Computational and Structural Biotechnology Journal 20 (2022) 3439–3448

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini review

Predicting 3D chromatin interactions from DNA sequence using Deep Learning

Robert S. Piecyk [a,1], Luca Schlegel [a,1], Frank Johannes [a,b,*]

[a] Department of Molecular Life Sciences, Technical University of Munich, Freising, Germany
[b] TUM Institute for Advanced Study, Garching, Germany

ARTICLE INFO

Article history:
Received 13 May 2022
Received in revised form 21 June 2022
Accepted 21 June 2022
Available online 25 June 2022

Keywords:
3D Chromatin Interaction
Deep Learning
Epigenetics
Genome folding
Chromosome conformation capture (3C)

ABSTRACT

Gene regulation in eukaryotes is profoundly shaped by the 3D organization of chromatin within the cell nucleus. Distal regulatory interactions between enhancers and their target genes are widespread and many causal loci underlying heritable agricultural or clinical traits have been mapped to distal cis-regulatory elements. Dissecting the sequence features that mediate such distal interactions is key to understanding their underlying biology. Deep Learning (DL) models coupled with genome-wide 3C-based sequencing data have emerged as powerful tools to infer the DNA sequence grammar underlying such distal interactions. In this review we show that most DL models have remarkably high prediction accuracy, which indicates that DNA sequence features are important determinants of chromatin looping. However, DL model training has so far been limited to a small set of human cell lines, raising questions about the generalization of these predictions to other tissue-types and species. Furthermore, we find that the model architecture seems less relevant for model performance than the training strategy and the data preparation step. Transfer learning, coupled with functionally curated interactions, appear to be the most promising approach to learn cell-type specific and possibly species- specific sequence features in future applications.

Contents

* Corresponding author at: Department of Molecular Life Sciences, Technical University of Munich, Freising, Germany.
E-mail address: frank@johanneslab.org (F. Johannes).

[1] These authors contributed equally to conceptualisation and writing of the paper. Robert S. Piecyk and Luca Schlegel are both shared first authors.

https://doi.org/10.1016/j.csbj.2022.06.047

1. Introduction

The spatial organization of chromatin in the nucleus of animal and plant cells plays important roles in genome regulation. It has central functions in processes such as DNA replication and repair, the spatial- and temporal patterning of gene expression, and in the silencing of transposable element (TE). In the past two decades, chromosome conformation capture (3C) coupled with next-generation sequencing has emerged as a powerful method to interrogate 3D chromatin interactions in a high-throughput fashion [1]. Among these, Hi-C was the first developed method [2]. It is designed to capture all chromatin interactions at the genome-wide scale, albeit at low resolution (see Fig. 1,2, [2]). By contrast, more recent methods, such as chromatin interaction analysis by paired-end tag sequencing (ChIA- PET) (see Fig. 2, [3]) or Hi-C coupled to ChIP-seq (HiChIP or PLAC-Seq) [4,5], generate high-resolution interaction maps, but are restricted to specific loci occupied by proteins that can be pulled down by ChIP (e.g. modified histones, transcription factors, and RNA polymerase II). Together, these techniques have generated unprecedented insights into the function of 3D chromatin organization of mammalian and more recently also in plant genomes. They have led to the systematic identification of Enhancer-Promoter Interactions (EPIs), Insulator Loops (e.g. CTCF-cohesin loop in human cells) and interactions mediated by specific transcription factors [6].

Numerous machine learning (ML) approaches have emerged in parallel to these technological developments [7]. Their general aim is to use 3C data as input to train sequence-based predictors of chromatin looping and to identify specific sequence features that may facilitate physical contacts between distal genomic regions. The most promising of such ML approaches are supervised Deep Learning (DL) methods. As in other areas of genome biology, DL methods provide the most accurate predictions, can handle large and complex amount of genomic data and automatically detect patterns or unanticipated genomic relationships [8]. DL methods thus provide a powerful framework for dissecting the causal determinants underlying 3D chromatin interactions and for providing testable hypotheses for experimental follow up.

Here we review the state-of-the-art in the use of DL methods to predict 3D chromatin interactions from DNA sequence (Table 1). We evaluate these models in terms of their objective, data preprocessing, architecture, training procedure and finally by their prediction performance. We find that data selection and preprocessing in combination with transfer learning appear to be more important for model performance than the choice of model architecture. Moreover, even though all these models perform very well on their test data, they have so far been mainly restricted to the analysis of specific human cell lines. Similar model-based approaches are thus urgently needed for a wider range of cell lines, tissues or species. This information could facilitate deeper insights into the evolutionary and developmental factors that impact chromatin looping biology. Moreover, experimental validation of model-based predictions is needed to assess the biological value of these approaches and for fine-tuning model architecture.

1.1. Sequencing methods

Chromosome conformation capture (3C) followed by high-throughput sequencing has emerged as a powerful experimental approach to probe chromatin interactions at the genome-wide scale. Hi-C and ChIA-PET are two variants of these measurement approaches, which have served as the main data input for most of the DL methods reviewed in Table 1. The general experimental workflows for Hi-C and ChIA-PET involve cross-linking and fragmentation of chromatin, the addition of biomarkers, ligation, purification and finally sequencing [44]. We detail these experimental steps in Fig. 2. Despite the popularity of Hi-C, there are a number of well-known limitations with this method. First, its

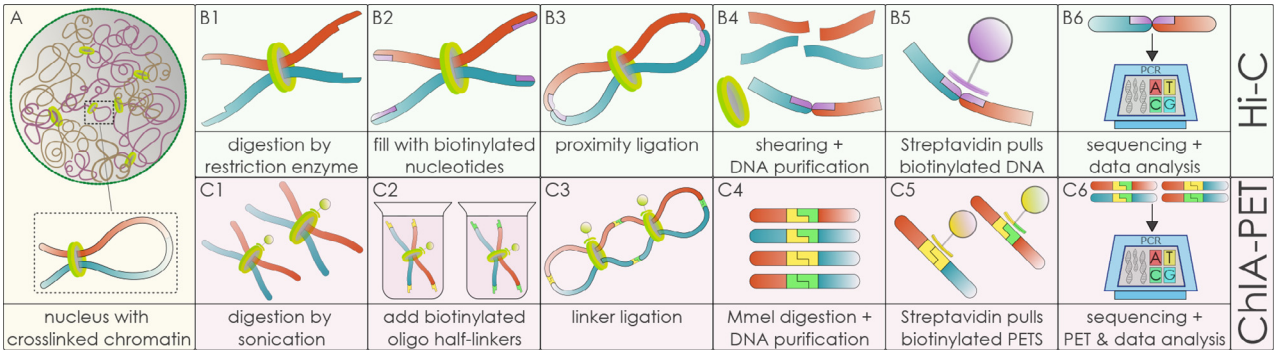

Fig. 1. (A), (B1) – (B5) Hi-C sequencing. A restriction enzyme buffer in combination with SDS solubilization enables the access to open cross-linked chromatin and removes any other substances. Next, a type II restriction endonuclease digests the accessible chromatin. HindIII enzyme detects and cleaves ~ 80 – 90% of all 5'-AAGCTT-3'sequences, which are filled with biotin-14-dCTP. Special dilute conditions favor proximity ligation and can be identified by its unique 5'-GCTAGC-3' NheI site. These chromatin ligation products are degraded by Proteinase K followed by several DNA purification steps [44]. (B6), (C6) DNA samples are mapped to a given DNA library with the correct reference genome including many quality steps. (A), (C1) - (C5) ChIA-PET sequencing. Formaldehyde stabilizes cross-linked DNA–protein complexes before sonication is used for digestion. ChIP is applied, while using the corresponding antibody for the protein of interest. The precipitate, which is enriched with the digested chromatin complexes, is divided into two separate locations with two different half-linker oligonucleotides. Both samples are mixed, which activates proximal half-linker and self-ligation. The restriction enzyme MmeI is added for digestion and DNA fragments from paired-end tags (PETs) are isolated. The final DNA sequences can be assigned uniquely by their ligation type. Self-ligated sequences are considered as chromatin looping with small distance, while mixed linkers referring to long distance base pairs, eventually on different chromosomes [49].

Table 1
Deep Learning algorithms for 3D chromatin interactions, sorted by architecture. All models are based on Convolution Neural Networks or Recurrent Neural Networks. [18,41,43,9,19,27,28,13,22,29,33,38,40,42,31,30] [11,12,14–17,20,21,23,25,26,32,34–37].

Method	Objective	Architecture	Training data	Organism	Auxiliary data	Input
EPIANN [9]	predict EPI	CNN	Hi-C [10]	human	annotations [11, 12]	anchor 1: 3 kb anchor 2: 2 kb
TransEPI [13]	predict EPI	CNN	Hi-C [10]	human	ChIP-seq [14] histone marks [12] DNase-seq [12] BENGI [15] annotations [16, 17]	DNA sequence: ~ 2.5 Mb + chromosome features
Rambutan [18]	predict 3D chromatin interaction	CNN	Hi-C [10]	human	DNase-seq [12]	anchor 1: 1 kb anchor 2: 1 kb
Akita [19]	predict 3D chromatin interaction	CNN	Hi-C [10] Micro-C [20, 21]	human and mouse	-	DNA sequence: ~ 1 Mb
DeepC [22]	predict 3D chromatin interactions	CNN transfer learning	Hi-C [10, 21]	human and mouse	annotations, ChIP-seq [23, 24, 25] ATAC-seq [23, 25, 24] DNase-seq [26, 24] histone marks [12, 24]	DNA sequence: 1 Mb + chromosome features
Orca [27]	multiscale 3D chromatin interaction	CNN transfer learning	Micro-C [20]	human	annotations[11, 12] DNase-seq [12] ChIP-seq [24]	DNA sequence: from 1 to 265 Mb
EPIsCNN [28]	predict EPI	CNN transfer learning	Hi-C [10]	human	annotations [11, 12]	anchor 1: 3 kb anchor 2: 2 kb
EPIsHilbert [29]	predict EPI	CNN transfer learning	Hi-C [10]	human	annotations [11, 12]	anchor 1: 3 kb anchor 2: 2 kb
EPIHC [30]	predict EPI	Multilayer Perception transfer learning	Hi-C [10]	human	annotations [11, 12] ChIP-seq [24] DNase-seq [14] RNA-seq [14] Methylation [24]	anchor 1: 3 kb anchor 2: 2 kb + chromosome features
ChiNN [31]	predict 3D chromatin interaction	CNN gradient tree boosting	Hi-C [10] ChIA-PET [32]	human	ChIP-seq [24] DNase-seq [24]	anchor 1: 1 kb anchor 2: 1 kb + chromosome features
DeepTact [33]	predict EPI/PPI	CNN + RNN	PCHi-C [34]	human	ChIP-seq [14] FANTOM5 [35] DNase-seq [24] TSS [36] CHiCAGO scores [37]	anchor 1: 1 kb anchor 2: 1 kb
DeepMILO [38]	predict 3D chromatin interaction	CNN + BLSTM	ChIA-PET [39, 32]	human	ChIP-seq [24]	anchor 1: 4 kb anchor 2: 4 kb
SEPT [40]	predict EPI	CNN + LSTM transfer learning	Hi-C [10]	human	annotations [11, 12] RNA-seq [24]	anchor 1: 3 kb anchor 2: 2 kb
SPEID [41]	predict EPI	CNN + LSTM transfer learning	Hi-C [10]	human	annotations [11, 12]	anchor 1: 3 kb anchor 2: 2 kb
EPIVAN [42]	predict EPI	CNN + BiGRU	Hi-C [10]	human	annotations [11, 12]	anchor 1: 3 kb anchor 2: 2 kb
EPI-DLMH [43]	predict EPI	CNN + BiGRU	Hi-C [10]	human	annotations [11, 12]	anchor 1: 3 kb anchor 2: 2 kb

resolution is limited by the choice of the digestion enzyme [45], so that spatially close chromatin interactions may be missed in genomic regions where the distribution of enzymatic cut sites is sparse. Second, Hi-C often does not capture long range interactions and omits many simultaneous promoter-enhancer interactions [46]. Third, false positive interactions are often detected because of spurious cross-linking and ligation. The first two limitations could be viewed as potential opportunities for DL methods, because computationally predicted loops could, in principle, be generated in genomic regions where the measurement technology has failed. However, the third limitation is disadvantageous for model training, where clean true positive (and true negative) loop sets are necessary. Newer experimental approaches, including ChIA-PET, Micro-C or Promoter Capture Hi-C (PCHi-C), try bypass many of these limitations. Micro-C uses micrococcal nuclease (MNase) as replacement for restriction enzymes to archive a higher resolution for short-distance interactions [47]. PCHi-C introduces an additional individual biotinylated RNA purification step for promoter rich fragments, to reduce the amount of ligation products before PCR sequencing is applied [48]. Unlike Hi-C, ChIA-PET combines 3C with chromatin immunoprecipitation (ChIP) followed by sequencing [3]. It performs chromatin fragmentation by sonication and uses a smarter biolinker concept [49]. Additionally, the use of ChIP enriches for chromatin contacts that harbor specific transcription factors, or other binding proteins, such as CTCF, and thus ensures a higher rate of true functional chromatin interactions [49]. However, as a trade-off ChIA-PET has relatively low sensitivity and requires a large amount of material [5], which could increase measurement variation due to cellular heterogeneity. Still, ChIA-PET is frequently applied and has thus been employed in the training of several DL algorithms (see Table 1).

2. Computational methods

2.1. Preprocessing

2.1.1. Preparation of training data

Measured (i.e. observed) chromatin interactions from Hi-C or ChIP-PET are the starting point for all DL methods reviewed here. However, DL methods differ in the way they use this information at the input stage. We can broadly distinguish between classification- and regression-based methods.

Classification-based methods (EPIANN, TransEPI, EPIsCNN, Rambutan, EPIsHilbert, EPIHC, ChINN, DeepTact, DeepMILO, SEPT, SPEID, EPIVAN, EPI-DLMH) typically take a list of discrete, interacting regions, called “anchors”, as input, which are treated as true

positives. The goal is to learn specific sequencing features within the anchors that may facilitate looping. The sizes of the input anchors can range from 2 kb to 3 kb, depending on the anchor type (see Table 1). In addition to these true positives, it is also necessary to supply true negative anchor pairs.

Most methods do this by "rewiring" the positive set of anchors in new ways; that is, they form *in silico*loops between anchors that have not been observed to interact in the original Hi-C or ChIP-PET data. These generated negatives are often matched for distances similar to those in the positive set, in order to avoid biases in model training (DeepMILO, DeepTACT, SEPT, SPEID). Alternatively, any distance information can be added as the auxiliary input in the training procedure itself (ChINN, Rambutan, TransEPI). Because of the substantial imbalance between positive and negative interactions, a data augmentation procedure is sometimes employed that adds arbitrary sub-sequences flanking the original anchors from the positive set (EPIANN, EPIsCNN, EPIsHilbert, SEPT, SPEID, EPIVAN, EPI-DLMH, EPIHC).

In contrast to classification-based approaches, regression-based methods (Akita, DeepC, Orca) take contact frequency maps as input, which are constructed at megabase (Mb) resolution. To this end, the genome is partitioned into non-overlapping virtual contigs, and interaction counts within ~ 1 Mb sub-sequences are taken to generate a two-dimensional frequency matrix. DNA sequences and corresponding frequency matrix are fed into the DL algorithm to predict factors affecting local interaction typologies (Akita, DeepC). Orca introduces, in addition to the sequence encoder, a multilevel cascading decoder to provide genome wide interaction on window sizes from 1 Mb to 265 Mb with resolution between 4 kb and 1024 kb.

2.1.2. Auxiliary data

Many classification-based methods further employ auxiliary data to filter anchor pairs prior to training. For example, methods such as EPIANN, TransEPI or SPEID (see Table 1), select anchors that map to annotated enhancers and promoters. Such filtering strategies reduce the genome-wide interaction landscape to a subset of functional regions, and (probably) reduce biases or measurement errors arising from the 3C assay itself. Clearly, such *a priori* filtering strategies are only sensible in applications involving high-quality genomes where sufficient epigenomic information is available and all regulatory elements are well annotated. This is certainly true for all DL methods to date, as they have been developed specifically for human genomic applications, and therefore benefit from the extensive ENCODE data resources [24].

Common auxiliary data modalities include RNA-seq (gene expression), DNase-seq (accessible regions), as well as ChIP-seq for various histone modifications, RNA polymerase, transcription factors or CTCF binding proteins. Using such functional data, anchors overlapping low-expressed transcripts from RNA-seq can be removed from the enhancer-promoter data, since they are less likely to be actively regulated by enhancers [13,40,50]. Similarly, anchors can be screened for chromatin accessible regions related to CTCF or RNA Pol II binding sites [31] to be enriched for active loops. Furthermore, the incorporation of functional data makes it possible to identify cell-type-specific chromatin interactions. This creates avenues for predicting and understanding looping biology underlying cell lineage determination.

As an alternative to *a priori* filtering, some methods integrate such auxiliary data directly into the model training, either in a pre-training step (DeepC, Orca, ChiNN) or in the full training procedure (TransEPI, EPIHC). Either way, it has been shown that the inclusion of auxiliary data into DL methods can improve algorithm performance (TransEPI, DeepC, Orca, ChiNN, EPIHC), and thus appears to be an important aspect of data preparation.

2.1.3. Sequence embedding

The nucleotide sequences used in DL model training need to first be converted into a machine readable format. One-hot encoding is a commonly used approach to represent sequences as binary vectors. The four nucleotides A, T, C and G are saved in four separate channels. A is represented as $[1,0,0,0]^T$, T as $[0,1,0,0]^T$, C as $[0,0,1,0]^T$ and G as $[0,0,0,1]^T$. The unknown nucleotide category *N* can be either removed (SPEID, EPIANN, SIMCNN), represented as $[0.25,0.25,0.25,0.25]^T$ (ChINN), or saved as the fifth dimension in matrix representation (DeepMILO). Moreover, one-hot encoding can be easily extended to capture the spatial relationship between anchors by converting a two-dimensional one-hot matrix into a three-dimensional matrix–vector representation by applying a space-filling Hilbert curve [51,52] (see EPIsHilbert).

Another group of methods uses instead a distributed representation of small DNA sequences, called k-mers. Dna2vec embedding [53] is based on a popular word2vec [54] natural language processing model. The algorithm takes distributed samples of variable-length k-mers to train a shallow two-layer neural network model. This aggregated model enables the user to perform a decomposition by k-mer length, followed by the selection of the best low-dimensional and high-quality vector representation used for sequence embedding. This increases computational efficiency as well as specificity in DL approaches (EPIVAN), and preserves the hidden information from the correlation between single small DNA sequences. It can be either trained on the whole genome or on a set of anchors.

2.2. Deep Learning approaches

We examined the model architectures of the DL methods in Table 1 in detail. The most frequent architecture uses a Convolutional neural network (CNN) in combination with a long short-term memory (LSTM) (see Fig. 2). CNNs, or simply convolutional networks (CNs), are a specific type of DL, which are based on artificial neural networks (ANNs) [55]. Models with a long short-term memory (LSTM) unit belong to the family of recurrent neural networks (RNNs) and are also a class of ANNs [56]. The family of LSTM units also comprises Gated Recurrent Units (GRU). It is a slightly simpler version with less parameters [57]. Both models often occur with additional connections in opposite directions within one layer [58]. In that case, they are called Bidirectional LSTM (Bi-LSTM) and Bidirectional-Gru (Bi-Gru). Due to the continuous refinement of 3C-based methods, especially Hi-C, ChIA-PET or other throughput techniques, large amounts of high-quality chromatin data is available. This data motivates the use of supervised versions of the introduced ANN models. If we consider Table 1, we notice that all models are using CNNs as core models and six of 16 additionally add an extra (Bi)-LSTM or Bi-Gru unit (DeepTact, DeepMilo, SEPT, SPEID, EPIVAN and EPI-DLMH). This observation could be explained by the historical applications of CNNs and LSTMs. As its name indicates, CNNs use a convolution operator, also called kernel function, to search for specific patterns in discrete grid-based topologies or sequential data [59]. The concept was originally developed for image or sound classification and is related to the neuronal connections of the human brain [60]. Section 2.1 explained the conversion of genomic sequences into a matrix or vector-based representation. This indicates the connection between image recognition and the computation of chromatin interaction since the input data of both problems can be represented by a matrix or linear vector combination that contain unknown patterns and possibly long-term dependencies. LSTM units are designed to capture long-term dependencies, due to shared parameters that function as a memory unit. One could say, if CNNs catch mostly local patterns, LSTMs act as a global

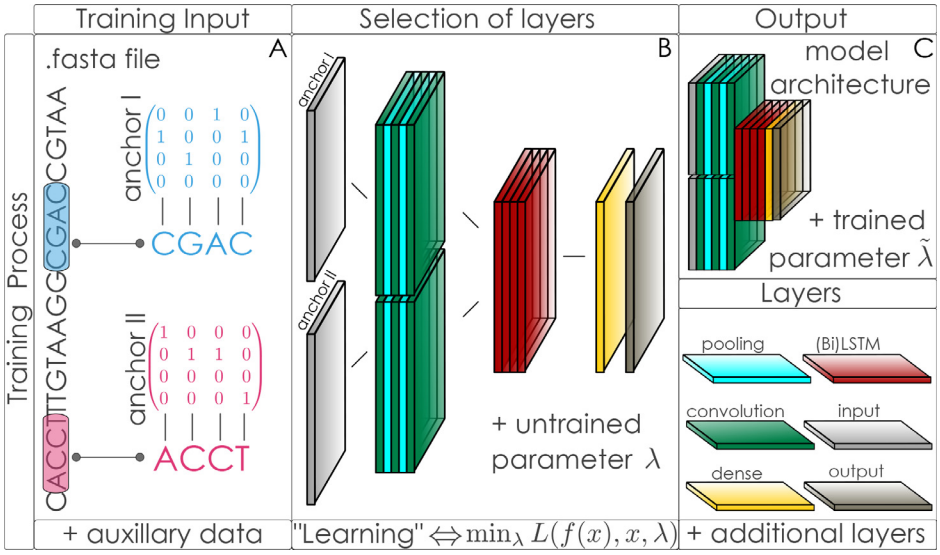

Fig. 2. Training procedure of a typical CNN + LSTM model. (A) During input data preprocessing, chromatin sequences are translated using a one-hot encoding technique. (B) Schematic representation of a commonly used architecture in chromatin interaction detection. Two input anchors are managed separately by CNN blocks, which are defined by the number of convolutional and pooling layers and several hyperparameters. Once the architecture, the arrangement of the layers and hyperparameters are selected, we start with an initial unbiased parameter distribution λ. The concept of 'training' refers to minimizing a certain loss function L with respect to lambda, which contains the input values x, the parameter set λ and all non-linear functions $f(x)$. This representation is drastically reduced and on a highly abstract level. Many additional decisions are necessary to define a full CNN model with LSTM units. (C) After training, we end up with a set of optimal parameters $\tilde{\lambda}$. This set of trained parameters in combination with the model architecture must be applied to a final test data for validation, before applying to completely new data.

observer. Therefore, LSTMs have been used mainly in time dependent problems, as they store crucial information over a long period of time. A typical application is speech recognition, where it can take a long time until a specific word or phrase is repeated [56]. A similar event occurs in promoter-enhancer interactions, which often span distances of up to several Mbs [61]. A common DL configuration as well as typical training data is illustrated in Fig. 2. Even though the layer-based representation shown in that figure is helpful for visualizing the overall model structure, it is important to keep the actual neural-based architecture in mind. All layers contain nodes that are connected through edges. On edges, linear transformations are applied, which provide all the weights or parameters. On nodes, preselected non-linear transformations $f = f(x)$ are necessary. The process of training is equivalent to minimizing a loss function L with respect to the corresponding parameters λ, which connect the layers. The output of the training process is this specific set of parameters $\tilde{\lambda}$ in combination with the layer structure and arrangement. During the training process, the model is validated with a subset of the training set (usually around 10%). Another 10% of the training data, the testing set, is saved and used for the final test run on unknown data. In the supervised setup, this test run provides all statistical measurements and quality values like accuracy, sensitivity, specificity and other conditions on the sample set.

2.3. Performance and training strategies

We sought to rank the prediction performance of the models reviewed in Table 1. The majority of the original studies provide values for the area under precision recall curve (AUPR) as statistical measurement of prediction accuracy. A higher AUPR value is indicative of better model performance. Since Akita, Orca and DeepC are regression based models, they did not published those values and are not included in the performance analysis. Nevertheless, since Akita and Orca are trained on the same Micro-C data set, it is possible to compare their Pearson correlation coefficient. On average, in Orca this correlation coefficient is $\sim 7.4\%$ higher for the H1 embryonic stem cells (H1-ESC) and $\sim 6.4\%$ higher for the human foreskin fibroblasts (HFF). Beside these, we exclude all algorithms that are not trained and tested on Hi-C [10] data, because mixed data modalities would render comparisons difficult.

The remaining models (EPIsHilbert, EPIVAN, EPIsCNN, EPI-DLMH, EPIsCNN, EPIVAN, EPIHC and EPIANN), which provide AUPR values, have been trained on a combination of six cell lines: K562 (mesoderm lineage cells from a leukemia patient), GM12878 (lymphoblastoid cells), HeLa-S3 (ectoderm lineage cells from a cervical cancer patient), HUVEC (umbilical vein endothelial cells), IMR90 (fetal lung fibroblasts) and NHEK (epidermal keratinocytes) [10], and employ a rich set of auxiliary functional data [62].

We observe four essential training strategies:

- Training on a specific cell line
- Training on a combination of all cell lines
- Model-based learning
- Cellular transfer learning (data based)

Model-based training often introduces an attention layer, to merge the acquired feature knowledge of cell-line specific submodels. Cellular transfer learning suggests two or more stages. First, all cell line data is used for training, and second, a specific cell-line is selected for another training procedure. The second cell-line specific training consumes little computational effort [28], since the model architecture remains the same and the primary trained parameters are used as initial weights. If cellular transfer learning has been applied, the process is repeated for all six cell lines separately to obtain a total of six AUPR values. Training and test-sets originate from the same cell line. The mean value of these six AUPR values is plotted in Fig. 3 with the range of all values indicated by the yellow interval. We observe that models using transfer learning tend to perform comparatively well. This could be an indicator that cell-specific features are important predictors [63]. Interestingly, we also see that the more sophisticated CNN + LSTM architectures are not among the top performers.

There are a number of caveats with this simple side-by-side performance evaluation. First, AUPR values are just one of several statistical measurement tools. The imbalanced training data set

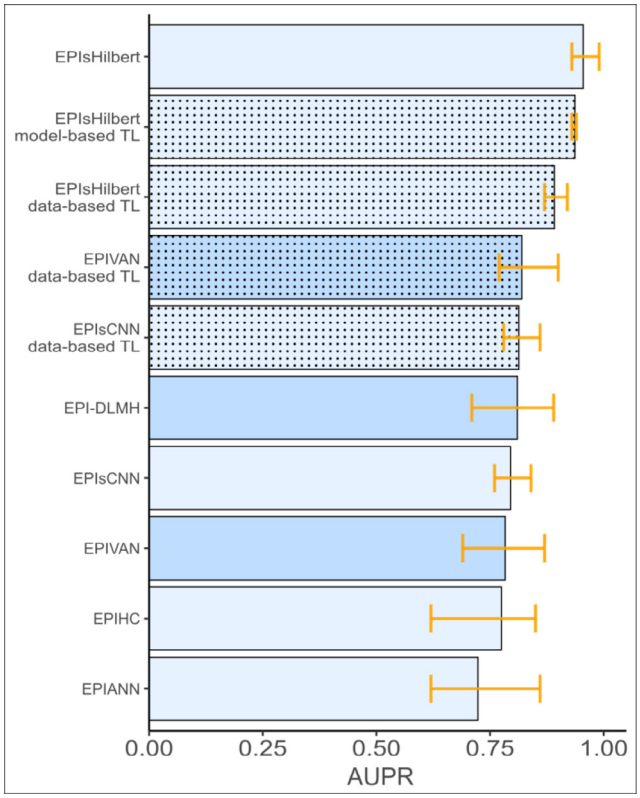

Fig. 3. Performance bar plot for models, which provided AUPR value for cell type specific training and testing. Light blue bars indicate a pure CNN model, dark blue refers to a CNN + RNN model. Transfer learning is represented by dotted bars. The yellow interval is defined by the minimum and maximum AUPR value.

contains approximate 20 times more true interactions. Since the AUPR value mainly composes true interactions [64], it is the most frequently used performance tool. Second, we also do not consider the computational effort and costs of the models, which is an essential factor for the efficiency.

2.4. Biological insights and applications

2.4.1. Genomic determinants of looping

DL models are undoubtedly powerful tools for predicting 3D chromatin looping based on DNA sequence. However, they also provide an intriguing framework for dissecting the underlying looping biology. A common approach in classification-based algorithms is to use *in silico* mutagenesis, whereby mutations are artificially introduced into specific anchor pairs, and then re-supplied to the trained DL model. The concept of this approach is presented in the Fig. 4. Mutations that lead to significant drops in the predicted interaction probability would suggest an important role in facilitating chromatin contacts, perhaps because they are located in crucial protein binding motifs. Algorithms such as DeepC, Akita, Orca and DeepMILO have extensively used this approach to assess the effect of specific deletions, single nucleotide polymorphisms (SNP) or structural variants. Akita, for instance, mutagenized a set of random regions within and near CTCF motifs. The results of this study showed the significant impact of SNPs on CTCF binding, either directly or by flanking cofactors. Additionally, a mouse-trained model from Akita was used to predict the effect of a 622 kb inversion at the enhancer locus Eph4A on 3D folding. The experimental studies observed that the inversion effected CTCF binding [65]. Using its predicted contact maps, Akita confirmed that an *in silico* inversion of the Eph4A locus lead to a loss of CTCF mediated insulator looping. However, *in silico* mutagenesis is a brute force approach. It is not optimal for probing the complete combinatorial mutation space within a given anchor sequence. This limitation may be crucial in situations where looping is facilitated by combinations of specific (and possible complex) motif sequences. As an alternative, DeepC employs a metric called saliency score [66], which quantifies the importance of every base pair and motif to the predicted interaction. This metric can be calculated as the scalar product of the model output gradient, with respect to the one-hot-encoded input sequence. Akita used this metric as well. They found saliency peaks at the CTCF motifs and active promoters, and hypothesized that mutations of these regions would affect chromatin architecture, and thus, gene expression, which can be investigated by expression quantitative trait locus (eQTL) studies. To test this, they used the set of cell-type specific eQTLs located in open chromatin or CTCF sites. They compared the saliency scores of these eQTLs and mutagenized random regions. The significantly higher saliency score for eQTLs revealed that this metric can be used for eQTL mapping when the expression changes are caused by chromatin architecture perturbation. Orca successfully predicted the influence of several structural variants in six studies while comparing the model output with experimental chromatin capture data.

Class Activation Maps (CAM) [67] is another approach to quantify the influence of sequence features on enhancer-promoter interaction. It visualizes three-dimensional vectors with a heatmap matrix, which represents the interaction occurrence and their spatial relationship. These association maps may be used to highlight sequence patterns leading to the chromatin interaction (EPIsHilbert). Another proposed method to detect motifs responsible for chromatin looping is based on the output of the first convolutional layer. This procedure can extract the best matching subsequences for each kernel, with respect to the model architecture. SEPT used this strategy to compute a position frequency matrix (PFM) and then, compare the PFM-related features with known TF motifs from HOCOMOCO database [68]. They found a set of potentially important regulatory elements, which are involved in transcription and cell-cycle regulation that may determine their role in chromatin looping. These results revealed that SEPT has the ability to learn cell-type specific patterns crucial in genome folding, which explains its relevance in transfer learning approaches.

Attention layers [69] can be used not only to merge cell-line specific submodels in transfer learning, but also to evaluate the impact of those features on prediction. EPIANN labeled each base in observed enhancer and promoter sequences with the corresponding marginal attention to highlight regulatory elements overlapping patterns crucial in predicting EPIs. It showed that the attention regions are usually highly correlated with other genomic annotations. This information can be further investigated to interpret the key patterns in chromatin interactions.

2.4.2. Screening of disrupting variants

Building on the above-describe approach, several studies have tried to test or identify specific loop-disrupting genetic variants and their impact on cancerogenesis [39]. Such genetic variants could result in the suppression, or in some cases even the activation, of chromatin loops. In the latter case, enhancer elements could be erroneously brought in physical contacts with proto-oncogenes and thus promote cancer progression [70,71]. The DL prediction framework provides a means to systematically screen through available GWAS or re-sequencing databases to identify putative causal variants underlying differential looping.

Using such an approach, DeepMILO was tested on two known deletions in T-Cell Acute Lymphoblastic Leukemia (T-ALL) patients at anchors containing oncogenes TAL1 and LMO2. The algorithm correctly predicted that these mutations should lead to insulator loop disruption. In addition, DeepMILO employed *in silico* mutagenesis and further predicted that a number of smaller

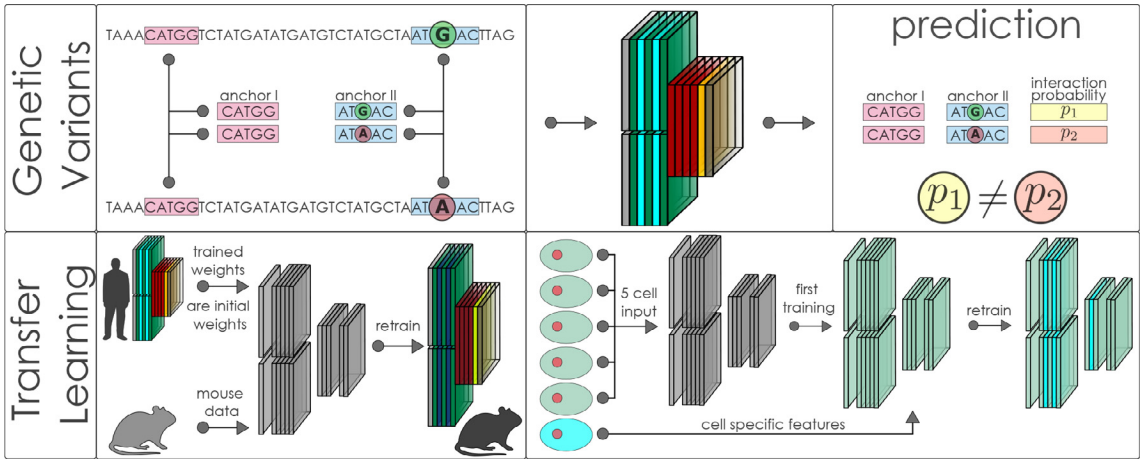

Fig. 4. Biological insights by Deep Learning. Deep Learning models can be used to predict chromatin interaction in combination with several nucleotide variants through *in silico* mutagenesis. A change in the loop probabilities p_1 and p_2 between sequence variants indicates the importance of the specific single nucleotide polymorphisms in chromatin looping. Transfer learning can be used to extend previously trained knowledge to different species or cell lines.

mutations, just outside the CTCF binding motif, may also affect chromatin interactions. These latter predictions provide concrete hypotheses for experimental follow-up. Conceptually similar approaches were taken by TransEPI and ChINN to identify putative loop-disrupting mutations in the context of neuronal disorder and chronic lymphocytic leukemia (CLL), respectively. Using chronic lymphocytic leukemia (CLL) patient data, ChINN was able to construct a patient-specific chromatin interaction profile, suggesting that such predictions could serve as a per-clinical tool to predict CLL disease risk. Orca predicted high structural impact regions (10 bp) consistently with Chip-seq data and confirmed reduced genome interactions through disrupting motifs like POU5F1::SOX2 or AP-1. Overall, these results suggest local and global effects of disrupting variants.

Many eQTLs identified in human population data have been shown to act in *trans*, that is, the QTL affects the expression state at distal genes, rather than locally. One possible mechanism for these distal interactions is chromatin looping [72]. Indeed, application of Akita on a set of eQTLs from GTEx (Genotype-Tissue Expression, [73]) whole blood samples revealed significantly higher disruption for SNP with greater causal probability, within and outside of CTCF motifs. This indicates the impact of CTCF and non-CTCF variants on genome folding. Hence, eQTL data sets can be used in biological validation of predicted interactions to reduce false positives from 3C-based data.

2.4.3. Cell-type and species specific chromatin interactions

Cross-cell line prediction poses a big challenge due to the presence of cell-type specificity [63,10]. Our review revealed that general DL models trained on all cell lines together, tend to perform poorly in capturing cell-specific events due to the relative disproportion between shared and private chromatin interactions. Conversely, most DL methods that are trained on each cell line separately, failed to identify general and cell-line specific interactions simultaneously. Approaches based on transfer learning perform much better in this setting. EPIsHilbert hypothesized that the effectiveness of these approaches is determined by the numerous common sequence patterns among all cell lines. To evidence this, they calculated the overlapping ratio of chromatin interactions between different tissues. It indicated that there are more common than private interactions. This knowledge can be used to create a model that has the ability to predict chromatin interactions on novel cell lines. SEPT used a feature extractor and domain discriminator to learn EPIs-related features and recognize cell-line specific patterns at the same time. This provides an opportunity to create an universal model, which is able to predict chromatin interactions not only on previously trained cell lines, but also on novel cell lines using general EPI features. Similarities between mammalian genome folding may allow us to predict species-specific differences in genome folding. Recent studies showed that ChAHP complexes lead to the disruption of insulator loops within mouse-specific B2 SINE elements [74,75]. To test this, Akita trained models on human and murine embryonic stem cell (ESC) Hi-C to show the impact of *in silico* mutation in these elements on CTCF binding. Comparison of these results confirmed that both models correctly predicted the disturbance of genome folding before and after mutagenesis of B2 SINE elements. This highlights the opportunity to use DL and transfer learning approaches in studies investigating species-specific regulatory strategies (see Fig. 4).

3. Discussion

Here, we have reviewed current supervised DL models for predicting 3D chromatin interactions from DNA sequence. We find that these methods have remarkably high prediction accuracy, which indicates that DNA sequence features are important determinants of chromatin looping. By examining the learned sequence features it is possible to uncover complex, combinatorial, sequence motifs that would otherwise be difficult to discover, even with elaborate experimental assays. Thus, DL models have the potential to provide novel insights into chromatin biology. Similarly, trained DL models can be used as a tool to identify loop disrupting genetic variants from population-level sequencing data. This type of information is highly relevant for understanding the genetic basis of regulatory variation underlying complex traits, both in a clinical as well as in an agricultural setting. Indeed, numerous genome-wide association studies have identified causal loci in non-coding regions of genomes, many of which appear to act as eQTL for distal target genes [76,77]. DL models could be used to assess if these non-coding loci and their targets are likely to interact physically, and whether the type of genetic variants seen at the eQTL locus is expected to cause differential looping. Understanding the mode of action of disease-associated non-coding variants can facilitate insights into disease etiology and potentially lead to novel treatment targets in biomedical applications.

Despite such exciting prospects, our review also revealed a number of key limitations with current DL methods. All models to date have been trained and tested on human data of six cell

lines from the ENCODE project [24], or a subset of those cell lines. While the use of a single, or a few, reference data set(s) enables comparisons of the algorithms among themselves, it reduces the generalizability of the trained models. This potential limitation is already apparent in DL models that trained on multiple cell lines and showed how sensitive the trained model can be to cell-type specific features (EPIsCNN, EPIsHilbert, EPIVAN, SEPT, TransEPI). It is therefore highly unlikely that current models readily extent to other *in vivo* tissues in humans, and much less across different species. Hence, a much broader range of training data sets is urgently needed.

From our perspective, a cross-species DL model would be highly interesting. For instance, it is well known that the biology underlying chromatin looping in mammals and plants differs fundamentally. Plants lack CTCF proteins and display several other key topological differences in the 3D organization of their genomes. In contrast to mammalian TADs structure, plant compartment domains tend to interact with each other at intra and inter-chromosome level [78,79]. Moreover, although mammalian loop domains are conserved through different species due to conservation of CTCF binding sites, plant compartments might differ for several plant species [80]. The molecular components underlying chromatin looping in plants is not fully resolved. Cross-species DL models would not only be able to identify conserved and divergence sequence determinants, but also reveal how the molecular mechanisms underlying chromatin looping have evolved. However, training the DL models reviewed here on other species may not be trivial. Many of the models rely heavily on auxiliary functional data (DNase-seq, RNA-seq, ChIP-seq, CTCF binding proteins, transcription factors, RNA polymerase) and high-quality genome annotations in the training procedure itself (see Table 1). While this type of data is useful for boosting model performance, it is not readily available in most non–human species. Thus, such auxiliary data would have to be generated first, or alternative training strategy would have been employed that rely less on such data sources. From a bioinformatics point of view, it may be tempting to focus on the development of the ANN structure. However, if we consider the performance metrics in Fig. 3, there is no clear indication that ANN structure has a significant impact, if it consists of some minimal criteria like two well performing CNN blocks. Zhuang et al. [28], for instance, shows that a simple CNN model has similar performance on the same training data as a CNN models coupled with a RNN. This even raises questions about the implementation of an LSTM unit altogether, even though there are decent biological interpretations of their function. Since the computational effort of LSTM or related units is very expensive compared to CNN blocks, they might fail in the long run due to lack of efficiency. On the other hand, the training process itself seems to have a very deep impact on prediction results. If we consider the AUPR values in Fig. 3, we observe that the most promising results are derived by training processes that use transfer learning strategies. This improvement in statistical measurements, can be explained biologically by cell line specific and general features along the chromatin. In mammalian cells, for example, it is known that CTCF bindings are conserved through all cell lines, but other interactions are cell line specific [81]. It should also be clear that the use of DL models cannot fully replace experimental data in revealing chromatin interactions. Experimental validation of predicted interactions should go hand-in–hand with model building. To date, there has been relatively little effort to perform such validations.

Author contributions

All authors contributed to the conceptualisation and writing of the paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

F.J. acknowledges support from the Technical University of Munich Institute for Advanced Study, funded by the German Excellence Initiative and the European Seventh Framework Programme under grant agreement No. 291763. R.S.P. was supported by the SFB Sonderforschungsbereich924 of the Deutsche Forschungsgemeinschaft (DFG). L.S. was supported by the DFG.

References

[1] Annette Denker and Wouter de Laat. The second decade of 3c technologies: detailed insights into nuclear organization. Genes Develop, 30(12), 1357–1382, jun 2016.
[2] Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M.A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326 (5950), 289–293, Oct 2009.
[3] Fullwood Melissa J, Ruan Yijun. ChIP-based methods for the identification of long-range chromatin interactions. J Cell Biochem 2009;107(1):30–9.
[4] Rongxin Fang, Miao Yu, Guoqiang Li, Sora Chee, Tristin Liu, Anthony D Schmitt, and Bing Ren. Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. Cell Res, 26(12), 1345–1348, Nov 2016.
[5] Maxwell R Mumbach, Adam J Rubin, Ryan A Flynn, Chao Dai, Paul A Khavari, William J Greenleaf, and Howard Y Chang. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nature Methods, 13(11), 919–922, sep 2016.
[6] Karin M. Too many transcription factors: positive and negative interactions. New Biol 1990;2:126–31.
[7] Hang Xu, Zhang Shijie, Yi Xianfu, Plewczynski Dariusz, Li Mulin Jun. Exploring 3d chromatin contacts in gene regulation: The evolution of approaches for the identification of functional enhancer-promoter interaction. Comput Struct Biotechnol J 2020;18:558–70.
[8] Eraslan Gökcen, Avsec Žiga, Gagneur Julien, Theis Fabian J. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet 2019;20 (7):389–403.
[9] Mao Weiguang, Kostka Dennis, Chikina Maria. Modeling enhancer-promoter interactions with attention-based neural networks. EPIANN 2017;11.
[10] Suhas S.P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and Erez Lieberman Aiden. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 159(7), 1665–1680, Dec 2014.
[11] A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol, 9(4): e1001046, apr 2011.
[12] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, Viren Amin, John W. Whitaker, Matthew D. Schultz, Lucas D. Ward, Abhishek Sarkar, Gerald Quon, Richard S. Sandstrom, Matthew L. Eaton, Yi-Chieh Wu, Andreas R. Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R. Alan Harris, Noam Shoresh, Charles B. Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R. David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J. Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K. Canfield, R. Scott Hansen, Rajinder Kaul, Peter J. Sabo, Mukul S. Bansal, Annaick Carles, Jesse R. Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R. Mercer, Shane J. Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C. Sallari, Kyle T. Siebenthall, Nicholas A. Sinnott-Armstrong, Michael Stevens, Robert E. Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E. Beaudet, Laurie A. Boyer, Philip L. De Jager, Peggy J. Farnham, Susan J. Fisher, David Haussler, Steven J.M. Jones, Wei Li, Marco A. Marra, Michael T. McManus, Shamil Sunyaev, James A. Thomson, Thea D. Tlsty, Li-Huei Tsai, Wei Wang, Robert A. Waterland, Michael Q. Zhang, Lisa H. Chadwick, Bradley E. Bernstein, Joseph F. Costello, Joseph R. Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A. Stamatoyannopoulos, Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. Nature, 518(7539), 317–330, Feb 2015.
[13] Ken Chen, Huiying Zhao, and Yuedong Yang. Capturing large genomic contexts for accurately predicting enhancer-promoter interactions. bioRxiv, 2021.
[14] Carrie A Davis, Benjamin C Hitz, Cricket A Sloan, Esther T Chan, Jean M Davidson, Idan Gabdank, Jason A Hilton, Kriti Jain, Ulugbek K Baymuradov, Aditi K Narayanan, Kathrina C Onate, Keenan Graham, Stuart R Miyasato,

Timothy R Dreszer, J Seth Strattan, Otto Jolanki, Forrest Y Tanaka, and J Michael Cherry. The encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res, 46(D1), D794–D801, Nov 2017.

[15] Moore Jill E, Pratt Henry E, Purcaro Michael J, Weng Zhiping. A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. Genome Biol 2020;21(1).

[16] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Silvia Carbonell Sala, Jacqueline Chrast, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Toby Hunt, Osagie G Izuogu, Julien Lagarde, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Jinuri Xu, Andrew Yates, Daniel Zerbino, Yan Zhang, Bronwen Aken, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Alexandre Reymond, Michael L Tress, and Paul Flicek. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res, 47(D1), D766–D773, Oct 2018.

[17] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38 (16):e164.

[18] Jacob Schreiber, Maxwell Libbrecht, Jeffrey Bilmes, and William Stafford Noble. Nucleotide sequence and dnasei sensitivity are predictive of 3d chromatin architecture. bioRxiv, 2017.

[19] Fudenberg Geoff, Kelley David R, Pollard Katherine S. Predicting 3d genome folding from DNA sequence with akita. Nat Methods 2020;17(11):1111–7.

[20] Nils Krietenstein, Sameer Abraham, Sergey V. Venev, Nezar Abdennur, Johan Gibcus, Tsung-Han S. Hsieh, Krishna Mohan Parsi, Liyan Yang, René Maehr, Leonid A. Mirny, Job Dekker, and Oliver J. Rando. Ultrastructural details of mammalian chromosome architecture. Molecular Cell, 78(3), 554–565.e7, May 2020.

[21] Boyan Bonev, Netta Mendelson Cohen, Quentin Szabo, Lauriane Fritsch, Giorgio L. Papadopoulos, Yaniv Lubling, Xiaole Xu, Xiaodan Lv, Jean-Philippe Hugnot, Amos Tanay, and Giacomo Cavalli. Multiscale 3d genome rewiring during mouse neural development. Cell, 171(3), 557–572.e24, Oct 2017.

[22] Ron Schwessinger, Matthew Gosden, Damien Downes, Richard C. Brown, A. Marieke Oudelaar, Jelena Telenius, Yee Whye Teh, Gerton Lunter, and Jim R. Hughes. DeepC: predicting 3d genome folding using megabase-scale transfer learning. Nature Methods, 17(11), 1118–1124, Oct 2020.

[23] M Ryan Corces, Jason D Buenrostro, Beijing Wu, Peyton G Greenside, Steven M Chan, Julie L Koenig, Michael P Snyder, Jonathan K Pritchard, Anshul Kundaje, William J Greenleaf, Ravindra Majeti, and Howard Y Chang. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nature Genet, 48(10), 1193–1203, Aug 2016.

[24] An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), 57–74, sep 2012.

[25] Damien J. Downes, Ron Schwessinger, Stephanie J. Hill, Lea Nussbaum, Caroline Scott, Matthew E. Gosden, Priscila P. Hirschfeld, Jelena M. Telenius, Chris Q. Eijsbouts, Simon J. McGowan, Antony J. Cutler, Jon Kerry, Jessica L. Davies, Calliope A. Dendrou, Jamie R.J. Inshaw, Martin S.C. Larke, A. Marieke Oudelaar, Yavor Bozhilov, Andrew J. King, Richard C. Brown, Maria C. Suciu, James O.J. Davies, Philip Hublitz, Chris Fisher, Ryo Kurita, Yukio Nakamura, Gerton Lunter, Stephen Taylor, Veronica J. Buckle, John A. Todd, Douglas R. Higgs, and Jim R. Hughes. An integrated platform to systematically identify causal variants and genes for polygenic human traits. Oct 2019.

[26] Ron Schwessinger, Maria C. Suciu, Simon J. McGowan, Jelena Telenius, Stephen Taylor, Doug R. Higgs, and Jim R. Hughes. Sasquatch: predicting the impact of regulatory SNPs on transcription factor binding from cell- and tissue-specific DNase footprints. Genome Res, 27(10), 1730–1742, sep 2017.

[27] Jian Zhou. Sequence-based modeling of genome 3d architecture from kilobase to chromosome-scale. May 2021.

[28] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. 2019

[29] Mingyang Zhang, Yujia Hu, and Min Zhu. EPIsHilbert: Prediction of enhancer-promoter interactions via hilbert curve encoding and transfer learning. Genes, 12(9):1385, sep 2021.

[30] Liu Shuai, Xu Xinran, Yang Zhihao, Zhao Xiaohan, Liu Shichao, Zhang Wen. EPIHC: Improving enhancer-promoter interaction prediction by using hybrid features and communicative learning. IEEE/ACM Trans Comput Biol Bioinf 2021:1.

[31] Fan Cao, Yu Zhang, Yichao Cai, Sambhavi Animesh, Ying Zhang, Semih Can Akincilar, Yan Ping Loh, Xinya Li, Wee Joo Chng, Vinay Tergaonkar, Chee Keong Kwoh, and Melissa J. Fullwood. Chromatin interaction neural network (ChINN): a machine learning-based method for predicting chromatin interactions from DNA sequences. Genome Biol, 22(1), Aug 2021.

[32] Zhonghui Tang, Oscar Junhong Luo, Xingwang Li, Meizhen Zheng, Jacqueline Jufen Zhu, Przemyslaw Szalaj, Pawel Trzaskoma, Adriana Magalska, Jakub Wlodarczyk, Blazej Ruszczycki, Paul Michalski, Emaly Piecuch, Ping Wang, Danjuan Wang, Simon Zhongyuan Tian, May Penrad-Mobayed, Laurent M. Sachs, Xiaoan Ruan, Chia-Lin Wei, Edison T. Liu, Grzegorz M. Wilczynski, Dariusz Plewczynski, Guoliang Li, and Yijun Ruan. CTCF-mediated human 3d genome architecture reveals chromatin topology for transcription. Cell, 163(7), 1611–1627, Dec 2015.

[33] Li Wenran, Wong Wing Hung, Jiang Rui. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. Nucleic Acids Res 2019;47(10). e60–e60.

[34] Javierre Biola M, Burren Oliver S, Wilder Steven P, Kreuzhuber Roman, Hill Steven M, Sewitz Sven, Cairns Jonathan, Wingett Steven W, Várnai Csilla, Thiecke Michiel J, Burden Frances, Farrow Samantha, Cutler Antony J, Rehnström Karola, Downes Kate, Grassi Luigi, Kostadima Myrto, Freire-Pritchett Paula, Wang Fan, Stunnenberg Hendrik G, Todd John A, Zerbino Daniel R, Stegle Oliver, Ouwehand Willem H, Frontini Mattia, Wallace Chris, Spivakov Mikhail, Fraser Peter, Martens Joost H, Kim Bowon, Sharifi Nilofar, Janssen-Megens Eva M, Yaspo Marie-Laure, Linser Matthias, Kovacsovics Alexander, Clarke Laura, Richardson David, Datta Avik, Flicek Paul. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell 2016;167(5):1369–84. e19.

[35] Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, Evgenia Ntini, Erik Arner, Eivind Valen, Kang Li, Lucia Schwarzfischer, Dagmar Glatz, Johanna Raithel, Berit Lilje, Nicolas Rapin, Frederik Otzen Bagger, Mette Jørgensen, Peter Refsing Andersen, Nicolas Bertin, Owen Rackham, A. Maxwell Burroughs, J. Kenneth Baillie, Yuri Ishizu, Yuri Shimizu, Erina Furuhata, Shiori Maeda, Yutaka Negishi, Christopher J. Mungall, Terrence F. Meehan, Timo Lassmann, Masayoshi Itoh, Hideya Kawaji, Naoto Kondo, Jun Kawai, Andreas Lennartsson, Carsten O. Daub, Peter Heutink, David A. Hume, Torben Heick Jensen, Harukazu Suzuki, Yoshihide Hayashizaki, Ferenc Müller, Alistair R.R. Forrest, Piero Carninci, Michael Rehli, and Albin Sandelin. An atlas of active enhancers across human cell types and tissues. Nature, 507(7493), 455–461, March 2014.

[36] Fiona Cunningham, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah E. Hunt, Sophie H. Janacek, Nathan Johnson, Thomas Juettemann, Andreas K. Kähäri, Stephen Keenan, Fergal J. Martin, Thomas Maurel, William McLaren, Daniel N. Murphy, Rishi Nag, Bert Overduin, Anne Parker, Mateus Patricio, Emily Perry, Miguel Pignatelli, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P. Wilder, Amonida Zadissa, Bronwen L. Aken, Ewan Birney, Jennifer Harrow, Rhoda Kinsella, Matthieu Muffato, Magali Ruffier, Stephen M.J. Searle, Giulietta Spudich, Stephen J. Trevanion, Andy Yates, Daniel R. Zerbino, and Paul Flicek. Ensembl 2015. Nucleic Acids Res, 43(D1), D662–D669, Oct 2014.

[37] Jonathan Cairns, Paula Freire-Pritchett, Steven W. Wingett, Csilla Várnai, Andrew Dimond, Vincent Plagnol, Daniel Zerbino, Stefan Schoenfelder, Biola-Maria Javierre, Cameron Osborne, Peter Fraser, and Mikhail Spivakov. CHiCAGO: robust detection of DNA looping interactions in capture hi-c data. Genome Biol, 17(1), jun 2016.

[38] Trieu Tuan, Martinez-Fundichely Alexander, Khurana Ekta. DeepMILO: a deep learning approach to predict the impact of non-coding sequence variants on 3d chromatin structure. Genome Biol 2020;21(1).

[39] Hnisz Denes, Weintraub Abraham S, Day Daniel S, Valton Anne-Laure, Bak Rasmus O, Li Charles H, Goldmann Johanna, Lajoie Bryan R, Fan Zi Peng, Sigova Alla A, Reddy Jessica, Borges-Rivera Diego, Lee Tong Ihn, Jaenisch Rudolf, Porteus Matthew H, Dekker Job, Young Richard A. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science 2016;351 (6280):1454–8.

[40] Jing Fang, Zhang Shao-Wu, Zhang Shihua. Prediction of enhancer-promoter interactions using the cross-cell type information and domain adversarial neural network. BMC Bioinform 2020;21(1).

[41] Shashank Singh, Yang Yang, Barnabás Póczos, and Jian Ma. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. Quant Biol, 7(2), 122–137, jun 2019.

[42] Zengyan Hong, Xiangxiang Zeng, Leyi Wei, and Xiangrong Liu. Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. Bioinformatics, sep 2019.

[43] Min Xiaoping, Ye Congmin, Liu Xiangrong, Zeng Xiangxiang. Predicting enhancer-promoter interactions by deep learning and matching heuristic. Briefings Bioinform 2020;22(4).

[44] Jon-Matthew Belton, Rachel Patton McCord, Johan Harmen Gibcus, Natalia Naumova, Ye Zhan, and Job Dekker. Hi-c: A comprehensive technique to capture the conformation of genomes. Methods, 58(3), 268–276, 2012. 3D chromatin architecture.

[45] Pal Koustav, Forcato Mattia, Ferrari Francesco. Hi-c analysis: from data generation to integration. Biophys Rev 2018;11(1):67–78.

[46] Schoenfelder Stefan, Furlan-Magaril Mayra, Mifsud Borbala, Tavares-Cadete Filipe, Sugar Robert, Javierre Biola-Maria, Nagano Takashi, Katsman Yulia, Sakthidevi Moorthy, Wingett Steven W, Dimitrova Emilia, Dimond Andrew, Edelman Lucas B, Elderkin Sarah, Tabbada Kristina, Darbo Elodie, Andrews Simon, Herman Bram, Higgs Andy, LeProust Emily, Osborne Cameron S, Mitchell Jennifer A, Luscombe Nicholas M, Fraser Peter. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res 2015;25(4):582–97.

[47] Hsieh Tsung-Han S, Weiner Assaf, Lajoie Bryan, Dekker Job, Friedman Nir, Rando Oliver J. Mapping nucleosome resolution chromosome folding in yeast by micro-c. Cell 2015;162(1):108–19.

[48] Schoenfelder Stefan, Javierre Biola-Maria, Furlan-Magaril Mayra, Wingett Steven W, Fraser Peter. Promoter capture hi-c: High-resolution, genome-wide profiling of promoter interactions. J Visual Exp 2018(136).

[49] Li Guoliang, Cai Liuyang, Chang Huidan, Hong Ping, Zhou Qiangwei, Kulakova Ekaterina V, Kolchanov Nikolay A, Ruan Yijun. Chromatin interaction analysis with paired-end tag (chia-pet) sequencing technology and application. BMC Genom 2014;15(12):S11.
[50] Ramsköld Daniel, Wang Eric T, Burge Christopher B, Sandberg Rickard. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput Biol 2009;5(12):e1000598.
[51] Anders S. Visualization of genomic data with the hilbert curve. Bioinformatics 2009;25(10):1231–5.
[52] Monowar Md. Anjum, Ibrahim Asadullah Tahmid, and M. Sohel Rahman. CNN model with hilbert curve representation of DNA sequence for enhancer prediction. Feb 2019.
[53] Patrick Ng. dna2vec: Consistent vector representations of variable-length k-mers. January 2017.
[54] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. January 2013.
[55] Valueva MV, Nagornov NN, Lyakhov PA, Valuev GV, Chervyakov NI. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. Math Comput Simul 2020;177:232–43.
[56] Hochreiter Sepp, Schmidhuber Jürgen. Long short-term memory. Neural Comput 1997;9(8):1735–80.
[57] F.A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: continual prediction with lstm. In 1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470), volume 2, pages 850–855 vol 2, 1999.
[58] Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Trans Signal Process 1997;45(11):2673–81.
[59] Goodfellow Ian, Bengio Yoshua, Courville Aaron. Deep Learning. MIT Press 2016. URL:http://www.deeplearningbook.org.
[60] Masakazu Matsugu, Katsuhiko Mori, Yusuke Mitari, and Yuji Kaneda. Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Networks, 16(5), 555–559, 2003. Advances in Neural Networks Research: IJCNN '03.
[61] Williamson Iain, Hill Robert E, Bickmore Wendy A. Enhancers: From developmental genetics to the genetics of common human disease. Dev Cell 2011;21(1):17–9.
[62] Whalen Sean, Truty Rebecca M, Pollard Katherine S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat Genet 2016;48(5):488–96.
[63] Jennifer E.F. Butler and James T. Kadonaga. Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. Genes Devel, 15(19), 2515–2519, Oct 2001.
[64] Saito Takaya, Rehmsmeier Marc. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE 2015;10(3):e0118432.
[65] Kraft Katerina, Magg Andreas, Heinrich Verena, Riemenschneider Christina, Schöpflin Robert, Markowski Julia, Ibrahim Daniel M, Acuna-Hidalgo Rocío, Despang Alexandra, Andrey Guillaume, Wittler Lars, Timmermann Bernd, Vingron Martin, Mundlos Stefan. Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations. Nat Cell Biol 2019;21(3):305–10.
[66] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. December 2013.
[67] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. December 2015
[68] Kulakovskiy Ivan V, Medvedeva Yulia A, Schaefer Ulf, Kasianov Artem S, Vorontsov Ilya E, Bajic Vladimir B, Makeev Vsevolod J. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. Nucleic Acids Res 2012;41(D1):D195–202.
[69] Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. Structured attention networks. February 2017.
[70] Lovén Jakob, Hoke Heather A, Lin Charles Y, Lau Ashley, Orlando David A, Vakoc Christopher R, Bradner James E, Lee Tong Ihn, Young Richard A. Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell 2013;153(2):320–34.
[71] Yichao Cai, Ying Zhang, Yan Ping Loh, Jia Qi Tng, Mei Chee Lim, Zhendong Cao, Anandhkumar Raju, Erez Lieberman Aiden, Shang Li, Lakshmanan Manikandan, Vinay Tergaonkar, Greg Tucker-Kellogg, and Melissa Jane Fullwood. H3k27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. Nature Commun, 12(1), Jan 2021.
[72] Grubert Fabian, Zaugg Judith B, Kasowski Maya, Ursu Oana, Spacek Damek V, Martin Alicia R, Greenside Peyton, Srivas Rohith, Phanstiel Doug H, Pekowska Aleksandra, Heidari Nastaran, Euskirchen Ghia, Huber Wolfgang, Pritchard Jonathan K, Bustamante Carlos D, Steinmetz Lars M, Kundaje Anshul, Snyder Michael. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. Cell 2015;162(5):1051–65.
[73] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, Susan Sullivan, Jason Bridge, Harold Magazine, John Syron, Johnelle Fleming, Laura Siminoff, Heather Traino, Maghboeba Mosavel, Laura Barker, Scott Jewell, Dan Rohrer, Dan Maxim, Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Eric Hudson, Kristin Feenstra, Leslie Sobin, James Robb, Phillip Branton, Greg Korzeniewski, Charles Shive, David Tabor, Liqun Qi, Kevin Groch, Sreenath Nampally, Steve Buia, Angela Zimmerman, Anna Smith, Robin Burges, Karna Robinson, Kim Valentino, Deborah Bradbury, Mark Cosentino, Norma Diaz-Mayoral, Mary Kennedy, Theresa Engel, Penelope Williams, Kenyon Erickson, Kristin Ardlie, Wendy Winckler, Gad Getz, David DeLuca, Daniel MacArthur, Manolis Kellis, Alexander Thomson, Taylor Young, Ellen Gelfand, Molly Donovan, Yan Meng, George Grant, Deborah Mash, Yvonne Marcus, Margaret Basile, Jun Liu, Jun Zhu, Zhidong Tu, Nancy J Cox, Dan L Nicolae, Eric R Gamazon, Hae Kyung Im, Anuar Konkashbaev, Jonathan Pritchard, Matthew Stevens, Timothèe Flutre, Xiaoquan Wen, Emmanouil T Dermitzakis, Tuuli Lappalainen, Roderic Guigo, Jean Monlong, Michael Sammeth, Daphne Koller, Alexis Battle, Sara Mostafavi, Mark McCarthy, Manual Rivas, Julian Maller, Ivan Rusyn, Andrew Nobel, Fred Wright, Andrey Shabalin, Mike Feolo, Nataliya Sharopova, Anne Sturcke, Justin Paschal, James M Anderson, Elizabeth L Wilder, Leslie K Derr, Eric D Green, Jeffery P Struewing, Gary Temple, Simona Volpi, Joy T Boyer, Elizabeth J Thomson, Mark S Guyer, Cathy Ng, Assya Abdallah, Deborah Colantuoni, Thomas R Insel, Susan E Koester, A Roger Little, Patrick K Bender, Thomas Lehner, Yin Yao, Carolyn C Compton, Jimmie B Vaught, Sherilyn Sawyer, Nicole C Lockhart, Joanne Demchok, and Helen F Moore. The genotype-tissue expression (GTEx) project. Nature Genet, 45(6), 580–585, May 2013.
[74] Lucas J.T. Kaaij, Fabio Mohn, Robin H. van der Weide, Elzo de Wit, and Marc Bühler. The ChAHP complex counteracts chromatin looping at CTCF sites that emerged from SINE expansions in mouse. Cell, 178(6), 1437–1451.e14, sep 2019.
[75] Diehl Adam G, Ouyang Ningxin, Boyle Alan P. Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. Nature Commun 2020;11(1).
[76] Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Seyhan Yazar, Harm Brugge, Roy Oelen, Dylan H. de Vries, Monique G.P. van der Wijst, Silva Kasela, Natalia Pervjakova, Isabel Alves, Marie-Julie Favé, Mawussé Agbessi, Mark W. Christiansen, Rick Jansen, Ilkka Seppälä, Lin Tong, Alexander Teumer, Katharina Schramm, Gibran Hemani, Joost Verlouw, Hanieh Yaghootkar, Reyhan Sönmez Flitman, Andrew Brown, Viktorija Kukushkina, Anette Kalnapenkis, Sina Rüeger, Eleonora Porcu, Jaanika Kronberg, Johannes Kettunen, Bernett Lee, Futao Zhang, Ting Qi, Jose Alquicira Hernandez, Wibowo Arindrarto, Frank Beutner, Peter A.C. 't Hoen, Joyce van Meurs, Jenny van Dongen, Maarten van Iterson, Morris A. Swertz, Marc Jan Bonder, Julia Dmitrieva, Mahmoud Elansary, Benjamin P. Fairfax, Michel Georges, Bastiaan T. Heijmans, Alex W. Hewitt, Mika Kähönen, Yungil Kim, Julian C. Knight, Peter Kovacs, Knut Krohn, Shuang Li, Markus Loeffler, Urko M. Marigorta, Hailang Mei, Yukihide Momozawa, Martina Müller-Nurasyid, Matthias Nauck, Michel G. Nivard, Brenda W.J.H. Penninx, Jonathan K. Pritchard, Olli T. Raitakari, Olaf Rotzschke, Eline P. Slagboom, Coen D.A. Stehouwer, Michael Stumvoll, Patrick Sullivan, Peter A.C. 't Hoen, Joachim Thiery, Anke Tönjes, Jenny van Dongen, Maarten van Iterson, Jan H. Veldink, Uwe Völker, Robert Warmerdam, Cisca Wijmenga, Morris Swertz, Anand Andiappan, Grant W. Montgomery, Samuli Ripatti, Markus Perola, Zoltan Kutalik, Emmanouil Dermitzakis, Sven Bergmann, Timothy Frayling, Joyce van Meurs, Holger Prokisch, Habibul Ahsan, Brandon L. Pierce, Terho Lehtimäki, Dorret I. Boomsma, Bruce M. Psaty, Sina A. Gharib, Philip Awadalla, Lili Milani, Willem H. Ouwehand, Kate Downes, Oliver Stegle, Alexis Battle, Peter M. Visscher, Jian Yang, Markus Scholz, Joseph Powell, Greg Gibson, Tõnu Esko, Lude Franke. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nature Genet, 53(9), 1300–1310, sep 2021.
[77] Matthew Weiser, Sayan Mukherjee, and Terrence S Furey. Novel distal eQTL analysis demonstrates effect of population genetic architecture on detecting and interpreting associations. Genetics, 198(3), 879–893, sep 2014.
[78] Dong Pengfei, Xiaoyu Tu, Chu Po-Yu, Lü Peitao, Zhu Ning, Grierson Donald, Baijuan Du, Li Pinghua, Zhong Silin. 3d chromatin architecture of large plant genomes determined by local a/b compartments. Molecular Plant 2017;10(12):1497–509.
[79] Dong Pengfei, Xiaoyu Tu, Liang Zizheng, Kang Byung-Ho, Zhong Silin. Plant and animal chromatin three-dimensional organization: similar structures but different functions. J Exp Bot 2020;71(17):5119–28.
[80] M. Jordan Rowley, Michael H. Nichols, Xiaowen Lyu, Masami Ando-Kuri, I. Sarahi M. Rivera, Karen Hermetz, Ping Wang, Yijun Ruan, and Victor G. Corces. Evolutionarily conserved principles predict 3d chromatin organization. Molecular Cell, 67(5), 837–852.e7, sep 2017.
[81] Jill M. Dowen, Zi Peng Fan, Denes Hnisz, Gang Ren, Brian J. Abraham, Lyndon N. Zhang, Abraham S. Weintraub, Jurian Schuijers, Tong Ihn Lee, Keji Zhao, and Richard A. Young. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell, 159(2), 374–387, 2014.