HiCDB: a sensitive and robust method for detecting contact domain boundaries

Fengling Chen
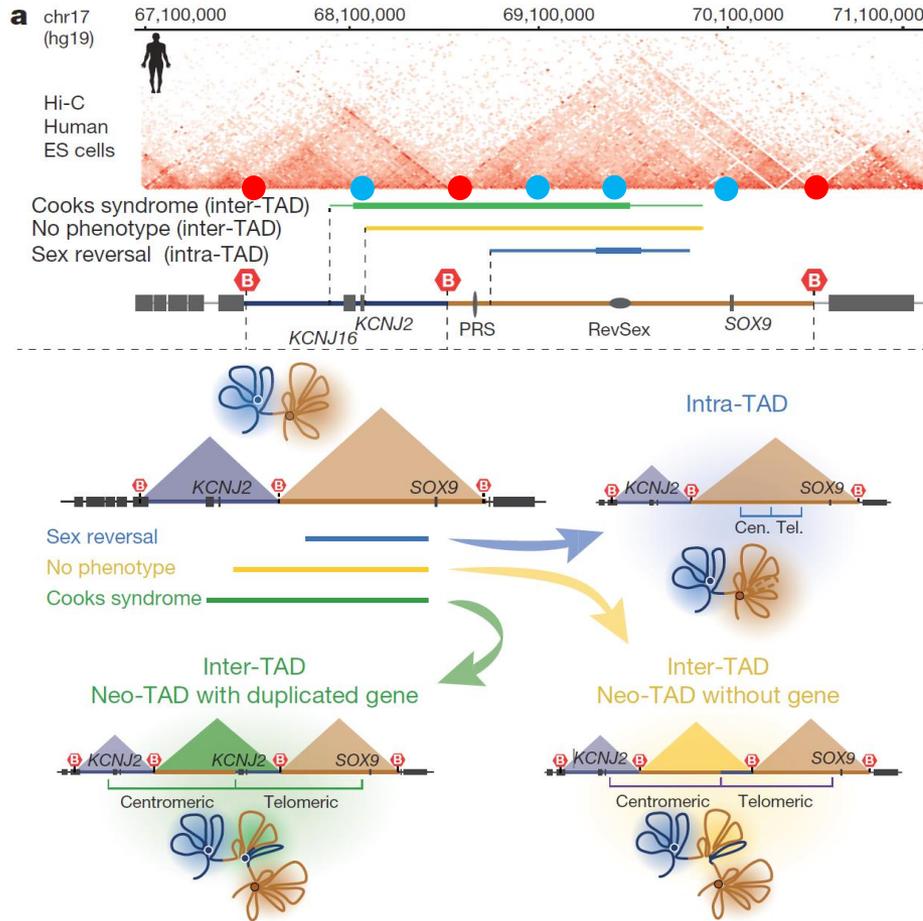Supervisor: Michael Q. Zhang & Yang Chen

MOE Key Laboratory of Bioinformatics
Bioinformatics Division and Center for Synthetic and Systems Biology, BNRist
Department of Automation, Tsinghua University

14 October 2018

Tsinghua University

contact domain boundary (CDB)

Why new CDB detecting methods ?

- existed methods were not fully evaluated.
- not updated for high resolution
- not robust enough to detect differential CDBs
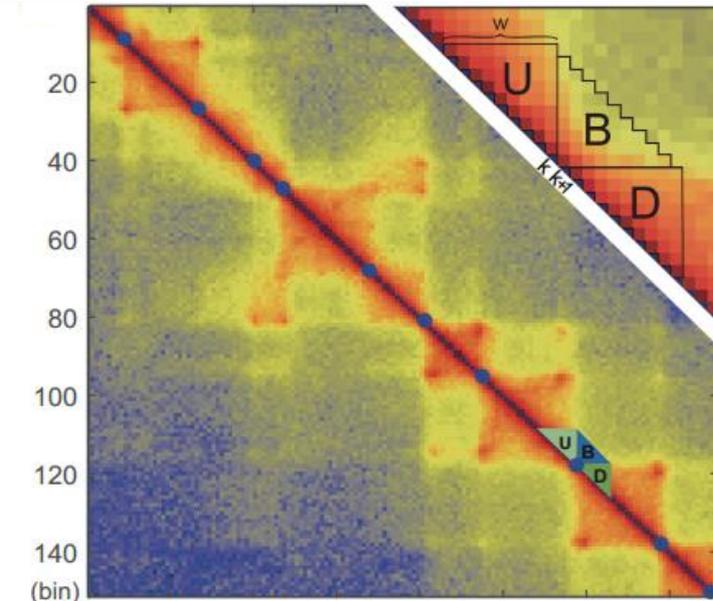- not specially designed to better detecting CDBs under highly-connected regions

genome-wide analysis of

epigenetic features of CDBs

function of cell-type specific CDBs

contact domain boundaries (CDBs) includes TAD boundaries and sub-TAD boundaries.

TADs： invariant/conserved; absolute insulated

sub-TADs : varied ; cell-type specific gene regulation; relatively insulated

Formation of new chromatin domains determines pathogenicity of genomic duplications. Franke et al. Nature. 2016

2. HiCDB Method

calculate RI under multiple window sizes

find local maximum peaks of average RI
generate local background by twice envelope

calculate LRI for each bin
decide cutoff of LRI

▼peak —1st envelope — 2nd envelope *CDB

- ✓ Calculate relative insulation (RI) under different window size

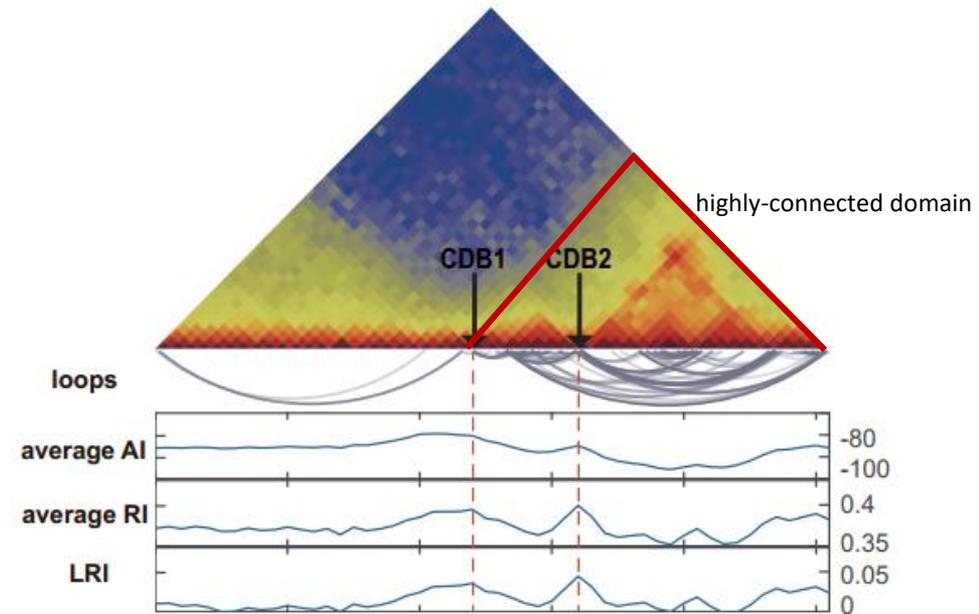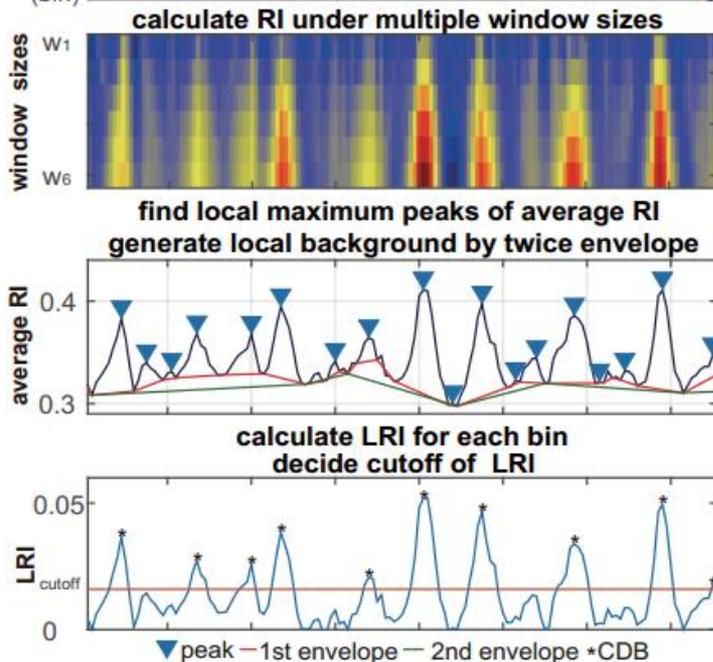$$RI(w,s) = \frac{U(w,s) + D(w,s) - B(w,s)}{U(w,s) + D(w,s) + B(w,s)}$$

- ✓ Peak detection on average RI

$$\overline{RI}(s) = \frac{1}{w_n - w_1} \sum_{w=w_1}^{w_n} RI(w,s)$$

- ✓ Remove background

$$LRI(s) = \overline{RI}(s) - lower\_envelope(lower\_envelope(\overline{RI}(s)))$$
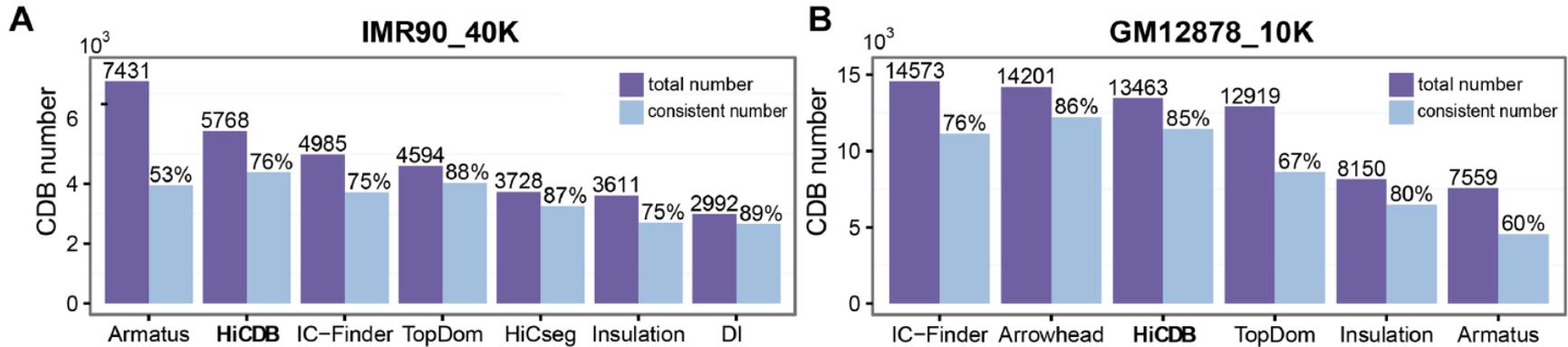
- ✓ Cut-off chosen (GSEA-like method)

highly-connected domain

loops

average AI

average RI

LRI

aRI or LRI can be compared genome-wide
aRI or LRI help to find CDBs under highly-connected domains.

清華大學
Tsinghua University
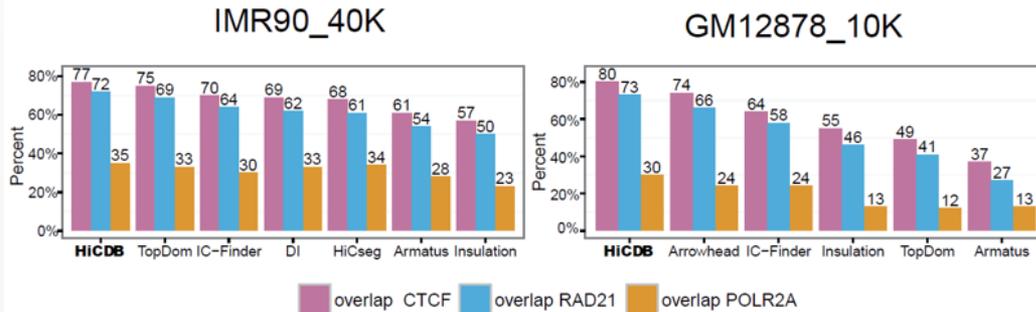
All comparisons were carried on at 40kb-resolution and 10kb-resolution datasets
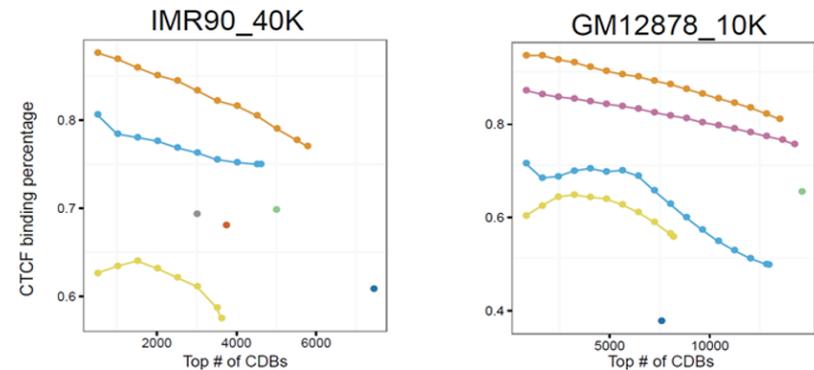
Consistency between all methods

CTCF/cohesin and active regulatory signal enrich on HiCDB detected CDBs
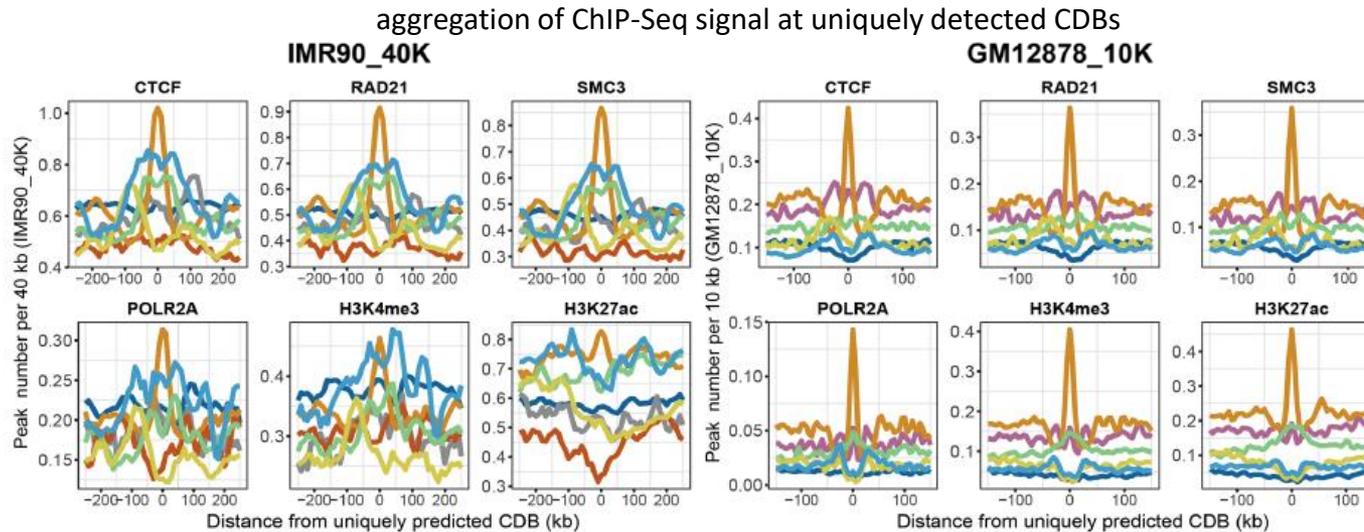

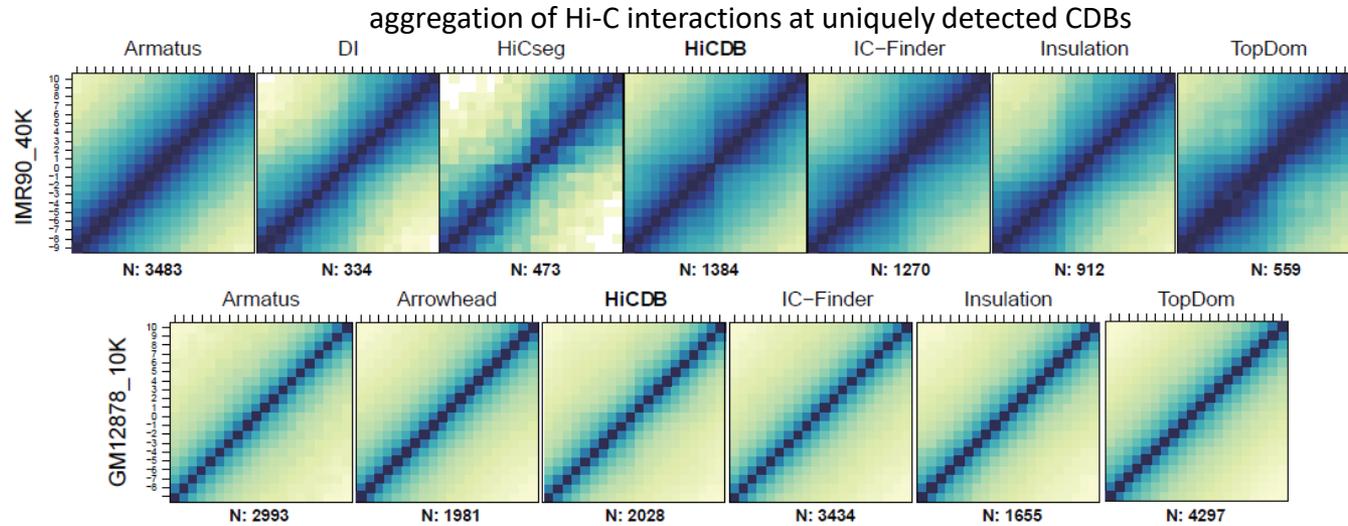Percentages of CDBs overlapped with protein ChIP-Seq

CTCF binding percentages of the HiCDB-detected CDBs were always the highest for different cutoffs

3. Compare with other methods

Evaluation of uniquely detected CDBs of each methods

aggregation of Hi-C interactions at uniquely detected CDBs

aggregation of ChIP-Seq signal at uniquely detected CDBs

HiCDB uniquely detected CDBs are meaningfull.

Tsinghua University

Method reproducibility at different cutoff

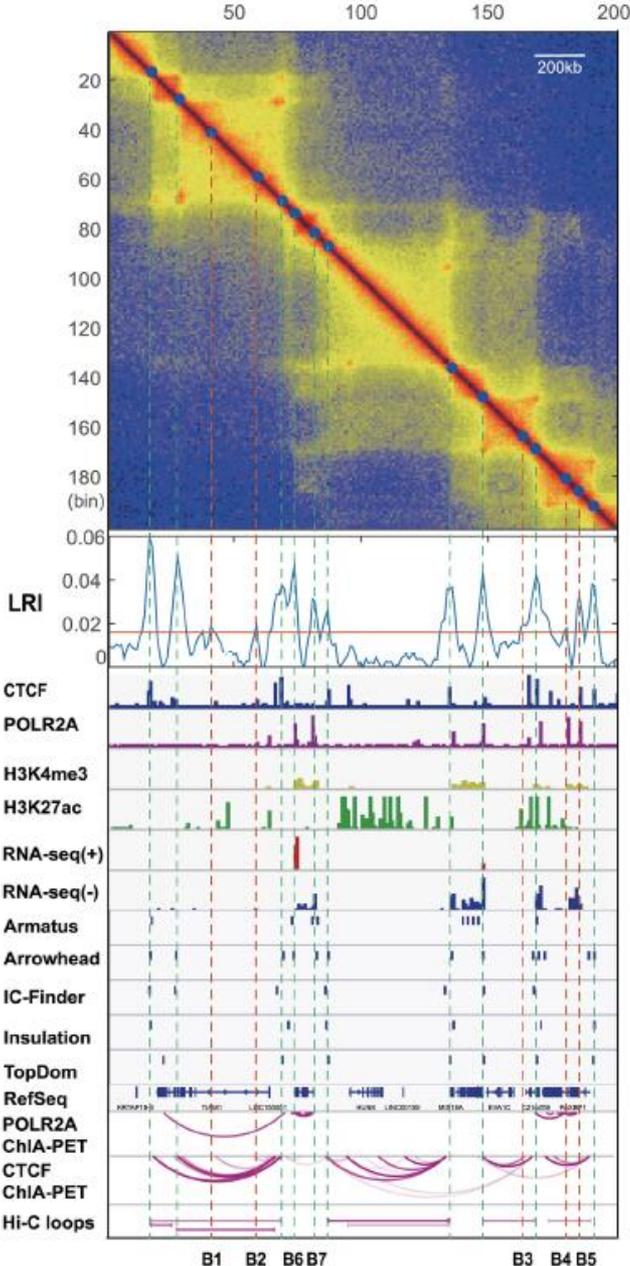hESC_40K   GM12878_10K
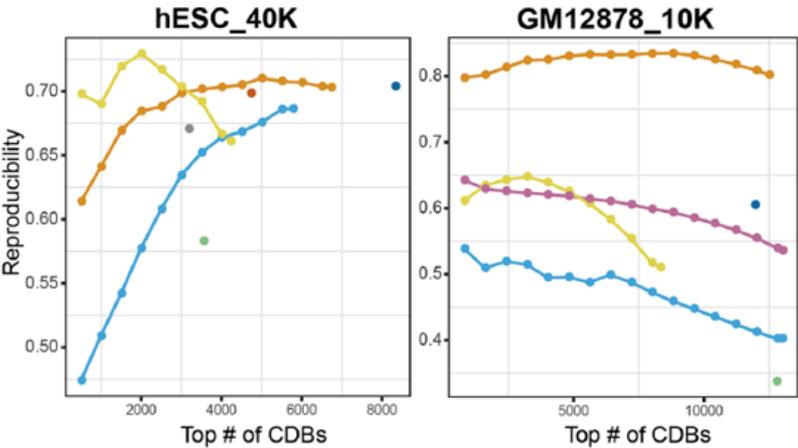
Sensitivity / Specificity
(use CDBs detected on high resolution Hi-C data as TRUE)
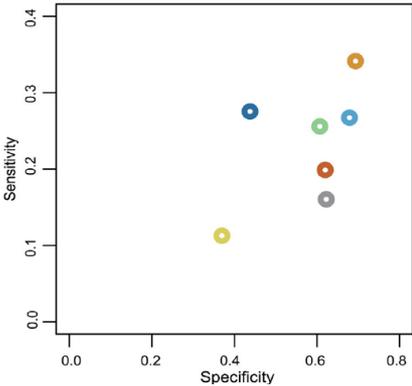
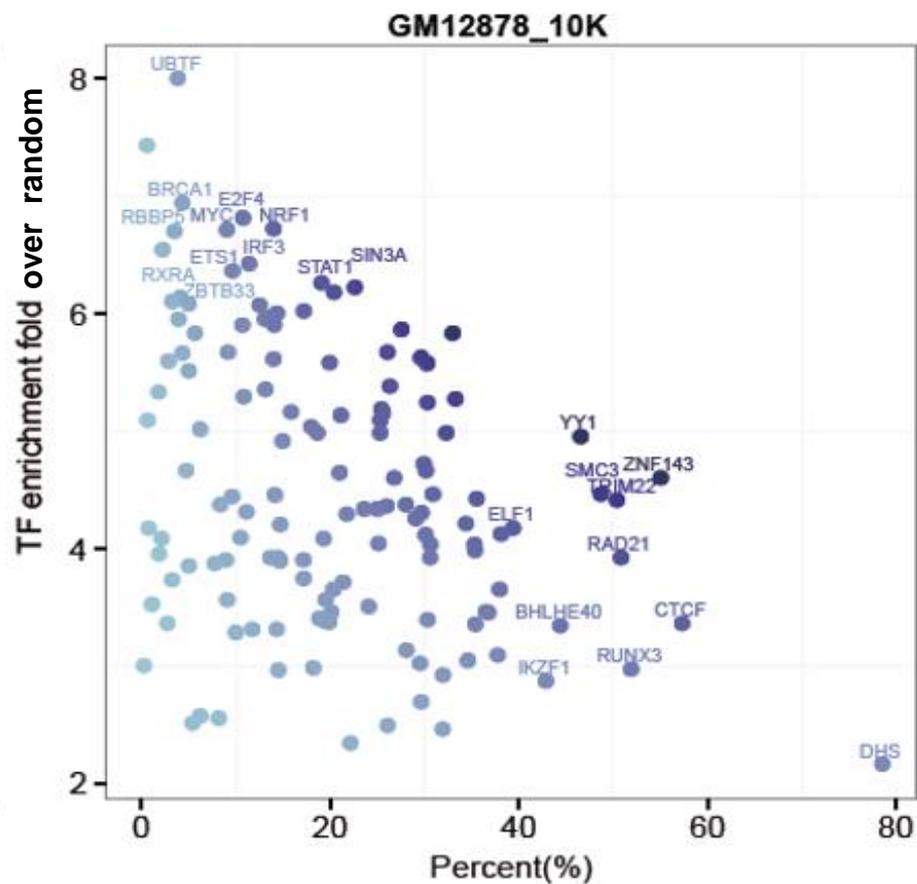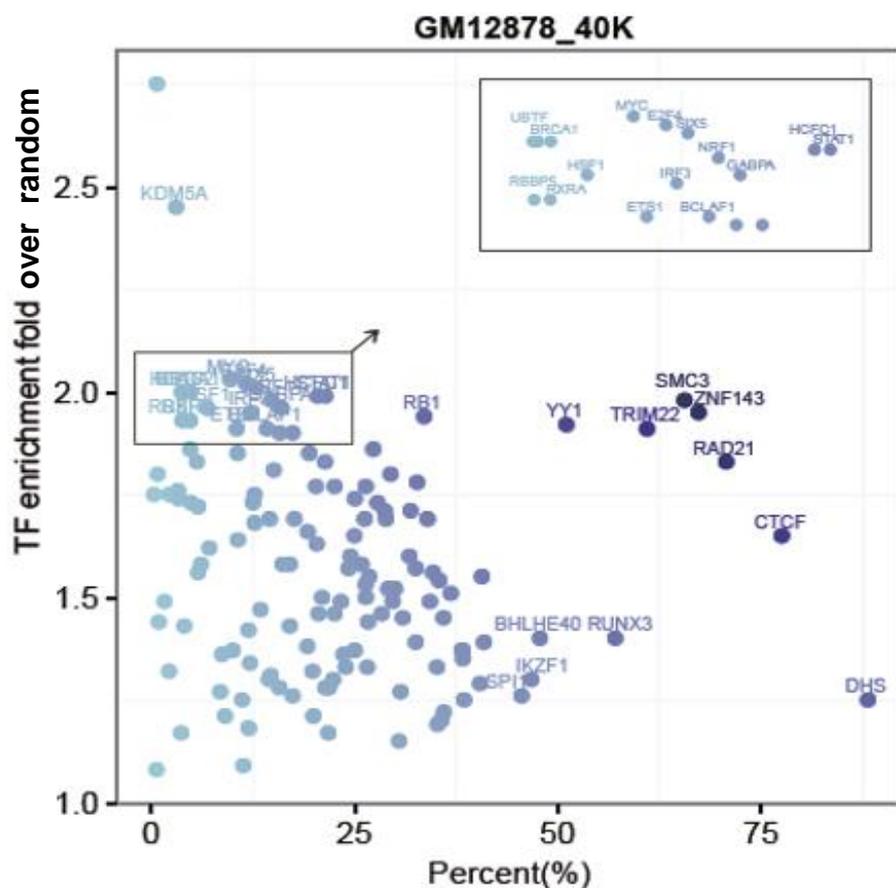● Armatus   ● Arrowhead   ● DI   ● HiCseg   ● **HiCDB**   ● IC-Finder   ● Insulation   ● TopDom

comprehensive metrics:

the CDB number, consistency, protein binding enrichment, robustness and time complexity.
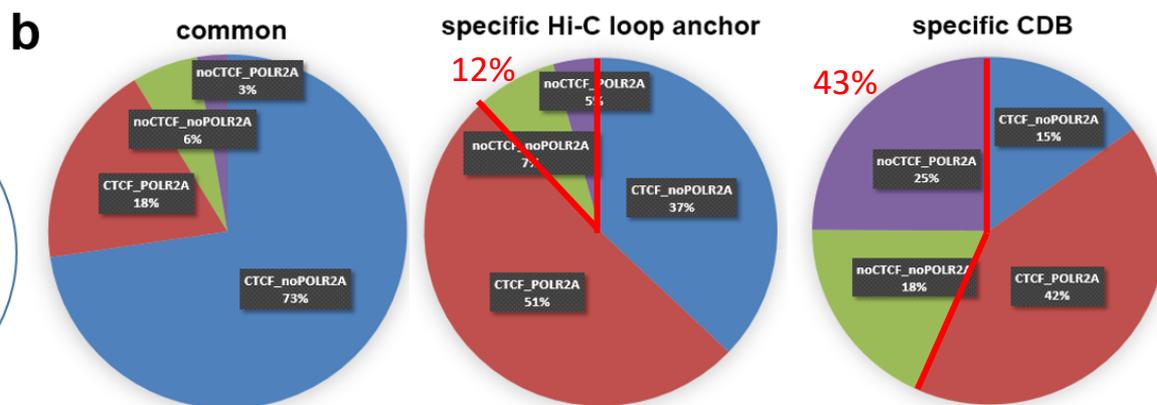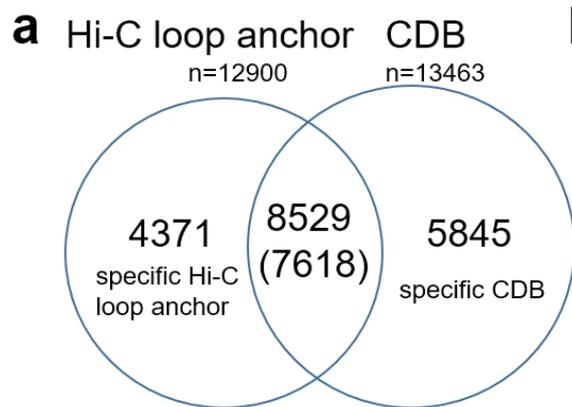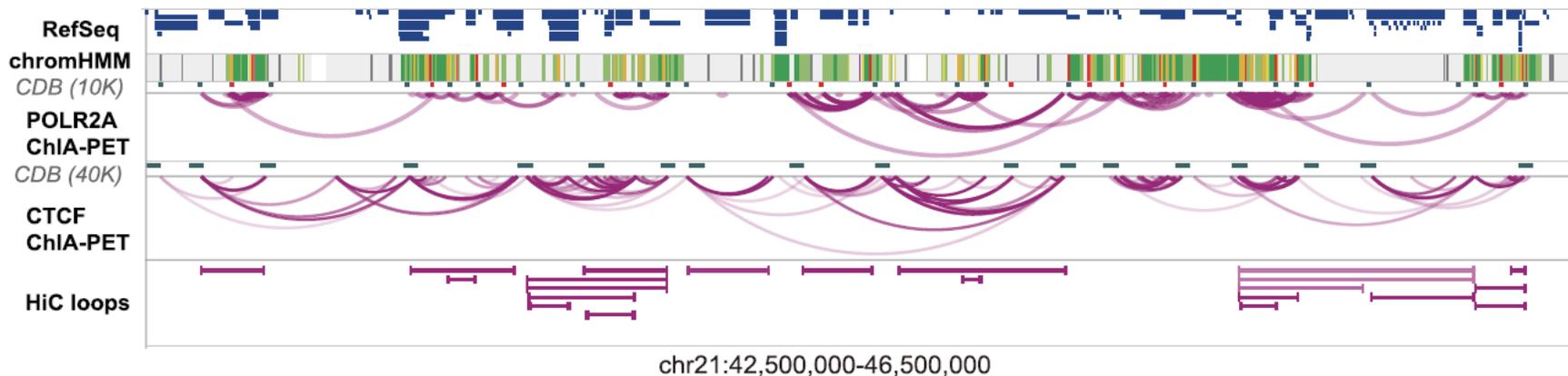
清華大學
Tsinghua University

CDB enriches structural regulator as well as cell-type-specific TFs

✓ TRIM22 is probably a structural regulator(GM12878 & MCF-7)

4. Epigenetic features of CDBs

CDBs cover more CTCF independent regulatory regions than Hi-C loops detected by HiCCUPS
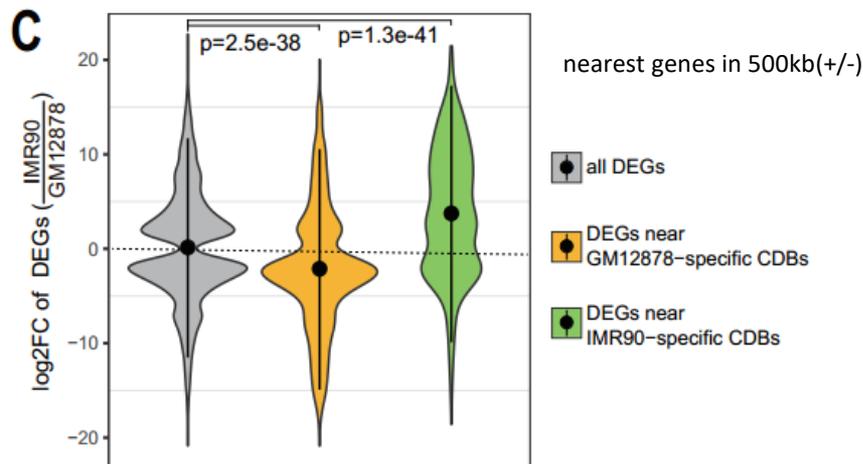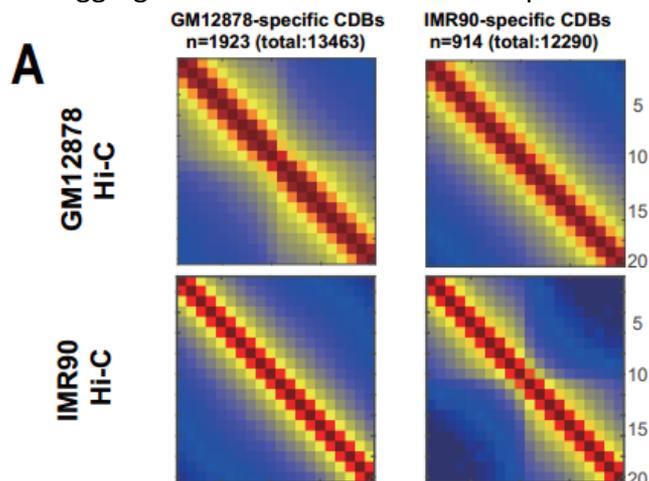

chr21:42,500,000-46,500,000

a Hi-C loop anchor / CDB

Hi-C loop anchor n=12900   CDB n=13463

4371 specific Hi-C loop anchor | 8529 (7618) | 5845 specific CDB

b

common

- noCTCF_POLR2A 3%
- noCTCF_noPOLR2A 6%
- CTCF_POLR2A 18%
- CTCF_noPOLR2A 73%

specific Hi-C loop anchor

- 12%
- noCTCF_POLR2A 5%
- noCTCF_noPOLR2A 7%
- CTCF_noPOLR2A 37%
- CTCF_POLR2A 51%

specific CDB

- 43%
- CTCF_noPOLR2A 15%
- noCTCF_POLR2A 25%
- noCTCF_noPOLR2A 18%
- CTCF_POLR2A 42%

The enrichment of CTCF independent loops and cell-type-specific TFs on CDBs strongly suggest that **the CDBs are not only structurally but also functionally related to transcription regulation.**

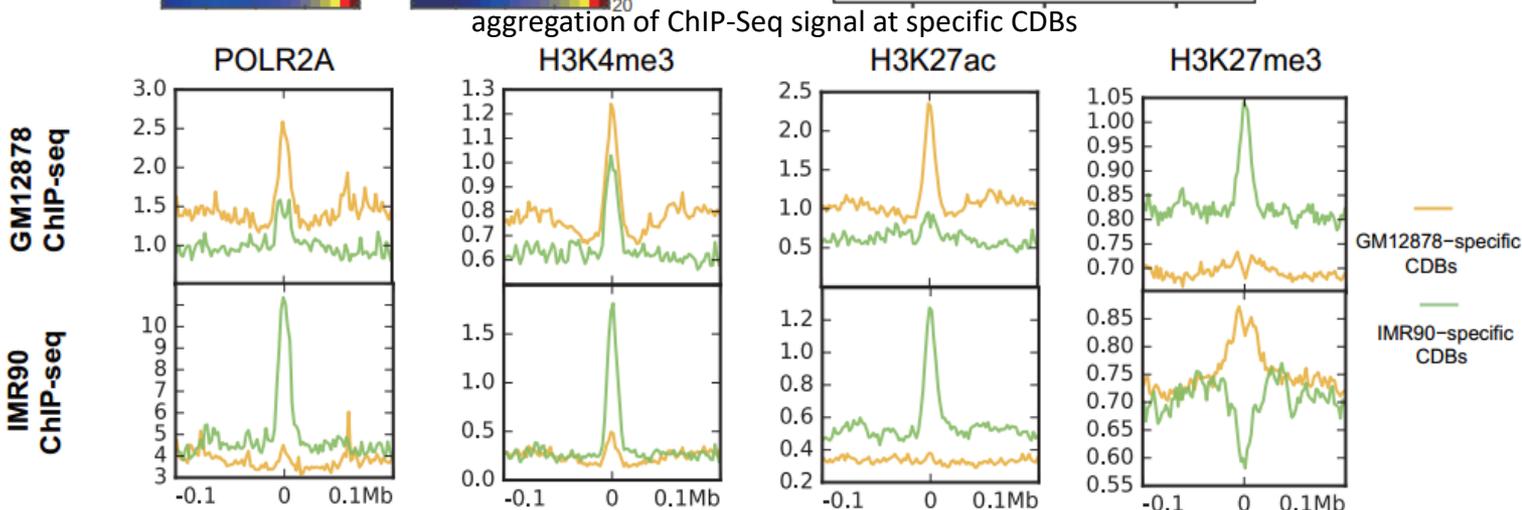# 5. Differential CDB detection

Emergence of CDBs are frequently associated with the activation of nearby cell-type-specific genes.

- ✓ cell-type-specific CDBs correlate with cell-type-specific histone modifications
- ✓ GM12878-specific CDBs：B cell activation, interferon-gamma-mediated signaling pathway etc.
- ✓ IMR90-specific CDBs：lung development etc.
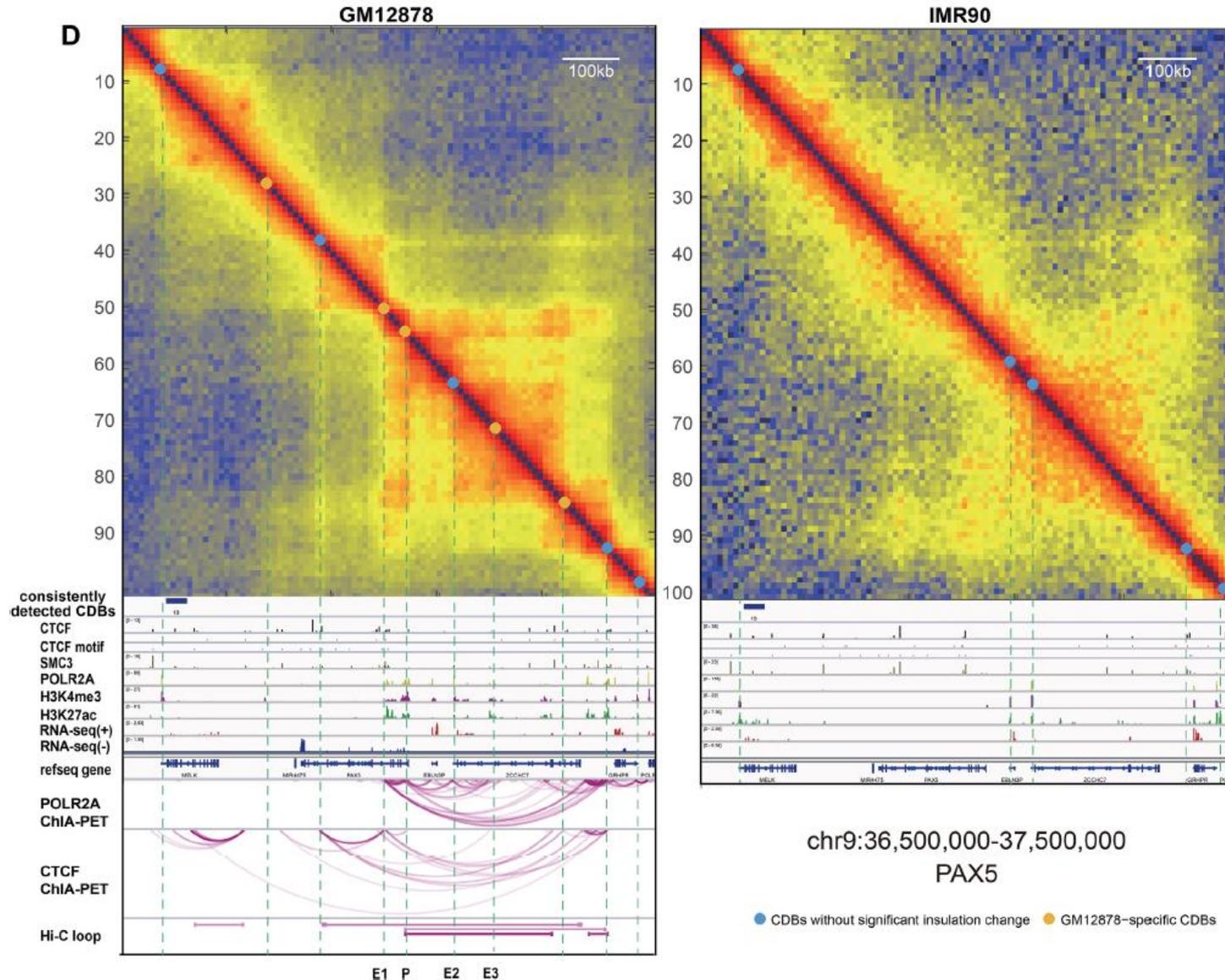

aggregation of Hi-C interactions at specific CDBs
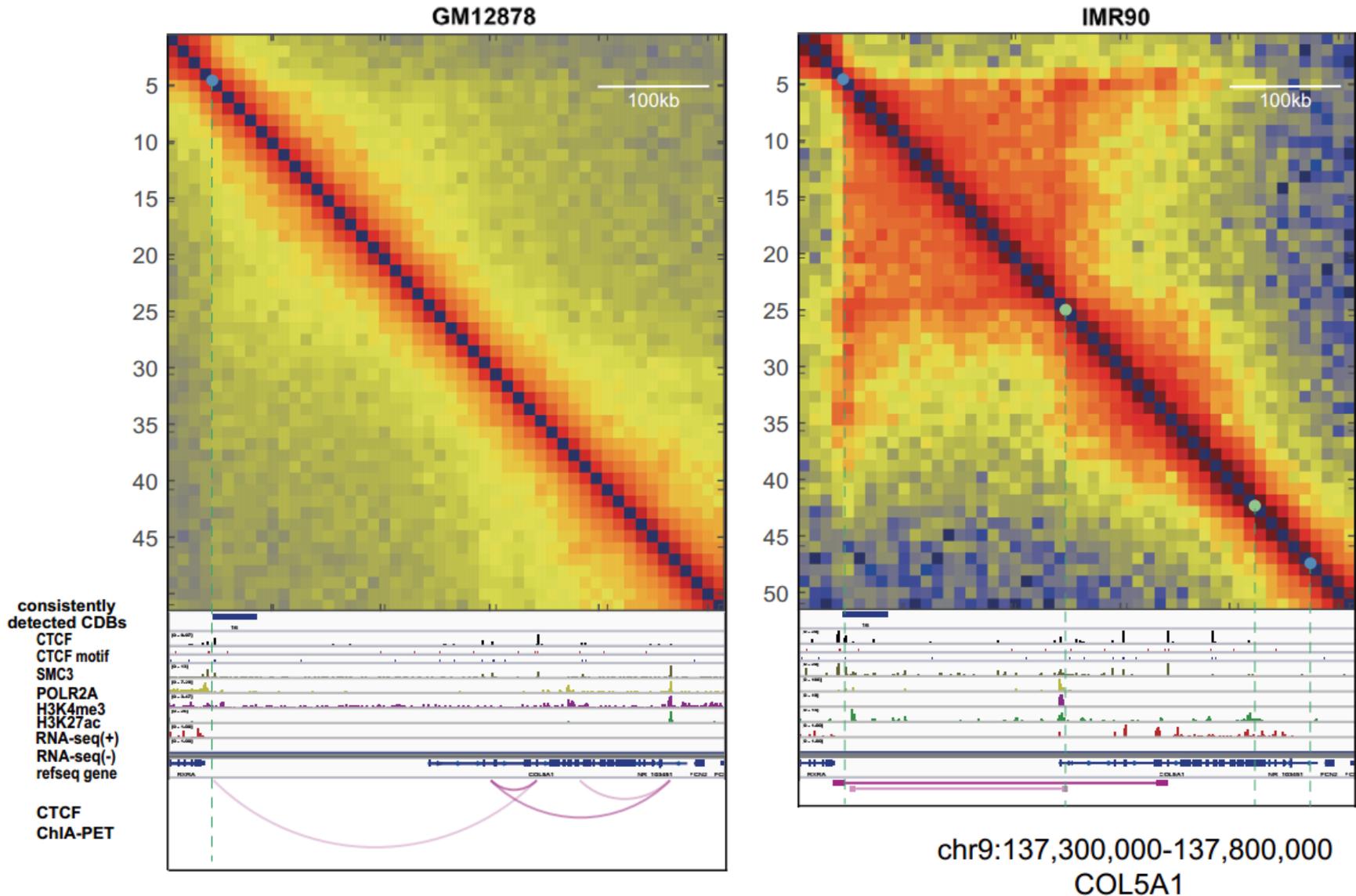
aggregation of ChIP-Seq signal at specific CDBs

Differential CDB example

Cell-type-specific CDBs correlate with cell-type-specific histone modification and gene activation.
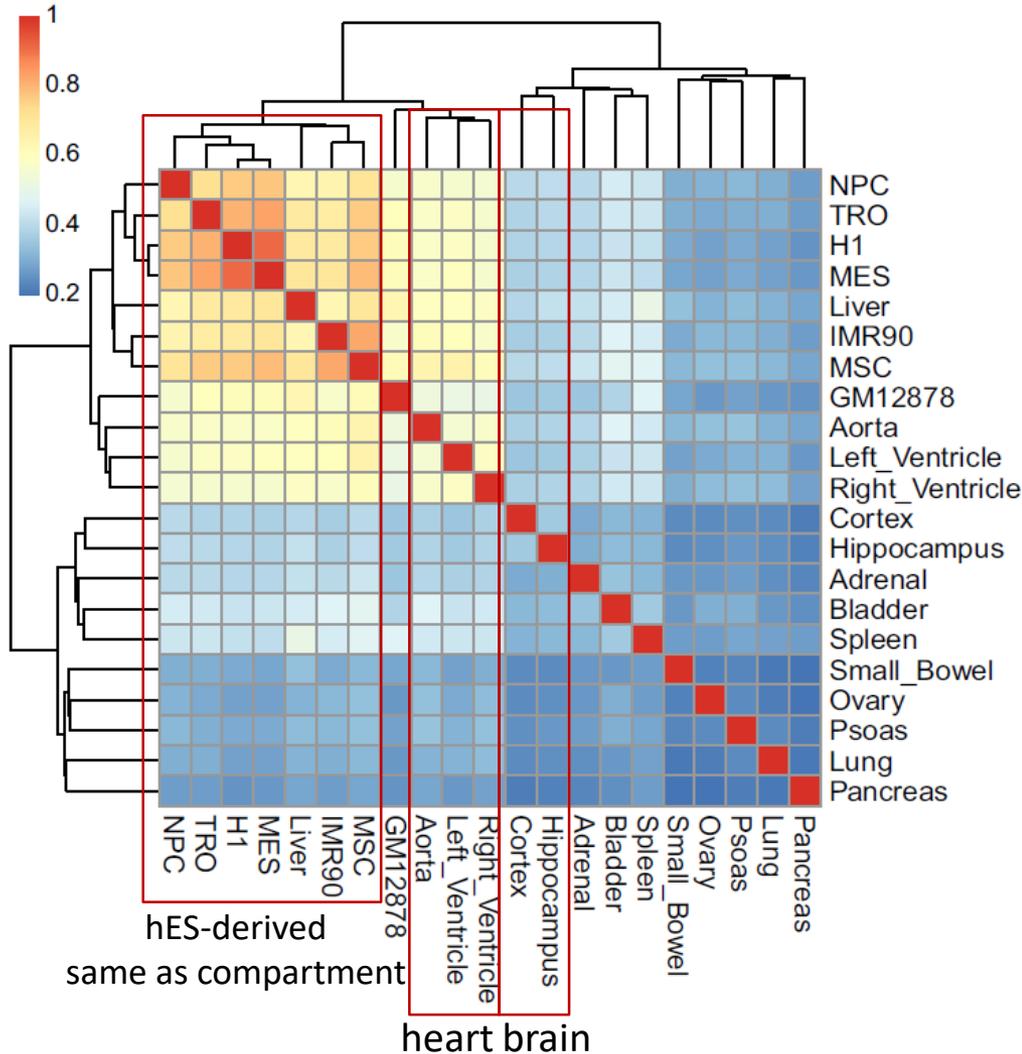
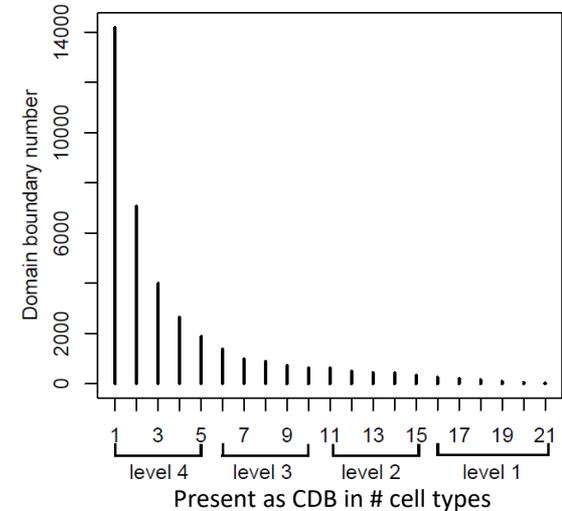**Differential CDB example**

Cell-type-specific CDBs correlate with cell-type-specific histone modification and gene activation.

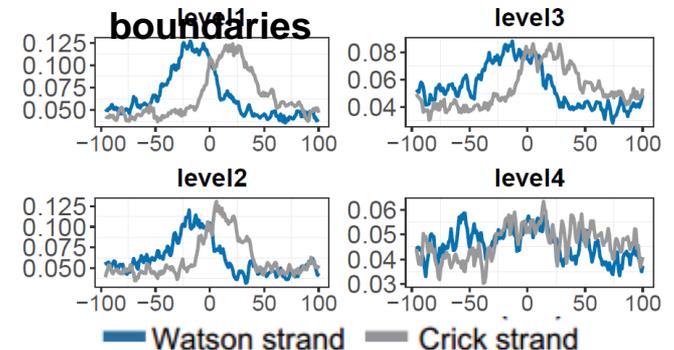**Spearman correlation of CDB aRI values**

hES-derived
same as compartment

heart brain

CDB consistency across 21 cell types Total CDBs: 37518; avg 6615
invariant CDBs (level 1): 2542
(available on github)

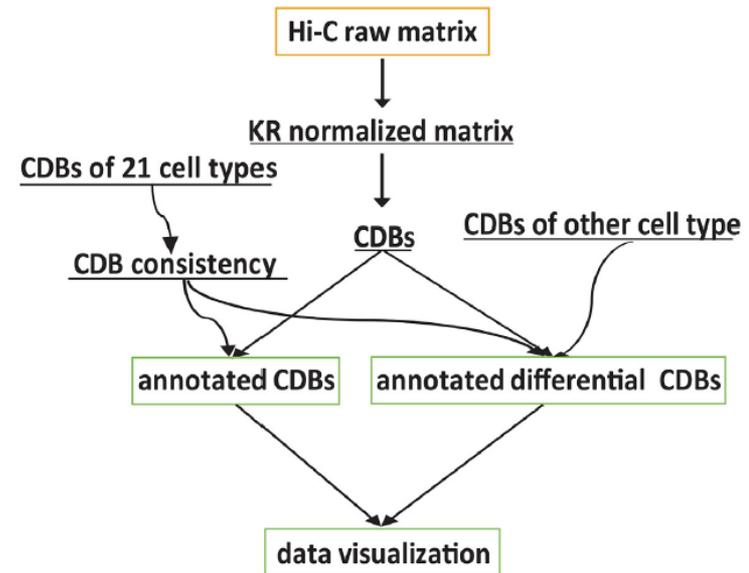Present as CDB in # cell types

CTCF motif around boundaries

Watson strand — Crick strand

CTCF motif is more divergent at invariant CDBs

A compendium of chromatin contact maps reveals spatially active regions in the human genome[J]. Schmitt A D, et al. Cell reports, 2016.

Take home messages

1. HiCDB is fast and sensitive which can be applied to multiple resolution.

2. Applying relative insulation, HiCDB is powerful in defining CDBs within highly-connected regions. Relative insulation is associated with gene regulation and histone modification.

3. Use HiCDB to find functional structure differences between conditions.

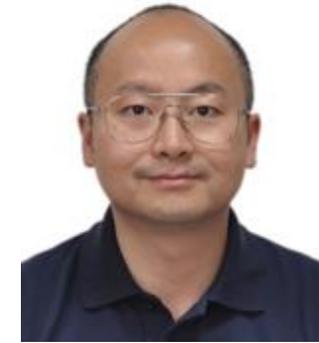4. Refer to our comprehensive comparisons to select boundary detecting tools.

https://github.com/ChenFengling/RHiCDB   (R)
https://github.com/ChenFengling/HiCDB   (matlab)

Overview of HiCDB package

Acknowledgement

2017.12

Thank you!

Fengling Chen 陈凤玲: Ph.D. candidate
Email: cfl15@mails.tsinghua.edu.cn
fchen@github

cut-off chosen

LRI

Peaks hit
CTCF motif
or not

Erichment
Score

order of local maximum peaks
10⁴
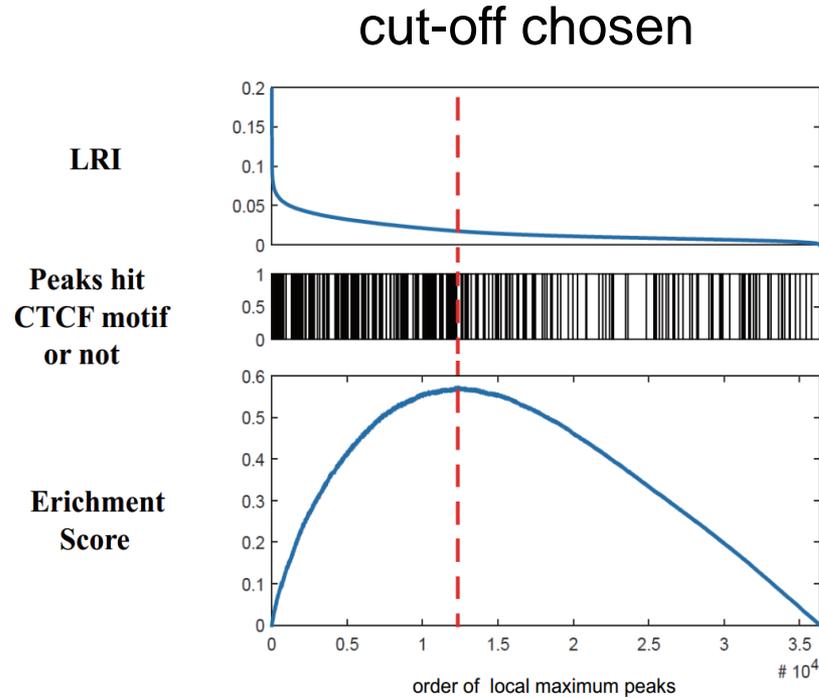
$$p(i) = \begin{cases} \dfrac{|LRI_i|}{N_{LRI}}, where\ N_{LRI} = \displaystyle\sum_{L_m \in S} LRI_m & L_i \in S \\[2em] -\dfrac{1}{N - N_{hit}} & L_i \notin S \end{cases}$$

$$ES(i) = \sum_{t=1}^{i} p(t)$$