Published online DD MM YYYY

Nucleic Acids Research, YYYY, Vol. xx, No. xx ***1–4***
doi:10.1093/nar/gkn000

LocalFinder: detecting local correlation and enrichment between epigenomic tracks

Pengfei Yin[1†], Puxuan Sun[1†], Yongwen Ding[1], Yanqiang Fu[1], and Jiankang Wang[1,2,*]

[1]School of Biomedical Sciences, Hunan University, Changsha 410082, China
[2]Shenzhen Research Institute, Hunan University, Shenzhen 518000, China

Received YYYY-MM-DD; Revised YYYY-MM-DD; Accepted YYYY-MM-DD

ABSTRACT

Local relationships between genomic tracks often encode key regulatory information, yet most existing methods focus on either local enrichment or local correlation, leaving their joint signal underused. Here we present LocalFinder, a tool that computes enrichment significance (ES) and harmonic-mean–weighted local correlation (HMC) profiles from a pair of genomic tracks. The ES profile sensitively identifies significantly different regions (SDRs), while HMC quantifies local concordance and further refines SDRs into biologically interpretable classes, such as totally vs partially depleted or reprogrammed vs pre-programmed regions. Benchmarking on synthetic and real datasets shows that (1) jointly leveraging local enrichment and local correlation consistently outperforms single-feature approaches (e.g., MACS2, enrichment only; StereoGene, correlation only) in distinguishing totally depleted, partially depleted, and unchanged regions; and (2) LocalFinder's ES+HMC provides the best two-feature integration, outperforming all three cross-tool pairings (ES+StereoGene, MACS2+StereoGene, and MACS2+HMC). LocalFinder is broadly applicable to diverse pairs of genomic tracks, whether they originate from distinct assays (e.g., ChIP-seq, ATAC-seq, Hi-C–derived tracks) or from bulk and single-cell datasets. It scales to large data collections, enabling systematic discovery of local regulatory changes and biological insights, and is freely available online.

INTRODUCTION

Genome-wide scale relationships between epigenomic features and transcription are informative—for example, H3K4me3 is broadly enriched at promoters of actively transcribed genes, linking a chromatin mark to gene activity (Cell 2007, PMID: 17512414). However, such global relationships can provide limited insight into regulatory mechanisms at specific loci, because a high overall correlation can mask sharp, functionally meaningful local differences. For instance, Maurano et al. reported a very high overall similarity between DNase-seq profiles in fetal heart and fetal kidney (R = 0.99), yet individual local regions exhibit markedly different accessibility, consistent with tissue-specific regulatory element usage (Science 2012, PMID: 22955828).

Importantly, once analysis shifts from global to local relationships, distinct local patterns often reflect distinct regulatory mechanisms. During glucocorticoid signaling, John et al. compared DNase-seq with GR ChIP-seq and identified pre-programmed sites where GR binds predominantly within pre-existing accessible chromatin, versus reprogrammed sites where GR binding coincides with newly induced accessibility, consistent with stimulus-dependent chromatin remodeling. These patterns carry different biological implications: pre-programmed binding highlights a dominant role of the baseline chromatin landscape in directing de novo factor occupancy, helping explain why the same glucocorticoid machinery can produce tissue-specific responses across cell types with different accessibility profiles; in contrast, reprogrammed sites represent a smaller set of loci where chromatin remodeling can create new potential occupancy sites, providing a mechanism for incremental, directional regulatory change and a form of cellular memory across differentiation (Nature Genetics 2011, PMID: 21258342). Likewise, under acute CTCF depletion, local comparisons between CTCF ChIP-seq and Hi-C–derived insulation/loop features reveal heterogeneous sensitivity: some loci become totally depleted of CTCF with pronounced loss of local insulation, whereas others are only partially depleted and retain weaker (but detectable) insulation. This distinction is biologically informative because insulation defects scale with the degree of CTCF loss—CTCF acts potently and dose-dependently in maintaining TAD insulation—so partially depleted loci can preserve boundary function, whereas near-complete depletion is required to trigger the most substantial folding defects (Cell 2017, PMID: 28525758).

Various tools have been developed to study genomic relationships, each focused on different types of analysis. Global correlation tools like DeepTools (NAR 2014, PMID: 24799436; NAR 2016, PMID: 27079975) and BEDTools (Bioinformatics 2010, PMID: 20110278) are used to analyze relationships across the entire genome, especially for continuous data (e.g., ChIP-seq and RNA-seq) or interval data, by calculating how genomic features co-vary across all positions. DeepTools offers a range of methods for calculating global correlations, with Pearson's correlation being the most commonly used method to assess relationships between tracks, such as histone modifications

and transcriptional activity. It also provides options for Spearman's rank correlation and distance-based methods to capture non-linear relationships when required. BEDTools, on the other hand, calculates global correlation by evaluating the overlap or co-occurrence of interval-based features across genomic tracks. It typically uses simple overlap analysis or Fisher's exact test to measure the statistical significance of these overlaps. StereoGene (Bioinformatics 2017, PMID: 29028265) is another global correlation tool that rapidly estimates correlations between pairs of genomic features using kernel-based correlation. It provides an efficient way to compute the correlation of continuous genomic profiles across the genome, offering a smoothed estimate of correlation using a Gaussian kernel. For local correlation, StereoGene also computes the local correlation (LC) of genomic features as a function of genomic position, providing a detailed view of how features co-vary at a finer scale. This allows researchers to analyze how relationships between features vary at specific loci, enabling insights into spatially resolved regulatory mechanisms. DeepTools is also useful for local correlation by computing correlation scores within sliding windows across the genome. For local enrichment, ChIPSeeker (Bioinformatics, 2015, PMID: 25765347), Homer (Mol Cell 2010, PMID: 20513432), and MACS2 (Nature Protocol 2012, PMCID: PMC3868217) identify enriched genomic regions for specific epigenomic marks like H3K27ac and H3K4me3, using statistical models to detect significant peaks, while DROMPA (Methods 2020, PMID: 32240773) and DeepTools focus on identifying regions with high chromatin feature densities, using background correction to account for genomic biases.

When calculating local enrichment and local correlation between genomic tracks, a key challenge is that regions with low values in both tracks can lead to random fluctuations, making the results unreliable. For example, in local enrichment tools like DROMPA, the ratio of the two tracks is used to calculate enrichment (Methods 2020, PMID: 32240773). However, the ratio of two small values can fluctuate significantly, introducing instability. Additionally, tools like and DROMPA assume that genomic regions follow a Poisson or Negative Binomial distribution, using background correction models to account for genomic noise. However, this approach has limitations, particularly when dealing with an abundance of zero values across the genome, which can undermine the reliability of peak calling in low-coverage regions. This is where the Zero-Inflated Negative Binomial (ZINB) distribution may offer improvements, as it can better model the excess zeros and handle low-coverage data more robustly (Genome Biology 2011, PMID: 21787385). MACS2 circumvents this issue by focusing on high-coverage candidate regions, but even these regions can contain valuable regulatory information. Furthermore, MACS2's null hypothesis tests for enrichment but lacks the ability to differentiate between de novo peaks and strengthened peaks, which can have distinct biological significance. Strengthened peaks typically represent regions with pre-programmed regulatory features, where the signal is consistently higher and correlates with established regulatory mechanisms. In contrast, de novo peaks may indicate reprogrammed binding regions, marking novel regulatory changes or cell type-specific events (Nature Genetics 2011, PMID: 21258342). Additionally, methods like TOBIS (NC2020, PMID: 32848148) use a floor value to filter out low-coverage regions, ensuring that unreliable data does not influence the results. For local correlation, StereoGene's LC approach addresses fluctuations in low-coverage regions by using coverage as a weight in the correlation calculation, reducing the impact of unreliable data and ensuring more accurate correlations in higher-coverage regions (Bioinformatics 2017, PMID: 29028265).

Current tools primarily focus on either local enrichment or local correlation to study the relationships between two genomic tracks. Here, we introduce LocalFinder, which leverages both local enrichment and local correlation. Unlike methods that focus solely on high-coverage regions, LocalFinder is capable of producing reliable results even in low-coverage regions by incorporating Zero-Inflated Negative Binomial (ZINB) modeling and a floor value for the enrichment significance (ES) track, and by applying a harmonic mean as a weight for local correlation to generate the HMC track. Benchmarking on synthetic and real datasets demonstrates that jointly using both features consistently outperforms single-feature approaches, with LocalFinder's ES and HMC further enhanced by these strategies. Additionally, LocalFinder excels at distinguishing between re-programmed and pre-programmed regions, as well as between partially depleted and totally depleted regions, which have distinct biological implications. Further applications highlight LocalFinder's broad applicability and its ability to systematically uncover local regulatory changes, providing valuable biological insights.

(1).

Text. Text. Text. Text. Text. Text. (2, 3).

MATERIALS AND METHODS

LocalFinder

LocalFinder calculates the harmonic mean correlation (HMC) and enrichment significance (ES) between two genomic tracks, and can optionally identify significant differential regions (SDRs).

Binned tracks Input tracks may be provided in BigWig, BedGraph, or BAM format. All tracks were first converted to BedGraph, and subsequently binned into equal-width bins (default bin size = 200 bp).

Normalization of two tracks For each bin i, we record two raw read-coverage values $x_{k,i}$ for both tracks ($k \in \{1,2\}$, $i \in \{1,\ldots,N\}$), and define the total coverage for track k as $T_k = \sum_{j=1}^{N} x_{k,j}$, where N is the total number of bins. LocalFinder supports three alternative normalization methods:

(i) **RPKM** — reads per kilobase per million

$$\tilde{x}_{k,i} = 10^6 \frac{x_{k,i}}{T_k L/1000}, \text{where } N \text{ is bin number.}$$

(ii) **CPM** — counts per million

$$\tilde{x}_{k,i} = 10^6 \frac{x_{k,i}}{T_k},$$

(iii) **SCALE** — downscale to the smaller library

$$\tilde{x}_{k,i} = x_{k,i} \frac{\min(T_1, T_2)}{T_k},$$

For simplicity of notation, we will henceforth denote the normalized values $\tilde{x}_{k,i}$ simply as $x_{k,i}$.

Local ES in a sliding window

1) Floor-correction of low-coverage bins Noise in the local ES arises from small-value ratios. To reduce noise, low-coverage bins are replaced with the p-th percentile (default $p = 90\%$):

$$x'_{k,i} = \max\big(x_{k,i}, \tau_k\big), \quad \text{where} \quad \tau_k = \text{Percentile}_{90}\big(\{x_{k,i}\}_{i=1}^{N}\big).$$

2) Window statistics Let $P = 2h_p + 1$ denote the peak size, a parameter that can be adjusted for different peak widths, and $W = 2h + 1$ denote the window size.
Peak-window means (for logFC):

$$\tilde{\mu}_{k,i} = \frac{1}{P} \sum_{j=i-h_p}^{i+h_p} x'_{k,j}.$$

Sliding-window means, variances, and dispersions (for SE): To estimate the standard error of the log fold-change in the absence of biological replicates, we approximate local sampling variability with a Negative Binomial (NB) variance model and estimate dispersion by method of moments, in the style of `edgeR` for count data. Here, the empirical variance $\sigma^2 k, i$ is computed within the local window, and the dispersion $\phi k, i$ quantifies overdispersion relative to a Poisson model (for which $\phi_{k,i} = 0$). Estimates are truncated at zero to enforce non-negativity:

$$\mu_{k,i} = \frac{1}{W} \sum_{j=i-h}^{i+h} x'_{k,j},$$

$$\sigma^2_{k,i} = \frac{1}{W} \sum_{j=i-h}^{i+h} \big(x'_{k,j} - \mu_{k,i}\big)^2,$$

$$\phi_{k,i} = \max\Big(\frac{\sigma^2_{k,i} - \mu_{k,i}}{\mu^2_{k,i}}, 0\Big).$$

8

3) Wald test for enrichment With a fold-change threshold $\text{FC}_{\text{thresh}} = 1.5$, we form a Wald statistic for each bin. The standard error of the log fold-change, SE_i, is obtained via the delta method under a NB variance model estimated locally within the sliding window. In practice, SE_i reflects two sources of variability: (i) a Poisson-like sampling component that decreases as the local mean increases, and (ii) an overdispersion component capturing extra-Poisson variability; the dispersion is constrained to be non-negative. When a percentile floor is applied, the floored window means are used in these calculations. This mirrors the spirit of `edgeR`'s Wald framework, but here the variance components and dispersion are estimated per-window to assess enrichment at each genomic bin.

8

$$\text{logFC}_i = \frac{\ln\big(\tilde{\mu}_{2,i} / \tilde{\mu}_{1,i}\big)}{\ln(\text{FC}_{\text{thresh}})},$$

$$\text{SE}_i = \sqrt{\frac{1}{\mu_{1,i}} + \frac{1}{\mu_{2,i}} + \phi_{1,i} + \phi_{2,i}},$$

$$Z_i = \frac{\text{logFC}_i}{\text{SE}_i},$$

$$p_i = 2(1 - \Phi(|Z_i|)).$$

where Φ is the standard normal cumulative distribution function (CDF).

8

4) ES track

$$\boxed{\text{ES}_i = \text{sgn}(\text{logFC}_i)\big(-\log_{10} p_i\big)}$$

Positive ES_i marks enrichment of track 2 over track 1, while negative ES_i marks depletion.

Local HMC in a sliding window

1) Local correlation The local correlation within the sliding window, by default computed as the Pearson correlation coefficient, is denoted r_i,

$$r_i = \frac{\sum_{j=i-h}^{i+h} (x_{1,j} - \mu_{1,i})(x_{2,j} - \mu_{2,i})}{\sqrt{\sum_{j=i-h}^{i+h} (x_{1,j} - \mu_{1,i})^2} \sqrt{\sum_{j=i-h}^{i+h} (x_{2,j} - \mu_{2,i})^2}},$$

where $\mu_{k,i} = \frac{1}{W} \sum_{j=i-h}^{i+h} x_{k,j}$. To reduce random noise, these correlations are averaged across the window $\bar{r}_i = \frac{1}{W} \sum_{j=i-h}^{i+h} r_j$.

2) Harmonic-mean weight Noise in the local HMC arises from small-value series. To reduce noise, local correlations are weighted by hormonic-mean of bin coverage.

$$w_i = \frac{\mu_{1,i}\mu_{2,i}}{\mu_{1,i}+\mu_{2,i}+\varepsilon}, \quad \text{where} \quad \varepsilon = 10^{-9}.$$

3) HMC The raw HMC is computed as $\mathrm{HMC}_i^* = \bar{r}_i w_i$. Let $Q_{0.9995}$ denote the 99.95$^{\text{th}}$ percentile of the set $\{\mathrm{HMC}_i^*\}_{i=1}^N$. Each score is then clipped to this cap and linearly rescaled:

$$\boxed{\mathrm{HMC}_i = \frac{\min\left(\mathrm{HMC}_i^*, Q_{0.999}\right)}{Q_{0.999}} \in [0,1]}$$

SDRs of two tracks SDRs are called from the signed significance track $\mathrm{ES}_i = \mathrm{sgn}(\mathrm{logFC}_i)(-\log_{10} p_i)$ using a user-defined p-value threshold. Significant bins of the same sign are merged into regions using a gap-tolerant procedure similar to peak merging in `MACS2`, with a default minimum run length ($m = 2$) requiring at least m consecutive bins above the significance threshold, and a default maximum gap size ($g = 0$). For each SDR, the HMC is calculated over significant bins only. Regions are ranked by their mean HMC: lower values correspond to nearly depleted or reprogrammed regions, whereas higher values indicate partially depleted or pre-programmed regions.

Speed and memory optimization To efficiently compute `HMC` and `ES` on large genomic datasets, LocalFinder adopts a chromosome-by-chromosome, multi-threaded strategy that minimizes memory footprint and reduces redundant computation. The core optimizations include:

(i) **Chromosome-wise parallelism**: Each chromosome is processed independently in separate threads, avoiding the need to load the entire genome into memory at once and enabling parallel execution.

(ii) **Vectorized operations with prefix sums**: Bin counts are converted to `NumPy` arrays, and prefix-sum arrays are precomputed for values, squares, and cross-products. This allows $O(1)$ retrieval of sliding-window means and variances, avoiding repeated summations in loops.

These optimizations collectively reduce both runtime and memory usage, enabling LocalFinder to process high-resolution genomic tracks efficiently on multi-core systems.

ACKNOWLEDGEMENTS

Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.

Conflict of interest statement. None declared.

REFERENCES

1. Author,A.B. and Author,C. (1992) Article title. *Abbreviated Journal Name*, **5**, 300–330.
2. Author,D., Author,E.F. and Author,G. (1995) *Book Title*. Publisher Name, Publisher Address.
3. Author,H. and Author,I. (2005) Chapter title. In Editor,A. and Editor,B. (eds), *Book Title*, Publisher Name, Publisher Address, pp. 60–80.
4. Author,Y. and Author,Z. (2002) Article title. *Abbreviated Journal Name*, **53**, 500–520.