Quantitative prediction of enhancer-promoter interactions

Polina Belokopytova, Miroslav Nuriddinov, Evgeniy Mozheiko, et al.

Genome Res. published online December 2, 2019
Access the most recent version at doi:10.1101/gr.249367.119

Quantitative prediction of enhancer-promoter interactions

P.S. Belokopytova[1,2], M.A. Nuriddinov[1], E.A. Mozheiko[1], D. Fishman[2] and V. Fishman[1,2,*]

1. Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

2. Novosibirsk State University, Novosibirsk, Russia

* correspondence should be addressed to Veniamin Fishman, minja-f@ya.ru

Abstract

Recent experimental and computational efforts provided large datasets describing 3-dimensional organization of mouse and human genomes and showed the interconnection between expression profile, epigenetic state, and spatial interactions of loci. These interconnections were utilized to infer spatial organization of chromatin, including enhancer-promoter contacts, from 1-dimensional epigenetic marks. Here we showed that predictive power of some of these algorithms is overestimated due to peculiar properties of biological data. We proposed an alternative approach, which allowed obtaining high-quality predictions of chromatin interactions using information on gene expression and CTCF-binding alone. Using multiple metrics, we confirmed that our algorithm could efficiently predict 3-dimensional architecture of both normal and rearranged genomes.

Introduction

Spatial interactions between promoters and their regulatory sequences are required to maintain a cell type-specific expression pattern (Rao et al. 2014). It's known enhancers do not necessarily regulate closest promoters, and enhancer-promoter (EP) interactions often span large genomic distances (Rao et al. 2014; Tang et al. 2015). Although enhancer targets can be directly identified when using high-resolution 3C-methods (Rao et al. 2014), these data is costly and currently available only for a small subset of cell types. Besides, experimental identification of enhancer targets does not provide mechanism explaining target selection.

Several computational tools were developed to address these challenges. Their task was to predict 3-dimensional EP interactions, based on data on 1-dimensional genetic and epigenetic marks (Buckle et al. 2018; Zeng et al. 2018; Whalen et al. 2016; Ibn-Salem and Andrade-Navarro 2018; Chiariello et al. 2016; Chen et al. 2016; Moore et al. 2015; Di Pierro et al. 2017; Kai et al. 2018; Zhang et al. 2018a; Zhu et al. 2016; Al Bkhetan and Plewczynski 2018; Fortin and Hansen 2015; Qi and Zhang 2019). All these tools fall into two categories: physical models and statistical approaches. Former rely on knowledge of polymer physics to build a physical model of chromatin, and optimize the model parameters to fit experimental (usually Hi-C) data (Chiariello et al. 2016; Buckle et al. 2018; Di Pierro et al. 2017). The optimized model can be used to infer spatial conformation of chromatin, including those regions containing EP interactions. In contrast, statistical methods do not imply any *a priori* knowledge of polymer physics, aiming to find consistent patterns in epigenetic data which would explain 3-dimensional contacts of loci (Zeng et al. 2018; Di Pierro et al. 2017; Whalen et al. 2016; Ibn-Salem and Andrade-Navarro 2018; Chen et al. 2016; Moore et al. 2015; Kai et al. 2018; Zhang et al. 2018a). Thus, statistical approaches are able to predict spatial contacts of chromatin even without complete knowledge of physical mechanisms underlying 3-dimensional organization of genome.

Here, we aimed to infer 3-dimensional interactions of chromatin, and particularly promoter-enhancer interactions, in normal and rearranged genomes, using available epigenetic data. We benchmarked existing statistical approach and found that its predictive power is overestimated due to peculiar properties of biological data. Thus, we have taken a challenge to develop a new machine-learning algorithm for quantitative prediction of genome architecture based on broadly available epigenetic datasets.

Results

TargetFinder fails to predict EP interactions. Our objective was to develop an algorithm for prediction of enhancer-promoter interactions in normal and rearranged genomes. For this aim, we decided to employ existing TargetFinder algorithm (Whalen et al. 2016), which is of particular interest, because of high accuracy, low false-discovery rate, and reproducibility, demonstrated by an analysis of several human cell types. Since several well-studied examples of chromosomal rearrangements causing changes of chromatin architecture are investigated using mouse models (Spielmann et al. 2018; Fishman et al. 2018), we aimed to extend TargetFinder algorithm for prediction of EP interactions in mouse cells.

We annotated promoters and enhancers as interacting and non-interacting using high resolution Hi-C data on mouse ES cells (Bonev et al. 2017) and collected a set of 24 genetic and epigenetic predictors. To construct our datasets, we used an original definition of “interacting” promoters and enhancers, proposed in TargetFinder paper (Whalen et al. 2016), i.e. promoter and enhancer were considered as interacting only if they were located in the anchors of a Hi-C loop. The accuracy of TargetFinder (measured by either precision, recall or F1-score on validation dataset) was lower than previously reported on human data (Table 1). We found out that changing ratio of interacting to non-interacting EP pairs from 1:20 to 1:1 increases F1-scores; however, obtained values were still below than reported previously (Whalen et al. 2016) (Table 1). We additionally run TargetFinder on mouse cortex and neural stem cells (NPC) data (Bonev et al. 2017) using 10 available epigenetic predictors. As on ES cells data, TargetFinder was not efficient on these datasets (Table 1).

To understand why TargetFinder fails to predict EP interactions, we re-processed original human data generating predictors, training and validation datasets for human GM12878 cells *de novo*. Running TargetFinder on these re-processed human datasets resulted in low F1-scores, with only small improvement comparing to mouse ES cells data (Table 1).

Comparing our protocol of data processing with the pipeline that was used to generate original TargetFinder datasets, we noticed the difference in composition of training and validation samples. In the original approach, EP pairs were randomly split to obtain training (~90% of data) and validation (~10% of data) datasets. Our pipeline randomly selects two chromosomes and designs all EP pairs on these chromosomes as validation dataset (~10% of all data), and the rest of EP pairs as training dataset. This difference in design of training and validation datasets is essential, because when we performed by-chromosome split of original TargetFinder data on human GM12878, F1-scores reduced a lot comparing to random-split, and become similar to those obtained for mouse data (Table 1). Moreover, when we used random-split strategy on mouse data, F1-scores increased substantially (Table 1). Thus, the TargetFinder performance strongly depends on design of training and validation datasets.

To explain observed effect of data splitting strategy on TargetFinder efficiency, we explored the structure of EP datasets. We found that ~70% of GM12878 promoters interact with multiple enhancers, which are located close to each other. Such EP pairs share a large portion of genomic region between promoter and enhancer (referred hereinafter as "window"). In general, overlaps of EP windows are frequent (>99% of all pairs share a window with at least one another pair) and overlap size is often large. Thus, epigenetic predictors characterizing window of these pairs are not independent, and EP-pairs with shared window should not be placed in both training and validation dataset (Fig. 1, A and B). As this happens when employing random-split strategy, TargetFinder could match overlapping samples in training and validation sets based on window information and then copy the information about interactions from the pair in training set to the pair in validation set. One should note that patterns of spatial contacts of neighboring genomic regions are correlative. Thus, interactions of two EP pairs, one from the training set and another from the validation set, are often similar if both promoters and enhancers are located nearby (i.e. if window overlap is high). To confirm this, for each EP pair we explicitly used interaction of the EP pair with the highest window overlap as a predictor, and obtained high F1-score (~0.9) for GM12878 cells. Moreover, for NPC and mouse cortex datasets, which contain approximately three times less interacting EP pairs than mouse ES cells and human GM12878 cells (Table 1), and therefore the lowest rate of overlapping windows, we obtained the lowest F1-scores in the random-split design.

To sum up, random-split strategy breaks assumption of independence of samples in training and validation datasets and thus results in overestimation of the predictive power of machine-learning algorithms. In contrast, when using by-chromosome splitting strategy, genomic regions never overlap between training and validation datasets, allowing unbiased estimation of algorithms efficiency. Considering that TargetFinder efficiency drops when using by-chromosome data splitting, we claim that this algorithm cannot reconstruct relations between 1D-genetic marks and 3D-genome organization. Moreover, while this manuscript was in preparation Xi and Beer (Xi and Beer 2018) independently concluded that local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy.

We note that other published algorithms also use random-split design of training and validation datasets (see discussion). Thus, our results highlighted specific peculiarities of biological data, which should be considered in future to prevent overfitting issues and wrong efficiency estimation of machine-learning approaches focused on predictions of chromatin interactions.

Enhancer-promoter interactions are quantitative rather than qualitative.

As we found that TargetFinder cannot efficiently predict EP interactions, we aimed to improve this algorithm. We considered following enhancements.

First, we decided that epigenetic marks not only between, but also outside of, promoter and enhancer should be considered. This makes sense in light of recently discovered mechanisms, underlying spatial organization of chromatin. For example, according to the loop extrusion model (Fudenberg et al. 2017), binding of CTCFs in the converged orientation outside of, but close to, EP pair will result in increased looping between promoter and enhancer. Based on the loop extrusion model, we also introduced orientation of CTCF sites as a predictor.

Second, we reconsidered the definition of the enhancer-promoter *interaction*. TargetFinder and other approaches (Li et al. 2019), define EP pair as interacting only if enhancer and promoter occur within anchors of a Hi-C loop. To benchmark this approach, we collected all interacting EP pairs for human monocytes based on SlideBase and GeneHancer databases (see methods for details). In addition to promotor-capture 3C data, these databases utilize information of co-expression of promoters and regulatory elements, their distance and other information to define interacting EP pairs. We used human monocytes data because these cells are represented in SlideBase and GeneHancer database and characterized by high-resolution Hi-C (Phanstiel et al. 2017), which allowed comparison of Hi-C loops and interacting EP pairs. We confirmed that contact frequencies between interacting EP pairs, as well as between loop anchors, are higher than average (Fig. 2 A,B). However, vast majority of interacting EP pairs do not overlap with loops, although they are often located within reasonable distance from loop anchors (Fig. 2 C). Similar results were obtained for human macrophages (Supplementary Fig. 1).

The set of EP pairs described in SlideBase and GeneHancer is probably not complete, and these databases (partially) rely on the 3C information to infer EP connections. The gold standard for identification of functional EP interactions is direct genetic screening. Such screenings are expensive and time-consuming, thus there is very limited number of experimentally validated enhancers, which prevents systematic genome-wide analysis of their relations to Hi-C loops. However, individual reports of genetically validated functional EP interactions support our general conclusion. For example, a recent study (Fulco et al. 2016) identified seven distal enhancers of *MYC* gene, which form two clusters located 0.16 and 1.9 Mb away from the *MYC* promoter, respectively. According to (Fulco et al. 2016), all of these enhancers affect *MYC* expression in K562 cells proportionally to the number of contacts between enhancer and *MYC* promoter in this cell type. However, we found that only *e6* and *e7* enhancers overlap Hi-C loops (Fig. 2 D). Moreover, out of five loops containing *MYC* promoter only one contains validated *MYC* enhancers. Altogether, this means that binary classification of EP interactions guided by location of Hi-C loop anchors may have poor predictive power. These observations are consistent with the recent model of enhancer-promoter communication (Furlong and Levine 2018), which suggests that loops and domains serve to decrease effective distance separating enhancers and promoters, but are not necessarily formed by EP pairs themselves.

Thus, we concluded that increased interaction frequency, rather than location within loop anchors, should be used to characterize EP interaction. As spatial interactions are quantitative, we aimed to design quantitative algorithm which predicts frequencies of spatial interactions between genomic loci in general, and EP interactions in particular.

Quantitative prediction of EP interactions using machine learning tool 3DPredictor. We used following biological information to predict EP and other genomic interactions: ChIP-seq profiles, describing chromatin binding of architectural proteins or histone modifications; RNA-seq profiles, describing gene expression levels; E1 values, classifying chromatin to active (A) and inactive (B) compartments; genomic distance, which is an essential factor of 3-dimensional contacts. We restricted our algorithm to the prediction of mid-range contacts (<= 1.5 Mb) since almost all EP interactions occur within this distance. To increase sample size and avoid overfitting, we included contacts of all loci, regardless of the presence of promoters and enhancers, into the training set. We always performed training and validation on different chromosomes, and never used chromosome number or genomic coordinates of loci as predictors, to prevent overfitting.

Using recently generated Hi-C and genomic data on mouse hepatocytes and human K562 and GM12878 cells, we compared several forms of predictors parametrization and performance of different machine learning algorithms (see Supplementary Note, Supplementary Fig. 9-12 and Supplementary Tables 2-3 for details). To estimate quality of predictions, we used Pearson's correlation, stratum adjusted correlation coefficient (SCC) (Yang et al. 2017a), mean squared error (MSE), mean absolute error (MAE) and mean relative error (MRE). As a result, we developed 3DPredictor, a machine-learning tool for computational prediction of chromatin interactions. Analyses of importance of different epigenetic features showed that information about cohesin and CTCF-binding, gene expression, chromatin accessibility and distance between loci has the greatest contribution to the prediction accuracy (see Supplementary Fig. 2, Supplementary Table 1 and Supplementary Note). Moreover, according to the feature importance analysis epigenetic characteristics of the region between interacting loci are essential for accurate prediction (Supplementary Fig. 2), which supports previously obtained (Whalen et al. 2016) conclusion that there is significant information relevant to looping interactions outside the interacting loci themselves.

Although various epigenetic information contributed to the prediction of chromatin contacts, it appeared that many predictors are interchangeable. We were able to generate accurate predictions of chromatin interactions in mouse hepatocytes (Pearson's R=0.92-0.95, SCC=0.53-0.72, MSE=0.0017-0.0082, MAE=0.0010-0.0015, MRE=0.52-1.74; Fig. 3 A-F), limiting input information to CTCF ChIP-seq data (including orientation of the occupied CTCF-sites), RNA-seq data and distance between loci, and using only one chromosome out of 20 for training. These results can be further improved using multiple chromosomes for training (Fig. 3 B-F). Orientation of CTCF-sites was among features with highest importance, and omitting this information impaired predictions of loops (Supplementary Fig. 2 and 3). Thus, we used information about CTCF binding, RNA-seq and genomic distance for all predictions in this paper.

Chromatin contacts are known to be moderately similar between cell types (Battulin et al. 2015; Rao et al. 2014). To find whether our predictions are cell type-specific, we first compared chromatin architecture of different cell types using aforementioned measures. In most cases, results obtained by 3DPredictor differ from real data less than cell types differ from each other (Fig. 3 B-F). For example, for 13 chromosomes 3DPredictor results, judged by mean average error, resemble experimental hepatocyte' Hi-C data more closely than experimental data derived from other studied cell types. For four more chromosomes (Chromosomes 1, 4, 6, 9 and 15) prediction errors were comparable with MAE obtained for different cell types, and on Chromosomes 19 and X predictions were worse than transferring contact counts from other cell types. Similar results were obtained for MSE and MRE. According to the SCC, 3DPredictor performs approximately at the level of intercellular differences, whereas according to the Pearson's correlation predictions were almost always more similar to the hepatocyte's data than other cell types.

We next compared Hi-C data of mouse hepatocytes and NPC and found that some genomic regions show apparently different 3D-organization in these cell types. In most cases, the differences were due to the presence of cell type-specific TADs, which borders coincide with cell type-specific CTCF sites, as was observed previously (Bonev et al. 2017; Rao et al. 2014). We utilized an insulation-based score to select genomic regions with cell type-specific chromatin architecture (see Methods for details), and run 3DPredictor for these regions using cell type-specific RNA-seq and CTCF ChIP-seq data. Predicted contact frequencies reflected cell type-specific genome organization (Fig. 4, A-C, Supplementary Fig. 4), and correlation of insulation scores of predicted and experimental data was much higher than between cell types (Fig. 4, D). We provided an example of accurate prediction of cell type-specific TAD boundary in NPC and hepatocytes in Fig. 4 B and C and Supplementary Fig. 4.

Finally, we run 3DPredictor on human GM12878 data. According to all metrics except SCC, predictions fit experimentally-derived Hi-C interactions better than data from other cell types, even when using single chromosome for training (Supplementary Fig. 5). At the same time, however,

transferring interaction frequencies from other cell types results in better SCC values comparing to the predictions with only one exception on the Chromosome 9, and, in general, SCC values obtained on human data were slightly lower than obtained on mouse data.

When focused on EP contacts, we found that for this specific set of interactions predictions accuracy was same as for other interactions. MRE of contact frequencies for interacting (according to SlideBase and GeneHancer databases) EP pairs was slightly lower than for all chromatin interactions predicted in monocytes, and MSE and MAE slightly higher (Fig. 5, A and Supplementary Fig. 6). In general, experimentally-derived contact frequencies of EP pairs in monocytes were highly correlated with predicted contact frequencies for corresponding loci in these cells (Fig. 5, B). We defined cell type-specific EP interactions (see methods), to examine whether 3DPredictor captures differences in EP interactions between cell types. As for the cell type-specific TADs, the difference between predicted and experimentally-measured EP interactions was smaller than between interactions of these enhancers and promoters measured in different cell types (Fig. 5, C). These results have also been confirmed using mouse data (Supplementary Fig. 7), and both low-resolution (25 kb, Fig. 5 and Supplementary Fig. 7) and high-resolution (5 kb, Supplementary Fig. 6 and 7) models.

We next used 3DPredictor trained on mouse hepatocytes data (single chromosome or half of the genome) to predict contact frequencies in mouse NPC (Fig. 6, A). Predictions of spatial interactions for the cell type that was not used for training appeared to be as good as when the same cell type was used for training and validation (Fig. 6, B-F). From the practical point of view, this indicates that our approach can be used to predict 3-dimensional genome organization, including EP contacts, in those cell types where 3C-data is not available. From the fundamental standpoint, these results show that principles of genome architecture are very similar in different cell types.

Comparing 3DPredictor with other models. There are several computational tools which could quantitatively predict short- and mid-range chromatin interactions (see discussion for comprehensive comparison of these tools). For example, MEGABASE+MiChroM (Di Pierro et al. 2017) predicts chromatin interactions at 50 kb resolution using information about epigenetic marks and CTCF-loops. Whereas modeling of CTCF-mediated looping interactions requires Hi-C data to infer loop anchors, use of the reduced MiChroM Hamiltonian lacking the term in that energy function that models the CTCF-mediated looping interactions can be used to predicted chromatin contacts without any experimental measurements of 3D genome organization (Di Pierro et al. 2017). We benchmarked 3DPredictor against this reduced MEGABASE+MiChroM model, and found that 3DPredictor significantly outperforms it, showing much higher SCC (Supplementary Table 5 and Supplementary Fig. 13 A-C). However, we wish to note that MEGABASE+MiChroM was originally developed to capture long-range interactions mediated by chromatin compartmentalization, and lack of information about CTCF-mediated loops could explain, at least partially, poor performance of short-range interactions prediction.

Qi and Zhang have recently proposed another model based on polymer physics to predict Hi-C interactions using epigenetic data (Qi and Zhang 2019). In contrast to the full MEGABASE+MiChroM model, Qi and Zhang do not use experimental 3C-information to define CTCF-mediated loop anchors, requiring only ChIP-seq data and genomic sequence to describe CTCF binding landscape. When employing the approach proposed by Qi and Zhang to infer chromatin contacts in GM128787 cells, we obtain better results comparing to the reduced MEGABASE+MiChroM model (Supplementary Tables 5, 7). However, performance of 3DPredictor on the same dataset was even higher, judged by SCC, Pearson's correlation, MSE and MAE (Supplementary Fig. 15 A-C and Supplementary Table 7). It is worth pointing out that the polymer models were developed with 3D structures in mind, and are useful for studying compartmentalization and higher-order contacts as well.

Statistical approach showing that CTCF-looping and gene expression could explain chromatin contacts in mammalian cells was recently proposed by (Rowley et al. 2017). This approach requires very limited amount of information as an input; however, similarly to the full MEGABASE+MiChroM model, it

cannot be used to predict chromatin interactions, because information about CTCF-looping should be extracted from experimental Hi-C data. For example, in the region of Chromosome 4 of GM12878 cells, analyzed by (Rowley et al. 2017), their model uses only 63 manually-selected CTCF sites, which comprise ~35.4% of all CTCF-bound sites in this region (Supplementary Fig. 16). Moreover, the (Rowley et al. 2017) approach requires Hi-C information to define pairs of interacting CTCF-sites. This information can not be trivially obtained from ChIP-seq data, because in some cases loops are formed between distal CTCF-bound sites, skipping the nearest CTCF-bound site in convergent orientation (see Supplementary Fig. 16 for representative examples and (Kai et al. 2018) for systematic analysis). Nevertheless, we compared 3DPredictor with the (Rowley et al. 2017) model and found that latter gives significantly better results (Supplementary Table 6 and Supplementary Fig. 14 A-C). Consistent with the fact that (Rowley et al. 2017) derived information about CTCF-mediated looping from the experimental data, their model captures experimental loops especially well (Supplementary Fig. 14 C). Although 3DPredictor does not require any experimental 3C-information, it also captures approximately half of the looping interactions (Supplementary Fig. 17), and predicted frequencies of contacts between loop anchors were higher than between other genomic regions (Supplementary Fig. 14 D).

Predicting effects of chromosomal rearrangements on 3-dimensional genome organization. One of the applications of enhancer targets prediction is understanding of EP rewiring after chromosomal rearrangements. There are several well-studied examples of pathological changes in EP contacts caused by deletions, inversions (Lupiáñez et al. 2015) or duplications (Franke et al. 2016). Recently, PRISMR (Bianco et al. 2018) was developed to resolve chromatin structure of rearranged genome. Although impressively accurate, PRISMR requires Hi-C data to optimize chromatin model parameters, which limits its usage to cell types with available Hi-C data or genomic regions with 3-dimensional structure conserved across cell types. 3DPredictor lacks these limitations, as we have shown that it can predict chromatin packaging of cell type-specific regions and previously unstudied cell types.

We employed recently generated 5C data describing mouse *Epha4* rearrangements to find whether 3DPredictor can infer ectopic interactions in the mutated genome. We re-analyzed 5C data generated from wild-type cells, as well as cells carrying homozygous deletion of ~1.5 Mb encompassing *Epha4* gene (Lupiáñez et al. 2015). This deletion (referred as *DelB* in (Bianco et al. 2018)) results in establishment of ectopic contacts between *Pax3* gene and *Epha4* enhancers cluster, which is associated with *Pax3* misexpression leading to brachydactyly.

We run 3DPredictor trained on mouse hepatocytes to infer 3-dimensional organization of the rearranged *Epha4* locus in hindlimb cells. We did not use any *a priori* knowledge of the 3-dimensional structure of wild-type *Epha4* locus in hindlimb cells, yet 3DPredictor results were very similar to experimental data (Fig. 7 A). We used method described in (Bianco et al. 2018) to find ectopic interactions in the rearranged locus. Out of 1561 interactions inferred from the experimental data, 589 were captured by 3DPredictor, including majority of interactions between *Pax3* gene and *Epha4* enhancers (Fig. 7 A, B). The overlap between real and predicted ectopic interactions was large and differed significantly from randomized control (Fig. 7 C, p-value $< 5\times10^{-6}$). This shows that our model successfully predicts ectopic interactions in the rearranged genome.

Discussion

Machine-learning approaches are actively employed to capture complex epigenetic signatures underlying chromatin contacts. As we have shown here, biological data may have specific structure, which should be accounted when designing computational experiments. For example, pairs of loci with overlapping windows partially share epigenetic environment and often display similar 3-dimensional architecture. This means that these regions can not represent independent samples in training and validation datasets, and correlations captured by machine-learning approaches do not reflect causation

underlying genome architecture if overlapping regions are present in both training and validation datasets..

We benchmarked TargetFinder because it's often cited as straightforward tool and employed for prediction EP interactions (Yang et al. 2017b; Zhu et al. 2018; Stricker et al. 2016; Atlasi and Stunnenberg 2017; Moorthy et al. 2017; Wu et al. 2018; Gudmundsson et al. 2017); however this is not the sole example of research that does not take account of this peculiarity of biological data. For instance, recently published EP2vec (Zeng et al. 2018) utilizes the same dataset as TargetFinder and constructs training and validation samples in the same way. Another tool aimed to predict CTCF loops CTCF-MP (Zhang et al. 2018a) does not take into consideration nested loops when employs window features. Although both EP2Vec and CTCF-MP can generate predictions without window information, performance of such a setup is lower: ~10% of accuracy drops when CTCF-MP *is* trained without DNase I and ChIP-seq window features and ~2-4% of F1-score drops when EP2Vec is trained without TargetFinder-derived window features.

In the recent preprint describing a tool for HiC-data prediction HiC-Reg (Zhang et al. 2018b) authors also show that sharing genomic regions between training and validation datasets improves prediction scores. However, the authors connect this observation to a chromosome-specific biological mechanism, which cannot be modeled when overlapping data is omitted from validation set. Whereas chromosome- and even region-specific mechanisms of DNA-packaging indeed exist (Jiang et al. 2017), and we also shown that prediction is better when multiple chromosomes are used to train the model, better results of intrachromosomal cross-validation are likely to originate from existence of overlapping regions. One should note, that although pairs with overlapping left or right anchors are not shared between training and validation datasets in HiC-Reg (so-called easy samples), authors do not exclude regions, which share a part of window between interacting anchors.

We next raised the question of definition of promoter-enhancer *interaction*. Currently, most of the studies use all 3C-interactions, which differ statistically from distance-adjusted background as functional EP interactions (Mishra and Hawkins 2017; Whalen et al. 2016). According to our results, functional interactions of promoters and enhancers do not fully overlap with Hi-C loops, and, probably, do not overlap completely with any other set of enriched interactions. Whereas spatial proximity is required for EP communication, it is not clear which spatial distance is necessary and sufficient to achieve functional interaction. For example, the recent study of *Shh-ZRS* TAD showed that nearly entire ~900 kb intra-TAD region can be activated by *ZRS* enhancer, although pronounced looping was observed only between *Shh* promoter and *ZRS* enhancer (Symmons et al. 2016). Removing *Shh-ZRS* TAD boundary reduces intra-TAD contact frequencies to the background level and disturbs *Shh* expression in the developing limbs; however, relocating enhancer closer to *Shh* promoter region restores expression pattern. These results indicate that background-level interactions within TAD might be sufficient to establish functional connections of promoter and enhancer. Moreover, recent paper reports that intra-TAD promoter regions often show significant level of interaction with TAD boundaries, and disruption of these interactions does not lead to changes of expression levels (Sun et al. 2019). To sum up, our view is that statistical increase of spatial contact frequencies, i.e. formation of loops, is important indicator of promoter-enhancer connectivity, but cannot be solely used to distinguish functional interactions. In accord with this, recent large-scale CRISPR assay of promoter-enhancer connections (Fulco et al. 2019) suggested quantitative “contact-by-activity” model of EP interaction. In this model, enhancer impact is quantitative and proportional to both promoter-enhancer proximity and enhancer activity. Whereas latter can be estimated using DNase I or ATAC-seq data available for many cell types, here we described 3DPredictor, which can be used to quantitatively predict spatial architecture of chromatin, including enhancer-promoter interactions, to supplement this ATAC-seq or DNA-seq data.

There are many methods published previously for prediction of TAD boundaries, Hi-C interactions and enhancer-promoter interactions (Xu et al. 2018). CITD (Chen et al. 2016), MEGABASE (Di Pierro et al. 2017), EpiTensor (Zhu et al. 2016) and model described in (Qi and Zhang 2019) can predict 3D-

interactions, including EP interactions, at various resolutions based on epigenetic data. However, these tools require large amount of epigenetic information: 5-10 patterns of histone modifications for CITD; 11 for MEGABASE; 12 for (Qi and Zhang 2019); 16 histone modifications and additional data for EpiTensor. CISD (Zhang et al. 2017) and PRISMR (Bianco et al. 2018) are able to infer chromatin contacts genome-wide as well, but require Hi-C data as an input. In contrast, 3DPredictor could make predictions when supplied with CTCF- and RNA-seq data only.

Chromatin simulations proposed by Rowley et al. (Rowley et al. 2017) also utilize CTCF and transcription data only; however, they require manual selection of interacting CTCF-sites based on Hi-C data, thus making it impossible to predict 3D-interactions from epigenetic data alone.

3DEpiLoop (Al Bkhetan and Plewczynski 2018), Lollipop (Kai et al. 2018), CTCF-MP (Zhang et al. 2018a), BART model (Huang et al. 2015) and other tools (Fortin and Hansen 2015), (Jenkinson et al. 2017), (Al Bkhetan and Plewczynski 2017) could predict specific chromatin features, such as TAD boundaries, A/B-compartments, CTCF-interactions or loops. However, in contrast to 3DPredictor these approaches 1.) Do not infer EP interactions 2.) Perform qualitative, rather than quantitative prediction (i.e. classification) 3.) Most of them require significantly more input information than 3DPredictor.

There are multiple computation tools design specifically to infer EP interactions, for example (Whalen et al. 2016; Zeng et al. 2018; Li et al. 2019; O'Connor et al. 2017; Cao et al. 2017; Hait et al. 2018). However, all these tools are fundamentally different from 3DPredictor, as they consider EP interaction as qualitative, rather than quantitative. Moreover, most of them require large amount of epigenetic data to make prediction, and performance of some of them (Whalen et al. 2016; Zeng et al. 2018) might be overestimated as discussed above.

To sum up, 3DPredictor is the unique tool, which allows predicting large set of interactions, including EP interactions, quantitatively, using only small amount of input epigenetic data.

It is essential that our model not only predicts chromatin interactions in normal genome, but could also capture ectopic interactions, which are formed as a result of chromosomal rearrangements. Currently, both experimental data describing 3D-genome alterations associated with known rearrangements and tools modeling spatial landscape of novel variants are limited. In the same time, others (Lupiáñez et al. 2015; Franke et al. 2016; Zepeda-Mendoza et al. 2017; Redin et al. 2017) and we (Gridina et al. 2018) have recently reported novel variations with unexpected pathological phenotype, which might be explained, at least partially, by changes of chromatin organization (Spielmann et al. 2018; Fishman et al. 2018). Future development and validation of models predicting chromatin contacts in rearranged genome is essential for better understanding of biomedical consequences of these rearrangements. Moreover, integrating chromatin interactions, derived from 3DPredictor, with enhancer activity information using the "activity-by-contact" model may allow precise estimation of transcriptional changes caused by structural variations.

Materials and Methods

Hi-C data processing.

Hi-C data for mouse hepatocytes (GSE95116) and cardiomyocytes (SRX2658510) were downloaded from NCBI and processed using Juicer (Durand et al. 2016b). Resulting .hic-files are deposed at genedev hic-file server (http://genedev.bionet.nsc.ru/site/hic_out/) under accessions "Hepat" and "CardioMyo". Hi-C data for mouse ES cells, NPC, cortex (Bonev et al. 2017), CH12.LX lymphocytes (Rao et al. 2014) and human GM12878, K562, IMR-90, NHEK (Rao et al. 2014), macrophages and monocytes (Phanstiel et al. 2017) are available at AidenLab hic-file server via Juicebox and Juicer Tools (Durand et al. 2016a, 2016b). All datasets were KR-normalized. For each Hi-C dataset, contacts were obtained at 25 or 5 kb resolutions using Juicer Tools *dump* command. To be able to perform comparisons between cell types, we normalized datasets dividing each contact by normalization coefficient *Coef,* which reflects average bin coverage:

$Coef = \frac{\sum_{i \in K} \quad \sum_{i<j<=N} \quad C_{i,j}}{N}$, where $C_{i,i}$- contacts between i-th and j-th bins, K - number of bins on Chromosome 1, *N* - number of bins in genome. To speed up *Coef* computation we only used bins of Chromosome 1, although this should not affect results as we use KR-normalized matrices where coverage of all bins are roughly equal.

Loops were called by Juicer Tools *HiCCUPS* command with default parameters using heatmaps at 25 or 5 kb resolution. K562 loops presented in Fig. 2 are from (Rao et al. 2014).

First eigenvector (E1) values of Hi-C matrixes were obtained using Juicer Tools *eigenvector* module.

5C data describing 3-dimensional organization of wild-time and mutated mouse *Epha4* locus in distal limb buds were downloaded from GEO: GSE92291. Data was processed by HiCPro (Servant et al. 2015) pipeline using mm10 genome.

The relative error of Hi-C contact counts shown at Supplementary Fig. 3 was estimated based on binomial distribution: $RE = \sqrt{\frac{1}{N}} \times 100\%$, where *N* is a number Hi-C reads between contacted loci. The average and standard deviation of relative errors were independently calculated for each genomic distance.

To estimate correlation of contact counts on different resolutions for Supplementary Fig. 9, we used data for Chromosome 10 of GM12878 cell line. We randomly choose 1000 loci pairs and calculated Pearson's correlation between KR-normalized contact frequencies on different resolutions. We aggregated calculations of 100 independent samplings by averaging to obtain final results.

Definition of promoters and enhancers

For human macrophages and monocytes enhancers were defined using SlideBase (http://slidebase.binf.ku.dk) database. This database is supported by the FANTOM5 consortium (http://fantom.gsc.riken.jp/data/, (Andersson et al. 2014)) and represents a map of human regulatory elements of each cellular state. It contains levels of enhancer expression based on CAGE sequencing of RNA isolated from every major human organ, over 200 cancer cell lines, 30 time courses of cellular differentiation, mouse developmental time courses and over 200 primary cell types. Thereby, an enhancer can be specific to a set of primary cells and organs (tissue samples) or can be broadly (or ubiquitously) expressed. We took into account enhancers displayed in more than 25% samples related to the target cell line.

Using GeneHancer database (https://www.genecards.org), we defined gene promoters regulated by given enhancer. GeneHancer is a database of genome-wide enhancer-to-gene and promoter-to-gene

associations, embedded in GeneCards. GeneHancer EP associations were generated using following information:

1. eQTLs (expression quantitative trait loci) from GTEx
2. Capture Hi-C EP long range interactions
3. Expression correlations between eRNAs and candidate target genes from FANTOM5
4. Cross-tissue expression correlations between a transcription factor interacting with an enhancer and a candidate target gene
5. GeneHancer-gene distance-based associations, scored utilizing inferred distance distributions. Associations include several approaches: (a) Nearest neighbors, where each GeneHancer is associated with its two proximal genes; (b) Overlaps with the gene territory (intragenic); (c) Proximity to the gene TSS (<2 kb). TSS proximity scores are boosted to elevate GeneHancer associations in the vicinity of the gene TSS.

The "true" interacting EP pairs of human monocytes and macrophages were calculated by combining the list of cell type-specific enhancers from SlideBase and list of enhancer-gene associations from GeneHancer.

When using TargetFinder pipeline on human data, we used the authors' definition of active promoters and enhancers, and obtained coordinates from https://github.com/shwhalen/targetfinder. For mouse data we first obtained promoters using TSS (Transcription Start Sites) downloaded from UCSC, and active enhancers based on annotations from (Bogu et al. 2016). Next, we defined interacting pairs as promoters and enhancers located within the anchors of one loop.

ChIP-seq data processing. All ChIP-seq data for human GM12878 and K562 cell lines were downloaded from https://github.com/shwhalen/targetfinder. ChIP-seq data for mouse hepatocytes (NCBI SRX2578761-SRX2578762), mouse NPC (NCBI SRX2636706-SRX2636707, ENCODE ChIP-seq data for forebrain embryo 13.5, GSE96107, GSE96107), mouse cortex (ENCODE ChIP-seq data for forebrain embryo 13.5, GSE96107, GSE96107) and human monocytes (ENCODE ENCSR000ATN) were downloaded from NCBI or ENCODE and processed using aquas pipeline (https://github.com/kundajelab/chipseq_pipeline). CTCF motif oritentation was defined using GemmeMotifs (van Heeringen and Veenstra 2011) software.

RNA-seq data processing. RNA-seq data for human GM12878 (ENCODE ENCFF212CQQ) and human K562 (ENCODE ENCFF026BMH) cell lines were downloaded from ENCODE. RNA-seq data for mouse hepatocytes (NCBI GSE95111) and mouse NPC (NCBI GSM2533845) were downloaded from NCBI. Data for human monocytes (NCBI SRX2785183) were downloaded from NCBI and processed using standard protocols with HISAT2 and StringTie (Pertea et al. 2016).

Prediction of 3-dimensional interaction frequencies.

For training purposes, all data was split into non-overlapping genomic intervals. Usually, we use one or several chromosomes for training, and other chromosomes for validation. To perform prediction genome-wide, we first used odd chromosomes for training and made predictions for contacts on even chromosomes, and then used even chromosomes for training and predicted contacts on odd chromosomes. If other is not mentioned, we used only CTCF and RNA-seq data for predictions. For all results except those described in Supplementary Note we used Gradient Boosting with parameters n_estimators=100, max_depth=9, subsample=0.7. Predictors parametrization and other details are explained in details in the Supplementary Note.

Estimating predictions efficiency

We used several metrics to choose the best model. Pearson's correlation is the most common metric; however, Pearson's correlation is dominated by dependence of contact frequencies from genomic distance. Thus, we also used the SCC metric (Yang et al. 2017a) which measures correlation of contact frequencies on each diagonal of Hi-C matrix independently, thereby neglecting the factor of

genomic distance. To reduce the effect of random noise, we smoothed Hi-C matrices before calculating SCC, as was suggested by (Yang et al. 2017a). All comparisons were carried out with the same noise smoothing parameter *h = 2* (see Supplementary Note and Supplementary Table 4 for justification of *h* value). In addition, to evaluate the model's quality, we used other metrics such as MSE, MAE and MRE.

To benchmark model predictions against transfer of contact counts from another cell type, we performed pairwise comparisons of mouse Hi-C data (CH12.LX lymphocytes, cortical neurons, cardiomyocytes cells and hepatocytes) using same metrics as described above. Similarly, we compared human NHEK, K562, IMR-90, GM12878 to benchmark predictions of human data.

Comparing 3DPredictor with other models.

Chromatin interactions for GM12878 cells predicted by MEGABASE+MiChroM were downloaded from Juicebox server (https://s3.amazonaws.com/hicfiles/external/ctbp_8_4_17/all_intra_megabase_michrom.hic). As these interactions were at 50 kb resolution, we predicted the same regions at 5 kb resolution and averaged data to obtain contacts at 50 kb.

Interaction frequencies for 53-75 Mb region on Chromosome 4 of GM12878 cells, predicted by (Rowley et al. 2017), as well as CTCF loop anchors were provided by V.G. Corces, M.J Rowley and M.H. Nichols (personal communication).

Interaction frequencies for 20-45 Mb region on Chromosome 1 of GM12878 cells, predicted by (Qi and Zhang 2019), were provided by Y. Qi and B. Zhang (personal communication).

Note that we used here SCC smoothing parameter equal to two for all comparisons, whereas (Qi and Zhang 2019) used SCC smoothing parameter values above five. Note that changing smoothing parameter does not affect results of 3DPredictor benchmark (see Supplementary Fig. E in Qi and Zhang and Supplementary Note, Supplementary Table 4 and Supplementary Table 7 in this paper for details of SCC smoothing parameter effect).

Defining cell type-specific TADs and cell type-specific EP interactions.

To define cell type-specific TAD boundaries, we utilized insulation-based score, which reflects depletion of contacts spanning putative TAD boundary, similarly to the approach used in (Fishman et al. 2019; Vietri Rudan et al. 2015; Sexton et al. 2012), but with some modifications. The schematic representation of current approach is shown in Supplementary Fig. 8. In particular, for each bin i of the Hi-C matrix A, we define four vectors a_L, b_L, a_R, b_R each containing N elements:

$$a_L = \left(A_{i-k,i-1}\right);\ b_L = \left(A_{i-k+2,i+1}\right);\ a_R = \left(A_{i-1;i+k+2}\right);\ b_R = \left(A_{i+1;i+k+4}\right); k = 1..N,$$

Where $A_{i,j}$ is number of contacts between bins *i* and *j*, and *N* is empirical constant which was equal to 5 in this study. Thus, for two bins a and b, surrounding the bin i, vectors a_L, b_L, a_R, b_R describe local (+/- *N* bins) contacts.

Then, the insulation score S_i of bin i was computed by dividing the frequency of contacts crossing bin i to the frequency of distance-matched contacts located downstream or upstream of i (Supplementary Fig. 8) and summing obtained ratios:

$$S_i = \sum_{j=1\ldots N} \frac{a_L^j}{b_L^j} + \sum_{j=1\ldots N} \frac{a_L^j}{b_L^j}$$

If for a bin *i* we observed a high (above empirically defined upper threshold) insulation-based score in one cell type and low (under empirically defined lower threshold) score in another cell type, then we considered the bin *i* as a center of cell type-specific region. We defined upper and lower thresholds based on distribution of insulation scores (Supplementary Fig. 18) so that upper threshold corresponded to the strong insulation and lower threshold was close to the natural noise of insulation in

Hi-C data. In particular, we used the following parameters to compare NPC and hepatocytes: bin size equal to 25 kb; *N* equal to five, which means that we used the interval +/-100 kb around putative boundary to calculate insulation score; upper threshold equal to 3×*N*=15; lower threshold equal to 2.4×*N*=12. With these parameters, we obtained 88 cell type-specific regions.

To estimate prediction accuracy for cell type-specific EP interactions, we compared differences between predicted and control data with differences between cell types and replicates. We characterized contacts of EP pairs by observed-to-expected contacts ratio (OE). The EP interactions were referred to as cell type-specific, if it's OE differ between replicates less than two times and differ between cell types more than two times:

$$\left|\log_2 \frac{OE_{rep1}}{OE_{rep2}}\right| < 1 \; and \; \left|\log_2 \frac{OE_{celltype1}}{OE_{celltype2}}\right| > 1.$$

To measure similarity of cell type-specific interactions in two samples (i.e. predicted and experimental data or two experimental samples), we calculated mean difference of OE values for corresponding interactions:

$$Similarity = mean\left(\left|\log_2 \frac{OE_1}{OE_2}\right|\right)$$

The cell type-specific interactions were obtained comparing mouse hepatocytes (combined data and replicates) with NPC (only combined data) and human K562 cell (combined data and by replicates) with monocytes (only combined data) on 5 kb and 25 kb resolution.

Analysis of looping contacts.

For quantitative comparison of interactions in loop anchors shown on Supplementary Fig. 17, we used loops derived from the experimental NPC Hi-C data using HiCCUPS. We next aimed to call loop anchors based on 3DPredictor data using HiCCUPS. However, HiCCUPS required normalization vectors to be provided in .hic-files, and since these vectors were not available for predicted data, we failed to annotate the loops automatically. Thus, we manually annotated all loops for both experimental and predicted Hi-C maps of Chromosome 5 of NPC cells (Supplementary Table 8,9) shows coordinates of the manually annotated loop anchors), and compared obtained data to estimate the number of correctly predicted loops.

Modeling chromosomal rearrangements.

To model chromosomal rearrangements, we used 3DPredictor trained on mouse hepatocyte cells. To generate predictors, we obtained CTCF (NCBI SRX1975285-SRX1975286) and RNA-seq (NCBI SRX1975216-SRX1975217) data from wild-type mouse hindlimb E11.5 cells. Next, we deleted all CTCF peaks and genes from the region [mm10]: Chr1:76392403-78064264, which corresponds to the deletion coordinates described in (Bianco et al. 2018). Resulting set of predictors was used to model all chromatin contacts within the region [mm10]: Chr1:70950000-81000000. To compare contact frequencies predicted by the model with experimental data, we defined ectopic interactions as described in (Bianco et al. 2018). We first generated normalized difference matrix between mutated and WT matrices. For this, we multiply the mutant matrix by a coefficient that equalizes the coverage of regions that are not involved in the mutation. Next, we subtracted the WT matrix from the mutated matrix. We normalized the difference matrix by dividing each sub-diagonal by the average number of reads observed at the corresponding genomic distance in WT data. After we get the normalized difference matrix, we find ectopic interactions for each sub-diagonal. Specifically, we filtered out the sub-diagonal elements, which were above 96th percentile of all sub-diagonal values, and calculate standard deviation of remaining values. All points, which differ from zero by more than three standard deviations, were considered ectopic.

Software availability

3DPredictor source code (https://github.com/labdevgen/3DPredictor) and Jupiter Notebook with the code used to reproduce TargetFinder results (https://github.com/labdevgen/targetFinderTests) are both freely available on GitHub and in Supplemental Code.

Authors contribution. V.F. proposed the study. P.B., D.F. and V.F. benchmarked *TargetFinder* algorithm. V.S. and P.B. wrote main parts of 3DPredictor code. E.M. implemented metrics of algorithm performance and cell type-specific TADs comparisons. V.S., P.B., E.M. and M.N. developed different predictor parameterizations and compared training parameters. M.N. prepared training and validation data and identified and analyzed cell type-specific EP interactions. P.B. performed *Epha4* locus modeling. All authors contributed to manuscript preparation.

Acknowledgments. This work was supported by Russian Foundation For Basic Research (RFBR) grant #18-29-13021 and Russian Science Foundation grant #17-74-10143. All computations were performed with support from the Computational Cluster of the Novosibirsk State University and Computational Nodes of the Institute of Cytology and Genetics (Budget Project #0324-2019-0041). We acknowledge scientific discussions with Dr. Nariman Battulin and Emil Valeev. We are thankful to V.G. Corces, M.J Rowley and M.H. Nichols, and Y. Qi and B. Zhang, who provided their predictions of GM12878 contacts. We also thankful to Michele Di Pierro, who pointed us to the MEGABASE+MiChroM predicted contacts on the Juicebox web-server.

Competing interests. The authors declare no competing interests.

References

Al Bkhetan Z, Plewczynski D. 2017. Multi-levels 3D Chromatin Interactions Prediction Using Epigenomic Profiles. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10352, pp. 19–28. doi:10.1007/978-3-319-60438-1_2.

Al Bkhetan Z, Plewczynski D. 2018. Three-dimensional Epigenome Statistical Model: Genome-wide Chromatin Looping Prediction. *Sci Rep* **8**: 5217. doi:10.1038/s41598-018-23276-8.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455–461. doi:10.1038/nature12787.

Atlasi Y, Stunnenberg HG. 2017. The interplay of epigenetic marks during stem cell differentiation and development. *Nat Rev Genet* **18**: 643–658. doi:10.1038/nrg.2017.57.

Battulin N, Fishman VS, Mazur AM, Pomaznoy M, Khabarova AA, Afonnikov DA, Prokhortchouk EB, Serov OL. 2015. Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach. *Genome Biol* **16**: 77. doi:10.1186/s13059-015-0642-0.

Bianco S, Lupiáñez DG, Chiariello AM, Annunziatella C, Kraft K, Schöpflin R, Wittler L, Andrey G, Vingron M, Pombo A, et al. 2018. Polymer physics predicts the effects of structural variants on chromatin architecture. *Nat Genet* **50**: 662–667. doi:10.1038/s41588-018-0098-8.

Bogu GK, Vizán P, Stanton LW, Beato M, Di Croce L, Marti-Renom MA. 2016. Chromatin and RNA Maps Reveal Regulatory Long Noncoding RNAs in Mouse. *Mol Cell Biol* **36**: 809–819. doi:10.1128/mcb.00955-15.

Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot JP, Tanay A, et al. 2017. Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**: 557–572.e24. doi:10.1016/j.cell.2017.09.043.

Buckle A, Brackley CA, Boyle S, Marenduzzo D, Gilbert N. 2018. Polymer Simulations of Heteromorphic Chromatin Predict the 3D Folding of Complex Genomic Loci. *Mol Cell* **72**: 786–797.e11.

doi:10.1016/j.molcel.2018.09.016.

Cao Q, Anyansi C, Hu X, Xu L, Xiong L, Tang W, Mok MTS, Cheng C, Fan X, Gerstein M, et al. 2017. Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nat Genet* **49**: 1428–1436. doi:10.1038/ng.3950.

Chen Y, Wang Y, Xuan Z, Chen M, Zhang MQ. 2016. De novo deciphering three-dimensional chromatin interaction and topological domains by wavelet transformation of epigenetic profiles. *Nucleic Acids Res* **44**: e106–e106. doi:10.1093/nar/gkw225.

Chiariello AM, Annunziatella C, Bianco S, Esposito A, Nicodemi M. 2016. Polymer physics of chromosome large-scale 3D organisation. *Sci Rep* **6**: 1–8. doi:10.1038/srep29775.

Di Pierro M, Cheng RR, Lieberman Aiden E, Wolynes PG, Onuchic JN. 2017. De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proc Natl Acad Sci* **114**: 12126–12131. doi:10.1073/pnas.1714980114.

Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. 2016a. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**: 99–101. doi:10.1016/j.cels.2015.07.012.

Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. 2016b. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**: 95–98. doi:10.1016/j.cels.2016.07.002.

Fishman V, Battulin N, Nuriddinov M, Maslova A, Zlotina A, Strunov A, Chervyakova D, Korablev A, Serov O, Krasikova A. 2019. 3D organization of chicken genome demonstrates evolutionary conservation of topologically associated domains and highlights unique architecture of erythrocytes' chromatin. *Nucleic Acids Res* **47**: 648–665. doi:10.1093/nar/gky1103.

Fishman VS, Salnikov PA, Battulin NR. 2018. Interpreting Chromosomal Rearrangements in the Context of 3-Dimentional Genome Organization: A Practical Guide for Medical Genetics. *Biochemistry (Mosc)* **83**: 393–401. doi:10.1134/S0006297918040107.

Fortin J-P, Hansen KD. 2015. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol* **16**: 180. doi:10.1186/s13059-015-0741-y.

Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, Kraft K, Kempfer R, Jerković I, Chan W-L, et al. 2016. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**: 265–269. doi:10.1038/nature19800.

Fudenberg G, Abdennur N, Imakaev M, Goloborodko A, Mirny LA. 2017. Emerging Evidence of Chromosome Folding by Loop Extrusion. *Cold Spring Harb Symp Quant Biol* **82**: 45–55. doi:10.1101/sqb.2017.82.034710.

Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, Kane M, Cleary B, Lander ES, Engreitz JM. 2016. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science (80)* **354**: 769–773. doi:10.1126/science.aag2445.

Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, Grossman SR, Anyoha R, Patwardhan TA, Nguyen TH, et al. 2019. Activity-by-Contact model of enhancer specificity from thousands of CRISPR perturbations. *bioRxiv* 529990. doi:10.1101/529990.

Furlong EEM, Levine M. 2018. Developmental enhancers and chromosome topology. *Science (80)* **361**: 1341–1345. doi:10.1126/science.aau0320.

Gridina MM, Matveeva NM, Fishman VS, Menzorov AG, Kizilova HA, Beregovoy NA, Kovrigin II, Pristyazhnyuk IE, Oscorbin IP, Filipenko ML, et al. 2018. Allele-Specific Biased Expression of the CNTN6 Gene in iPS Cell-Derived Neurons from a Patient with Intellectual Disability and 3p26.3 Microduplication Involving the CNTN6 Gene. *Mol Neurobiol* **55**: 6533–6546. doi:10.1007/s12035-017-0851-5.

Gudmundsson J, Thorleifsson G, Sigurdsson JK, Stefansdottir L, Jonasson JG, Gudjonsson SA, Gudbjartsson DF, Masson G, Johannsdottir H, Halldorsson GH, et al. 2017. A genome-wide association study yields five novel thyroid cancer risk loci. *Nat Commun* **8**: 14517. doi:10.1038/ncomms14517.

Hait TA, Amar D, Shamir R, Elkon R. 2018. FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer–promoter map. *Genome Biol* **19**: 56. doi:10.1186/s13059-018-1432-2.

Huang J, Marco E, Pinello L, Yuan G. 2015. Predicting chromatin organization using histone marks. *Genome Biol* **16**: 162. doi:10.1186/s13059-015-0740-z.

Ibn-Salem J, Andrade-Navarro MA. 2018. Computational Chromosome Conformation Capture by Correlation of ChIP-seq at CTCF motifs. *bioRxiv* 257584. doi:10.1101/257584.

Jenkinson G, Pujadas E, Goutsias J, Feinberg AP. 2017. Potential energy landscapes identify the information-theoretic nature of the epigenome. *Nat Genet* **49**: 719–729. doi:10.1038/ng.3811.

Jiang Y, Loh YHE, Rajarajan P, Hirayama T, Liao W, Kassim BS, Javidfar B, Hartley BJ, Kleofas L, Park RB, et al. 2017. The methyltransferase SETDB1 regulates a large neuron-specific topological chromatin domain. *Nat Genet* **49**: 1239–1250. doi:10.1038/ng.3906.

Kai Y, Andricovich J, Zeng Z, Zhu J, Tzatsos A, Peng W. 2018. Predicting CTCF-mediated chromatin interactions by integrating genomic and epigenomic features. *Nat Commun* **9**: 4221. doi:10.1038/s41467-018-06664-6.

Li W, Wong WH, Jiang R. 2019. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Res* **47**: e60. doi:10.1093/nar/gkz167.

Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, et al. 2015. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* **161**: 1–14. doi:10.1016/j.cell.2015.04.004.

Mishra A, Hawkins RD. 2017. Three-dimensional genome architecture and emerging technologies: Looping in disease. *Genome Med* **9**: 1–14. doi:10.1186/s13073-017-0477-2.

Moore BL, Aitken S, Semple CA. 2015. Integrative modeling reveals the principles of multi-scale chromatin boundary formation in human nuclear organization. *Genome Biol* **16**: 110. doi:10.1186/s13059-015-0661-x.

Moorthy SD, Davidson S, Shchuka VM, Singh G, Malek-Gilani N, Langroudi L, Martchenko A, So V, Macpherson NN, Mitchell JA. 2017. Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Res* **27**: 246–258. doi:10.1101/gr.210930.116.

O'Connor T, Bodén M, Bailey TL. 2017. CisMapper: predicting regulatory interactions from transcription factor ChIP-seq data. *Nucleic Acids Res* **45**: e19. doi:10.1093/nar/gkw956.

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* **11**: 1650–1667. doi:10.1038/nprot.2016.095.

Phanstiel DH, Van Bortle K, Spacek D, Hess GT, Shamim MS, Machol I, Love MI, Aiden EL, Bassik MC, Snyder MP. 2017. Static and Dynamic DNA Loops form AP-1-Bound Activation Hubs during Macrophage Development. *Mol Cell* **67**: 1037–1048.e6. doi:10.1016/j.molcel.2017.08.006.

Qi Y, Zhang B. 2019. Predicting three-dimensional genome organization with chromatin states. *PLOS Comput Biol* **15**: e1007024. doi:10.1371/journal.pcbi.1007024.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014. A 3D Map of the Human Genome at Kilobase Resolution Reveals

Principles of Chromatin Looping. *Cell* **159**: 1665–1680. doi:10.1016/j.cell.2014.11.021.

Redin C, Brand H, Collins RL, Kammin T, Mitchell E, Hodge JC, Hanscom C, Pillalamarri V, Seabra CM, Abbott MA, et al. 2017. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nat Genet* **49**: 36–45. doi:10.1038/ng.3720.

Rowley MJ, Nichols MH, Lyu X, Ando-Kuri M, Rivera ISM, Hermetz K, Wang P, Ruan Y, Corces VG. 2017. Evolutionarily Conserved Principles Predict 3D Chromatin Organization. *Mol Cell* **67**: 837–852.e7. doi:10.1016/j.molcel.2017.07.022.

Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, Heard E, Dekker J, Barillot E. 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16**: 259. doi:10.1186/s13059-015-0831-x.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**: 458–472. doi:10.1016/j.cell.2012.01.010.

Spielmann M, Lupianez DG, Mundlos S. 2018. Structural variation in the 3D genome. *Nat Rev Genet* **19**: 453–467. doi:10.1038/s41576-018-0007-0.

Stricker SH, Köferle A, Beck S. 2016. From profiles to function in epigenomics. *Nat Rev Genet* **18**: 51–66. doi:10.1038/nrg.2016.138.

Sun F, Chronis C, Kronenberg M, Chen XF, Su T, Lay FD, Plath K, Kurdistani SK, Carey MF. 2019. Promoter-Enhancer Communication Occurs Primarily within Insulated Neighborhoods. *Mol Cell* **73**: 250–263.e5. doi:10.1016/j.molcel.2018.10.039.

Symmons O, Pan L, Remeseiro S, Aktas T, Klein F, Huber W, Spitz F. 2016. The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances. *Dev Cell* **39**: 529–543. doi:10.1016/j.devcel.2016.10.015.

Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J, Ruszczycki B, et al. 2015. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163**: 1611–1627. doi:10.1016/j.cell.2015.11.024.

van Heeringen SJ, Veenstra GJC. 2011. GimmeMotifs: A de novo motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics* **27**: 270–271. doi:10.1093/bioinformatics/btq636.

Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S. 2015. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Rep* **10**: 1297–1309. doi:10.1016/j.celrep.2015.02.004.

Whalen S, Truty RM, Pollard KS. 2016. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* **48**: 488–496. doi:10.1038/ng.3539.

Wu Y, Zeng J, Zhang F, Zhu Z, Qi T, Zheng Z, Lloyd-Jones LR, Marioni RE, Martin NG, Montgomery GW, et al. 2018. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat Commun* **9**: 918. doi:10.1038/s41467-018-03371-0.

Xi W, Beer MA. 2018. Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy ed. W.S. Noble. *PLOS Comput Biol* **14**: e1006625. doi:10.1371/journal.pcbi.1006625.

Xu T, Zheng X, Li B, Jin P, Qin Z, Wu H. 2018. A comprehensive review of computational prediction of genome-wide features. *Brief Bioinform*. doi:10.1093/bib/bby110.

Yang T, Zhang F, Yardımci GG, Song F, Hardison RC, Noble WS, Yue F, Li Q. 2017a. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res* **27**: 1939–1949. doi:10.1101/gr.220640.117.

Yang Y, Zhang R, Singh S, Ma J. 2017b. Exploiting sequence-based features for predicting enhancer–

promoter interactions. *Bioinformatics* **33**: i252–i260. doi:10.1093/bioinformatics/btx257.

Zeng W, Wu M, Jiang R. 2018. Prediction of enhancer-promoter interactions via natural language processing. *BMC Genomics* **19**: 84. doi:10.1186/s12864-018-4459-6.

Zepeda-Mendoza CJ, Ibn-Salem J, Kammin T, Harris DJ, Rita D, Gripp KW, MacKenzie JJ, Gropman A, Graham B, Shaheen R, et al. 2017. Computational Prediction of Position Effects of Apparently Balanced Human Chromosomal Rearrangements. *Am J Hum Genet* **101**: 206–217. doi:10.1016/j.ajhg.2017.06.011.

Zhang H, Li F, Jia Y, Xu B, Zhang Y, Li X, Zhang Z. 2017. Characteristic arrangement of nucleosomes is predictive of chromatin interactions at kilobase resolution. *Nucleic Acids Res* **45**: 12739–12751. doi:10.1093/nar/gkx885.

Zhang R, Wang Y, Yang Y, Zhang Y, Ma J. 2018a. Predicting CTCF-mediated chromatin loops using CTCF-MP. *Bioinformatics* **34**: i133–i141. doi:10.1093/bioinformatics/bty248.

Zhang S, Chasman D, Knaack S, Roy S. 2018b. In silico prediction of high-resolution Hi-C interaction matrices. *bioRxiv* 406322. doi:10.1101/406322.

Zhu G, Deng W, Hu H, Ma R, Zhang S, Yang J, Peng J, Kaplan T, Zeng J. 2018. Reconstructing spatial organizations of chromosomes through manifold learning. *Nucleic Acids Res* **46**: e50–e50. doi:10.1093/nar/gky065.

Zhu Y, Chen Z, Zhang K, Wang M, Medovoy D, Whitaker JW, Ding B, Li N, Zheng L, Wang W. 2016. Constructing 3D interaction maps from 1D epigenomes. *Nat Commun* **7**: 10812. doi:10.1038/ncomms10812.

Cell type	Predictors	Loops	Interacting EP pairs	Train / validation split	F1-score	
					Interacting : Non-interacting 1:20*	Interacting : Non-interacting 1:1
Mouse ES cells	24	9091	1602	this paper	**0.015**	0.56
				original	**0.82**	0.91
Mouse cortex	10	9972	625	this paper	**0.19**	0.69
				original	**0.42**	0.79
Mouse NPC	10	9360	635	this paper	**0.16**	0.71
				original	**0.46**	0.77
Human GM12878	100	9448	2113	this paper	**0.039**	0.61
				original	**<u>0.77</u>**	0.89

Table 1. Effect of train/validation split strategy on *TargetFinder* efficiency.

Figure legends

Figure 1. Promoter-enhancer pairs with overlapping windows in training and validation datasets. A. Schematic illustration showing how information could be shared between training and validation datasets because of overlapping EP windows. **B**. Distribution of distances between boundaries of overlapping EP windows. For each EP pair we found the window of another EP pair so, that the distance between boundaries of their windows (*d=D1+D2*) is minimal. Histogram shows distribution of the obtained values of *d*.

Figure 2. Hi-C loops do not provide complete information about interacting EP interactions. A,B Distribution of row (A) and distance-normalized (B) contact frequencies for interacting EP pairs and loop anchors in human monocytes. **C.** Number of interacting EP pairs overlapping loops in monocyte' data. Red line - number of EP pairs overlapping any Hi-C loop anchors or located within distance not more than *X* kb of them, shown as a function of *X*. Gray line and grey area represent average and 3 standard deviations of 100 randomized controls. **D.** Chromatin interactions within the region on human Chromosome 8 containing five experimentally validated *MYC* enhancers (yellow lines, *e1-e7*) and Hi-C loops (blue squares). Although both enhancers and loops were identified in the same cell type (K562 cells) they show little overlap.

Figure 3. 3DPredictor efficiently reconstructs spatial interactions based on CTCF-occupancy, expression and genomic distance. A. Representative region of mouse Chromosome 2 showing predicted and experimentally-derived Hi-C interactions in mouse hepatocytes. **B-F.** Various metrics of 3DPredictor accuracy for each chromosome of mouse hepatocytes. Circles represent comparison between two replicas; red squares show comparisons between hepatocytes and other cell types. Red triangles display 3DPredictor results obtained using single Chromosome 5 for training; data obtained when validating on the same chromosome marked with asterisks. Blue triangles show results of 3DPredictor trained on 10 chromosomes (results for even chromosomes obtained using model trained on odd chromosomes and *vice versa*)

Figure 4. 3DPredictor accurately reconstructs cell type-specific chromatin organization. A. Representative region on Chromosome 3 showing different 3D-organization in mouse hepatocytes and NPC. Cell type-specific TAD boundary is shown by arrow. **C and D**. Comparison of 3DPredictor results with experimental hepatocyte (C) or NPC (D) data for the same region of Chromosome 3 **D.** Insulation scores in 88 NPC cell type-specific regions correlate with insulation scores calculated based on predicted contacts significantly better, than with insulation scores based on experimental hepatocyte' data (p-value 4×10^{-6}).

Figure 5. Accurate prediction of promoter-enhancer interaction frequencies. A. Prediction accuracy of contact frequencies of EP pairs defined as "interacting" in monocytes according to SlideBase and GeneHancer databases ("*EP*"), and all other pairs of loci ("*all except EP*"). **B.** Scatter-plot displaying predicted (Y-axes) and experimentally-measured (X-axes) contact frequencies for interacting EP pairs. **C.** Distribution of the similarity scores for cell type-specific EP interactions in different cell types (K562 vs Monocytes) or experimental and predicted data (K562 experimental vs K562 predicted and monocytes experimental vs monocytes predicted). See methods for definition of cell type-specific EP interactions and similarity scores. Data on A-C provided for 25 kb resolution.

Figure 6. 3DPredictor accurately reconstructs genome organization of novel cell type. A. Example of mouse NPC Hi-C contact map derived from experimental data (above diagonal) or obtained using 3DPredictor trained on hepatocyte' contacts and provided with epigenetic data relevant for NPC. **B-F.** MRE (B), SCC (C), MAE (D), Person correlation (E) or MSE (F) measurements of 3DPredictor accuracy for training and validation on same (green and blue lines) or different (red line) cell types. **G.** Legend for plots B-F.

Figure 7. 3DPredictor captures 3-dimensional organization of rearranged genomic regions. A. Contact map of mouse *DelB* locus, carrying homozygous deletion of ~1.5 Mb, with experimentally-measured contacts in the top, and 3DPredictor modeling results in the bottom. White lines correspond to contacts of the deleted locus. Note ectopic interactions between *Pax3* and *Epha4* TADs (indicated by arrows). These ectopic interactions are even better visible on **B**, where the same region is plotted and only those interactions, which differ between WT and *DelB* by more than three standard deviations, are kept. On **A**, the color indicates contact counts, whereas on **B** the color indicates significance of differences between WT and *DelB* data. **C** shows sizes of observed (red vertical bar) and expected (blue bars) overlaps between experimental and predicted ectopic interactions.

A

Hi-C loop = interaction

E1 E2 Window P1 Window E3 E4

Random split of EP pairs

Train

P1 E3 Yes

P1 E2 No

Validation

? P1 E1

? P1 E4

Match train/validation based on shared window

P1 E3 Yes

P1 E2 No

? P1 E1

? P1 E4

Use infromation from training set to identify interacting EP pairs

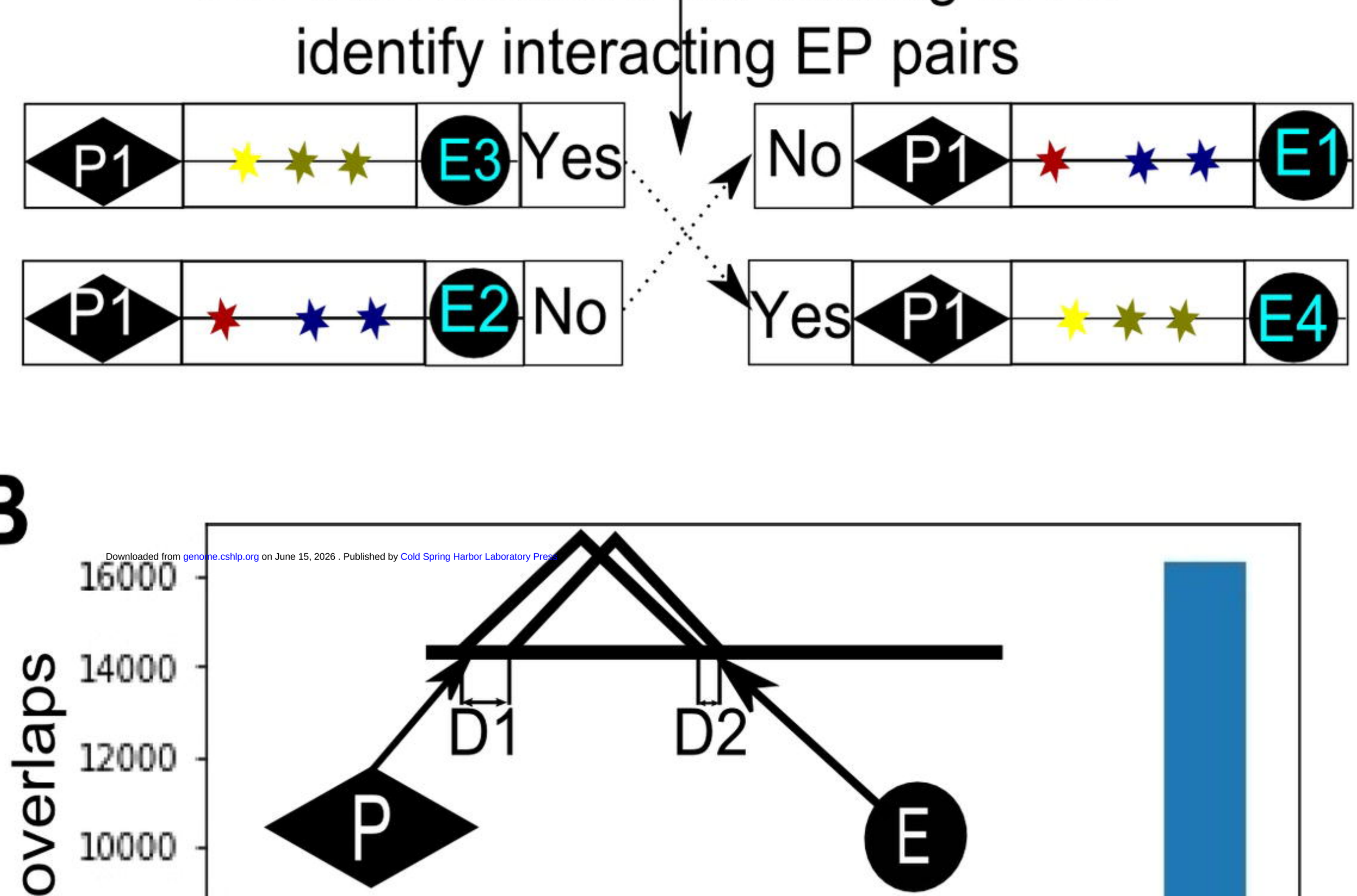
P1 E3 Yes

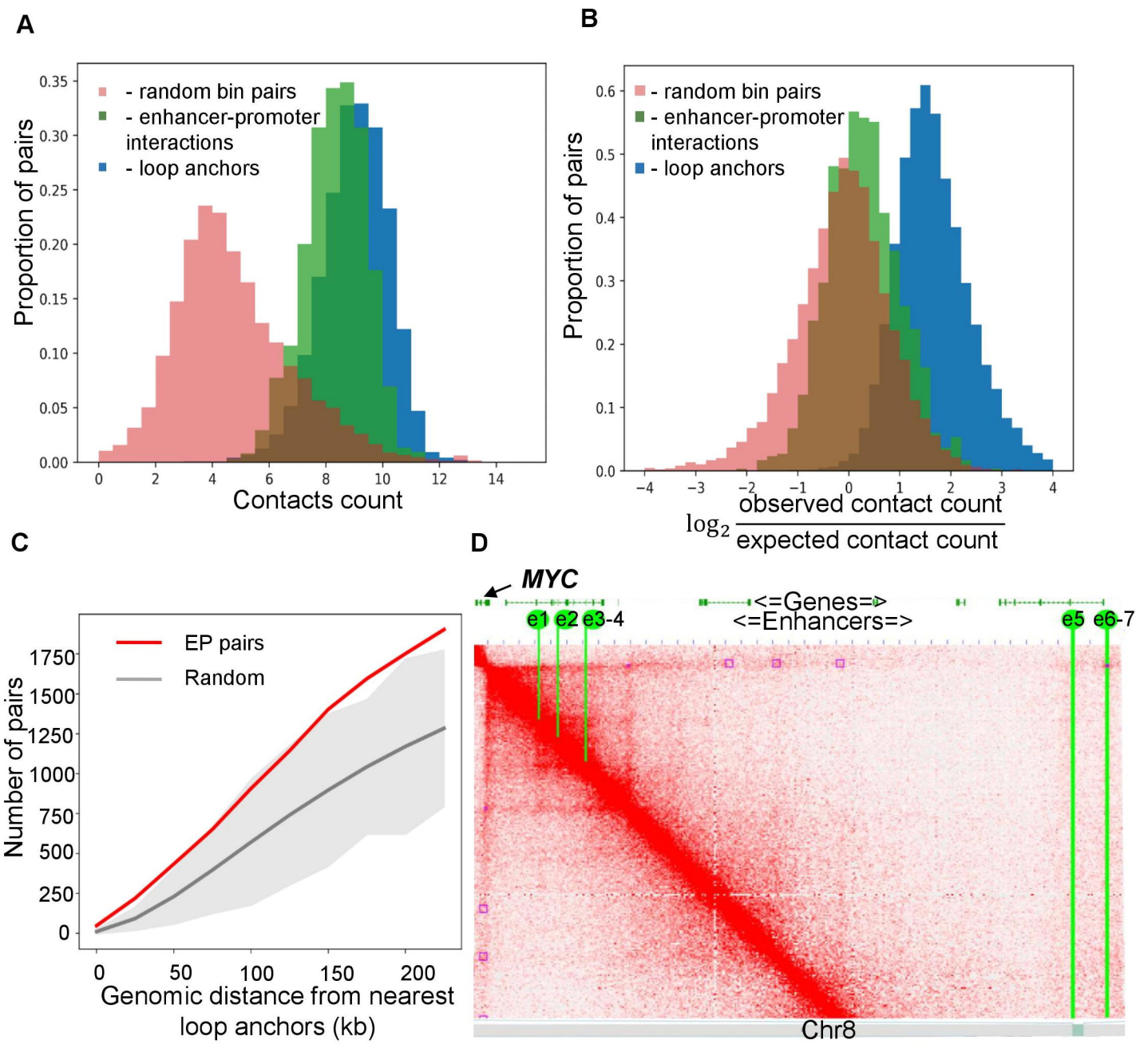
P1 E2 No

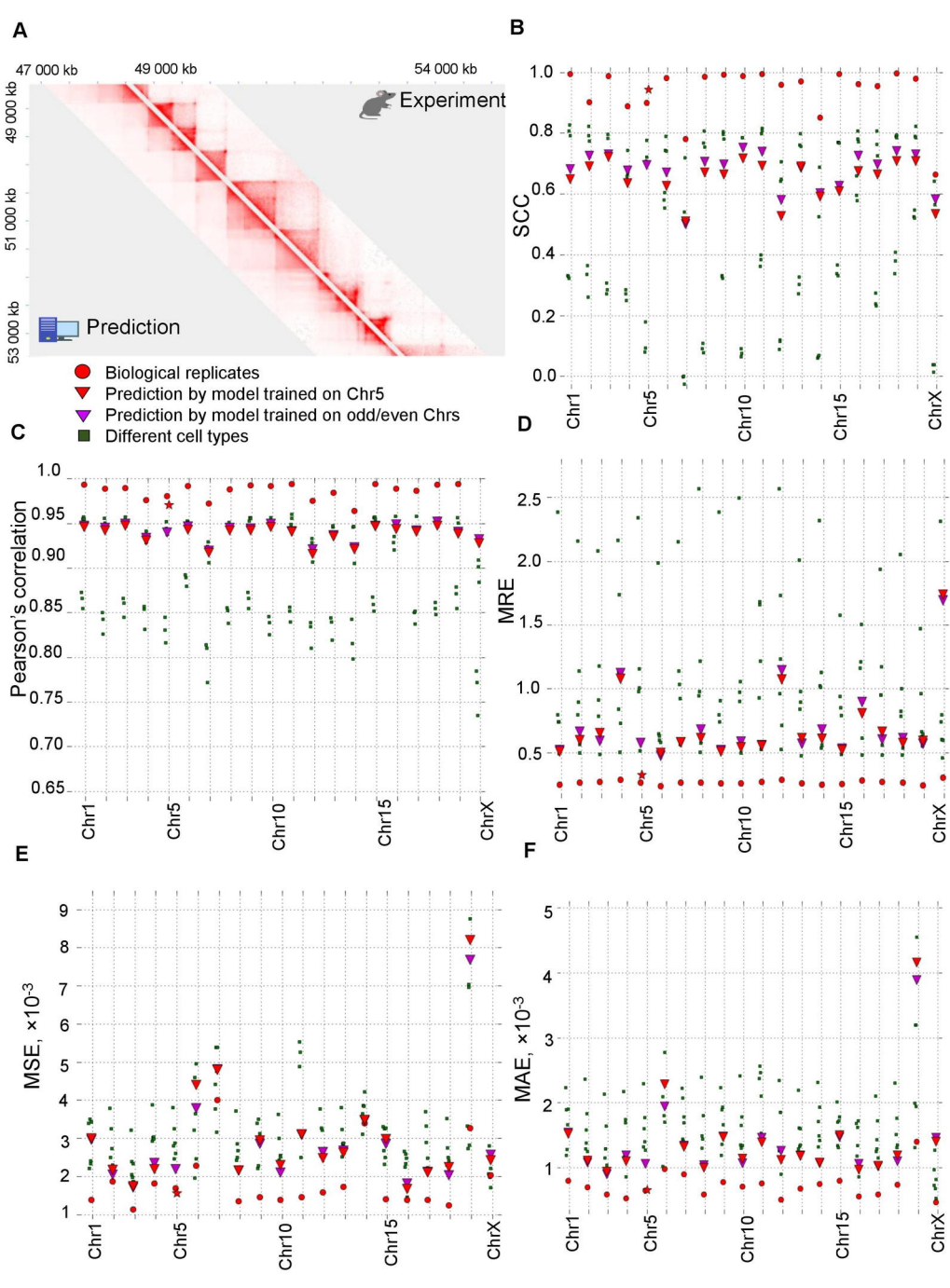
No P1 E1

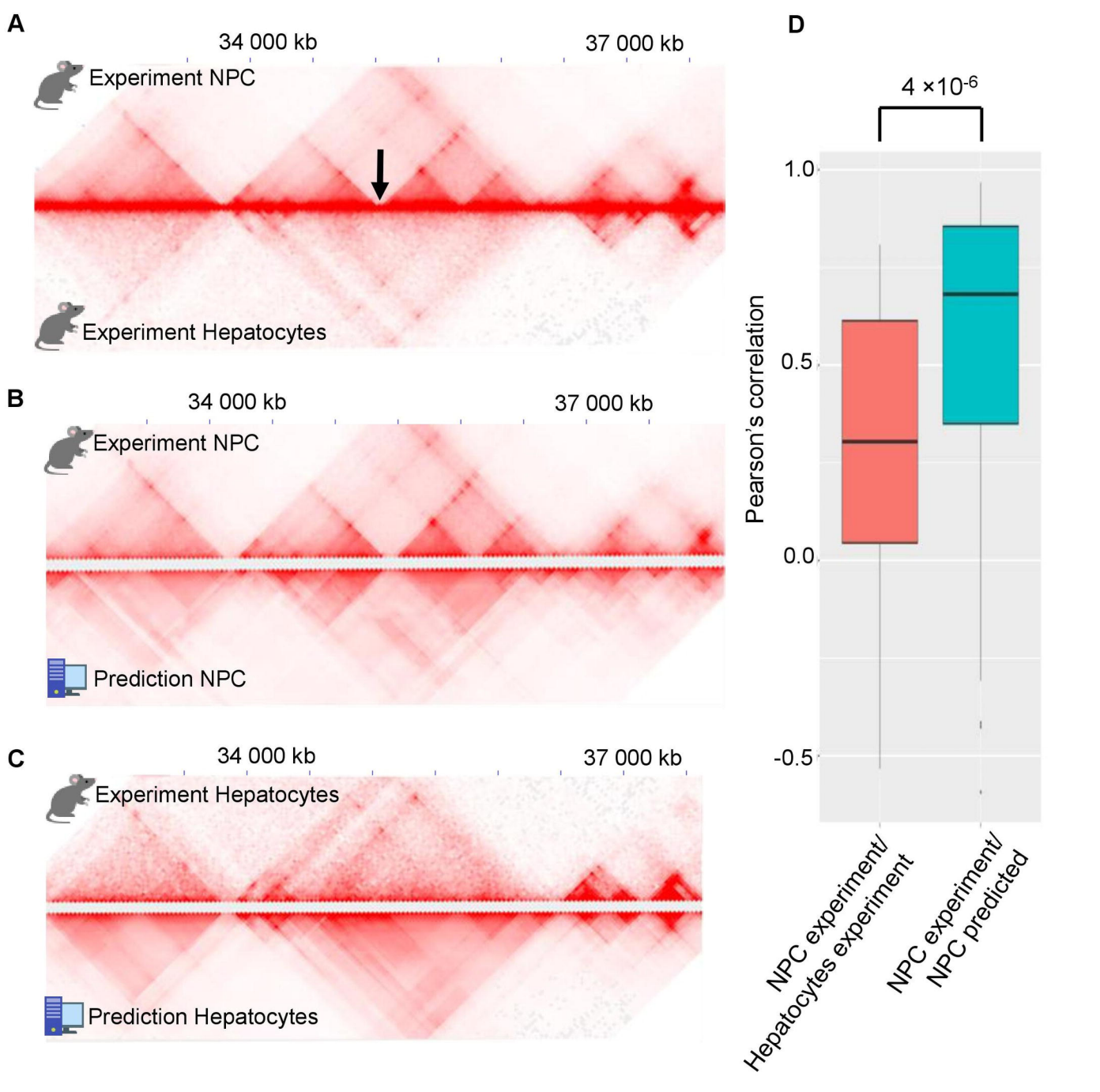
Yes P1 E4

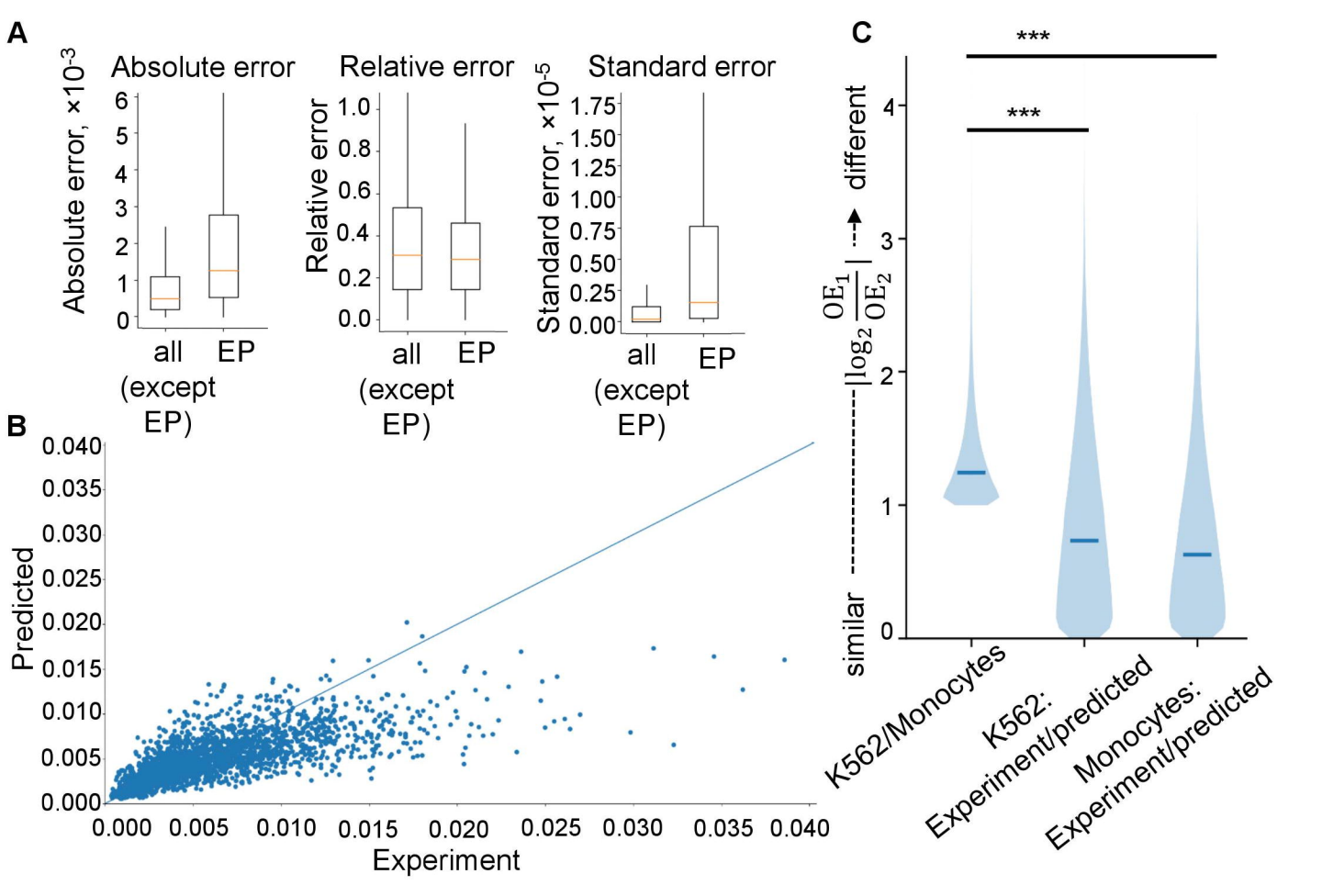
B

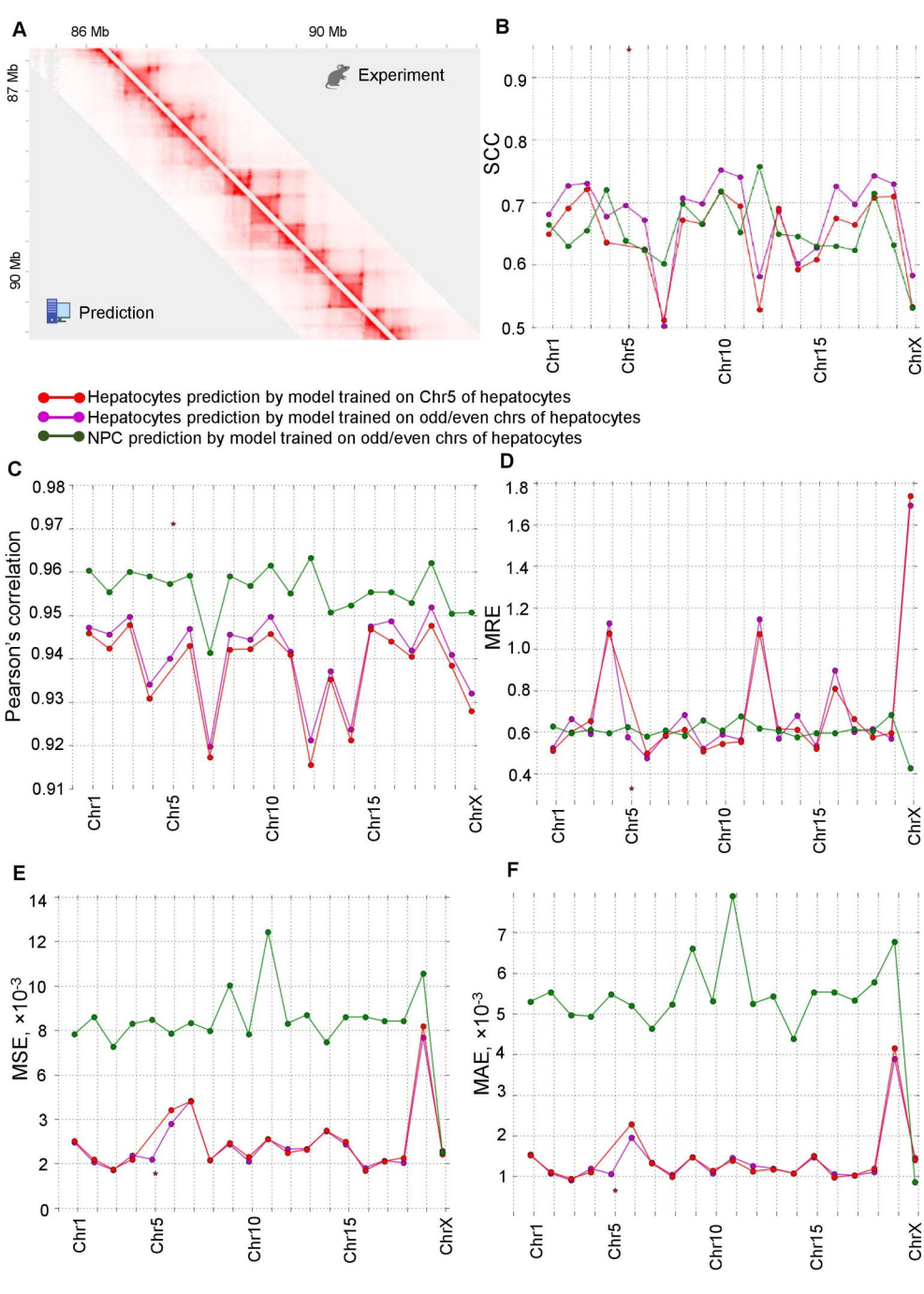


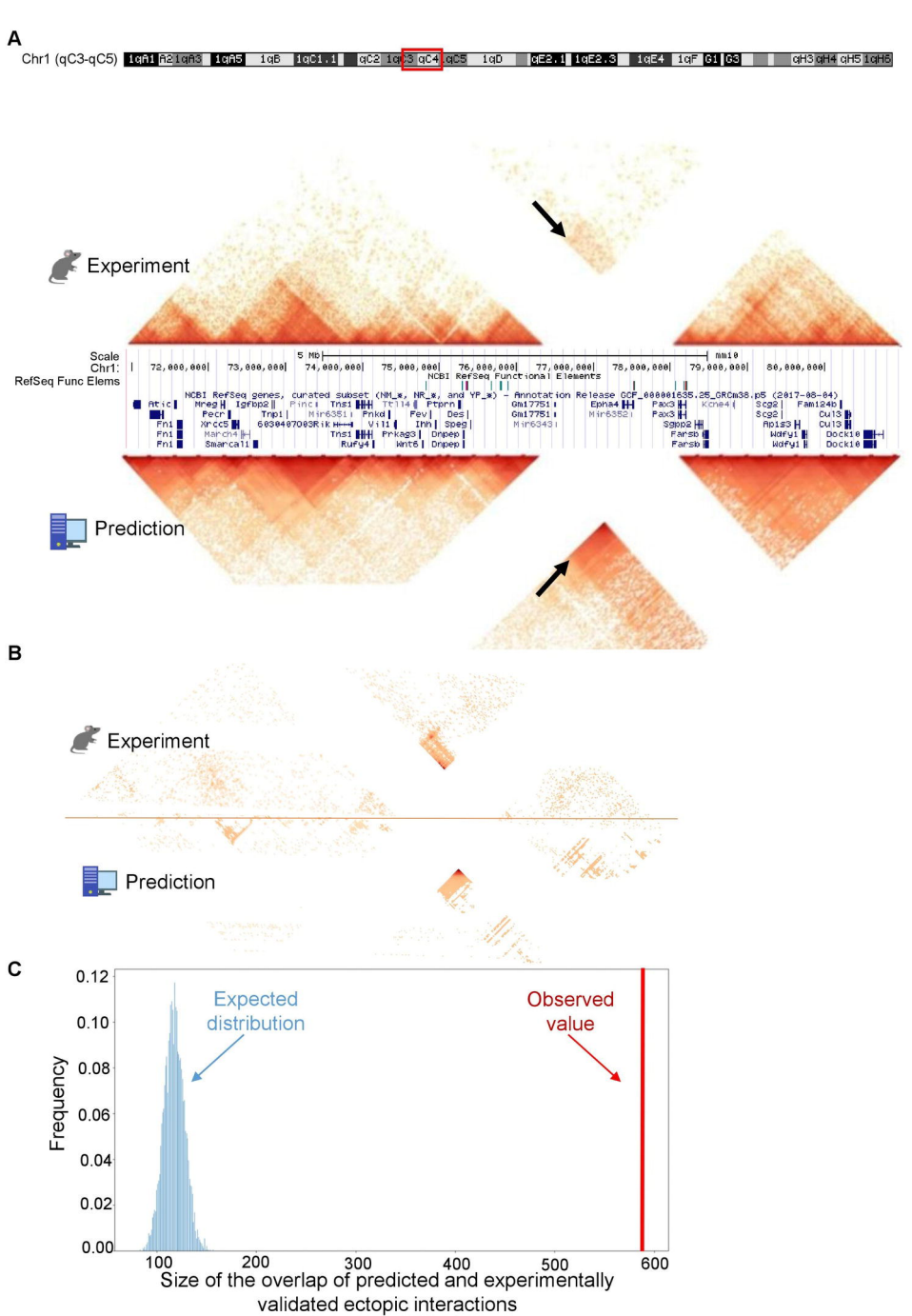

A
- random bin pairs
- enhancer-promoter interactions
- loop anchors
Proportion of pairs
Contacts count
B
- random bin pairs
- enhancer-promoter interactions
- loop anchors
Proportion of pairs
log2 observed contact count / expected contact count
C
EP pairs
Random
Number of pairs
Genomic distance from nearest loop anchors (kb)
D
MYC
<=Genes=>
<=Enhancers=>
e1 e2 e3-4
e5 e6-7
Chr8


A
47 000 kb
49 000 kb
54 000 kb
49 000 kb
51 000 kb
53 000 kb
Experiment
Prediction
Biological replicates
Prediction by model trained on Chr5
Prediction by model trained on odd/even Chrs
Different cell types
B
SCC
C
Pearson's correlation
D
MRE
E
MSE, ×10-3
F
MAE, ×10-3
Chr1
Chr5
Chr10
Chr15
ChrX


A
34 000 kb
37 000 kb
Experiment NPC
Experiment Hepatocytes
B
34 000 kb
37 000 kb
Experiment NPC
Prediction NPC
C
34 000 kb
37 000 kb
Experiment Hepatocytes
Prediction Hepatocytes
D
4 ×10-6
1.0
0.5
0.0
-0.5
Pearson's correlation
NPC experiment/ Hepatocytes experiment
NPC experiment/ NPC predicted


A
Absolute error
Absolute error, ×10-3
6
5
4
3
2
1
0
all
EP
(except EP)
Relative error
Relative error
1.0
0.8
0.6
0.4
0.2
0.0
all
EP
(except EP)
Standard error
Standard error, ×10-5
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
all
EP
(except EP)
B
Predicted
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
Experiment
C
***
***
similar
|log2 OE1/OE2|
different
4
3
2
1
0
K562/Monocytes
K562: Experiment/predicted
Monocytes: Experiment/predicted


A
86 Mb
90 Mb
87 Mb
90 Mb
Experiment
Prediction
B
SCC
0.9
0.8
0.7
0.6
0.5
Chr1
Chr5
Chr10
Chr15
ChrX
Hepatocytes prediction by model trained on Chr5 of hepatocytes
Hepatocytes prediction by model trained on odd/even chrs of hepatocytes
NPC prediction by model trained on odd/even chrs of hepatocytes
C
Pearson's correlation
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
Chr1
Chr5
Chr10
Chr15
ChrX
D
MRE
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
Chr1
Chr5
Chr10
Chr15
ChrX
E
MSE, ×10-3
14
12
10
8
6
3
2
0
Chr1
Chr5
Chr10
Chr15
ChrX
F
MAE, ×10-3
7
6
5
4
3
2
1
Chr1
Chr5
Chr10
Chr15
ChrX


A
Chr1 (qC3-qC5)
Experiment
Prediction
B
Experiment
Prediction
C
Expected distribution
Observed value
Frequency
0.12
0.10
0.08
0.06
0.04
0.02
0.00
100
200
300
400
500
600
Size of the overlap of predicted and experimentally validated ectopic interactions