Annual Review of Biomedical Data Science

Computational Approaches for Understanding Sequence Variation Effects on the 3D Genome Architecture

Pavel Avdeyev and Jian Zhou

Lyda Hill Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, Texas, USA; email: Jian.Zhou@utsouthwestern.edu

Annu. Rev. Biomed. Data Sci. 2022. 5:183–204

First published as a Review in Advance on May 10, 2022

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-102521-012018

Keywords

sequence variations, chromatin organization, machine learning, sequence-based models

Abstract

Decoding how genomic sequence and its variations affect 3D genome architecture is indispensable for understanding the genetic architecture of various traits and diseases. The 3D genome organization can be significantly altered by genome variations and in turn impact the function of the genomic sequence. Techniques for measuring the 3D genome architecture across spatial scales have opened up new possibilities for understanding how the 3D genome depends upon the genomic sequence and how it can be altered by sequence variations. Computational methods have become instrumental in analyzing and modeling the sequence effects on 3D genome architecture, and recent development in deep learning sequence models have opened up new opportunities for studying the interplay between sequence variations and the 3D genome. In this review, we focus on computational approaches for both the detection and modeling of sequence variation effects on the 3D genome, and we discuss the opportunities presented by these approaches.

1. INTRODUCTION

The 3D architecture of the chromosome and the process of chromatin folding are increasingly well understood, in large part due to the recent development of technologies for studying them in unprecedented detail. High-resolution genome-wide chromatin interaction maps have revealed the multiscale chromatin organization with very fine resolution. Such techniques and datasets have provided the foundation for understanding the genomic sequence dependencies of genome architecture at all spatial scales. The 3D genome architecture is highly specific to genomic location and expected to be strongly dependent on genomic sequence. The 3D genome structure also in turn determines the environment of sequence-based regulatory processes, making it essential for understanding sequence variation effects (1–3). Sequence variations can have profound impacts on the 3D genome architecture, often with pathogenic consequences (1).

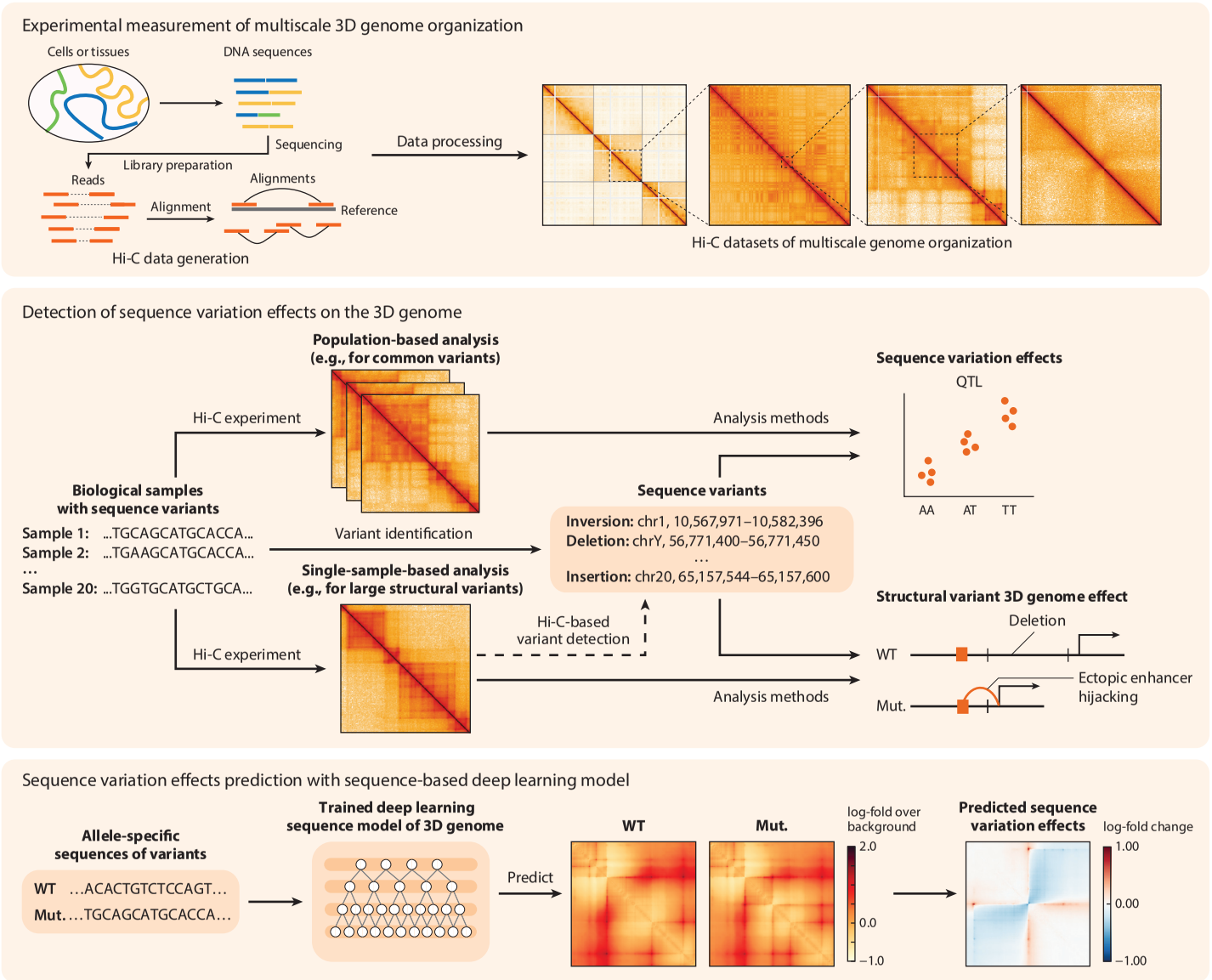
The scale and complexity of the genomic dataset for measuring the 3D genome present unprecedented opportunities for understanding the interplay between genomic sequence and architecture, which drives the development of computational methods for these purposes. Computational methods have been critical for both detecting sequence variation effects from experimental datasets and modeling the 3D genome dependencies on the sequence. In particular, computational sequence-based modeling approaches have now allowed high-throughput prediction of the impact of sequence variation on the 3D genome for any variant. Moreover, the sequence-based models can assist in understanding the mechanisms of the 3D genome by generating experimentally testable hypotheses. The improvement and applications of these computational methods will likely accelerate our understanding of the interplay between the 3D genome and the genomic sequence. In this review, we discuss the current computational approaches and opportunities in advancing the understanding of the interplay between genome sequence and 3D genome architecture (**Figure 1**). We have also included an introduction to relevant background information for techniques and 3D genome organization to help understand the 3D genome data. We aim to make this review accessible for readers from either computational or experimental backgrounds.

2. TECHNIQUES AND CURRENT UNDERSTANDING OF THE 3D GENOME ARCHITECTURE

2.1. Technologies for Measuring 3D Genome Architecture

The recent introduction and development of high-throughput sequencing–based techniques for measuring chromatin organization, represented by chromosome conformation capture (3C)-based methods (4–8), have revolutionized our understanding of the multiscale 3D genome structure. 3C technologies detect the spatial proximity of DNA sequences in the 3D space based on their increased likelihood of ligation. The first step of 3C-based methods typically involves the formaldehyde cross-linking that fixes the positions of DNA sequences that are spatially proximal in the nucleus. The second step is the fragmentation of the chromatin, followed by in situ ligation and removal of cross-links. The last step involves the determination of pairwise interactions using ligation junctions with either polymerase chain reaction or sequencing technologies. Here we briefly review several representative high-throughput sequencing–based experimental techniques (**Table 1**). For a more detailed overview of different 3C-based approaches, readers are referred to References 2, 9, and 10.

Hi-C (high-throughput 3C) combines 3C with high-throughput sequencing and allows analyzing genome contacts across the entire genome. Currently, it is the most widely used technique to assess chromatin architecture and has undergone dynamic improvements to achieve higher resolution. Micro-C (micrococcal nuclease chromosome conformation assay) (11–13) uses micrococcal nuclease digestion to obtain up to mononucleosome resolution in the detection of



(Caption appears on following page)

Figure 1 (*Figure appears on preceding page*)

Overview of computational methods for understanding sequence and sequence variation effects on the 3D genome. Experimental techniques for measuring 3D genome structure, especially chromosome conformation capture–based techniques such as Hi-C, provide multiscale 3D genome organization measurement covering diverse types of structures and interactions. Such data and techniques enable both computational methods for the detection of sequence variation effects from Hi-C datasets (through QTL-based analysis of population-level 3D genome datasets or analysis of large structural variant effects from Hi-C datasets of single samples) and computational modeling of sequence variation effects (through training a deep learning sequence model that predicts the 3D genome from sequence). Abbreviations: Hi-C, high-throughput chromosome conformation capture; Mut., mutant; QTL, quantitative trait locus; WT, wild-type.

chromatin interactions, which allows for resolving differences in genome interaction patterns among small genomic intervals near or below the kilobase level. Another improvement to the Hi-C technique, called Hi-C 3.0, combines the advantages of Micro-C for chromosomal loop detection and Hi-C for detecting stronger chromatin compartments (14). The recently introduced MCC (Micro-Capture-C) method enables studying sequence interactions on base pair resolution from individual viewpoints (15), breaking the barrier of 1-kilobase (kb) resolution. ChIA-PET (chromatin interaction analysis with paired-end tag) (16) and, more recently, HiChIP (Hi-C chromatin immunoprecipitation) (17) and PLAC-seq [poly(A)-ClickSeq] (18) combine 3C techniques with chromatin immunoprecipitation and help to analyze the contact points of sequences bound by a protein of interest.

In addition to improving resolution, scHi-C (single-cell Hi-C) has been developed for detecting pairwise contact interactions from individual isolated cells (19, 20). The single-cell approach has the potential to reveal 3D genome structures not visible by regular Hi-C, as the latter pools measurements from a population of cells. Thus, the single-cell approach has the advantage of

Table 1 Example of technologies that can be used to assay 3D genome organization

Bulk or single-cell	Type of interactions[a]	Number of regions involved in interactions[b]	Examples	Description
Bulk (typically millions of cells)	One-way	One-versus-all	MCC, NG-Capture-C, 4C-seq	Detect sequence interactions for a single viewpoint location with up to base pair resolution
		All-versus-all	Hi-C, Hi-C 3.0, Micro-C	Detect sequence interactions across the entire genome
			ChIA-PET, Hi-ChIP, PLAC-seq	Detect interactions between sequences bound by a protein of interest
	Multiway	All-versus-all	GAM, SPRITE, Pore-C	Detect genome-wide higher-order interactions within the nucleus
Single-cell	One-way	All-versus-all	scHi-C	Detects 3D genome organization in rare cell types and at specific stages of the cell cycle
	Multiway		scSPRITE	Detects multiway interactions within an individual cell

[a]One-way interactions involve a single interaction between a pair of regions; multiway interactions involve all interactions among three or more regions.
[b]One-versus-all interactions include interactions connecting one region with all regions; all-versus-all interactions include interactions connecting all regions with all regions.

Abbreviations: 3C, chromosome conformation capture; 4C-seq, circular 3C sequencing; ChIA-PET, chromatin interaction analysis with paired-end tag; GAM, genome architecture mapping; Hi-C, high-throughput 3C; Hi-ChIP, Hi-C chromatin immunoprecipitation; MCC, Micro-Capture-C; Micro-C, micrococcal nuclease chromosome conformation assay; NG, next-generation; sc, single-cell; PLAC-seq, poly(A)-ClickSeq; SPRITE, split-pool recognition of interactions by tag extension.

enabling the study of the 3D organization of the genome in rare cell types and at specific stages of the cell cycle (21). However, as a tradeoff for single-cell specificity, this technique detects many fewer contacts than regular Hi-C and currently does not provide comparable resolution.

More recently, methods have been developed that detect the spatial proximity of multiple sequences using ligation-free approaches. Such interactions are often called multiway or higher-order interactions. Genome architecture mapping (GAM) technology (22) obtains information about chromatin structure from sequencing a collection of thin cryo-sectioned nuclear profiles. Another approach, SPRITE (split-pool recognition of interactions by tag extension), can map both RNA and DNA interactions simultaneously (23). SPRITE involves tagging cross-linked interacting molecules with a unique barcode through repeated splitting and pooling procedures. Interactions among different genomic regions are then identified by sequencing and matching reads containing the same barcode to identify colocalized sequences (24). In silico simulations suggest that to obtain high-resolution contact maps SPRITE requires a smaller number of cells than Hi-C, while GAM requires more and has the added advantage of detecting contacts above the 1-megabase (Mb) distance scale (25).

Current sequencing-based assays mostly rely on short-read sequence technology and can have difficulty resolving the 3D genome structure of highly repetitive regions. The recent development of long-read sequencing-based techniques and their application to 3D genome measurement (26) have the potential to address the mappability problems within repetitive regions and provide more information about chromatin interactions within them.

In parallel to sequencing-based techniques, high-throughput imaging–based methods have been developed to provide single-cell and in situ measurements. While we focus on sequencing-based techniques in this review because of the abundance of high-quality datasets generated with these technologies and computational methods developed for these datasets, we refer interested readers to previous reviews of these techniques (10).

2.2. The Multiscale Organization of 3D Genome

With the data generated from the abovementioned techniques, we have now gained a much better understanding of the hierarchical organization of the 3D genome. In typical interphase cells, at kilobase and larger spatial scales, the 3D genome structure can often be described by domain-type structures. These include, but are not limited to, entire regions of varying lengths such as topologically associating domains (TADs) and chromatin compartments, as well as loop-type structures that have interactions between two genomic locations. Both domain- and loop-type structures involve multiple mechanisms that act at different spatial scales (2, 10).

The most prominent character of the 3D genome organization at a length scale of 40 kb–1 Mb is that the genome is spatially organized into sub-megabase-scale domains (i.e., TADs) (27–29). Sequences within the same TAD interact with each other much more frequently than they do with segments located in different domains. TADs often show a nested structure whereby large TADs can be further subdivided into smaller domains (sometimes called sub-TADs), and algorithms have been developed to identify TADs in a hierarchical fashion (30). In contrast, most loop structures span distances similar to or less than the size of a TAD. The most common loop structure is the CTCF-cohesin-dependent loop, which is closely related to TADs and observed at TAD boundaries (31, 32), and cohesin complex–dependent loop extrusion is likely the mechanism for TAD formation (33–37). Other types of loops include Polycomb-mediated contacts (38) and loops between transcriptional regulatory elements like promoter–enhancer, promoter–promoter, and enhancer–enhancer loops, which have been observed both together with and independent of CTCF-cohesin-dependent loops (39, 40). The loop structures can play a role in gene regulation

by bringing *cis*-regulatory elements such as enhancers and promoters into close spatial proximity (41).

At larger scales (>1 Mb), the most prominent feature of genome organization is the chromatin compartment, which is often separated into A and B compartments. Regions in each compartment preferentially interact with the same compartment; that is, regions in compartment A interact mostly with other compartment A regions, and similarly for compartment B, forming a plaid-like pattern in Hi-C contact maps (42). Chromatin compartment A is enriched in active genes and open chromatin while compartment B is relatively gene depleted and enriched in closed chromatin (27, 42, 43). A and B compartments can also be more finely divided into six different subcompartments (31, 44, 45). What drives the process of compartment formation is still largely unknown, especially at the sequence level, even though active gene expression is associated with and has been suggested to play a role in compartment A formation (3).

It is important to note that even though compartmentalization was originally discovered at larger spatial scales, more recent studies with higher-resolution data show that chromatin compartmentalization does happen at TAD scales, and smaller domains with the characteristic compartment-like interaction pattern can be observed (13, 46). A recent study with a very high-coverage Hi-C dataset (30 billion) reported that the median length of compartment domains is only 12.5 kb (46). Thus, chromatin compartmentalization likely shapes the local 3D genome jointly with other mechanisms.

Finally, chromosomes occupy distinct spaces in the nucleus called chromosome territories (42, 47). The relative positions of chromosomes with respect to each other are assumed to be largely random. However, there appears to be a tendency for some chromosomes to be located near the center of the nucleus, while others tend to be located more peripherally (also known as radial positioning) (48, 49).

While the exact roles of the 3D genome in gene regulation are still not fully understood, with these discoveries of 3D sequence organization it has become clear that chromatin architecture is significant for spatial gene positioning and gene regulation (3, 41, 50, 51). 3D structures influence essential biological functions such as transcription, replication, and DNA repair (52, 53). Thus, modifications in the 3D organization due to sequence mutations may lead to alterations of these processes that sometimes can severely impact human health (54). For example, duplication of a TAD boundary at the *SOX9* locus causes neo-TAD formation and leads to Cooks syndrome and nail aplasia (55, 56). More detailed information about 3D chromosome organization and its impact on various biological processes has been extensively reviewed before and we refer the reader to References 2, 41, and 57. In the next sections, we discuss variations of DNA sequence and their impact on 3D genome architecture.

3. SEQUENCE VARIATIONS AND THEIR IMPACT ON 3D GENOME ARCHITECTURE

Sequence variation refers to differences in genomic sequences across individuals or biological samples that can be inherited or arise from spontaneous mutations. The size of DNA sequence variation ranges from point mutations, also known as single-nucleotide polymorphisms (SNPs), to large-scale chromosomal events, also known as structural variants (SVs) (58, 59). SVs are extremely diverse in length; the term "SVs" is generally used to refer to variants from approximately 10 base pairs (bp) to multiple megabases in length or even to whole-chromosome abnormalities. SVs also include numerous types of events such as insertions, deletions, translocations, inversions, tandem and segmental duplications, mobile element insertions, and copy number variations (60, 61). Complex genome rearrangement that involves more than three genome breakpoints can also

happen and is common in cancer (62, 63). Sequence variations vary greatly in their frequency in the human population, ranging from common variants that are typically inherited to de novo germline mutations that are often unique to the individual (64). Common variants usually have small effects on health since variants with a strong negative impact on fitness tend to undergo negative selection and are consequently less frequent in the population. However, the high population frequency of common variants allows for information to be pooled across a population, thus enabling more sensitive detection of their effects. On the other end of the spectrum, rare variants can have a greater impact, although most of them are not impactful. Sequence mutations can also occur after fertilization and result in somatic mutations that exist only in certain cell lineages, as is the case for the vast majority of cancer mutations.

Sequence variations can now be systematically discovered through genome sequencing approaches and many have been found to play a crucial role in complex human diseases such as cancer, intellectual disability, autism, major depressive disorder, and type 2 diabetes mellitus (65–70). SNPs are the most common form of human genome variation (64, 71, 72), and the associations between common variants that are mostly SNPs and various forms of human phenotypic variations have been analyzed in genome-wide association studies (GWAS), in which millions of sequence variants across the genomes of many individuals are analyzed to identify genotype–phenotype associations. From GWAS, more than 200,000 associations of genome-wide significance have been reported between sequence variants and common diseases and traits, and the vast majority of them are located in the noncoding genome (73).

Widespread application of whole-genome high-throughput sequencing technologies for the detection of genetic variants has also shown large amounts of SVs present in the human population. For example, the 1000 Genomes Project detected 68,818 SVs after analyzing 2,504 human genomes (64). Large SVs have been observed in nearly all types of cancer at every stage of the disease progression (74). For example, Aganezov et al. (67) showed that a large number of novel SVs that were missed by previous sequencing technologies can be detected in breast cancer cell lines by using long-read sequencing technologies. SVs are expected to have a larger impact than the smaller genome variations and have often been linked to pathogenic outcomes. For a more detailed review of sequence variations, methods for their discovery, and their impact on human health, readers are referred to References 58–60 and 75.

While high-throughput sequencing technologies have accelerated the detection of sequence variations and linked many to diseases and traits, understanding their molecular-level effects and phenotypic consequences, including their impact on the 3D genome, remains challenging. Due to the relevance of 3D genome architecture on genome function and regulation and the potentially strong impact of sequence variation on the 3D genome, it thus becomes very important to understand sequence variations from the perspective of the 3D genome organization and study how sequence variation may change chromatin architecture. In the next section, we discuss different computational approaches whose main goal is to facilitate discoveries of different sequence variation effects on 3D genome architecture.

4. COMPUTATIONAL METHODS FOR DETECTION OF SEQUENCE VARIANT EFFECTS ON 3D GENOME FROM EXPERIMENTAL DATA

Impacts of SVs on 3D genome architecture have been directly established by experimental data. In this section, we briefly review the discoveries of disease-causing SVs and their impact on 3D genome, and then discuss computational approaches to detect effects from genomic variants observed in biosamples when experimental data are available, including detecting common variant effects from a population of samples and detecting large SV effects from individual samples.

Several early studies experimentally examined the 3D genome effects of SVs implicated in diseases. Targeted 3C techniques are typically used in these studies, including capture Hi-C, which generates a high-resolution contact map for a subset of genomic regions, and 4C-seq (circular 3C sequencing), which measures interactions with a selected genomic location of interest. One of the early findings was that several disease-causing SVs at the *EPHA4* locus involve misexpression of different genes due to 3D genome reorganization (76). A deletion near the *PAX3* gene results in the fusion of two neighboring TADs (also called TAD fusion), leading to an ectopic interaction between *PAX3* and an enhancer cluster, which results in the abnormal expression and congenital limb malformation called brachydactyly (76). A different inversion at this *EPHA4* locus results in the same enhancer cluster being relocated into the neighborhood of *WNT6*, causing abnormal limb development (76). This mechanism of reorganizing parts of TADs that do not naturally belong together and causing enhancer adaptation was called TAD shuffling (1). Another different type of 3D genome disruption was the creation of a new domain due to duplication (also called neo-TADs). For example, a duplication of noncoding DNA near *SOX9* leads to the formation of a neo-TAD that contains *KCNJ2*. This causes *KCNJ2* to interact with duplicated enhancers originally in the *SOX9* TAD, leading to Cooks syndrome (55). Subsequent studies also demonstrated that TAD disruption represents a common phenomenon in cancer development (77, 78). We refer interested readers to a previous review (1) for more discussion of the impacts of disease-causing SVs on chromatin 3D architecture. All these examples have established that the disruption of chromatin organization by SVs is an important genetic mechanism for disease.

Beyond targeted studies of variants of clinical interest, to systematically understand the genetic architecture of the 3D genome and accelerate the discovery of disease-relevant variants, researchers need computational methods that can unbiasedly and systematically detect the impact of sequence variants on 3D genome organization. When it is feasible to obtain both experimental genome-wide 3D genome measurements and genotype information, several approaches have been developed for using experimental data to detect large-scale variant effects on the 3D genome (**Figure 2**). For common variants, powerful approaches can be applied to detect the quantitative trait locus (QTL) using a population of paired genotype and 3D genome data. Computational methods have been developed to systematically identify SVs with strong impact in a population and analyze the Hi-C data to reconstruct their effects. In later sections, we introduce computational methods that predict the 3D genome effects of variants using only the genomic sequence (**Figures 3** and **4**), which is complementary to the experimental data–based approaches which we introduce below.

4.1. Population-Based Quantitative Trait Locus Analysis of the 3D Genome Phenotypes

QTL analysis is a statistical genetics approach that identifies significant associations between genotypes (most commonly, SNPs) and quantitative phenotypes, which can be high-dimensional molecular-level phenotypes such as gene expression. For each identified QTL variant, usually one allele (e.g., T) is significantly associated with a higher or lower score of the quantitative phenotype compared to the other allele (e.g., G). With a large enough sample, QTL analysis has the power to identify even variants with weak effects. Gorkin et al. (79) performed a QTL analysis to systematically characterize the 3D genome organization–associated variants across a population of unrelated individuals. Even though their sample was moderate in size (20 individuals and 11 unrelated Yoruba individuals were used for QTL discovery), the study demonstrated the feasibility of detecting QTLs associated with 3D chromatin conformation. The authors used lymphoblastoid cell line Hi-C data derived from individuals whose genetic variation has been cataloged by the 1000 Genomes Consortium.

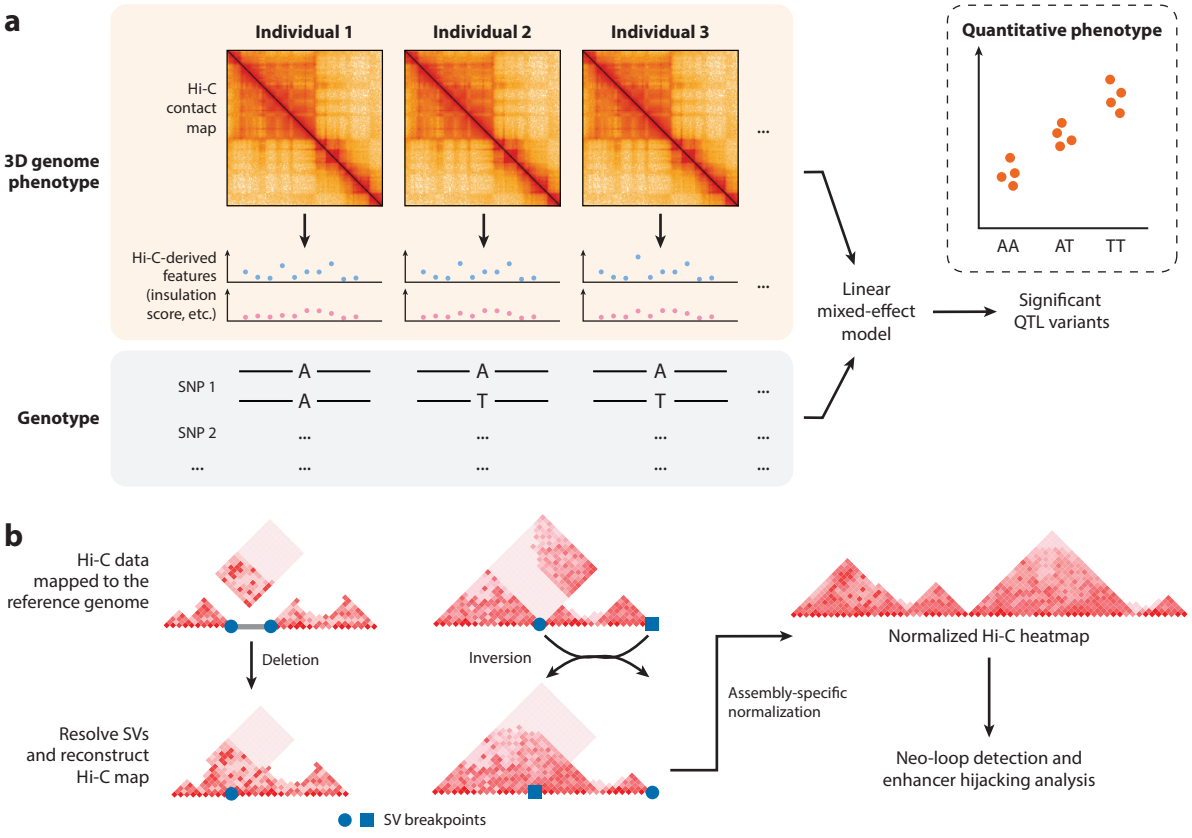


Figure 2

Computational approaches for detecting SV effects on the 3D genome. (*a*) A population-based QTL approach can be applied when the genotype and 3D genome datasets are available for a population of individuals. Multiple quantitative phenotypes can be extracted from the 3D genome dataset, such as pairwise contacts, insulation scores, and directionality index. Such quantitative scores can be modeled with genotype data through the standard linear mixed-effect modeling approach for QTL detection. (*b*) The detection of large SV effects on the 3D genome from individual samples. Resolving SVs based on Hi-C data and reconstructing the Hi-C map allow for large-scale identification and interpretation of SV effects. Additional steps can be taken to normalize for differences in coverage and identify neo-loop formation and enhancer hijacking events. Abbreviations: Hi-C, high-throughput chromosome conformation capture; QTL, quantitative trait locus; SNP, single-nucleotide polymorphism; SV, structural variant.

To define the molecular phenotypes for their QTL analysis, in addition to a 40-kb-resolution contact matrix, Gorkin et al. extracted several features for each 40-kb bin from Hi-C data: the insulation score, which measures the interaction spanning two opposite sides of the position of interest; the directionality index, which measures imbalance in the directionality of contacts from a given region of interest (i.e., upstream or downstream); the first principal component of the contact matrix, which measures chromatin compartmentalization; and the frequently interacting regions score, which is defined to measure the amount of nonself interaction for each region. For computational procedures of processing and normalizing Hi-C-type data, we refer readers to previous reviews on this topic (80, 81). Each of these features was used as a 3D genome phenotype for further QTL analysis and resulted in the identification of several types of QTLs. For each phenotype, a subset of 40-kb bins was first selected to reduce the search space, and SNPs located within the selected bins were tested for association with the phenotype using a linear mixed-effect

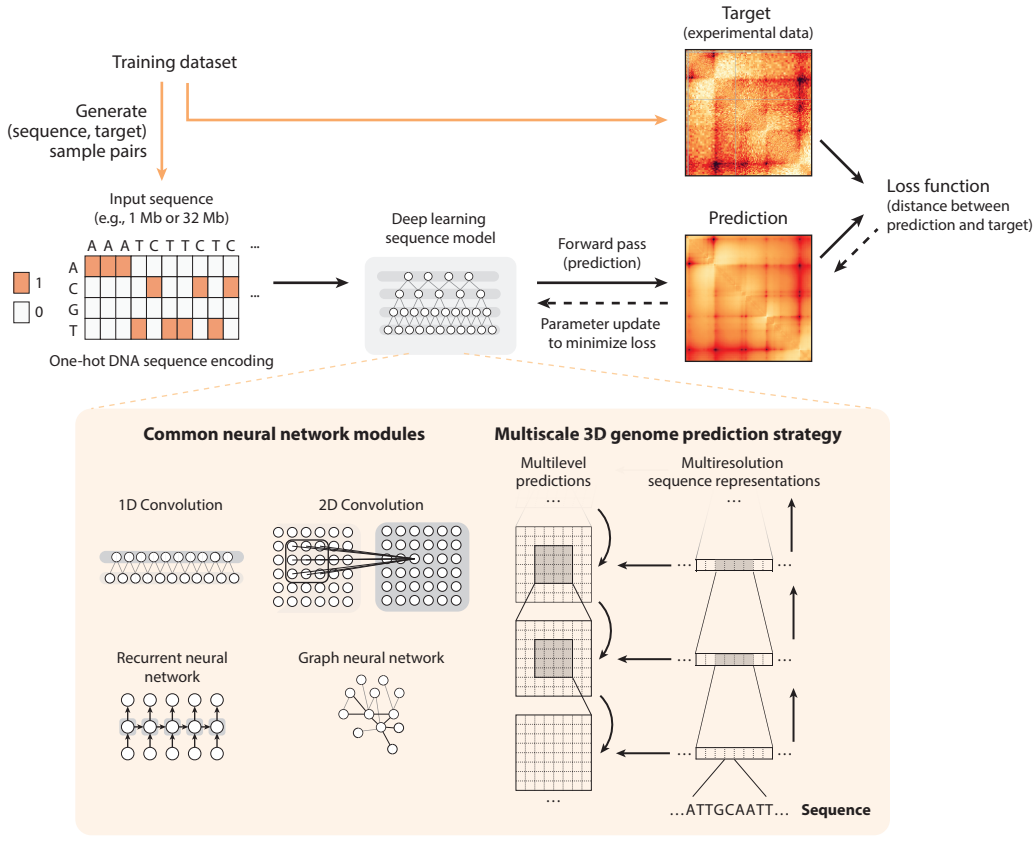


Figure 3

Sequence-based prediction of 3D genome structure. Training the sequence models of the 3D genome requires a training dataset that contains or can generate pairs of input (DNA sequences) and corresponding target [typically a Hi-C (high-throughput chromosome conformation capture) contact matrix–like representation of genome interaction]. The DNA sequence is typically encoded numerically by one-hot encoding. The deep learning sequence models take the sequence as input and predict the corresponding 3D genome interactions in the range of the sequence. A loss function will be defined to compute the loss from the prediction and the target, which typically measures the distance between prediction and target. The sequence model is trained by optimization algorithms that minimize the loss using the gradient of the loss function.

model, which is a standard approach for QTL analysis (82). Overall, this study illustrated that the QTL approach can be effectively applied to identify genetic variation within the population that influences multiple features of 3D chromatin conformation. Even though currently the cost and difficulty of collecting large Hi-C datasets still limit its application to a larger number of individuals and its ability to detect variant effects with greater sensitivity, such a population-based approach could help resolve the contribution of the 3D genome to common variants' associations with disease and traits identified from GWAS.

4.2. Large-Scale Detection of Structural Variant Effects on the 3D Genome from a Single Sample

On the other end of the variant effect spectrum, for variants with strong effects on the 3D genome such as large SVs that are often rare, 3D genome effects can be detected with even a single sample. This is because the effect of large SVs is typically dramatic and involves broad genomic regions,

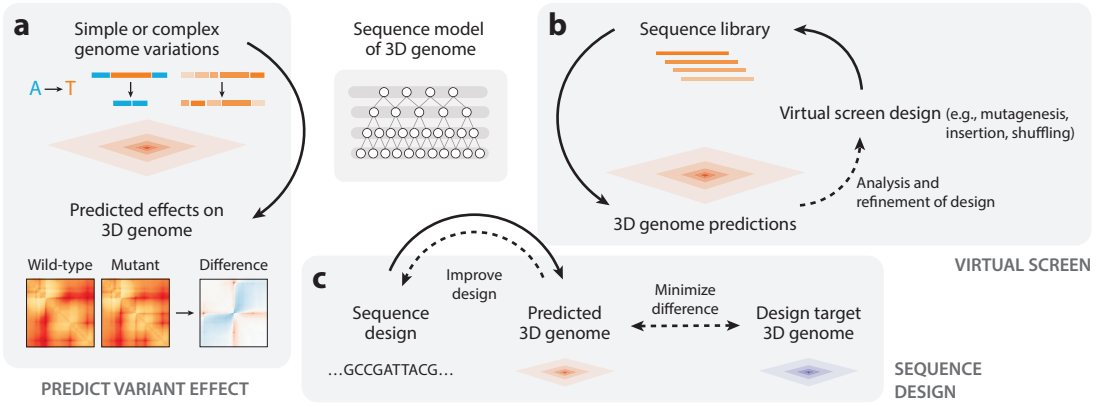


Figure 4

Examples of applications of sequence models of the 3D genome, including predicting variant effects on the 3D genome (*a*), performing virtual genetic screens to probe the sequence-based mechanism of 3D genome organization (*b*), and designing sequences with the desired 3D genome properties (*c*).

such as the gain of strong interactions between genomic locations that were previously not interacting. Analysis of individual SVs has been clearly demonstrated in the disease SV studies discussed above. However, identifying SVs and quantitatively measuring their effects on a large scale is still challenging and requires the development of computational algorithms. Fortunately, the spatial proximity information measured by Hi-C is directly informative for SV detection, and a correctly reassembled genome can be verified with Hi-C contact signals decreasing smoothly with genomic distance (83–85). Resolving the SV then allows for direct reconstruction of the genome interaction patterns in the rearranged genome, when contacts spanning the breakpoints mostly originate from the chromosome that carries the SVs, which is typical for large SVs.

Genome-wide integrative detection of simple or complex SVs and the identification of their effects on 3D genome reorganization, including enhancer hijacking events based on Hi-C data, have been made possible by computational frameworks such as NeoLoopFinder (85). The method starts from Hi-C data and SV breakpoints as input and analyzes SV effects on the 3D genome organization in several steps. The method reconstructs the Hi-C matrix for the rearranged genomic regions by resolving the correct order of the genome fragments using a shortest-path algorithm based on the Hi-C data. Normalization steps are performed before and after the reconstruction to correct for the variation of Hi-C contact signal strength due to copy number or allele frequency. With the reconstructed and normalized Hi-C map, NeoLoopFinder can then identify the newly formed chromatin loops by the SVs and annotate the enhancer hijacking events based on the enhancer histone mark profile such as H3K27ac.

While we expect that further improvements of experimental and computational methods will significantly expand our ability to measure the sequence variations effects on the 3D genome, current costs and the need for biological samples that carry the variant of interest still limit the scalability of experimental data–based measurement of 3D genome effects. However, computational modeling of sequence dependencies for the 3D genome that we discuss below will allow sequence variant effects to be predicted on a large scale in order to prioritize variants of interest for further study, and such modeling can be applied more broadly since it does not require biological samples carrying the sequence variations. Combining computational modeling with the experimental approach can provide a powerful means of generating and testing hypotheses about sequence-based mechanisms of 3D genome organization.

5. SEQUENCE-BASED PREDICTION OF 3D GENOME ORGANIZATION

The availability of genome-wide, high-resolution 3D genome structure datasets has allowed machine learning approaches to capture sequence–3D genome structure dependencies. Machine learning has the potential to learn complex sequence pattern dependencies in a quantitative manner that are otherwise challenging to resolve. If the machine learning models have adequately captured the causal dependencies from sequence to 3D structure, they can be applied to predict the structural consequences of any sequence change and can help researchers understand the sequence basis of 3D genome organization. Specifically, such models allow for the prediction of any sequence variant's effects on the 3D genome organization in the following manner: Two sequences are generated that differ only by a genomic variant, where each sequence carries one of the alleles, then the 3D organization of each sequence is predicted from the model, and finally the predicted influences of the two variants on chromatin architecture are compared. Additionally, these models allow any mutation to the input sequence to be introduced and their influences analyzed, making them highly flexible tools for the study of the 3D genome. In this section, we discuss the machine learning methods that have been developed for sequence-based prediction of 3D genome structure.

While the predictive power of a machine learning model does not generally guarantee that it is capturing the causal relationships between the input and the predictive target (or that its predictions reflect causal consequences), using the genomic sequence as input has a unique advantage in learning causal relationships. This is because the genomic sequence in most scenarios is likely the cause rather than the consequence of downstream biological processes including 3D genome organization. Genomic sequences remain mostly unchanged postconception and are thus affected only through the process of evolution. Thus, a sequence-based model likely has the advantage of capturing causal dependencies, even though it is still important to rigorously evaluate its ability to predict causal consequences of sequence change. In contrast, other genomic features such as epigenomic profiles and gene expression are often directly affected by the 3D genome organization. Machine learning models that use these data as input thus may be using the consequences of 3D genome organization to predict its cause and are thus not ideal for inferring causal relationships (86–92).

Therefore, here we focus on predictive methods for the 3D genome organization that are sequence based (i.e., that use only DNA sequences as input). Many 3D genome prediction methods also include other types of input data, such as chromatin profiles and gene expression (86–92). While these methods are useful for other applications such as inferring 3D genome organization in cell types for which such data are not available, we do not cover these methods in detail because they are not designed for predicting sequence and sequence variation effects.

Machine learning approaches that we consider here learn to extract patterns in the input (such as genomic sequences) to make predictions about the target (in the form of Hi-C contact matrices). The model training process usually requires a training dataset, an optimization algorithm, and a loss function for measuring error between targets and predictions (**Figure 3**). A typical example of a training dataset for sequence-based 3D genome prediction includes a DNA sequence as input and a Hi-C contact map for the corresponding region as output. The training process iterates over a training dataset and learns model parameters by minimizing the error function through stochastic gradient descent or related optimization algorithms. To evaluate the model performance, researchers need to apply the model on input data not used during model training and compare its predictions with the targets.

Many machine learning models for sequence-based prediction of chromatin structure are deep learning–based models. Deep learning models can generally be described as differentiable

nonlinear functions with learnable parameters that transform the input into output predictions (93, 94). They are often composed of sequential modular building blocks called layers. The first layer receives the initial input, the last layer outputs the final prediction, and the outputs of internal layers can be considered as representations of the input data that are informative for the prediction task after training. A typical learnable layer executes a linear transformation of the output from the previous layer followed by a nonlinear transformation called the activation function (93, 94). Common neural network layers or modules with learnable parameters include convolutional layers, fully connected layers, recurrent neural networks, transformers, and graph neural networks. Other layers do not contain learnable parameters and represent a specific transformation such as maximum pooling or dropout (93, 94). Deep learning models that primarily use convolutional layers as learnable layers are also called convolutional neural networks. The design of deep learning models is highly flexible and allows for different structures, as long as the entire computation is differentiable. Training deep learning models usually requires defining a loss function, such as a function that computes a score to measure the difference between the prediction and the experimental observation; parameters in such deep learning models are optimized to minimize the loss function during the training process. We refer readers to deep learning textbooks for more about the topic (93, 94).

Deep learning models for genomic sequences were first introduced in DeepBind (95) and DeepSEA (96) for tasks such as predicting transcription factor binding, RNA–protein binding, chromatin accessibility, and histone marks. Both methods use convolutional network architectures. For example, DeepSEA utilized a hierarchy of 1D convolutional layers and maximum pooling layers to allow for a wider context sequence window of 1,000 bp. The use of 1D convolution is based on the fact that genomic sequences have one spatial dimension that satisfies translational invariance (unlike 2D images, which have two dimensions that satisfy translational invariance)—in other words, sequence patterns observed at different locations of the sequence share similar interpretations. Other neural network architectures for predicting chromatin properties have also been proposed, such as Basset, which also uses a convolutional network architecture (97), and DanQ, which introduced a recurrent neural network layer (98). Deep learning sequence models have since been improved to utilize even longer sequences, notably Basenji (∼131-kb input with a receptive field of ∼40 kb) (99) and SpliceAI (10-kb input) (100), both of which used many layers of dilated convolution layers to increase the size of the receptive field. Sequence models based on convolutional network architecture have also been developed to make simultaneous predictions for multiple positions in the sequence, including at the single–base pair level (99, 101).

Modeling chromatin organization from sequence can also be formulated as a sequence-based machine learning problem. Early sequence-based methods such as PEP (102) predicted chromatin interactions between two genomic regions, using two separate sequences at each region as input. The PEP model used sequence features with boosted decision trees, and many later methods have adopted the deep learning sequence modeling approach, combining the representation of two sequences as input, including EPIANN (103), SPEID (104), DeepTACT (105), SEPT (106), DeepMILO (107), and ChiNN (108). All these methods take two DNA sequences as input and predict their interaction as a binary event using neural networks. The length of input DNA sequences can range from one base pair to several kilobases. Since these methods take only a pair of sequences as input that are relatively short, consider the range of the interactions, and ignore the sequence between the input sequences, their performance has been limited because the sequence in-between can play a major role in determining chromatin interactions such as TAD boundaries.

Since 3D genome organization can involve very broad genomic regions, it is necessary to utilize very long input sequences. DeepC (109) and Akita (110) were the first 3D genome prediction

methods that used up to ~1-Mb sequences as input. Importantly, both of these methods use the whole sequence to make all-to-all predictions for the ~1-Mb input region in genomic bins of length 5 kb (DeepC) or 2 kb (Akita). The all-to-all predictions are naturally represented as matrices similar to Hi-C contact matrices, which we here term genome interaction maps. Both approaches use hierarchical convolutional network architecture and maximum pooling and dilated convolution layers to support a broad sequence context. DeepC uses 1D convolutional layers and a fully connected layer as the output layer, and Akita uses 2D convolutional layers on top of 1D convolutional layers. Akita converts internal representations from 1D convolutional layers to 2D (the number of positions times the number of positions) by computing the average of representations from all pairs of spatial positions and applies multiple 2D convolutional layers to compute the prediction. In addition, DeepC uses a transfer learning approach, whereby a network trained on chromatin profile prediction tasks similar to DeepSEA was taken without the final layers and additional layers were then added to train the network for the 3D genome interaction tasks. These methods have shown high concordance between predicted and experimental genomic contact maps. These models can capture a variety of chromatin organization patterns at submegabase scale such as Hi-C peaks (or dots or loops) and TAD boundaries.

Models such as DeepC and Akita that use sequences up to 1 Mb for sequence-based prediction are still not sufficient for predicting larger-scale organization of the 3D genome or considering the larger sequence context. The deep learning method Orca (111) further extended 3D genome prediction to the whole-chromosome scale using a multiscale modeling approach. Orca can predict interactions for up to 256 Mb of sequence, which is sufficient to model the longest human chromosome and predict interchromosomal interactions. Orca generates a series of predictions that progressively zoom into finer regions to generate predictions at multiple scales and resolutions. This is enabled by a 1D convolutional network multiresolution sequence encoder that generates sequence representations at multiple resolutions, and a series of cascading 2D convolutional network decoders that predict genome contact matrices from the corresponding level of sequence representation and the upper-level prediction. Multiresolution genome interaction map predictions can be obtained by training multiple decoders that utilize sequence representations at different resolutions, and information from a large sequence context can be passed down to finer-level predictions. The multiscale modeling approach enables predictions from the kilobase scale to the whole-chromosome scale and is in principle capable of modeling all scales of interactions measured by Hi-C-type techniques.

All three methods have shown high concordance between the predicted and experimental genomic contact maps. These models can capture a variety of chromatin organization patterns such as chromatin loops and TAD boundaries. Overall, Akita, DeepC, and Orca are expected to capture complex DNA sequence patterns that regulate 3D genome organization.

Altogether, current sequence-based methods provide a foundation for multiscale 3D genome organization modeling from sequence. However, these approaches still have limitations and room for further improvements. Some genome interaction maps predicted by these tools still diverge from experimental measurement, and these discrepancies cannot be justified by experimental noise. Additionally, all machine learning–based methods face challenges in learning sequence-based dependencies if such sequence patterns appear in very few or only one genomic locus. Thus, further development of neural network architectures and experimental techniques will help advance sequence-based methods toward higher resolution and cover more diverse cell types and species. We expect current and future sequence models to expand our abilities to prioritize sequence variants, understand mechanisms behind genome folding, model enhancer–promoter interactions, and more. In the next section, we discuss these aspects in more detail.

6. OPPORTUNITIES AND CHALLENGES PRESENTED BY COMPUTATIONAL METHODS OF 3D GENOME PREDICTION

We envision that the capability to accurately predict 3D genome organization from sequence will accelerate experimental efforts on a broad range of topics, from the impact of sequence variations on genome folding to the sequence determinants of genome folding across spatial scales. Since computational predictions generally have an advantage in their scalability and the input for computational models can be in principle any sequence, they are flexible and adaptable and allow for diverse applications. Deep learning models that capture sequence–3D genome dependencies likely implicitly model the mechanism of 3D genome organization, thus allowing them to be used as a hypothesis-generation engine that can provide high-quality testable hypotheses. Here we discuss a few applications of such predictions, including both explored and new directions: predicting effects of SVs, probing the sequence-based mechanism of 3D genome organization, and designing genomic sequences with the desired structural properties (**Figure 4*a–c***).

SVs can have profound impacts on 3D genome structure, which have been demonstrated to cause severe diseases, as discussed above. It is still costly to measure the effect of SVs on a large scale; thus, prediction from computational methods offers the ability to analyze a large number of genomic variants, including SVs, and prioritize ones of interest for further experimental study. Akita, DeepC, and Orca have all demonstrated the ability to predict the consequences of structural variations for 3D chromosome organization (**Figure 4*a***). To show this possibility, both DeepC and Akita predicted the effects of an ∼25-kb genetically engineered deletion at the *Lmo2* locus (109, 110). Sequence models can in principle analyze any kind of sequence variation, with the only hard limitation being the length of the sequence input. Akita and DeepC can predict from a sequence up to 1 Mb in length, and Orca allows an even broader range of SVs by chromosome-scale input and multiscale prediction (109–111). Orca was tested on more than a dozen SVs ranging from 0.3 kb to 80 Mb in size, for which the impact on the 3D architecture was also experimentally measured (55, 76–78, 112, 113). All three methods showed high concordance with experimental observations of chromatin organization, demonstrating that sequence models are promising tools for understanding SVs' effect on the 3D genome. For further development, because sequence model–based prediction can be applied to arbitrarily complex variants as long as the DNA sequence can be constructed computationally, we expect it to be applicable to predicting the effects of complex genome rearrangement events. Sequence models will likely also be useful for predicting the effects of combinations of variants on the same haplotype. Continued validation of model prediction capability is necessary as more SVs' 3D genome effect data become available. Sequence model predictions can also in turn help SV detection from experimental Hi-C data by providing the expected genome interaction map to help detect deviation from expectation due to genome rearrangement.

Sequence models can be applied to predict 3D genome effects of any mutation, which allow for applications beyond observed variants. Predicting effects of any mutation from sequence is often called in silico mutagenesis, and it has been widely used to interpret *cis*-regulatory sequence patterns underlying specific regulatory properties predicted by deep learning sequence models (96, 97, 99, 101), including sequence models for 3D genome organization (109–111). Building upon the idea of in silico mutagenesis, sequence models can also be considered as an "in silico genome observatory" platform to perform virtual genetic screen experiments in order to generate and test hypotheses of sequence-based 3D genome organization (111) (**Figure 4*b***). Similar to genetic screen experiments, one can design the library of mutations introduced according to the problem of interest, introduce the mutations computationally by generating sequences that carry those mutations, and then predict 3D genome effects using sequence models. This approach can

be a valuable hypothesis-generation engine. Due to the advantages of computational approaches, the virtual screen is fast and flexible and allows for fast iterations to improve the design. For example, a series of virtual insertion and permutation screens with Orca has been used to propose a sequence-based model of chromatin compartments, whereby the transcription start sites are a key driver of compartment A, and compartment B can form passively without any specific sequence pattern from extended sequence without compartment A activity (111). At a more local submegabase scale, DeepC, Akita, and Orca detected the importance of CTCF binding sites and Orca additionally suggested that the vast majority of motifs with the strongest tier of 3D genome impact either overlap CTCF motifs or lie within a small neighborhood of them. Akita reported that CTCFL motifs have the second largest effect on chromatin organization. All methods also predict boundaries not associated with CTCF sites, and Orca-based analysis suggested that non-CTCF motifs with moderate 3D genome effects show strong cell-type specificity.

These sequence models also enable the design of sequences with certain 3D genome properties, allowing engineered sequences, by edits or from scratch, to possess any desired properties that can be predicted from sequence (**Figure 4*c***). Sequence design guided by predictive sequence models has already been explored for other applications (114–117). For example, recent work explored combining a predictive model with a generative model for generating a regulatory DNA sequence that is capable of driving expression (118). Such approaches can in principle be developed for sequence models of the 3D genome. Thus, the development of sequence models opens up the possibility of utilizing the sequence models to design sequences with a desired 3D genome structure.

Finally, while sequence models have so far demonstrated high-quality predictions when given unseen sequences, rigorous evaluations with experimental data are required for any new applications. Moreover, it is worth noting that supervised learning models can produce erroneous predictions when given out-of-distribution inputs that are distinct from the training sequence. The power of sequence models to generalize for sequences with large deviations from the natural statistics of human genomic sequences (e.g., mononucleotide or dinucleotide frequencies) is still unknown and a subject for future studies.

7. CONCLUSIONS AND OUTLOOK

Understanding how genomic sequence affects the 3D genome is critical for deciphering the mechanisms of sequence variations underlying various diseases and provides a unique window into the mechanisms of 3D genome organization. The continued development of computational methods will be pivotal in the study of genomic variation effects on the 3D genome, including detecting simple and complex structural variation, measuring sequence variation effects from experimental 3D genome data, and directly modeling the 3D genome from sequence. In particular, we envision that modeling the 3D genome variation effects will accelerate hypothesis generation, help prioritize experimental efforts, and enable that analysis of sequence variation on a large scale. The computational models will continue to improve with methodological improvements, likely from both deep learning and biophysical modeling. The constant improvement of experimental techniques will also enable higher-resolution predictions and more comprehensive evaluation of model capabilities. We expect such developments will improve our understanding of the genetic basis of 3D genome organization and its relevance to diseases, including those caused by rare genetic variations. Sequence-based variant effect prediction also has the potential to assist in the diagnosis of diseases caused by genome variants, especially large SVs.

The development of user-friendly and open-source software is important to enable wider application of sequence-based models of the 3D genome. We strongly believe in the importance

of releasing open-source code, models, and tutorials through public repositories. Moreover, because code repositories are typically intended for experienced users familiar with command line usage and often require a basic level of programming expertise, for the methods to be accessible by the broader research community, user-friendly interfaces that require little or no installation such as web servers will be important, especially as more diverse applications are enabled. Continued advancement in modeling capability and accessibility will make computational approaches a driving force for illuminating the interplay and interactions between genomic sequence and the 3D genome architecture.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported by the Cancer Prevention and Research Institute of Texas grant RR190071, the National Institutes of Health grant DP2GM146336, and the UT Southwestern Endowed Scholars program.

LITERATURE CITED

1. Spielmann M, Lupiáñez DG, Mundlos S. 2018. Structural variation in the 3D genome. *Nat. Rev. Genet.* 19(7):453–67
2. Bonev B, Cavalli G. 2016. Organization and function of the 3D genome. *Nat. Rev. Genet.* 17(11):661–78
3. van Steensel B, Furlong EEM. 2019. The role of transcription in shaping the spatial organization of the genome. *Nat. Rev. Mol. Cell Biol.* 20(6):327–37
4. Cullen KE, Kladde MP, Seyfred MA. 1993. Interaction between transcription regulatory regions of prolactin chromatin. *Science* 261(5118):203–6
5. Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. *Science* 295(5558):1306–11
6. Van De Werken HJG, Landan G, Holwerda SJB, Hoichman M, Klous P, et al. 2012. Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat. Methods* 9(10):969–72
7. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, et al. 2006. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* 16(10):1299–309
8. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, et al. 2006. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* 38(11):1348–54
9. Kempfer R, Pombo A. 2020. Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.* 21(4):207–26
10. Jerkovic I, Cavalli G. 2021. Understanding 3D genome organization by multidisciplinary methods. *Nat. Rev. Mol. Cell Biol.* 22(8):511–28
11. Hsieh T-HS, Weiner A, Lajoie B, Dekker J, Friedman N, Rando OJ. 2015. Mapping nucleosome resolution chromosome folding in yeast by Micro-C. *Cell* 162(1):108–19
12. Hsieh T-HS, Fudenberg G, Goloborodko A, Rando OJ. 2016. Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome. *Nat. Methods* 13(12):1009–11
13. Krietenstein N, Abraham S, Venev SV, Abdennur N, Gibcus J, et al. 2020. Ultrastructural details of mammalian chromosome architecture. *Mol. Cell* 78(3):554–65.e7
14. Akgol Oksuz B, Yang L, Abraham S, Venev SV, Krietenstein N, et al. 2021. Systematic evaluation of chromosome conformation capture assays. *Nat. Methods* 18(9):1046–55
15. Hua P, Badat M, Hanssen LLP, Hentges LD, Crump N, et al. 2021. Defining genome architecture at base-pair resolution. *Nature* 595(7865):125–29

16. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, et al. 2009. An oestrogen-receptor-α-bound human chromatin interactome. *Nature* 462(7269):58–64
17. Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, et al. 2016. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* 13(11):919–22
18. Fang R, Yu M, Li G, Chee S, Liu T, et al. 2016. Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res.* 26(12):1345–48
19. Nagano T, Lubling Y, Yaffe E, Wingett SW, Dean W, et al. 2015. Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. *Nat. Protoc.* 10(12):1986–2003
20. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, et al. 2013. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 502(7469):59–64
21. Nagano T, Lubling Y, Várnai C, Dudley C, Leung W, et al. 2017. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* 547(7661):61–67
22. Beagrie RA, Scialdone A, Schueler M, Kraemer DCA, Chotalia M, et al. 2017. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* 543(7646):519–24
23. Quinodoz SA, Ollikainen N, Tabak B, Palla A, Schmidt JM, et al. 2018. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* 174(3):744–57.e24
24. Koch L. 2018. Chromosome interaction hubs around nuclear bodies. *Nat. Rev. Genet.* 19:470–71
25. Fiorillo L, Musella F, Conte M, Kempfer R, Chiariello AM, et al. 2021. Comparison of the Hi-C, GAM and SPRITE methods using polymer models of chromatin. *Nat. Methods* 18(5):482–90
26. Ulahannan N, Pendleton M, Deshpande A, Schwenk S, Behr JM, et al. 2019. Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure. bioRxiv 10.1101/833590. **https://doi.org/10.1101/833590**
27. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, et al. 2012. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* 148(3):458–72
28. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, et al. 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485(7398):381–85
29. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, et al. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485(7398):376–80
30. Norton HK, Emerson DJ, Huang H, Kim J, Titus KR, et al. 2018. Detecting hierarchical genome folding with network modularity. *Nat. Methods* 15(2):119–22
31. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159(7):1665–80
32. Nasmyth K, Haering CH. 2009. Cohesin: its roles and mechanisms. *Annu. Rev. Genet.* 43:525–58
33. Riggs AD. 1990. DNA methylation and late replication probably aid cell memory, and type I DNA reeling could aid chromosome folding and enhancer function. *Philos. Trans. R. Soc. Lond. B* 326(1235):285–97
34. Alipour E, Marko JF. 2012. Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res.* 40(22):11202–12
35. Sanborn AL, Rao SSP, Huang S-C, Durand NC, Huntley MH, et al. 2015. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *PNAS* 112(47):E6456–65
36. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. 2016. Formation of chromosomal domains by loop extrusion. *Cell Rep.* 15(9):2038–49
37. Rao SSP, Huang SC, St. Hilaire BG, Engreitz JM, Perez EM, et al. 2017. Cohesin loss eliminates all loop domains. *Cell* 171(2):305–20.e24
38. Bantignies F, Roure V, Comet I, Leblanc B, Schuettengruber B, et al. 2011. Polycomb-dependent regulatory contacts between distant Hox loci in *Drosophila*. *Cell* 144(2):214–26
39. Palstra R-J, Tolhuis B, Splinter E, Nijmeijer R, Grosveld F, de Laat W. 2003. The β-globin nuclear compartment in development and erythroid differentiation. *Nat. Genet.* 35:190–94
40. Schoenfelder S, Sexton T, Chakalova L, Cope NF, Horton A, et al. 2010. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet.* 42(1):53–61

41. Pombo A, Dillon N. 2015. Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.* 16(4):245–57
42. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950):289–93
43. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. 2011. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* 30:90–98
44. Wijchers PJ, Krijger PHL, Geeven G, Zhu Y, Denker A, et al. 2016. Cause and consequence of tethering a subTAD to different nuclear compartments. *Mol. Cell* 61(3):461–73
45. Wang S, Su J-H, Beliveau BJ, Bintu B, Moffitt JR, et al. 2016. Spatial organization of chromatin domains and compartments in single chromosomes. *Science* 353(6299):598–602
46. Gu H, Harris H, Olshansky M, Eliaz Y, Krishna A, et al. 2021. Fine-mapping of nuclear compartments using ultra-deep Hi-C shows that active promoter and enhancer elements localize in the active A compartment even when adjacent sequences do not. bioRxiv 2021.10.03.462599. **https://doi.org/10.1101/2021.10.03.462599**
47. Lichter P, Cremer T, Borden J, Manuelidis L, Ward DC. 1988. Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries. *Hum. Genet.* 80(3):224–34
48. Cremer T, Cremer M. 2010. Chromosome territories. *Cold Spring Harb. Perspect. Biol.* 2(3):a003889
49. Sun HB, Shen J, Yokota H. 2000. Size-dependent positioning of human chromosomes in interphase nuclei. *Biophys. J.* 79(1):184–90
50. Bickmore WA, van Steensel B. 2013. Genome architecture: domain organization of interphase chromosomes. *Cell* 152(6):1270–84
51. Sexton T, Cavalli G. 2015. The role of chromosome domains in shaping the functional genome. *Cell* 160(6):1049–59
52. Therizols P, Illingworth RS, Courilleau C, Boyle S, Wood AJ, Bickmore WA. 2014. Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. *Science* 346(6214):1238–42
53. Gonzalez-Sandoval A, Towbin BD, Kalck V, Cabianca DS, Gaidatzis D, et al. 2015. Perinuclear anchoring of H3K9-methylated chromatin stabilizes induced cell fate in *C. elegans* embryos. *Cell* 163(6):1333–47
54. Lupiáñez DG, Spielmann M, Mundlos S. 2016. Breaking TADs: how alterations of chromatin domains result in disease. *Trends Genet.* 32(4):225–37
55. Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, et al. 2016. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* 538(7624):265–69
56. Lettice LA, Daniels S, Sweeney E, Venkataraman S, Devenney PS, et al. 2011. Enhancer-adoption as a mechanism of human developmental disease. *Hum. Mutat.* 32(12):1492–99
57. Krijger PHL, de Laat W. 2016. Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. Cell Biol.* 17:771–82
58. Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 12(5):363–76
59. Ho SS, Urban AE, Mills RE. 2020. Structural variation in the sequencing era. *Nat. Rev. Genet.* 21(3):171–89
60. Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. *Nat. Rev. Genet.* 7(2):85–97
61. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. 2018. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* 19(6):329–46
62. Zhang F, Carvalho CMB, Lupski JR. 2009. Complex human chromosomal and genomic rearrangements. *Trends Genet.* 25(7):298–307
63. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, et al. 2020. Patterns of somatic structural variation in human cancer genomes. *Nature* 578(7793):112–21
64. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* 526(7571):75–81

65. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, et al. 2007. Strong association of de novo copy number mutations with autism. *Science* 316(5823):445–49
66. de Vries BBA, Pfundt R, Leisink M, Koolen DA, Vissers LELM, et al. 2005. Diagnostic genome profiling in mental retardation. *Am. J. Hum. Genet.* 77(4):606–16
67. Aganezov S, Goodwin S, Sherman RM, Sedlazeck FJ, Arun G, et al. 2020. Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Res.* 30(9):1258–73
68. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* 14(2):125–38
69. Hyde CL, Nagle MW, Tian C, Chen X, Paciga SA, et al. 2016. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat. Genet.* 48(9):1031–36
70. Zhao W, Rasheed A, Tikkanen E, Lee J-J, Butterworth AS, et al. 2017. Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nat. Genet.* 49(10):1450–57
71. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921
72. Int. HapMap Consort. 2005. A haplotype map of the human genome. *Nature* 437(7063):1299–320
73. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, et al. 2019. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47(D1):D1005–12
74. Martincorena I, Campbell PJ. 2015. Somatic mutation in cancer and normal cells. *Science* 349(6255):1483–89
75. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. 2019. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* 20(8):467–84
76. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, et al. 2015. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* 161(5):1012–25
77. Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, et al. 2016. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* 351(6280):1454–58
78. Weischenfeldt J, Dubash T, Drainas AP, Mardin BR, Chen Y, et al. 2017. Pan-cancer analysis of somatic copy-number alterations implicates *IRS4* and *IGF2* in enhancer hijacking. *Nat. Genet.* 49:65–74
79. Gorkin DU, Qiu Y, Hu M, Fletez-Brant K, Liu T, et al. 2019. Common DNA sequence variation influences 3-dimensional conformation of the human genome. *Genome Biol.* 20:255
80. Dekker J, Marti-Renom MA, Mirny LA. 2013. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* 14(6):390–403
81. Pal K, Forcato M, Ferrari F. 2019. Hi-C analysis: from data generation to integration. *Biophys. Rev.* 11:67–78
82. Lynch M, Walsh B. 1998. *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer
83. Dixon JR, Xu J, Dileep V, Zhan Y, Song F, et al. 2018. Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* 50:1388–98
84. Wang S, Lee S, Chu C, Jain D, Kerpedjiev P, et al. 2020. HiNT: a computational method for detecting copy number variations and translocations from Hi-C data. *Genome Biol.* 21:73
85. Wang X, Xu J, Zhang B, Hou Y, Song F, et al. 2021. Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. *Nat. Methods* 18(6):661–68
86. He B, Chen C, Teng L, Tan K. 2014. Global view of enhancer-promoter interactome in human cells. *PNAS* 111(21):E2191–99
87. Roy S, Siahpirani AF, Chasman D, Knaack S, Ay F, et al. 2015. A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Res.* 43(18):8694–712
88. Zhu Y, Chen Z, Zhang K, Wang M, Medovoy D, et al. 2016. Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.* 7:10812
89. Whalen S, Truty RM, Pollard KS. 2016. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* 48(5):488–96
90. Cao Q, Anyansi C, Hu X, Xu L, Xiong L, et al. 2017. Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nat. Genet.* 49(10):1428–36

91. Belokopytova PS, Nuriddinov MA, Mozheiko EA, Fishman D, Fishman V. 2020. Quantitative prediction of enhancer-promoter interactions. *Genome Res.* 30:72–84
92. Zhang S, Chasman D, Knaack S, Roy S. 2019. In silico prediction of high-resolution Hi-C interaction matrices. *Nat. Commun.* 10:5449
93. Goodfellow I, Bengio Y, Courville A. 2016. *Deep Learning*. Cambridge, MA: MIT Press
94. Murphy KP. 2022. *Probabilistic Machine Learning: An Introduction*. Cambridge, MA: MIT Press
95. Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33(8):831–38
96. Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* 12(10):931–34
97. Kelley DR, Snoek J, Rinn JL. 2016. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* 26(7):990–99
98. Quang D, Xie X. 2016. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* 44(11):e107
99. Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. 2018. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* 28(5):739–50
100. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, et al. 2019. Predicting splicing from primary sequence with deep learning. *Cell* 176(3):535–48.e24
101. Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, et al. 2021. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* 53(3):354–66
102. Yang Y, Zhang R, Singh S, Ma J. 2017. Exploiting sequence-based features for predicting enhancer-promoter interactions. *Bioinformatics* 33(14):i252–60
103. Mao W, Kostka D, Chikina M. 2017. Modeling enhancer-promoter interactions with attention-based neural networks. bioRxiv 10.1101/219667. **https://doi.org/10.1101/219667**
104. Singh S, Yang Y, Póczos B, Ma J. 2019. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quant. Biol.* 7(2):122–37
105. Li W, Wong WH, Jiang R. 2019. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Res.* 47(10):e60
106. Jing F, Zhang S-W, Zhang S. 2020. Prediction of enhancer-promoter interactions using the cross-cell type information and domain adversarial neural network. *BMC Bioinform.* 21:507
107. Trieu T, Martinez-Fundichely A, Khurana E. 2020. DeepMILO: a deep learning approach to predict the impact of non-coding sequence variants on 3D chromatin structure. *Genome Biol.* 21:79
108. Cao F, Zhang Y, Cai Y, Animesh S, Zhang Y, et al. 2021. Chromatin Interaction Neural Network (ChINN): a machine learning-based method for predicting chromatin interactions from DNA sequences. *Genome Biol.* 22:226
109. Schwessinger R, Gosden M, Downes D, Brown RC, Oudelaar AM, et al. 2020. DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat. Methods* 17(11):1118–24
110. Fudenberg G, Kelley DR, Pollard KS. 2020. Predicting 3D genome folding from DNA sequence with Akita. *Nat. Methods* 17(11):1111–17
111. Zhou J. 2021. Sequence-based modeling of genome 3D architecture from kilobase to chromosome-scale. bioRxiv 10.1101/2021.05.19.444847. **https://doi.org/10.1101/2021.05.19.444847**
112. Laugsch M, Bartusel M, Rehimi R, Alirzayeva H, Karaolidou A, et al. 2019. Modeling the pathological long-range regulatory effects of human structural variation with patient-specific hiPSCs. *Cell Stem Cell* 24(5):736–52.e12
113. Gröschel S, Sanders MA, Hoogenboezem R, de Wit E, Bouwman BAM, et al. 2014. A single oncogenic enhancer rearrangement causes concomitant *EVI1* and *GATA2* deregulation in leukemia. *Cell* 157(2):369–81
114. Linder J, Bogard N, Rosenberg AB, Seelig G. 2019. Deep exploration networks for rapid engineering of functional DNA sequences. bioRxiv 864363. **https://doi.org/10.1101/864363**
115. Brookes D, Park H, Listgarten J. 2019. Conditioning by adaptive sampling for robust design. arXiv:1901.10060 [cs.LG]

116. Gupta A, Kundaje A. 2019. Targeted optimization of regulatory DNA sequences with neural editing architectures. bioRxiv 10.1101/714402. **https://doi.org/10.1101/714402**
117. Schreiber J, Lu YY, Noble WS. 2020. *Ledidi: designing genomic edits that induce functional activity*. Paper presented at Int. Conf. Mach. Learn. (ICML 2020), Virtual, Jul. 17
118. Zrimec J, Fu X, Muhammad AS, Skrekas C, Jauniskis V, et al. 2021. Supervised generative design of regulatory DNA for gene expression control. bioRxiv 10.1101/2021.07.15.452480. **https://doi.org/10.1101/2021.07.15.452480**