Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Dynamix identifies accurate chromatin stripes by Bayesian Normal Dynamic Linear Model on Hi-C signal probabilities

Soomin Han

hsm0511@snu.ac.kr

Independent Researcher

Research Article

Keywords: Chromatin stripes, Bayesian Normal Dynamic Linear Model, Zero-Inflated Negative Binomial

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11721766/v1

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

Additional Declarations: No competing interests reported.

Dynamix identifies accurate chromatin stripes by Bayesian Normal Dynamic Linear Model on Hi-C signal probabilities.

Soomin Han*

Independent Researcher, Suwon, South Korea

*Corresponding Author : hsm0511@snu.ac.kr

Abstract

Background: Chromatin stripes, characterized by interactions between a localized genomic region (the stripe anchor) and a broader, continuous genomic segment (the stripe length range), are topological features visible in Hi-C interaction matrices. Existing chromatin stripe-calling algorithms often mislocalize stripes into regions of low interaction probability by operating on Hi-C normalized matrices without proper bias removal, highlighting the need to introduce more precise tool.

Results: Here, I present Dynamix, a software designed to call chromatin stripes through a two-step statistical framework. First, it computes the signal probabilities of genomic interactions using a Zero-Inflated Negative Binomial and Negative Binomial mixture expectation-maximization algorithm, while adjusting for genomic distance, GC ratio, and mappability biases. Second, utilizing this signal probability matrix, it identifies stripe anchors and length ranges via a Bayesian Normal Dynamic Linear Model. Dynamix detects the largest number of stripes, and these stripes accurately capture enhancer-promoter interactions across both their anchors and length ranges, successfully uncovering long-range enhancer-promoter interactions even beyond topologically associating domain boundaries, compared to all evaluated tools.

Conclusions: Dynamix provides a precise tool for identifying chromatin stripes, facilitating a deeper investigation of dynamic loop formation and enhancer-promoter interactions in cellular environments.

Keywords: Chromatin stripes, Bayesian Normal Dynamic Linear Model, Zero-Inflated Negative Binomial.

Background

Functional regulatory elements separated by large genomic distances frequently come into spatial contact to regulate gene expression [1]. These interactions are visible as architectural stripes in Hi-C matrix representing contact frequencies between pairs of chromatin regions quantified using reads from ligation-based assays [2, 3]. Chromatin stripe comprises two interacting genomic regions: a localized region as the stripe anchor, and a broader segment as the stripe length range [4]. These stripes typically arise from cohesin-mediated loop extrusion stalled by convergent CTCF motifs, enhancer-promoter interactions, and TAD boundaries [5, 6]. Accurate stripe detection is essential for understanding the perturbation of insulated loops leading to pathogenic alterations such as dysregulation of oncogenes [7–9].

Despite their significance, the precise detection remains challenging. While stripe anchors may appear as local maxima in Hi-C matrix, elevated contact counts often originate from short genomic distances, high GC content or poor mappability [10]. Furthermore, mapping

the interacting length range is difficult because high contact frequencies at short genomic distances overshadow distant, genuine interactions with lower contact counts.

Existing tools, such as Quagga [11] and Stripenn [12] exhibit limitations. Operating on bias-uncontrolled matrices, both methods frequently mislocalize stripes to low-probability regions. Furthermore, Quagga overestimates widths due to fixed thresholds, whereas Stripenn tends to detect most stripes originating at the main diagonal.

Accurate stripe calling requires operating on a signal probability matrix adjusting biases. However, constructing such a matrix requires heavy likelihood calculations. Although ZIPHiC [13] and FastHiC [14] utilized mixture models to decouple signals from noise, they rely on intensive hidden Markov random field (HMRF) or Bayesian MCMC calculations with high runtimes, alongside heuristic approximations of a noise component weight in ZIPHiC. Meanwhile, HiCDCPlus [15] calculates p-values of interactions rapidly using a negative binomial GLM, but its model relies on outlier filtering that can distort dispersion parameters, and the lack of a zero-inflated component can reduce accuracy in sparse datasets.

Here, I introduce Dynamix, a software that precisely detects chromatin stripe anchors and interacting length ranges. Dynamix efficiently constructs probability matrix by adjusting for distance, GC, and mappability effects using a Zero-Inflated Negative Binomial and Negative Binomial mixture expectation-maximization algorithm. The model separates signal and noise distributions while estimating parameters, such as structural zero weights and noise proportions by likelihood maxmization. Upon these signal probabilities, Dynamix captures stripe boundaries using a Bayesian normal dynamic linear model (NDLM) [16] that models the latent states of signal probabilities with coordinate-specific parameters in an efficient runtime. Consequently, compared to existing methods, Dynamix successfully identifies stripes across diverse genomic distances, exhibiting a higher density of enhancer-promoter interactions across both anchors and length ranges.

Results

Dynamix overview

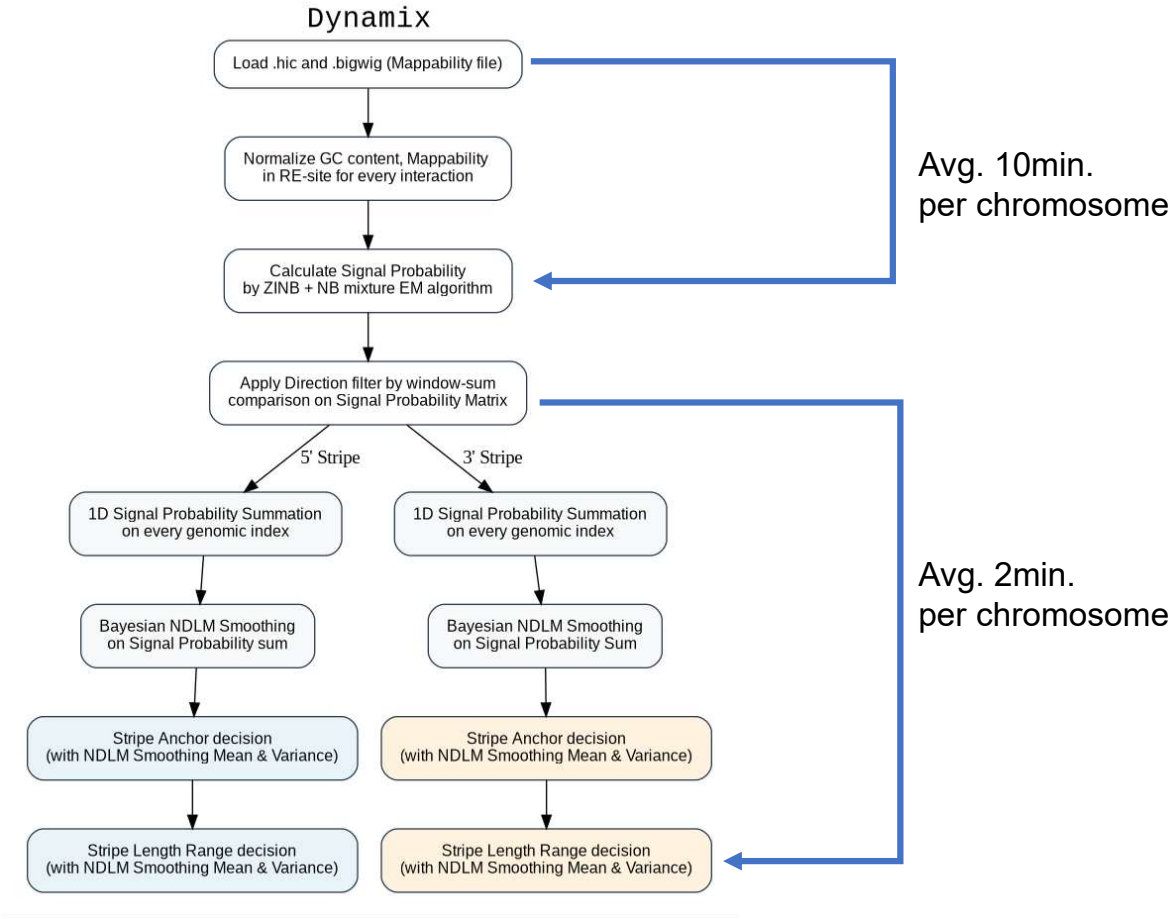


Fig. 1 Overview of Dynamix workflow including a two-step statistical framework with computationally efficient runtimes.

Dynamix identifies stripes through a two-step statistical framework (Fig. 1): first, calculating signal probabilities of interactions using a Zero-Inflated Negative Binomial (ZINB) and Negative Binomial mixture EM, and second, calling stripes via a Bayesian Normal Dynamic Linear Model (NDLM).

The pipeline loads Hi-C and mappability files to compute and normalize GC content and mappability within restriction enzyme sites for each interaction. Using distance, GC, and mappability as covariates, it iteratively converges ZINB-NB mixture distributions by a covariate-dependent GLM. From the resulting signal probability matrix, a directional filter is applied by comparing window sums and 1D summation is followed. Bayesian NDLM filtering and smoothing are performed on 1D probability sums to determine stripe anchors and select the densest continuous length ranges using coordinate-specific smoothing parameters (detailed in Methods). Averaging 12 minutes per chromosome, Dynamix achieves high computational efficiency across millions of interaction pairs (Fig. 1).

Dynamix calls greater EP associated stripes and high CTCF stripes

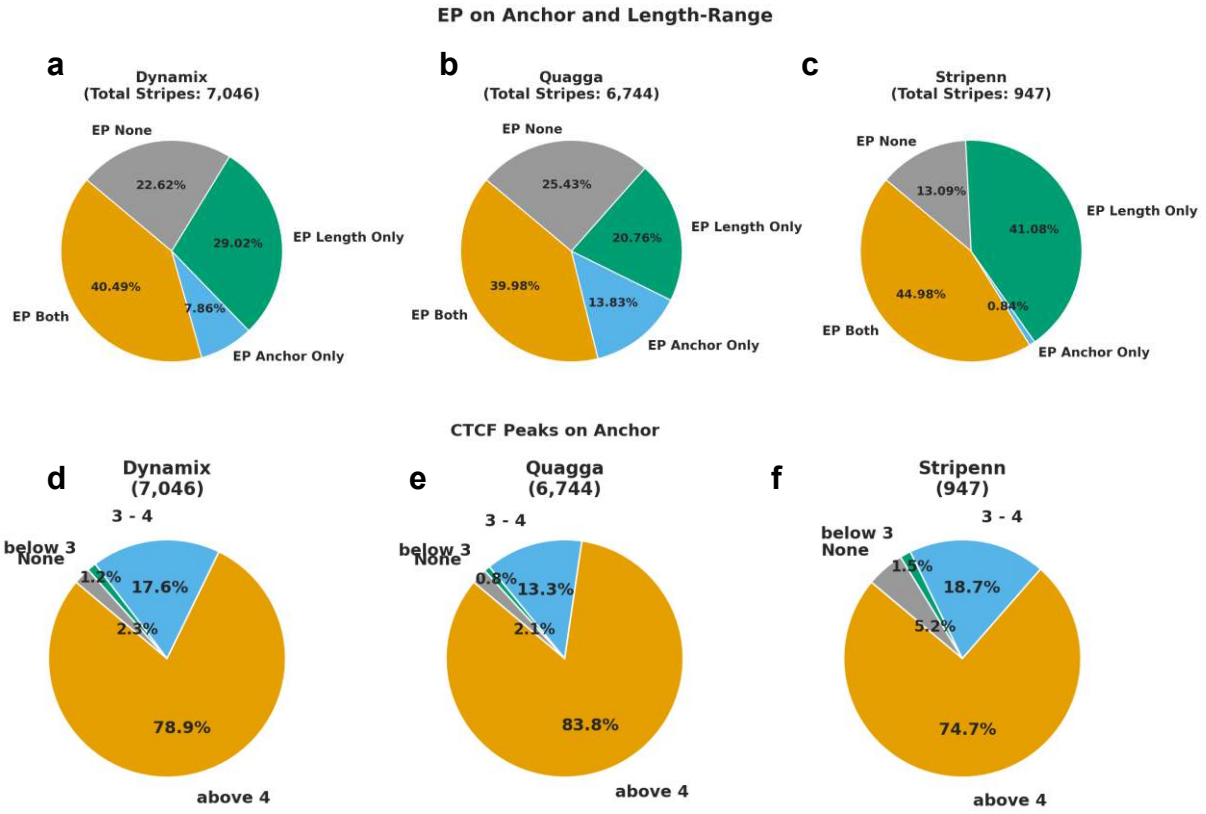


Fig. 2 Dynamix exhibits higher E-P associations alongside strong CTCF stripes.
a-c Proportion of E-P regions within stripe anchors or length-ranges identified by Dynamix, Quagga, and Stripenn (total stripe counts: 7,046, 6,744, and 947, respectively).
d-f Proportion of CTCF q-value peak categories within stripe achors for the three tools.

Dynamix calls the highest number of stripes (7,046), compared to Quagga [11] (6,744) and Stripenn [12] (947). It exhibits a higher proportion and count of ‘EP Both’ stripes—enhancer-promoter regions at both the anchor and stripe length range— and ‘EP Length range Only’ stripes than Quagga, while showing a lower fraction of ‘EP None’ (Fig. 2a, b). While Dynamix detects 7.4-fold more total stripes than Stripenn, it maintains a comparable fraction of ‘EP Both’ and yields higher counts across all EP categories (Fig. 2a, c). Although Stripenn exhibits a higher ratio for 'EP Length Only', this is attributable to its longer stripe length (2.2-fold greater than Dynamix's) (Fig. 3b), inflating the probability of capturing an EP region. Because the 'EP Anchor Only' group lacks EP regions within the interacting length range, stripes in this category are likely to have mislocated length ranges. Notably, Dynamix exhibits a lower proportion of such stripes than Quagga.

Furthermore, Dynamix exhibits strong CTCF enrichment at stripe anchors (Fig. 2d-f): 78.9% of its stripes feature high CTCF peaks and only 3.7% show low or no CTCF peaks. The proportion of high CTCF stripes in Dynamix is higher than that of Stripenn and slightly lower than Quagga. Notably, since Quagga's median stripe width is wider than that of Dynamix (Fig. 3a), this likely contributes to Quagga's higher high-CTCF proportion.

Dynamix captures variable stripe lengths with compact widths

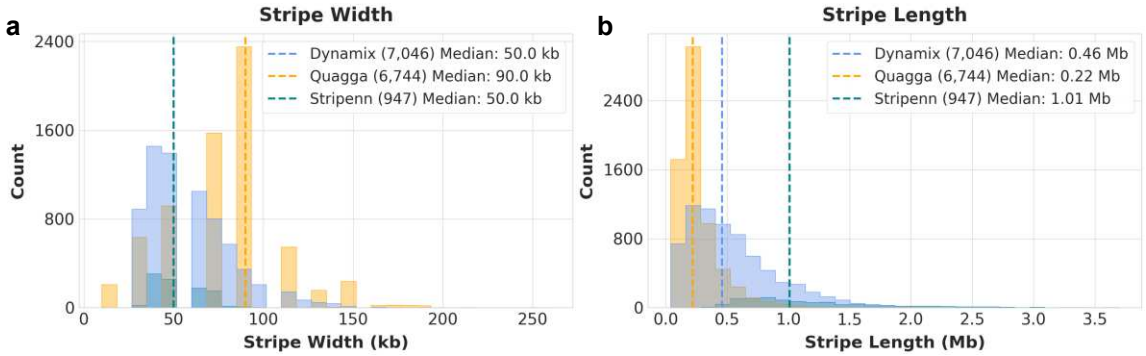


Fig. 3 Dynamix captures diverse stripe lengths while maintaining compact widths
a Stripe width distributions and median values across the three tools.
b Stripe length distributions and median values across the three tools.

Dynamix detects stripes with a median length of 460kbp (Fig. 3b)—longer than Quagga but shorter than Stripenn. While individual loops typically span shorter than 460kbp in single cells [17], Hi-C measures ensemble averages across cell populations. A single anchor (e.g., containing a plus CTCF motif) often associates with multiple downstream minus-strand motifs across different cells. With these accumulate, the resulting stripe appears longer than an individual loop size, making Dynamix's length distribution more accurately reflect this dynamic heterogeneity. Furthermore, Dynamix exhibits a narrower stripe width than Quagga (median 50kbp) (Fig. 3a), a compactness driven by Bayesian NDLM's coordinate-specific parameters rather than static thresholds.

Dynamix identifies accurate Stripe Anchor

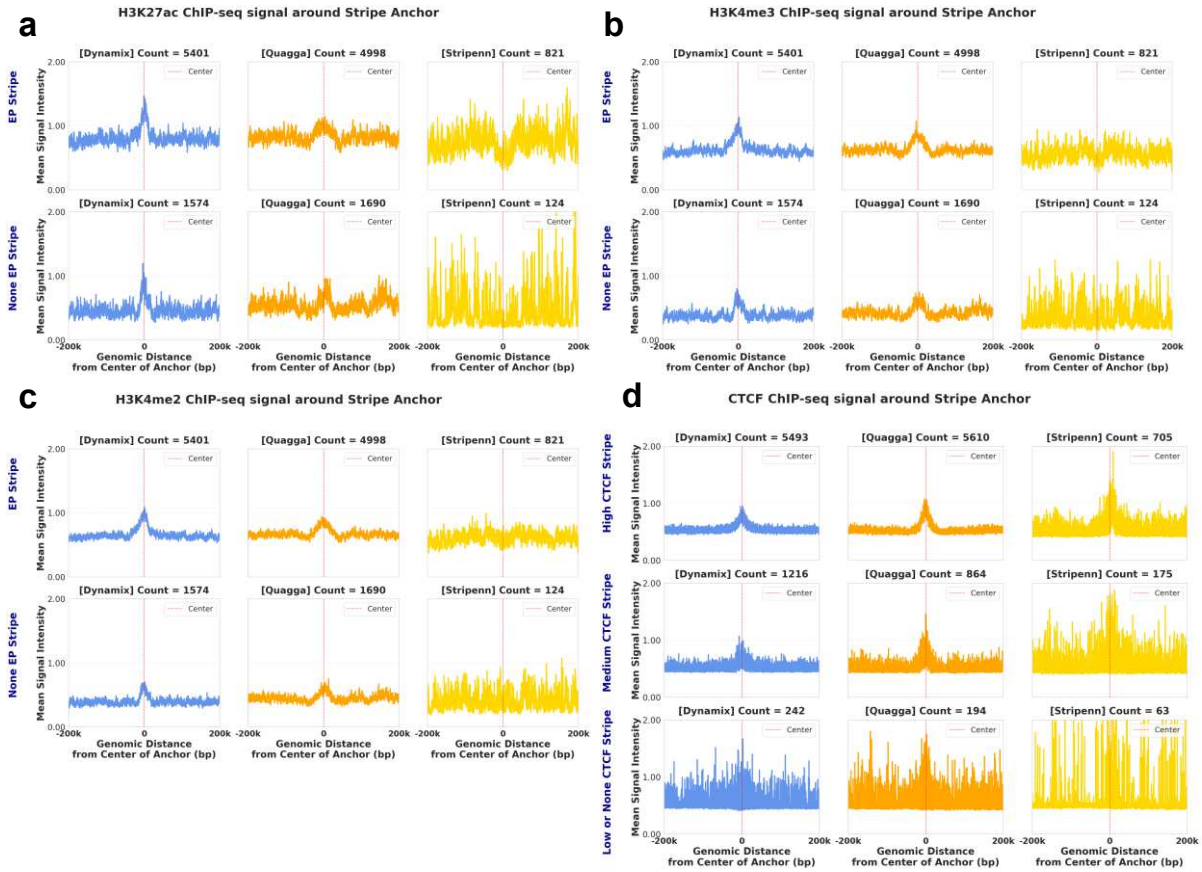


Fig. 4 Dynamix identifies accurate stripe anchors with robust ChIP-seq signal enrichment.
a-c Histone modification (H3K27ac, H3K4me3, and H3K4me2) ChIP-seq signal pileups around stripe anchors for EP and non-EP stripes across the three tools.
d CTCF ChIP-seq signal pileups around stripe anchors categorized by high, median, and low or absent CTCF levels across the three tools.

For EP stripes, Dynamix exhibits the highest ChIP-seq signal peaks at anchors for active histone marks (H3K27ac, H3K4me3, and H3K4me2) [18], the largest EP versus non-EP signal difference among evaluated methods (Fig. 4a-c), particularly for H3K27ac. This indicates that Dynamix more accurately identifies stripe anchors associated with enhancer-promoters compared to other methods. Furthermore, CTCF-associated stripes of Dynamix display robust enrichment at anchors in the high-CTCF category, progressively attenuating in the medium- and low/none-CTCF stripes (Fig. 4d).

Dynamix precisely captures Stripe-Length-Range

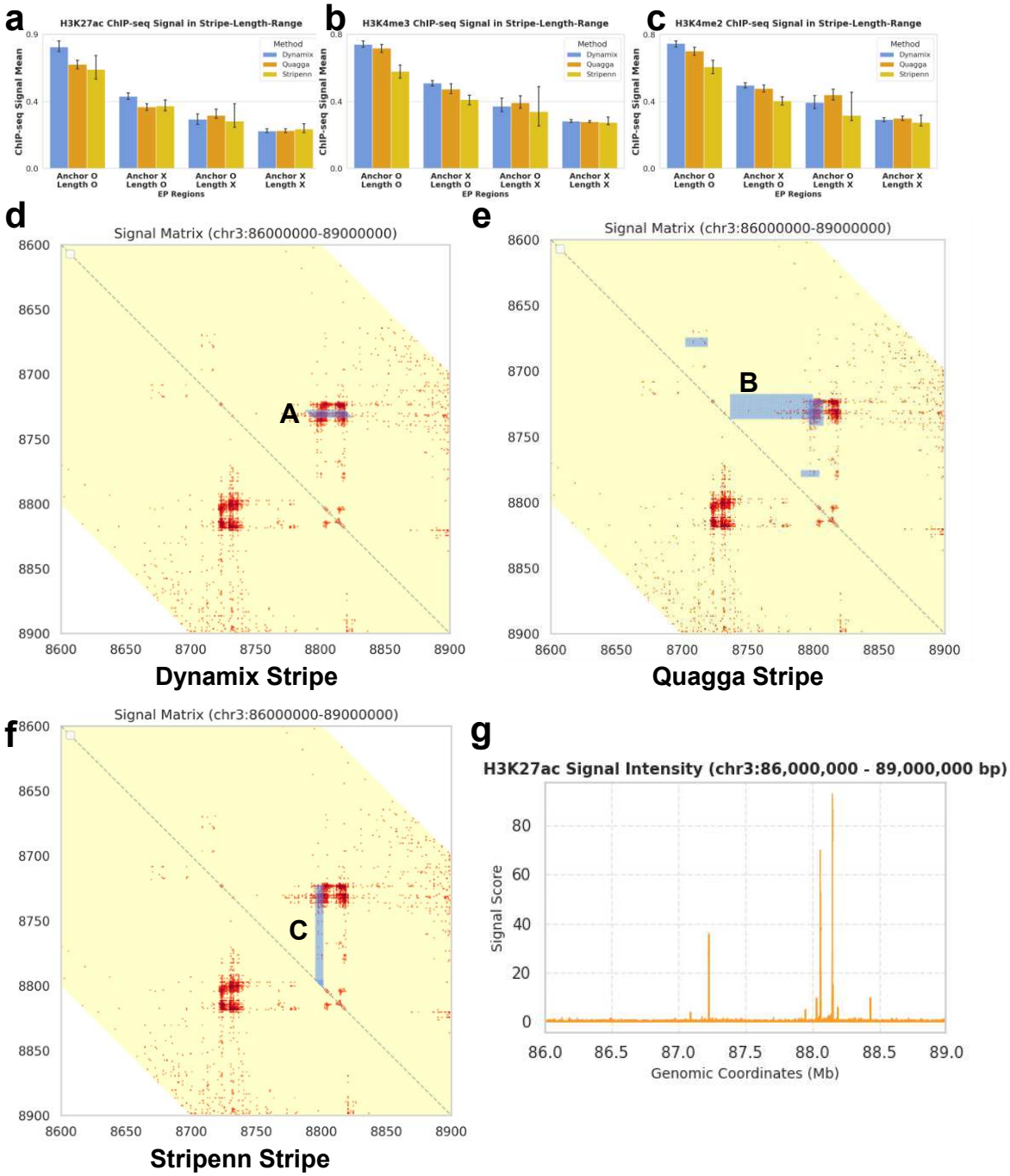


Fig. 5 Dynamix precisely captures Stripe-Length-Range
a-c Median values of mean histone modification (H3K27ac, H3K4me3, and H3K4me2) ChIP-seq signals on stripe length ranges for EP region categories.
d-f Stripe visualizations on signal probability matrices in the genomic region (chr3:86,000,000–89,000,000) for each tool.
g H3K27ac ChIP-seq signals in the genomic region (chr3:86,000,000–89,000,000).

Dynamix accurately detects not only stripe anchors but also the corresponding stripe length ranges. Dynamix EP stripes exhibit the highest histone marks ChIP-seq Length-Range signal means across categories containing EP regions within the stripe length range ('Anchor O / Length O' and 'Anchor X / Length O') (Fig. 5a-c). Despite having a median

length 2.1-fold greater than Quagga (Fig. 3b), Dynamix maintains higher ChIP-seq signal intensity without dilution. Conversely, for the 'Anchor O / Length X' and 'Anchor X / Length X' groups, Dynamix shows lower ChIP-seq length-range signal, further validating the precision of its length-range localization with active E-P interactions.

As an example of Dynamix's precise length ranges, Dynamix Stripe A accurately captures H3K27ac ChIP-seq peaks at chr3, 87.2 Mb and 88.1 Mb (Fig. 5g), with its anchor around 87.2 Mb and its length range spanning about 88.0 Mb to 88.2 Mb (Fig. 5d). In contrast, Quagga Stripe B includes low-probability start regions due to an extended length range (87.2–88.2 Mb) (Fig. 5e), while Stripenn Stripe C initiates from the main diagonal with marginal overlaps in the end region (Fig. 5f). Dynamix's precision stems from bias-corrected signal probability map, which avoids the close-distance, high-count distortions seen in bias-unremoved matrices.

Dynamix Captures Stripes with High Signal Probability

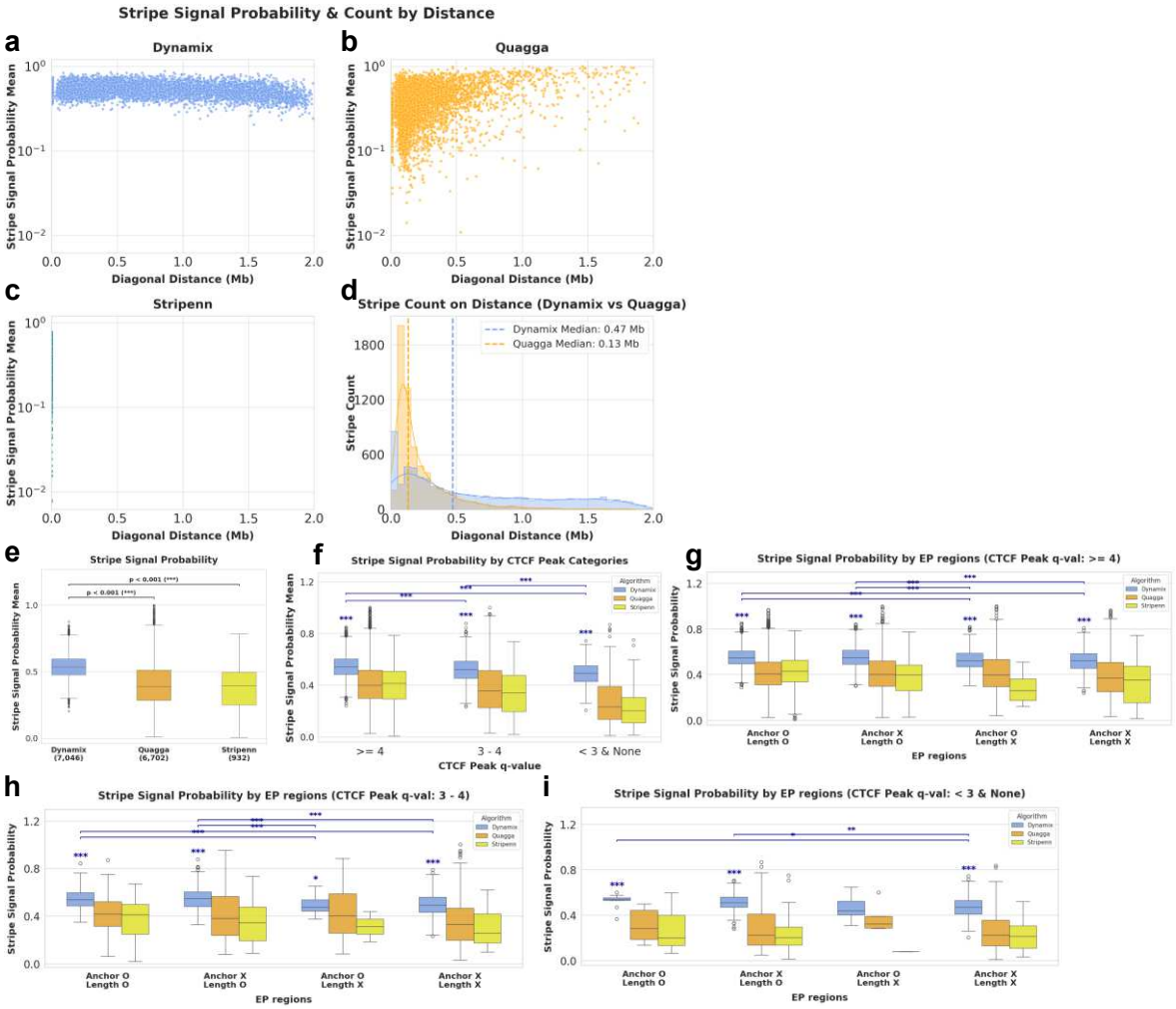


Fig. 6 Dynamix Captures Stripes with High Signal Probability
a-c Scatter plots of stripe signal probability means by diagonal distance of stripe for each tool
d Distribution of stripe counts by diagonal distance for Dynamix and Quagga.
e Box plots of stripe signal probability means across all stripes for each tool, with statistical significance between Dynamix and other tools evaluated by t-tests (denoted by asterisks).
f Box plots of stripe signal probability means grouped by CTCF peak intensity categories. Statistical significance from pairwise t-tests between groups within Dynamix, as well as comparisons across tools (Dynamix vs. Quagga and Stripenn) within each category, is indicated by asterisks.
g-i Box plots of stripe signal probability means grouped by enhancer-promoter (EP) region categories within high (g), medium (h), and low or absent (i) CTCF stripe groups. Statistical significance from pairwise t-tests between groups within Dynamix, as well as comparisons across tools (Dynamix vs. Quagga and Stripenn) within each category, is indicated by asterisks.

Dynamix maintains high stripe signal probabilities (mean $P_{unfiltered}(i,j)$ per stripe; Eq.11) across all diagonal distances, with only 0.3% falling below stripe signal probabilty 0.3, compared to 28.2% for Quagga and 31.9% for Stripenn (Fig. 6a-d). Across all CTCF and E-P categories, Dynamix achieves statistically significantly higher signal probabilities (Student's t-test; Fig. 6e-i). These probabilities reflect biological trends, decreasing with weaker CTCF intensity (Fig. 6f) and showing higher values for 'EP Both' and 'EP Length Only' groups over 'None EP' and ‘EP Anchor Only’ categories (Fig. 6g-i) (to eliminate effects of CTCF, EP stripes were stratified within matching CTCF peak groups). Notably, even 'None EP' stripes in low CTCF categories maintain a high median probability of 0.47 (Fig. 6i). Rather than indicating them as false positives, this group likely includes stripes embedded within long-range loops including E-P regions just outside the stripe boundaries.

Dynamix robustly identifies long-range interacting stripes

Dynamix detects numerous long-range stripes with enriched EP proportions and strong histone modification signals

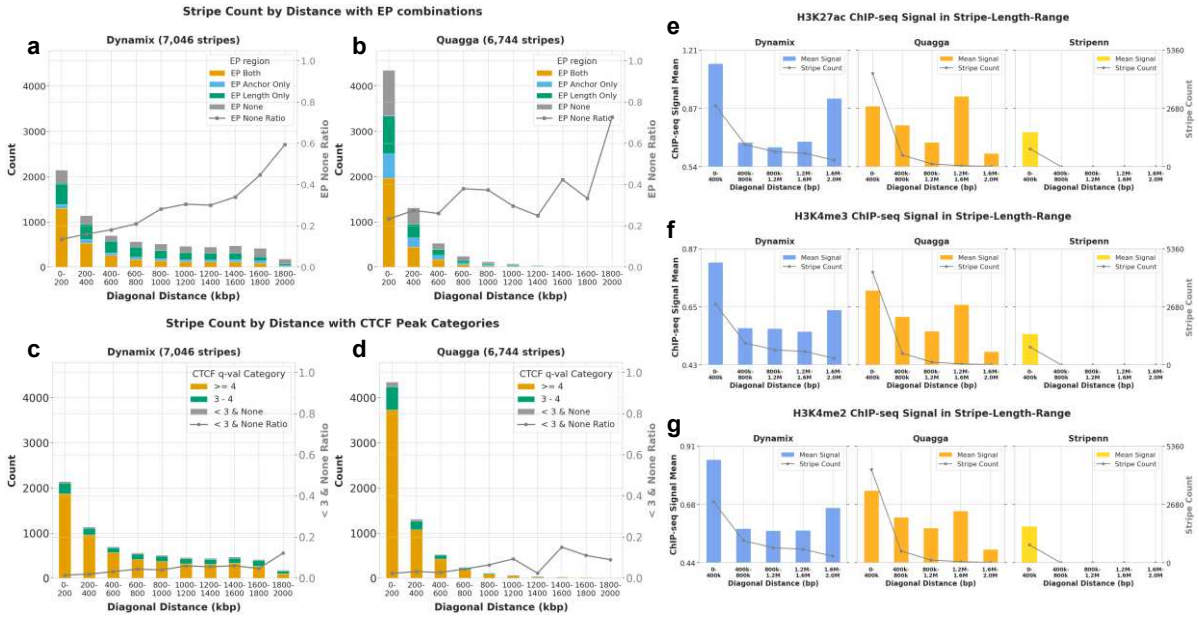


Fig. 7 Dynamix robustly identifies long-range interacting stripes
a,b Number of stripes grouped by diagonal distance bins (1 bin = 200 kb), with EP region categories shown as stacked bars and the non-EP stripe ratio indicated by the curve, for Dynamix (a) and Quagga (b).
c,d Number of stripes grouped by diagonal distance bins, with CTCF peak intensity categories shown as stacked bars and the low- or non-CTCF stripe ratio indicated by the curve, for Dynamix (c) and Quagga (d).
e-g Mean histone modification ChIP-seq signals (H3K27ac, H3K4me3, and H3K4me2, respectively) in stripe length ranges grouped by diagonal distance bins (1 bin = 400 kb) for EP stripes across the tools. The gray curved line represents the stripe count in each bin.

Dynamix exhibits a balanced distance distribution (56.1% < 600 kb; ~7% across extended bins), unlike Quagga's short-range concentration (91.5% < 600 kb) and Stripenn's 0 kb restriction (Fig. 7a-d) (The diagonal distance is defined as the minimum interaction distance $|i-j| * resolution$ across all interaction pairs (i,j) comprising a stripe). This broader distance achieved by Dynamix better reflects the accumulation of chromatin loop variability across cell populations [7]. Furthermore, Dynamix shows lower non-E-P proportions across most distance intervals with robust sample sizes (>400 stripes/bin) and keeping low/non-CTCF ratios under 0.1 across all bins except the farthest (Fig. 7a-d).

Dynamix's active histone mark ChIP-seq signals in EP stripe length range remain robust across distances, showing an intermediate dip and a farthest-bin resurgence (driven by shorter stripe lengths in farthest-bin) (Fig. 7e-g). This reflects the tendency of long-range interactions in extended loops to feature less compact E-P regions [19]. Furthermore, Dynamix yields highest mean ChIP-seq signals at short ranges where the majority of stripes by alternative methods are concentrated. Although Quagga exhibits a higher signal at 1,200–1,600 kb, Dynamix detects 12.8-fold more stripes in this region, revealing that Quagga's elevated mean stems from a sparse subset of stripes captured in bias-unadjusted matrices (Fig. 7e-g). Despite maintaining average lengths twice those of Quagga, Dynamix sustains high ChIP-seq signals across all bins, confirming the accurate capture of long-range cross-TAD E-P interactions [17].

Dyanmix exclusively captures a long-range enhancer-promoter interaction at the *HDAC6* locus



Fig. 8 Dynamix accurately captures long-range enhancer-promoter interactions.
a-c Stripe visualizations on signal probability matrices in the genomic region (chrX:47,000,000–49,500,000) for each tool.
d Comparison of interaction loops in stripes (Dynamix Stripes A–D and Quagga Stripes E–G), H3K27ac ChIP-seq signals, and candidate promoter and enhancer positions validated by CRISPRi-FlowFISH in K562 cells, along with RefSeq genes at the bottom, all within the genomic region (chrX:47,000,000–49,500,000).

Dynamix captures long-range enhancer-promoter interactions between the *HDAC6*

promoter (around chrX:47,000 kb) and enhancer (around chrX:49,000 kb) (Fig. 8a, d), supported by H3K27ac signals and K562 CRISPRi-FlowFISH validation [20]. Specifically, Dynamix stripe A captures E-P interactions spanning ~1,800 kb, with stripes B and C capturing ~1,400 kb interactions within high signal probability regions. These distances exceed the average TAD size of 880kb [21], reflecting the existence of cross-TAD E-P interaction [17]. In contrast, Quagga stripe E, F and G fail to extend beyond 500 kb where long-range E-P interactions reside and Stripenn stripe H is restricted to 1,000 kb interactions, due to their reliance on unadjusted matrices rather than probability maps (Fig. 8b-d).

Dynamix Stripe directions strongly match CTCF motif orientation

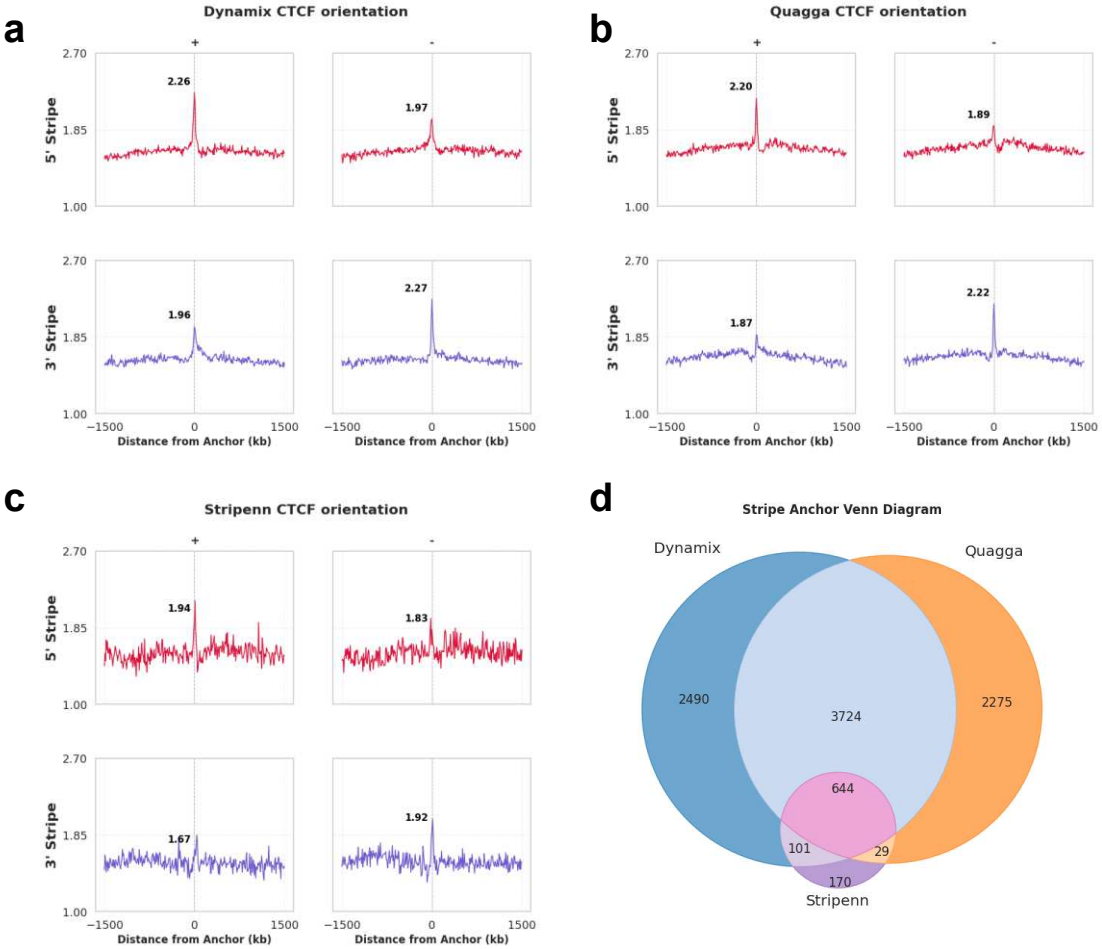


Fig. 9 Dynamix Stripe directions strongly match CTCF motif orientation
a-c Average CTCF plus (+) and minus (-) motif counts around stripe anchors for 5' and 3' stripes across each tool.
d Venn diagram illustrating the overlap of stripe anchors across the tools, with an intersection defined within the stripe width range using a 2 kb tolerance.

Dynamix exhibits the highest enrichment of matching CTCF motifs at stripe anchors (e.g., 2.26 for 5' stripes on the plus strand), showing a clear distinction from the opposite strand (Fig. 9a-c). This strand-matched capability reflects *in vivo* loop architecture, where anchoring motifs bridge downstream elements [7]. Furthermore, Dynamix achieves the highest proportion of uniquely identified anchors among evaluated tools while maintaining the highest enrichment of matching CTCF motifs (Fig. 9d).

Signal probabilities of Dynamix are reliable by robust ZINB-NB model

Model Coefficient	ZINB component significance (Pval < 0.05)	NB component significance (Pval < 0.05)	Total number of coefficients (n)
Intercept	100.00%	100.00%	46
Distance (ns1~ns6)	100.00%	100.00%	276
GC ratio	91.30%	100.00%	46
Mappability	34.78%	82.61%	46

Table 1 Statistical significance of coefficients in the converged ZINB–NB mixture EM model.
Proportions of statistically significant coefficients (defined as p < 0.05) and total coefficient counts across model components.

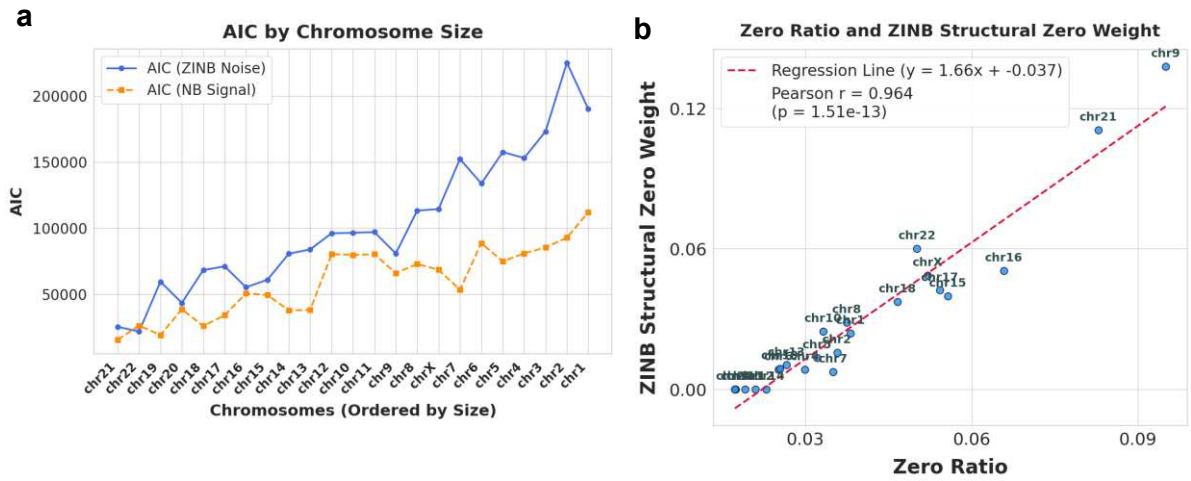


Fig. 10 Signal probabilities of Dynamix are reliable by robust ZINB-NB model
a AIC values of the ZINB noise distribution and NB signal distribution, sorted by chromosome size.
b Correlation between the structural zero weight of the ZINB component and zero count ratio across chromosomes, overlaid with a linear regression line (red dashed line, y = 1.66x - 0.37), Pearson's correlation coefficient (r = 0.964), and p-value (p = 1.51e-13).

The robust convergence of the ZINB-NB mixture EM model across all 23 chromosomes validates the credibility of Dynamix's signal probabilities (Table 1). Coefficient analysis reveals strong significance for both signal (NB) and noise (ZINB) components, with distance coefficients achieving 100% significance, whereas low-count noise (ZINB) exhibits weaker linear trends in GC (91.4%) and mappability (34.8%) due to count characteristics. Additionally, GC coefficients are consistently positive, while mappability coefficients vary across chromosomes (Supplementary Table 1).

The upward trend of the Akaike Information Criterion (AIC) [22] across chromosome sizes reflects model robustness (Fig. 10a). Larger chromosomes accumulate more residuals between expected and observed counts, and ZINB exhibits a higher AIC than NB due to additional discrepancies from zero-count interactions. Furthermore, the structural zero weight w_{zero} (Eq. 7) of the ZINB model strongly correlates with the empirical zero count ratio $(\mathrm{R}^2 = 0.929,\ \mathrm{F} = 275.4)$ (Fig. 10b), confirming model convergence. Detailed parameters of the mixture model across all chromosomes are provided in Supplementary Table 1.

Methods

Dynamix identifies chromatin stripes using a Bayesian normal dynamic linear model (NDLM) [16] applied to the summed signal probabilities of interactions that are calculated by an expectation-maximization (EM) algorithm based on a zero-inflated negative binomial (ZINB) and negative binomial (NB) mixture model.

1. Calculate signal probabilites by ZINB-NB mixture EM

First, 1% of the data is randomly and independently sampled from both zero-count and non-zero-count groups for 201 genomic distance categories from 0 bin to 200 bin (1 bin = 10kb, the default resolution).

For these selected interactions (i, j), spanning genomic regions $from\ i * resolution\ to\ (i + 1) * resolution\ (bp)$ and $from\ j * resolution\ to\ (j + 1) * resolution\ (bp)$, the expected counts μ_{ij} are modeled by negative binomial regression-based generalized linear model of genomic distance, GC ratio, mappability score. The distance effect is modeled using a natural spline of $\log(D_{ij} + 2)$ with six degrees of freedom.

$$\log(\mu_{ij}) = \beta_0 + \sum_{k=1}^{6} \beta_k * ns_k(\log(D_{ij} + 2)) + \beta_7 * \log(gc_{ij}) + \beta_8 * \log(map_{ij}) \quad (1)$$

gc_{ij} : Normalized log-sum of GC ratios in the restriction enzyme cut sites for interaction (i, j), obtained from HiCDCPlus [15].

map_{ij} : Normalized log-sum of mappability score in the restriction enzyme cut sites for interaction (i, j), obtained from HiCDCPlus.

x_{ij} : Hi-C contact read counts for interaction (i, j)

The probabilities for the ZINB component P_{ZINB} and NB component P_{NB} are calculated by incorporating the structural zero weight $w_{zero,prev}$ and the ZINB component weight w_{prev} from the previous EM iteration, combined with the negative binomial likelihood parameterized by μ_{ij} and dispersion parameter θ.

$$P(x_{ij}|ZINB) = \begin{cases} w_{zero,prev} + (1 - w_{zero,prev}) * NB(0; \mu_{ij}^{ZINB}, \theta_{ZINB}) & if\ x_{ij} = 0 \\ (1 - w_{zero,prev}) * NB(x_{ij}; \mu_{ij}^{ZINB}, \theta_{ZINB}) & if\ x_{ij} \neq 0 \end{cases} \quad (2\text{-}1)$$

$$P_{ZINB} = P(x_{ij}|ZINB) * w_{prev} \quad (2\text{-}2)$$

$$P_{NB} = (1 - w_{prev}) * NB(x_{ij}; \mu_{ij}^{NB}, \theta_{NB}) \quad (x_{ij} \neq 0) \quad (3\text{-}1)$$

At each iteration, the posterior probability vector v, representing the component membership, is computed for every interaction.

$$v = \begin{cases} (1, 0) & if\ x_{ij} = 0 \\ \left(\frac{P_{ZINB}}{P_{ZINB}+P_{NB}}, \frac{P_{NB}}{P_{ZINB}+P_{NB}}\right) & if\ x_{ij} \neq 0 \end{cases} \quad (4)$$

To calculate the structural zero weight in ZINB group, the posterior probability vector v_{zin}—membership in the structural zero versus NB component—is computed for every interaction, with non-zero counts assigned (0, 1) since they cannot originate from structural zeros.

$$v_{zin} = \begin{cases} \left(\frac{w_{zero,prev}}{C_{ij}}, \ \frac{(1-w_{zero,prev}) * NB\left(0; \mu_{ij}^{ZINB}, \theta_{ZINB}\right)}{C_{ij}} \right) & if\ x_{ij} = 0 \\ (0,1) & if\ x_{ij} \neq 0 \end{cases} \quad (5\text{-}1)$$

$$C_{ij} = w_{zero,prev} + \left(1 - w_{zero,prev}\right) * NB\left(0; \mu_{ij}^{ZINB}, \theta_{ZINB}\right) \quad (5\text{-}2)$$

The component weight w and structural zero weight w_{zero} are updated using vector elements from v and v_{zin}.

$$w = mean(v_1) \quad (6)$$

$$w_{zero} = \frac{\sum v_{zin,1} * v_1}{\sum v_1} \quad (7)$$

To estimate the coefficients β and dispersion parameter θ by optimizing negative binomial likelihoods in Equations (2-1) and (3-1)—which are used to calculate P_{ZINB} and P_{NB} —distinct fit weights for the likelihoods are applied across interactions as follows:

$$fit\ weight\ for\ P_{ZINB} = \frac{v_{1,prev} * v_{zin,2,prev}}{\sum (v_{1,prev} * v_{zin,2,prev})} \quad (2\text{-}3)$$

$$fit\ weight\ for\ P_{NB} = \frac{v_{2,prev}}{\sum v_{2,prev}} \quad (3\text{-}2)$$

($v_{1,prev}$, $v_{2,prev}$, $v_{zin,2,prev}$ are $v_1, v_2, v_{zin,2}$ in previous iteration)

Parameters are updated via the EM algorithm until convergence:

$$Q = \log(w * v_1 * P_{ZINB}) + \log((1-w) * v_2 * P_{NB}) \quad (8)$$

$$convergence\ condition: \frac{|Q - Q_{prev}|}{|Q_{prev}|} < 10^{-5} \quad (9)$$

Upon convergence, the background ZINB distribution is established with parameters β, θ_{ZINB}, and w_{zero} obtained from the final iteration and estimated from the 1% sampled data. The expected mean counts μ_{ij}^{ZINB} are computed for the entire dataset from Eq. 1 and p-values are calculated.

$$p\ value = \begin{cases} 1 & if\ x_{ij} = 0 \\ (1 - w_{zero}) * P\left(X \geq x_{ij} \mid \mu_{ij}^{ZINB}, \theta_{ZINB}\right) & if\ x_{ij} \neq 0 \end{cases} \quad (10)$$

Signal Proability

The p-values are logit-transformed, and fitted with a two-component Gaussian mixture model (GMM) via EM. The component with the smaller mean (μ_{signal}) defines the signal group, yielding the unfiltered signal probability $P_{unfiltered}(i,j)$.

$$P_{unfiltered}(i,j) = \begin{cases} 0 & if\ p\ value = 1 \\ 1 & if\ p\ value = 0 \\ P(X \geq logit(p\ value) \mid \mu_{signal}, \sigma_{signal}) & if\ 0 < p\ value < 1 \end{cases} \quad (11)$$

Thresholding $P_{unfiltered}(i,j)$ below 0.5 eliminates minor probabilities to obtain final stripe-calling probability $P(i,j)$.

$$P(i,j) = \begin{cases} 0 & if\ P_{unfiltered}(i,j) < 0.5 \\ P_{unfiltered}(i,j) & if\ P_{unfiltered}(i,j) \geq 0.5 \end{cases} \quad (12)$$

2. Calling stripes by Bayesian NDLM

For each interaction (i, j) ($i \leq j$), a direction number dir_{ij} is assigned by comparing the row- and column-wise window sums of $P(i,j)$.

$$dir_{ij} = \begin{cases} 1 & if\ \sum_{k=j-w}^{j+w} P(i,k) > \sum_{k=i-w}^{i+w} P(k,j) \\ 2 & if\ \sum_{k=j-w}^{j+w} P(i,k) < \sum_{k=i-w}^{i+w} P(k,j) \\ 3 & if\ \sum_{k=j-w}^{j+w} P(i,k) = \sum_{k=i-w}^{i+w} P(k,j) \end{cases} \quad (13\text{-}1)$$

$$w = \frac{25*10^4}{resolution} \quad (13\text{-}2)$$

5' and 3' stripe are calculated from row- and column-sums using specific direction criteria.

$$S_{5prime}(i) = \sum_{k=i}^{i+d} P(i,k) * I(dir_{ik} \in \{1,3\}) \quad (14)$$

$$S_{3prime}(i) = \sum_{k=i-d}^{i} P(k,i) * I(dir_{ki} \in \{2,3\}) \quad (15)$$

$$d = \frac{200*10^4(\max distance)}{resolution} \quad (16)$$

For $S_i \in \{S_{5prime}(i), S_{3prime}(i)\}$, Bayesian inference of a first-order polynomial normal dynamic linear model (NDLM) is computed along forward and reverse genomic bin orders:

$$S_i = \theta_i + h_i, \quad h_i \sim N(0,H) \quad (17)$$

$$\theta_i = \theta_{i-1} + q_i, \quad q_i \sim N(0,Q) \quad (18)$$

$$\theta_0 \sim N(m_0, C_0) \quad (19)$$

$$H = 10 * var(S_i)$$

$$Q = \frac{H}{sm}$$

$$m_0 = mean(S_i)$$

$$C_0 = 10^4$$

θ_i : Latent state of the signal-probability-sum at bin i (genomic region spanning $from\ i * resolution\ to\ (i+1) * resolution\ (bp)$)

h_i : observation noise at bin i

q_i : system noise at bin i

H : observation variance

Q : system variance

sm : smoothing factor (default = 150).

m_0, C_0 : prior mean and variance for the initial state θ_0

The prior variance C_0 was set to 10^4 for non-informative prior, ensuring the posterior distribution of θ_i is driven predominantly by data rather than prior assumptions after fitting the NDLM.

For the reverse NDLM, proceeding backward from the end of the genomic index, the

equations (from N down to 1) are as follows and use the same parameter values as the forward NDLM.

$$S_i = \theta_i + h_i, \quad h_i \sim N(0, H) \tag{20}$$

$$\theta_i = \theta_{i+1} + q_i, \quad q_i \sim N(0, Q) \tag{21}$$

$$\theta_{N+1} \sim N(m_{N+1}, C_{N+1}) \tag{22}$$

$$m_{N+1} = m_0, \quad C_{N+1} = C_0 \tag{22-2}$$

where N represents the maximum genomic bin index of the region (spanning from $N * resolution\ to\ (N+1) * resolution$).

Starting from the prior distributions of θ_0 and θ_{N+1}, Dynamix performed Bayesian NDLM filtering to obtain the filtering means and variances for each θ_i. Bayesian NDLM smoothing was conducted using these filtering results (see Supplementary Note 1 for detailed equations). Final smoothing distributions for both directions are denoted as:

$$\theta_i^{forward} \sim N(a_i^{forward}, R_i^{forward}) \tag{23}$$

$$\theta_i^{reverse} \sim N(a_i^{reverse}, R_i^{reverse}) \tag{24}$$

Then the distribution of θ_i was obtained by combining the parameters from both directions as follows:

$$\theta_i \sim N(a_i,\ R_i) \tag{25}$$

$$a_i = \frac{a_i^{forward} + a_i^{reverse}}{2} \tag{26}$$

$$R_i = \frac{R_i^{forward} + R_i^{reverse}}{4} \tag{27}$$

The p-value of S_i is calculated from the distribution in Eq. (25) as follows:

$$P(X \geq S_i \mid a_i, R_i) \tag{28}$$

The stripe candidate probability γ_i for S_i is computed based on the GMM algorithm (Eq. 11) using the logit-transformed p-values.

Stripe Candidates

Stripe candidate indices (SC) are obtained by taking the intersection of the indices where γ_i>0.5 and the local maximum indices of the NDLM smoothing distribution (a_is).

$$Stripe\ candidates\ SC = \{\, i \mid \gamma_i > 0.5 \,\} \cap \{\, i \mid i \in LocMax(a_i) \,\} \tag{29}$$

Stripe width range (anchor)

For Stripe Candidate index $k \in SC$, the stripe width range is defined from bin index $st\ to\ (ed+1)$, where st and ed are the first indices extending outward from stripe candidate index k such that $S_{st-1} < CutVal_k$ and $S_{ed+1} < CutVal_k$. Here, $CutVal_k$ is derived from the NDLM smoothing distribution (Eq. 25):

$$P(a_k \leq X \leq CutVal_k \mid a_k, R_k) = \varepsilon * P(a_k \leq X \leq S_k \mid a_k, R_k) \tag{30}$$

$$\varepsilon = 0.1\ (default)$$

$$Stripe\ Width\ Range = st * resolution\ to\ (ed+1) * resolution\ (bp) \tag{31}$$

Stripe length range

Score vector SV is computed for each stripe width range to decide stripe length. First, a signal probability block is constructed around the stripe width range, where rows represent the stripe width $(ed - st + 1)$ and number of columns is $(d + 1)$(Eq. 16). The matrix values are filtered for those satisfying $dir_{ij} \in \{dir, 3\}$ $(dir = 1\ for\ 5'stripe\ calling\ and\ dir = 2\ for\ 3'stripe\ calling)$. Next, column-wise means are computed to obtain a raw score vector of length $d + 1$. Then SV is built by subtracting bs from raw score vector, where bs is derived from parameters the NDLM smoothing distribution (Eq. 25) as follows:

$$P(X \leq bs) = \delta, \quad X \sim N(\mu_{pixel}, \sigma^2_{pixel}) \qquad (32)$$

$$\delta = 0.95\ (default)$$

$$\mu_{pixel} = \frac{1}{width} \sum_{i=st}^{ed} \frac{a_i}{\sum_{j=i}^{i+d} I(dir_{ij} \in \{dir, 3\})} \qquad (33)$$

$$\sigma^2_{pixel} = \frac{1}{width^2} \sum_{i=st}^{ed} \frac{R_i}{\left(\sum_{j=i}^{i+d} I(dir_{ij} \in \{dir, 3\})\right)^2} \qquad (34)$$

$$raw\ score\ vector\ SV_j^{raw} = \frac{\sum_{k=st}^{ed} P(k,j) * I(dir_{kj} \in \{dir, 3\})}{\sum_{k=st}^{ed} I(dir_{kj} \in \{dir, 3\})} \qquad (35)$$

$$SV_j = SV_j^{raw} - bs \qquad (36)$$

$$score\ vector\ SV = (SV_1, SV_2, \ldots, SV_{d+1}) \qquad (37)$$

From the beginning of the SV, the cumulative score is updated. If the running sum drops below zero, the score is reset, and the search resumes from the next index [11]. The genomic interval with the maximum cumulative score is designated as the stripe length range.

$$stripe\ length\ range\ (head, tail) = argmax_{1 \leq s \leq e \leq (d+1)} \sum_{m=s}^{e} SV_m \qquad (38)$$

Stripe p-value

Then stripe score is averaged SV^{raw} (Eq. 35) in length range and p-value of stripe is calculated from the distribution of score. (Eq. 32)

$$stripe\ score = mean(SV^{raw}[head{:}\ tail]) \qquad (39)$$

$$pvalue_{stripe} = P(X \geq stripe\ score \mid \mu_{pixel}, \sigma^2_{pixel}) \qquad (40)$$

Stripe Calling Parameter Settings

For Dynamix, stripe calling was performed using default parameters which were automatically applied without specifying command-line arguments. The default parameters are as follows:

(resolution=10000, window_size =510,000, down_sample_ratio=0.01, smoothing_factor=150, var_prior_ndlm=10000, threshold_rowsums_em_p=0.5, peak_prominence=0.1, threshold_mean_to_peak=0.1, max_stripe_width=300,000, threshold_baseline_p=0.95, threshold_p_value_stripe=0.005)

The final Dynamix stripes (pass=True in output file) were filtered using length-dependent p-value thresholds: p< 0.005^2 for lengths $< 60kbp$ and p<0.005 for lengths $\geq 60kbp$.

For Quagga [11], the command line was executed using the following parameters:

(norm=balanced, threshold=0.15, resolution=10000, max_range=2,000,000,

min_length=50,000, min_distance=2,000,000, max_width=150,000, window_size=7, centromere_file=None, nstrata_blank=5, sigma=1, rel_height=0.3, gabor_freq=0.1)

The final stripes were obtained by filtering with p-value < 0.05.

For Stripenn [12], the parameters are set to (-m 0.95,0.96,0.97,0.98,0.99), and the final stripes were obtained from the filtered results (result_filtered.tsv).

Stripe annotation

Stripes were annotated as Enhancer-Promoter stripes if their chromatin states—ChromHMM [23] with active regulators (TssA, TssAFlnk, TxFlnk, EnhG, Enh, BivFlnk, EnhBiv) [24] —overlapped with either the stripe-width range (anchor) or the stripe-length range. Stripe width overlaps used a $\pm$50 kbp tolerance window, whereas length overlaps used no extension.

Stripes were categorized into High ($\geq$ 4), Medium ($3-4$), and Low or None (< 3 or absent) CTCF classes based on the maximum q-value of CTCF ChIP-seq peaks within their width ranges (anchors) allowing a $\pm$50 kbp tolerance window.

Stripe visualization on Signal Probability matrix

$P(i,j)$ (Eq. 12) are displayed with higher values in deeper red in signal probability matrix, while pixels within the stripes are highlighted in blue (Fig. 5d-f, Fig. 8a-c). Because the matrix is symmetric to the main diagonal (gray dashed line), the stripes are displayed exclusively in the upper triangular (i < j).

ChIP-seq signal analysis

To evaluate ChIP-seq signal pileups around anchors, a $\pm$200 kbp window was centered at the midpoint of the anchor for all stripes (Fig. 4). The signals at each relative position were averaged across stripes. For the ChIP-seq signal analysis within stripe length ranges, the y-axis values in Fig. 5a-c represent the median of the mean ChIP-seq signal in the length range. Fig. 7e-g displays the corresponding mean values.

CRISPRi-FlowFISH data and Refseq genes

In Fig. 8d, CRISPRi-FlowFISH enhancer-promoter positions in K562 cell within the genomic region (chrX:47,000,000–49,500,000) were obtained from Supplementary Table 3a of Fulco et al. [20] for entries where the 'Significant' column is True. Promoters and enhancers are color-coded based on the 'Element name' column. RefSeq genes in the figure were annotated using the Bioconductor packages GenomicRanges, TxDb.Hsapiens.UCSC.hg38.knownGene and org.Hs.eg with gene symbols mapped from Entrez Gene IDs.

CTCF motif orientation analysis

To analyze CTCF motif pileups around stripe anchors, a $\pm$1,500 kbp window centered at the midpoint of each width range was established (Fig. 9a-c) and divided into 10 kbp bins. For each bin, the numbers of plus and minus strand CTCF motifs—identified by the FIMO [25]—were separately summed and averaged in total stripes.

Stripe Anchor Venn Diagram

Intersections among the tools were evaluated based on the stripe width ranges (anchors) with a $\pm$2 kbp tolerance window (Fig. 9d).

Discussion

Operating on a bias-removed signal probability matrix and applying a Bayesian Normal Dynamic Linear Model (NDLM) [16] to it, Dynamix achieves high sensitivity in capturing both short- and long-range interacting stripes enriched for enhancer-promoter interactions, unlike existing tools.

Dynamix achieves computational efficiency by constructing the signal probability matrix using a ZINB-NB mixture EM model in an average of 10 minutes per chromosome, and identifies high-probability stripe regions using a Bayesian NDLM—a novel implementation for stripe-calling tools—in an average of 2 minutes per chromosome.

Dynamix identifies the highest number of EP stripes with accurate localization, supported by biological evidence. Compared to alternative tools, both the anchors and length ranges of Dynamix stripes exhibit the highest mean ChIP-seq signals for active E-P markers, even though their median length is more than twice that of Quagga.

Notably, Dynamix effectively captures long-range interacting stripes. While over 90% of stripes by other tools originate within a diagonal distance less than 600 kb, only 56% of Dynamix stripes fall into this short-range, with the rest distributed across diverse distances. Dyanmix long-range EP stripes exhibit high ChIP-seq signals of histone markers and uniquely capture experimentally validated long-distance promoter-enhancers, such as those at the *HDAC6* locus, which align closely with K562 CRISPRi-FlowFISH data [20].

The versatility in width, length, and diagonal distance of Dynamix stripes stems from the Bayesian NDLM which utilizes genomic index-specific parameters to identify stripe boundaries. Notably, its median length exceeds the average size of CTCF or E-P loops by approximately 150 kb, reflecting the accumulation of diverse, cell-to-cell loop variations within cell populations [7].

A limitation of this study is that Dynamix was not evaluated on Micro-C or single-cell Hi-C datasets. Applying the tool to these high-resolution and single-cell platforms to further analyze cell-type specific chromatin stripes is a key direction for future work.

Conclusions

By integrating a two-step statistical framework, Dynamix provides a robust computational tool for advanced comparative studies in 3D genomics—such as contrasting chromatin stripes between healthy and diseased cell types and investigating aberrant loop formations or chromatin re-wiring that dysregulte disease-associated genes.

List of abbreviations

ZINB: Zero-inflated negative bionomial

NB: Negative bionomial

NDLM: Normal Dynamic Linear Model

EP: Enhancer-promoter

CTCF: CCCTC-binding factor

TAD: topologically associating domain

Data Availiability

Hi-C data of the human B-lymphoblastoid cell line GM12878 was obtained from the 4D Nucleome (https://data.4dnucleome.org), and the mappability file was acquired from the UCSC Genome Browser. Dynamix stripes were identified using these datasets, and Quagga [11] and Stripenn [12] stripes were also generated using the same Hi-C data. For analyses, ChIP-seq datasets for histone modifications and transcription factors (CTCF, H3K27ac, H3K4me3, and H3K4me2), along with the ChromHMM [23] file for E-P annotation, were acquired from the ENCODE. Chromosome FASTA files were downloaded from the UCSC Genome Browser, and the CTCF motif meme file was retrieved from the JASPAR. Utilized files and accession numbers are provided in Supplementary Table 2.

Software availability

The source code and the documentation for the Dynamix R package are freely available on Github at https://github.com/soominhan511/Dynamix.

Funding

The author received no specific funding for this work.

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The author declares no competing interests.

Consent for publication

Not applicable.

References

1. Hua P, Badat M, Hanssen LLP, Hentges LD, Crump N, Downes DJ, et al. Defining genome architecture at base-pair resolution. Nature. 2021;595:125–9. https://doi.org/10.1038/s41586-021-03639-4.

2. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. Formation of Chromosomal Domains by Loop Extrusion. Cell Rep. 2016;15:2038–49. https://doi.org/10.1016/j.celrep.2016.04.085.

3. Vian L, Pękowska A, Rao SSP, Kieffer-Kwon K-R, Jung S, Baranello L, et al. The

Energetics and Physiological Impact of Cohesin Extrusion. Cell. 2018;173:1165-1178.e20. https://doi.org/10.1016/j.cell.2018.03.072.

4. Banigan EJ, Van Den Berg AA, Brandão HB, Marko JF, Mirny LA. Chromosome organization by one-sided and two-sided loop extrusion. eLife. 2020;9:e53558. https://doi.org/10.7554/eLife.53558.

5. Sridhar A, Farr SE, Portella G, Schlick T, Orozco M, Collepardo-Guevara R. Emergence of chromatin hierarchical loops from protein disorder and nucleosome asymmetry. Proc Natl Acad Sci USA. 2020;117:7216–24. https://doi.org/10.1073/pnas.1910044117.

6. Davidson IF, Bauer B, Goetz D, Tang W, Wutz G, Peters J-M. DNA loop extrusion by human cohesin. Science. 2019;366:1338–45. https://doi.org/10.1126/science.aaz3418.

7. Barrington C, Georgopoulou D, Pezic D, Varsally W, Herrero J, Hadjur S. Enhancer accessibility and CTCF occupancy underlie asymmetric TAD architecture and cell type specific genome topology. Nat Commun. 2019;10:2908. https://doi.org/10.1038/s41467-019-10725-9.

8. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015;161:1012–25. https://doi.org/10.1016/j.cell.2015.04.004.

9. Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science. 2016;351:1454–8. https://doi.org/10.1126/science.aad9024.

10. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet. 2011;43:1059–65. https://doi.org/10.1038/ng.947.

11. Feng F, Moran S, Hansen AS, Zhang X, Liu J. Statistically rigorous and computationally efficient chromatin stripe detection with Quagga. Genome Res. 2026;36:159–68. https://doi.org/10.1101/gr.280132.124.

12. Yoon S, Chandra A, Vahedi G. Stripenn detects architectural stripes from chromatin conformation data using computer vision. Nat Commun. 2022;13:1602. https://doi.org/10.1038/s41467-022-29258-9.

13. Osuntoki IG, Harrison A, Dai H, Bao Y, Zabet NR. ZipHiC: a novel Bayesian framework to identify enriched interactions and experimental biases in Hi-C data. Bioinformatics. 2022;38:3523–31. https://doi.org/10.1093/bioinformatics/btac387.

14. Xu Z, Zhang G, Wu C, Li Y, Hu M. FastHiC: a fast and accurate algorithm to detect long-range chromosomal interactions from Hi-C data. Bioinformatics. 2016;32:2692–5. https://doi.org/10.1093/bioinformatics/btw240.

15. Sahin M, Wong W, Zhan Y, Van Deynze K, Koche R, Leslie CS. HiC-DC+ enables systematic 3D interaction calls and differential analysis for Hi-C and HiChIP. Nat Commun. 2021;12:3366. https://doi.org/10.1038/s41467-021-23749-x.

16. Prado R, Ferreira MAR, West M. Time Series: Modeling, Computation, and Inference. 2nd edition. Boca Raton: Chapman and Hall/CRC; 2021. https://doi.org/10.1201/9781351259422.

17. Hsieh T-HS, Cattoglio C, Slobodyanyuk E, Hansen AS, Darzacq X, Tjian R. Enhancer-promoter interactions and transcription are largely maintained upon acute loss of CTCF, cohesin, WAPL or YY1. Nat Genet. 2022;54:1919–32. https://doi.org/10.1038/s41588-022-01223-8.

18. Calo E, Wysocka J. Modification of enhancer chromatin: what, how, and why? Mol Cell. 2013;49:825–37. https://doi.org/10.1016/j.molcel.2013.01.038.

19. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489:109–13. https://doi.org/10.1038/nature11279.

20. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nat Genet. 2019;51:1664–9. https://doi.org/10.1038/s41588-019-0538-0.

21. Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically associating domains. Sci Adv. 2019;5:eaaw1668. https://doi.org/10.1126/sciadv.aaw1668.

22. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr. 1974;19:716–23. https://doi.org/10.1109/TAC.1974.1100705.

23. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9:215–6. https://doi.org/10.1038/nmeth.1906.

24. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30. https://doi.org/10.1038/nature14248.

25. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27:1017–8. https://doi.org/10.1093/bioinformatics/btr064.

Supplementary Files

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

- SupplementaryNote1.pdf
- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx