ENT3C: an entropy-based similarity measure for contact matrices

Xenia Lainscsek and Leila Taher[1]

[1]Institute of Biomedical Informatics, Graz University of Technology

Abstract

Motivation:

3C-sequencing technologies have shed light on the profound importance of 3D genome organization in cellular function. The resulting contact matrices are extremely sparse and susceptible to technical- and sequence-based biases, making their comparison challenging. The development of reliable, robust and efficient methods for quantifying similarity between contact matrix is crucial for investigating variations in the 3D genome organization between different cell types or under different conditions, as well as evaluating the reproducibility of 3C-based experiments.

Results:

ENT3C is a novel method for quantifying contact matrix similarity. Specifically, ENT3C compares the entropy signals representing the complexity of the patterns along the diagonals of two contact matrices. The entropy signal itself can also be used for biologically relevant analyses.

Availability: `https://github.com/xX3N1A/ENT3C`

Hi-C | mirco-C | entropy

Correspondence: *leila.taher@tugraz.at*

Introduction

Chromosome conformation capturing (3C-based) (1) has been pivotal in the discovery of hallmarks of 3D genome organization such as compartments, compartmental domains, topologically associating domains and chromatin loops, that each play distinct roles in regulating gene expression. Specifically, Hi-C (2–4) and micro-C (5) assays ligate DNA fragments in 3D proximity to one another and use deep sequencing to infer chromatin interaction frequencies. These data are then typically summarized in a *contact matrix* whereby the genome is binned into fixed-sized loci and each entry in the matrix is the estimated interaction between a pair of loci. Contact matrices are sparse, and suffer from low signal-to-noise ratios. For example, a low value in the matrix could result from a true lack of contact between the corresponding genomic loci or simply from insufficient sampling. High matrix values may reflect frequent chromatin interactions, or merely spurious ligation products associated with the adopted chromatin fragmentation and ligation protocols (6). In addition, sequence-driven noise in contact matrices arises from nucleotide composition and mappability. Because it helps mitigate stochastic fluctuations or experimental artifacts by emphasizing consistent patterns from which biologically meaningful features like A/B compartments can be determined, contact matrices are often transformed into Pearson correlation matrices where each entry is the correlation coefficient between the corresponding row and column in the contact matrix (2–4).

Quantifying the similarity between Hi-C contact matrices is essential for understanding how the 3D genome organization differs among various cell lines or under different biological conditions. Although the scarcity of the data and the intricate nature of artifacts (7, 8) render this task difficult, the recognition of inherent geometric shapes and patterns has been facilitated by the representation of contact matrices as images and networks (9). Similar images will share geometric layouts of local self-similarity (10). Recognizing the analogy, Ardakany et al. (11) developed Selfish, a tool that relies on local structural "self-similarity" to quantify the similarity of two contact matrices. Since, on average, the chromatin interaction frequencies between two genomic loci decrease as their distance increases (2)–resulting in a higher frequency of interactions along the contact matrix diagonal–Selfish restricts its computations to submatrices defined by sliding a window along the contact matrix diagonals. Contact matrices can also be regarded as networks where genomic loci are nodes and the contact frequencies are edges. Thus, GenomeDISCO (12) uses random walks in a network representation of the matrices to smooth them across a range of levels before computing their L1 distance, and HiC-spector (13) quantifies the similarity between contact maps applying spectral graph theory. A gold standard has yet to be established (see Yardımcı et al. (8) for an extensive review).

Our method, which we call "Entropy 3C" (ENT3C), is motivated from both the image processing and network analysis perspectives (9). Like Selfish, our approach to quantifying contact matrix similarity draws upon insights from image processing. The abundant and intertwined features that characterize an image such as its texture, edges, spatial correlations, luminescence and frequency distributions are difficult to disentangle and all contribute to its overall complexity. In particular, the amount of information in an image increases with increasing complexity, which is reflected in higher information entropy. Consistently, entropy has proven valuable in image analysis, aiding in tasks like face detection, character recognition, and parameter determination for augmented reality (14). Not surprisingly, the concept of entropy has also found application in network theory (15). Building on these results, ENT3C detects local changes in the signal near the diagonal of a contact matrix based on the von Neumann information entropy and recent work concerning entropy quantification of Pearson correlation matrices (16). The von Neumann entropy is a definition of entropy used to describe quantum systems that is equivalent to the classical Shannon entropy under certain conditions. Given a contact matrix, ENT3C computes the entropy of smaller Pearson-transformed submatrices extracted along its main diagonal. We define the sim-

DRAFT

ilarity between two contact matrices as the Pearson correlation between their entropy signals. We show that our similarity score is robust to both sequencing depth and contact matrix binning resolution and can perfectly distinguish between contact matrices derived from different cell lines and those derived from biological replicates of the same cell line. We also demonstrate the potential of these entropy signals to investigate intra- and intercellular differences in 3D chromatin interactions.

Methods

Concepts of entropy form the basis for ENT3C. ENT3C compares two contact matrices by calculating an entropy signal S along their diagonals. Specifically, ENT3C computes the von Neumann entropy (17), a generalization of the classical Shannon entropy to quantum mechanical systems defined as:

$$S(\boldsymbol{\rho}) = -tr(\boldsymbol{\rho}\log\boldsymbol{\rho}) \quad \textbf{(1)}$$

where $\boldsymbol{\rho}$ is a density matrix representing a quantum mechanical system that is in a mixed state described by an ensemble of pure states with different probabilities. A density matrix is Hermitian (i.e., equal to its own complex conjugate), positive semi-definite (i.e., its eigenvalues are greater than or equal to zero), and has unit trace (i.e., its diagonal elements sum to one). A contact matrix can be transformed into a Pearson correlation matrix in which an element i,j corresponds to the Pearson correlation coefficient between the ith row and jth column of the original matrix (2). Given a Pearson correlation matrix $\mathbf{P}$ of dimension $N \times N$, $\frac{\mathbf{P}}{N}$ has been shown to satisfy all necessary conditions for a density matrix (16).

Entropy signal S**.** To obtain the entropy signal S for a contact matrix $\mathbf{A}$ of dimensions $N \times N$, ENT3C extracts submatrices $\mathbf{a}$ of dimensions $n \times n$ along its diagonal. By default, n is determined as $n = \lfloor N/c + 0.5 \rfloor$, where $\lfloor \quad \rfloor$ is the floor function and the chromosome-split c represents the number of submatrices into which the contact matrix is partitioned. Next, ENT3C takes the logarithm of each entry in a, sets nan values to the logarithm of the minimum value of $\mathbf{a}$, and transforms the resulting matrix into a Pearson correlation matrix $\mathbf{P}$ using the MATLAB function `corrcoef(a,'rows','complete')`. The total number of submatrices WN, and in turn the number of data points in S, is determined by N, n and the number of bins to the next matrix, i.e., the window shift WS: $WN = 1 + \lfloor (\frac{N-n}{WS}) \rfloor$. The analogue of the quantum mechanical density matrix is then given by $\boldsymbol{\rho} = \mathbf{P}/n$ and we can compute the entropy in the equivalent form of Eq.1 $S(\boldsymbol{\rho}) = \sum_j \lambda_j \log \lambda_j$, where λ_j is the jth eigenvalue of $\boldsymbol{\rho}$ and computed with MATLAB's `eig()` function. Values for which $\lambda \leq 0$ were ignored. Each contact matrix $\mathbf{A}$ yields a signal of entropy values $S = (s_i)$ of length c, where each element s_i represents the entropy of a submatrix $\mathbf{a_i}$.

Similarity score Q**.** ENT3C scores the similarity between two contact matrices $\mathbf{A}$ and $\mathbf{B}$ as the Pearson correlation coefficient r between the entropy signals $S_{\mathbf{A}}$ and $S_{\mathbf{B}}$. In this context, we refer to r as Q. When comparing contact matrices, bins were excluded from the calculation if they were empty in any of the matrices of interest.

We evaluated ENT3C's performance by its ability to distinguish contact matrices obtained from biological replicates of the same cell line (SCs) from those derived from different cell lines (nonSCs) as given by the average similarity scores between all considered cell lines and chromosomes, $\overline{Q}_{SC}$ and $\overline{Q}_{nonSC}$ respectively. Furthermore, we define the range $d = |\overline{Q}_{SC} - \overline{Q}_{nonSC}|$.

Hi-C and micro-C data sets. Processed Hi-C data for the A549, G401, and LNCaP cell lines with identifiers ENCSR444WCZ, ENCSR079VIJ, and ENCSR346DCU, respectively, were obtained from the ENCODE Portal (https://www.encodeproject.org/, last accessed in November, 2023). Specifically, we downloaded the *pairs* files (ENCODE4 v1.10.0 GRCh38) for each of the two available replicates. Processed micro-C data for the H1-hESC and HFFc6 cell lines with identifiers 4DNES21D8SP8 and 4DNESWST3UBH, respectively, were obtained from the 4D Nucleome Data Portal (https://data.4dnucleome.org/, last accessed in November, 2023). We downloaded *contact list-replicate (pairs)* files for the two H1-hESC and three HFFc6 replicates.

Aggregating technical and biological replicates. Technical replicate pairs files of the same biological replicates were merged to construct biological replicate contact matrices for each cell line using *pairtools*'s (v1.0.2) `merge` function (18). Similarly, all pairs files corresponding to the same cell line were merged to construct contact matrices of pooled biological replicates.

Binning and generating contact matrices. Pairs were combined into 40 kb and 5 kb binned contact matrices using cooler's (v0.8.11) (19) `cload pairs` function and recursively coarsened to lower resolutions with the `zoomify` function. Contact matrices were stored in HDF5 cooler files/containers.

Downsampling interactions. Pairs files were downsampled to contain a fixed number of interactions with *pairtool*'s (v1.0.2) (18) `sample` function. To account for variations in library size, the pairs files were downsampled to contain $30e6$ interactions, which is the minimum number of interactions still covered by the lowest sequenced sample (LNCaP BR2; Supplementary Table S1-S2).

Identifying high and low complexity regions within a contact matrix. Submatrices corresponding to the highest and lowest entropy values in S were specifically inspected in all cell lines for 40 kb binned (pooled biological replicates) contact matrices on chromosome 14 (with parameters $n = 300$, $WS = 10$ and $WN_{MAX} = 1000$). The same analysis was performed for all autosomes of HFFc6.

Quantifying contact matrix similarity with ENT3C. We demonstrated ENT3C's ability to distinguish between 11 SC

DRAFT

and 48 nonSC contact matrices by computing their average pairwise similarity scores $\overline{Q}_{SC}$ and $\overline{Q}_{nonSC}$. For this analysis, ENT3C was run with a submatrix dimension $n = 300$ and $WS = 1$, and no restriction on the maximum number of generated entropy data points WN_{max} on 40 kb contact matrices derived from pairs files that included: (1) $30e6$ interactions; and (2) all interactions.

Testing the sensitivity of ENT3C's parameters n and WS. We tested the sensitivity of ENT3C's parameters on 40 kb binned contact matrices of chromosome 14. First, we fixed $WN_{max} = \infty$, $WS = 1$ and let $n \in \{50, 263, 476, 689, 902, 1,114, 1,327, 1,540, 1,743\}$. Secondly, we fixed $n = 300$ and varied $WS \in \{1, 63, 126, 188, 251, 313, 375, 438, 500\}$. Matrices were derived from downsampled pairs files (30e6 interactions) to ensure comparability.

Benchmarking ENT3C in regard to sequencing depth. We demonstrated the stability of our similarity measure to the sparsity of the contact matrix by randomly downsampling the pairs files of each biological replicate from which 40 kb contact matrices were derived to contain a fixed number of total interactions ($5e6$, $10e6$, $20e6$, $25e6$, $30e6$, $60e6$, $12e7$, $24e7$, $4e8$, and $8e8$), provided the original pairs file contained a sufficient number of interactions (Supplementary Table S1-S2), and computing $\overline{Q}_{SC}$ and $\overline{Q}_{nonSC}$. For this analysis, the chromosome-split was fixed to $c = 7$ to account for the chromosome-dependent dimensions of the contact matrices and $WN_{MAX} = 1000$; WS was set to 1, but when needed increased until maximum number of windows evaluated was WN_{MAX}.

Benchmarking ENT3C in regard to binning resolution. We tested ENT3C's sensitivity to binning resolution by computing $\overline{Q}_{SC}$ and $\overline{Q}_{nonSC}$ for 8 different binning resolutions (5, 10, 25, 40, 50, 100, and 500 kb). Only samples sequenced sufficiently deep were analyzed for resolutions below 40 kb (Supplementary Table S2); thus, HFFc6, H1-hESC and G401 were considered at 25 kb, and only the two micro-C data sets (HFFc6 and H1-hESC) were considered below 25 kb. The pairs files of each biological replicate were aggregated into 5 kb contact matrices and coarsened with *cooler's* `zoomify` function. To obtain sufficient data points in S (WN), we set c depending on the binning resolution: $c = 150$ for 5 kb, $c = 100$ for 10 kb, $c = 25$ for 25 kb, $c = 7$ for 40 kb and 50 kb, $c = 6$ for 100 kb, $c = 3$ for 500 kb, and $c = 2$ for 1 Mb. The maximum window number was set to $WN_{MAX} = 2000$; WS was set to 1, but when needed increased until the maximum number of windows evaluated was WN_{MAX}.

Comparisons to state-of-the art methods. HiC-Spector (`run_reproducibility_v1.py`) (13), QuASARRep (20), GenomeDISCO (v1.0.0) (12), and HiCRep (v1.6.0) (13) were run using the wrapper in the `3DChromatin_ReplicateQC` github repository (version 1.12.2) and the provided default parameters file. We converted the cool files of each contact matrix into the input format required by `3DChromatin_ReplicateQC` using `HiCExplorer`'s (v3.7.2) (21) `hicConvertFormat` function with options "`-inputFormat cool -outputFormat ginteractions`" and subsequently rearranging the columns. As HiC-Spector, QuASAR and GenomeDisco implement internal normalization algorithms, we called `hicConvertFormat` with the option "`-load_raw_values`" to extract unbalanced matrices from the cooler containers. The balanced matrices from the output of the cooler `balance function` with option "`-max-iters 300`" were used as input for HiCRep. Selfish (11) (MATLAB implementation committed on Feb 15, 2019) accepts contact matrices in Rao (hic) format and its two parameters k and c are determined empirically. Parameter k is directly proportional to the blocksize of the submatrices which are slid over the diagonal and must be chosen such that the resulting submatrices enclose structures likely to be preserved between cell lines, such as TADs; c depends on the Hi-C protocol. We fixed $k = 100$ as the authors did in their original publication, since our contact matrices were of the same resolution. As Hi-C and micro-C protocols are different, we initially set $c = 10$ and iterated for $c = c - 1$ to find the largest integer where similarity score for biological replicates (SCs) reached 0.9 as suggested by the authors. All tools were run following the author's recommendations to the best of our abilities, using `3DChromatin_ReplicateQC`'s parameters file where applicable. ENT3C was run with parameters $c = 7$, $WS = 1$ unless $WN_{MAX} > 1000$. Tests were performed on contact matrices 1) containing all possible interactions, and 2) derived from pairs files downsampled to a total of $30e6$ interactions. Performance of each method was determined by their respective average similarity scores across biological replicates of the same cell line (SCs) and different cell lines (nonSCs) (all referred to as $\overline{Q}_{SC}$ and $\overline{Q}_{nonSC}$, respectively, independently of the underlying similarity measure) and their separating margin d.

Identifying high and low complexity regions within and between two contact matrices. To identify the most similar genome-wide regions between HFFc6 and H1-HESC (biological replicate 1) we fit a linear model using R's `lm` function to the entropy signals of their contact matrices at 5, 10, 25, 40 and 50 kb resolution. Then we selected the regions closest to the linear fit (below the 0.3% quantile) and extracted the Ensembl IDs of protein-coding genes with transcription start sites falling within these regions. ENT3C was run with parameters $c = 150$ for 5 kb, $c = 100$ for 10 kb, $c = 25$ for 25 kb, $c = 7$ for 40 kb and 50 kb, $c = 6$ for 100 kb, $c = 3$ for 500 kb, $c = 2$ for 1 Mb. The maximum window number was set to $WN_{MAX} = 2000$; WS was set to 1, but when needed increased until the maximum number of windows evaluated was WN_{MAX}. Annotation data were obtained from BioMart (GRCh38.p14). Functional enrichment was performed to determine enriched biological process gene ontology (22) terms using `clusterProfiler`'s (23) `compareCluster` (v.4.10) function with parameter '`qvalueCutoff=0.05`'.

DRAFT

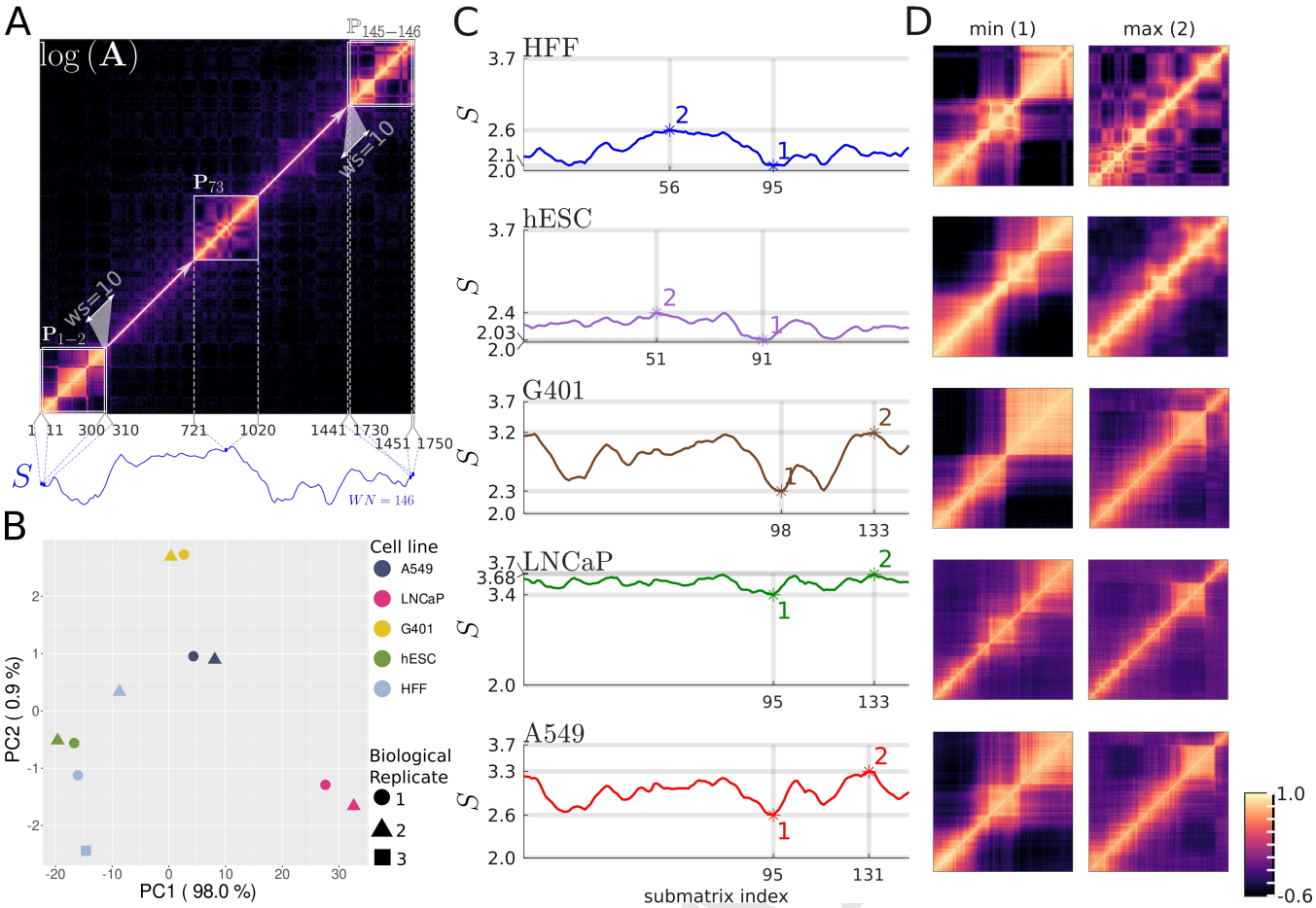

Fig. 1. ENT3C computes cell-line specific entropy signals S from the diagonal of a contact matrix and identifies low and high complexity regions.(**A**) Exemplary depiction of ENT3C derivation of the entropy signal S of the contact matrix $\mathbf{A}$ of chromosome 14 binned at 40 kb of the HFFc6 cell line (biological replicate 1). ENT3C's parameters: submatrix dimension $n = 300$, window shift $WS = 10$, maximum number of data points in S, $WN_{MAX} = \infty$, were used, resulting in 146 submatrices along the diagonal of the contact matrix. For subsequent Pearson-transformed submatrices $\mathbf{P}$ along the diagonal of $\log \mathbf{A}$, ENT3C computes the von Neumann entropies $S_i(\mathbf{P_i})$; the resulting signal S is shown in blue under the matrix. The first two ($\mathbf{P}_{1-2}$), the middle ($\mathbf{P}_{73}$), and the last two Pearson submatrices ($\mathbf{P}_{145-146}$) are shown. (**B**) PCA plot of S of each sample for chromosome 14 cluster highly along PC1 (98% variance explained). The most outlying sample is biological replicate 2 of HFFc6. (**C**) ENT3C signals S of contact matrices $\mathbf{A}$ of chromosome 14 in all five cell lines using the same ENT3C parameters as in (A). Minimum (1) and maximum (2) S values overlap in some cases. (**D**) Submatrices corresponding to minimum and maximum entropy values in (C) correspond to differences in pattern complexity. Autosome-wide analysis for HFFc6 is given in Supplementary Fig S2

Results

Variations in local entropy patterns indicate changes in 3D genome organization. Within a cell, regions in the genome characterized by complex 3D organization will be reflected by more intricate patterns in their Hi-C/micro-C contact matrices. The degree of complexity of these patterns can be quantified by entropy. Given an $N \times N$ contact matrix, ENT3C generates an entropy signal S by calculating the von Neumann entropy (17) of $n \times n$ Pearson-transformed submatrices along the diagonal, each shifted by WS (Fig. 1A, Supplementary Fig. S1). Supporting our hypothesis that variations in local entropy patterns indicate changes in cell-line specific 3D organization, S resulting from biological replicates of the same cell line tended to be more clustered along principal component 1 than those of different cell lines (Fig. 1B; Methods). To illustrate how S can capture differ ences in the complexity of different regions of a contact matrix, we identified the regions of minimum, mean and maximum entropy values $S_{min}, S_{med}, S_{max}$ of chromosome 14 for biological replicates of five different cell lines (Fig. 1C). Indeed, we found that higher entropy values reflected more intricate patterns in the corresponding contact matrices, particularly in HFFc6 (Fig. 1D, Supplementary Fig. S2). Interestingly, we identified nearly the same low complexity region for each cell line, suggesting this region's transcriptional landscape is conserved between cell lines. Moreover, HFFc6 and H1-hESC as well as G401, LNCaP and A549 displayed similar high complexity regions.

ENT3C's similarity score accurately captures contact matrix similarity. ENT3C quantifies the similarity between two contact matrices by computing the Pearson correlation coefficient between their entropy signals, to which we further refer as the *similarity score* Q. Remarkably, ENT3C effectively distinguished between SC and non-SC contact matrices (Methods) with averages of $\overline{Q}_{SC} = 0.92$ and $\overline{Q}_{nonSC} = 0.35$, respectively, and its discrimination power remained almost the same upon downsampling the interactions to account for variations in library size ($\overline{Q}_{SC} = 0.95$ and $\overline{Q}_{nonSC} = 0.35$;

Methods; Fig. 2A, Supplementary Fig. S3). By only analyzing submatrices near the diagonal, ENT3C captures signals primarily arising from dominant patterns in the contact matrix. This renders notably insensitive to sampling depth and binning resolution, as well as robust to noise. To verify robustness to the sequencing depth of a Hi-C/micro-C experiment, we applied ENT3C to contact matrices comprising different numbers of interactions (Methods; Supplementary Table S2), and found that ENT3C consistently distinguished between SC and nonSC contact matrices in all cases (Supplementary Fig. S4A). Similarly, we tested ENT3C's robustness to the binning resolution of the contact matrix for matrices generated at 8 binning resolutions ranging from 100 kbp to 5 kbp (Methods). Once more, ENT3C consistently distinguished between contact matrices of SC and nonSC in all cases (Supplementary Fig. S4B). ENT3C has only two parameters, the submatrix dimension n and the window shift WS. To show that ENT3C is only marginally dependent upon parameter choice, we computed Q_{SC} and Q_{nonSC} for various values for n and WS (Methods; Supplementary Fig. S4C-D). Although the separating margin d did decrease due to the loss of information, ENT3C's similarity measure still correctly distinguished SC from nonSCs even for extremely small matrix dimensions ($n = 50$; Supplementary Fig. S4C). For the largest submatrix dimension ($n = 1743$), S had merely $WN = 9$ data points, rendering the Pearson correlation coefficient between two signals a poor and unreliable similarity measure. When testing the window shift (WS) parameter, ENT3C failed at $WS \geq 313$, very close to the submatrix dimension chosen for the analysis ($n = 300$; Supplementary Fig. S4D). This can be explained by the omission of presumably relevant segments along the diagonal of the contact matrix. Note that a submatrix dimension of $1,743$ encompasses nearly the entire contact matrix (1,753 bins at 40 kb) and results only in 10 data points in S.

ENT3C competes well with other methods quantifying 3C contact matrix similarity. ENT3C's performance is similar to that of other state-of-the-art methods. Moreover, ENT3C shows a wider scoring range d (Table 1). To ensure comparability, we performed comparisons both on the contact matrices containing all valid interaction pairs as well as on contact matrices in which the interaction pairs file was downsampled to $30e6$ genome-wide interactions (Methods). For the intact and downsampled contact matrices, averaged over all autosomes, ENT3C, HiCRep and Selfish achieved perfect separation between SC and nonSC samples, whereby ENT3C's similarity scores displayed the highest average separation between SC and nonSC samples ($d = 0.53$) (Supplementary Fig. S5 and Table 1). A similar trend was seen when the contact matrices were derived from downsampled pairs files. ENT3C resulted in the highest average separation between SC and nonSC samples ($d = 0.60$) (Figure 2B & Supplementary Fig. S6 and Table 1). In addition, the similarity scores of ENT3C, HiCRep, Selfish and GenomeDISCO agree in many cell-line specific differences. For example, when examining the LNCaP cell line, it is evident that the differences to the HFFc6 and H1-hESC cell lines are more

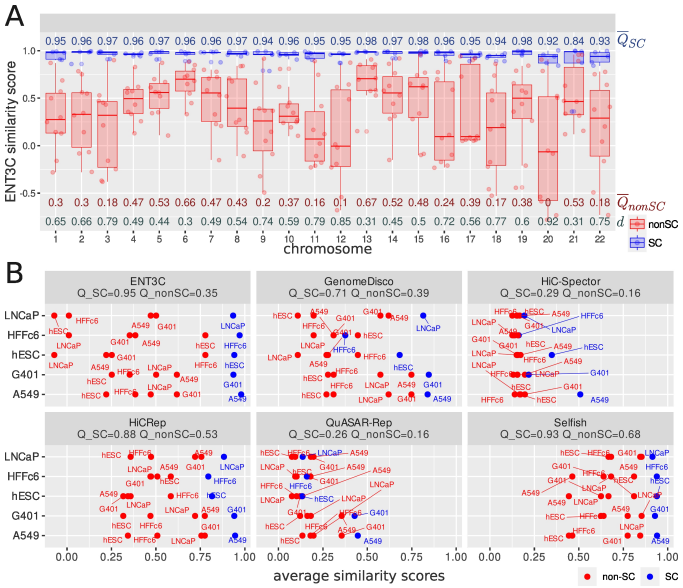

Fig. 2. ENT3C can distinguish SCs from nonSCs. (**A**) Genome-wide boxplots of ENT3C's similarity scores between SC and nonSC contact matrix. Contact matrices were generated from pairs files downsampled to contain 30e6 interactions. $\overline{Q}_{SC}$ and $\overline{Q}_{nonSC}$ for each chromosome are shown at the top and bottom, respectively. The separation range $d = |\overline{Q}_{SC} - \overline{Q}_{nonSC}|$ is shown below $\overline{Q}_{nonSC}$. (**B**) Comparison of mean similarity scores across different methods for contact matrices derived from downsampled pairs file to contain 30e6 total interactions.

pronounced than those between G401 and A549 (First rows ENT3C, HiCRep, GenomeDisco, HiCRep and Selfish in Figure 2B). Similar trends were observed for other cell types. It is noteworthy that Selfish's similarity scores remained especially robust disregardless to downsampling. Surprisingly, ENT3C's score is higher for the downsampled datasets.

ENT3C uses entropy to reveal regions of varying complexity between cells. To demonstrate how ENT3C can be used to compare the 3D chromatin organization of different cell lines, we identified regions with the most similar entropy values between the HFFc6 and H1-HESC cell-lines at resolutions between 5-50 kb (Supplementary Fig. S7-S8, Methods). Genes in these regions were associated with biological processes involved in structure and organization (nucleosome assembly, intermediate filament organization) or DNA-templated transcription initiation, which are vital for a majority of cells. Interestingly, many such processes were resolution-specific, which is consistent with distinct gene regulation mechanisms being captured at different resolutions (24).

Discussion

Most state-of-the-art methods for quantifying the similarity between two contact matrices apply dimensionality-reducing transformations aimed at modeling and reducing noise and artifacts before computing a similarity score (8). Similarity scores are calculated as the correlation of valid values from the transformed matrices (20), the weighted sum of distance-stratified Pearson correlation coefficients (13), the difference in random walks in a network representation of the matrices (8, 12, 20, 25), or by spectral decomposition of a network

DRAFT

Table 1. Comparison of ENT3C to other methods.
$\overline{Q}_{SC}, \overline{Q}_{nonSC} \ldots$ average SC and nonSC similarity scores on nonperturbed contact matrices
$\overline{Q}_{SC}(30e6), \overline{Q}_{nonSC}(30e6) \ldots$ average similarity scores on contact matrices derived from pairs files downsampled to 30e6 interactions
corresponding ranges $d = \overline{Q}_{SC}, \overline{Q}_{nonSC}$

	$\overline{Q}_{SC}$	$\overline{Q}_{SC}$ (30e6)	$\overline{Q}_{nonSC}$	$\overline{Q}_{nonSC}$ (30e6)	d	d (30e6)
ENT3C	0.92	0.95	0.34	0.35	0.59	0.60
HiC-Spector	0.32	0.29	0.17	0.16	0.15	0.13
HiCRep	0.95	0.88	0.59	0.53	0.36	0.35
QuASAR-Rep	0.70	0.26	0.48	0.16	0.22	0.10
Selfish	0.93	0.93	0.68	0.68	0.25	0.25
GenomeDisco	0.80	0.71	0.43	0.39	0.37	0.32

representation of the matrices (13), among other methods. ENT3C uses a radically different approach based on the von Neumann entropy (17). Contact matrices can be interpreted among others as images or networks, both of which have been analyzed in other contexts by various definitions of entropy. Moreover, 3D chromatin organization as well as 3C contact matrices have long been associated with fractals (e.g. (26–30)) further motivating the application of entropy measures to their analysis (31, 32). The von Neumann entropy is the quantum-mechanical analogue of the classical Shannon entropy and both agree with each other under certain conditions, but is more suitable for analyzing contact matrices because it considers both the positions and the distribution of values in a matrix. ENT3C achieved the largest average separation range between contact matrices derived from biological replicates of the same cell line (SC) and those derived from different cell lines (nonSC) when compared to other state-of-the-art methods. Although such tools were run with default settings and could perform better upon parameter optimization, this result is relevant because a similarity score with a wide separation range can not only make a binary distinction between SC and nonSC derived contact matrices, but yield a continuous representation of the 3D organization across diverse cell lines. Interestingly, we observed that method performance tended to be chromosomal-dependent, indicating that the various approaches –and potentially those that we may have inadvertently omitted given the rapidly evolving nature of this field– pick up on different signals in the contact matrices and could be combined into an even stronger similarity metric. ENT3C disregards any interactions beyond a submatrix dimension from the diagonal. While we might overlook important cell-line differences in the process, the sparsity of long range interactions in current Hi-C and micro-C datasets and gained computational efficiency, justifies this strategy. As data quality increases, ENT3C could easily be modified to take into account longer range interactions. In general, ENT3C's performance is only marginally dependent on the dimension of its input parameters, i.e., the dimension of the submatrices n and the size of the window shift. Naturally, as fewer data points are generated, a gradual decrease in the information contained in the signal becomes apparent and the similarity metric breaks down. By default, ENT3C automatically sets a suitable value of n based on the dimension of the entire contact matrix. The computationally most demanding step of ENT3C is computing the Pearson matrix and eigenvalues; nonetheless after loading the contact matrix and excluding empty bins, computing S for chromosome 1 of HFF (biological replicate 1) binned at 40 kb (1753 bins) takes merely ~ 10s using parameters ($n = 300$, $WS = 1$, $WN_{MAX} = \infty$) (MATLAB version 9.14.0.2337262 (R2023a) Update 5 on an AMD© Ryzen 9 3900×12-core processor×24; Ubuntu 20.04.6 LTS). ENT3C's high performance and computational efficiency also renders it ideal for training and testing models predictive of 3D chromatin organization. By tracking subtle differences in the complexity of local patterns along the diagonals of two contact matrices, ENT3C is not only able to distinguish SCs from nonSCs, but can also be used to investigate the 3D genome organization of one or more cell types in more detail to reveal potentially biologically significant features. The entropy values are difficult to interpret, but the upper value of the von Neumann entropy is confined by the (sub)matrix dimension, and we observed clear differences in the complexity pattern of the submatrices yielding the highest and lowest value in S. Likewise, regions exhibiting most similar entropy values between two cell lines were associated with specific biological processes which varied depending on the resolution of the contact matrices, confirming that distinct regulatory mechanisms are captured by different resolutions. Entropy has long been established beyond its original definitions in thermodynamics and statistical mechanics, across a diverse array of fields. We have demonstrated its potential in 3D genomics by applying it to the challenging problem of quantifying contact matrix similarity. ENT3C provides an efficient and robust similarity measure to assess the reproducibility of Hi-C and micro-C experiments, crucial for coming to cell-type- and condition-dependent conclusions.

DRAFT

References

1. Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, February 2002.
2. 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, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of Long-Range interactions reveals folding principles of the human genome. *Science*, October 2009.
3. Zheng Wang, Renzhi Cao, Kristen Taylor, Aaron Briley, Charles Caldwell, and Jianlin Cheng. The properties of genome conformation and spatial gene interaction and regulation networks of normal and malignant human cell types. *PLoS One*, 8(3):e58793, March 2013.
4. Ji Hyun Bak, Min Hyeok Kim, Lei Liu, and Changbong Hyeon. A unified framework for inferring the multi-scale organization of chromatin domains from Hi-C. *PLoS Comput. Biol.*, 17(3):e1008834, March 2021.
5. Nils Krietenstein, Sameer Abraham, Sergey V Venev, Nezar Abdennur, Johan Gibcus, Tsung-Han S Hsieh, Krishna Mohan Parsi, Liyan Yang, René Maehr, Leonid A Mirny, Job Dekker, and Oliver J Rando. Ultrastructural details of mammalian chromosome architecture. *Mol. Cell*, 78(3):554–565.e7, May 2020.
6. Betul Akgol Oksuz, Liyan Yang, Sameer Abraham, Sergey V Venev, Nils Krietenstein, Krishna Mohan Parsi, Hakan Ozadam, Marlies E Oomen, Ankita Nand, Hui Mao, Ryan M J

Genga, Rene Maehr, Oliver J Rando, Leonid A Mirny, Johan H Gibcus, and Job Dekker. Systematic evaluation of chromosome conformation capture assays. *Nat. Methods*, 18(9): 1046–1055, September 2021.
7. Tao Yang, Xi He, Lin An, and Qunhua Li. Methods to assess the reproducibility and similarity of Hi-C data. *Hi-C Data Analysis*, pages 17–37, 2022.
8. Galip Gürkan Yardımcı, Hakan Ozadam, Michael E G Sauria, Oana Ursu, Koon-Kiu Yan, Tao Yang, Abhijit Chakraborty, Arya Kaul, Bryan R Lajoie, Fan Song, Ye Zhan, Ferhat Ay, Mark Gerstein, Anshul Kundaje, Qunhua Li, James Taylor, Feng Yue, Job Dekker, and William S Noble. Measuring the reproducibility and quality of Hi-C data. *Genome Biol.*, 20(1):1–19, March 2019.
9. Andrea Raffo and Jonas Paulsen. The shape of chromatin: insights from computational recognition of geometric patterns in Hi-C data. *Brief. Bioinform.*, 24(5):bbad302, August 2023.
10. Eli Shechtman and Michal Irani. Matching local Self-Similarities across images and videos. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8, 17-22 June 2007, Minneapolis, MN, USA, 2007.
11. Abbas Roayaei Ardakany, Ferhat Ay, and Stefano Lonardi. Selfish: discovery of differential chromatin interactions via a self-similarity measure. *Bioinformatics*, 35(14):i145–i153, July 2019.
12. Oana Ursu, Nathan Boley, Maryna Taranova, Y X Rachel Wang, Galip Gurkan Yardimci, William Stafford Noble, and Anshul Kundaje. GenomeDISCO: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics*, 34(16):2701–2707, August 2018.
13. Koon-Kiu Yan, Galip Gürkan Yardimci, Chengfei Yan, William S Noble, and Mark Gerstein. HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps. *Bioinformatics*, 33(14):2199–2201, July 2017.
14. Amelia Carolina Sparavigna. Entropy in image analysis. *Entropy*, 21(5):502, May 2019.
15. Cristopher G S Freitas, Andre L L Aquino, Heitor S Ramos, Alejandro C Frery, and Osvaldo A Rosso. A detailed characterization of complex networks using information theory. *Sci. Rep.*, 9(1):16689, November 2019.
16. H Felippe, A Viol, D B de Araujo, M G E da Luz, F Palhano-Fontes, H Onias, E P Raposo, and G M Viswanathan. Threshold-free estimation of entropy from a pearson matrix. *EPL*, 141(3):31003, February 2023.
17. J. von Neumann. Thermodynamik quantenmechanischer gesamtheiten. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, pages 273–291, 1927.
18. Open2C, Nezar Abdennur, Geoffrey Fudenberg, Ilya M Flyamer, Aleksandra A Galitsyna, Anton Goloborodko, Maxim Imakaev, and Sergey V Venev. Pairtools: from sequencing data to chromosome contacts. *bioRxiv*, February 2023.
19. Nezar Abdennur and Leonid A Mirny. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*, 36(1):311–316, January 2020.
20. Michael E G Sauria and James Taylor. QuASAR: Quality assessment of spatial arrangement reproducibility in Hi-C data. *bioRxiv*, page 204438, 2017.
21. Fidel Ramírez, Vivek Bhardwaj, Laura Arrigoni, Kin Chung Lam, Björn A Grüning, José Villaveces, Bianca Habermann, Asifa Akhtar, and Thomas Manke. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.*, 9(1):189, January 2018.
22. Gene Ontology Consortium. The gene ontology knowledgebase in 2023. *Genetics*, 224(1), May 2023.
23. Tianzhi Wu, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, Lang Zhou, Wenli Tang, Li Zhan, Xiaocong Fu, Shanshan Liu, Xiaochen Bo, and Guangchuang Yu. clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)*, 2(3):100141, August 2021.
24. Daniel M Ibrahim and Stefan Mundlos. The role of 3D chromatin domains in gene regulation: a multi-facetted view on genome organization. *Curr. Opin. Genet. Dev.*, 61:1–8, April 2020.
25. Tao Yang, Feipeng Zhang, Galip Gürkan Yardımcı, Fan Song, Ross C Hardison, William Stafford Noble, Feng Yue, and Qunhua Li. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.*, 27(11):1939–1949, November 2017.
26. A Grosberg, Y Rabin, S Havlin, and A Neer. Crumpled globule model of the three-dimensional structure of DNA. *EPL*, 23(5):373–378, August 1993.
27. Alexander Y Grosberg. Extruding loops to make loopy globules? *Biophys. J.*, 110(10): 2133–2135, May 2016.
28. Leonid A Mirny. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.*, 19(1):37–51, January 2011.
29. James G McNally and Davide Mazza. Fractal geometry in the nucleus. *EMBO J.*, 29(1):2, January 2010.
30. Simone Pigolotti, Mogens H Jensen, Yinxiu Zhan, and Guido Tiana. Bifractal nature of chromosome contact maps. *Phys. Rev. Res.*, 2(4):043078, October 2020.
31. Patricio Venegas-Aravena, Enrique G Cordaro, and David Laroze. Natural fractals as irreversible disorder: Entropy approach from cracks in the semi Brittle-Ductile lithosphere and generalization. *Entropy*, 24(10), September 2022.
32. O. Zmeskal and P.D. Michal Vesely. Entropy of fractal systems. *Comput. Math. Appl.*, 66(2): 135–146, 2013.

DRAFT