nature methods

ARTICLES

https://doi.org/10.1038/s41592-020-0960-3

DeepC: predicting 3D genome folding using megabase-scale transfer learning

Ron Schwessinger[1,2,3], Matthew Gosden[1], Damien Downes[1], Richard C. Brown[3], A. Marieke Oudelaar[1,2], Jelena Telenius[2], Yee Whye Teh[4], Gerton Lunter[2,3] ✉ and Jim R. Hughes[1,2] ✉

Predicting the impact of noncoding genetic variation requires interpreting it in the context of three-dimensional genome architecture. We have developed deepC, a transfer-learning-based deep neural network that accurately predicts genome folding from megabase-scale DNA sequence. DeepC predicts domain boundaries at high resolution, learns the sequence determinants of genome folding and predicts the impact of both large-scale structural and single base-pair variations.

Most genetic variants associated with common diseases affect gene regulatory regions distal to target genes[1,2]. Genome three-dimensional (3D) structure is central to mediating these functional interactions, but its intricately convoluted and large-scale nature renders it challenging to understand and predict. Proposed machine learning and polymer modeling approaches to predict 3D genome structure have produced promising results, but none effectively integrates across resolutions. Methods that use information at the base-pair level focus on window-to-window-based predictions[3–5], while methods that incorporate a large genomic context do so by coarse segregation into genomic features[6–8] or polymer beads[9,10], thus compromising their ability to predict the impact of variation at base-pair resolution.

We propose that to accurately predict topologically associated domains (TADs) a model needs to capture sequence patterns across large genomic distances. Regulatory elements can interact over megabase (Mb) distances and boundary elements lying in between may alter chromatin contacts substantially. A chromatin interaction model should thus integrate information at the megabase scale. However, to predict the impact of genetic variation the model must also learn to interpret DNA sequence at base-pair resolution. Since 3D genome interactions are determined by genomic regulatory elements such as CCCTC-binding factor (CTCF) bound domain boundaries[11], a model that has learned the grammar of regulatory elements could help guide the prediction of 3D genome structure.

Based on these ideas, we developed deepC, a deep neural network that bridges the gap from base pairs to TADs. DeepC uses a transfer learning approach and tissue-specific Hi-C data to train models that predict genome folding from megabase windows of DNA sequence (Fig. 1a). The trained models can then be used to predict chromatin domain boundaries at high resolution and to identify the sequence determinants of genome folding. They allow us to predict the impact of genetic variants from large structural variations down to single-nucleotide polymorphisms (SNPs).

Results

A deep learning model for predicting chromatin interactions from megabase-scale DNA. We encode Hi-C data as a vector of pairwise interaction values between 5-kilobase (kb) genomic bins at distances of up to ~1 Mb (Fig. 1b). DeepC learns to predict these contact frequencies taking as input the underlying ~1-Mb window of DNA sequence. The deepC network architecture is constructed from a convolutional module with maximum pooling that has proved powerful for predicting chromatin features from DNA sequence[12,13]. This is followed by a dilated convolutional module that excels at incorporating large-scale context while maintaining resolution[14–16]. Finally, a fully connected layer integrates the detected patterns over 1 Mb of DNA sequence to predict chromatin folding.

We found two factors to be crucial for deepC's effective learning and generalization. First, we percentile-normalize the raw contact frequency signal in Hi-C data by genomic distance (Extended Data Fig. 1 and Methods), termed the 'skeleton'. This normalization reveals informative longer-range interactions and enhances the contrast at domain boundaries. Second, we use transfer learning[17] (Fig. 1a and Extended Data Fig. 2), a concept that has proved powerful in deep learning applications for image analysis and natural language processing. In a first phase of training, the initial convolutional module learns to predict a compendium of chromatin features such as open chromatin regions and CTCF binding sites across cell types[12,18,19]. Next, the convolutional module is stripped of the fully connected layer responsible for interpretation. Only the learned sequence patterns are transferred to the second training phase where they are refined, and the dilated module and fully connected layer are trained ab initio to predict chromatin interactions. The same weights pretrained on a chromatin feature compendium across cell types are used for transfer learning irrespective of the cell type of the Hi-C data source.

We trained deepC models on seven human[20] and one mouse[21] Hi-C datasets with different sequencing depths and at different resolutions (Supplementary Figs. 1–4). We focused our analysis on the primary GM12878 (~3.6 billion reads) and K562 (~1.3 billion reads) data, training models at 5-kb resolution. DeepC yields smooth but detailed predictions that resolve the hierarchical nature of TADs and insulated domains (Fig. 1c and Supplementary Fig. 1). In a cross-validation scheme across all chromosomes in GM12878 (Fig. 1d), deepC achieves an average, distance-stratified Pearson correlation between predictions and Hi-C skeleton of ~0.36 on raw skeleton data and ~0.57 when applying a small smoothing filter to the discrete and noisy skeleton (~0.28 and ~0.51 in K562, Supplementary Fig. 5). We compared deepC to a recently proposed

[1]MRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [2]MRC WIMM Centre for Computational Biology, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [3]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [4]Department of Statistics, University of Oxford, Oxford, UK. ✉e-mail: gerton.lunter@well.ox.ac.uk; jim.hughes@imm.ox.ac.uk

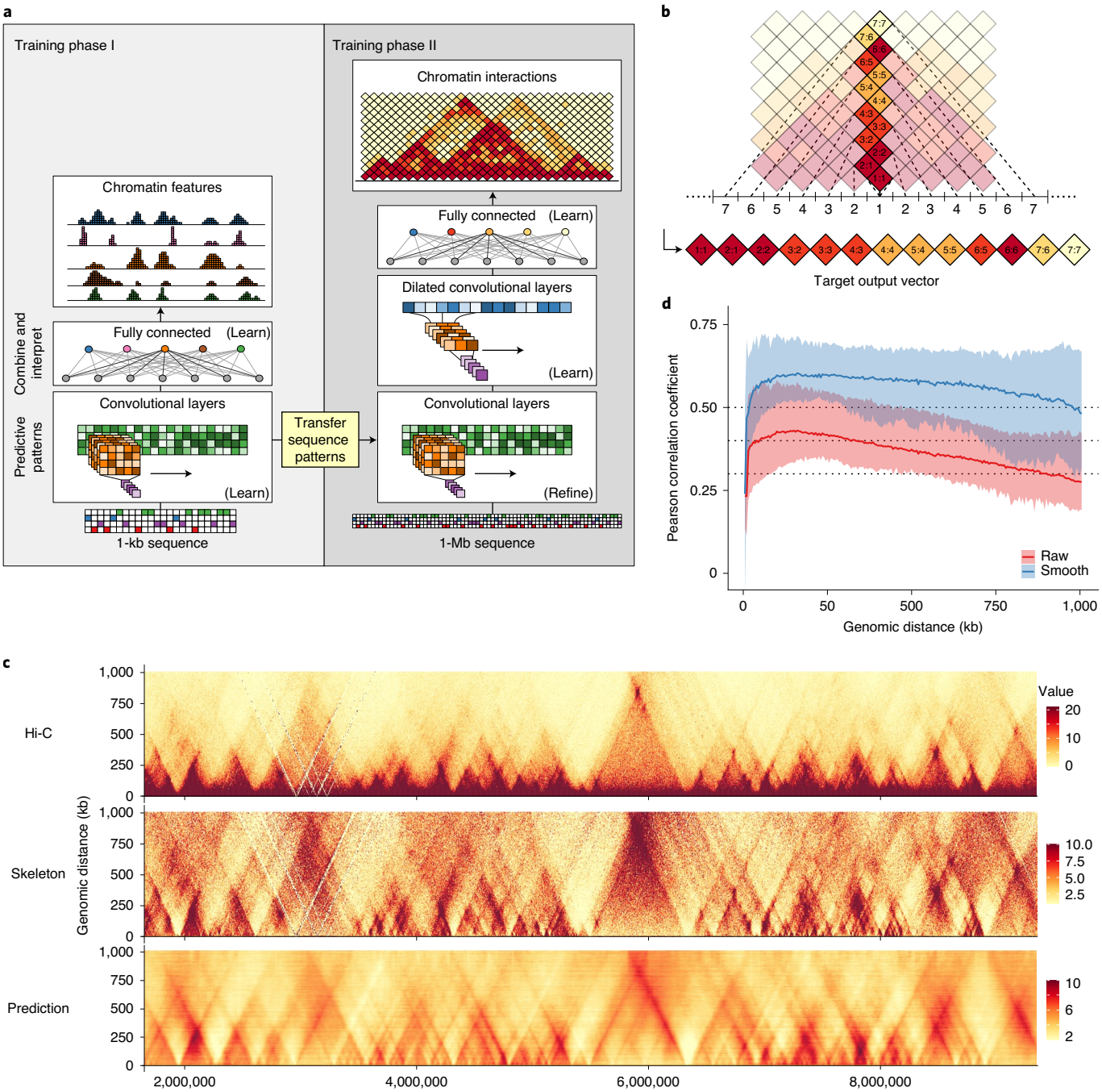


Fig. 1 | Predicting Hi-C interactions from DNA sequence. a, Overview of the deepC architecture and training workflow. **b**, Encoding of Hi-C data as target vector for prediction given a 1-Mb window of DNA sequence. **c**, Comparison of Hi-C data, the derived Hi-C skeleton and the interactions predicted from DNA sequence using deepC. Shown is a ~7-Mb region on hold-out chromosome 17. The color-coded values represent the normalized Hi-C, the skeleton transformed and the predicted interaction frequency. **d**, Distance-stratified Pearson correlation between the Hi-C skeleton and the deepC predictions in a cross-validation scheme across all chromosomes. Solid lines indicate the mean correlation value and the area indicates the space between the maximum and the minimum values over all chromosomes. Red shows the correlation with the raw and blue with the (5 × 5) mean filter smoothed skeleton values. Dotted lines are at 0.3, 0.4 and 0.5.

random-forest-based method HiC-Reg[8], which predicts chromatin interactions up to 1-Mb distance using chromatin features of the interacting windows and the window in between rather than using DNA sequence as input. We observed that deepC predicts the domain structure of unseen chromosomes more accurately (Supplementary Figs. 6 and 7).

Although we focused our main analysis on the deeply sequenced Hi-C data from Rao et al.[20] we hypothesized that deepC is able to predict chromatin interactions after training on data with substantially lower sequencing depth. To this end, we trained GM12878 models with Hi-C data downsampled from originally ~2.6 billion to 1 billion, 100 million and 10 million valid Hi-C contacts (Supplementary Fig. 8). The 1-billion and 100-million contact models still learned to predict chromatin structure, with a mean Pearson correlation of 0.46 and 0.4 between the 1-billion and 100-million models on hold-out chromosomes 16 and 17, respectively.

In contrast, the 10-million model failed to learn chromatin structure. While deepC can predict chromatin interactions from less deeply sequenced samples, we do note that dedicated methods have been proposed for increasing the resolution of Hi-C maps that take as input low-resolution maps directly[22,23] rather than predicting from DNA sequence.

We train separate models for each Hi-C dataset derived from a different cell type. These distinct models learn tissue-specific chromatin interactions (Extended Data Fig. 3 and Supplementary Fig. 9). We also trained a deepC model to predict chromatin interactions in multiple cell types jointly, but we found that the jointly trained network captured the tissue-specific Hi-C patterns less well compared to the individually trained models (Supplementary Fig. 10).

Validating deepC with high-sensitivity chromosome confirmation capture

We next sought to validate deepC predictions with an independent set of chromatin interactions. To this end, we used NG Capture-C[24] (Methods), which generates high-resolution interaction data from targeted viewpoints and identifies chromatin interactions at higher sensitivity than Hi-C. We captured the interactions of 220 viewpoints in two cell types (GM12878 and K562) covering 81 CTCF sites and 139 sites lying within insulated domains, not overlapping with regulatory elements to capture the domain structure. We observed good agreement between the predicted domain structure and interaction peaks in the NG Capture-C tracks (Extended Data Fig. 4), showing that deepC is capable of predicting true biophysical boundaries that are evident in these sensitive 3C assays but poorly captured by the original Hi-C, especially at lower sequencing depth.

Although deepC effectively captures the positions of boundary elements, some aspects of interactions are not captured fully by the model. When comparing the virtual4C track from the Hi-C skeleton and the predictions and distance-normalized NG Capture-C tracks from CTCF viewpoints (Supplementary Fig. 11) we saw that the NG Capture-C correlated more strongly with the Hi-C skeleton than with the predictions (Fig. 2a; Pearson correlation in GM12878, 0.59 versus 0.30; K562, 0.55 versus 0.37). This was due to the tendency of deepC predictions to de-emphasize the characteristic punctate nature of signal at the apex of interacting CTCF elements. This may be explained by an inability of deepC to model the detailed characteristics of the loop extrusion mechanism such as cohesin processivity, which may not be encoded in the local DNA sequence and may be more dependent on factors such as nuclear concentration of extruding factors[25]. In contrast, for intradomain viewpoints deepC correlates equally well with NG Capture-C data as NG Capture-C correlates with the Hi-C skeleton (Fig. 2a; GM12878, 0.30 versus 0.36; K562, 0.46 versus 0.42). Therefore, even though deepC is predicting interactions from sequence in these instances it performs as well as when comparing two different experimental sources of 3C data.

Taken together, these analyses indicated that deepC is capable of modeling the DNA encoded signals that determine the activity and position of boundary elements. To test this, we called boundaries within 1 Mb from the NG Capture-C viewpoints using the Hi-C data, the skeleton and the deepC predictions (Supplementary Figs. 12 and 13) using the established insulation-score-based approach[26] with parameters for high-resolution calling (Methods). We then compared the called boundaries from these three sources with the high-sensitivity NG Capture-C 3C data to quantify the enrichment of chromatin interactions (Fig. 2b and Supplementary Fig. 14). We observed clear enrichment over the deepC predicted boundaries indicating that they on average represent biophysical barriers to genome interactions. In contrast, the boundaries called directly from the Hi-C data and from the Hi-C skeleton showed less pronounced enrichment, suggesting that calling directly from the data, on average, captures boundaries less effectively at the available sequencing depths. We confirmed these results with boundaries called using TopDom[27] (Supplementary Fig. 15), a TAD caller that showed best overall robustness in a recent bench-marking study[28].

To visualize the coherence of the deepC predictions and called boundaries across specific loci at a sensitivity higher than the available Hi-C data, we used Tiled-C[29] (Methods), which generates Hi-C-like data for specific loci at high sensitivity and resolution. We performed Tiled-C for a selection of loci where deepC predicted fine-grained boundaries (Fig. 2c and Supplementary Fig. 16) or cell-type-specific patterns (Extended Data Fig. 3 and Supplementary Fig. 9). In line with the enrichment analysis, we confirmed that, when called at high resolution, deepC boundaries are evident and align well with the boundaries in this highly sensitive 3C data. The added benefit for boundary calling is particularly striking in the comparatively lower-coverage K562 Hi-C data (~1.3 billion reads) (Fig. 2c and Supplementary Fig. 16).

When comparing the overall structure of deepC predictions to the Hi-C and Tiled-C data we observed that deepC tends to predict interdomain interactions in the form of more pronounced stripes and dots, some of which are only faintly detectable in Hi-C and Tiled-C data and some appear new (Figs. 1c and 2c and Supplementary Fig. 9). This suggests that deepC tends to underestimate the insulation between domains. Future refinements to the model architecture might be able to better capture the interdomain insulation. The effect appears amplified when comparing skeleton transformed to raw data as necessary for Tiled-C.

Dissecting the sequence determinants of genome folding

DeepC allows us to dissect the sequence determinants of genome folding at base-pair resolution. To estimate the relative importance of every base pair for predicting chromatin interactions we employed the saliency score as a computationally efficient method adapted from image analysis[30]. The saliency score estimates how much the interaction prediction depends on each single base pair by calculating the gradient of the model output with respect to the sequence input (Methods). The saliency score predicts important regions and highlights transcription factor motifs within them (Extended Data Fig. 5).

Genome wide, we identify sharp saliency peaks at CTCF sites and broader saliency peaks at active promoters (Fig. 3a). As bases with high saliency scores mark positions predicted to be important for chromatin architecture, we hypothesized that mutations within these regions would be enriched for those affecting gene expression. To test this, we retrieved 6,607 GM12878 cell-type-specific eQTLs (GTEx v.7) that are located in open chromatin (DNase-seq) or CTCF sites (CTCF ChIP–seq, ENCODE), and are thus likely to lie in regulatory elements. We found that these eQTLs have significantly higher saliency scores than SNPs randomly resampled from the same regions ($P < 1 \times 10^{-85}$ using a two-sample Kolmogorov–Smirnov test) (Supplementary Fig. 17). This indicates that the deepC saliency score can be used to fine map eQTLs when expression changes are mediated through an impact on chromatin architecture.

A long-standing question has been which functional elements within the genome underlie the patterns of genome folding? To investigate this, we performed an in silico deletion screen of all active elements genome wide and used deepC to assess their importance for chromatin interactions (Fig. 3b). As expected, we found that deleting CTCF sites as well as enhancers and promoters with proximal CTCF binding has the strongest average predicted impact. We also found promoter and enhancers without proximal CTCF binding sites to be important, with deletions of promoters on average having a stronger effect. In addition, deletions of promoters and enhancers with strong activity-associated histone marks have a higher predicted impact than those without such marks.

Our analysis indicates that, in addition to known factors such as CTCF binding and orientation, active regulatory elements, in particular promoters, are critical elements for effectively predicting

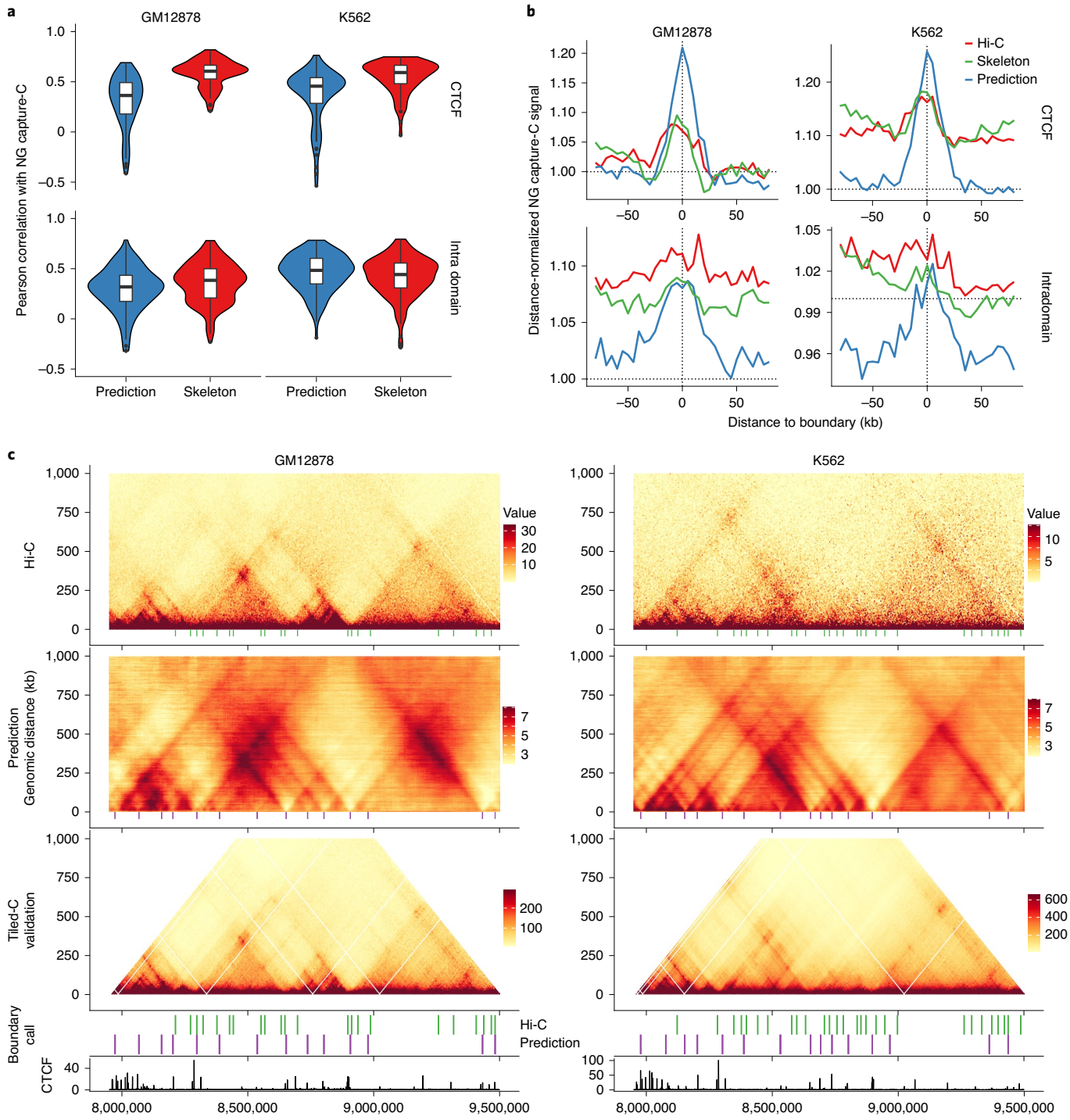


Fig. 2 | Validation of deepC predictions. a, Comparison of the correlation between the validation NG Capture-C profiles and the virtual4C profiles derived from the Hi-C skeleton (red); and the deepC prediction map (blue) from all viewpoints in two cell types. Compared are $n = 81$ CTCF and $n = 139$ intradomain viewpoints. Boxplots, median middle thick line, 25th and 75th percentile left and right hinge, respectively, whiskers stretch up 1.5 times the interquartile range. **b**, Meta-profiles of the average NG Capture-C signal over domain boundaries called at high resolution from the Hi-C data, the skeleton and the deepC predicted interaction map. Shown is the mean distance-normalized NG Capture-C signal relative to the boundary center. The labels CTCF and Intradomain refer to the NG Capture-C viewpoint fragments. These were designed to overlap either CTCF sites or to lie within insulated domains but not overlap with regulatory genomic elements, so as to capture the domain structure and not the interactions of specific genomic elements. **c**, Shown are Hi-C data, the deepC predicted interaction map and the Tiled-C high-sensitivity map over a locus on chr17, a hold-out chromosome. The color-coded values represent the interaction frequency in normalized Hi-C, the prediction, and as observed with Tiled-C. Boundaries called at high resolution from Hi-C (green) and deepC predictions (purple) are aligned under the respective map and the Tiled-C map. Cell-type-specific CTCF ChIP-seq tracks are visualized below. For contrast, Hi-C and Tiled-C data were bounded between 5 and 95% of coverage and deepC predictions were bounded between a 2 and 8 predicted regression score.

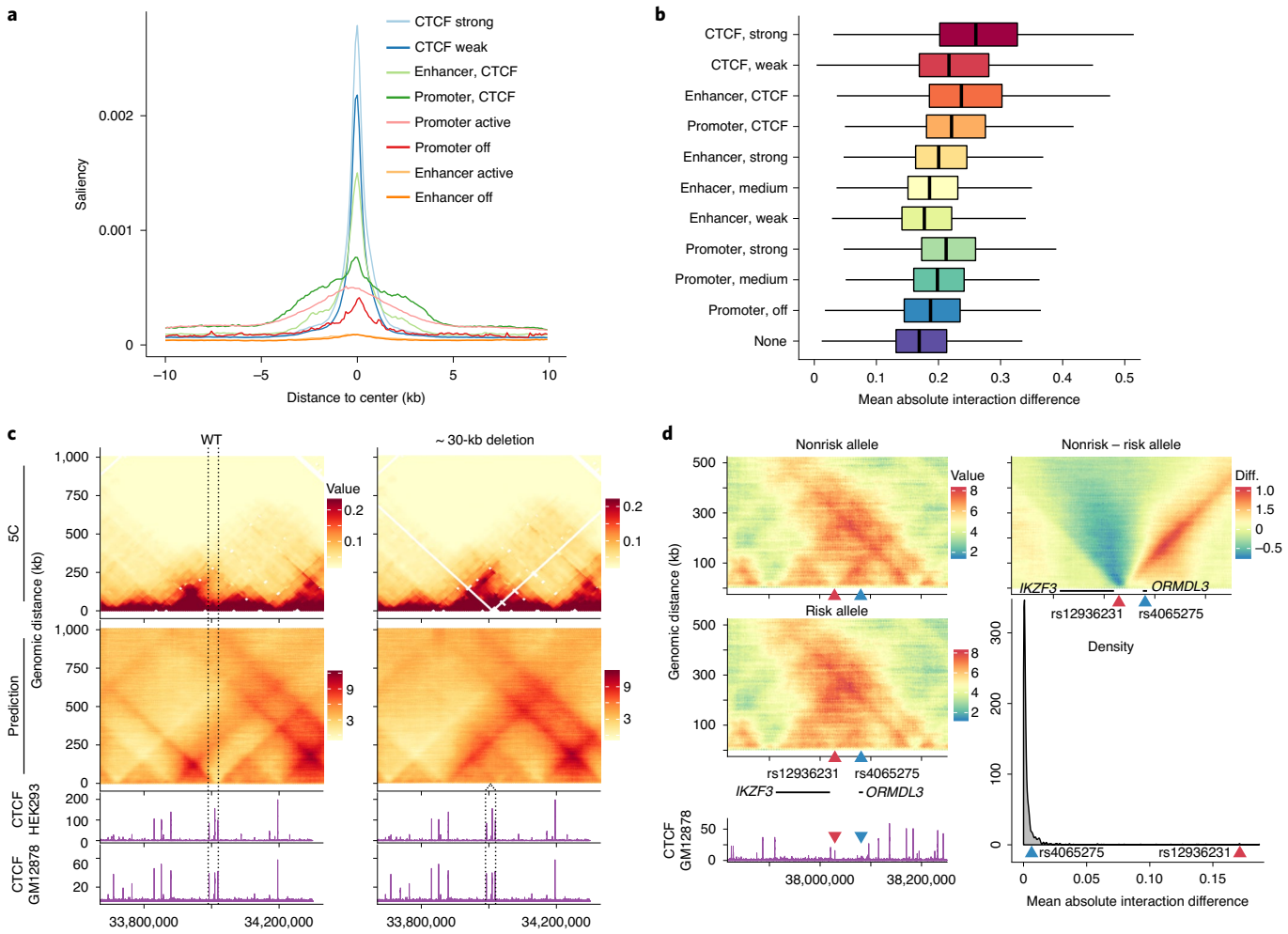


Fig. 3 | DeepC for dissecting the determinants of genome folding and predicting the impact of variation. a, Mean saliency enrichment in 10-bp bins over different classes of regulatory elements genome wide in GM12878 cells. **b**, Mean predicted chromatin interaction difference between reference and variant when deleting all open chromatin and CTCF sites and shuffled background sequences individually, genome wide in GM12878 cells. Differences were quantified as the mean absolute interaction difference per bin-to-bin interaction. Each deletion was classified based on the overlap with genome segmentation classes. A total of $n = 96{,}805$ was analyzed. Boxplots show the median as thick lines, and the 25th and 75th percentiles as left and right hinges, respectively. The whiskers stretch up to 1.5 times the interquartile range away from the respective whisker. Outliers outside this range were omitted from the plotting. **c**, Effect of deleting a ~30-kb fragment containing four CTCF sites. Shown is the effect as validated by 5C data in wild-type (WT) and mutant in HEK293 cells by Hnisz et al.[31], and the predicted effect using the GM12878 model. CTCF ChIP-seq tracks from HEK293 and GM12878 cells are visualized below. Dotted lines mark the deleted fragment. The color-coded values represent the normalized 5C and the predicted interaction frequency, respectively. The mutant deepC prediction was shifted to be centered on the deleted site. **d**, Predicting the effect of two asthma associated SNPs with the GM12878 model. Shown is the prediction for nonrisk and the risk allele, the differential map (nonrisk–risk) and a GM12878 CTCF ChIP-seq track. Location of the SNPs is indicated by triangles (red, rs12936231; blue, rs4065275). Black lines indicate the location of *ORMDL3* and *IKZF3*, and only those two genes are highlighted for clarity. The color-coded values represent the predicted interaction frequency and the difference (diff.) between nonrisk and risk. Comparing the predicted SNP effects against 1,000 randomly sampled SNPs within CTCF sites places rs4065275 in the top 11% and rs12936231 in the top 1% of predicted mean absolute interaction difference.

genome interactions. Furthermore, deepC predicts boundaries not associated with CTCF sites. Taken together, this indicates that deepC has learned aspects of a complex grammar of genome folding beyond CTCF motifs and their relative orientation, and suggests a causal role for regulatory elements in defining and stabilizing 3D genome structure.

Predicting the impact of sequence variation on genome folding

To test deepC's ability to predict the impact of sequence variation on genome folding we used two well-characterized examples from the literature. Hnisz et al. showed that a ~30-kb CRISPR-mediated deletion in human embryonic kidney 293T (HEK293T) cells, encompassing four CTCF sites at the *LMO2* locus, leads to a local rearrangement of the chromatin structure as confirmed by 5C[31]. Computationally reproducing this deletion in GM12878 (Fig. 3c) and K562 cells (Supplementary Fig. 18) recapitulates the domain fusion observed by Hnisz and colleagues. Furthermore, using deepC we computationally deleted the CTCF sites individually, predicting that no single deletion alone is sufficient for causing the rearrangement (Supplementary Fig. 19), suggesting multiple redundant boundaries at this region.

Crucially, our sequence-based model can predict the impact of single base-pair variants. To demonstrate this, we tested two asthma-risk associated SNPs[32] shown to impact *ORMDL3* expression in immune cells. Schmiedel et al. demonstrated that the SNPs

impact CTCF binding sites, disrupting enhancer–promoter interactions of *ORMDL3* in CD4-positive T cells. DeepC predictions recapitulate the loss of a boundary element, insulating *ORMDL3* from downstream interactions (Fig. 3d). When testing the individual SNPs (Supplementary Fig. 20a,b), deepC predicts the rs12936231 risk allele to have a strong effect on genome folding (mean absolute interaction difference 0.176). In contrast, although the rs4065275 risk allele suggests a boundary creating effect, the predicted strength is weak (0.006). To put these predicted SNP effects into context, we compared them to the in silico deletion screen effects (Supplementary Fig. 20c). The effect of rs12936231 lies above the 25th percentile of the 500-bp, weak CTCF site deletions. In addition, we sampled 1,000 SNPs from CTCF sites (Fig. 3d) as well as from promoters, enhancers and background sequences (Supplementary Fig. 20d). In comparison to the sampled 1,000 CTCF SNPs, rs4065275 lies within the top 11% and rs12936231 within the top 1%. Taken together, deepC prioritizes rs12936231 as the likely causal variant. DeepC predicts this effect in GM12878 (immortalized B cells) but not in K562 (myeloid leukemia cell line with erythroid characteristics) or IMR90 (human embryonic lung fibroblasts), pointing to a potential lymphoid specific effect (Supplementary Fig. 21).

Discussion

Mammalian chromatin architecture folds at the megabase and sub-megabase scales, constraining distal regulatory interactions within TADs and smaller insulated domains[11,33,34]. Ultimately, chromatin interactions are encoded in the DNA sequence through an intricate interplay of protein binding sites and other sequence determinants. Understanding the link between individual sequences and large-scale chromatin interactions at base-pair resolution is a key challenge for understanding chromatin architecture and its role in gene regulation[35,36]. We developed deepC to traverse the gap between base-pair sequences and megabase structures. DeepC is a sequence-based deep learning model that predicts chromatin interactions from DNA sequence while integrating a sequence context of megabase scale. This scale of analysis is necessary for accurate prediction of chromatin interactions, which in turn allows for the determination of the elements driving these interactions and assessment of the mutations disrupting them.

We found deepC models to yield substantially more accurate predictions when we preseeded the model with hidden layers optimized to predict a compendium of chromatin features. This allows deepC to predict intricate chromatin interactions even when trained on low-depth, low-resolution Hi-C data. By validating the results with NG Capture-C and Tiled-C, each of which are 3C methods capable of extreme depth and sensitivity, we showed that the deepC approach effectively increases the resolution of Hi-C data at which domain boundaries can be called and interpreted. When deriving interaction profiles from single viewpoints via virtual4C we observed that deepC predictions from sequence are comparable to but do not outperform the original Hi-C map. Specifically, deepC tends to predict the pinnacles of domains at a lower resolution than is observed in Hi-C. Improvements to the data encoding and the network architecture may help to address this limitation.

The deepC predictions exhibit stripes more frequently than Hi-C data and Tiled-C maps, which show more dot-shaped interaction patterns in corresponding locations. However, both stripes and dot interactions can be explained by the mechanics of loop extrusion, depending on the concentration of CTCF and frequency and location of cohesin loading[25], which are not explicitly included in a sequence-based model. Preseeding increases the predictive ability of the network, but only to the extent of current knowledge and the available assays for genomics features that mark functional regions in the genome and that can be used to learn predictive sequence patterns. The effect of this is demonstrated by deepC's comparatively poor prediction of chromatin domains in larger regions associated with the nuclear lamina (see the large domain on the left of Extended Data Fig. 3 and Supplementary Fig. 9). These domains are generally repressive and are refractory to the formation of open chromatin or even the binding of protein such as CTCF. The tissue-specific annotation of such domains across cell types is too sparse at the moment for effective training and it is currently not clear which chromatin modifications mediate this repressive behavior. However, it is straightforward to incorporate relevant new genomic data into the proposed framework as they become available.

We demonstrated that deepC can be used to fine map the sequence determinants of chromatin architecture at base-pair resolution and link these with effects on gene expression. Additionally, our genome-wide deletion screen of potential regulatory elements sheds light on the mechanics of chromatin interactions. It confirmed that CTCF binding site deletions are most likely to cause strong chromatin interaction changes. DeepC also indicates that both promoters and enhancers contribute to genome folding, in addition to CTCF. Generally, promoter deletions have a higher predicted effect on genome organization than enhancer deletions, and we find that deletions of enhancer and promoters associated with active chromatin marks have a higher predicted impact than those without such marks. Our observations are in line with findings from orthologous methods, which find CTCF binding, open chromatin, active histone marks and RNA-seq to be most predictive[4–8,10]. The finding that identifying active promoters and enhancers, in addition to CTCF binding sites, is required to accurately predict 3D genome structures indicates that these elements play an important role in establishing these structures, possibly via recruitment of its components and by actively stabilizing certain loops.

We believe deep learning-based genome folding predictions will facilitate chromatin architecture research. In a parallel study, Fudenberg et al.[37] have developed an alternative model (Akita) to accurately predict interactions in megabase-scale loci. DeepC and Akita have a similar convolutional module as network base, but vary notably in the remaining network structure as well as the data encoding and training scheme. We believe that future comparative study and consolidation between these advances will bring further insights into genome function.

Here we present deepC as a valuable tool for dissecting the functional elements that shape chromatin architecture and for predicting the impact of sequence changes from single base pair to structural variants. Furthermore, deepC represents a step toward predictive models of gene regulation that integrate the intricate and long-ranged chromatin landscape of mammalian genomes.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-020-0960-3.

Received: 28 January 2020; Accepted: 20 August 2020;
Published online: 12 October 2020

References

1. Hindorff, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106**, 9362–9367 (2009).
2. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
3. Schreiber, J., Libbrecht, M., Bilmes, J. & Noble, W. S. Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture. Preprint at *bioRxiv* https://doi.org/10.1101/103614 (2017).
4. Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* **48**, 488–496 (2016).

5. Li, W., Wong, W. H. & Jiang, R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Res.* **47**, e60 (2019).
6. Qi, Y. & Zhang, B. Predicting three-dimensional genome organization with chromatin states. *PLoS Comput. Biol.* **15**, e1007024 (2019).
7. Belokopytova, P. S., Nuriddinov, M. A., Mozheiko, E. A., Fishman, D. & Fishman, V. Quantitative prediction of enhancer–promoter interactions. *Genome Res.* **30**, 72–84 (2020).
8. Zhang, S., Chasman, D., Knaack, S. & Roy, S. In silico prediction of high-resolution Hi-C interaction matrices. *Nat. Commun.* **10**, 5449 (2019).
9. Buckle, A., Brackley, C. A., Boyle, S., Marenduzzo, D. & Gilbert, N. Polymer simulations of heteromorphic chromatin predict the 3D folding of complex genomic loci. *Mol. Cell* **72**, 786–797.e11 (2018).
10. Bianco, S. et al. Polymer physics predicts the effects of structural variants on chromatin architecture. *Nat. Genet.* **50**, 662–667 (2018).
11. Hnisz, D., Day, D. S. & Young, R. A. Insulated neighborhoods: structural and functional units of mammalian gene control. *Cell* **167**, 1188–1200 (2016).
12. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* **12**, 931–934 (2015).
13. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).
14. Kelley, D. R. et al. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* **28**, 739–750 (2018).
15. Yu, F. & Koltun, V. Multi-scale context aggregation by dilated convolutions. Preprint at http://arxiv.org/abs/1511.07122 (2015).
16. van den Oord, A. et al. WaveNet: a generative model for raw audio. Preprint at https://arxiv.org/abs/1609.03499 (2016).
17. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? *Adv. Neural Inf. Process. Syst.* **4**, 3320–3328 (2014).
18. Bernstein, B. E. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
19. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
20. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
21. Bonev, B. et al. Multiscale 3D genome rewiring during mouse neural development. *Cell* **171**, 557–572.e24 (2017).
22. Zhang, Y. et al. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nat. Commun.* **9**, 750 (2018).
23. Liu, Q., Lv, H. & Jiang, R. hicGAN infers super resolution Hi-C data with generative adversarial networks. *Bioinformatics* **35**, i99–i107 (2019).
24. Davies, J. O. J. et al. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat. Methods* **13**, 74–80 (2016).
25. Fudenberg, G. et al. Formation of chromosomal domains by loop extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
26. Crane, E. et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240–244 (2015).
27. Shin, H. et al. TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.* **44**, e70 (2015).
28. Zufferey, M., Tavernari, D., Oricchio, E. & Ciriello, G. Comparison of computational methods for the identification of topologically associating domains. *Genome Biol.* **19**, 217 (2018).
29. Oudelaar, A. M. et al. Dissection of the 4D chromatin structure of the α-globin locus through in vivo erythroid differentiation with extreme spatial and temporal resolution. Preprint at *bioRxiv* https://doi.org/10.1101/763763 (2019).
30. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: visualising image classification models and saliency maps. In *Proc. 2nd International Conference on Learning Representations (ICLR 2014) Workshop Track* (2013).
31. Hnisz, D. et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016).
32. Schmiedel, B. J. et al. 17q21 asthma-risk variants switch CTCF binding and regulate IL-2 production by T cells. *Nat. Commun.* **7**, 13426 (2016).
33. Robson, M. I., Ringel, A. R. & Mundlos, S. Regulatory landscaping: how enhancer-promoter communication is sculpted in 3D. *Mol. Cell* **74**, 1110–1122 (2019).
34. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin domains: the unit of chromosome organization. *Mol. Cell* **62**, 668–680 (2016).
35. Marti-Renom, M. A. et al. Challenges and guidelines toward 4D nucleome data and model standards. *Nat. Genet.* **50**, 1352–1358 (2018).
36. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018).
37. Fudenberg, G., Kelley, D. R. & Pollard, K. S. Predicting 3D genome folding from DNA sequence with Akita. *Nat. Methods* https://doi.org/10.1038/s41592-020-0958-x (2020).

Methods

Chromatin feature data. As human chromatin feature compendiums, the ENCODE[18] and Roadmap[19] chromatin data used in DeepSEA[12] were used. Narrow peak calls (hg19) for 918 experiments were downloaded. The data were supplemented with additional erythroid linage data. Five sets of ATAC-seq data from Corces et al.[38], two DNase-seq experiments[39] and ten ATAC-seq and one CTCF ChIP–seq experiments from Downes et al. along with in-house erythroid differentiation[40] were used. All data used are listed in Supplementary Table 1. All additional data were aligned to hg19 using the NGseqBasic pipeline[41]. Peaks were called with macs2 (ref. [42]) (default parameters, -q 0.01). The peak signals were aggregated following the procedure described in Zhou et al.[12]. In brief, the genome was split into 200-bp bins. All peak calls were intersected with these bins. If a bin overlaps a peak call to at least 50% (100 bp), the bin was labeled as belonging to that dataset class. All genomic bins that do not intersect with at least one peak call were discarded. Only autosomes were used for all analysis.

Mouse chromatin data were retrieved from ENCODE[18]. Histone modification peak calls were downloaded from the ENCODE data portal. For DNase-seq, ATAC-seq and transcription factor ChIP–seq data, the aligned .bam files were downloaded and peak called with macs2 (ref. [42]) as described above. Replicates were collapsed into unions. All mouse data used are listed in Supplementary Table 1.

Hi-C data. Publicly available, deeply sequenced Hi-C data from Rao et al.[20] were used. The available 5- and 10-kb resolution intrachromosomal contacts maps of seven cell lines (and 1-kb data from GM12878) were downloaded and normalized using the provided KRnorm factors. Four replicates of mouse embryonic stem cell data[21] were retrieved as raw fastqs from (GSE96107).

Hi-C encoding for deep learning. The genome was divided into bins matching the bin size of the respective Hi-C data resolution used for training (1, 5 and 10 kb). For every stretch of DNA of size 1 Mb plus bin size in bp (for example, 1,005,000 for 5-kb bins), the chromatin interactions associated with the window were assigned as squares in a vertical, zig-zag pole over the center of the sequence window (see Fig. 1b). Every square encodes the Hi-C interactions observed between two bin-sized windows of increasing distance (up to 1 Mb away). By sliding the large DNA stretch over a chromosome with a bin-sized increment, this encoding recovers the chromosome wide Hi-C map up to an interaction distance of 1 Mb. Regions with a median interaction count along this pole of 0 where excluded from training. The Hi-C data were percentile normalized across individual chromosomes for every interaction distance in bin sizes. It is of particular interest to resolve high levels of Hi-C interactions at high resolution and only a low percentage of chromatin interactions is expected to yield strong interactions at larger distances such as the corners of TAD triangular structures. Thus, the percentile normalization was designed to better resolve these high interaction levels at larger distances by using uneven percentiles in a pyramid-like scheme (from low to high 2×20%, 4×10%, 4×5%; see Extended Data Fig. 1). The identifier of the respective pyramid percentile (1–10) was stored. The chromatin interaction network was then trained to predict the percentile identifier as a regression problem (see below).

Deep neural network architectures and training. A two-step training process with transfer learning (Extended Data Fig. 2) was used. First, a convolutional neural network was trained to predict chromatin features from 1 kb of DNA sequence, using the compendium of 936 datasets described above. The principle network architecture was adapted from DeepSEA[12]. Five convolutional layers (hidden units 300, 600, 600, 900, 900; filter widths 8, 8, 8, 4, 4, 4) with ReLU activation, maximum pooling (widths 4, 5, 5, 5, 2) and dropout (rate 0.2) were used followed by a fully connected layer with sigmoid activation to output individual probabilities for each chromatin feature class (multilabel classification). The network parameters were trained by minimizing the sum of the binary cross entropies using the ADAM optimizer (epsilon 0.1) in batches of 100. Batch size, dropout rate, learning rate and filter size were optimized by grid search.

Second, a chromatin interaction network was trained to predict Hi-C interaction from DNA sequence. The chromatin interaction network takes as input 1 Mb + 1× Hi-C bin size (bp), (for example, 1,005,000 for a 5-kb bin network). The first module consists of five convolutional layers, with ReLU, maximum pooling and dropout with the dimensions and hyperparameters matching the chromatin feature network. The hidden weights were initialized by seeding with the weights of the trained chromatin feature network from step one. All chromatin features were used for pretraining and the same weights were used for seeding the chromatin interaction network training independent of the Hi-C data cell type.

The second module is a series of ten dilated, gated one-dimensional convolutional layers with residuals[16]. Gated convolutional layers require training of double the amount of filter parameters but have the potential of modeling more complex functions through their multiplicative units. The residual units allow information to propagate more easily through the network without having to necessarily pass through convolutions[43]. One hundred hidden units were used, and dilation rates were increased exponentially to reach the full sequence context in the last layer (1, 2, 4, 8, 16, 32, 64, 128, 256, 1). The dilated layers were followed by a fully connected layer. Outputs are the predicted interaction strengths (in units matching the percentile normalization). The model was trained with ADAM (epsilon 0.1) to minimize the sum-of-squares error between the outputs and the true percentiles. Graphics processing unit (GPU) memory limited us to using a batch size of one. Hidden units (for dilated layers), dropout rate, learning and ADAM epsilon were optimized using grid search.

Network training, computational resources and limitations. For both training procedures, the data were split into training, validation and test set based on chromosomes. For the chromatin feature network, chr11 and chr12 were used for validation and chr15, chr16 and chr17 for testing. For the chromatin interaction network, to increase the number of training examples the same validation chromosomes were used but only chr16 and chr17 were used as test chromosomes.

All models were trained on NVIDIA Titan V cards with 12 GB of video memory. Training on smaller cards is possible but slower. The final models have ~60 M parameters. Scaling the models to larger DNA inputs will likely benefit from network pruning or a refined architecture.

Fully training the chromatin feature network required 14 epochs with about 8 h per epoch. The training set order was reshuffled after every epoch. To minimize the number of times large chunks of DNA sequence had to be loaded into memory the network was trained on one chromosome at a time. Within a chromosome, the order in which training batches were drawn was random. We observed that the chromatin interaction network, when seeded with the pretrained weights in the first convolutional filters, converged quickly after training on ~3–6 chromosomes and only marginally improved after training for an entire epoch or longer. Models were trained for one full epoch as we have not observed notable improvement after training for longer, and the limited batch size as well as the network complexity make training slow. For cross-validation, we trained multiple iterations holding out different chromosomes from training.

While training networks is only feasible with GPU support, predictions with trained models can be run on the central processing unit only. For example, predicting the impact of a variant requires ~5 min with GPU and ~2 h with only central processing unit support.

Predicting changes in chromatin interactions. For calculating differences in chromatin interactions, the interactions over the reference sequence were predicted for every position that is within 1 Mb (plus 1× Hi-C bin size) of the sequence variant. This matches the spatial reach of the respective models. The reference sequence was then modified to match the sequence variant of interest. After predicting the chromatin interactions over the variant sequence, the difference was quantified by calculating the absolute difference between reference and variant prediction at each interaction bin and summarized as the mean absolute difference over all covered interactions.

Distance-stratified correlation. The Pearson correlation coefficient was calculated between the Hi-C skeleton and the deepC predicted regression score. Note that the skeleton percentiles are discrete (percentile tag 1–10), while the regression score is continuous. The Hi-C skeleton is noisy even at very deep sequencing depths (for example, GM12878). Therefore, a small mean filter was employed using a 5×5 window to smooth the skeleton. The distance-stratified correlation was calculated between the prediction and the raw or the smoothed skeleton.

Comparison against HiC-Reg. The available CrossChrom predictions, trained on chr14 and predicted on chr17, were downloaded from the Zenodo repository linked to the HiC-Reg article[8]. HiC-Reg predictions were distance normalized as described above.

Realigning and downsampling Hi-C data. The primary GM12878 replicate was realigned from raw fastqs using HiCPro[44]. The valid Hi-C interactions were downsampled to achieve 1 billion, 100 million and 10 million valid interactions, respectively. Mouse embryonic stem-cell Hi-C data were aligned and processed from raw fastq data using HiCPro[44].

Selection of validation capture probes. A total of 220 viewpoints were selected for validating the deepC predictions, specifically selecting genomic locations where the Hi-C data and deepC predictions differed in detail or where the deepC predicted structures were only very faintly noticeable in the Hi-C data. Two sets were designed, one targeting 81 CTCF sites and one targeting 139 intradomain viewpoints that lie within a distinct Hi-C/deepC domain but are not intersecting with any potential functional elements. Capture probes were designed using CapSequm (http://apps.molbiol.ox.ac.uk/CaptureC/cgi-bin/CapSequm.cgi), filtering out repetitive probe regions as described in the online documentation. For the final probe design see Supplementary Table 2.

Cell culture and fixation. Human GM12878 lymphocyte cell lines were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research, and cultured in RPMI 1640 supplemented with 15% FBS, 2 mM L-glutamine and 100 U ml^{-1} pen-strep at 37 °C in a 5% CO_2 incubator. K562 cells were supplied by the WIMM transgenics facility. Cells were maintained in RPMI 1640 media supplemented with 10% FCS at 37 °C in a 5% CO_2 incubator. Both cell

types were fixed and processed using the same protocol. Cells were resuspended at 1×10^6 cells per ml and fixed at room temperature with 2% v/v formaldehyde for 10 min. Fixation was quenched with 120 mM glycine. Cells were washed with ice-cold PBS. Cells were resuspended in cold lysis buffer (10 mM Tris, 10 mM NaCl, 0.2% Igepal CA-630 and complete proteinase inhibitor (Roche)) and snap frozen to −80 °C. See Reporting Summary for additional details.

3C library preparation. The 3C libraries were prepared as described previously[24] with the following modifications: centrifugations were performed at 500 relative centrifugal force (rcf), thermomixer incubations were set to 500 r.p.m. and, following ligation, chromatin was pelleted by centrifugation (15 min, 4 °C, 500 rcf) and the supernatant discarded. To increase sequencing depth and minimize PCR duplicates, experiments were performed in technical triplicates and four unique adapters were used per replicate.

NG Capture-C. Double capture was performed as described previously[24] with biotinylated oligonucleotides (IDT xGen Lockdown Probes) in two pools (Supplementary Table 2) with 3 pg of each oligonucleotide per 3C library. The generated NG Capture-C libraries were sequenced using Illumina sequencing platforms (V2 chemistry, 150-bp paired-end reads) and data collected using the NextSeq System Suite (v.2). To resolve even subtle changes in chromatin interaction domains at high resolution, the libraries were deeply sequenced (GM12878 CTCF, 128 million; GM12878 intradomain, 118 million reads; K562 CTCF, 302 M; K562 intradomain, 289 million reads). All technical replicates were merged for the analysis.

Tiled-C. Tiled-C generates Hi-C-like data focused on loci of interest at greater depth by using an oligonucleotide capture enriching a 3C library for viewpoints tiled over the regions of interest. Tiled oligonucleotides were designed using the design approach and tool described in Oudelaar et al.[29] (https://oligo.readthedocs.io/en/latest/). A panel of double-stranded capture oligonucleotides from Twist Bioscience (custom probes for next-generation sequencing target enrichment) was used. The Tiled-C procedure was performed as described in Oudelaar et al. using the 3C libraries from GM12878 and K562 cells. All biological and technical replicates were merged for the analysis. For a summary of the regions of interest and designed probes, see Supplementary Table 3.

NG Capture-C analysis. NG Capture-C data were mapped, quality controlled and visualized using CcseqBasicS[45] following the procedure described previously[24].

Tiled-C analysis. Tiled-C data were mapped and quality controlled using the tiled mode of CcseqBasicS described above. Each region was normalized using the iterative correction and eigenvector decomposition (ICE) implementation of HiCPro with default parameters.

Distance-normalized NG Capture-C tracks. To compare them to the Hi-C skeleton, NG Capture-C tracks were normalized for distance dependence per viewpoint. The number of interactions with each restriction enzyme fragment was extracted and the distance to the viewpoint recorded. The interactions were then normalized for the total number of *cis* interactions for the respective viewpoint. Pooling this information across all viewpoints, we observed that the distance decay approximately follows a log–log linear trend when we split the data into three distance bins (close, intermediate and far). The distance thresholds for these bins were empirically optimized for every NG Capture-C set (see Supplementary Table 4), excluding all interactions closer than 2.5 kb to the respective viewpoint. The distance decay is then approximated by three linear regression fits, one for each distance bin. The distance-normalized interactions were calculated per viewpoint by dividing the observed *cis*-normalized interactions with the expected interactions from the linear fit at the respective distance.

Insulation-score boundary calling. Interaction domain boundaries were called using the Hi-C data, the Hi-C skeleton and the deepC predicted interactions using an insulation-score-based approach that was adapted from Crane et al.[26] . Using the 5-kb bin-sized data, the mean insulation-score profile was calculated based on a 25-kb window to allow for a more intricate boundary call. The first derivative, or delta vector, of the insulation-score profile was approximated using a one-dimensional Sobel operator. Zero crossings in this delta vector represent local minima and maxima of the insulation score. Maxima were discarded. The remaining boundaries were further filtered by calculating the approximation of the second derivative of the insulation profile using the same procedure described above. The height of this delta2 vector reflects the change in delta, with sharper boundaries having a higher delta2 score. Boundaries with a delta2 score smaller than 0.1 were removed and the remaining boundaries were stratified based on their delta2 score.

TopDom. TopDom was retrieved from the gitHub implementation (https://github.com/HenrikBengtsson/TopDom). Boundaries were called using the window parameters 5, 10 and 20. Boundaries between all types of called domain were used.

Distance-normalized NG Capture-C signal over boundaries. The mean, distance-normalized NG Capture-C signals over boundaries were calculated. In NG Capture-C tracks from single viewpoints, domain boundaries can be subtle and become harder to detect the further away from the viewpoint they are located. Therefore, boundaries further then 1 Mb away from a viewpoint were excluded. The mean normalized Capture-C signal over boundaries relative to their center was calculated.

Virtual4C from Hi-C skeleton and deepC maps. By extracting all interacting windows with a viewpoint of interest, Hi-C data can be transformed into virtual4C profiles. For this work, virtual4C profiles were derived from the Hi-C skeleton and the deepC predictions yielding distance-normalized profiles. Virtual 4C profiles from the Hi-C skeleton and deepC predictions were compared to distance-normalized NG Capture-C tracks by calculating the respective Pearson correlation of all interactions within 1 Mb from a given viewpoint. Because the skeleton percentiles are discrete and punctuate, a running mean smoothing window of 25 kb was applied. In contrast, the deepC predictions are smooth and therefore no additional smoothing was applied.

Chromatin segmentation. GM12878 and K562 chromatin data were downloaded from the ENCODE data portal (see Supplementary Table 5). Filtered alignments to hg19 were downloaded and replicates were merged. Peaks were called using macs2 (ref. [42]) with default settings and -q 0.01. Deeptools[46] was used to create bigwig coverage tracks. DNase-seq and CTCF ChIP–seq peaks were merged to a union set merging peaks within 10 bp of each other using bedtools[47] (bedtools merge -d 10). Union peaks were formatted to 600-bp elements centered on the peaks. Deeptools was then used to extract the read coverage for each chromatin dataset over each peak union element. For this, elements were extended to 1,000 bp to better capture flanking histone modifications. Using the derived count matrix, chromatin classes were segmented using GenoSTAN[48] running on the elements rather than entire chromosome stretches. The hidden Markov model was trained using the Poisson log-normal distributions. Twelve classes were fitted and merged into 11 classes based on the similarity of the chromatin signatures. The classes were manually curated and classified into promoter, enhancer and CTCF sites with varying activity levels based on H3K27ac coverage.

Saliency score. Adapted from image analysis[30], the saliency score serves as a proxy for the importance of every base pair to the interaction predictions. Explicitly, the saliency score was calculated as the dot product of the gradient of the model output with respect to the sequence input and the one-hot encoded DNA sequence input. This effectively masks the impact of nonpresent bases. For a given window, the saliency score relates to the interaction pole on the center. To visualize saliency tracks, the sequence window was moved in bin-sized steps and the saliency per base pair was averaged over all sequence windows (sized 1 Mb plus bin size) that include the respective base pair. To simplify visualization and interpretation, the absolute value of the saliency score was used. Metaplots were computed with deepTools.

eQTL data analysis. EBV transformed lymphocyte specific eQTLs were retrieved from GTEx (v.7 accessed from the GTEx portal, 1 March 2019). A union of DNase-seq and CTCF ChIP–seq peaks was created using bedtools merge. The eQTL SNPs were filtered for intersection with the union of GM12878 open chromatin and CTCF peaks. Indels were removed. A background SNP set was constructed by shuffling the eQTL SNPs on the respective same chromosome and forcing them to stem from within the union peaks (bedtools shuffle -chrom -incl). Absolute saliency scores of the SNP bases derived from the 5-kb resolution GM12878 model were extracted. Empirical cumulative distributions were derived and tested for significance using a two-sample Kolmogorov–Smirnov test (*R*, ks.test, reshuffled SNP saliency ($n = 6{,}607$) versus eQTL set ($n = 6{,}607$) saliency, alternative hypothesis is 'less').

Deletion screen. Separately, GM12878 DNase-seq and CTCF ChIP–seq peaks were merged if multiple peaks were found within 1.5 kb of each other (bedtools merge -d 1500). Peaks were extended to at least 300 bp. All DNase peaks that overlapped with CTCF peaks were removed. For every remaining CTCF ($n = 45{,}635$) and DNase ($n = 47{,}320$) site, the impact on chromatin interactions on deleting the respective site was predicted by the 5-kb GM12878 model. Chromatin classes were assigned based on overlap with the GenoSTAN chromatin segmentation described above. In addition, $n = 3{,}850$ background sites were selected by shuffling all CTCF and open chromatin sites on chr16, forcing no overlap (bedtools shuffle -chrom -noOverlapping).

5C data. Processed 5C data from Hnisz et al.[31] were downloaded, binned into 5-kb bins and visualized.

SNP sampling. One thousand random SNPs each were sampled from strong CTCF sites, strong promoters and strong enhancers, as classified by the chromatin segmentation procedure described above. For a background set, 1,000 SNPs were sampled from the 400-bp regions flanking these regulatory elements while avoiding

any overlap with other elements. SNPs positions were sampled using bedtools shuffle and variant bases were randomly selected from the three bases not present in hg19 at the respective position.

Statistics and replication. Statistical analysis was performed in R. Statistical tests are described in the relevant subsection of the Methods. NG Capture-C and Tiled-C experiments were performed once, and technical replicates were pooled for maximum read depth (see Reporting Summary for additional details).

Additional software and packages. All neural networks were implemented in python (v.3.5) and tensorflow[49] (developed under v.1.8.0).

Additional tools.

- samtools[50] (v.1.3)
- FastQC (v.0.11.4) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
- Bowtie[51] (v.1.1.2)

Additional R packages.

- cowplot (v.0.6.2, https://github.com/wilkelab/cowplot)
- GenomicRanges[52] (v.1.30.3)
- ggplot2 (ref. [53]) (v.3.1.0)
- RcolorBrewer (v.1.1.1–2, https://cran.r-project.org/web/packages/RColorBrewer/index.html

- rtracklayer[54] (v.1.30.4)
- tidyverse (v.1.3.0) (https://www.tidyverse.org)
- zoo[55] (v.1.8.1)

Additional Python libraries.

- numpy[56] (v.1.16.4)
- h5py (v.2.9.0, http://www.h5py.org)
- pysam (v.0.15.2, https://github.com/pysam-developers/pysam)

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Hi-C data from Rao et al. is available under GSE63525. Chromatin feature data from ENCODE, Roadmap and other publicly available data are listed in detail with accession numbers in Supplementary Table 1. Additional ENCODE data used for chromatin segmentation and visualization are listed with accession numbers in Supplementary Table 5. Tiled-C and NG Capture-C validation data are available under the Gene Expression Omnibus superseries GSE137437. Source data are provided with this paper.

Code availability

All code for training and employing deepC networks as well as trained models are available at https://github.com/rschwess/deepC; all code for training and employing chromatin feature networks is available at https://github.com/rschwess/deepHaem

References

38. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
39. Schwessinger, R. et al. Sasquatch: predicting the impact of regulatory SNPs on transcription factor binding from cell- and tissue-specific DNase footprints. *Genome Res.* **27**, 1730–1742 (2017).
40. Downes, D. J. et al. An integrated platform to systematically identify causal variants and genes for polygenic human traits. Preprint at *bioRxiv* https://doi.org/10.1101/813618 (2019).
41. Telenius, J., Consortium, T. W. & Hughes, J. R. NGseqBasic—a single-command UNIX tool for ATAC-seq, DNaseI-seq, Cut-and-Run, and ChIP–seq data mapping, high-resolution visualisation, and quality control. Preprint at *bioRxiv* https://doi.org/10.1101/393413 (2018).
42. Zhang, Y. et al. Model-based analysis of ChIP–Seq (MACS). *Genome Biol.* **9**, R137 (2008).
43. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Multimed. Tools Appl.* **77**, 10437–10453 (2015).
44. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
45. Telenius, J. M. et al. CaptureCompendium: a comprehensive toolkit for 3C analysis. Preprint at *bioRxiv* http://doi.org/10.1101/2020.02.17.952572 (2020).
46. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
47. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
48. Zacher, B. et al. Accurate promoter and enhancer identification in 127 ENCODE and roadmap epigenomics cell types and tissues by GenoSTAN. *PLoS ONE* **12**, e0169249 (2017).
49. Abadi, M. et al. TensorFlow: a system for large-scale machine learning. In *Proc. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)* 265–284 (2016).
50. Li, H. et al. The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
51. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
52. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
53. Wickham, *H ggplot2: Elegant Graphics for Data Analysis* (Springer, 2009).
54. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).
55. Zeileis, A. & Grothendieck, G. Zoo: S3 infrastructure for regular and irregular time series. *J. Stat. Softw.* **14**, 1–27 (2005).
56. van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).

Acknowledgements

We thank R. Beagrie for help in refining the manuscript. This work was supported by the MRC (grant no. MC_UU_00016/14 to J.R.H.) and the Wellcome Trust via Strategic Award (no. 106130/Z/14/Z to J.R.H.) and Institutional Strategic Support Fund (reference no. 105605/Z/14/Z to J.R.H.). The Wellcome Trust Genomic Medicine and Statistics PhD Program (grant nos. 203728/Z/16/Z to R.S. and 203141/Z/16/Z to R.C.B.). The Stevenson Junior Research Fellowship at University College, Oxford (to A.M.O). G.L. is supported by the Wellcome Trust supporting award (no. 090532/Z/09/Z). Y.W.T. is supported by the European Research Council under the European Union's Seventh Framework Program (grant no. FP7/2007-2013) ERC grant agreement no. 617071. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by the NCI, NHGRI, NHLBI, NIDA, NIMH and NINDS.

Author contributions

R.S., G.L. and J.R.H. conceived the project. R.S., R.C.B., Y.W.T. and G.L designed the neural network architectures. R.S. optimized and trained the neural networks and performed downstream analysis. R.S., M.G., D.D., A.M.O. and J.R.H. designed and evaluated the validation strategy. M.G. performed NG Capture-C experiments. D.D. performed Tiled-C experiments. R.S., A.M.O. and J.T. performed bioinformatic analysis of NG Capture-C and Tiled-C. R.S. performed integrative analysis and prepared the figures. R.S., G.L and J.R.H. wrote the manuscript with inputs from all authors.

Competing interests

The authors declare no competing interests.

Additional information

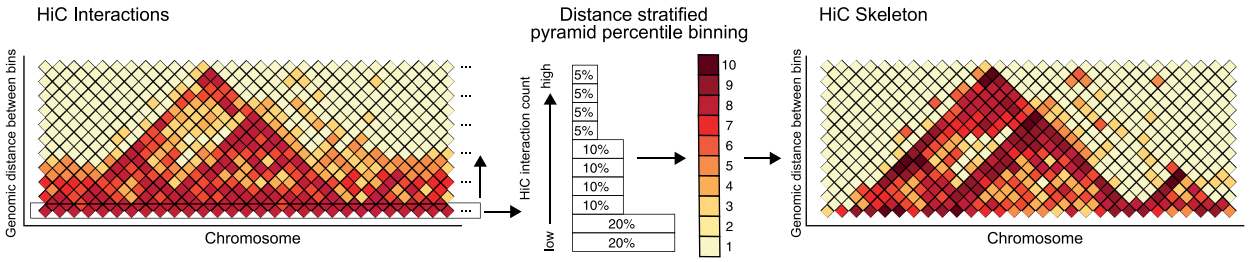
Extended data is available for this paper at https://doi.org/10.1038/s41592-020-0960-3.

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-020-0960-3.

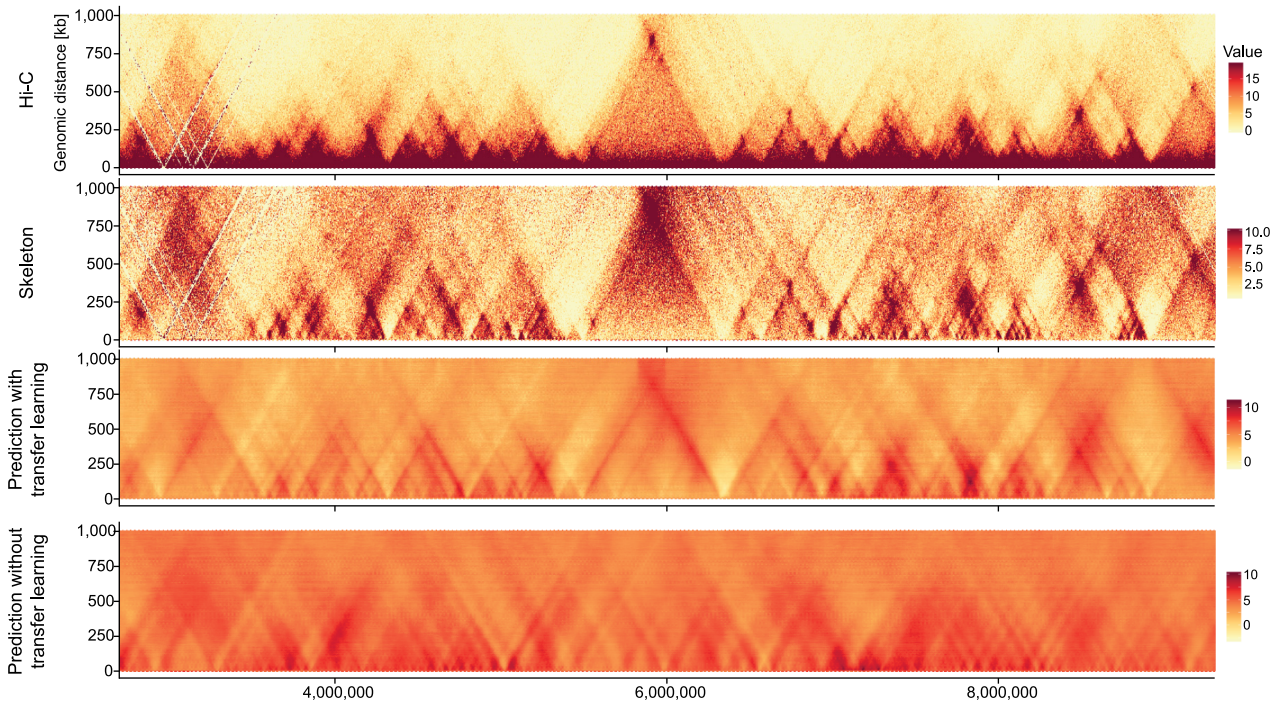
Correspondence and requests for materials should be addressed to G.L. or J.R.H.

Peer review information Lin Tang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

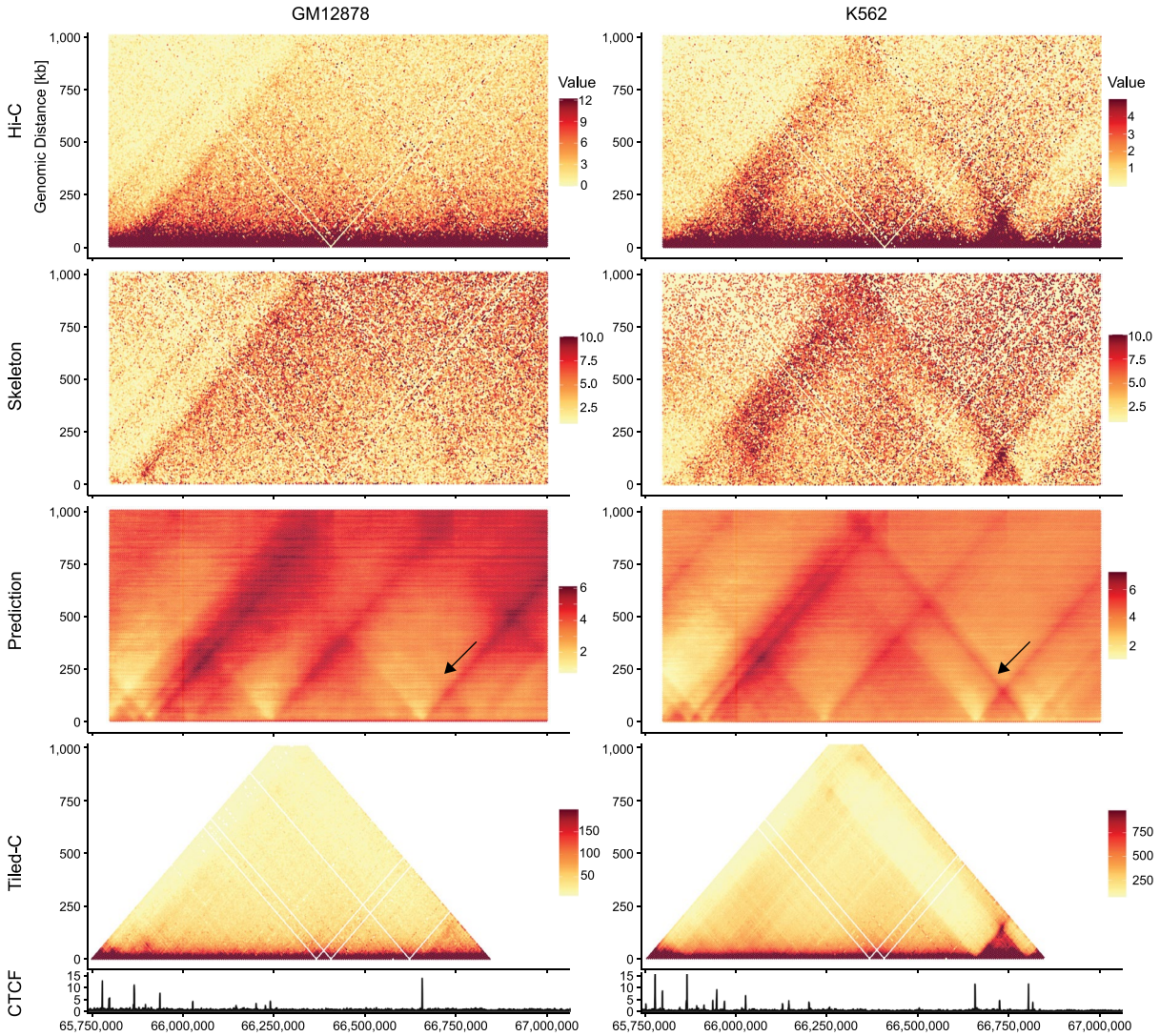
Reprints and permissions information is available at www.nature.com/reprints.



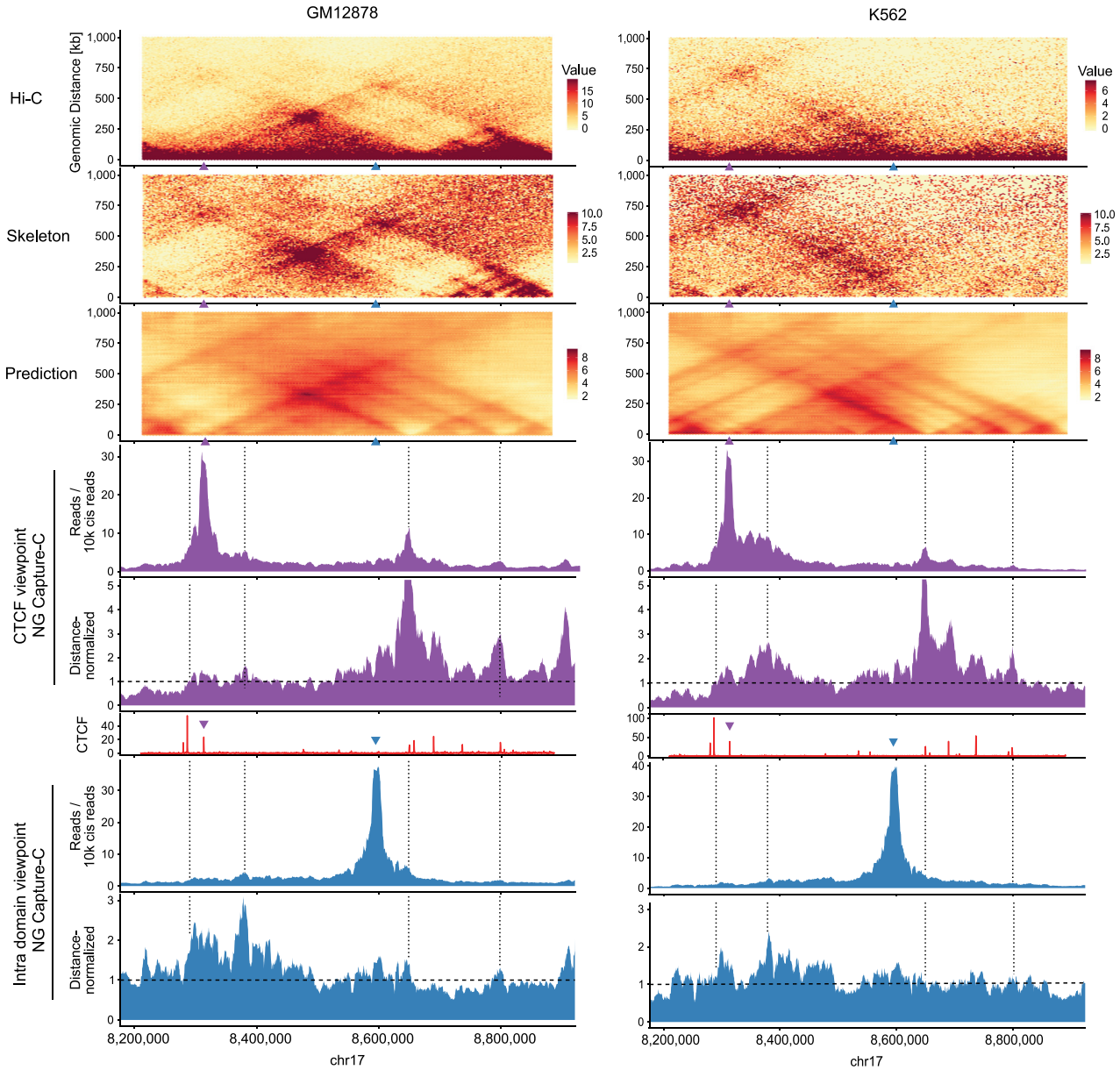
Extended Data Fig. 1 | Percentile normalizing Hi-C data for deep learning. The Hi-C interactions are percentile-binned in a distance-stratified manner. For every genomic distance, in steps equal to the bin size, the Hi-C signal is split into unequal percentiles ranging from 20 % bottom to 5 % top. The percentiles are attributed the values 1 to 10 yielding the Hi-C skeleton. The unequal percentile sizes ensure a finer distinction of the differences at the high Hi-C interaction value range, while minor differences in the low interaction value range are squished. Effectively, this procedure reduces the proximity signal and enhances domains and domain boundaries.



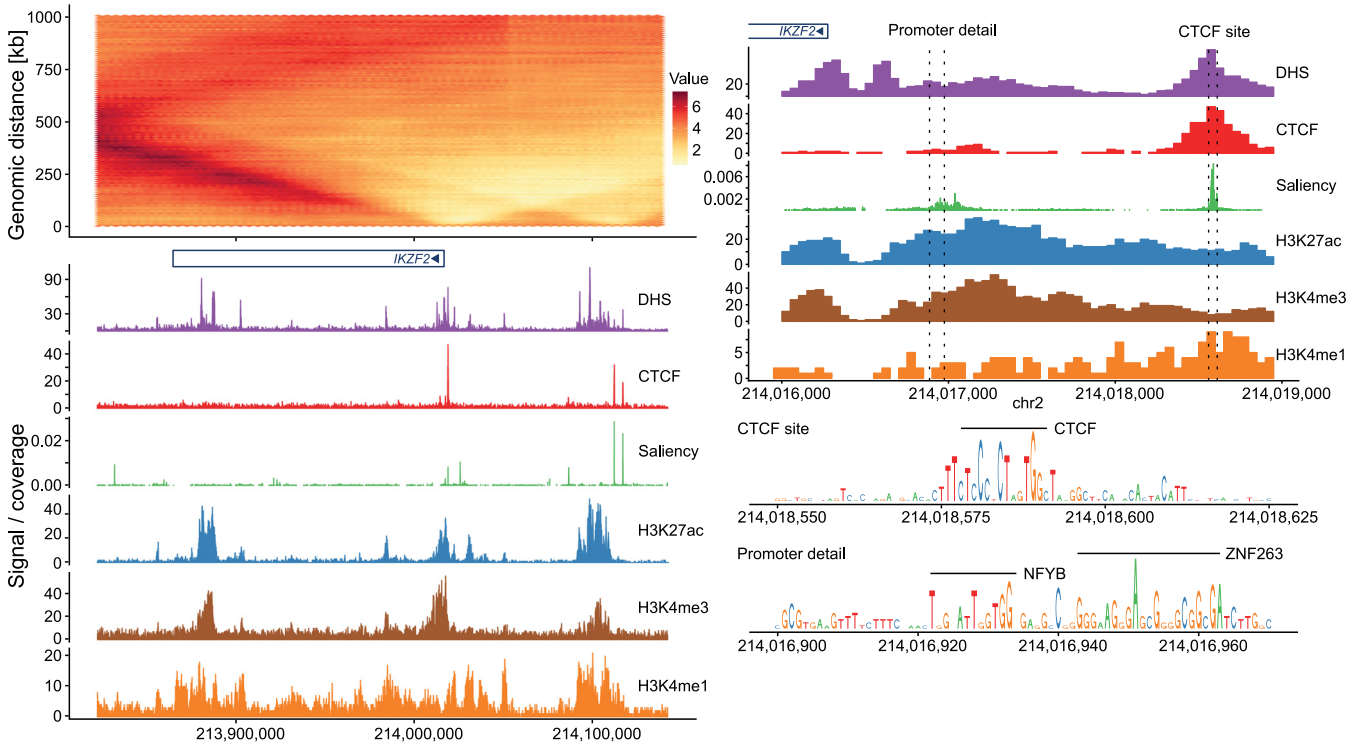
Extended Data Fig. 2 | Comparison of deepC training with and without transfer learning. Training a deepC model with the same architecture but without pre-seeding the lower convolutional layers with the chromatin feature model weights results in the emergence of triangular structures. Their positioning however does not match with the Hi-C structures. In contrast, with pre-seeding the predicted domains overlap well with the Hi-C skeleton.



Extended Data Fig. 3 | Tissue-specific deepC predictions. Shown is a region on chromosome 2 around the *MEIS1* locus. DeepC predicts a small domain with insulation to the upstream regions (black arrow) in a tissue specific manner. The domain is only visible in K562 Hi-C data and matches with tissue-specific CTCF binding. Tiled-C confirms the tissue-specific domain. For contrast, Tiled-C data were bounded between the 5 and 95 percentiles.



Extended Data Fig. 4 | NG Capture-C validation of deepC predictions. a) Example region with overlap of GM12878: Hi-C, skeleton and deepC prediction; NG Capture-C tracks, distance-normalized NG Capture-C tracks and CTCF ChIP-seq track (red). Shown is a CTCF viewpoint (purple triangle) and an intra domain viewpoint (blue triangle) not overlapping with any active elements. Dashed lines in the distance-normalized NG Capture-C tracks indicate the expected interaction value. Dotted black lines highlight deepC prediction details that correspond to boundaries in the NG Capture-C tracks. **b**) K562 data of the same region.



Extended Data Fig. 5 | Mapping important features for genome folding. Shown are GM12878 deepC predictions over the *IKZF2* locus (**a**) on chromosome 2 and focused on the *IKZF2* promoter (**b**). Aligned are DHS as well as ChIP-seq tracks for CTCF and histone modifications. Shown in green is the saliency score which is a proxy for the importance every base has in predicting the chromatin interactions of that region. The saliency score shows sharp peaks overlapping CTCF binding sites and broader peaks overlapping active gene promoters. Resolving the saliency score at base-pair resolution (b) highlights CTCF and general transcription factor binding motifs.

nature research

Corresponding author(s): Jim R. Hughes, Gerton Lunter

Last updated by author(s): Jul 9, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data was collected with commercially available software: Sequencing: NextSeq System Suite (v2)
Data analysis	Genome wide sequencing data analysis: FASTQC(v0.11.4), Bowtie(v1.1.2), samtools (v1.3), deeptools (v2.4.2), MACS (v2.0.10), bedtools (v2.25.0), CCseqBasicS (https://github.com/Hughes-Genome-Group/CCseqBasicS), HiCPro (v2.11.1) Deep Learning related: Python (v3.5), Tensorflow (v1.8), numpy (1.16.4), h5py (v2.9.0), pysam (0.15.2) Integrative analysis and visualization: R(v3.4.4) and packages: cowplot (v0.6.2), GenomicRanges (v1.30.3), GenoSTAN (STAN v2.6.0), ggplot2 (v3.1.0), RColorBrewer (v1.1.1-2), rtracklayer (v1.30.4), tidyverse (v1.3.0), zoo (v1.8.1) Developed Code: All code for training and employing deepC networks as well as trained models are available under: https://github.com/rschwess/deepC All code for training and employing chromatin feature networks is available under: https://github.com/rschwess/deepHaem

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Hi-C data from Rao et al. 2015 is available under GSE63525
NG Caputre-C and Tiled-C data were deposited under GSE137437
The publicly available DNase-seq, ATAC-seq and ChIP-seq data used, their accesion codes and web links to peak calls used are listed in Supplementary Table 1.
Additional ENCODE data used for chromatin segmentation and visualization are listed with accession numbers in Supplementary Table 5

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For NG Capture-C and Tiled-C validation, owing to the sequencing costs and depth of sequencing required to resolve validation data at high resolution, single biological replicates have been split into technical replicates that were merged to enhance data resolution. Sequencing costs were prohibitive for performing replications of NG Capture-C and Tiled-C for this study. Sample size, as in the number of probe locations was estimated with the respective expected sequencing coverage per probe in mind.
Data exclusions	No data have been excluded.
Replication	NG Capture-C and Tiled-C have been performed once and were split into three technical replicates each purely for parallelization. Technical replicates were merged to enhance data resolution.
Randomization	Randomization was not relevant to this study as no treatment and control conditions exist. Where in silico screens did not involve sampling genomic elements exhaustively, elements were sampled randomly (uniform) under constraints (no overlap with other elements).
Blinding	Blinding was not appropriate for data analysis because non-biased computational approaches were employed

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*

Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access and import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used — *Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*

Validation — *Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s) — GM12878 - NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research
K562 - were supplied by the WIMM transgenics facility

Authentication — We declare that the cell lines were not authenticated.

Mycoplasma contamination — Cell lines were obtained from the Coriell Institute and the WIMM transgenics facility which implement regular mycoplasma tests. Upon receiving, the cell lines were not tested again for mycoplasma contamination.

Commonly misidentified lines (See ICLAC register) — No commonly misidentified cell lines were used.

Palaeontology

Specimen provenance — *Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).*

Specimen deposition — *Indicate where the specimens have been deposited to permit free access by other researchers.*

Dating methods — *If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals — *For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.*

Wild animals — *Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*

Field-collected samples — *For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*

Ethics oversight — *Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics — *Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic*

Population characteristics	*information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s) — *Specify: functional, structural, diffusion, perfusion.*

Field strength — *Specify in Tesla*

Sequence & imaging parameters — *Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*

Area of acquisition — *State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software — *Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*

Normalization — *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*

Normalization template — *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*

Noise and artifact removal — *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring — *Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings — *Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*

Effect(s) tested — *Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016) — *Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

Correction — *Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity — *Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*

Graph analysis — *Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*

Multivariate modeling and predictive analysis — *Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*