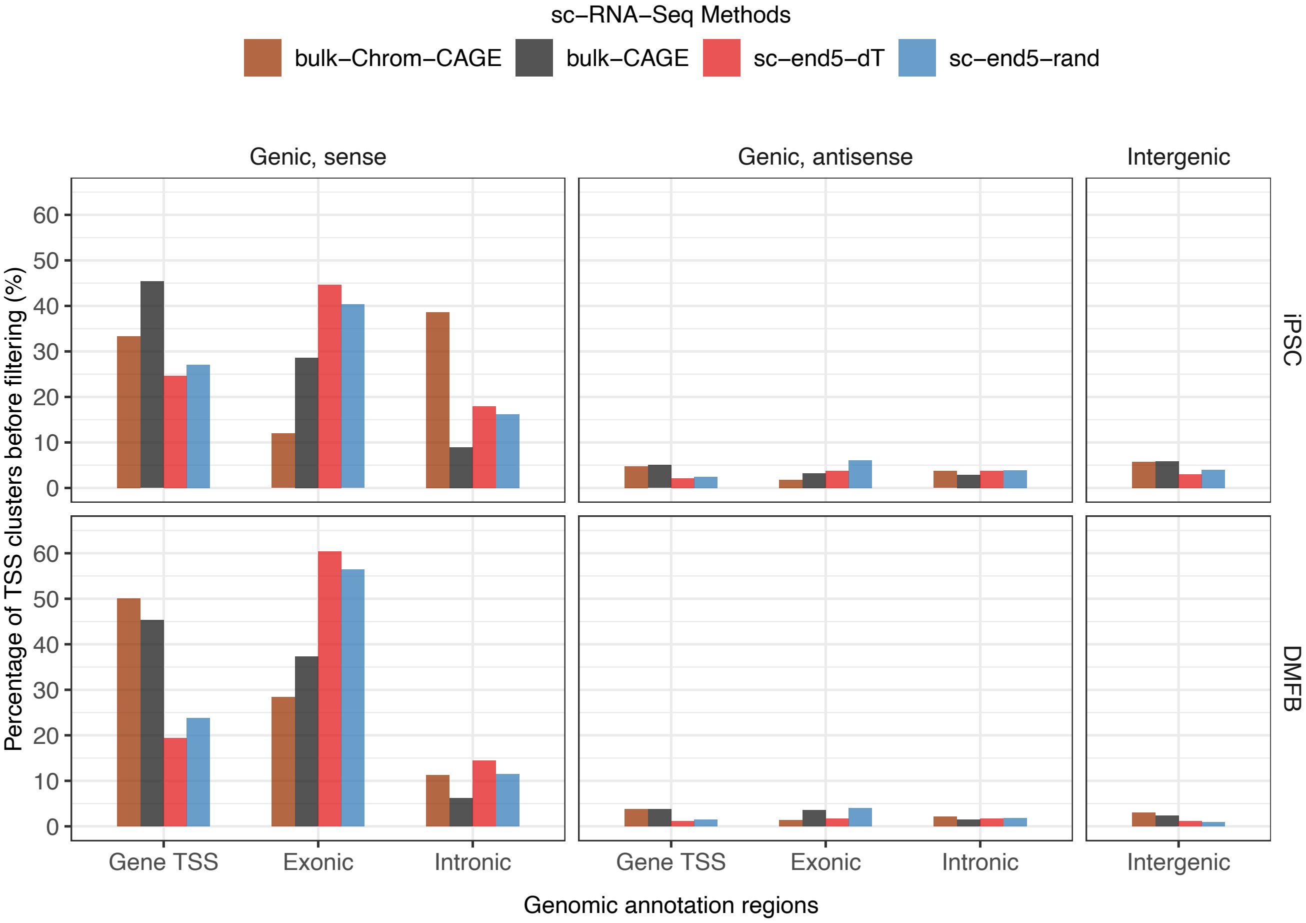
Supplementary Fig. 3: Integration of sc-RNA-seq methods. UMAP of sc-RNA-seq methods in iPSC and DMFB before integration (*left*), after *Seurat* CCA integration (*right*) demonstrating the ability to batch correct between different sequencing methods allowing the integration of sc-end5-seq datasets with existing sc-end3-seq resources.


SCAFE (Single Cell Analysis of Five-prime Ends)
Cell-barcode UMI read alignments (*.bam)
Cellranger count (mapping & counting)
Individual sample
Pooling multiple samples
I
bam_to_ctss (extract Capped TSS [ctss])
Read 5'end with first nucleotide information (*.ctss.bed)
I
remove_strand_invader (remove strand invaders)
Read 5'end with strand invasion information (*.ctss.bed)
pool (pool ctss across samples)
P
Read 5'end pooled from multiple samples (*.ctss.bed)
Legends
SCAFE tools
External tools
SCAFE data
External data
SCAFE output
SCAEF input
External data flow
SCAEF data exchange
SCAFE individial workflow
SCAFE pool workflow
Paraclu (clustering)
cluster (define TSS clusters)
Raw TSS clusters with minimum count cutoffs (*.bed)
Caret (regression)
filter (filter TSS clusters)
Genuine TSS clusters with logistic probability cutoffs (*.bed , *.tsv)
annotate (define & annotate tCREs)
Annotated tCREs with genomic annotations (*.bed, *.tsv)
count (count UMI within tCREs)
Count matrix tCRE UMI count per cell (*.tsv)
Cicero (coactivity)
link (link tCREs by coactivity)
tCRE links and networks based on coactivity (*.bed, *.tsv)

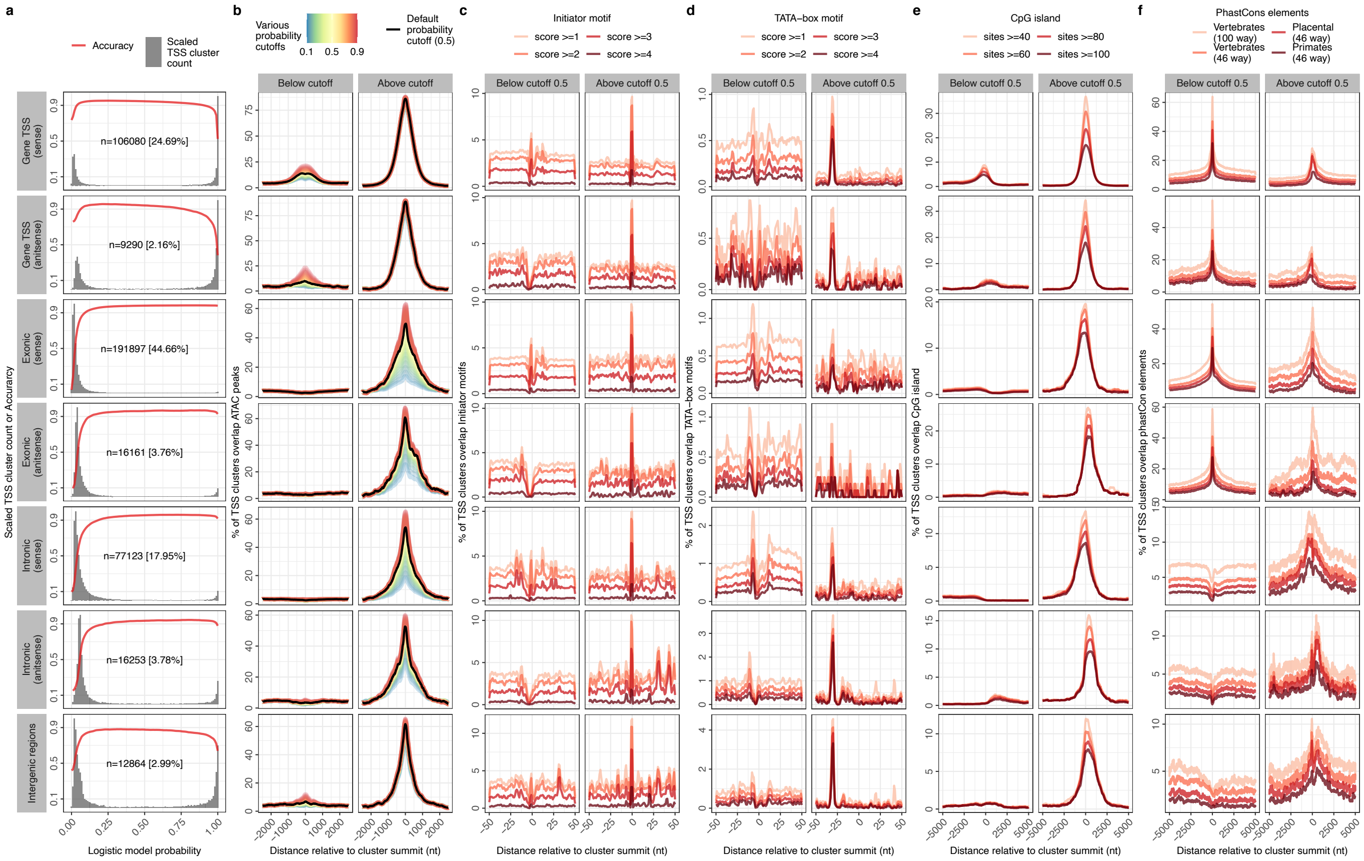
Supplementary Fig. 4: Overview of *SCAFE* tool suite. *SCAFE* consists of a set of *perl* programs for processing of sc-5end-seq data. Major tools are listed here, for all tools please visit https://github.com/chung-lab/scafe. *SCAFE* accepts read alignment in. *bam* format from standard 10x Genomics^{TM} tool *Cellranger*. Tool *bam_to_ctss* extracts the 5′ position of reads, taking the 5′ unencoded-Gs into account. Tool *remove_strand_invader* removes read 5′ends that are strand invasion artifacts by aligning the TS oligo sequence to the immediate upstream sequence of the read 5′end. Tool *cluster* performs clustering of read 5′ends using 3rd-party tool *Paraclu.* Tool *filter* extracts the properties of TSS clusters and performs multiple logistic regression to distinguish genuine TSS clusters from artifacts. Tool *annotate* define tCREs by merging closely located TSS clusters and annotate tCREs based on their proximity to known genes. Tool *count* counts the number of UMI within each tCRE in single cells and generates a tCRE-Cell UMI count matrix. *SCAFE* tools were also implemented workflows for processing of individual samples or pooling of multiple samples.



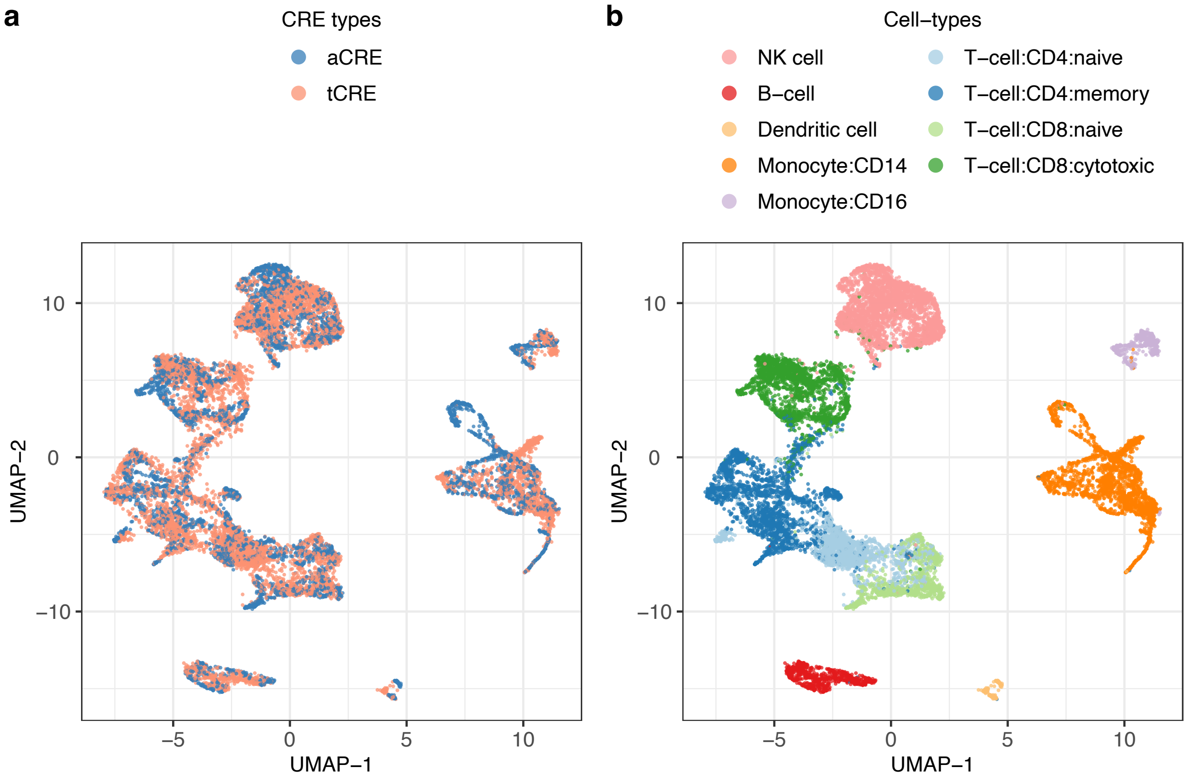
Supplementary Fig. 5: Detection of strand invasion artefacts. a, Rationale of strand invasion detection. The immediate upstream sequence of the read 5′end were aligned with TS oligo sequence. Number of upstream non-Gs was calculated from the first 3nt of the immediate upstream sequence. Edit distance was calculated from the last 10nt of the alignment. The shown example has 2 edit distances and 1 upstream non-Gs. **b,** Extent of strand invasion artefacts in various sc-RNA-seq methods. Maximum edit distance of 5 (*vertical dotted line*) and 2 upstream non-Gs (middle column) is chosen as the threshold to define strand invasion artefacts. At this threshold, the extent of strand invasion artefacts is consistently higher in sc-end5-rand (*blue*), compared to sc-end5-dT (*red*), in both DMFB and iPSC. sc-end3-dT (*green*) serves as a negative control of the random genomic background.



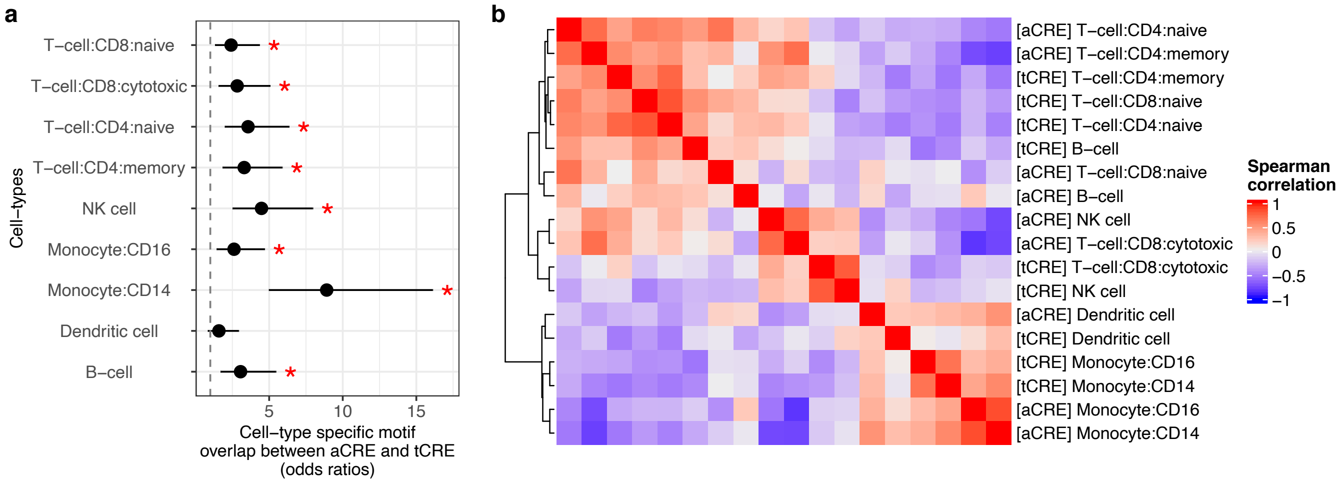
Supplementary Fig. 6: Genomic distribution of unfiltered TSS clusters. Unfiltered TSS clusters were assigned to various genic and intergenic annotations, based on their intersection with GENCODE annotation, in specific hierarchical orders (See Methods). In both DMFB and iPSC, a large fraction of TSS clusters were assigned to the exonic and intronic regions in sense orientation, compared to that in the antisense orientation. This could be attributed to the “exon painting” artefacts as discussed, which could be filtered by considering various properties of the TSS clusters.



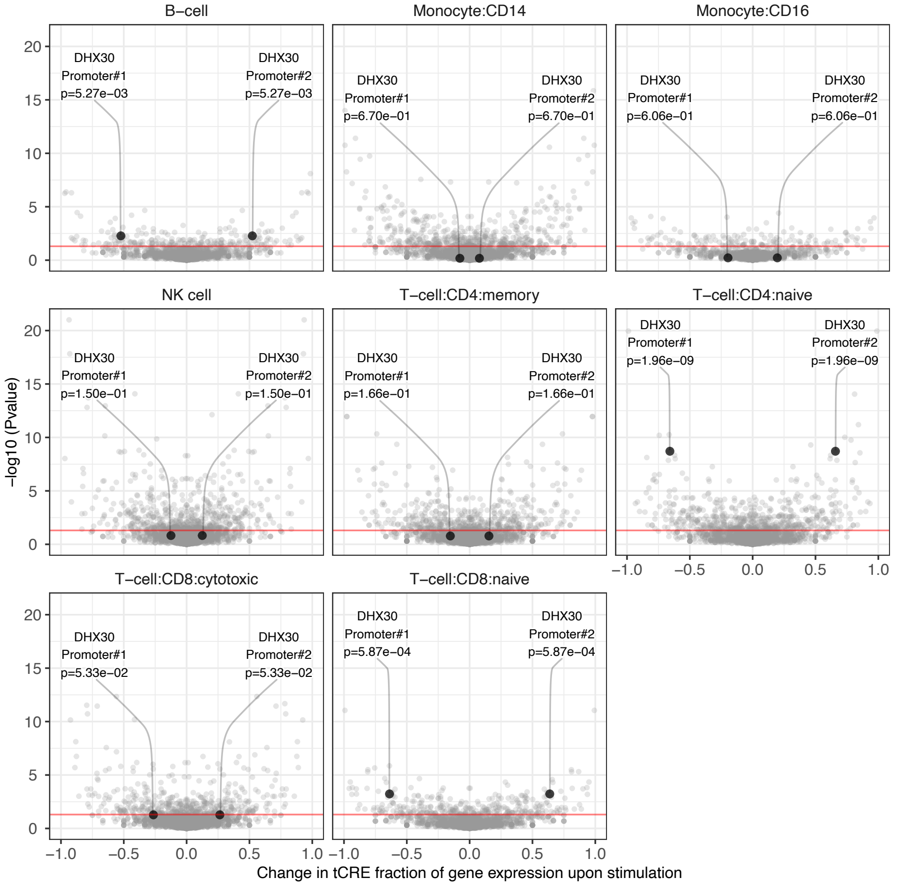
Supplementary Fig. 7: Properties of logistic model probability cutoffs for identification of genuine TSS clusters. a, Proportion of TSS clusters and accuracy along logistic model probability cutoffs. "n" and "%" refers to the number and percentage of TSS clusters in the category. **b,** Chromatin accessibility around summit of TSS clusters along logistic model probability thresholds. **c**,**d**,**e**,**f**, Distribution of Initiator motif, TATA-box motif, CpG island and PhastCons elements, respectively, around summit of TSS clusters below and above logistic model probability 0.5. Initiator motif and TATA-box motif were predicted on hg19 using *HOMER* (http://homer.ucsd.edu/homer/motif/). CpG island and PhastCons elements were downloaded from UCSC table browser (https://genome.ucsc.edu/). "Score" in *c* and *d* refers to score of motif prediction from *HOMER*. "Sites" in *e* refers to number of CG dinucleotides. In *f*, 100 ways and 46 ways refer to multiple alignments of 100 and 46 species respectively. Vertebrates, Placental and Primates refer to the scope of species used to define PhastCons elements. Initiator motif and TATA-box motif are, as expected, enriched at ~0nt and ~ –30nt, respectively, of the TSS cluster above below cutoff 0.5. The enrichment of PhastCons elements at the center of the "Gene TSS" and "Exonic" TSS clusters below cutoff 0.5 can be attributed to their overlap with exon regions, which are relative more conserved than intronic and intergenic regions.



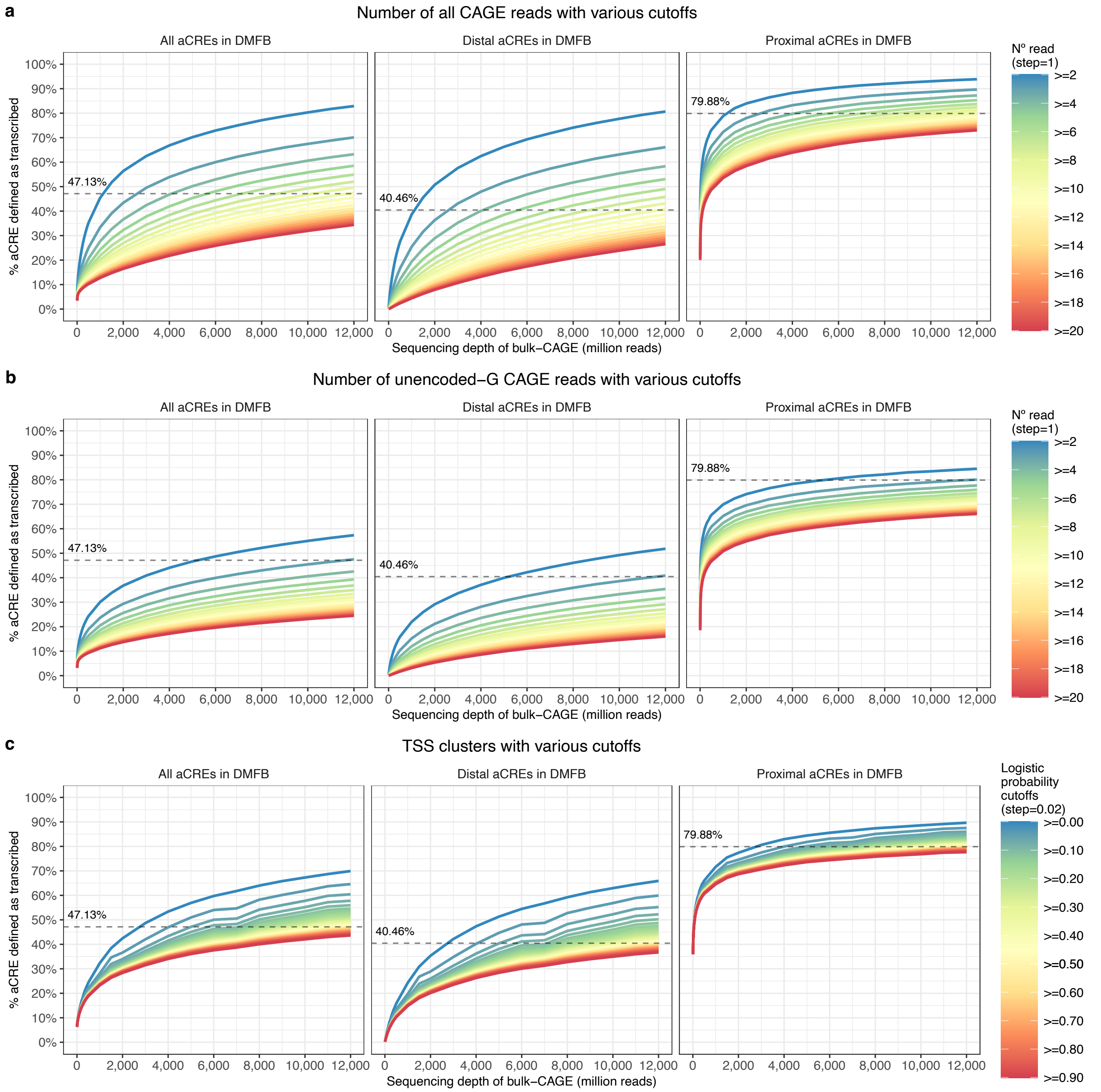
Supplementary Fig. 8: Integration of tCRE and aCRE. UMAP of tCRE and aCRE cells after integration by *Seurat CCA*. Colored by technology (*left*) and cell type annotation (*right*), cell type labels have been transferred from tCRE to aCRE.



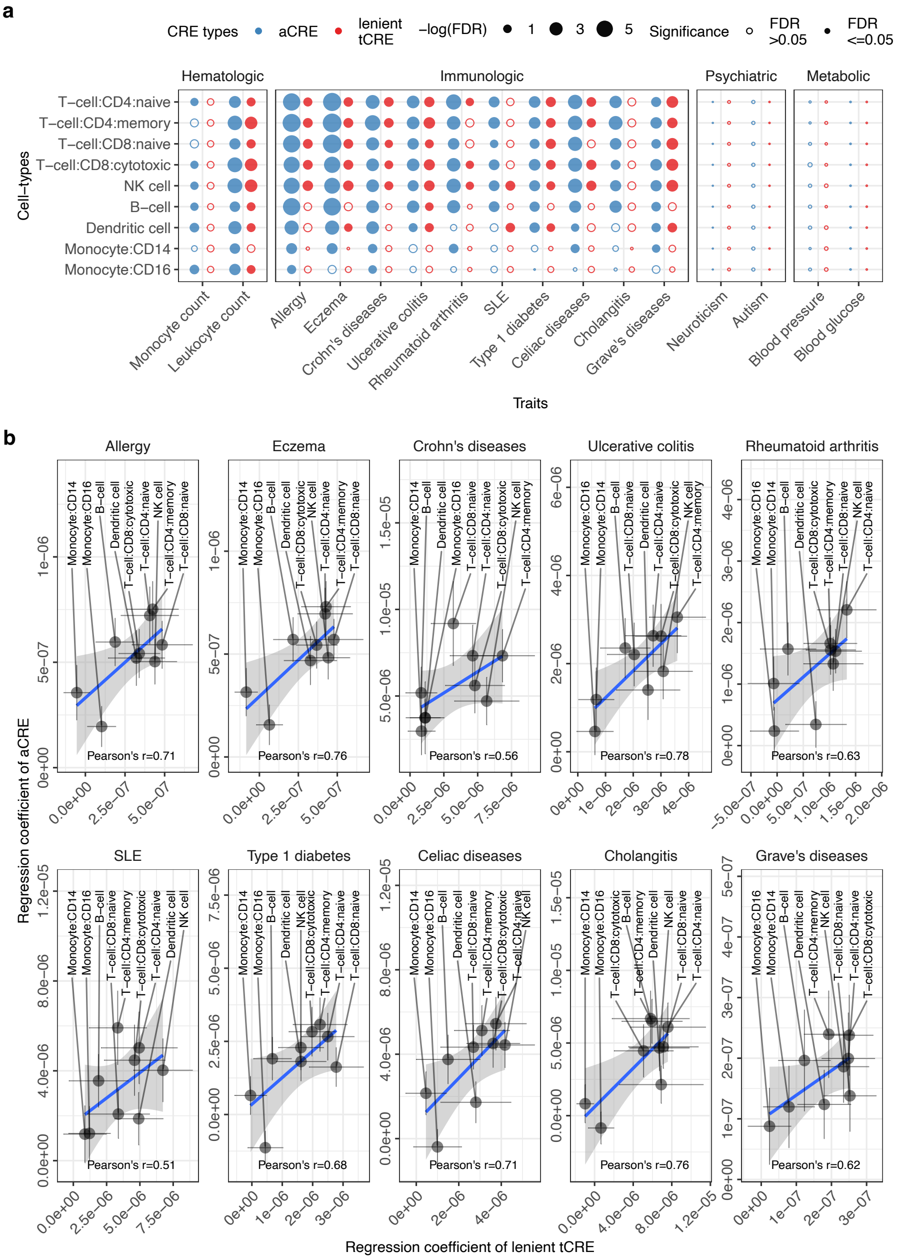
Supplementary Fig. 9: Cell in tCRE and aCRE. a, Fisher's exact test for odds ratio of overlap in top 80 cell-type specific TF motifs calculated with *ChromVAR* in aCRE and tCRE. *asterisk*, P<0.05; *error bars*, 95% confidence interval; **b,** Heatmap of common cell type specific motif activity from (a) averaged per cell type (spearman correlation).



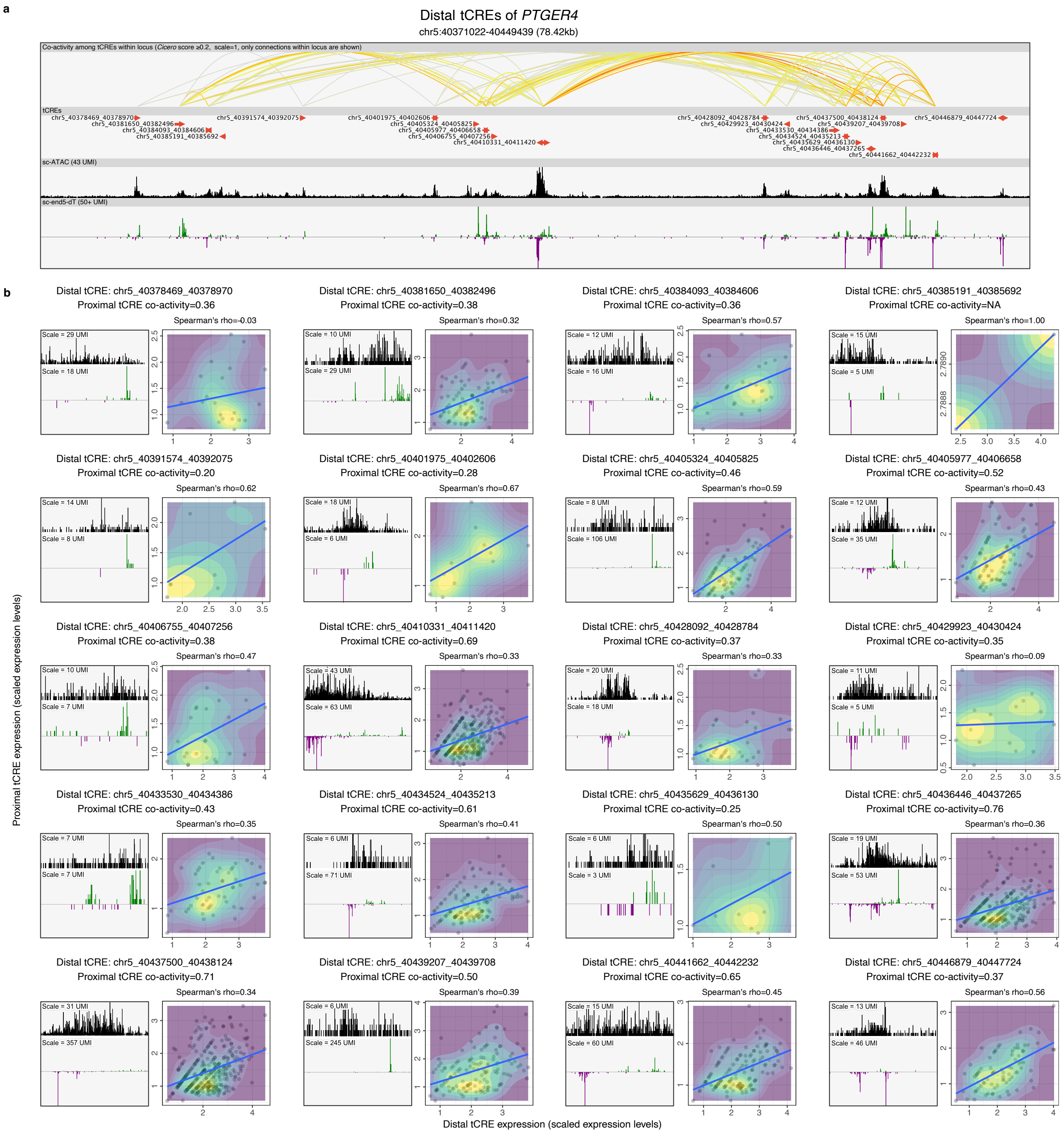
Supplementary Fig. 10: Alternative promoters in 5′end sc-RNA-seq. Volcano Plot for genes with multiple proximal tCRE corresponding to alternative promoters, change in mean fraction of gene expression in metacells from each tCRE after stimulation (X-axis), $-\log_{10}(P)$ of Mann-Whitney U test for change in tCRE usage between metacells (Y-axis). Labeled example tCRE of the *DHX30* gene. Switching from Promoter#1 to Promoter#2 occurs significantly upon stimulation in naive CD4 T-cells naive CD8 T-cells and B-cells.



Supplementary Fig. 11: Percentage aCRE that are transcribed in DMFB. Estimating the percentage of aCREs that are transcribing using pooled CAGE libraries of DMFB at unprecedented sequencing depth based on **a,** number of all CAGE reads at TSS summit within aCRE, **b,** number of unencoded-G CAGE reads at TSS summit within aCRE, or **c,** highest logistic probability of TSS clusters within aCRE. *Dashed line*, estimate of transcribed aCRE % at highest sequencing depth (i.e. 12,000M) based on TSS clusters with default logistic probability cutoffs (i.e. 0.05).



Supplementary Fig. 12: Heritability enrichment in stimulation-responsive CREs. a, Enrichment of heritability in stimulation-responsive CREs in various cell-types. Solid circles, significant enrichments with FDR <0.05. **b,** Ranking of cell-type relevance to diseases based on heritability enrichment. Regression coefficient, from the analysis in **(a)**, can be interpreted as the extent of heritability enrichment, and thus cell-type relevance. Error bars, standard error of the estimate. Blue line and grey shade, linear regression mean and 95% confidence intervals.



Distal tCRE: chr5_40378469_40378970
Proximal tCRE co-activity=0.36
Spearman's rho=-0.03

Distal tCRE: chr5_40381650_40382496
Proximal tCRE co-activity=0.38
Spearman's rho=0.32

Distal tCRE: chr5_40384093_40384606
Proximal tCRE co-activity=0.36
Spearman's rho=0.57

Distal tCRE: chr5_40385191_40385692
Proximal tCRE co-activity=NA
Spearman's rho=1.00

Distal tCRE: chr5_40391574_40392075
Proximal tCRE co-activity=0.20
Spearman's rho=0.62

Distal tCRE: chr5_40401975_40402606
Proximal tCRE co-activity=0.28
Spearman's rho=0.67

Distal tCRE: chr5_40405324_40405825
Proximal tCRE co-activity=0.46
Spearman's rho=0.59

Distal tCRE: chr5_40405977_40406658
Proximal tCRE co-activity=0.52
Spearman's rho=0.43

Distal tCRE: chr5_40406755_40407256
Proximal tCRE co-activity=0.38
Spearman's rho=0.47

Distal tCRE: chr5_40410331_40411420
Proximal tCRE co-activity=0.69
Spearman's rho=0.33

Distal tCRE: chr5_40428092_40428784
Proximal tCRE co-activity=0.37
Spearman's rho=0.33

Distal tCRE: chr5_40429923_40430424
Proximal tCRE co-activity=0.35
Spearman's rho=0.09

Distal tCRE: chr5_40433530_40434386
Proximal tCRE co-activity=0.43
Spearman's rho=0.35

Distal tCRE: chr5_40434524_40435213
Proximal tCRE co-activity=0.61
Spearman's rho=0.41

Distal tCRE: chr5_40435629_40436130
Proximal tCRE co-activity=0.25
Spearman's rho=0.50

Distal tCRE: chr5_40436446_40437265
Proximal tCRE co-activity=0.76
Spearman's rho=0.36

Distal tCRE: chr5_40437500_40438124
Proximal tCRE co-activity=0.71
Spearman's rho=0.34

Distal tCRE: chr5_40439207_40439708
Proximal tCRE co-activity=0.50
Spearman's rho=0.39

Distal tCRE: chr5_40441662_40442232
Proximal tCRE co-activity=0.65
Spearman's rho=0.45

Distal tCRE: chr5_40446879_40447724
Proximal tCRE co-activity=0.37
Spearman's rho=0.56

Supplementary Fig. 13: Distal tCRE activity at the *PTGER4* locus. a, Overview of the distal tCREs in close proximity to *PTGER4*. Twenty distal tCREs were shown. Co-activity among these 20 tCREs, with *Cicero* co-activity score ≥0.2, is represented by the color of the arcs. Only coactivity among tCRE within the view was shown. Resting and stimulated PBMC data were pooled in the sc-ATAC-seq and sc-end5-dT tracks. *Green* and *blue bars* in the sc-end5-dT track represent the forward and reverse strand signal. The view was generated in the Zenbu genome browser with modifications. **b**, Individual distal tCREs and their coactivity with *PTGER4* proximal tCRE. For each distal tCRE in **(a)**, a zoom-in view at the locus is shown. The scale of the signal bars is indicated as UMI counts. Expression of individual distal tCRE and the *PTGER4* proximal tCRE within single cells are plotted. Only cells with non-zero values in both tCREs are plotted. *Blue line*, mean of linear regression.