PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

An Epigenomic fingerprint of human cancers by landscape interrogation of super enhancers at the constituent level

Xiang Liu[1], Nancy Gillis[2], Chang Jiang[3], Anthony McCofie[1], Timothy I. Shaw[1], Aik-Choon Tan[4], Bo Zhao[5], Lixin Wan[3], Derek R. Duckett[6], Mingxiang Teng[1]*

1 Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, Florida, United States of America, **2** Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, Florida, United States of America, **3** Department of Molecular Oncology, Moffitt Cancer Center, Tampa, Florida, United States of America, **4** Department of Oncological Sciences, Huntsman Cancer Institute, The University of Utah, Salt Lake City, Utah, United States of America, **5** Division of Infectious Disease, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States of America, **6** Department of Drug Discovery, Moffitt Cancer Center, Tampa, Florida, United States of America

* mingxiang.teng@moffitt.org

Citation: Liu X, Gillis N, Jiang C, McCofie A, Shaw TI, Tan A-C, et al. (2024) An Epigenomic fingerprint of human cancers by landscape interrogation of super enhancers at the constituent level. PLoS Comput Biol 20(2): e1011873. https://doi.org/10.1371/journal.pcbi.1011873

Editor: Ilya Ioshikhes, CANADA

Received: July 11, 2023

Accepted: January 30, 2024

Published: February 9, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1011873

Data Availability Statement: ChIP-seq data in this study were obtained from Gene Expression Omnibus (GEO) with accession ID GSE143653. Other public data were summarized in S1 Table. All analysis code in this manuscript were documented

Abstract

Super enhancers (SE), large genomic elements that activate transcription and drive cell identity, have been found with cancer-specific gene regulation in human cancers. Recent studies reported the importance of understanding the cooperation and function of SE internal components, i.e., the constituent enhancers (CE). However, there are no pan-cancer studies to identify cancer-specific SE signatures at the constituent level. Here, by revisiting pan-cancer SE activities with H3K27Ac ChIP-seq datasets, we report fingerprint SE signatures for 28 cancer types in the NCI-60 cell panel. We implement a mixture model to discriminate active CEs from inactive CEs by taking into consideration ChIP-seq variabilities between cancer samples and across CEs. We demonstrate that the model-based estimation of CE states provides improved functional interpretation of SE-associated regulation. We identify cancer-specific CEs by balancing their active prevalence with their capability of encoding cancer type identities. We further demonstrate that cancer-specific CEs have the strongest per-base enhancer activities in independent enhancer sequencing assays, suggesting their importance in understanding critical SE signatures. We summarize fingerprint SEs based on the cancer-specific statuses of their component CEs and build an easy-to-use R package to facilitate the query, exploration, and visualization of fingerprint SEs across cancers.

Author summary

Super enhancers are large genomic elements comprised of multiple enhancers working together to drive gene transcription. They play a crucial role in defining cell identity and act as drivers of oncogenic gene expression in cancer cells. Characterizing cancer-specific super enhancer signatures can reveal transcriptional deregulation associated with cell

on GitHub (https://github.com/tenglab/cSEAdb_plos_code). In addition, numerical values to generate manuscript graphs and histograms were documented in S1 Data. To ensure result reproducibility, we deposited the computational framework of SE fingerprint identification for NCI-60 cancers as a protocol: https://dx.doi.org/10.17504/protocols.io.kxygx38wzg8j/v1.

Funding: This work was supported by NIH grants R03DE030580 (MT), R01CA262530 (DRD), R01CA255398 (LW), R01AI123420 (BZ), P30CA076292 (Biostatistics and Bioinformatics Shared Resource at Moffitt Cancer Center) and Moffitt Bio2 Pilot Grant (MT). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

origin and malignant transformation. Here, we generated a high-resolution fingerprint of super enhancers across 60 cancer cell lines through statistical modeling of both active and inactive components inside super enhancers. Our study revealed that cancer-specific super enhancer components are highly informative in delineating the identity of cancer cells. Our findings further revealed that cancer-specific active components exhibit stronger enhancer activities compared to non-cancer-specific components, suggesting the importance of studying the functional divergence inside super enhancers across different cancer types. Finally, we generated a database of cancer-specific super enhancer signatures for 28 cancer types with a companion computational tool to facilitate the query, exploration, and visualization of these signatures across cancers.

Introduction

The diversity in oncogenesis mechanisms across cancers and their subtypes is largely underlain by cancer molecular profiles. Numerous molecular signatures, either pan-cancer-involved or cancer-specific, are reported in genetics and epigenetics studies. These key cancer-specific signatures oftentimes act differently across cancers, including somatic mutations, DNA methylation, and dysregulated transcription of oncogenes [1–6]. Together, these signatures provide complementary knowledge in understanding divergent oncogenic mechanisms. Super enhancers (SE), a group of large genomic elements highly corresponding to cell identities, have critical functions in cancer gene regulation [7–9]. Specifically, some SEs have dominant roles in driving tumor progression compared to other genetic signatures [9–11]. For instance, multiple SEs can significantly amplify the expression of oncogene MYC and promote tumorigenesis under different mechanisms across cancers [12,13]. Evaluating cancer-specific SE signatures thus provides critical insights toward fully dissecting the divergence of cancer mechanisms.

Pan-cancer analysis has strength in prioritizing cancer-specific signatures [1,14–18]. Currently, SE presence and absence is assessed across the NCI-60 cell panel [19] and other cells [20]; however, the binary characterization ignores variation in the SE structure, such as drifting or shrinkage of SE occupancies on the genome. These differences in effects provide important functional information, especially given the dynamic size of most SEs. Indeed, we previously showed that a significant portion of SEs are involved in structural alterations between cancers [21]. The functional effects of SEs are at least in part driven by their constituent components (i.e., constituent enhancers), which have variable contributions to SEs' overall functions [22–26]. This existing knowledge suggests an urgent need to interrogate cancer-specific SE signatures at a higher resolution by zooming into SEs at the constituent level.

In this study, we revisit SE profiles using the pan-cancer dataset (i.e., H3K27Ac ChIP-seq) of the NCI-60 cell panel [19]. To achieve a high-resolution interrogation, we focus on comparing individual constituent enhancer (CE) activities instead of the presence and absence of the whole SE. We implement a mixture model to automatically discriminate active CEs from inactive CEs which enhances the identification of cancer-specific SE signatures at the CE level. We investigate cancer-cell-specific compared to non-specific active CEs in the context of SE activity to provide insight on the functional impacts of the CEs across cancers. We generate an SE-based signature fingerprints for human cancers and build an R tool to facilitate the exploration of these signatures.

Materials and methods

Data acquisition

Raw H3K27ac ChIP-seq of NCI-60 human cancer cell lines were downloaded from GEO repositories with accession ID GSE143653 [19]. Quality-controlled chromatin contact files of POLR2A ChIA-PET and STARR-seq for four cancer cell lines (A549—Lung Cancer, HCT116—Colorectal Cancer, K562—Leukemia, and MCF7—Breast Cancer) were downloaded from ENCODE [27] with accession numbers (ChIA-PET: ENCFF946FGU, ENCFF246ZKR, ENCFF511QFN, and ENCFF377SXL; STARR-seq: ENCFF646OQS, ENCFF428KHI, ENCFF045TVA, and ENCFF826BPU). PRO-seq of A549 were download from ENCODE with accession number of ENCFF719NYS and ENCFF454FJE. PRO-seq of K562 were downloaded from GEO with accession number of GSM1480327 [28]. Transcription factor ChIP-seq data was downloaded from ENCODE (S1 Table). Normalized gene expression profiles for 1450 cancer cell lines were downloaded from DepMap data portal [29]. DepMap gene expression values were transformed to percentile-based (0 to 100 percentile) measurements for each gene to ease the comparison of gene expression levels across genes.

SE profile identification and normalization

Raw H3K27Ac ChIP-seq data of the NCI-60 cancer cell lines was aligned to human genome hg38 using Bowtie2 [30], followed with peak calling by MACS2 [31]. SEs were estimated using ROSE for each ChIP-seq sample [32,33]. Peaks overlapped with gene promotors (upstream 3k bp to downstream 1k bp) and blacklist regions [34] were excluded from SE estimation. Default parameters were applied in these tools. To create a unified SE candidate list across the NCI-60 cell lines, SE regions from all samples (60 cell lines x 2 replicates = 120 samples) were merged if at least 25% width overlapping is detected between SE regions. Here, requiring a smaller overlapping window decreases the risk of merging two consecutive but independent SEs, while a larger overlapping window increases the risk of over-fragmenting the same SE into multiple separate SEs. 25% was selected as the maximum cutoff with which no obvious drop was observed for the median widths of the merged SE lists (**S1 Fig**). It is noted that the median width and total number of the merged SEs vary less than 0.2% and 3.4%, respectively, with the overlap cutoffs ranged between 0 to 25%. This suggests the cutoff has limited effects on the final SE list. To create a unified list of candidate CEs, CEs with at least 25% width overlapping are merged following guidelines by previous publications [35,36]. Similarly, we observed limited effects of cutoff selection on the final CE list, as the median width and total number of the merged CEs vary within 8.6% and 6.5%, respectively, with the cutoffs ranged between 0 to 25%. Other enhancers were unified with the same merging parameters for downstream normalization purpose.

In order to compare SE/enhancer activities across cancer samples, we first quantified genome-wide enhancer activities based on ChIP-seq read signals at the unified enhancer regions using featureCount [37]. A matrix of enhancer activities for all cancer samples were generated. Replicate samples were aggregated together resulting in an enhancer activity matrix for 60 cancer cell lines. Enhancer activities were normalized to adjust sequencing depths across cancer cell lines using the RLE method implemented in DESeq2 package [38]. To facilitate downstream model fitting, we replace zero signals in the normalized matrix with imputated values. In brief, for a zero signal at the enhancer X in a cell line Y, we replaced it as the half of minimum signals of all non-zero enhancers in the cell line Y. We repeated this process for all zero signals at different enhancers in different cell lines. The normalized and imputated CEs' activity matrix were applied in downstream analysis.

Mixture model to discriminate active from inactive CEs

A mixture model was developed at each CE to automatically estimate active and inactive cancer cell lines. We hypothesize each candidate CE has two states, the active and inactive. High H3K27Ac ChIP-seq signals correspond to active states while the low signals indicate inactive or weak activities. A two-component mixture model was fit to determine an optimal activity cutoff separating ChIP-seq signals between the active and inactive states.

We assumed the log2-transformed distribution for each CE activity consisted of a mixture of two normal distributions. First, to facilitate model fitting of individual CEs, we estimated the model priors based on genome-wide CEs (**S2 Fig**). Activities of genome-wide CEs were fit with global mixture models for each cell line (using R package *mixtools* [39]) to estimate the mean and standard deviations of the high and low activity mixtures. These values were averaged across cell lines to generate overall active and inactive consensuses and applied as priors in fitting models for individual CEs. CE occurrences with zero signals before imputation were excluded from the global models to ensure robust fitting. Second, we fit two-component normal-distributed mixture models for each CE using their activities across 60 cell lines. Basically, the means and standard deviations of active and inactive components for individual CEs were estimated. The probability of each activity value belonging to the active component was estimated. Expectation-Maximization algorithms was used to control model converging at no more than 1e-8 change of log-likelihood [40]. Third, for a CE activity in a fit mixture model, if its probability of belonging to the active component was larger than 0.5, we assigned this activity as active; otherwise, we assigned it as inactive.

Cancer/Cell-specific CE/SE identification

We selected cell-specific CEs if their active or inactive prevalences are below a threshold across 60 cancer cell lines (**S3 Fig**). For a given CE, active prevalence means the frequency of being active across cell lines, while inactive prevalence means the frequency of being inactive. We determined the prevalence threshold as follows. First, for a candidate prevalence threshold between 0 and 0.5, we selected cell-specific CEs with prevalence less than this threshold and performed hierarchical clustering of 60 cancer cell lines using the activities of the selected CEs. Second, the clustering results were compared to the true cell cluster information (i.e. cancer types) using the variation of information distances [41]. Variation of information is a metric to evaluate the similarity of two clustering results for the same groups of samples. The smaller the variation of information, the more similar the two clusterings are. Third, the variation of information for all candidate prevalence thresholds between 0 to 0.5 (at a step of 0.01) were calculated. Higher prevalence thresholds resulted in more selected CEs and smaller variation of information (i.e. better clustering). To balance the clustering efficiency and the cell specificity of the selected CEs, we selected the optimal prevalence threshold (~0.21) as the one corresponding to the inflection point on the curve plotted between variation of information and prevalence cutoffs. Here, inflection point is where the slope of smoothed curve equal to -1 on the plot. Finally, CEs with a active prevalence less than 0.21 were classified as cell-specific active CEs; CEs with a inactive prevalence less than 0.21 were classified as cell-specific inactive CEs.

Next, cell-specific CEs were used to define cancer-specific CEs with the rationale that cancer-specific CEs should be present in at least two cell lines for the same cancer type. For cancer types with only one cell line in the NCI-60 panel, cell-specific CEs were selected as cancer-specific CEs. Finally, SEs containing the cancer-specific CEs were summarized as cancer-specific SEs. As a result, SEs can be cancer-specific for different cancers depending on which cancers the CE components show specificity.

Activity evaluation of the specific and non-specific CEs

To test functional differences between cell-specific CEs and non-cell-specific CEs, we evaluated enhancer activity from independent sequencing assays including STARR-seq, PRO-seq, PRO-cap, GRO-seq, and GRO-cap (S1 Table). Four cancer cell lines with public data available were chosen for analysis: A549, K562, HCT-116, and MCF7. CEs that are specific-active, specific-inactive and non-specific-active in these cell lines were quantified for enhancer activity based on these assays. In detail, quality-controlled coverage signals of these assays were downloaded from ENCODE and GEO portals [27, 28]. Enhancer activity was quantified based on the total signal coverage at CEs using the *multiBigwigSummary* (with—binsize = 10 parameter) function from deepTools [42]. Both forward and reverse strand signals were aggregated together for nascent RNA sequencing datasets. Per-kilo-base enhancer activity was calculated using signals at CEs divided by CE widths and timed 1000.

Results

Super enhancer encodes epigenomic diversities across cancers

Super enhancers (SE) are involved in the cancer-specific regulation of essential genes in human cancers. The same SEs might function differently across cancers in terms of which and how gene targets are regulated [20,21,43]. To understand the SE divergence across human cancers, we revisited SE profiles for 28 cancer types based on the public H3K27Ac ChIP-seq data of the NCI-60 cancer cell lines [19]. In total, 11,100 genomic regions were identified as candidate SEs that showed significantly enriched H3K27Ac signals in at least one of the cancers (**Methods**). The widths of the SE regions ranged from 0.5Kb to 1.2Mb. The number of SEs varied dramatically across cancers (**S4 Fig**) with only a small portion (~38%) of SEs detected in more than 3 cancer types, confirming cancer-specific roles of SEs [44].

Recently, we and others demonstrated the functionally critical structures inside SEs [21,22,24], motivating the study of SE divergence across cancers at a higher resolution, i.e., at the constituent enhancer (CE) level. We applied our tool *DASE* [21] to compare SE constituent differences between pair-wise cell lines (1770 comparisons across 60 cell lines, with an average of 9530 CEs per pair and 7 CEs per SE). For a given SE, *DASE* statistically evaluated the CE patterns between cell lines by considering both the divergence between cell lines and the consistency of two replicates within each cell line. It reported differential SEs with statistical significance and differential types (e.g. activity or constituent change). On average, the majority of SEs (~72%, FDR <0.05 & at least one differential type detected) showed significant differences between pair-wise cell lines (**Fig 1A**). Such differences consist of the dismissal of a few CEs, the location drift of the SEs, and the activity discharging for the entire SE across cancers (**Fig 1B**).

To understand CE divergence across cancers, we first excluded lowly-active CEs (i.e., low H3K27Ac ChIP-seq coverage) that contribute $<3\%$ of overall activity to their SE (**Fig 1C**). We removed these CEs in downstream analysis for two reasons: 1) lowly-active CEs provide minor contribution to the nomination of SE presence as defined by main SE detection tools [32,33,45]; 2) low H3K27Ac signals may correspond to ChIP-seq artifacts due to background noise and biases. The resulted active CEs exhibit skewed distribution among cancer cell lines with more CEs active in less cancers (**Fig 1D**), consistent with cancer-specific functions of SEs. In addition, a small portion of CEs (460, or 0.95%) are consistently expressed across all cancer cell lines, suggesting their pan-cancer roles. Interestingly, the list of SEs (429, or 3.86%) containing at least one consistently expressed CE is significantly larger than the list of SEs that are shared intactly across all cancers (2, or 0.02%). This indicates that these SEs, although varied

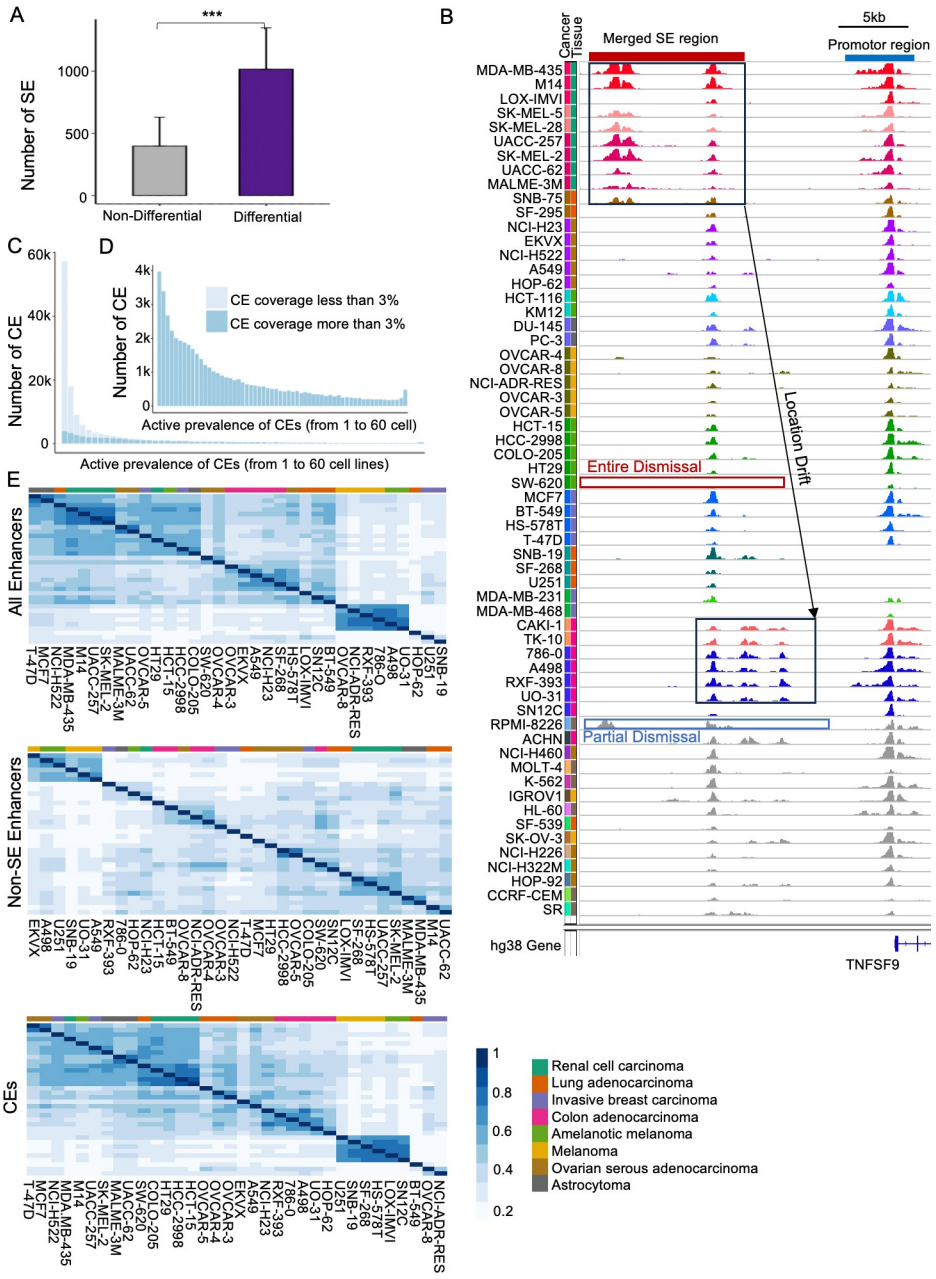

Fig 1. Summary of SE discrepancies at the CE level across cancers. A) Number of differential and non-differential SEs between pair-wise comparisons of NCI-60 cell lines. *Differential SEs were identified with at least one types* of differential changes (*FDR<0.05)* by *DASE* [21]. Error bars indicate standard deviation. (*** paired t-test p-value < 0.0001). **B)** An example of SE divergence across cancers involving dismissal and shifting of CEs. **C)** Frequency of CEs with different active prevalence across 60 cancer cell lines. **D)** The same as **C** but only based on CEs with high contribution to SE activity defined as H3K27Ac coverage greater than 3% of the total SE coverage. **E)** Hierarchical clustering of cancer cell lines based on the H3K27Ac-based activities of all enhancers (left), CEs with coverage greater than 3% of the SE total (middle), and non-SE enhancers (right).

https://doi.org/10.1371/journal.pcbi.1011873.g001

partially across cancers, express strong activity at their core CEs that might be critical to cancer regulation. Dissecting the core and unique CEs across cancers will maximize our understanding of SEs' functions.

With the elevated resolution of SE activities at the CE level, we evaluated how the SE landscape can encode cancer identities. We performed unsupervised clustering using H3K27Ac ChIP-seq signals for cancers containing more than three cell lines at three different sets of genome-wide enhancers: all enhancers, the selected CEs above, and non-SE enhancers (**Fig 1E**). We estimated clustering similarities by the three enhancer sets based on alignment score [46] (lower scores mean higher similarity). Interestingly, the selected CEs show similarity or even superior clustering of cancer samples to all enhancers (alignment score of 0.009), while non-SE enhancers provide insufficient power in grouping similar cancer cells (alignment score of 0.12 to all enhancers). This indicates the capability of SEs to encode cancer identities based on their divergent CE activities. On the contrary, it also demonstrates that the non-SE enhancers exhibit heavier discrepancies in a cell-type-specific instead of cancer-specific manner. It is noted that the slight improvement of clustering by the selected CEs over all enhancers might be attributed to the exclusion of non-cancer-specific signals (the non-SE enhancers) and other noise signals (the lowly-active CEs in **Fig 1C**).

Modeling super enhancer activities at the constituent level

To understand which CEs are involved in cancer-specific gene regulation, we compared CE activities across cancers. Previously, active CEs were defined by significant H3K27Ac ChIP-seq enrichment using peak calling algorithms for each cell line (**Fig 1D**). However, this approach might ignore critical CEs with low or marginal H3K27Ac ChIP-seq signals. First, ad hoc statistical cutoff (e.g., p-value$<$0.05) during peak calling brings artificial selection bias over marginal signals. Second, active consensus of enhancers by peak calling was built on genome-wide enhancer signals within a sample. It could underestimate small-size enhancers of which absolute activities are usually lower (**Fig 1B**). Third, peak calling with individual samples doesn't account for sample disparities, such as sequencing depth. A low-depth sequencing sample has less power to discriminate low CE signals from background noises (**S5 Fig**).

To provide an unbiased analysis of active CEs across cancer samples, we implemented two-component mixture models to estimate the active consensus for individual CEs (**Methods**). For a given CE region, H3K27Ac-based activity discrepancies are assumed among active samples due to reasons such as ChIP-seq measuring variability [47] (**Fig 2A**). Such discrepancies, however, should be smaller than those between the active and inactive samples. The mixture model examines the activity distribution of the CEs across all samples and converges the activities into two groups corresponding to the active and inactive states. In this case, the active consensus was built automatically based on the individual characteristics of the given CEs. We implemented one model for each CE separately. To ensure robust modeling, we normalized CE activity across the NCI-60 cell lines by scaling ChIP-seq coverage based on genome-wide enhancers (**Methods**). Zero signals were replaced with the half of sample-wise minimum signals (**Methods**).

Overall, most CEs (>99%) showed clear activity discrepancies between active and inactive cancer cell lines, while the rest of ~400 CE candidates expressed either pan-cancer activity or no activity across all samples (**Fig 2B and 2C**). The activity cutoffs to separate active and inactive samples estimated by mixture models varied across CEs (**Fig 2B**), confirming the importance of defining active CE consensus individually. Although the overall prevalence of active CEs (**S6 Fig**) does not change much from that based on peak calling (**Fig 1D**), numerous differences (> 11% of CE instances) exist in predictions of which samples these CEs are active (**Fig 2D**). 95 CE candidates with extremely low activity were identified as inactive across all samples (**Fig 2C**), suggesting they are among the false positive identifications from the peak calling method.

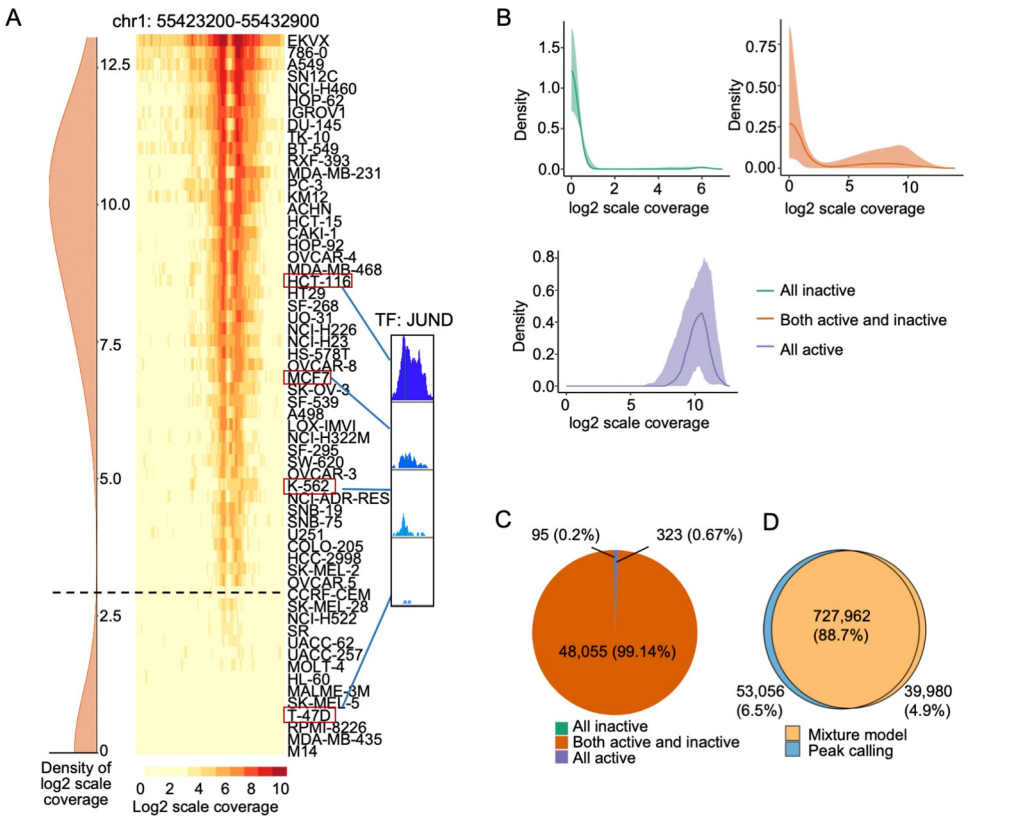

Fig 2. Mixture models discriminating active from inactive CEs. A) A CE example showing H3K27ac ChIP-seq coverage across 60 cell lines: left–mixture model-learned signal distribution of H3K27ac coverage; middle—coverage heatmap; right—a translation factor with binding signals correlated to the CE activity in four cell lines. **B)** Three types of CE activity densities by mixture models. Green group corresponds to CEs inactive across all cell lines; orange group corresponds to CEs active in some cell lines but inactive in others; purple group corresponds to CEs active in all cancer cell lines. The lines are the median values of the densities and the bands indicate 5% to 95% quantiles. **C)** The number of CEs belonging to each group in **B**. **D)** Comparisons of active CEs between mixture model and peak calling identification.

https://doi.org/10.1371/journal.pcbi.1011873.g002

Model-based states of active CEs provide improved functional interpretation

To evaluate model performance, we compared the estimated active CEs to those generated by peak calling. The active CEs by our model showed higher normalized H3K27Ac signals (log2 mean at 9.1) than those by peak calling (log2 mean at 5.1) and were comparable to the active CEs detected by both methods (**Fig 3A**). Further examination indicated the newly identified active CEs using the mixture models were those underestimated in single samples but with clear activities when compared across samples. POLR2A ChIA-PET sequencing can identify chromatin interactions between enhancers and promoters [48]. It allows us to functionally evaluate CE activities in regulating target genes. Using public POLR2A ChIA-PET datasets of four cancer cell lines [27], we found that the active CEs only by the mixture models interacted more frequently with cis-regions and linked to more target genes, compared to those by peak calling only (**Fig 3B and 3C**). More importantly, the linked target genes of active CEs by the mixture models showed higher expression levels compared to those by peak calling only (**Fig 3D**). Here, each gene's expression levels were quantified based on its expression ranking among 1450 cell lines documented in the DepMap data portal [29] (**Methods**). The strong

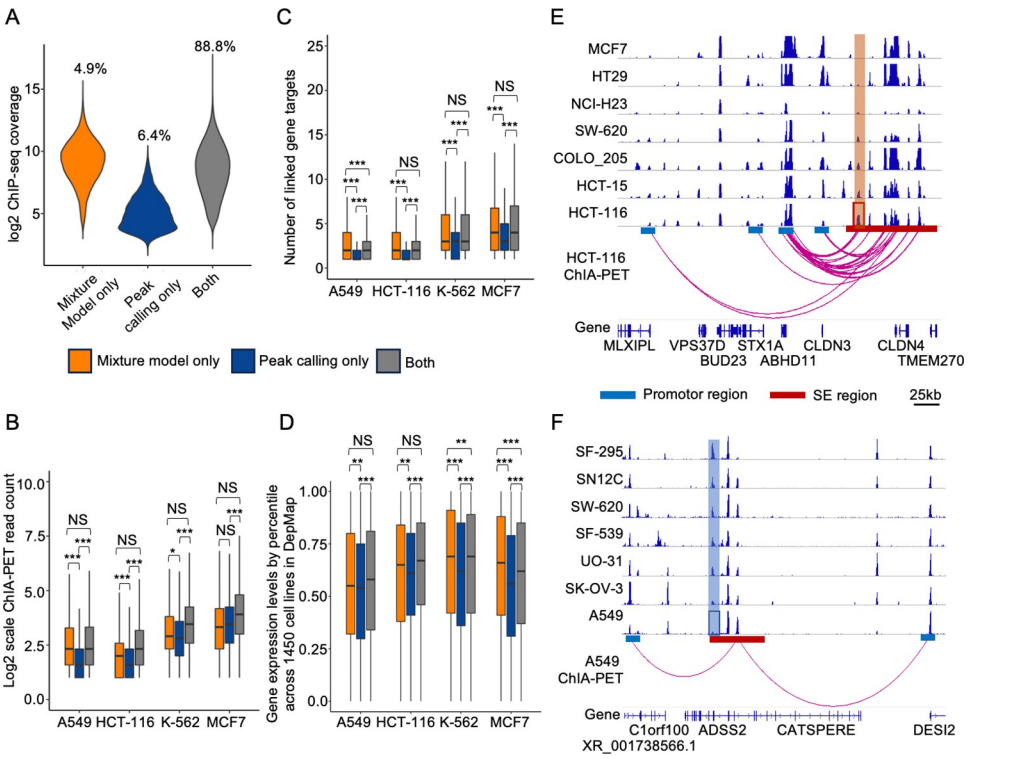

Fig 3. Mixture-model-based CE states provide better interpretation of CE activity. A) H3K27Ac ChIP-seq signals at active CEs predicted by different methods: only mixture model identified (orange), only peak calling identified (blue), and by both mixture model and peak calling (grey). Percentages of data points for each group are shown. **B-D)** Regulatory activities of the different groups of active CEs in **A**, as indicated by the number of associated ChIA-PET interactions (**B**), their linked targeted genes (**C**), and the expression of linked targeted genes (**D**) using ENCODE ChIA-PET datasets of four cancer cell lines. Outliers are ignored in the boxplots. Significance based on one-side Wilcoxon Rank Sum test is indicated: NS, non-significant; *, <0.05; **, <0.001; *** < 0.0001. **E-F)** Examples illustrating the improved sensitivity and specificity in detecting true active (E) and inactive (F) CEs by mixture models compared to peak calling. The red square is enhancer region under-estimated in HCT-116 (E) cell line and the blue square is enhancer region over-estimated in A549 (F) cell line by peak calling methods.

https://doi.org/10.1371/journal.pcbi.1011873.g003

activities of CEs only by mixture models are comparable to the active CEs identified by both mixture models and peak calling. This indicates that the mixture models were capable of prioritizing the highly active enhancers for individual cell lines.

We further demonstrated the improvement of false negative/positive predictions by modeling individual CEs across samples. For example, a candidate CE was identified as inactive in the HCT-116 cell line by peak calling due to marginal activity compared to other enhancers (Fig 3E). However, when examined across samples, the activity is stronger than those from other cell lines at the same location. Moreover, the CE also has strong evidence linking to the target gene ABHD11 in the HCT-116 cell line as illustrated by ChIA-PET data. This indicates the CE is functionally active in the HCT-116 cell line and highlights our model's ability to rescue the false negative calls made by the standard peak calling analysis. On the contrary, Fig 3F showed a candidate CE that was identified as active in the A549 cell line by peak calling. However, its activity is marginal and much weaker compared to the same location in other cell lines. Moreover, it shows no evidence of linking to target genes, suggesting no functional activity in the A549 cell lines. More examples are illustrated in S7 **and** S8 **Figs**. Together, our systematic mixture-model-based assessment across multiple cell lines reclassified these previously inaccurate predictions of CE states.

Cancer/cell-type-specific super enhancer signatures

We determined cancer-specific CEs by evaluating their prevalence across cancers as well as their capability to encode cancer identities. In brief, we used a variation of information [41] to estimate the grouping consistency between real cancer types and sample clustering by the selected CEs under a prevalence cutoff (**Methods**). A smaller variation of information indicates better consistency, or in other words, higher capability in encoding cancer identities by the selected CEs. A prevalence cutoff at 0.21 was chosen as the optimal to balance low variation of information and low prevalence (**Fig 4A**, **Methods**). It is noted that this threshold was also applied to select the inactive occurrences (i.e., CEs specifically lost in cancers).

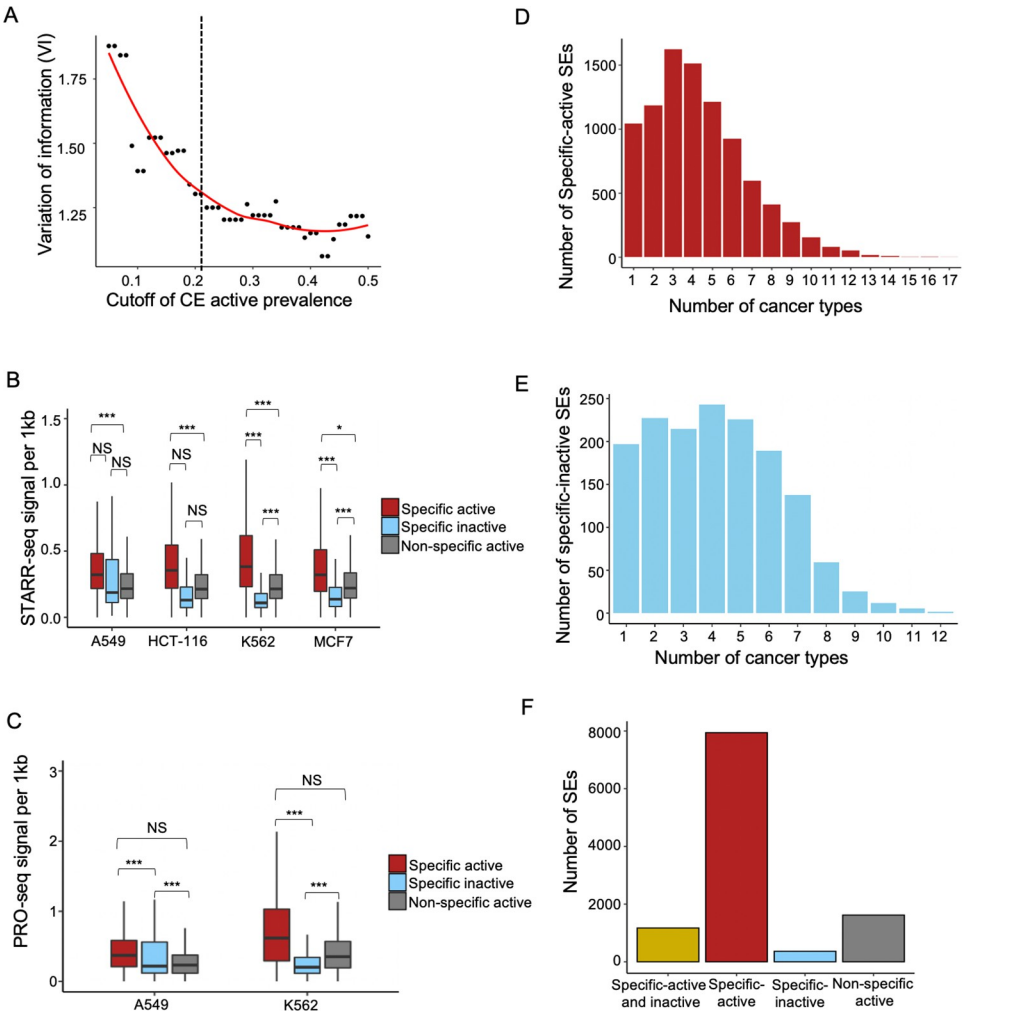

Fig 4. Cancer/cell-specific SE signatures. A) The capability of encoding cancer identities based on top cell-specific CEs. Running cutoffs of CE active prevalences were used to select top CEs. A cutoff of 0.21 was selected to define the final set of cell-specific CEs in downstream analysis. **B-C)** Per-kilobase enhancer activity measured by STARR-seq (B) and PRO-seq (C) for three types of CEs: cell-specific active (red); cell-specific inactive (blue) and non-specific active (grey). Significance based on student's t-test is indicated: NS, non-significant; *, <0.05; **, <0.001; *** < 0.0001. **D-E)** Number of SEs holding specific-active (D) or specific-inactive (E) cancer-specific CEs in different number of cancer types. Here, one SE may be counted multiple times depending on statuses of its multiple CEs. **F)** Number of SEs holding different types of cancer-specific CEs: SEs with both active and inactive cancer-specific CEs (gold); only active CEs (red); only inactive CEs (blue); SEs with no cancer-specific CEs (grey).

https://doi.org/10.1371/journal.pcbi.1011873.g004

We evaluated the true enhancer activity of the different groups of CEs (i.e., the specific-active, specific-inactive, and non-specific-active) in four selected cell lines based on independent sequencing assays. Two main types of assays were analyzed using public data [27,28], namely the massively parallel reporter assays (STARR-seq, **Fig 4B**) and nascent RNA sequencing (Pro-seq etc., **Figs 4C and S9**). STARR-seq is a reporter-gene-based assay implemented by constructing plasmid libraries that are introduced in cells to evaluate enhancer activity at millions of candidate DNA sequences [49]. Nascent RNA sequencing can evaluate enhancer activities by measuring the local transcriptional programs at enhancers [50–52]. Normalized sequencing signals were downloaded and quantified at individual CEs for the selected cell lines (**Methods**). The specific-active CEs showed the highest per-base activity while the specific-inactive CEs presented the lowest, and the observations were reproducible between STARR-seq and Pro-seq (**Fig 4B and 4C**). This suggests the critical regulatory roles of the specific-active SEs in the corresponding cell lines. Of note, we did not observe similar differences between the three CE groups when CE width was not justified (**S10 Fig**), in which case CE activity was confounded by the genomic sizes of CEs and wider CEs were overestimated by sequencing signals.

We further considered cancer-specific CEs as those specified in at least two of the cell lines for a given cancer type. We examined which SEs contain these specific CEs, both active and inactive. Overall, the number of SEs containing specific-active CEs (9114) is larger than the number of SEs holding specific-inactive CEs (1539) (**Fig 4D and 4E**). 1173 SEs hold specific-active and specific-inactive CEs simultaneously (**Fig 4F**), suggesting their versatile roles in different cancers. Moreover, 8,387 SEs contain multiple cancer-specific CEs that show activeness in different cancers (more than 2 cancer types). In other words, these SE are cancer-specific in multiple cancers according to the statuses of its different CE components. Together, the majority of SEs contain cancer-specific CEs, suggesting improved sensitivity of our methods in interrogating cancer-specific SE signatures.

A tool to visualize SE fingerprints across cancers

We summarized cancer-specific SEs based on their CE statuses for all cancer types in the NCI-60 cell panel (**Fig 5A**). As illustrated (**Figs 1E** and **4A**), these SEs are highly informative and can act as fingerprints in encoding cancer identities. We further implemented an R package *cSEAdb (https://github.com/tenglab/cSEAdb)*, which facilitates the query, exploration, and visualization of these SE fingerprints. The R package allows queries of cancer-specific SEs using four types of information: cancer types, cancer cell types, gene targets of SEs, or SE locations. For a query by cancer or cell types, the package returns the list of cancer-specific active and inactive SEs as well as an overview of fingerprint SEs across the genome (**Fig 5B and 5C**). For a query by gene targets, any nearby SE in at least one of the NCI-60 cells will be returned. Further zooming into individual SEs is allowed based on all query types above. When zooming into an SE or querying by an SE location, the details of SE activity are visualized with color-coded highlights of its cancer-specific CE components across cancers (**Fig 5D**).

Discussion

In this study, we systematically investigated super enhancer (SE) profiles across human cancers based on NCI-60 cancer cell lines and generated cancer-specific SE signatures at the resolution of constituent enhancers (CE). Through computational modeling of the varied CE activities across cell lines and individual CEs, we provided an improved estimation of active CEs for cancers. We evaluated the estimations using multiple functional assays associated with enhancers, including POLR2A ChIA-PET, STARR-seq, PRO-seq etc., and demonstrated that the modeling process improves the functional interpretation of SE activity in cancers. We further

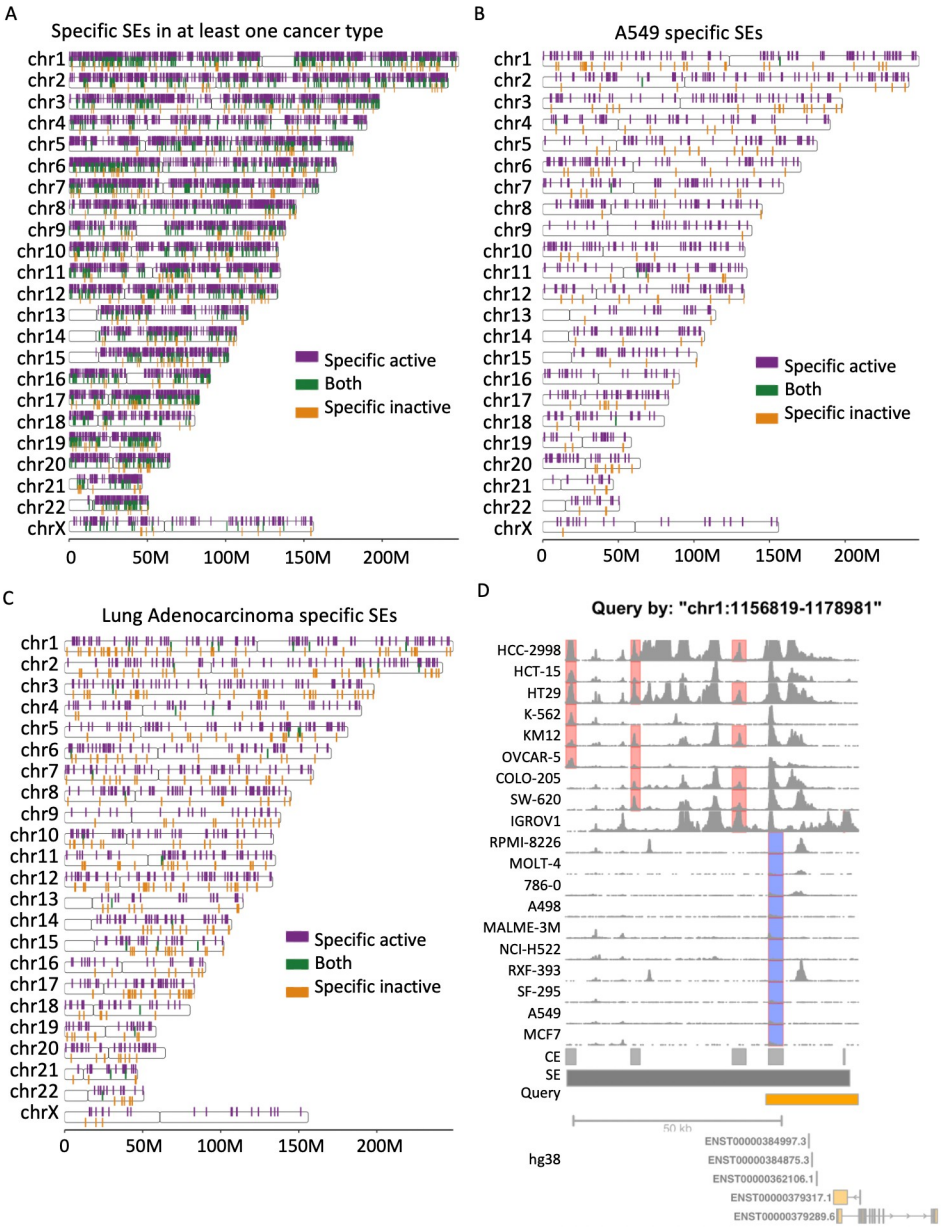

Fig 5. A tool for exploring fingerprint SE signatures. A) Genome-wide SE landscape with cancer-specific CEs across all studied cancer types. **B)** Genome-wide SE landscape with cell-specific CEs in A549 cell line. **C)** Genome-wide SE landscape with cancer-specific CEs in lung adenocarcinoma. **D)** H3K27Ac ChIP-seq signals for a SE example with cancer-specific active (red) and inactive (blue) CEs highlighted.

https://doi.org/10.1371/journal.pcbi.1011873.g005

showed that cancer-specific SE signatures encode cancer identities and have uniquely strong per-base activity in the corresponding cancers. To facilitate community exploration of our identified SE signatures, we built an easy-to-use R package to document, query and visualize SEs in cancers. Together, our algorithms provide a novel insight to model and interpret cancer-specific fingerprint SE signatures (S11 Fig).

It is interesting to observe that cell-specific CEs exhibit the strongest per-base instead of whole-enhancer activity. On one hand, this suggests the imperative evidence that cell-specific CEs are among the key regulators in the corresponding cell lines. Generating cancer-specific

CEs can provide insights to the essential epigenomic signatures in cancers. On the other hand, it highlights the importance of unravelling CE activity with its genomic size. Enhancer regions are usually involved with multiple transcription factor binding. Enhancer size is likely reflecting the complexity of local collaborations between transcription factors instead of regulatory activity [53, 54]. The per-base activity thus may better represent the true regulatory activity induced by key transcription factors.

Cell line models may exhibit different SE profiles compared to cancer tissues [43], introducing a possible challenge to the translation of these findings. However, cancer tissues are usually heterogeneous, containing immune cells, normal cells and cancer clones at different progression stages. A future application of this work will be comparison of these cell line-based SE profiles to the deconvoluted tumor SE profiles. However, this is challenging given the limited data that exists on tissue-based single-cell-type enhancer activity. One possible first step may be use of the recent single cell CUT&Tag experiments data [55–57]. Here, we optimized the utility of our findings by use of multiple cell lines of the same cancer types when building our cancer-specific SE signatures, thus providing an additional layer of robustness toward reproducible cancer-specific signatures for a given cancer type. It is noted that only one cell line is available for half of the cancer types in the NCI-60 panel, cancer-specific signatures from these cancers might confound with cell-specific or tissue-specific signatures. Our next plan is to add more batch-effect controlled data from other cancer and normal tissue cell lines to improve our fingerprint signatures.

Our study is the first to provide cancer-specific-inactive signatures across cancers. Interestingly, we observed that their frequency is lower than the cancer-specific-active signatures. More precise interpretation of the functional impact of these signatures is currently challenging, but we suspect that the specific disappearances of these enhancers may correspond to missing activities of transcription factors at these loci. Further investigations are needed to fully understand the function of these cancer-specific lost signatures. In addition, we anticipate that future functional annotation of both specific active and inactive CEs will promote the use of the presented resource here by a broader community of the field. For instance, the highlighted CE in S8B Fig is actually a cell-specific active signature in the K562 cell line (**S12 Fig**). Compared to A549 cells, this CE exhibits unique regulation to target genes ANKRD9 and RCOR1 in K562 cells. Full characterization of such unique regulation will lead to a better understanding of SE-involved cancer mechanisms.

The proposed mixture model is robust in identifying the few uniformly inactive CE elements that were classified as active by peak calling. These CEs showed extremely weak activities with no linking to any genes based on ChIA-PET, suggesting the accuracy of our mixture models. We identified pan-cancer active core CEs for all cancers, which may provide critical insight into functional knowledge of routine cancer cell maintenance by SE-associated regulation. Although beyond the scope of this study, our mixture model can also be applied to estimate the non-SE enhancer activity states across samples.

In summary, our pan-cancer analysis demonstrates improved accuracy in identifying active CEs and elevated sensitivity in detecting cancer-specific SE signatures. The R package will facilitate the cancer-specific exploration of potential therapeutic targets in epigenomics.

Supporting information

S1 Fig. Median width and total number of merged SEs and CEs based on different overlap cutoffs. Dashed line indicates 25% overlap cutoff. a. SE median width b. Total number of SEs c. CE median width d. Total number of CEs.
(PDF)

S2 Fig. Genome-wide priors of the inactive and active enhancer groups estimated across 60 cancer cell lines. Data points indicate the estimated means of the lower and higher mixtures in individual cell lines.
(PDF)

S3 Fig. Flowchart to define cell/cancer-specific CEs.
(PDF)

S4 Fig. Number of SEs identified in different cancers.
(PDF)

S5 Fig. Vertical comparison of H3K27Ac signals across cancers refines enhancer activity consensus in individual samples. Left: normalized ChIP-seq signals recovers enhancer activity in MDA-MB-231 cell line; right: raw signals indicate no active enhancers at the same location due to low ChIP-seq coverage in MDA-MB-231.
(PDF)

S6 Fig. Overall prevalence of active CEs identified by mixture models and peak calling.
(PDF)

S7 Fig. Extra examples illustrating the improved sensitivity in detecting true active CEs by mixture models compared to peak calling. (a). Red shaded square is an enhancer region under-estimated in A549 cell line by peak calling methods but identified as active by mixture model. This region shows strong enhancer activity in A549 compared to other cancer cell lines and presents regulatory interactions with target gene TRMT5 based on ChIA-PET data. (b). Similar to a) but for another enhancer region regulating SIX4 gene in A549 cell line. (c). Similar to a) but for another enhancer region regulating GNAI3 gene in HCT-116 cell line. (d). Similar to a) but for another enhancer region regulating HID1, MRPL58 and KCTD2 in HCT-116 cell line.
(PDF)

S8 Fig. Extra examples illustrating the improved specificity in detecting true inactive CEs by mixture models compared to peak calling. Blue shaded square is an enhancer region over-estimated in A549 cell line by peak calling methods but identified as inactive by mixture model. This region shows weak enhancer activity in A549 compared to other cancer cell lines and presents no regulatory interactions with any genes based on ChIA-PET data. (b). Similar to a) but for another enhancer region in A549 cell line. (c). Similar to a) but for another enhancer region in HCT-116 cell line. (d). Similar to a) but for another enhancer region in MCF7 cell line.
(PDF)

S9 Fig. Enhancer activity from extra nascent RNA sequencing datasets for K562, including GRO- cap, Pro-cap and GRO-seq, from the ENCODE and GEO data repositories.
(PDF)

S10 Fig. STARR-seq and PRO-seq based CE activity without normalization of CE width. a. STARR-seq. b. PRO-seq.
(PDF)

S11 Fig. Flowchart of identifying fingerprint SE signatures across cancers.
(PDF)

S12 Fig. An example of interpreting cancer-specific active CEs. Highlighted in blue square is a cell- specific active CE in K562 but inactive in A549. This CE links to promoters of two

gene targets, ANKRD9 and RCOR1, suggesting its cell-specific regulation in K562.
(PDF)

S1 Table. Accession IDs and meta information for the publicly downloaded data in this manuscript.
(XLSX)

S1 Data. Numberical values to generate manuscript graphs and histograms.
(XLSX)

Author Contributions

Conceptualization: Derek R. Duckett, Mingxiang Teng.

Formal analysis: Xiang Liu.

Funding acquisition: Bo Zhao, Lixin Wan, Derek R. Duckett.

Methodology: Xiang Liu, Nancy Gillis, Chang Jiang, Anthony McCofie, Timothy I. Shaw, Mingxiang Teng.

Software: Xiang Liu, Anthony McCofie.

Supervision: Aik-Choon Tan, Bo Zhao, Lixin Wan, Derek R. Duckett, Mingxiang Teng.

Writing – original draft: Xiang Liu, Mingxiang Teng.

Writing – review & editing: Xiang Liu, Nancy Gillis, Chang Jiang, Timothy I. Shaw, Aik-Choon Tan, Bo Zhao, Lixin Wan, Derek R. Duckett, Mingxiang Teng.

References

1. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007; 446(7132):153–8. https://doi.org/10.1038/nature05610 PMID: 17344846; PubMed Central PMCID: PMC2712719.
2. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020; 578(7793):94–101. Epub 20200205. https://doi.org/10.1038/s41586-020-1943-3 PMID: 32025018; PubMed Central PMCID: PMC7054213.
3. Steele CD, Abbasi A, Islam SMA, Bowes AL, Khandekar A, Haase K, et al. Signatures of copy number alterations in human cancer. Nature. 2022; 606(7916):984–91. Epub 20220615. https://doi.org/10.1038/s41586-022-04738-6 PMID: 35705804; PubMed Central PMCID: PMC9242861.
4. Fernandez AF, Assenov Y, Martin-Subero JI, Balint B, Siebert R, Taniguchi H, et al. A DNA methylation fingerprint of 1628 human samples. Genome Res. 2012; 22(2):407–19. Epub 20110525. https://doi.org/10.1101/gr.119867.110 PMID: 21613409; PubMed Central PMCID: PMC3266047.
5. Zhao SG, Chen WS, Li H, Foye A, Zhang M, Sjostrom M, et al. The DNA methylation landscape of advanced prostate cancer. Nat Genet. 2020; 52(8):778–89. Epub 20200713. https://doi.org/10.1038/s41588-020-0648-8 PMID: 32661416; PubMed Central PMCID: PMC7454228.
6. Fang C, Wang Z, Han C, Safgren SL, Helmin KA, Adelman ER, et al. Cancer-specific CTCF binding facilitates oncogenic transcriptional dysregulation. Genome Biol. 2020; 21(1):247. Epub 20200915. https://doi.org/10.1186/s13059-020-02152-7 PMID: 32933554; PubMed Central PMCID: PMC7493976.
7. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, et al. Super-enhancers in the control of cell identity and disease. Cell. 2013; 155(4):934–47. Epub 2013/10/15. https://doi.org/10.1016/j.cell.2013.09.053 PMID: 24119843; PubMed Central PMCID: PMC3841062.
8. Bradner JE, Hnisz D, Young RA. Transcriptional Addiction in Cancer. Cell. 2017; 168(4):629–43. https://doi.org/10.1016/j.cell.2016.12.013 PMID: 28187285; PubMed Central PMCID: PMC5308559.
9. Sengupta S, George RE. Super-Enhancer-Driven Transcriptional Dependencies in Cancer. Trends Cancer. 2017; 3(4):269–81. Epub 20170412. https://doi.org/10.1016/j.trecan.2017.03.006 PMID: 28718439; PubMed Central PMCID: PMC5546010.

10. Hnisz D, Schuijers J, Lin CY, Weintraub AS, Abraham BJ, Lee TI, et al. Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. Mol Cell. 2015; 58(2):362–70. Epub 20150319. https://doi.org/10.1016/j.molcel.2015.02.014 PMID: 25801169; PubMed Central PMCID: PMC4402134.

11. Wang C, Zhang L, Ke L, Ding W, Jiang S, Li D, et al. Primary effusion lymphoma enhancer connectome links super-enhancers to dependency factors. Nat Commun. 2020; 11(1):6318. Epub 2020/12/11. https://doi.org/10.1038/s41467-020-20136-w PMID: 33298918; PubMed Central PMCID: PMC7726151.

12. Bahr C, von Paleske L, Uslu VV, Remeseiro S, Takayama N, Ng SW, et al. A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. Nature. 2018; 553(7689):515–20. Epub 20180117. https://doi.org/10.1038/nature25193 PMID: 29342133.

13. Schuijers J, Manteiga JC, Weintraub AS, Day DS, Zamudio AV, Hnisz D, et al. Transcriptional Dysregulation of MYC Reveals Common Enhancer-Docking Mechanism. Cell Rep. 2018; 23(2):349–60. Epub 2018/04/12. https://doi.org/10.1016/j.celrep.2018.03.056 PMID: 29641996; PubMed Central PMCID: PMC5929158.

14. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 2018; 173(2):291–304 e6. Epub 2018/04/07. https://doi.org/10.1016/j.cell.2018.03.022 PMID: 29625048; PubMed Central PMCID: PMC5957518.

15. Chen H, Li C, Peng X, Zhou Z, Weinstein JN, Cancer Genome Atlas Research Network, et al. A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples. Cell. 2018; 173(2):386–99 e12. https://doi.org/10.1016/j.cell.2018.03.027 PMID: 29625054; PubMed Central PMCID: PMC5890960.

16. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018; 173(2):371–85 e18. https://doi.org/10.1016/j.cell.2018.02.060 PMID: 29625053; PubMed Central PMCID: PMC6029450.

17. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell. 2018; 173(2):321–37 e10. https://doi.org/10.1016/j.cell.2018.03.035 PMID: 29625050; PubMed Central PMCID: PMC6070353.

18. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020; 578(7793):82–93. Epub 20200205. https://doi.org/10.1038/s41586-020-1969-6 PMID: 32025007; PubMed Central PMCID: PMC7025898.

19. Gopi LK, Kidder BL. Integrative pan cancer analysis reveals epigenomic variation in cancer type and cell specific chromatin domains. Nat Commun. 2021; 12(1):1419. Epub 2021/03/05. https://doi.org/10.1038/s41467-021-21707-1 PMID: 33658503; PubMed Central PMCID: PMC7930052.

20. Lidschreiber K, Jung LA, von der Emde H, Dave K, Taipale J, Cramer P, et al. Transcriptionally active enhancers in human cancer cells. Mol Syst Biol. 2021; 17(1):e9873. https://doi.org/10.15252/msb.20209873 PMID: 33502116; PubMed Central PMCID: PMC7838827.

21. Liu X, Zhao B, Shaw TI, Fridley BL, Duckett DR, Tan AC, et al. Summarizing internal dynamics boosts differential analysis and functional interpretation of super enhancers. Nucleic Acids Res. 2022; 50(6):3115–27. Epub 2022/03/03. https://doi.org/10.1093/nar/gkac141 PMID: 35234924; PubMed Central PMCID: PMC8989535.

22. Huang J, Li K, Cai W, Liu X, Zhang Y, Orkin SH, et al. Dissecting super-enhancer hierarchy based on chromatin interactions. Nat Commun. 2018; 9(1):943. Epub 2018/03/07. https://doi.org/10.1038/s41467-018-03279-9 PMID: 29507293; PubMed Central PMCID: PMC5838163.

23. Amaral PP, Bannister AJ. Re-place your BETs: the dynamics of super enhancers. Mol Cell. 2014; 56(2):187–9. Epub 2014/11/07. https://doi.org/10.1016/j.molcel.2014.10.008 PMID: 25373538.

24. Allahyar A, Vermeulen C, Bouwman BAM, Krijger PHL, Verstegen M, Geeven G, et al. Enhancer hubs and loop collisions identified from single-allele topologies. Nat Genet. 2018; 50(8):1151–60. Epub 20180709. https://doi.org/10.1038/s41588-018-0161-5 PMID: 29988121.

25. Kai Y, Li BE, Zhu M, Li GY, Chen F, Han Y, et al. Mapping the evolving landscape of super-enhancers during cell differentiation. Genome Biol. 2021; 22(1):269. Epub 20210915. https://doi.org/10.1186/s13059-021-02485-x PMID: 34526084; PubMed Central PMCID: PMC8442463.

26. Li T, Jia L, Cao Y, Chen Q, Li C. OCEAN-C: mapping hubs of open chromatin interactions across the genome reveals gene regulatory networks. Genome Biol. 2018; 19(1):54. Epub 20180424. https://doi.org/10.1186/s13059-018-1430-4 PMID: 29690904; PubMed Central PMCID: PMC5926533.

27. Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. 2020; 48(D1):D882–D9. Epub 2019/11/13. https://doi.org/10.1093/nar/gkz1062 PMID: 31713622; PubMed Central PMCID: PMC7061942.

28. Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat Genet. 2014; 46 (12):1311–20. Epub 20141110. https://doi.org/10.1038/ng.3142 PMID: 25383968; PubMed Central PMCID: PMC4254663.

29. Ghandi M, Huang FW, Jane-Valbuena J, Kryukov GV, Lo CC, McDonald ER, 3rd, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019; 569(7757):503–8. Epub 20190508. https://doi.org/10.1038/s41586-019-1186-3 PMID: 31068700; PubMed Central PMCID: PMC6697103.

30. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9(4):357–9. Epub 2012/03/06. https://doi.org/10.1038/nmeth.1923 PMID: 22388286; PubMed Central PMCID: PMC3322381.

31. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9(9):R137. Epub 2008/09/19. https://doi.org/10.1186/gb-2008-9-9-r137 PMID: 18798982; PubMed Central PMCID: PMC2592715.

32. Loven J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell. 2013; 153(2):320–34. Epub 2013/04/16. https://doi.org/10.1016/j.cell.2013.03.036 PMID: 23582323; PubMed Central PMCID: PMC3760967.

33. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013; 153(2):307–19. Epub 2013/04/16. https://doi.org/10.1016/j.cell.2013.03.035 PMID: 23582322; PubMed Central PMCID: PMC3653129.

34. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep. 2019; 9(1):9354. Epub 2019/06/30. https://doi.org/10.1038/s41598-019-45839-z PMID: 31249361; PubMed Central PMCID: PMC6597582.

35. Mantsoki A, Parussel K, Joshi A. Identification and Characterisation of Putative Enhancer Elements in Mouse Embryonic Stem Cells. Bioinform Biol Insights. 2021; 15:1177932220974623. Epub 20210209. https://doi.org/10.1177/1177932220974623 PMID: 33623376; PubMed Central PMCID: PMC7876754.

36. Sethi A, Gu M, Gumusgoz E, Chan L, Yan KK, Rozowsky J, et al. Supervised enhancer prediction with epigenetic pattern recognition and targeted validation. Nat Methods. 2020; 17(8):807–14. Epub 20200729. https://doi.org/10.1038/s41592-020-0907-8 PMID: 32737473; PubMed Central PMCID: PMC8073243.

37. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014; 30(7):923–30. Epub 2013/11/15. https://doi.org/10.1093/bioinformatics/btt656 PMID: 24227677.

38. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15(12):550. Epub 2014/12/18. https://doi.org/10.1186/s13059-014-0550-8 PMID: 25516281; PubMed Central PMCID: PMC4302049.

39. Benaglia T, Chauveau D, Hunter DR, Young DS. mixtools: An R Package for Analyzing Finite Mixture Models. J Stat Softw. 2009; 32(6):1–29. WOS:000271534100001.

40. Muthen B, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics. 1999; 55(2):463–9. https://doi.org/10.1111/j.0006-341x.1999.00463.x PMID: 11318201.

41. Meila M. Comparing clusterings by the variation of information. Lect Notes Artif Int. 2003; 2777:173–87. https://doi.org/10.1007/978-3-540-45167-9_14 WOS:000185937100013.

42. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research. 2016; 44(W1):W160–W5. https://doi.org/10.1093/nar/gkw257 WOS:000379786800027. PMID: 27079975

43. Wang X, Cairns MJ, Yan J. Super-enhancers in transcriptional regulation and genome organization. Nucleic Acids Res. 2019; 47(22):11481–96. https://doi.org/10.1093/nar/gkz1038 PMID: 31724731; PubMed Central PMCID: PMC7145697.

44. Zhang X, Choi PS, Francis JM, Imielinski M, Watanabe H, Cherniack AD, et al. Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. Nat Genet. 2016; 48(2):176–82. Epub 20151214. https://doi.org/10.1038/ng.3470 PMID: 26656844; PubMed Central PMCID: PMC4857881.

45. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010; 38(4):576–89. Epub 2010/06/02. https://doi.org/10.1016/j.molcel.2010.05.004 PMID: 20513432; PubMed Central PMCID: PMC2898526.

46. Galili T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics. 2015; 31(22):3718–20. Epub 20150723. https://doi.org/10.1093/bioinformatics/btv428 PMID: 26209431; PubMed Central PMCID: PMC4817050.

47. Teng M, Du D, Chen D, Irizarry RA. Characterizing batch effects and binding site-specific variability in ChIP-seq data. NAR Genom Bioinform. 2021; 3(4):lqab098. Epub 2021/10/19. https://doi.org/10.1093/nargab/lqab098 PMID: 34661103; PubMed Central PMCID: PMC8515842.

48. Li G, Fullwood MJ, Xu H, Mulawadi FH, Velkov S, Vega V, et al. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol. 2010; 11(2):R22. Epub 20100225. https://doi.org/10.1186/gb-2010-11-2-r22 PMID: 20181287; PubMed Central PMCID: PMC2872882.

49. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013; 339(6123):1074–7. Epub 20130117. https://doi.org/10.1126/science.1232542 PMID: 23328393.

50. Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science. 2013; 339(6122):950–3. https://doi.org/10.1126/science.1229386 PMID: 23430654; PubMed Central PMCID: PMC3974810.

51. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008; 322(5909):1845–8. Epub 20081204. https://doi.org/10.1126/science.1162228 PMID: 19056941; PubMed Central PMCID: PMC2833333.

52. Kruesi WS, Core LJ, Waters CT, Lis JT, Meyer BJ. Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. Elife. 2013; 2:e00808. Epub 20130618. https://doi.org/10.7554/eLife.00808 PMID: 23795297; PubMed Central PMCID: PMC3687364.

53. Panigrahi A, O'Malley BW. Mechanisms of enhancer action: the known and the unknown. Genome Biol. 2021; 22(1):108. Epub 2021/04/17. https://doi.org/10.1186/s13059-021-02322-1 PMID: 33858480; PubMed Central PMCID: PMC8051032.

54. Spitz F, Furlong EE. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012; 13(9):613–26. Epub 20120807. https://doi.org/10.1038/nrg3207 PMID: 22868264.

55. Kaya-Okur HS, Wu SJ, Codomo CA, Pledger ES, Bryson TD, Henikoff JG, et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat Commun. 2019; 10(1):1930. Epub 2019/05/01. https://doi.org/10.1038/s41467-019-09982-5 PMID: 31036827; PubMed Central PMCID: PMC6488672.

56. Bartosovic M, Kabbe M, Castelo-Branco G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. Nat Biotechnol. 2021; 39(7):825–35. Epub 20210412. https://doi.org/10.1038/s41587-021-00869-9 PMID: 33846645; PubMed Central PMCID: PMC7611252.

57. Wu SJ, Furlan SN, Mihalas AB, Kaya-Okur HS, Feroze AH, Emerson SN, et al. Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. Nat Biotechnol. 2021; 39(7):819–24. Epub 20210412. https://doi.org/10.1038/s41587-021-00865-z PMID: 33846646; PubMed Central PMCID: PMC8277750.