

Identifying Robust Single-Cell 3D Genome Domain Boundaries with CellRoB

Yusen Ye

ysye@xidian.edu.cn

Xidian University <https://orcid.org/0000-0001-8687-1058>

Shiji Liu

Xidian University

Lin Gao

Xidian University <https://orcid.org/0000-0001-6396-0787>

Article

Keywords: 3D genome, single-cell, single-cell chromatin domain boundaries

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11342175/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

1 **Identifying Robust Single-Cell 3D Genome Domain**

2 **Boundaries with CellRoB**

3 Shiji Liu¹, Yusen Ye^{1*}, Lin Gao^{1*}

4

5 ¹ School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, China.

6

7 *To whom correspondence should be addressed. Email: ysye@xidian.edu.cn; lgao@mail.xidian.edu.cn.

8

9 **Abstract**

10 Extreme sparsity in single-cell 3D genome data obscures topologically associating domain
11 (TAD) boundaries and limits the characterization of structural heterogeneity. Here, we
12 introduce CellRoB, a computational framework that integrates single-cell boundary
13 detection with cross-cell alignment to improve boundary identification while preserving cell-
14 specific variation. Benchmarks on mouse brain Dip-C and human GM12878 scMicro-C
15 data demonstrate improved accuracy and robustness to contact dropout across resolutions
16 from 40 kb to 5 kb. CellRoB further defines and characterizes rare boundaries shared by
17 only a small fraction of cells, revealing low-frequency structural variation obscured by
18 population averaging. These boundaries carry cell-type information and exhibit distinct
19 functional associations. Applied to multimodal mouse embryo data, boundary-strength
20 profiles resolve two epiblast and primitive streak (EPI) subpopulations not separated in the
21 RNA-derived embedding, with transcriptional differences suggesting distinct
22 developmental tendencies. Together, CellRoB enables reliable single-cell TAD boundary
23 identification, reveals biologically informative rare boundaries, and establishes boundary
24 strength as a structural feature for resolving cellular heterogeneity.

25

26 **Key words:** 3D genome, single-cell, single-cell chromatin domain boundaries

27

28 Introduction

29 The three-dimensional organization of chromatin contributes to transcriptional regulation,
30 DNA replication and chromosome segregation^{1–8}. By bringing distant genomic elements
31 into spatial proximity, chromatin folding facilitates interactions between regulatory elements
32 and their target genes^{9–10}. Chromosome conformation capture technologies, particularly
33 Hi-C, have revealed multiple levels of genome organization, including A/B
34 compartments^{4,11}, topologically associating domains (TADs)^{12–14} and chromatin loops^{4,15–}
35 ¹⁶. Compartments reflect large-scale segregation of active and inactive chromatin, whereas
36 loops represent specific point-to-point interactions that often mediate specific regulatory
37 contacts^{4,11,15–20}. At an intermediate scale, chromatin domains such as TADs and their
38 hierarchical variants provide a mesoscale organizational framework that partitions the
39 genome into regulatory units^{6,12–14,21–24}. Among these architectural features, domain
40 boundaries are of particular importance, as they insulate neighboring domains and
41 constrain regulatory interactions within defined genomic intervals^{21,25}. Disruption of
42 domain boundaries can lead to aberrant enhancer-promoter communication and is strongly
43 associated with altered gene expression and disease phenotypes^{6,26–27}.

44 Single-cell 3D genome technologies enable the investigation of chromatin
45 organization beyond population averages, but extreme sparsity and technical noise make
46 domain boundaries difficult to identify reliably^{28–31}. Weak boundary signals may remain
47 undetected, while uncertainty in boundary positions complicates comparisons across cells.
48 Consequently, observed boundary differences can reflect both biological heterogeneity and
49 technical variability. Computational approaches, including scHiCluster³², Higashi³⁵,
50 DeTOKI³³, as well as clustering and optimization-based methods such as DeDoc2³⁶,
51 HiCS³⁷ have advanced single-cell chromatin analysis through contact imputation,
52 representation learning and domain detection. However, recovering domain structures
53 does not by itself establish a consistent boundary representation across cells. Even when
54 information is shared during imputation or representation learning, positional discrepancies
55 among downstream boundary calls can hinder systematic comparison of boundary
56 occurrence and strength.

57 This challenge is particularly important for boundaries shared by only a small fraction
58 of cells. Aggregating single-cell contact maps strengthens recurrent structural signals but
59 can obscure low-frequency boundaries, limiting access to potentially informative cellular
60 variation. Conversely, analyzing boundary calls independently makes it difficult to
61 determine whether nearby calls represent a shared boundary or distinct structural features.
62 A framework that improves cross-cell comparability while retaining cell-specific information
63 is therefore needed to distinguish widely conserved boundaries from rare boundaries and
64 examine their biological associations. Such a representation would also enable
65 assessment of whether quantitative boundary variation can reveal cellular states beyond
66 those resolved by transcriptomic analysis.

67 Here, we present CellRoB, a computational framework for robust identification and
68 comparison of single-cell chromatin domain boundaries. CellRoB integrates contact-map
69 imputation and single-cell boundary detection with a population-aware alignment
70 procedure. It models recurrent boundary regions and aligns boundary positions by

minimizing an insulation-score loss, improving cross-cell comparability while retaining cell-specific structural variation. Benchmarks on mouse brain Dip-C and human GM12878 scMicro-C data^{38–39} demonstrate improved boundary detection accuracy and robustness to contact dropout and bin-size variation, with evaluations spanning 40 kb to 5 kb.

Beyond boundary detection, CellRoB enables the definition and systematic characterization of rare boundaries alongside weakly and strongly conserved boundaries. These classes exhibit distinct epigenetic and functional associations, with rare boundaries retaining information about cell identity. Applied to paired chromatin conformation and transcriptomic profiles from mouse embryos⁴⁰, CellRoB-derived boundary-strength profiles resolve two subpopulations within the epiblast and primitive streak (EPI) group that are not separated in the RNA-derived embedding. Their transcriptional differences suggest distinct developmental tendencies, highlighting the value of quantitative boundary variation for characterizing cellular states. Together, CellRoB links reliable single-cell boundary identification to the discovery of low-frequency structural variation and finer characterization of cellular heterogeneity.

Results

1. Performance evaluation of CellRoB on Dip-C datasets

To benchmark CellRoB against existing approaches, we applied it to single-cell Dip-C data from Tan et al.³⁹ for chromosome 6 at 40 kb resolution. Because no universally accepted gold standard exists for domain boundary annotation in 3D genomics, we adopted multiple complementary evaluation criteria to systematically assess algorithmic performance. First, we examined the enrichment of known boundary-associated factors. CTCF, a canonical insulator protein, was highly enriched around CellRoB-predicted boundaries (Figure 2a). Moreover, three major boundary-associated factors exhibited stronger enrichment at CellRoB boundaries than at those predicted by alternative methods, supporting the biological validity of our predictions (Figure 2a). Second, we assessed the consistency of boundary detection between structurally similar cells. For each cell, we identified its most similar counterpart based on cell embedding and formed cell pairs accordingly. We then compared the similarity of boundaries and domain sets within these pairs using the Jaccard Index (JI) and Measure of Concordance (MoC). CellRoB boundaries showed the highest JI and MoC values, indicating that they better preserve cross-cell consistency of domain boundaries and thus enhance comparability across cells (Figure 2b). Furthermore, when using boundaries as cell features for embedding, UMAP visualization demonstrated that CellRoB preserved cellular heterogeneity without over-smoothing (Figure S2c). Third, we tested reproducibility by generating two replicate Dip-C datasets in which 5% of contacts were randomly removed from each cell. For each cell, we compared boundaries detected in the replicate datasets. CellRoB and DeDoc2 achieved the highest reproducibility (JI and MoC), followed by HiCS, Higashi, and DeTOKI (Figure 2c). Finally, we evaluated robustness to noise by randomly dropping out 5%, 10%, 20%, and 40% of contacts from the chromatin contact matrices. For each cell, we constructed a reference boundary set by taking the predictions from all methods on the full data (see Methods), and quantified predictor accuracy using F1 score. CellRoB achieved the highest F1 scores across all

noise levels, followed by DeDoc2 (Figure 2d, Figure S2a,b).

Taken together, these results demonstrate that CellRoB can stably and accurately identify biologically meaningful domain boundaries from single-cell Hi-C data while maintaining cross-cell comparability.

2. Performance evaluation of CellRoB on high-resolution scMicro-C datasets

With the continuous advancement of single-cell Hi-C technologies, the contact number per cell has reached the million level, enabling high-resolution analyses of single-cell chromatin architecture. However, the robustness of existing domain boundary detection methods under such high-resolution settings has not been systematically evaluated. To further assess the robustness of CellRoB, we benchmarked its performance against representative methods on high-resolution scMicro-C datasets.

A recently developed technique, scMicro-C, is capable of resolving 3D genome structures at kilobase resolution³⁸. We processed the scMicro-C datasets at multiple binsizes (40 kb, 30 kb, 20 kb, 10 kb, and 5 kb) to evaluate the robustness of various boundary detection algorithms under different resolutions for the downstream analysis.

We found that CellRoB demonstrated strong robustness to varying binsizes. Specifically, the number of domain boundaries predicted by HiCS, DeDoc2, and DeTOKI increased exponentially as the bin size decreased, whereas CellRoB and Higashi did not show this pattern; for Higashi, this is expected because a predefined TAD-length parameter constrains the number of boundaries detected per cell (Figure 3a). Among all tested methods, CellRoB and Higashi achieved the highest MoC, followed by HiCS, DeTOKI, and DeDoc2 (Figure 3b,c). Furthermore, CellRoB exhibited consistently stronger enrichment of boundary-associated biological features than all other predictors, even at the finest 5-kb resolution (Figure 3d, Figure S4a). Similar to the Dip-C results, in the scMicro-C dataset, boundaries with higher conservation exhibited clear enrichment of architectural proteins such as CTCF, whereas rare boundaries did not (Figure S4b).

We next performed Gene Ontology (GO) enrichment analysis on genes associated with weakly and strongly conserved boundaries in the scMicro-C dataset, in which all cells were synchronized in the G1 phase. These boundary classes were defined according to their occurrence frequencies across cells. The associated genes were significantly enriched in rRNA processing, ribosome biogenesis, and ncRNA processing. These functions are consistent with the biosynthetic activity of G1-phase cells as they prepare for DNA replication in the subsequent S phase (Figure 3e).

These cross-resolution and biological validation results collectively indicate that CellRoB maintains stable performance and biological interpretability as the resolution increases, demonstrating its robustness for high-resolution single-cell 3D genome data.

3. CellRoB reveals domain boundary heterogeneity and its association with cellular identity

Aggregating single-cell Hi-C data into pseudo-bulk contact maps strengthens shared structural signals but can obscure boundaries present in only a small fraction of cells. By improving boundary comparability across individual cells, CellRoB enables direct

155 characterization of boundary sharing and variation within a cell population.

156 To evaluate performance, we further compared the conserved domain boundaries
157 identified by CellRoB with those detected by TopDom⁴¹, Insulation Score (IS)¹², and
158 MSTD⁴² from pooled single-cell Hi-C data. Boundaries detected by at least three methods
159 were defined as benchmark references (see method). We subsampled the Dip-C dataset
160 at varying cell numbers (10, 25, 50, 100, 200, 500, and 1000 cells). For each sampling
161 depth, we computed the F1 score, precision, and recall for CellRoB, TopDom, IS, and
162 MSTD against the benchmark. Across all sampling levels, CellRoB consistently achieved
163 the highest F1 score (Figure 4a). When only 10 cells were used, CellRoB detected slightly
164 fewer boundaries than MSTD; however, its F1 score, precision, and recall were all
165 significantly higher (Figure 4a, Figure S3a,b,c). Although TopDom exhibited marginally
166 higher precision at the 25- and 50-cell subsamples, it identified substantially fewer total
167 boundaries (Figure S3a,b). Moreover, the boundaries predicted by CellRoB showed
168 stronger enrichment for multiple canonical boundary-associated features, including CTCF,
169 H3K4me3, and HK genes (Figure 4c).

170 Because CellRoB operates directly on single-cell Hi-C data, it can detect domain
171 boundaries present in only a small subset of cells, which are typically obscured in bulk Hi-
172 C analyses. Compared with TopDom, CellRoB detected more boundaries overall and
173 showed a markedly improved ability to capture rare boundaries within the cell population
174 (Figure 4b). We then grouped the aligned boundaries into rare, weakly conserved, and
175 strongly conserved classes according to their occurrence frequencies across cells. These
176 classes exhibited distinct epigenetic enrichment patterns, with rare boundaries showing
177 limited enrichment for CTCF, H3K4me3, and HK genes (Figure 4d).

178 We treated domain boundaries corresponding to each conservation class as features
179 on UMAP, to evaluate whether different levels of boundary conservation could capture cell-
180 type-specific features. The embedding revealed that all three conservation classes
181 enabled some degree of separation among cell types, indicating that even rare boundaries
182 carry biologically meaningful information. We next performed GO enrichment analysis for
183 genes associated with each boundary class. Genes near rare boundaries were enriched
184 in terms related to nervous system development and neuronal differentiation (Figure 4e).
185 Genes near weakly conserved boundaries were enriched in cellular response to stimulus,
186 inflammatory response, and regulation of cell population proliferation (Figure 4f). In
187 contrast, genes adjacent to strongly conserved boundaries were predominantly involved in
188 cell division, chromosome segregation, and nuclear division (Figure 4f). Consistent with
189 these patterns, the cell-type specificity of flanking genes increased as boundary
190 conservation decreased.

191 Together, these results demonstrate that CellRoB robustly identifies shared domain
192 boundaries and improves the recovery of rare boundaries compared with pooled-data
193 approaches. The distinct functional associations and cell-type information captured by
194 different boundary classes highlight the value of boundary occurrence frequency for
195 characterizing chromatin heterogeneity.

4. CellRoB-derived boundaries resolves previously unrecognized developmental subpopulations

With the growing availability of single-cell sequencing technology and the emergence of multimodal technologies that jointly profile chromatin conformation and gene expression, it is now possible to explore how 3D genome architecture contributes to cellular differentiation. We applied CellRoB to the HiRES dataset, which includes matched single-cell Hi-C and RNA-seq profiles from the mouse brain⁴⁰.

HiRES transcriptomic data revealed two major developmental lineages: a neuroectodermal trajectory and a mesenchymal trajectory, both originating from epiblast and primitive streak (EPI) cells. Notably, in the RNA-derived cell embedding, EPI was identified as a single homogeneous population. In contrast, when we constructed cell embeddings using domain boundary features detected by CellRoB, not only were the same two developmental trajectories recapitulated, but EPI cells were further subdivided into two distinct subpopulations (Figure 5a,b, Figure S5a).

To explore the relationship between these two EPI subtypes and the subsequent developmental fates, we performed differential gene expression analysis. Interestingly, the differentially expressed genes between the two subtypes showed limited overlap with the highly variable genes used in the transcriptomic embedding (Figure S5b,c), potentially explaining why the RNA-based approach failed to resolve the heterogeneity within EPI.

Although the subtype-to-lineage correspondence remains unclear, GO enrichment analysis of genes upregulated in each subtype provided informative clues. Genes upregulated in Subtype 1 were enriched in terms related to nervous system development, particularly *neural tube closure*, a key process during neural tube formation (Figure 5c). This observation is consistent with the presence of neural tube cells in the HiRES dataset, suggesting that Subtype 1 may represent an early neural precursor state. In contrast, Subtype 2 displayed enrichment for genes involved in non-neural developmental processes, such as *spermatogenesis* and *limb development* (Figure 5d), implicating a potential bias toward mesenchymal or non-neural lineages.

Another noteworthy finding emerged when we assessed the nature of the domain boundary features. When using a binary representation (i.e., presence or absence of domain boundaries), EPI cells could not be subdivided into two subtypes, mirroring the result from RNA data. However, when we incorporated boundary strength quantified via insulation scores at each domain boundary, CellRoB successfully resolved the two EPI subtypes. Subtype-level analysis revealed that differences in domain boundaries between subtypes manifest predominantly as changes in boundary strength, rather than through boundary loss or formation (Figure S5d). This result highlights that the key distinguishing factor between these two subpopulations lies not in the presence or absence of domain boundaries, but in the subtle modulation of boundary strength, suggesting a quantitative remodeling of 3D genome architecture during early lineage bifurcation.

Discussion

Single-cell 3D genome assays reveal chromatin variation that is obscured by population averaging, yet data sparsity makes domain boundaries difficult to identify and

compare reliably. Recurrent boundaries may be detected at slightly different genomic positions across cells, complicating the distinction between technical uncertainty and biological heterogeneity. Our results suggest that some apparent boundary differences reflect positional variability within shared boundary regions. Accounting for this variability is therefore important for establishing comparable structural features across cells and interpreting their biological significance. CellRoB addresses this challenge by representing recurrent boundary calls as peak-like regions and aligning them through the minimization of an insulation-score loss. This population-aware approach improves cross-cell comparability while retaining variation in boundary occurrence and strength. Benchmarks on Dip-C and scMicro-C data support improved boundary detection accuracy and robustness to contact dropout and genomic resolution, including analyses at resolutions from 40 kb to 5 kb. These results highlight the value of explicitly modeling boundary positions across cells in addition to recovering sparse contact signals. Because increased consistency alone does not establish biological accuracy, the accompanying epigenomic enrichment and reproducibility analyses provide complementary support for the inferred boundaries.

An important contribution of CellRoB is the definition and systematic characterization of rare boundaries shared by only a small fraction of cells. These low-frequency features can be obscured in pooled contact maps and difficult to recognize when boundary positions are inconsistent across cells. By distinguishing rare, weakly conserved and strongly conserved boundaries, CellRoB enables analysis of how boundary occurrence relates to cellular function. In mouse brain data, genes near rare boundaries were enriched for nervous system development and neuronal differentiation, whereas those near strongly conserved boundaries were associated with cell division and chromosome segregation. Rare-boundary features also retained information about cell identity. Together, these findings suggest that low-frequency boundaries capture a biologically informative component of chromatin heterogeneity, although their weaker enrichment for canonical boundary-associated factors warrants further investigation.

Beyond boundary occurrence, our findings demonstrate the value of quantitative boundary strength for characterizing cellular states. In the HiRES mouse embryo dataset, boundary-strength profiles separated the RNA-annotated EPI group into two subpopulations that were not resolved in the RNA-derived embedding or using binary boundary features. This result indicates that variation in insulation strength can reveal structural heterogeneity without requiring boundary gain or loss. The distinct transcriptional programs of these subpopulations suggest different developmental tendencies, supporting the use of boundary strength as a complementary feature for studying early cellular diversification. Their precise relationships to subsequent lineages, however, remain to be established.

Despite these advantages, CellRoB depends on input coverage and preprocessing choices, including imputation, genomic bin size and insulation-window size. Sharing information across cells may attenuate genuine structural differences, particularly in sparsely sampled populations. Its representation of recurrent boundary regions may also merge nearby but biologically distinct boundaries or inadequately capture complex positional distributions. In addition, rare-boundary classification depends on operational

frequency thresholds and the composition of the sampled population; a boundary that is rare overall may be common within a particular cell type. Systematic sensitivity analyses, cell-type-aware comparisons and independent validation will be important for assessing these effects. Because single-cell boundary annotations lack a definitive gold standard, complementary experimental measurements will also be needed to establish the physical and functional significance of inferred boundaries.

Overall, Overall, CellRoB provides a framework for connecting robust single-cell domain-boundary identification with the analysis of structural heterogeneity. Its ability to characterize low-frequency boundaries and quantify boundary-strength variation expands the information available for studying cellular identity and developmental states. Future integration with chromatin accessibility, architectural protein occupancy and imaging-based measurements could further clarify how these boundary features relate to gene regulation and cell fate.

Methods

Single-cell 3D genome Data

We used three single-cell 3D genome datasets for evaluation. First, we used the Dip-C dataset, with an average of 401,000 chromatin contacts per cell. For downstream analysis, the top 1,812 cells with more than 250,000 contacts were selected to ensure data quality. Second, we used multimodal data from HiRES, which simultaneously profiles chromatin contacts and gene expression in the same cells. Each cell contained on average 304,124 Hi-C contacts and 27,897 RNA UMIs from 6,966 genes, and 4,273 cells with more than 250,000 Hi-C contacts were retained for integrated analysis. Third, we included a scMicro-C dataset to further evaluate CellRoB on an independent high-resolution single-cell chromatin contact dataset.

Preprocess Contact Probability for each Chromosome of the individual cell

We first applied Higashi to single-cell Hi-C contact matrices binned at 500 kb to obtain a cell-cell similarity matrix $\mathbf{W}$, where W_{ij} denotes the similarity between cell i and cell j . For each cell, we retained only its top $k = 15$ nearest neighbors and set all other similarity values to zero, i.e., $W_{ij} = 0$.

Such similarity matrices can also be derived from other structural features that distinguish cell types or from other omics data. The matrix $\mathbf{W}$ captures phenotypic proximity between cells at a coarse resolution. We then used $\mathbf{W}$ to impute the 40 kb contact map of each cell as follows:

$$\mathbf{H}_i^{\text{imp}} = \frac{\mathbf{H}_i^{\text{raw}} + \sum_{j=1}^N W_{ij} \mathbf{H}_j^{\text{raw}}}{1 + \sum_{j=1}^N W_{ij}}, \quad (1)$$

where $\mathbf{H}_i^{\text{raw}} \in \mathbb{R}^{B \times B}$ is the observed 40 kb contact map of cell i , N is the total number of cells, and B is the number of genomic bins in the contact map.

Initial detection of single-cell domain boundaries

Inspired by a previously proposed fast density-based clustering algorithm for grouping data

points⁴³, we leveraged its strategy for identifying cluster centers to delineate domain boundaries along each chromosome in individual cells (Figure 1). Specifically, two indices were defined for each genomic bin.

1) Insulation strength. The insulation strength $\rho(i)$ of the i -th genomic position was defined as

$$\rho(i) = \frac{I_i^{\text{intra}} - I_i^{\text{inter}}}{I_i^{\text{intra}} + I_i^{\text{inter}}}, \quad (2)$$

where

$$I_i^{\text{intra}} = \sum_{p,q \in [i-w, i-1]} C(p, q) + \sum_{p,q \in [i+1, i+w]} C(p, q), \quad (3)$$

$$I_i^{\text{inter}} = \sum_{p \leq i-w-1} \sum_{q \geq i+w+1} C(p, q), \quad (4)$$

where $C(p, q)$ denotes the Hi-C contact count between genomic bins p and q , and $w = 15$, corresponding to 600 kb at 40 kb resolution.

2) Minimum distance to a bin with higher insulation strength. For each bin i , the minimum distance between bin i and any other bin j with higher insulation strength was defined as

$$\delta(i) = \min_{\substack{j: \rho(j) > \rho(i) \\ |j-i| \leq M}} |j-i|, \quad M = \frac{20 \text{ Mb}}{\text{bin size}}. \quad (5)$$

If no such bin j exists within the range $\pm M$, we set $\delta(i) = M$.

Let

$$\rho_{\max} = \max_i \rho(i), \quad \delta_{\max} = \max_i \delta(i). \quad (6)$$

The two indices were then normalized as

$$\rho'(i) = \frac{\rho(i)}{\rho_{\max}}, \quad \delta'(i) = \frac{\delta(i)}{\delta_{\max}}, \quad (7)$$

and combined to obtain

$$\eta(i) = \rho'(i) \delta'(i). \quad (8)$$

Thus, $\rho'(i), \delta'(i) \in [0,1]$, and $\eta(i)$ ranks genomic bins according to both insulation strength and spatial separation.

All bins were ranked in descending order of $\eta(i)$, yielding $\gamma(i) \in \{1, \dots, N\}$, where N is the total number of genomic bins. The rank was normalized as

$$\gamma'(i) = \frac{\gamma(i)}{N}. \quad (9)$$

The reflection point r was determined by

$$r = \arg \min_{1 \leq i \leq N} [\eta(i) + \gamma'(i)]. \quad (10)$$

This reflection point balances a high combined score with a low rank. Bins satisfying

$$\eta(i) > \alpha \eta(r), \quad \alpha = \frac{40 \text{ kb}}{\text{bin size}}, \quad (11)$$

were selected as the initial single-cell TAD boundary positions for the corresponding chromosome.

This procedure integrates local insulation contrast ρ , spatial separation δ , and a reflection-point cutoff r to identify significant domain boundaries in individual cells.

Defining the separation landscape of domain boundaries

1) Raw boundary signal. For each genomic bin x , the raw boundary signal was defined as

$$S_{\text{raw}}(x) = B_x, \quad x = 1, \dots, L, \quad (12)$$

where B_x denotes the number of single cells exhibiting a TAD boundary at genomic bin x , and L is the total number of genomic bins along the chromosome.

2) Signal smoothing. The raw boundary signal was smoothed by convolution with a uniform kernel:

$$S_{\text{smooth}} = S_{\text{raw}} * h, \quad (13)$$

where $*$ denotes the convolution operation and

$$h[m] = \frac{1}{M}, \quad m = 0, \dots, M-1, \quad M = \left\lceil \frac{40 \text{ kb}}{\text{bin size}} \right\rceil. \quad (14)$$

Equivalently, the smoothed signal at genomic bin x can be written as

$$S_{\text{smooth}}(x) = \sum_{m=0}^{M-1} h[m] S_{\text{raw}}(x-m). \quad (15)$$

3) Local valley detection. The set of local valleys in the smoothed boundary signal was defined as

$$\mathcal{V} = \{x \mid S_{\text{smooth}}(x) < S_{\text{smooth}}(x-1), S_{\text{smooth}}(x) < S_{\text{smooth}}(x+1)\}. \quad (16)$$

These local valleys partition the single-cell boundary positions into distinct groups.

Cross-cell boundary alignment

For each group G , we sought a single consensus bin $\hat{x}_G \in G$ that minimized the total squared loss over only those cells with an existing boundary in G .

Let

$$\mathcal{C}_G = \{c \mid \exists i \text{ such that } b_i^{(c)} \in G\} \quad (17)$$

denote the set of cells containing a boundary within group G . For each $c \in \mathcal{C}_G$, let i_c denote the position of the corresponding single-cell boundary. The loss for a candidate consensus bin $x \in G$ was defined as

$$L_G(x) = \sum_{c \in \mathcal{C}_G} [\rho(c, i_c) - \rho(c, x)]^2, \quad (18)$$

and the consensus boundary position was determined by

$$\hat{x}_G = \arg \min_{x \in G} L_G(x). \quad (19)$$

Here, $b_i^{(c)}$ denotes a single-cell boundary located at genomic bin i in cell c , and

$\rho(c, i)$ denotes the insulation strength of genomic bin i in cell c .

Boundary Conservation Classification

For each aligned boundary, we define its conservation frequency as the proportion of cells in which the boundary is detected. Let n_i denote the number of cells containing boundary i , and let N denote the total number of cells. The conservation frequency is calculated as $f_i = n_i/N$. Boundaries are categorized as rare ($1\% \leq f_i < 10\%$), weakly conserved ($10\% \leq f_i < 50\%$), or strongly conserved ($f_i \geq 50\%$).

Jaccard Index

Let $B(U)$ and $B(V)$ denote the sets of genomic bin coordinates corresponding to domain boundaries in cells U and V , respectively. The Jaccard index was then defined as

$$J(U, V) = \frac{|B(U) \cap B(V)|}{|B(U) \cup B(V)|} \quad (20)$$

where $|\cdot|$ denotes the cardinality of a set. The Jaccard index ranges from 0, indicating that no boundaries are shared, to 1, indicating that the two boundary sets are identical.

Measure of Concordance

Let

$$U = \{U_i\}_{i=1}^c \quad \text{and} \quad V = \{V_j\}_{j=1}^s$$

be two sets of TAD-like domains identified from an $n \times n$ Hi-C contact matrix, where c and s denote the numbers of domains in U and V , respectively. We define

$$p_i = \frac{|U_i|}{n}, \quad q_j = \frac{|V_j|}{n}, \quad r_{ij} = \frac{|U_i \cap V_j|}{n} \quad (21)$$

where $|U_i|$ and $|V_j|$ denote the numbers of genomic bins contained in domains U_i and V_j , respectively, and $|U_i \cap V_j|$ denotes the number of bins shared by the two domains.

Adapting the clustering-concordance metric proposed by Zufferey et al. ⁴⁴, we treated each domain as a cluster of genomic bins. Using the normalized domain overlaps r_{ij} , the measure of concordance (MoC) was defined as

$$\text{MoC}(U, V) = \begin{cases} 1, & \text{if } c = s = 1, \\ \frac{1}{\sqrt{cs}-1} \left(\sum_{i=1}^c \sum_{j=1}^s \frac{r_{ij}^2}{p_i q_j} - 1 \right), & \text{otherwise.} \end{cases} \quad (22)$$

Benchmark Construction

First, we applied each TAD-calling method to the same Hi-C contact map and collected the domain boundaries identified by each method. We then constructed a genome-wide boundary score profile using a boundary-voting strategy.

Let M denote the total number of TAD-calling methods. For each genomic bin b , the boundary score was defined as

$$S(b) = \sum_{m=1}^M \mathbb{I}(b \in B_m), \quad (23)$$

where B_m denotes the set of boundaries identified by method m , and $\mathbb{I}(\cdot)$ is the indicator function. Thus, $S(b)$ represents the number of methods that detected a boundary at genomic bin b .

Consecutive genomic bins with nonzero boundary scores, $S(b) > 0$, were merged into candidate boundary regions. For a candidate region R , its method-support score was defined as

$$S(R) = \sum_{m=1}^M \mathbb{I}(B_m \cap R \neq \emptyset). \quad (24)$$

Candidate regions supported by more than 60% of all methods,

$$\frac{S(R)}{M} > 0.6, \quad (25)$$

were designated as benchmark boundary regions.

Precision, Recall and F1-Score

Let $\mathcal{B}$ denote the set of benchmark boundary regions, and let $N = |\mathcal{B}|$ be the number of contiguous nonzero-vote regions that satisfy the predefined length and voting thresholds. Suppose method A identifies M boundaries in total, of which m overlap at least one benchmark boundary region. Among the N benchmark regions, let n denote the number that overlap at least one boundary identified by method A .

The precision and recall of method A were defined as

$$\text{Precision} = \frac{m}{M}, \quad \text{Recall} = \frac{n}{N}. \quad (26)$$

The F1-score, defined as the harmonic mean of precision and recall, was calculated as

$$\text{F1} = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \frac{\frac{mn}{MN}}{\frac{m}{M} + \frac{n}{N}} = \frac{2mn}{mN + nM}. \quad (27)$$

This metric provides a balanced assessment of the accuracy and completeness of method A in recovering the benchmark boundary regions.

Dataset availability

We used three single-cell 3D genome datasets for evaluation. First, we used the Dip-C dataset, with an average of 401,000 chromatin contacts per cell. For downstream analysis, the top 1,812 cells with more than 250,000 contacts were selected to ensure data quality. Second, we used multimodal data from HiRES, which simultaneously profiles chromatin contacts and gene expression in the same cells. Each cell contained on average 304,124 Hi-C contacts and 27,897 RNA UMIs from 6,966 genes, and 4,273 cells with more than 250,000 Hi-C contacts were retained for integrated analysis. Third, we included a scMicro-C dataset to further evaluate CellRoB on an independent high-resolution single-cell chromatin contact dataset.

The Dip-C dataset is available from NCBI GEO under accession GSE146397 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE146397>).

The HiRES dataset is available from NCBI GEO under accession GSE223917

(<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE223917>).

The scMicro-C dataset is available from NCBI GEO under accession GSE281150 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE281150>).

Code availability

The CellRoB software is available on GitHub at <https://github.com/liushiji0511/CellRoB>.

Acknowledgements

This work was supported by National Natural Science Foundation of China grant (no. 62573335 to Y.Y., nos. 62132015, 62350087 and 62550005 to L.G.), and a grant from the Fundamental Research Funds for the Central Universities (no. QTZX23074 to Y.Y.). We also thank the Key Laboratory of Computational Bioinformatics of Xi'an at Xidian University.

Author contributions

S.L. developed and implemented the CellRoB algorithm and performed the computational experiments and data analysis. Y.Y. and L.G. provided methodological guidance and supervised the study.

Computing interests

The authors declare no competing interests.

References

1. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nature Reviews Genetics* **17**, 661–678 (2016).
2. Misteli, T. The self-organizing genome: principles of genome architecture and function. *Cell* **183**, 28–45 (2020).
3. Dekker, J. & Mirny, L. The 3D genome as moderator of chromosomal communication. *Cell* **164**, 1110–1121 (2016).
4. Rao, S.S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
5. Schoenfelder, S. & Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics* **20**, 437–455 (2019).
6. Lupiáñez, D.G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
7. Pope, B.D. et al. Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**, 402–405 (2014).
8. Hirano, T. Condensin-based chromosome organization from bacteria to vertebrates. *Cell* **164**, 847–857 (2016).
9. Dekker, J. Gene regulation in the third dimension. *Science* **319**, 1793–1794 (2008).
10. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
11. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions

reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).

12. Dixon, J.R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).

13. Nora, E.P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).

14. Sexton, T. et al. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148**, 458–472 (2012).

15. Jin, F. et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).

16. Tang, Z. et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* **163**, 1611–1627 (2015).

17. Dixon, J.R. et al. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).

18. Fortin, J.P. & Hansen, K.D. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biology* **16**, 180 (2015).

19. Mifsud, B. et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics* **47**, 598–606 (2015).

20. Javierre, B.M. et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369–1384 (2016).

21. Dixon, J.R., Gorkin, D.U. & Ren, B. Chromatin domains: the unit of chromosome organization. *Molecular Cell* **62**, 668–680 (2016).

22. Phillips-Cremins, J.E. et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281–1295 (2013).

23. Symmons, O. et al. Functional and topological characteristics of mammalian regulatory domains. *Genome Research* **24**, 390–400 (2014).

24. Fraser, J. et al. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Molecular Systems Biology* **11**, MSB156492 (2015).

25. Chang, L.H., Ghosh, S. & Noordermeer, D. TADs and their borders: free movement or building a wall? *Journal of Molecular Biology* **432**, 643–652 (2020).

26. Lupiáñez, D.G., Spielmann, M. & Mundlos, S. Breaking TADs: how alterations of chromatin domains result in disease. *Trends in Genetics* **32**, 225–237 (2016).

27. Franke, M. et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).

28. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).

29. Stevens, T.J. et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59–64 (2017).

30. Nagano, T. et al. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61–67 (2017).

31. Ramani, V. et al. Massively multiplex single-cell Hi-C. *Nature Methods* **14**, 263–

535 266 (2017).

536 32. Zhou, J. et al. Robust single-cell Hi-C clustering by convolution-and random-
537 walk-based imputation. *Proceedings of the National Academy of Sciences* **116**,
538 14011–14018 (2019).

539 33. Li, X., Zeng, G., Li, A. & Zhang, Z. DeTOKI identifies and characterizes the
540 dynamics of chromatin TAD-like domains in a single cell. *Genome Biology* **22**,
541 217 (2021).

542 34. Liu, E. et al. Identifying TAD-like domains on single-cell Hi-C data by graph
543 embedding and changepoint detection. *Bioinformatics* **40**, btae138 (2024).

544 35. Zhang, R., Zhou, T. & Ma, J. Multiscale and integrative single-cell Hi-C analysis
545 with Higashi. *Nature Biotechnology* **40**, 254–261 (2022).

546 36. Li, A., Zeng, G., Wang, H., Li, X. & Zhang, Z. DeDoc2 Identifies and
547 Characterizes the Hierarchy and Dynamics of Chromatin TAD-Like Domains in
548 the Single Cells. *Advanced Science* **10**, 2300366 (2023).

549 37. Ye, Y., Zhang, S., Gao, L., Zhu, Y. & Zhang, J. Deciphering hierarchical chromatin
550 domains and preference of genomic position forming boundaries in single mouse
551 embryonic stem cells. *Advanced Science* **10**, 2205162 (2023).

552 38. Wu, H., Zhang, J., Tan, L. & Xie, X.S. Single-cell Micro-C profiles 3D genome
553 structures at high resolution and characterizes multi-enhancer hubs. *Nature*
554 *Genetics* **57**, 1777–1786 (2025).

555 39. Tan, L. et al. Changes in genome architecture and transcriptional dynamics
556 progress independently of sensory experience during post-natal brain
557 development. *Cell* **184**, 741–758 (2021).

558 40. Liu, Z. et al. Linking genome structures to functions by simultaneous single-cell
559 Hi-C and RNA-seq. *Science* **380**, 1070–1076 (2023).

560 41. Shin, H. et al. TopDom: an efficient and deterministic method for identifying
561 topological domains in genomes. *Nucleic Acids Research* **44**, e70 (2016).

562 42. Ye, Y., Gao, L. & Zhang, S. MSTD: an efficient method for detecting multi-scale
563 topological domains from symmetric and asymmetric 3D genomic maps. *Nucleic*
564 *Acids Research* **47**, e65 (2019).

565 43. Rodriguez, A. & Laio, A. Clustering by fast search and find of density peaks.
566 *Science* **344**, 1492–1496 (2014).

567 44. Zufferey, M., Tavernari, D., Oricchio, E. & Ciriello, G. Comparison of
568 computational methods for the identification of topologically associating domains.
569 *Genome Biology* **19**, 217 (2018).

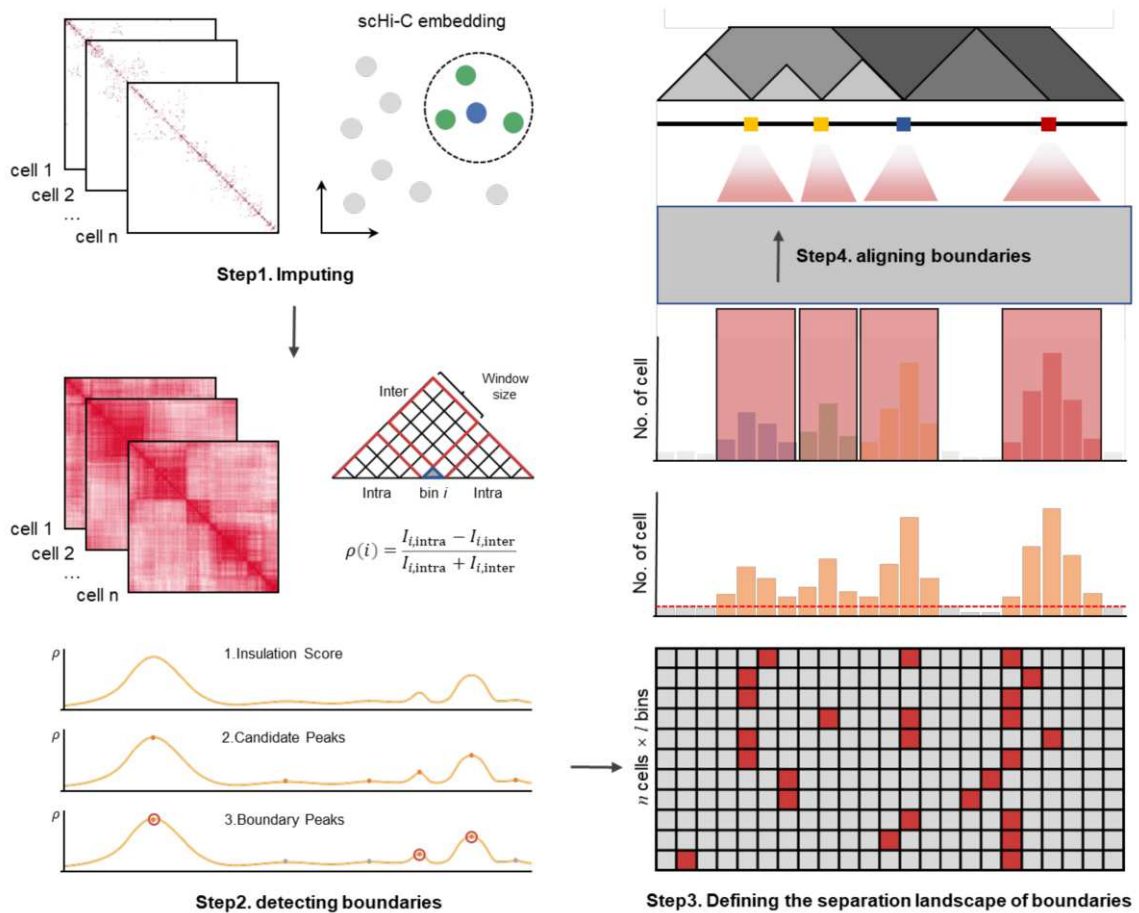


Fig. 1 | Overview of the CellRoB framework. a, Higashi embeddings project raw single-cell Hi-C contact maps into a low-dimensional space to identify structurally similar cells and impute missing contacts. **b**, Insulation scores are computed for each imputed cell to identify individual domain boundaries. **c**, Peak regions of single-cell boundaries are delineated genome-wide using an insulation-score threshold followed by Gaussian fitting. **d**, Boundaries within the same peak are aligned across cells by minimizing an insulation-score loss, yielding a consensus boundary set with improved intercellular consistency.

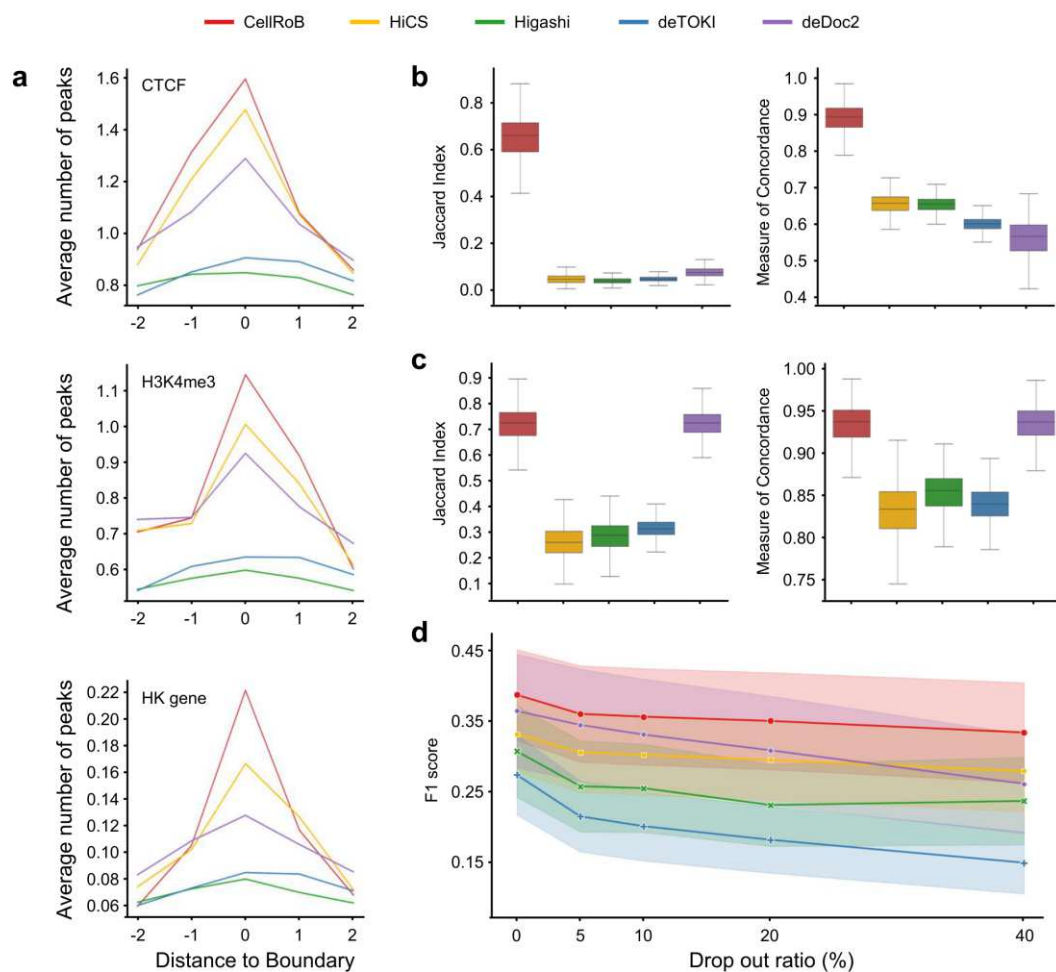


Fig. 2 | Performance evaluation of CellRoB on the Dip-C dataset. a, Comparative analysis of epigenomic enrichment at boundaries identified by different methods. **b**, Comparison of boundary-detection consistency between nearest-neighbor cell pairs. **c**, Comparison of boundary-detection consistency between biological replicate pairs. **d**, Boundary-detection performance of different methods under increasing dropout rates in single-cell Hi-C data.

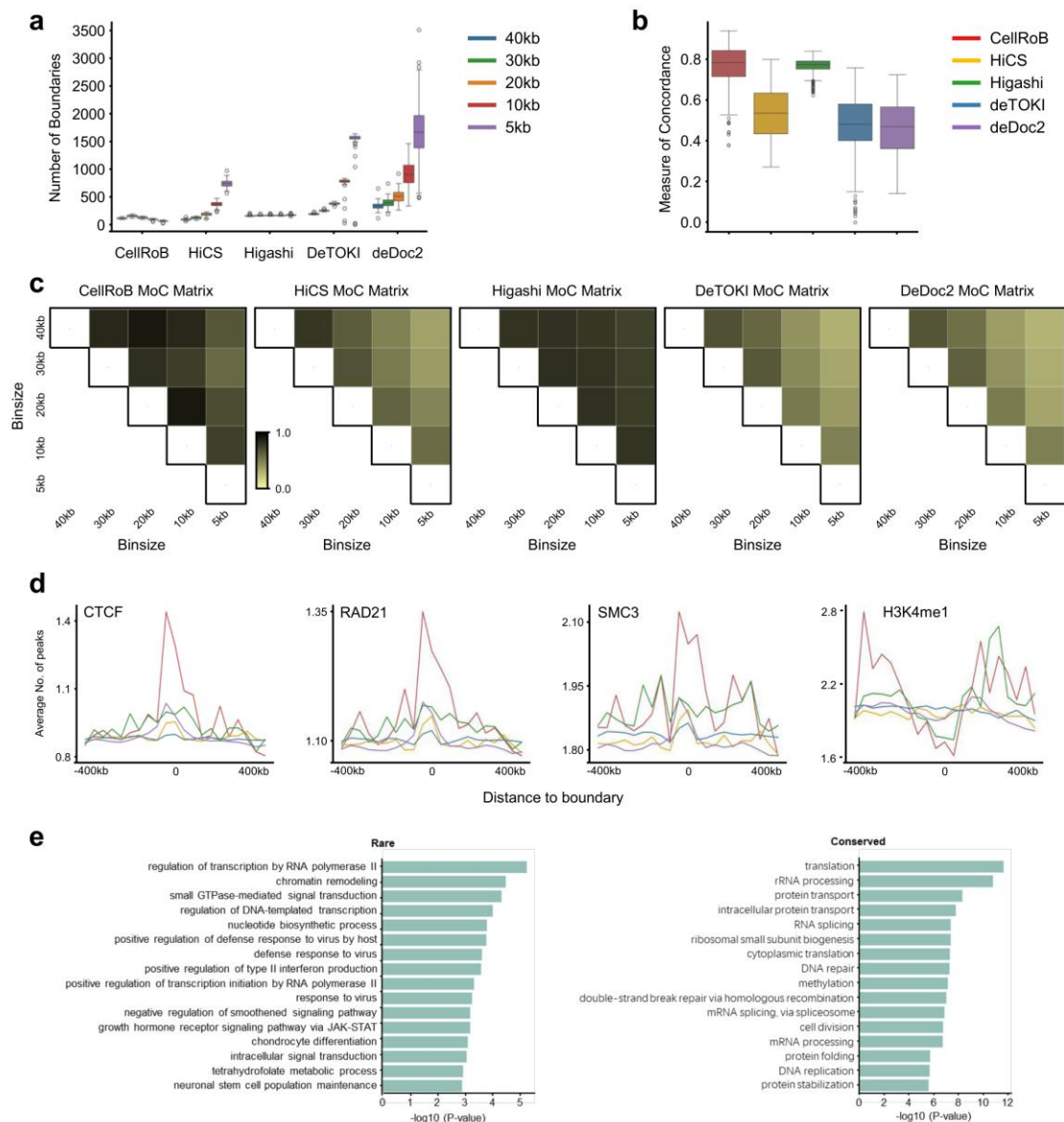


Fig. 3 | CellRoB identifies shared single-cell domain boundaries and reveals conservation-dependent biological patterns. a, Boundary-detection robustness of different methods when subsampling different numbers of cells. **b**, Comparison of CellRoB and TopDom in detecting rare domain boundaries. **c**, Epigenomic enrichment at boundaries identified by different methods. **d**, Epigenomic enrichment at CellRoB-identified boundaries stratified into rare, weakly conserved, and strongly conserved classes. **e**, UMAP visualizations of Dip-C cells using features derived from rare, weakly conserved, and strongly conserved domain boundaries. **f**, Gene Ontology enrichment analysis of genes associated with the three boundary-conservation classes, showing distinct functional enrichment patterns.

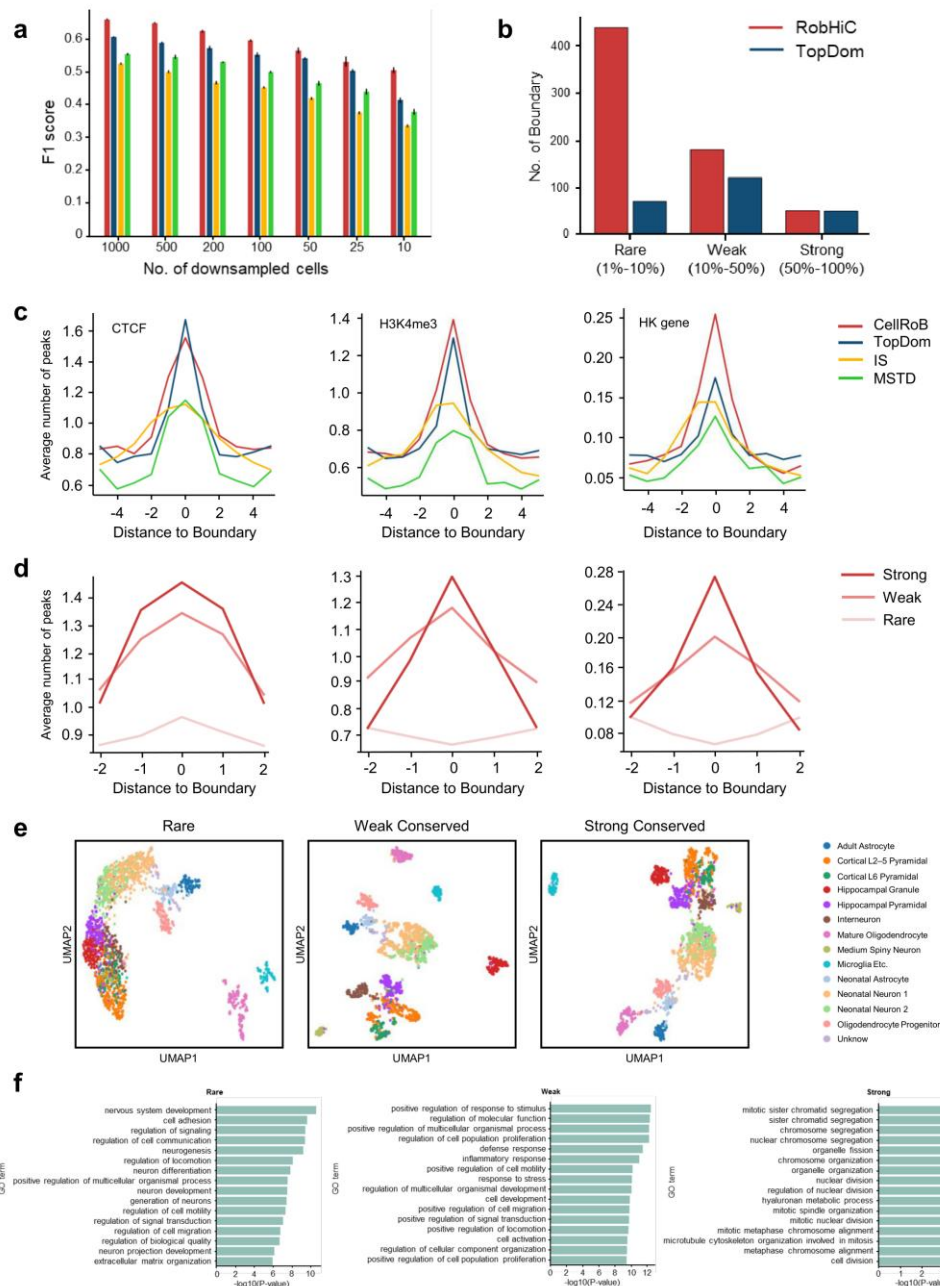


Fig. 4 | CellRoB identifies shared single-cell domain boundaries and reveals conservation-dependent biological patterns. **a**, Robustness comparison across methods when subsampling different numbers of cells. **b**, Comparison of CellRoB and TopDom in detecting rare low-frequency domain boundaries. **c**, Epigenomic enrichment at boundaries identified by different methods. **d**, Epigenomic enrichment at CellRoB-identified boundaries stratified into rare, weakly conserved, and strongly conserved classes. **e**, UMAP visualizations of Dip-C cells using features derived from rare, weakly conserved, and strongly conserved domain boundaries. **f**, Gene Ontology enrichment analysis of genes associated with the three boundary-conservation classes, revealing distinct biological functions.

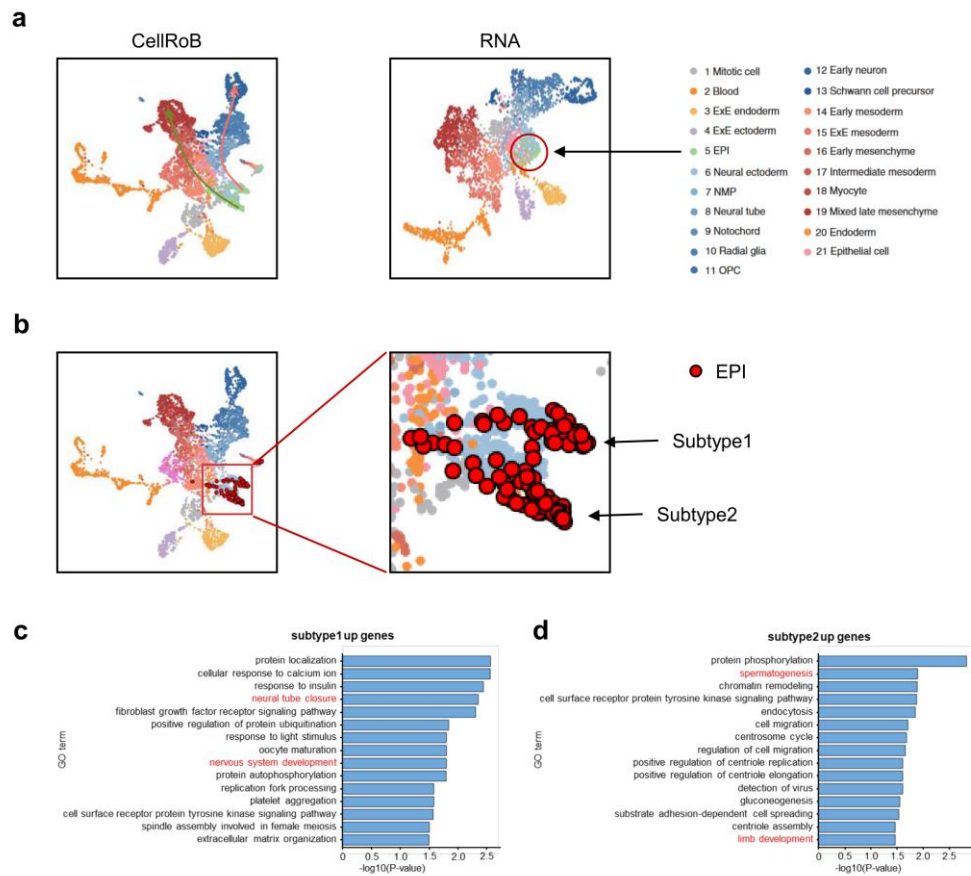


Fig. 5 | CellRoB-defined boundary features refine the characterization of cell developmental trajectories. a, UMAP visualizations of single-cell embeddings derived from CellRoB boundary features and RNA expression profiles. **b**, UMAP visualization highlighting the CellRoB-defined subpopulations of EPI cells. **c**, Gene Ontology enrichment analysis of genes upregulated in EPI Subtype 1. **d**, Gene Ontology enrichment analysis of genes upregulated in EPI Subtype 2.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryTables.xlsx](#)
- [SupplementaryNotes.docx](#)