Predicting 3D genome architecture directly from the nucleotide sequence with DNA-DDA

Xenia Lainscsek and Leila Taher[1,✉]

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

3D genome architecture is characterized by multi-scale patterns and plays an essential role in proper cellular function. Chromatin conformation capturing experiments have revealed many properties underlying 3D genome architecture. However, they are complex, costly, and time consuming, and therefore only a limited number of cell types have been examined using these techniques. Increasing effort is being directed towards deriving computational methods that can predict chromatin conformation and associated structures. Here we present DNA-DDA, a purely sequence-based method that relies on nonlinear dynamics to predict genome-wide chromatin contacts and A/B compartments. We show that the DNA sequence and A/B compartment labels of a 20Mbp-long region of the human genome is sufficient to predict the 3D architecture of the entire genome. Although this is a proof-of-concept study, our method shows promise in elucidating the mechanisms responsible for genome folding as well as modeling the impact of genetic variation on 3D genome architecture and the processes regulated thereby.

Hi-C | 3D genome organization | nonlinear dynamics

Introduction

The hierarchical three-dimensional (3D) architecture of the genome allows linearly distal loci to interact with one another, thereby impacting genome function. Chromosome conformation capturing techniques, in particular high-throughput chromosome conformation capture (Hi-C) (1, 2), have enabled us to catalog genomic interactions (*contacts*) and features of the 3D genome architecture in various cell types. Notably, Lieberman-Aiden et al. (1) reported that at the resolution of 100kbp to Mbp, chromosomes can be partitioned into A and B compartments, corresponding to eu- and heterochromatin respectively. Hi-C assays have revolutionized our understanding of the 3D architecture of the genome, but they are expensive, time-consuming and expertise-demanding. Therefore, Hi-C data are only available for a limited number of human cells, and substantial effort has been put into deriving predictive computational models.

Predictive models fall into two main categories: polymer models and statistical, particularly machine learning (ML), models. Polymer modeling applies the principles of polymer physics to understand the biophysical properties and arrangement of the DNA. Specifically, experimentally observed Hi-C contacts were early shown to coincide with predictions made by a knot-free condensed polymer model known as the "fractal globule" (1, 3). On the basis of these promising findings, more sophisticated polymer models have been developed to explain various scale-dependent folding features of chromatin, such as A/B compartments and topologically associating domains (TADs) (4–6). Polymer models provide fundamental insights into the physical and molecular underpinnings of chromatin folding, however, due to our ignorance of the details of the system, no model exists that explains all its aspects. In contrast to polymer models, ML algorithms do not attempt to model the underlying processes but rather search for and learn interdependencies between properties, such as between the DNA sequence and/or epigenetic modifications, and the genomic contacts (7). Although ML models can achieve high prediction accuracies, the logic behind their predictions is difficult to explain. Furthermore, ML-based approaches are often limited in their ability to predict long range contacts and/or do not generalize well. An extensive review of currently available tools based on both polymer and statistical modeling was undertaken by Belokopytova and Fishman (8).

While epigenomic data are abundantly available for many cell types, purely sequence-based models enable us to study the consequences that sequence variations have on genome architecture and thus the phenotype, as derangement of 3D structure has been associated with a wide range of diseases (9). Here we present DNA-DDA, a method that relies on the principles of nonlinear dynamics to predict A/B compartments from the DNA sequence alone. Nonlinear dynamical systems have fractals and chaos among their paradigms, both widespread in biological systems (10–12). The genome-wide contact maps obtained from Hi-C experiments have been shown to be (bi-)fractal (13), and their intricate self-similar patterns bear the hallmarks of chaos.

Chaos manifests itself as the seemingly random behavior of a deterministic process which is hypersensitive to extremely small fluctuations in initial conditions (14). A deterministic dynamical system can be described by its current *state* (i.e., the system variables' current values) and a system of differential equations (i.e., *rules*) which dictate how the state changes over time. A system's state space (15, 16) comprises the set of all possible states, i.e., the solutions to the system of differential equations for every possible set of initial conditions. A trajectory in the state space is the sequence of states resulting from a particular set of initial conditions. The state space of a dynamical system can be investigated temporally in regard to its dynamical properties, such as the way in which neighboring trajectories with similar initial conditions diverge over time (17), or spatially in regard to a trajectory's geometric properties, such as its fractal dimension (18). A bridge between these two perspectives is ergodic theory (19–

DRAFT

21), which studies the statistical properties of a dynamical system. Trajectories of an ergodic system will eventually cover the entire state space so that under certain conditions, the time average of a function along a trajectory is related to the space average for almost all initial conditions (15).
The relationships between the nonlinear dynamical system governing the DNA polymer and the fractal nature of Hi-C contact maps motivated us to adapt a nonlinear time series analysis technique, delay differential analysis (DDA) (22), to the prediction of A/B compartments. DDA is a nonlinear time series classification framework that has been applied most extensively in epilepsy research (23–25). It is based on embedding theory (26), which states that under certain conditions, knowledge of a single variable of the high dimensional dynamical system providing a good or a global observability of the state space, can be sufficient to obtain overall underlying information about the state space (27). We speculate that the nucleotides in the DNA sequence, are highly informative of the 3D conformation of the DNA polymer. DDA relates the time-delayed and derivative embeddings of a measurement variable in a (nonlinear) functional form. In contrast to classical ML, DDA uses sparse models with a small number of features (typically 1-4), making it robust to overfitting and easily generalizable to new data (22). DDA has recently been extended to dynamical ergodicity-DDA (DE-DDA) (28), which defines a measure for assessing dynamical similarity. We hypothesize that loci close in 3D space will share certain nonlinear dynamical properties. Therefore, inspired by the ergodic hypothesis, we examine the ensemble and time averages of dynamical information inherent in the time series representation of the DNA sequences as described by DE-DDA, to infer their proximity in 3D space.

Methods

Pre-processing of Hi-C data sequencing data. Raw FASTQ files from Hi-C experiments involving four cell lines were downloaded from the Gene Expression Omnibus database (Table 1; (2, 4, 29)) and mapped to the human reference genome (GRCh38/hg38) using `bowtie-2` (v.2.4.1; (33)) with options `-reorder` and `-very-sensitive-local`.

Hi-C contact maps. The contact matrix of each autosome was generated from the mapped reads using the *HiCExplorer* pipeline (v.3.7.2; (34–36)). `hicBuildMatrix` was called with parameters "`-binSize 5000 -minMappingQuality 10 -restrictionSequence RS -danglingSequence DS`", where `DS` and `RS` are the restriction and dangling sequences listed in Table 1. The resulting Hi-C matrix was balanced with the algorithm introduced by Knight and Ruiz (37) using `hicCorrectMatrix correct`; the "`-filterThreshold`" parameter was chosen based on the histogram produced by `hicCorrectMatrix diagnostic_plot` (Supplementary Table 1). From this matrix, a contact map at the resolution of 100 kbp was derived using `hicMergeMatrixBins` with parameter "`-numBins 20`". For comparability, the contact map obtained in this manner was scaled to the 0 to 1 range with `hicNormalize -normalize norm_range`. An entry in the resulting matrix $\mathbf{H} = (h_{i,j})$ represents the contact frequency between genomic bins i and j.

DNA-DDA. Let $x(t)$ be a dynamical sequence of length $L+1$ where t represents increments in time or space. A nonlinear DDA model has the following general functional form:

$$\dot{x} = \sum_{k=1}^{K} a_k \prod_{n=1}^{N} x_{\tau_n}^{m_{n,k}} + \rho \tag{1}$$

where $\dot{x} = \dot{x}(t)$ is the derivative of the original time (or space) series $x = x(t)$, K denotes the number of monomials, N is the number of delay embeddings $x_{\tau_n} = x(t-\tau_n)$ contained in each monomial with delay τ_n, $m_{n,k} \in \mathbf{N}_0$ is the order of nonlinearity of the nth delay embedding in the kth monomial, a are the fitting coefficients, and ρ is the least squares error of the model (Equation 3) ((28) and citations therein).
For a dynamical sequence $x(t)$, Equation 1 can be written as the overdetermined system of equations

$$\dot{\vec{x}} = \mathbf{M}\vec{\alpha} \tag{2}$$

where each element in $\dot{\vec{x}}$ is the center derivative at each time (or space) point $(\dot{x}(t), \dot{x}(t+1), \ldots, \dot{x}(t+L))$, $\mathbf{M}$ a matrix where each column represents a monomial (or delay embedding $x_{\tau_n}^{m_{n,k}}$) at each time (or space) point (row), and $\vec{\alpha} = (a_1, a_2, a_3)$ are the fitting coefficients of the model. The fitting coefficients a_i are estimated for the input data using singular value decomposition (SVD), and together with the least squares error

$$\rho = \sqrt{\frac{1}{L}\left(\dot{x} - \sum_{k=1}^{K} a_k \prod_{n=1}^{N} x_{\tau_n}^{m_{n,k}}\right)^2} \tag{3}$$

make up the classifying feature set $F = (\boldsymbol{\alpha}, \rho)$. Furthermore, Equation 1 can be solved for I dynamical input sequences $\vec{x}(t) = (x_1(t), x_2(t), \ldots, x_I(t))$ either individually (single-trial (ST) DDA; Figure 1 C i)) or simultaneously (cross trial (CT) DDA; Figure 1 C ii)) by extending the over-determined system of equations of Equation 2 to

$$\begin{pmatrix} \dot{\vec{x}}_1 \\ \dot{\vec{x}}_2 \\ \vdots \\ \dot{\vec{x}}_I \end{pmatrix} = \begin{pmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \\ \vdots \\ \mathbf{M}_I \end{pmatrix} \vec{\alpha}. \tag{4}$$

ST and CT DDA features have recently been combined by (28) in a way which allows testing for dynamical similarity between two dynamical input sequences $x_i(t)$ and $x_j(t)$. Here the mean of the ST error is representative of the temporal average, and the CT error is representative of the ensemble average. In accordance with the ergodic hypothesis, if $x_i(t)$ and $x_j(t)$ are dynamically similar, the mean of the ST errors $\left(\overline{\rho_{s_{i,j}}} = \frac{\rho_{s_i} + \rho_{s_j}}{2}\right)$ and the CT error $\left(\rho_{c_{i,j}}\right)$ will also be

DRAFT

Cell type	Hi-C data		H3K4me1 ChIP-Seq		RS	DS
GM12878	GSE63525	(2), (29)	GSE29611; GSM733772	(30)	GATC	GATC
K562	GSE63525	(2), (29)	GSE31755; GSM788085	(31)	GATC	GATC
hESC	GSE35156	(4)	GSE29611; GSM733687	(30)	AAGCTT	GATC
IMR90	GSE35156	(4)	GSE103589; GSM2775001	(32)	AAGCTT	GATC

Table 1. Data sets
RS: restriction sequence; DS: dangling sequence.

similar and therefore, their quotient close to one. Dynamical ergodicity (DE) DDA $\epsilon_{i,j}$ is defined as

$$\epsilon_{i,j} = \left| \frac{\overline{\rho_{s_{i,j}}}}{\rho_{c_{i,j}}} - 1 \right|. \quad (5)$$

The smaller $\epsilon_{i,j}$, the more dynamical similar the two signals under investigation, i and j, are.

DNA-DDA Contact maps. **Numeric representation of DNA sequences**
To convert the DNA sequence into a numeric sequence that can be used as input for DDA we chose the 1D DNA walk representation in which the walker takes a step upwards if the nucleotide is a pyrimidine and downwards if it is a purine (38). DNA walks have been used for instance to identify long-range power-law correlations, periodicities and various fractal properties of genomic sequences (12, 39, 40).

DE-DDA classifying feature set computation
The human reference genome sequence was partitioned into 100kbp non-overlapping bins, and the sequence of each bin was represented as a numeric sequence using the aforementioned 1D DNA walk. For a pair of bins i, j, we computed the ST- and CT-errors ρ_{s_i}, ρ_{s_j} and $\rho_{c_{i,j}}$ using the a `C` implementation of DDA provided by the author (28). This executable takes as input a time series and parameters listed in Supplementary Table 2 and outputs ST- and CT- classifying feature set (a_1,a_2,a_3,ρ). We combined the errors ρ_{s_i}, ρ_{s_j} and $\rho_{c_{i,j}}$ into DE-DDA $\epsilon_{i,j}$ (Eq. 5), which we propose as an estimation of the interaction probability between the two pairs of genomic bins. We repeated this process for all bin pairs to obtain the DDA matrix $\mathbf{D} = (d_{i,j})$ of a given chromosome.

Matrix Post-processing
We hypothesize that bins with higher interaction probability will be similar in certain dynamical properties. Thus, we inverted each DNA-DDA matrix $\mathbf{D}$ so that highest values are mapped to the lowest and vice versa. To enhance outlier entries in the matrix, the logarithm of each non-zero value in $\mathbf{D}$ was taken. Afterwards, zero-valued entries were set to the lowest value of the log-transformed $\mathbf{D}$. The matrices were saved in the file format of `HOMER` (41) and then converted to `h5` with *HiCExplorer*'s `hicConvertFormat` function. The resulting DNA-DDA matrices were normalized to the 0 to 1 range with `hicNormalize -normalize norm_range`.

Compartment calling. We obtained the locations of gaps in the reference genome assembly from the UCSC table browser (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/gap.txt.gz) (31) and excluded them from the Hi-C and DDA matrices before proceeding with the analysis. To call compartments, we first derived the Pearson correlation matrices $\mathbf{C_H} = (c_{\mathbf{H}_{i,j}})$ and $\mathbf{C_D} = c_{\mathbf{D}_{i,j}}$ from the normalized Hi-C matrix $\mathbf{H}$ and computationally derived DNA-DDA matrices $\mathbf{D}$ as described by (1). An entry $c_{\mathbf{H}_{i,j}}$ ($c_{\mathbf{D}_{i,j}}$) refers to the Pearson correlation coefficient between the i-th row and j-th column of $\mathbf{H}$ ($\mathbf{D}$).
We then called chromosome compartments by applying principal component analysis (PCA) using MATLAB®'s `pca` function, to the Pearson correlation matrices $\mathbf{C_H}$ and $\mathbf{C_D}$. More specifically, the sign of the ith PC classifies regions either to the A or B compartment. Bins enclosing contromere locations (obtained from the UCSC table browser (31)) and adjacent empty bins in the HiC-maps which disrupted PCA were excluded from this calculation (Supplementary Table 5). From here on, we refer to the PC obtained from $\mathbf{C_H}$ as $\text{PC}_{\text{Hi-C}}$ and to the PC obtained from $\mathbf{C_D}$ s $\text{PC}_{\text{DNA-DDA}}$. $\text{PC}_{\text{DNA-DDA}}$ was smoothed with a sliding window of 5 using MATLAB®'s `movmean()` function before continuing with the following steps. Extreme outliers in the PCs (values more than three scaled MAD from the median) were replaced with nearest element that is not an outlier using MATLAB®'s `filloutliers(PC,'nearest')` function and each PC was normalized to zero mean and unit variance.
We used ChiP-Seq data for H3K4me1, an epigenetic modification associated with open chromatin (42), to determine the PC most likely to correspond to the compartments. More specifically, among the first three principal components (PCs), we selected the one with the strongest (maximum absolute value of) Pearson correlation with H3K4me1 density (30) to be representative of A/B compartments (Supplementary Table 6). If the Pearson correlation was negative, the PC was multiplied by -1. Values above 0 were assigned to the A compartment and scaled to $[0, 0.5]$; values below 0 were assigned to the B compartment and scaled to $[0.5, 1]$. In exceptional instances, a different PC with approximately the same Pearson correlation coefficient was deemed more likely to be

DRAFT

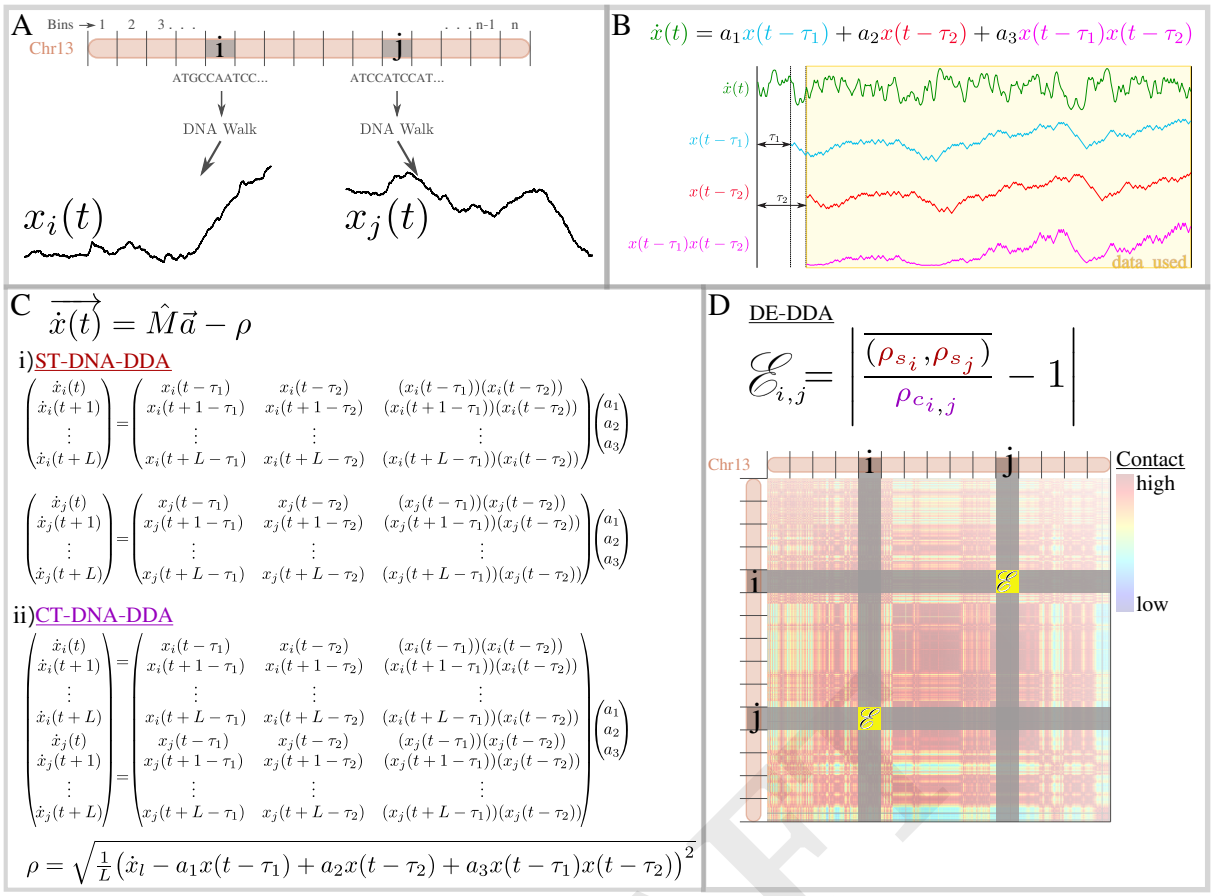

Fig. 1. Steps taken to derive the entries of a DNA-DDA contact matrix $\mathbf{D}$. A) the DNA sequences of exemplary two bins i and j are represented as random walks $x_i(t)$ and $x_j(t)$ B) data used in DDA considering the delay pair τ_1 and τ_2 and the model $\dot{x} = a_1 x_{\tau_1} + a_2 x_{\tau_2} + a_3 x_{\tau_1} x_{\tau_2}$ C) resulting over-determined system of equations for i) ST- and ii) CT-DDA and model error ρ D) DE-DDA $\epsilon_{i,j}$ is the matrix entry $\mathbf{D}_{i,j}$ and $\mathbf{D}_{j,i}$. It is defined by the quotient of the mean of the ST-DDA errors $\overline{(\rho_{s_i}, \rho_{s_j})}$ and CT-DDA errors $\rho_{c_{i,j}}$

associated with the compartments upon inspection, and used instead (Supplementary Table 6).

Structure selection and testing. A key difference to traditional ML approaches is that DDA models are not updated iteratively (i.e., they do not learn). Instead, an exhaustive sweep is performed over a list of putative models to search for those best suited to discriminate the dynamics of interest; this step is called *structure selection*.
The functional form of a DDA model is dictated by the overall system and obtained data type. The data type most extensively studied using DDA is EEG, for which a particular functional form therefor has been established (23–25, 43, 44). Typically, most terms in Equation 1 are set to zero to reduce the chances of overfitting. This work is a proof of concept to explore the feasibility of applying DDA to genomics data. Thus, we decided to use a simple, symmetric model with only quadratic degree of nonlinearity:

$$\dot{x} = a_1 x_{\tau_1} + a_2 x_{\tau_2} + a_3 x_{\tau_1} x_{\tau_2} \qquad \textbf{(6)}$$

Symmetric models require half the computational effort of non-symmetric ones to compute the classifying feature set $\{a_1, a_2, a_3, \rho\}$ for a range of delays between r_1 and r_2; $\tau_1, \tau_2 \in \{r_1, r_2\}$.

Next, in a supervised structure selection step, we searched for the delay pair τ_1, τ_2 that best captured A/B compartments in each of the four cell-types. Specifically, this was done using a 20Mbp region on chromosome 22 (chr22:16200000-36200001, Supplementary Figure 1). The performance of each delay pair was measured as the Pearson correlation coefficient between $PC_{\text{Hi-C}}$ and $PC_{\text{DNA-DDA}}$ in this region.
The four obtained DNA-DDA models were tested on all 100kbp genomic loci of all human autosomes with various performance measures (Pearson correlation coefficient r_{PC} between $PC_{\text{Hi-C}}$ and $PC_{\text{DNA-DDA}}$, area under the roc curve `AUC`, accuracy `ACC` and F1-score `F1` for classifying A/B compartments). We would like to emphasize that once we determined the DNA-DDA model for each cell type (Supplementary Table 3), all subsequent analyses and results were performed on never before seen data (with the exception of chr22:16200000-36200001).

Comparison to other methods. We compared DNA-DDA to two other methods that predict A/B compartments from sequence-based features or the sequence directly (Table 2). The features of the "Sequence-based Annotator of chromosomal compartments by Stacked Artificial Neural Networks" (SACSANN; (45)) are derived from GC content, transposable elements (TE), and predicted transcription factor bind-

DRAFT

	Method	Input	Target	Training/SS Data	Test Data	Nr. of Data Sets
SACSANN	random forest, ANN	number of TEs, TFBSs, GC content	compartments	21 chromosomes	1 chromosome	8
ABCNet	CNN	sequence	compartments	21 chromosomes	1 chromosome	28
DNA-DDA	DDA	sequence	genome-wide contacts, compartments	200 tandem 100kb bins on chr22	chromosomes 1-22	4

Table 2. Comparison of DNA-DDA to other methods
SS: structure selection; ANN: artificial neural network; CNN: convolutional neural network; Nr. of Data Sets: overall number of Hi-C data sets used to train/test each method (Supplementary Table 7).

ing sites (TFBS) within 100 kbp bins. It retains the 100 most important species/cell-type specific features based on a random forest predictor and then trains two stacked artificial neural networks (ANN) to classify input vectors either to the A or B compartment. The A/B Compartment Network (ABCNet; (46)) is a deep convolutional artificial neural network that takes a one-hot encoding of the sequence within a 100 kpb bin and extracts features by passing it through two layers convolutional kernels, an average pooling layer, and a fully connected dense layer to output a single value representing the predicted PC for each given bin.

Both SACSANN and ABCNet have been tested on numerous data sets (Table 2). We compared the performance of DNA-DDA on the GSE35156 data set to that reported by SACSANN's and ABCNet's authors on the same dataset. In addition, we compared the *overall* performance of the methods, i.e., the average AUC or ACC across various data sets. In the case of SACSANN, we considered all examined data sets ("Summary (ROC AUC score) SACSANN" in supplemental_file_S3.xls at https://github.com/BlanchetteLab/SACSANN/tree/master/supplemental_files/, last accessed in July 2022). For ABCNet we restricted the comparison to the ACC reported for human data sets (i.e., average μ across all "secondary" data sets in Table III in (46)).

Results

	Performance measure	hESC (GSE35156)	overall
SACSANN	AUC	0.81	0.83
DNA-DDA	AUC	0.82	0.81
ABCNet	ACC	0.75	0.80
DNA-DDA	ACC	0.75	0.76

Table 3. Comparison of DNA-DDA to other methods
hESC (GSE35156): average performance across all chromosomes on the GSE35156 Hi-C dataset. overall: average performance across multiple (possibly only partially) overlapping Hi-C datasets (see Methods and Supplementary Table 7)

DNA-DDA models were derived to predict A/B compartments in four different cell lines directly from the DNA sequence. DNA-DDA is a statistical approach derived from DDA, a method traditionally used on time series data based on nonlinear dynamics. DDA models relate the numerical derivatives of the input data to their time-delayed versions in a nonlinear functional form, of which the fitting weights and model error $\{a_1, a_2, a_3, \rho\}$ are used to access information about the underlying dynamical system. With simplicity and efficiency in mind, we chose the functional form given in Equation 6. In principle the analysis can be summarized in two steps: (1) determine the model parameters (delay pair τ_1, τ_2 in Equation 6) by supervised structure selection; and (2) estimate the classifying feature set $\{a_1, a_2, a_3, \rho\}$ by solving the over-determined system of equations with SVD (Figure 1) for model parameters τ_1, τ_2.

Structure selection. To find the delay pairs τ_1, τ_2 in Equation 6 that best capture the A/B compartments in each cell-type, we ran the model with all possible combinations of values for τ_1, τ_2 on a 20Mbp-long region of chromosome 22 (Methods). Briefly, we partitioned this region into 200 100kbp-long bins, estimated the interaction probability between each pair of bins i and j as $\epsilon_{i,j}$ (Equation 5), built a DNA-DDA contact map $\mathbf{D}$, and called A/B compartments using PCA to obtain $\text{PC}_{\text{DNA-DDA}}$. For each cell type, we then chose the delay pair that resulted in the highest Pearson correlation coefficient r_{PC} between $\text{PC}_{\text{DNA-DDA}}$ and $\text{PC}_{\text{Hi-C}}$ (Table 2). This resulted in four different delay pairs, one for each cell type. Pearson correlation coefficients between $\text{PC}_{\text{DNA-DDA}}$ and $\text{PC}_{\text{Hi-C}}$ ranged between 0.21 (for hESC) and 0.49 (for GM12878).

Testing. The DNA-DDA correlation matrices $\mathbf{C_D}$ exhibited strikingly similar patterns to the experimentally obtained Hi-C correlation matrices $\mathbf{C_H}$ (Figure 2 and Supplementary Figures 7 to 10). Moreover, DNA-DDA achieves exceptional performances on never before seen genome (Table 4 over all chromosomes and cell lines: $\overline{AUC} = 0.81$ $\overline{ACC} = 0.76$ $\overline{F1} = 0.73$ $\overline{r}_{PC} = 0.56$), especially considering the mere 20Mbp region that was used to determine the model/delay pair combination. The model/delay pair combination achieves the highest performance in the K562 cell line ($\overline{AUC} = 0.84$, $\overline{ACC} = 0.78$, $\overline{F1} = 0.76$, and $\overline{r}_{PC} = 0.62$), and the lowest performance in the IMR90 cell line ($\overline{AUC} = 0.75$, $\overline{ACC} = 0.72$, $\overline{F1} = 0.68$, and $\overline{r}_{PC} = 0.46$).

In general, supervised methods are limited by the uncertainty

DRAFT

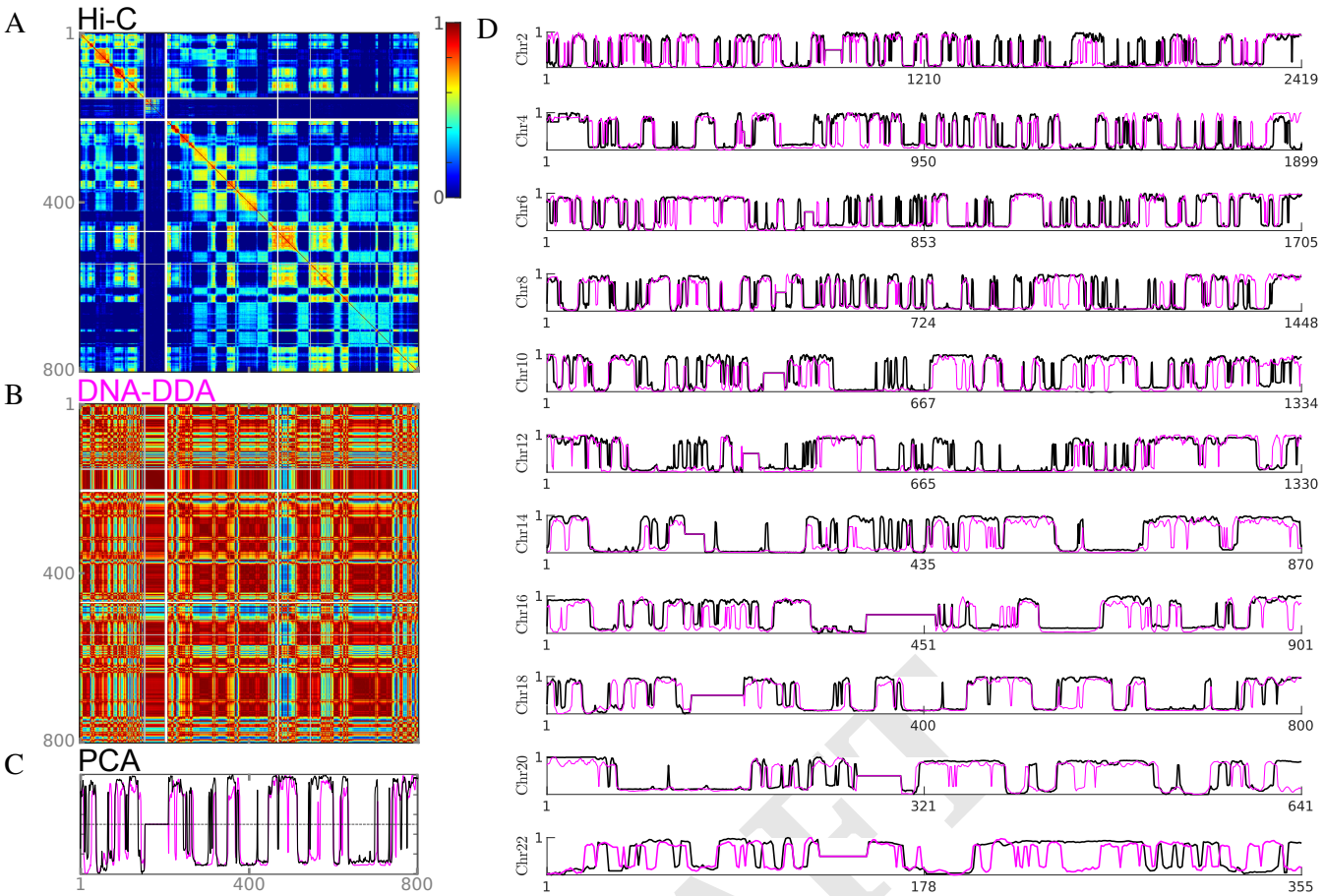

Fig. 2. DNA-DDA results The Hi-C (A) and DNA-DDA (B) Pearson correlation matrices of the K562 cell line for an example hold out chromosome 18 exhibit strikingly similar patterns. Resulting principal components (C) ($PC_{\text{DNA-DDA}}$ and $PC_{\text{Hi-C}}$) used to define the switching between A/B compartments are in very strong correlation to one another ($r_{PC} = 0.71$). D) shows the performance of this DNA-DDA model over all even numbered chromosomes for the K562 cell line

of the labels derived from the experimental data. It is known that identifying A and B compartments by PCA of the Pearson correlation matrix derived from the contact map is suboptimal and misses many features, especially at higher resolutions. This is particularly apparent in the IMR90 data set (Supplementary Figure 5) or particular chromosomes (e.g., chr22).

Comparison to other methods. We compared DNA-DDA to SACSANN(45) and ABCNet(46). To our knowledge, these are the only two methods that are directly comparable to DNA-DDA in that they predict A/B compartments from the DNA sequence or sequence-based features only. DNA-DDA competes well with other methods predictive of A/B compartments when assessed on the same Hi-C data set (GSE35156) as well as overall (Table 3). Indeed, all three methods showed very similar AUC (ACC) scores. Note the vast difference in the size of the *training* (pendant to structure selection in DNA-DDA) and test data between the methods in Table 2.

Discussion and Conclusion

DNA-DDA is a novel purely sequence-based method for predicting the A/B compartments in a cell-type specific manner. DNA-DDA is based on DDA, a nonlinear dynamical classification tool that has previously been applied to time series data and we adapted to analyze genomic data. We determined the model parameters (τ_1, τ_2 in Equation 6) using human chromosome 22 and tested it on the remaining human autosomes. The achieved performances on never-before-seen data were remarkable, competing strongly with existing methods whilst requiring far less data to derive the models.

Predicting long range interactions is challenging and computationally extremely expensive (47). Many factors must be taken into account when comparing two models of chromosome architecture, including the the amount and types of input features (e.g. nucleotide sequence, GC content, TFBs), the target output (e.g. A/B compartments, contacts at various resolutions, TADs), the data sets used for training, validation and testing, and computational effort. Akita (48) and DeepC (49) are both convolutional neural network (CNN) based models that can predict genomic contacts at a resolution of up to 1Mbp from the sequence alone. It will be exciting to see the development of these methods distal contacts > 1 Mbp and perhaps a merging between such data-driven and our nonlinear dynamics based approach will be critical in advancing our understanding of genome function. ABCNet (46) and SACSANN (45) on the other hand, are also CNN-based models and can predict genome-wide A/B com-

DRAFT

ChrNr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	μ	σ
GM12878																								
AUC	0.85	0.81	0.91	0.84	0.81	0.84	0.76	0.82	0.83	0.76	0.90	0.84	0.81	0.88	0.77	0.83	0.78	0.83	0.81	0.81	0.81	0.70	0.82	0.00
ACC	0.81	0.73	0.83	0.77	0.74	0.78	0.72	0.76	0.81	0.70	0.81	0.75	0.77	0.79	0.69	0.80	0.74	0.79	0.78	0.77	0.78	0.69	0.76	0.00
F1	0.78	0.70	0.81	0.72	0.70	0.76	0.72	0.72	0.76	0.69	0.79	0.71	0.76	0.77	0.68	0.76	0.75	0.74	0.79	0.75	0.76	0.71	0.74	0.00
r_{PC}	0.67	0.53	0.75	0.59	0.54	0.61	0.49	0.56	0.66	0.45	0.70	0.59	0.58	0.65	0.46	0.63	0.52	0.61	0.56	0.57	0.60	0.39	0.58	0.01
K562																								
AUC	0.89	0.88	0.91	0.88	0.84	0.87	0.85	0.85	0.72	0.84	0.87	0.88	0.84	0.92	0.80	0.86	0.75	0.89	0.82	0.85	0.86	0.67	0.84	0.00
ACC	0.84	0.80	0.83	0.80	0.77	0.80	0.76	0.79	0.74	0.74	0.78	0.81	0.79	0.83	0.71	0.83	0.71	0.84	0.77	0.81	0.81	0.65	0.78	0.00
F1	0.81	0.76	0.80	0.76	0.72	0.78	0.72	0.76	0.67	0.73	0.75	0.78	0.76	0.81	0.69	0.80	0.70	0.82	0.80	0.80	0.78	0.65	0.76	0.00
r_{PC}	0.72	0.66	0.74	0.64	0.59	0.68	0.60	0.63	0.53	0.58	0.63	0.70	0.63	0.75	0.51	0.68	0.44	0.73	0.60	0.66	0.67	0.31	0.62	0.01
IMR90																								
AUC	0.81	0.76	0.80	0.82	0.75	0.77	0.67	0.71	0.73	0.71	0.79	0.84	0.82	0.81	0.68	0.83	0.71	0.75	0.72	0.75	0.80	0.59	0.75	0.00
ACC	0.76	0.69	0.76	0.75	0.69	0.70	0.66	0.66	0.75	0.66	0.73	0.78	0.77	0.74	0.65	0.81	0.68	0.72	0.70	0.73	0.78	0.59	0.72	0.00
F1	0.70	0.64	0.71	0.72	0.61	0.66	0.55	0.62	0.66	0.64	0.72	0.78	0.76	0.69	0.61	0.75	0.69	0.67	0.69	0.70	0.72	0.62	0.68	0.00
r_{PC}	0.56	0.42	0.57	0.55	0.42	0.46	0.31	0.37	0.46	0.39	0.54	0.60	0.60	0.53	0.30	0.63	0.38	0.47	0.42	0.45	0.58	0.20	0.46	0.01
hESC																								
AUC	0.90	0.73	0.93	0.81	0.73	0.76	0.88	0.75	0.90	0.76	0.86	0.96	0.82	0.88	0.85	0.74	0.80	0.81	0.75	0.81	0.83	0.73	0.82	0.00
ACC	0.85	0.68	0.86	0.75	0.67	0.71	0.84	0.71	0.89	0.68	0.81	0.91	0.76	0.82	0.77	0.75	0.74	0.76	0.72	0.78	0.77	0.71	0.77	0.00
F1	0.83	0.66	0.83	0.74	0.64	0.69	0.82	0.69	0.83	0.68	0.78	0.89	0.73	0.79	0.75	0.66	0.73	0.74	0.74	0.77	0.75	0.73	0.75	0.00
r_{PC}	0.75	0.45	0.78	0.58	0.44	0.51	0.74	0.49	0.77	0.46	0.67	0.86	0.56	0.72	0.61	0.47	0.54	0.54	0.44	0.60	0.58	0.42	0.59	0.02

Table 4. Performances of DNA-DDA

partments but not individual contacts.

Orca (50) is the first other purely sequence based model that predicts large-scale genome organization from kilobases to whole chromosome scale. It was trained (all chromosomes except for chr8-10) and validated (chr8) on two of the most highest resolution microC datasets to date and acheives 0.78-0.85 Pearson correlation on hold out test chromosomes (chr9-10). Orca accuracy was demonstrated in a comprehensive study in which experimentally observed impacts of particular sequence variants were verified.

The power of methods like Orca or DNA-DDA lies in their intermediate step of generating the entire contact matrix before calling A/B compartments, which can be used in the exact same manner as Hi-C maps. Although Orca currently predicts contact maps for higher resolutions, and has been more thoroughly tested in terms of biological function, DNA-DDA shows promising results whilst tackling the problem from a different angle and requiring far less training data. It would be interesting to see what information about 3D genome organization a combined effort ML based and nonlinear dynamics based methodologies could recover. In principle, in the same manner as presented here for input sequences of $1Mbp$, DNA-DDA could easily be modified to predict higher resolution contact maps and can therefore also be used to investigate other levels of chromatin organization, such as TADs, chromatin loops, and enhancer-promoter pairs. Finally, the fact that DDA-models can be derived from such a little amount of supervised data could enable the prediction of genome-wide interactions in other cell types from a very limited amount of chromosome conformation capturing data (e.g., generated with 5C).

DNA-DDA performs similarly to state-of-the-art methods which use a larger portion of the genome for training and/or input features other than solely the sequence itself (Table 3). For instance, SACSANN (45) uses counts of sequence-derived features (e.g. TEs, TFBSs). Although Prost et al. called compartments with an a different software (HOMER (41)), when tested on the same Hi-C data set, the AUC between SACSANN and DNA-DDA was nearly identical (0.81 vs 0.82 respectively). SACSANN shows robust performance across different cell types and species, however its input features need to be first identified in the genome of interest. ABCNet (46) uses only the sequence as input and achieves slightly higher performances than DNA-DDA.

We want to emphasize that the DNA-DDA models presented in this publication highly unlikely represent an optimum. This work is a proof of concept meant to illustrate the potential in applying nonlinear dynamics based approaches in the field of genomics. Many aspects of our analysis could improved. For example, the functional form of the model was chosen based only on simplicity and computational efficiency. In previous work applying DDA, the general assumption that the overall functional form of a DDA model tends to be specific to the data type used to monitor the system of interest (e.g., EEG, ECG, DNA sequence), while the delay pairs are sensitive to the question we ask about the system. Furthermore, we observed that $\mathrm{PC}_{\text{DNA-DDA}}$ is highly dependent on the way the the DNA-DDA contact matrix $\mathbf{D}$ is post processed (e.g. filtering, logarithmic transformation, etc.). In the future, we aim to derive a DDA model which captures nonlinear features inherent specifically to the DNA sequence irrespective of our specific goal of predicting A/B compartments. A large-scale exhaustive sweep of model-delay-pair combinations such as described in (23) could be implemented to optimize model-delay pair combinations. Also, determining the DNA-DDA model parameters τ_1 and τ_2 on larger regions exhibiting higher diversity across cells will most likely result in better cell-type specificity.

Even so, the models found on the 200 100kbp bins on the arbitrarily chosen chromosome 22 performed exceptionally well across hold out chromosomes. The sparsity of data used for structure selection to determine the model parameters τ_2, τ_2 provides certainty of not over-fitting and highlights the generalizability and global modeling capabilities of DDA. Finally, we inferred A/B compartments from the DNA-DDA contact matrix based on the principal component (PC) method proposed by Lieberman-Aiden et al. (1). We acknowledge that this method is not optimal. In fact, Rao et al. (2) showed that genomes segregate into at least six subcompartments, each exhibiting a distinctive pattern of genomic

DRAFT

and epigenetic features. A large improvement in the structure selection step would be to select delay-pairs based on the similarity of the DNA-DDA and Hi-C contact matrices themselves. This is not a trivial task as standard correlation measures like the Pearson correlation or Spearman correlation can produce false conclusions about the similarity of genomic contact matrices (51). As mentioned above, since all relevant 3D structures could be extracted from interaction matrices (at various resolutions), we strongly believe that choosing the delay-pairs in this manner would boost cell-type specificity tremendously.

The DNA sequence plays a fundamental role in the formation and maintenance of 3D chromatin architecture, which in turn is a central orchestrator of the gene regulatory network. It is indisputable, that the sequence contains key underlying features contributing to genome folding, however to what extent it alone can be used to gain access to all 3D interactions and relevant structures remains an open question. Our findings strongly support the hypothesis that the DNA sequence represents a highly observable variable of 3D genome architecture. Our preliminary results suggest the potential of this exciting new approach to predict chromatin interactions and we plan to improve our method with regards to structure selection as described above. Furthermore, DNA-DDA will be implemented to generate contact maps for higher resolutions. We believe DDA could also be suitable for deriving a computationally inexpensive model which directly classifies the genomic structural feature or interest (eg. compartments, TADs, loops or enhancer-promoter pairs), without the intermediate step of having to compute the entire contact map. As already demonstrated by Akita and DeepC, purely sequence based approaches have the enormous advantage of being able to model the effects of genetic variation on 3D structure, a task far to elaborate to be experimentally investigated. We plan to generate genome-wide contact matrices from perturbed sequences to investigate impacts of genetic variation and to understand more about the mechanisms imperative to functioning genome architecture.

Acknowledgements

We thank Dr. Claudia Lainscsek for the fruitful discussions and sharing the `C` implementation of DDA, which was used as a basis for DNA-DDA.

Data availability

Exemplary code used to predict contacts at 100kbp resolution of chromosome 22 can be found at https://github.com/xX3N1A/DNA-DDA.

Bibliography

1. E. Lieberman-Aiden, L. van Berkum N, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, R. Lajoie B, J. Sabo P, O. Dorschner M, R. Sandstrom, B. Bernstein, A. Bender M, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, A. Mirny L, S. Lander E, and Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–93, 2009.
2. S.S. Rao, M.H. Huntley, N.C. Durand, E.K. Stamenova, I.D. Bochkov, J.T. Robinson, A.L. Sanborn, I. Machol, A.D. Omer, E.S. Lander, and E.L Aiden. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–80., 2014.
3. A.U. Grosberg, S.K. Nechaev, and E.I. Shakhinovich. The role of topological constraints in the kinetics of collapse of macromolecules. *J. Phys. France*, 49(11):2095–2100, 1988.
4. J.R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J.S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–80, 2012.
5. R. Zhou and Y. Q. Gao. A dna sequence based polymer model for chromatin folding. *International journal of molecular sciences*, 22(3):1328, 2021.
6. J. Nuebler, G. Fudenberg, M. Imakaev, N. Abdennur, and L. A. Mirny. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proceedings of the National Academy of Sciences of the United States of America*, 115(29):E6697–E6706, 2018.
7. Y. Qi and B. Zhang. Predicting three-dimensional genome organization with chromatin states. *PLOS Computational Biology*, 15(6):e1007024, 2019.
8. P. Belokopytova and V. Fishman. Predicting genome architecture: Challenges and solutions. *Frontiers in genetics*, 11:617202, 2021.
9. A. Krumm and Z. Duan. Understanding the 3d genome: Emerging impacts on human disease. *Seminars in cell & developmental biology*, 90:62–77, 2019.
10. H. Degn, A.V. Holden, and L.F. Olsen. Chaos in biological systems. In H. Degn, A.V. Holden, and L.F. Olsen, editors, *NATO Advanced Research Workshop on "Chaos in Biological Systems" December 8-12, 1986*, Dyffryn House, St. Nicholas, Cardiff, U. K., 1987.
11. C. Letellier. *Chaos in Nature 2nd Edition*. WORLD SCIENTIFIC, Singapore, 2013.
12. B. Hewelt, H. Li, M.K. Jolly, P. Kulkarni, I. Mambetsariev, and R. Salgia. The dna walk and its demonstration of deterministic chaos-relevance to genomic alterations in lung cancer. *Bioinformatics*, 35(16):2738–2748, 2019.
13. S. Pigolotti, M.H. Jensen, Y. Zhan, and G Tiana. Bifractal nature of chromosome contact maps. *Phys. Rev. Research*, 2(4):043078, 2020.
14. Edward Lorenz. Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, 20(2): 130–148, 1963.
15. Henri Poincaré. Sur le problème des trois corps et les équations de la dynamique. *Acta Math*, 13:1–270, 1890.
16. Henri Poincaré. *Méthodes nouvelles de la mécanique céleste*. Gauthier-Villars et fils, Paris, 1854-1912.
17. A.M. Lyapunov. *The General Problem of the Stability of Motion*. PhD dissertation, University of Moscow, 1892.
18. B.B. Mandelbrot. *Les objets fractals: forme, hasard et dimension*. Flammarion, Paris, 1975.
19. Ludwig Boltzmann. *Vorlesungen über Gastheorie. Bd. 2*. Verlag von Johann Ambrosius Barth, Leipzig, 1889.
20. George D. Birkhoff. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences*, 17(12):656–660, 1931.
21. J. v. Neumann. Proof of the quasi-ergodic hypothesis. *Proceedings of the National Academy of Sciences*, 18(1):70–82, 1932.
22. C. Lainscsek and T.J Sejnowski. Delay differential analysis of time series. *Neural Computation*, 23(3):594–614, 2014.
23. C. Lainscsek, J. Weyhenmeyer, S.S. Cash, and T.J Sejnowski. Delay differential analysis of seizures in multichannel electrocorticography data. *Neural Computation*, 29(12): 3181–3218, 2017.
24. C. Lainscsek, C.E. Gonzalez, A.L. Sampson, S.S. Cash, and T.J. Sejnowski. Causality detection in cortical seizure dynamics using cross-dynamical delay differential analysis. *Chaos*, 29:101103, 2019.
25. C. Lainscsek, R. Rungratsameetaweemana, S.S. Cash, and T.J Sejnowski. Cortical chimera states predict epileptic seizures. *Chaos*, 29:121106, 2019.
26. Floris Takens. Detecting strange attractors in turbulence. In David Rand and Lai-Sang Young, editors, *Dynamical Systems and Turbulence, Warwick 1980*, pages 366–381, Berlin, Heidelberg, 1981. Springer Berlin Heidelberg.
27. I. Aguirre and C. Letellier. Investigating observability properties from data in nonlinear dynamics. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 83, 2020.
28. C. Lainscsek, S.S. Cash, T.J Sejnowski, and J. Kurths. Dynamical ergodicity dda reveals causal structure in time series. *Chaos*, 31:103108, 2021.
29. A.L. Sanborn, S.S. Rao, S.C. Huang, N.C. Durand, M.H. Huntley, A.I. Jewett, I.D. Bochkov, D. Chinnappan, A. Cutkosky, J. Li, K.P. Geeting, A. Gnirke, A. Melnikov, D. McKenna, E.K. Stamenova, E.S. Lander, and E.L. Aiden. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A*, 112(47):E6456–65, 2015.
30. ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.
31. D. Karolchik, Hinrichs A.S., Furey T.S., Roskin K.M., Sugnet C.W., D. Haussler, and Kent W.J. The ucsc table browser data retrieval tool. *Nucleic Acids Res.*, 1(32(Database issue):D493-6), 2004.
32. A.J. Parry, M. Hoare, D. Bihary, and R. et. al. Hänsel-Hertsch. Notch-mediated non-cell autonomous regulation of chromatin structure during senescence. *Nat Commun*, 9(1):1840, 2018.
33. Langmead Ben and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9:357–359, 2012.
34. Joachim Wolff, Leily Rabbani, Ralf Gilsbach, Gautier Richard, Thomas Manke, Rolf Backofen, and Björn A Grüning. Galaxy hicexplorer 5: a web server for reproducible hi-c, capture hi-c and single-cell hi-c data analysis, quality control and visualization. *Nucleic Acids Research*, 48(W1):W177–W184, 2020.
35. Joachim Wolff, Vivek Bhardwaj, Stephan Nothjunge, Gautier Richard, Gina Renschler, Ralf Gilsbach, Thomas Manke, Rolf Backofen, Fidel Ramírez, and Björn A Grüning. Galaxy hicexplorer: a web server for reproducible hi-c data analysis, quality control and visualization. *Nucleic Acids Research*, 46(W1):W11–W16, 2018.
36. 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. *Nature Communications*, 9: 189, 2018.
37. A.P. Knight and D. Ruiz. A fast algorithm for matrix balancing. *IMA J. Numer. Anal.*, 33,

DRAFT

2007.
38. C.K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H.E. Stanley. Long-range correlations in nucleotide sequences. *Nature*, 356:168–170, 1992.
39. C.-K. Peng, S.V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H.E. Stanley. Fractal landscape analysis of dna walks. *Physica A: Statistical Mechanics and its Applications*, 191(1-4):25–29, 1992.
40. J.A. Berger, S.K. Mitra, M. Carli, and A. Neri. Visualization and analysis of dna sequences using dna walks. *Journal of the Franklin Institute*, 341(1):37–53, 2004.
41. S. Heinz, C. Benner, N. Spann, E. Bertolino, Y.C. Lin, P. Laslo, J.X. Cheng, C. Murre, H. Singh, and C.K. Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Mol Cell*, 38:576–89, 2010.
42. M. Tsompana and M.J. Buck. Chromatin accessibility: a window into the genome. *Epigenetics & Chromatin*, 7(33), 2014.
43. C. Lainscsek, M. E. Hernandez, J. Weyhenmeyer, T. J. Sejnowski, and H Poizner. Nonlinear dynamical analysis of eeg time series distinguishes patients with parkinson's disease from healthy individuals. *Frontiers in neurology*, 4(200), 2013.
44. A.L. Sampson, C. Lainscsek, C.E. Gonzalez, I. Ulbert, O. Devinsky, D. Fabó, J.R. Madsen, E. Halgren, S.S. Cash, and T.J. Sejnowski. Delay differential analysis for dynamical sleep spindle detection. *Journal of neuroscience methods*, 316:12–21, 2019.
45. J. Prost, C. Cameron, and M. Blanchette. Sacsann: identifying sequence-based determinants of chromosomal compartments. *bioRxiv*, 2020.
46. M. Kirchhof, C.JF. Cameron, and S.C. Kremer. End-to-end chromosomal compartment prediction from reference genomes. In *IEEE BIBM 2021 (online)*, Houston, TX, USA, 2021.
47. H. Tao, H. Li, K. Xu, H. Hong, S. Jiang, G. Du, J. Wang, Y. Sun, X. Huang, Y. Ding, F. Li, X. Zheng, H. Chen, and Bo X. Computational methods for the prediction of chromatin interaction and organization using sequence and epigenomic profiles. *Brief Bioinform*, 22 (5):bbaa405, 2021.
48. G. Fudenberg, D.R. Kelley, and K.S. Pollard. Predicting 3d genome folding from dna sequence with akita. *Nat Methods*, 17:1111–1117, 2020.
49. R. Schwessinger, M. Gosden, D. Downes, R.C. Brown, Oudelaar A.M., Y.W. Telenius, J. Teh, G. Lunter, and J.R. Hughes. Deepc: predicting 3d genome folding using megabase-scale transfer learning. *Nat Methods*, 17:1118–1124, 2020.
50. J. Zhou. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat Genetics*, 54:725–734, 2022.
51. T. Yang, F. Zhang, GG. Yardımcı, F. Song, R.C. Hardison, W.S. Noble, F. Yue, and Q. Li. Hicrep: assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient. *Genome Res*, 11:1939–1949, 2017.

DRAFT