Identification of Cancer-specific Constituent Elements inside Super-enhancers (cSEAdb)

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

[1]Moffitt Cancer Center; [2]Huntsman Cancer Institute, The University of Utah; [3]Brigham and Women's Hospital and Harvard Medical School

xiang.liu

JAN 04, 2024

OPEN ACCESS

DOI: dx.doi.org/10.17504/protocols.io.kxygx38wzg8j/v1

Protocol Citation: Xiang Liu, Nancy Gillis, Chang Jiang, Anthony McCofie, Timothy I Shaw, Aik-Choon Tan, Bo Zhao, Lixin Wan, Derek R Duckett, Mingxiang Teng 2024. Identification of Cancer-specific Constituent Elements inside Super-enhancers (cSEAdb). **protocols.io** https://dx.doi.org/10.17504/protocols.io.kxygx38wzg8j/v1

Protocol status: Working
We use this protocol and it's working

Created: Jan 03, 2024

Last Modified: Jan 04, 2024

PROTOCOL integer ID: 92932

DISCLAIMER

ABSTRACT

Super enhancers (SE) are large genomic elements composed of multiple constituent enhancers. As super enhancers are key regulators associated to cancer mechanisms, identifying cancer-specific super enhancer signatures improves our understanding of cancer-associated gene regulation. This protocol aims to provide a computational framework to identify cancer-specific super enhancer signatures at their constituent levels, using public H3L27Ac ChIP-seq data of the NCI-60 cancer cell panel. The protocol covers from data acquisition, pre-processing, statistical modeling and cancer-specific signature identification. It also provides links of scripts toward building an R data object for the storage, management and query of these signatures.

Funders Acknowledgement:
NIH
Grant ID: R03DE030580
NIH
Grant ID: R01CA262530
NIH
Grant ID: R01CA255398
NIH
Grant ID: R01AI123420
Moffitt Cancer Center
Grant ID: P30CA076292

Enhancer and super-enhancer identification in each cancer cell line/sample

1 H3K27Ac ChIP-seq data acquisition

Download raw H3K27ac ChIP-seq data of NCI-60 human cancer cell lines from GEO repositories with accession ID GSE143653. Fastq files of 60 cancer cell lines with two H3K27Ac replicates and one INPUT replicate should be downloaded from the SRA repository linked to this GEO accession.

Dataset

NCI-60 ChIP-seq GEO link NAME

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE143653 LINK

Dataset

NCI-60 ChIP-seq SRA link NAME

https://www.ncbi.nlm.nih.gov/sra?term=SRP241932 LINK

CITATION

Gopi LK, Kidder BL (2021). Integrative pan cancer analysis reveals epigenomic variation in cancer type and cell specific chromatin domains.. LINK
https://doi.org/10.1038/s41467-021-21707-1

2 Enhancer and super-enhancer (SE) identification for each downloaded sample

2.1 Align raw ChIP-seq reads using Bowtie2

Align raw H3K27Ac ChIP-seq data of the NCI-60 cancer cell lines to human genome hg38 using Bowtie2.

CITATION

Langmead B, Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2..
LINK
https://doi.org/10.1038/nmeth.1923

Software

Bowtie2 NAME

http://bowtie-bio.sourceforge.net/bowtie2/index.shtml SOURCE LINK

Generate parallel bowtie2 alignment using the following script:

https://github.com/tenglab/cSEAdb_plos_code/blob/main/chip_seq/01_make_bowtie2_wlist.sh

Use the command below to implement Bowtie2 parallel alignment for all ChIP-seq samples.

Command

GNU Parallel

```
parallel -j 4 -k < $BOWTIE2_WLIS
T
```

2.2 **Remove sequencing duplicate reads and sort the remaining reads**

Use the following customized script to remove duplicated reads (minimizing potential PCR artifacts in the sequencing data) and sort sequencing reads (for downstream peak calling) in the aligned files.

https://github.com/tenglab/cSEAdb_plos_code/blob/main/chip_seq/02_sam_to_sort_bam.py

2.3 **Call peaks (enhancer candidates) of H3K27Ac data using MACS2**

Call peaks based on the aligned BAM files using tool MACS2. The called peaks are treated as potential enhancer candidates for downstream analysis and for super enhancer identification.

CITATION

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS (2008). Model-based analysis of ChIP-Seq (MACS)..
LINK
https://doi.org/10.1186/gb-2008-9-9-r137

Generate parallel peak calling using the following script.

https://github.com/tenglab/cSEAdb_plos_code/blob/main/chip_seq/03_make_macs2_wlist.sh

Use the command below to implement MACS2 parallel peak calling for all ChIP-seq samples.

Command

GNU Parallel

```
parallel -j 4 -k < $MACS2_WLIST
```

2.4 Convert file formats of MACS2 output for SE analysis

In order to call SEs on each ChIP-seq sample, some output files of MACS2 need to be re-formatted for fit SE calling software. Particularly, the ".narrowPeak" files from MACS2 have to be converted to ".gff" format using the following command.

Command

Convert .narrowPeak to .gff files for ROSE

```
awk '{print $1"\t"$4"\t""\t"$2"\t"$3"\t""\t"$6"\t""\t"$4}' $PROJECT_DIR/macs_out/*_peaks.narrowPeak > $PROJECT_DIR/macs_out/*_peaks.gff
```

2.5 SE identification using ROSE

For each cell line/sample, call super enhancers using the tool ROSE (https://hpc.nih.gov/apps/ROSE.html) based on files generated in the previous steps, including BAM files of H3K27Ac and INPUT, gff files and hg38 genome build.

CITATION

Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes..

LINK

https://doi.org/10.1016/j.cell.2013.03.035

CITATION

Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, Bradner JE, Lee TI, Young RA (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers..

LINK

https://doi.org/10.1016/j.cell.2013.03.036

Generate parallel super enhancer calling using the following script.

https://github.com/tenglab/cSEAdb_plos_code/blob/main/chip_seq/05_make_rose_wlist.sh

Use the command below to implement ROSE parallel peak calling for all ChIP-seq samples.

Command

GNU Parallel

```
parallel -j 4 -k < $ROSE_WLIST
```

Identification of cancer-specific super enhancer elements, documented in *cSEAdb*

3 Remove H3K27Ac ChIP-seq peaks overlapping with gene promotors and ChIP-seq blacklist regions

To avoid biases of promoter H3K27Ac signals, remove ChIP-seq peaks overlapping with gene promotors and blacklist regions using the following scripts. Promoter regions were defined as upstream 3000bp to downstream 1000bp surrounding genes' transcription start sites.

https://github.com/tenglab/cSEAdb_plos_code/blob/main/analysis/00_remove_gene_region_enhancer.R

CITATION

Amemiya HM, Kundaje A, Boyle AP (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome..
LINK
https://doi.org/10.1038/s41598-019-45839-z

4 Merge regions of different sample to generate a unique list of candidate enhancers and SEs

Merge enhancers and SEs across all cancer cell line samples using the scripts below. Basically, candidate enhancers and SEs from different samples might overlap with each other. The overlapped regions might indicate the same enhancer/SE regions across different cancer cells. We suggest only to merge consecutive regions with an overlap width at 25%.

https://github.com/tenglab/cSEAdb_plos_code/blob/main/analysis/01_merge_enhancer_se.R

CITATION

Mantsoki A, Parussel K, Joshi A (2021). Identification and Characterisation of Putative Enhancer Elements in Mouse Embryonic Stem Cells..
LINK
https://doi.org/10.1177/1177932220974623

CITATION

Sethi A, Gu M, Gumusgoz E, Chan L, Yan KK, Rozowsky J, Barozzi I, Afzal V, Akiyama JA, Plajzer-Frick I, Yan C, Novak CS, Kato M, Garvin TH, Pham Q, Harrington A, Mannion BJ, Lee EA, Fukuda-Yuzawa Y, Visel A, Dickel DE, Yip KY, Sutton R, Pennacchio LA, Gerstein M (2020). Supervised enhancer prediction with epigenetic pattern recognition and targeted validation..
LINK
https://doi.org/10.1038/s41592-020-0907-8

5 Quantify ChIP-seq signals for each constituent enhancers (CE) inside all SEs

First, quantify ChIP-seq signals for all candidate enhancers using featureCount, including those inside SEs and normal enhancers. This will generate a signal matrix with rows as enhancers and columns as samples. Signals of the two replicates for each cancer cell line should be summed together. In other words, each cell line should be put as one column in the signal matrix. Second, signal matrix is normalized using the normalization method documented in DESeq2. Third, a subset of rows, representing only enhancers inside SE candidates are subtracted further for downstream cancer-specific analysis.
The customized code for the steps above is documented here.

https://github.com/tenglab/cSEAdb_plos_code/blob/main/analysis/02_se_enhancer_signal.R

CITATION

Liao Y, Smyth GK, Shi W (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features..
LINK
https://doi.org/10.1093/bioinformatics/btt656

CITATION

Love MI, Huber W, Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2..
LINK
https://doi.org/

6 **Imputation of Zero signals in the CE signal matrix**

To ensure downstream mixture model fitting, zero signals in the signal matrix are imputed using the half of non-zero minimum signals in the signal matrix for each cell line . The following code can be used for the imputation. Here, se_ce_tmp is the signal matrix.

Command

Imputation of zero signals

```
se_ce <- se_ce_tmp %>% mutate_at(c(2:61),~replace(., . == 0, min(.[.>0], na.rm = TRUE)/2)
)
```

7 **Filter CEs with extreme low ChIP-seq coverage within their SEs**

To ensure robust comparison of CEs across cancer cell lines, CEs with no more than 3% of coverage within their belonged SEs in any of the studied cell lines are excluded from downstream analysis. The filtering process is implemented with the following script. Basically, some rows (low signal values) of the signal matrix are removed.

https://github.com/tenglab/cSEAdb_plos_code/blob/main/analysis/03_ce_filter.R

8 **Mixture model to identify active cell line from inactive cell line for each CE**

Apply customized mixture model scrips (below) to model each CE. For each row of the signal matrix, the mixture model takes the 60 signals values from 60 cell lines and defines a probability for each value to be considered as active. In other words, the 60 instances of the same CE across 60 cell lines will be assigned as either active or inactive depending on whether the corresponding probability is larger or smaller than 0.5

(https://github.com/tenglab/cSEAdb_plos_code/blob/main/analysis/em_mixture.R) t
https://github.com/tenglab/cSEAdb_plos_code/blob/main/analysis/03_mixture_model.R

9 **Identify cancer-cell-specific CEs/SEs**

Based on the frequency of being active across 60 cell lines for a given CE, cell-specific CEs are identified as less frequent but informative in encoding cancer cell identity. SEs holding these cell-specific CEs are concluded as cell-specific SEs.

In addition, the frequency of being inactive is also considered to identify cell-specific inactive CEs/SEs.

The below script is implemented to evaluate the frequency of active/inactive, the capability of encoding cell identity and the selection of cell-specific CEs and SEs.

https://github.com/tenglab/cSEAdb_plos_code/blob/main/analysis/04_cell_specific.R

10 **Identify cancer-specific CEs/SEs**

Cancer-specific CEs are identified based on the frequency of CEs being cell-specific for a given cancer type. In brief, if a cancer type has no less than 2 cell lines studied by the NCI-60 panel, a CE will be identified as cancer-specific only if it is cell-specific in at least 2 of the cell lines for this cancer type. For cancer types holding only on cell line, cell-specific CEs will be assigned as cancer-specific CEs.

SEs holding these cancer-specific CEs are concluded as cancer-specific SEs. In addition, cancer-specific inactive CEs/SEs are also identified.

The below script is implemented to identify cancer-specific SEs/CEs.

https://github.com/tenglab/cSEAdb_plos_code/blob/main/analysis/05_cancer_specific.R

11 **Create a data repository for cancer/cell-specific CEs/SEs - *cSEAdb***

To ease the use of the knowledge generated from the analysis, this protocol also provides scripts to generate R data object to store the cancer/cell-specific CEs and SEs.

The below script is implemented to generate such data object.
https://github.com/tenglab/cSEAdb_plos_code/blob/main/analysis/06_create_final_db_object.R

In addition, there is an R package (*cSEAdb*) available to explore, query and visualize the data object. For example, for a SE region, the cancer-specific CEs can be highlighted for a given cancer cell line or cancer type.

https://github.com/tenglab/cSEAdb