VSS-Hi-C: Variance-stabilized signals for chromatin 3D contacts

Faezeh Bayat and Maxwell Libbrecht*
Department of Computing Science, Simon Fraser University, Burnaby BC, Canada

Abstract: Chromatin conformation capture assays output a value for each pair of positions that quantifies their strength of interaction in the nucleus. Hi-C signals lack the property of variance stabilization. That is, a difference between 0 and 200 reads usually has a very different statistical importance from a difference between 2000 and 2200 reads. This deficiency impedes analysis and visualization of Hi-C data.

Here, we propose an approach called VSS-Hi-C that normalizes Hi-C data to produce variance-stabilized signals. It does so by learning the empirical relationship between the mean and variance of Hi-C values and transforming these values such mean and variance of the data are no longer dependent to each other. We show that VSS-Hi-C outperforms the other transformation approaches in stabilizing the variance of the data and using VSS-Hi-C transformed signals improves downstream analysis such as identifying chromosomal subcompartments. Using VSS-Hi-C additionally eliminates the need for complex methods such as negative binomial modeling for downstream analysis and enables expressive visualization of Hi-C data.

1 Introduction

The 3D conformation of the genome plays a central role in many cellular functions including gene expression and DNA replication [1, 2, 3, 4]. 3D interactions occur as a result of biological functions including promoter-enhancer interactions, polymer looping, nuclear compartmentalization, phase separation and others [5, 6, 7, 1, 8]. Chromatin conformation capture assays such as Hi-C measure chromosome conformation by quantifying the interactions between pairs of loci in the genome. These sequencing-based assays output a read count for each pair of genomic loci, indicating the loci's strength of interaction.

Like all biological data, Hi-C contains sources of experimental variability that must be mitigated and normalized to enable accurate analysis. Many approaches have been developed that aim to normalize for sources noise and bias in Hi-C data including mappability, restriction site density, G/C content, read depth, 1D distance, random polymer looping and others [9, 10, 11, 12, 13].

One source of experimental variability in Hi-C data is not addressed by existing normalization methods: interaction counts have a nonuniform mean-variance relationship. For example, a contact pair having 0 reads in one sample and 200 in the other sample is usually considered a more significant difference than a pair having 2000 reads in one and 2200 in the other. More generally, lower interaction counts tend to have lower variability than higher interaction counts. This nonuniform mean-variance relationship poses a challenge to downstream analyses. For instance, considering the difference in interaction counts between different samples is a poor measure of the difference in interaction strength.

To address this nonuniform mean-variance relationship in Hi-C data, some methods, such as those used for detecting significant chromatin interactions, employ techniques like negative binomial distribution modeling [14, 15, 16, 17]. However, due to the complexity of optimizing and implementing negative binomial models, many approaches employ Gaussian-based models and implicitly assume a uniform mean-variance relationship. This issue also afflicts any analysis that employs mean-squared error (MSE) metric for quantifying performance because this metric is equivalent to likelihood of a uniform-variance Gaussian model. Examples of Gaussian- or MSE-based models include identifying topologically associating domains (TAD) [18], identifying genome compartments or subcompartments [19, 20], enhancing Hi-C data resolution [21], detecting loops [22, 23] and identifying promoter-enhancer interactions [24].

To attempt to mitigate the nonuniform mean-variance relationship, many of the Gaussian-based methods employ transformations like $\log(\log(x+c)$ for a constant c (usually 1)) or inverse hyperbolic sine transformations ($\text{asinh}(x) = \log(x + \sqrt{x^2+1})$). These transformations stabilize the mean-variance relationship under specific assumptions (Methods), but we found that these assumptions are violated, resulting in nonuniform mean-variance relationship (Results). However, these transformations assume a specific relationship between the mean and variance of the data (Methods).

Accounting for mean-variance relationship is also crucial for visualization. Large outliers inherent to nonuniform variance often dominate the viewing scale of plots. This problem is of particular importance to Hi-C data, which is typically visualized as a heatmap, where the choice of color scale can radically change a plot's interpretation (Fig 1a). More generally, Euclidean distance in a 2D plot corresponds to the log likelihood of difference in a uniform-variance Gaussian model. Although these problems can be partially

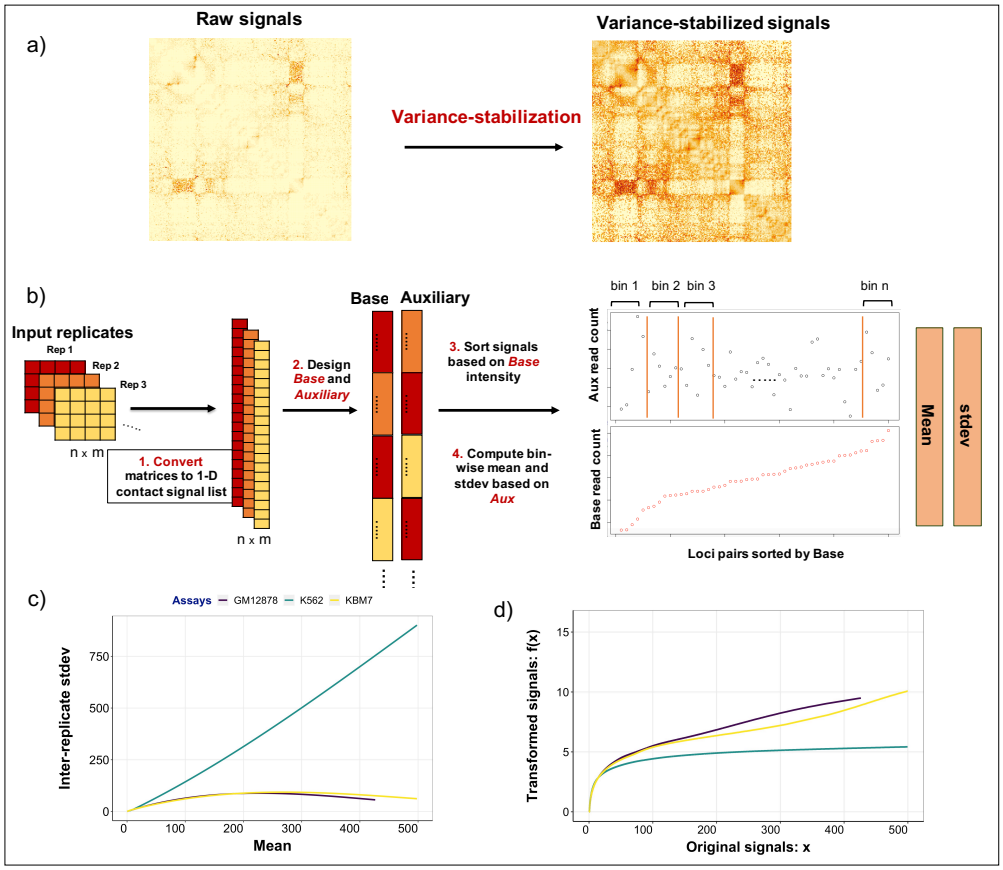

Figure 1: **General schematic of the VSS-Hi-C method.** (a) Intensity heatmap of Hi-C matrix before and after variance stabilization. Hi-C data is GM1278 observed/expected contact matrix for chr21:250000-480000 (hg19) at 100 kb resolution. Color indicates signal strength; signal strength was mapped uniformly to the yellow-red color palette such that yellow and red correspond to minimum and maximum contact strength respectively. In raw signals, due to a small number of large outliers, only a small number of pixels receive dark red colors. (b) VSS-Hi-C uses two or more replicates to learn the empirical mean-variance relationship of the input data (Methods). (c) Learned mean-variance relationships for several data sets. Horizontal and vertical axes denote mean and standard deviation respectively. Note that the mean-variance relationship differs across data sets, indicating that each requires a different transformation. (d) Learned transformation functions. Horizontal and vertical axes indicate input and output values respectively.

mitigated by carefully choosing a maximum viewing range, doing so corresponds to a crude crude linear+flat transformation.

For many other data sets such as RNA-seq, researchers precisely mitigate this mean-variance relationship using a variance-stabilizing transformation (Methods). We recently demonstrated the importance of doing so for 1D genomic data sets like ChIP-seq and developed a tool called VSS for this purpose [25].

Here, we extend our previous work and propose a method called VSS-Hi-C that stabilizes the variance of Hi-C contact strength. This method learns the empirical mean-variance relationship of the Hi-C matrices and transforms the Hi-C contact strength using a transformation based on this learned mean-variance relationship. We show that VSS-Hi-C transformed matrices have a fully stabilized mean-variance relationship, in contrast to other transformation methods. Moreover, we illustrate that variance-stabilized signals are beneficial for downstream analyses like identifying topological domains and subcompartments.

2 Methods

2.1 Hi-C data. We acquired three in-situ Hi-C data sets for GM12878, K562 and KBM7 cell lines generated by Aiden-Lieberman group [1] (GEO accession number GSE63525. Also available at ENCODE consortium with ENCODE accession numbers ENCSR968KAY, ENCSR545YBD and ENCSR987EPR, respectively.). The mentioned Hi-C interactions were mapped to the hg19 reference genome In this study, we use two measures of contact strength [1, 26]: (1) The "observed" contact matrix, which represents the raw read counts of the interactions between different genomic regions, and (2) The "observed over expected" (O/E) contact matrix, defined by dividing each observed signal by the average for pairs at the given 1D distance. This step accounts for the increased number of interactions between nearby genomic loci. We considered both 10kb and 100kb resolution contact matrices in the evaluations in this study.

Following previous work [1], normalization of Hi-C contact matrices are performed using the Knight and Ruiz [27] matrix-balancing algorithm in all the analyses performed in this study. After extracting the raw and O/E matrices, we used the Juicer tool [28] for extracting the KR-normalized contact matrices for Hi-C data. KR normalization approach makes sure that sums of each row and column of the contact matrices are equal, which accounts for sources of bias such as G/C content and mappability.

2.2 Repli-seq and ChIP-seq data. We acquired 6 phase Repli-seq data from the ENCODE consortium (encodeproject.org) from GM12878 and K562 cell lines with ENCODE accession numbers ENCSR218XP and ENCSR591OX respectively.

Moreover, we obtained ten different histone modification ChIP-seq data (H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me3, H4K20me1) from GM12878 and K562 cell lines from Roadmap Epigenomics data portal (`https://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidated/macs2signal/foldChange/`). We also acquired ChIP-seq data targeting H2A.Z and DNase from GM12878 and K562 cell lines. These ChIP-seq data sets quantified according to fold enrichment, defined as the ratio of observed read count over that of an Input control.

2.3 Identifying the mean-variance relationship. Our variance-stabilizing transformation depends on determining the mean–variance relationship for the input Hi-C datasets. We identify this relationship using multiple replicates of the same experiment. We do so for each pair of chromosomes separately.

We first convert the Hi-C matrices to contact signal list vectors. That is, for a Hi-C contact matrix $X_{p\times q}$, we obtain a contact signal list x of length pq. Consider there are M distinct replicates for an experiment. We define two vectors, $x^{(\text{base})}$ and $x^{(\text{aux})}$ that capture replicated contact signal lists. Specifically, for each distinct pair of 1-D $x^{(i)}, x^{(j)}$ where $i \neq j$, we concatenate $x^{(i)}$ to $x^{(\text{base})}$ and $x^{(j)}$ to $x^{(\text{aux})}$. Thus, $x^{(\text{base})}$ and $x^{(\text{aux})}$ are each list of length $(\frac{N}{r})^2 M(M-1)$, for M replicates and 1-D Hi-C contact signal list of length $(\frac{N}{r})^2$ in a genome with N genomic positions and Hi-C matrix resolution r. Base-aux pairs $\left(x_i^{(\text{base})}, x_i^{(\text{aux})}\right)$ represent every possible pair of replicated Hi-C contact signal. Consider $x_i^{(\text{base})}$ and $x_i^{(\text{aux})}$ as observed contact signal at position i. For a given 1-D Hi-C contact signal list, our model imagines that there is an unknown distribution of contact strength in every position i. Let the mean $\mu_i = \text{mean}(x_i)$ for a given x_i contact signal. We further assume that there is a $\text{var}(x_i) = \sigma(\mu_i)^2$ relationship between mean and variance of these distributions. We are interested in identifying the $\sigma(\mu)$ mean-variance relationship. For estimating the $\sigma(\mu)$ relationship, note that x_i is an unbiased estimate of μ_i, and that $(x_i^{(\text{base})} - x_i^{(\text{aux})})^2$ is an unbiased estimate of $\sigma(\mu_i)^2$.

For identifying the mean-variance relationship, we employ the approach that we introduced in our previous study [25]. Briefly, we first sort the contact intensities $i \in \{1 \ldots (\frac{N}{r})^2 M(M-1)\}$ by the value of $x_i^{(\text{base})}$. Then, we define a binning approach for dividing the contact signal intensities into equally-spaced bins each containing b contact intensity. For each bin j, we compute μ_j and σ_j^2 as $\mu_j = 1/b \sum_{i \in I_j} x_i^{(\text{aux})}$ and $\sigma_j^2 = 1/b \sum_{i \in I_j} (x_i^{(\text{aux})} - \mu_j)^2$ where $I_j \subseteq \{1 \ldots (\frac{N}{r})^2 M(M-1)\}$ are the set of contact tracks in bin j. We employ the weighted average of the bins to increase the robustness of these estimates by defining

$$\bar{\sigma}_j^2 = \frac{\sum_{i=j-w}^{j+w} 2^{-b|j-w|/\beta} \sigma_i^2}{\sum_{i=j-w}^{j+w} 2^{-b|j-w|/\beta}}. \tag{1}$$

That is, for a bin j, we consider the weighted average of $2w+1$ nearby neighbor bins which we assign $2^{-bk/\beta}$ weight coefficient for bin $j+k$. We define the window size w as $w = -\beta \log(0.01)/b \log(2)$ which considers the bins with weight at least 0.01. Here, we define β as the bandwidth parameter which controls the effectiveness of the distant bins. Meaning that lower values of β would force the smoothing process to concentrate on smaller number of nearby bins while higher number of β lets the more distant bins to have contribution on the bin j's weighted average.

To account for the bias-variance trade-off, we have considered multiple combination of bin number, b, and bandwidth, β. Meaning that, larger values for this parameters would include more observations by considering more distant bins as well for calculating $\bar{\sigma}_j$ which leads to higher variance. On the other hand, lower values for b and β would concentrate on smaller number of adjacent bins which leads to more similar position in I_j set and consequently, less variance among the bins.

To fit a curve to the identified mean-variance trend, we used a smoothing spline function implemented in R which implements a regularized regression over the natural spline basis.

Also, to identify the optimum number for parameters b and β, we have done a hyperparameter search (Figure 3).

2.4 Calculating variance-stabilized signals. Variance-stabilized contact strength signals can be computed using

$$t(x) = \int_0^x \frac{1}{\hat{\sigma}(u)} du, \tag{2}$$

for a learned mean-variance relationship, where x is an untransformed contact strength signal and $\hat{\sigma}(u)$ is the learned standard deviation for a contact signal with mean u.

This transformation stabilizes the variance among the contact signal matrices, meaning that for each pair of loci i, $\mathrm{var}(t(x_i))$ is constant.

2.5 Validation of experiments. In the experiments that are performed on inter-chromosomal contact matrices, VSS-Hi-C models are trained on contact matrices between chromosomes 3 and 4 and tested on contact matrices between chromosomes 1 and 2. For the subcompartment experiment analysis, VSS-Hi-C models are trained on inter-chromosomal contact matrices between chromosomes 7 and 8 and tested on the odd-even matrices. In the analyses that require intra-chromosomal contact matrices, VSS-Hi-C models are trained on chromosome 2 intra-chromosomal contact matrices and tested on chromosome 1 intra-chromosomal contact matrices.

2.6 Alternative transformations. To stabilize the variance of data, existing approaches mainly employ a log or arcsinh transformation. These transformation stabilize the variance where there is a certain mean-variance relationship in the data [29]. Specifically, $\log(x)$ is variance-stabilizing when $\sigma(\mu) = s\mu$ for some constant s, and $\mathrm{arcsinh}(x)$ is variance-stabilizing when $\sigma(\mu) = s\sqrt{\mu^2+1}$ [30]. In addition to log or arcsinh transformation for stabilizing the variance, there is another approach called Haar-Fisz transformation [31] used by [9] to stabilize the variance of Hi-C contact profiles. Briefly, assume that vector $v = (v_0, v_1, ..., v_{N-1})$ for $v_i \geq 0$ where $N = 2^J$ needs to be variance stabilized. Haar-Fisz algorithm works as follows: It first takes the Haar Discrete Wavelet Transformation of the vector v by defining f_i^j as:

$$f_i^j = \begin{cases} 0, & \text{if } s_i^j = 0 \\ \frac{d_i^j}{\sqrt{s_i^j}}, & \text{otherwise.} \end{cases}$$

where $d_i^j = (s_{2i}^{j-1} - s_{2i+1}^{j-1})/2$ and $s_i^j = (s_{2i}^{j-1} - s_{2i+1}^{j-1})/2$. It recursively computes f_i^js for all $i \in 0, ..., (2^{J-j}-1)$ where $J \in 1, ...J$ and sets $s_0 = v$. Finally, to get the Haar-Fisz transformation operator—λ—, it applies the inverse Haar Discrete Wavelet Transformation to the transformed vector $(s^J, F^J, ..., F^1)$ to extract the vector u where $u = \lambda v$ [31]. We used "hft" funtion in R to apply haar-fisz transformation on Hi-C contact profile.

2.7 Variance instability evaluation. Following our previous work [32], we employ a variance-instability metric to evaluate whether a transformation achieves a uniform mean-variance relationship. Consider $t(x_i^{(\text{base})})$ and $t(x_i^{(\text{aux})})$ as transformed contact strength signals at the ith position in 1D contact signal list. We first sort the signals by increasing values of $t(x_i^{(\text{base})})$. Then, we divide all positions in track profile into B bins each containing $b = 10000$ signal values. We define the variance-instability metric as

$$\text{variance-instability}(t) = \frac{1}{\sigma_1^2 \sigma_2^2} \text{var}(v_{1:B}), \tag{3}$$

where $v_j = \sum_{i \in \text{bin } j} \left(t(x_i^{(\text{base})}) - t(x_i^{(\text{aux})}) \right)^2$ is the mean squared difference between replicates for positions in bin j, and σ_1 and σ_2 are the standard deviation of $t(x^{(\text{base})})$ and $t(x^{(\text{aux})})$ respectively.

The $\frac{1}{\sigma_1^2 \sigma_2^2}$ normalization factor considers for variance of the transformed contact strength signals so that for any constant α, $t(x)$ and $\alpha t(x)$ have the same variance-instability value. Large values for the variance-instability metric indicates that variance of the signals are unstable.

2.8 Evaluation through clustering interchromosomal Hi-C contacts to assign labels to subcompartments. To evaluate the utility of VSS-Hi-C-transformed signals, we evaluated how useful they are for defining subcompartments. Genomic regions can be categorized in 3 main clusters including active, inactive centromer-proximal and inactive centromer-distal [13]. Further, another study [1] shows that genomic regions can be segregated into at least six distinct subcompartments which are associated with different patterns of histone marks.

We used the method of Rao et al. [1] to identify subcompartments from an input Hi-C contact matrix. Briefly, to assign different labels to distinct subcompartments in the Hi-C matrix, we separated odd and even chromosomes to eliminate the intra-chromosomal contact profiles. We constructed a 100kb resolution odd-even matrix which consists of inter-chromosomal contact matrices, such that loci from even chromosomes appear on the columns of the matrix and rows of the matrix represents the loci from the odd chromosomes. To cluster different loci based on contact patterns, we employed the GaussianHMM clustering algorithm on this odd-even matrix. For extracting clusters for loci in the odd chromosomes, we considered rows of the matrix as samples and columns of the matrix as features and then performed the clustering algorithm. Following this step, for identifying the clusters for loci in the even chromosomes, we transposed the matrix and did the same procedure. The parameter k defines the number of clusters that contact patterns can be categorized in. Rao et al. [1] found that k-means and hierarchical clustering approaches had the same results as GaussianHMM algorithm.

Having assigned distinct labels to subcompartments for different genomic loci, we adopted a variance-explained metric from [33] to quantify the signal variance that is explainable by predicted labels. Let B be the resolution of the odd-even Inter-chromosomal Hi-C matrix. We first use the binning approach to divide genomic signals into equally-spaced bins with length B and calculate the average of signals for each bin. Let $l_i \in \{1..k\}$ (k distinct labels) be the assigned subcompartment label to each genomic loci i. Also, consider s_i as the average of signals in bin i. For each distinct label l, we compute the mean of the signals with the same label indicated by μ_l and for each genomic loci i, we define predicted signal vector as $s_i^p = \mu_{l_i}$.

We defined the variance-explained metric [33] for any given transformation as

$$\text{variance-explained}(t) = \frac{var(s_i) - var(s_i - s_i^p)}{var(s_i)} \tag{4}$$

which represents the difference between total variance and the residuals of the prediction variance. Higher variance-explained value (bounded by [0,1]) indicates more agreement between signals and predicted subcompartment labels.

In this study, we have constructed a 100kb resolution inter-chromosomal odd-even matrix which has chromosomes 1,3 and 5 in the rows of the matrix and chromosomes 2,4 and 6 in the columns of the matrix. We also applied k-means clustering algorithm for k number of clusters in which $k \in \{3, 4, 5, 6, 7\}$.

2.9 Evaluation relative to topologically-associating domain (TAD) enrichment for structural proteins and histone marks. Identifying self-interacting genomic regions known as topologically-associating

domains (TADs) is of great importance, as their rearrangement or disruption can cause disease by affecting the expression of the adjacent genes [1, 34, 35, 36]. Following previous work [36] on evaluating the TAD callers, we used fold change metric to evaluate if TAD boundaries are more enriched in ChIP-seq peaks than their adjacent flanking positions. To evaluate the effectiveness of the variance-stabilized signals, we followed TAD evaluation paper [36] to evaluate how useful they are in identification of TAD boundaries. Zufferey et al. [36] start assessing the peak enrichment by dividing the genomic regions into equally-spaced bins and computing the average number of peaks in each bin. They calculated fold change as $FC = (\text{peak}/\text{background}) - 1$, where "peak" is the average number of peaks in a surrounding region around the TAD boundary (10kb radius from the boundary) and "background" is the average number of peaks in 5kb intervals in two regions of length 100kb which are located 400kb apart from TAD boundary.

For the sake of consistency, we derived the proteins and corresponding ChIP-seq peak files from [36]. We acquired peak files on GM12878 cell line for chromatin insulator protein CTCF and two core subunits of cohesin complex including SMC3 and RAD21 from ENCODE consortium (encodeproject.org). For deriving a peak set for CTCF, intersection of four experiments with encode accession numbers ENCSR000DRZ, ENCSR000DKV, ENCSR000DZN and ENCSR000AKB were used. For RAD21, intersection of peaks from experiments ENCSR000BMY and ENCSR000EAC were used. Also, for SMC3, peaks were derived from experiment ENCSR000DZP. Moreover, we acquired peak files on K562 cell line for CTCF, SMC3 and RAD21 from experiments ENCFF002CEL, ENCFF483CZB and ENCFF002CXU, respectively.

Zufferey et al. [36] compared different approaches for identifying TADs considering multiple criteria including robustness to resolution and normalization, cost-effectiveness, concordance with other TAD callers, enrichment for biological features and computational efficiency. Considering all metrics, TopDom TAD-caller [37] has been able to satisfy all criteria. In this study, we used TopDom R package [38] to identify the TADs in Hi-C contact matrices. Briefly, TopDom method falls in the category of TAD-callers that uses linear score for each bin that summarizes distribution of contacts in Hi-C contact profile introduced by the stated bin. The only parameter used in TopDom is the window size that selects different bin size for binSignals for identifying TADs. We used the recommended window size $w = 5$ for identifying the TADs.

In addition to critical role of cohesin bindings or CTCF in TAD formation, it is also important to see if TADs are enriched for any specific histone marks. Studies have shown that TADs are either enriched for H3K27me3 marks or H3K36me3 marks. To evaluate the utility of VSS-Hi-C transformed matrices in enrichment of TADs for any of the mentioned histone marks, we followed the methodology introduced by [36]. Briefly, Zufferey et al. [36] uses fold change over control ChIP-seq signals for H3K27me3 and H3K36me3 marks to quantify the ratio of the signals between these two histone marks. Consider m as the average size of the TADs identified by a TAD caller. They divide the ChIP-seq signals into equally-spaced intervals which each interval has the length of $0.1m$. Then, for each interval, they calculate LR, the log10 ratio between H3K27me3 and H3K36me3 signals. Next, for quantifying if a identified TAD is significant, for each TAD, they computed the average observed LR. Then, they shuffled all the LR values for all TADs 10 times so they have a distribution of randomized average LR values derived from each TAD. Finally, they employ Benjamini-Hochberg procedure to consider the false discovery rate (FDR) from the TAD-specific p-values that were computed by comparing the within-TAD's LR with the derived distribution of the shuffled LRs. Fraction of those TADs that have BH corrected p-value smaller than 0.1 are reported as the significant TADs identified by the TAD-caller. Following [36], we acquired H3K27me3 and H3K36me3 ChIP-seq signals on GM12878 from experiments with encode accession numbers ENCSR000DRX and ENCSR000DRW. We also used H3K27me3 and H3K36me3 ChIP-seq signals on K562 from experiments with ENCODE accession numbers ENCSR000AKQ and ENCSR000AKR, respectively.

3 Results

3.1 Differences between replicates are stabilized after transformation. To evaluate whether existing units for Hi-C contact matrices have stable variance, we identified the mean-variance trend for a number of existing data sets (Figure 1c).

We found that the variance of the Hi-C matrices has a strong dependence on the mean (Figure 1c). This means that pairs with low interaction intensity have smaller variance across replicates, in comparison to pairs with higher interactions where the variance of the data is higher (Figure 1c). Moreover, it has been shown that the mean-variance trend does not match the expected relationship assumed by the currently-used

$\log(x+1)$ and $\text{asinh}(x)$ transformations (Methods). This means that neither of these transformations fully stabilize the variance of the Hi-C contact matrices (Figure 1c). Furthermore, the mean-variance relationship differs between data sets, meaning that no single transformation can stabilize variance for all data sets.

Following our previous work on stabilizing the variance in sequencing-based genomic assays [25], we evaluated the consistency of variance of mean squared between-replicate differences using a variance-instability metric (Methods). Lower values of this metric indicates that given transformation stabilizes the variance in a given data set. We evaluated the performance of different transformation approaches on both observed and observed over expected (O/E) contact matrices and on both inter-and intra-chromosomal contacts. We found that signals transformed by VSS-Hi-C have lower variance instability (more stabilized) than the other transformation approaches in all cases (Figure 2a). While Haar-Fisz performs comparably to VSS-Hi-C on interchromosomal and O/E intrachromosomal signals, VSS-Hi-C greatly outperforms all other approaches on raw intrachromosomal signals. This results from the fact that VSS-Hi-C directly stabilizes variance, whereas other transformations use ad-hoc heuristics that attempt to do so.

3.2 Transformed signals improve subcompartment calling. To evaluate whether variance-stabilized Hi-C matrices can improve subcompartment calling, we quantified the agreement between predicted subcompartments and genome regulatory activities and replication timing.

Replication time is defined as the order within S phase that different segments of DNA are replicated. It has been shown that there is a correlation between replication timing and chromosome structure [39]. That is, for example, inactive chromatin domains replicate later than the active domains [40]. Moreover, the correlation between genome compartmentalization and regulatory activities such as histone modification enrichment have been shown in [1, 20, 33, 41].

Following [33], to evaluate the agreement between genome subcompartments called using transformed Hi-C data with replication time and histone modification respectively, we used variance-explained metric (See section 2.8). Briefly, we called subcompartments using k-means clustering, following [1], and evaluated the degree to which these subcompartments agree with 1D data sets (Methods).

We found that signals transformed by VSS-Hi-C have better agreement between Repli-seq data and predicted subcompartments for odd-even matrix in O/E contact matrices (SI Figure 4). Moreover, we have shown that VSS-Hi-C performs better than the other transformation methods in observed matrices when $k = 6$, while for the other number of labels, all transformation approaches perform similarly. There is very small agreement between Repli-seq signals and predicted labels when using untransformed signals (Figure 2b).

Similarly, when comparing subcompartments to 1D epigenomic data, VSS-Hi-C outperforms other approaches in O/E contact matrices (SI Figure 5), although it does not improve the agreement between 1D signals and predicted labels significantly in the observed contact matrices except for $k = 7$ (Figure 2c).

To evaluate the utility of transformed Hi-C data for identifying topologically-associating domains (TADs), we applied two evaluation metrics defined in [36] (Section 2.9). The first metric evaluates whether identified TAD boundaries are enriched for the known structural proteins SMC3 and RAD21 (Section 2.9). We found that VSS-Hi-C outperforms log and Haar-Fisz transformations according to this metric, although it is outperformed by asinh (Figure 2e).

The second metric evaluates whether identified TADs are uniformly activated or repressed as measured by the ratio of the transcription-associated histone mark H3K36me3 to the repression-associated mark H3K27me3 (Section 2.9). We found that all signals perform similarly in this evaluation (Figure 2d).

These results indicate that the choice of signals has a large impact on TAD calling, although variance-stabilized signals are not clearly superior according to these metrics. Notably, the TAD caller in question, TopDom, was originally designed to be used with untransformed data; it is likely that a TAD caller designed to take advantage of variance-stabilized data could achieve even better performance.

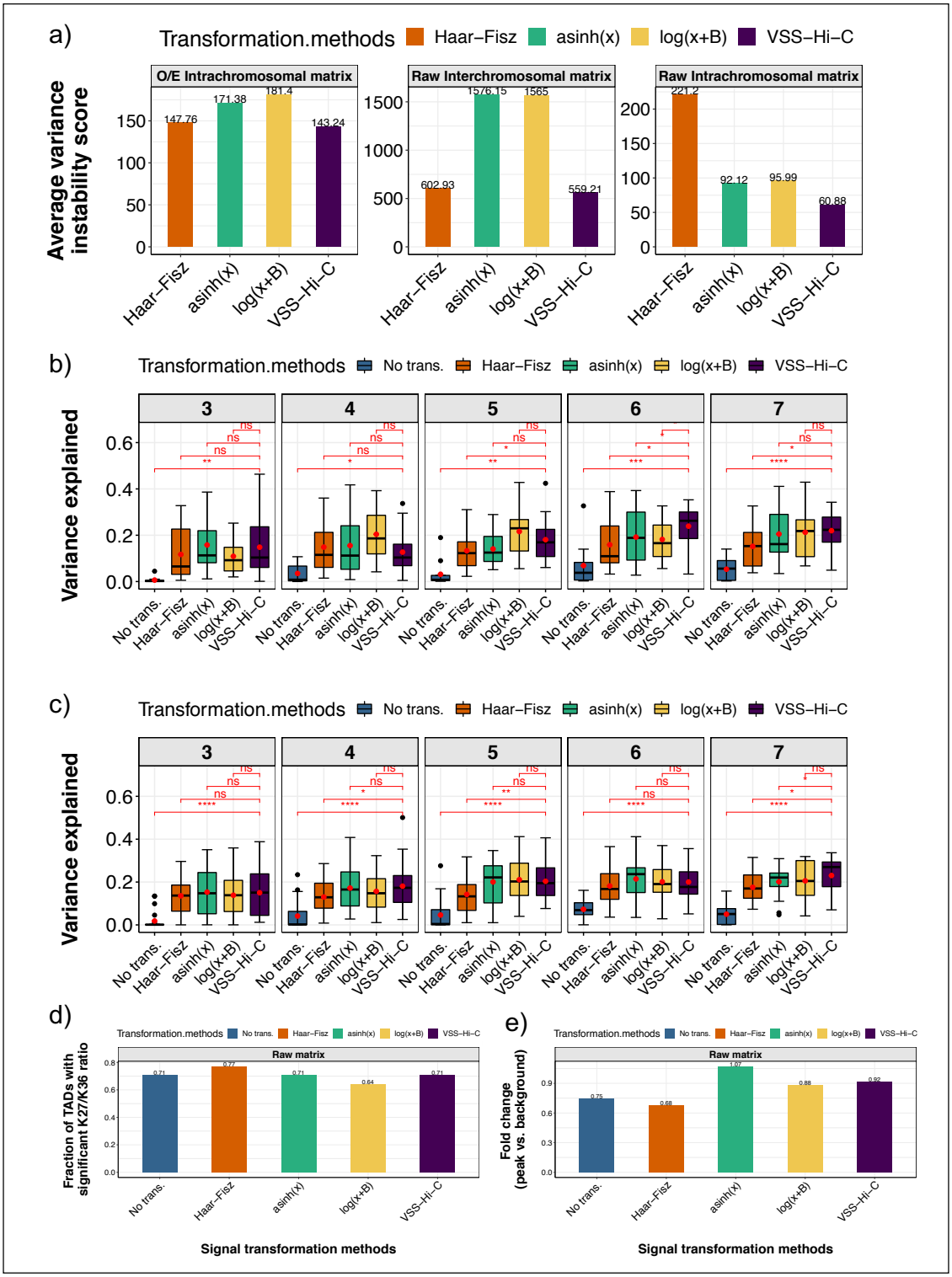

Figure 2: a) Average variance instability of transformed signals. Variance instability value for untransformed signals has been omitted as their value would distort the vertical axis. Signal variance-explained in subcompartment callers in b) Repli-seq data c) ChIP-seq data in observed matrices. Topologically associated domain enrichment in d) histone marks e) structural proteins (Methods).

References

[1] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[2] Thomas Cremer and Christoph Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews genetics*, 2(4):292–301, 2001.

[3] Wendy A Bickmore. The spatial organization of the human genome. *Annual review of genomics and human genetics*, 14:67–84, 2013.

[4] Tom Sexton, Heiko Schober, Peter Fraser, and Susan M Gasser. Gene regulation through nuclear organization. *Nature structural & molecular biology*, 14(11):1049–1055, 2007.

[5] Erica M. Hildebrand and Job Dekker. Mechanisms and functions of chromosome compartmentalization. *Trends in Biochemical Sciences*, 45(5):385–396, 2020.

[6] Ferhat Ay, Timothy L Bailey, and William Stafford Noble. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome research*, 24(6):999–1011, 2014.

[7] Julian Banerji, Sandro Rusconi, and Walter Schaffner. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2):299–308, 1981.

[8] Miklos Gaszner and Gary Felsenfeld. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature Reviews Genetics*, 7(9):703–713, 2006.

[9] Yoli Shavit et al. Combining a wavelet change point and the bayes factor for analysing chromosomal interaction data. *Molecular BioSystems*, 10(6):1576–1585, 2014.

[10] Axel Cournac, Hervé Marie-Nelly, Martial Marbouty, Romain Koszul, and Julien Mozziconacci. Normalization of a chromosomal contact map. *BMC genomics*, 13(1):1–13, 2012.

[11] Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R Lajoie, Job Dekker, and Leonid A Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods*, 9(10):999–1003, 2012.

[12] Ming Hu, Ke Deng, Siddarth Selvaraj, Zhaohui Qin, Bing Ren, and Jun S Liu. HiCNorm: removing biases in Hi-C data via poisson regression. *Bioinformatics*, 28(23):3131–3133, 2012.

[13] Eitan Yaffe and Amos Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, 43(11):1059, 2011.

[14] Mark Carty, Lee Zamparo, Merve Sahin, Alvaro González, Raphael Pelossof, Olivier Elemento, and Christina S Leslie. An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. *Nature communications*, 8(1):1–10, 2017.

[15] Tong Liu and Zheng Wang. scHiCNorm: a software package to eliminate systematic biases in single-cell Hi-C data. *Bioinformatics*, 34(6):1046–1047, 2018.

[16] Haipeng Xing, Yingru Wu, Michael Q Zhang, and Yong Chen. Deciphering hierarchical organization of topologically associated domains through change-point testing. *BMC bioinformatics*, 22(1):1–23, 2021.

[17] Cheynna Crowley, Yuchen Yang, Yunjiang Qiu, Benxia Hu, Armen Abnousi, Jakub Lipiński, Dariusz Plewczyński, Di Wu, Hyejung Won, Bing Ren, et al. FIREcaller: Detecting frequently interacting regions from Hi-C data. *Computational and structural biotechnology journal*, 19:355–362, 2021.

[18] Wenbao Yu, Bing He, and Kai Tan. Identifying topologically associating domains and subdomains by Gaussian mixture model and proportion test. *Nature communications*, 8(1):1–9, 2017.

[19] Yuchuan Wang, Yang Zhang, Ruochi Zhang, Tom van Schaik, Liguo Zhang, Takayo Sasaki, Daniel Peric-Hupkes, Yu Chen, David M Gilbert, Bas van Steensel, et al. SPIN reveals genome-wide landscape of nuclear compartmentalization. *Genome biology*, 22(1):1–23, 2021.

[20] Kyle Xiong and Jian Ma. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nature communications*, 10(1):1–12, 2019.

[21] Yan Zhang, Lin An, Jie Xu, Bo Zhang, W Jim Zheng, Ming Hu, Jijun Tang, and Feng Yue. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nature communications*, 9(1):1–9, 2018.

[22] M Jordan Rowley, Axel Poulet, Michael H Nichols, Brianna J Bixler, Adrian L Sanborn, Elizabeth A Brouhard, Karen Hermetz, Hannah Linsenbaum, Gyorgyi Csankovszki, Erez Lieberman Aiden, et al. Analysis of Hi-C data using SIP effectively identifies loops in organisms from c. elegans to mammals. *Genome research*, 30(3):447–458, 2020.

[23] Abbas Roayaei Ardakany, Halil Tuvan Gezer, Stefano Lonardi, and Ferhat Ay. Mustache: multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation. *Genome biology*, 21(1):1–17, 2020.

[24] G Ron, Y Globerson, D Moran, and T Kaplan. Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. nat commun. 2017; 8: 2237.

[25] Faezeh Bayat and Maxwell Libbrecht. VSS: variance-stabilized signals for sequencing-based genomic signals. *Bioinformatics*, 06 2021. btab457.

[26] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.

[27] Philip A Knight and Daniel Ruiz. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, 33(3):1029–1047, 2013.

[28] Neva C Durand, Muhammad S Shamim, Ido Machol, Suhas SP Rao, Miriam H Huntley, Eric S Lander, and Erez Lieberman Aiden. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems*, 3(1):95–98, 2016.

[29] Maurice S Bartlett. The use of transformations. *Biometrics*, 3(1):39–52, 1947.

[30] George EP Box. Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335, 1953.

[31] Piotr Fryzlewicz and Guy P Nason. A Haar-Fisz algorithm for Poisson intensity estimation. *Journal of computational and graphical statistics*, 13(3):621–638, 2004.

[32] Faezeh Bayat and Maxwell W Libbrecht. VSS: Variance-stabilized units for sequencing-based genomic signals. *bioRxiv*, 2020.

[33] Neda Shokraneh Kenary. *Unsupervised annotation of regulatory domains by integrating functional genomic assays and Hi-C data*. PhD thesis, Applied Sciences: School of Computing Science, 2020.

[34] Anders S Hansen, Iryna Pustova, Claudia Cattoglio, Robert Tjian, and Xavier Darzacq. Ctcf and cohesin regulate chromatin loop stability with distinct dynamics. *Elife*, 6:e25776, 2017.

[35] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

[36] Marie Zufferey, Daniele Tavernari, Elisa Oricchio, and Giovanni Ciriello. Comparison of computational methods for the identification of topologically associating domains. *Genome biology*, 19(1):1–18, 2018.

[37] Hanjun Shin, Yi Shi, Chao Dai, Harianto Tjong, Ke Gong, Frank Alber, and Xianghong Jasmine Zhou. TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic acids research*, 44(7):e70–e70, 2016.

[38] Henrik Bengtsson, Hanjun Shin, Harris Lazaris, Gangqing Hu, and Xianghong Zhou. *R Package TopDom: An Efficient and Deterministic Method for Identifying Topological Domains in Genomes*, 2020. R package version 0.9.1-9000.

[39] Juan Carlos Rivera-Mulia and David M Gilbert. Replication timing and transcriptional control: beyond cause and effect—part iii. *Current opinion in cell biology*, 40:168–178, 2016.

[40] Stefan Schoenfelder and Peter Fraser. Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics*, 20(8):437–455, 2019.

[41] Maxwell W Libbrecht, Ferhat Ay, Michael M Hoffman, David M Gilbert, Jeffrey A Bilmes, and William Stafford Noble. Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell-type-specific expression. *Genome research*, 25(4):544–557, 2015.