

Polymer physics predicts the effects of structural variants on chromatin architecture

Simona Bianco ^{1,9}, Darío G. Lupiáñez ^{2,3,4,8,9}, Andrea M. Chiariello ^{1,9}, Carlo Annunziatella^{1,9}, Katerina Kraft^{2,3}, Robert Schöpflin⁵, Lars Wittler⁶, Guillaume Andrey ², Martin Vingron ⁵, Ana Pombo ⁷, Stefan Mundlos ^{2,3,4*} and Mario Nicodemi ^{1*}

Structural variants (SVs) can result in changes in gene expression due to abnormal chromatin folding and cause disease. However, the prediction of such effects remains a challenge. Here we present a polymer-physics-based approach (PRISMR) to model 3D chromatin folding and to predict enhancer-promoter contacts. PRISMR predicts higher-order chromatin structure from genome-wide chromosome conformation capture (Hi-C) data. Using the *EPHA4* locus as a model, the effects of pathogenic SVs are predicted *in silico* and compared to Hi-C data generated from mouse limb buds and patient-derived fibroblasts. PRISMR deconvolves the folding complexity of the *EPHA4* locus and identifies SV-induced ectopic contacts and alterations of 3D genome organization in homozygous or heterozygous states. We show that SVs can reconfigure topologically associating domains, thereby producing extensive rewiring of regulatory interactions and causing disease by gene misexpression. PRISMR can be used to predict interactions *in silico*, thereby providing a tool for analyzing the disease-causing potential of SVs.

Technology-based approaches for the quantification of chromatin contacts have shown that mammalian genomes are folded in a highly controlled manner and that the resulting 3D-configuration directly influences gene regulation¹. Hi-C, a genome-wide variant of chromosome conformation capture (3C) technologies², has shown that mammalian genomes are organized in topologically associating domains (TADs), large genomic regions that display a high degree of interaction and that are largely conserved across species, cell types, and tissue types^{3–6}. TADs are separated by boundaries that constrain interactions between enhancers and their target genes. The 3D folding of the genome and especially the organization of TADs can be disrupted by genomic rearrangements, such as deletions, duplications, or inversions, collectively called structural variants (SVs)^{7–10}. SVs can result in a rewiring of enhancer–promoter contacts, gene misexpression, and disease. However, it is currently difficult to predict such ectopic interactions without performing extensive 3C studies in cells or tissues carrying the rearranged chromosomes.

Polymer models have been successfully employed to dissect the 3D organization of chromosomes. In one strategy, chromatin is described as beads on a chain, and their folding is fitted to represent an average 3D structure that matches Hi-C interaction frequencies^{11–13}.

3D chromatin conformations can also be derived using models from polymer physics and previous knowledge of chromatin factors, notably CTCF and other transcription factors^{14–22}. Here we focus on the strings and binders switch (SBS) polymer model¹⁹, which has been previously shown to recapitulate Hi-C and FISH data to a high degree^{14,17}. The SBS model considers the interactions between chromatin filaments and cognate molecular binders. A chromatin filament is represented as a self-avoiding string of beads (Fig. 1a). The string contains inert (gray) beads, which do not interact with other factors apart from steric hindrance, and bridging beads, which are binding sites for cognate binding molecules that can form loops.

Based on the SBS polymer model, we developed a simulated annealing Monte Carlo optimization procedure, named PRISMR (polymer-based recursive statistical inference method), to infer the minimal factors that shape chromatin folding and its equilibrium 3D structure under the laws of physics, without a priori assumptions and with no additional or tunable parameters. PRISMR scans through the space of all polymer models to find the minimum of a cost function that takes into account the distance between the input Hi-C matrix and the analogous contact matrix derived by polymer thermodynamics for the given model, and an additional Bayesian term to reduce overfitting by penalizing overestimation of binding site number and type (see Methods and Supplementary Fig. 1). The output of PRISMR at the cost-function convergence is the optimal polymer model of the genomic region of interest, with the minimum number and type of required binding sites, to reproduce the experimental Hi-C matrix (Fig. 1b and see Methods). Although PRISMR models are derived from Hi-C pairwise contacts, they can be used to derive any further aspect of folding (such as the ensemble of 3D conformations that the given locus assumes, its higher-order contacts, or the physical distances of genes and regulatory regions) and to predict the effect of rearrangements (Fig. 1c).

To test the predictive value of PRISMR, we chose the *EPHA4* locus, a key developmental region associated with different types of limb malformations⁷. In a previous study, we showed by circular chromosome conformation capture sequencing (4C-seq) that deletions, inversions, and duplications cause distinct phenotypes (brachydactyly, syndactyly, polydactyly) by altering the chromatin organization of the locus, thereby causing rewiring of enhancer–promoter contacts and gene misexpression⁷. To study the

¹Dipartimento di Fisica, Università di Napoli Federico II, and INFN Napoli, Complesso di Monte Sant'Angelo, Naples, Italy. ²Max Planck Institute for Molecular Genetics, RG Development and Disease, Berlin, Germany. ³Institute for Medical and Human Genetics, Charité – Universitätsmedizin Berlin, Berlin, Germany. ⁴Berlin-Brandenburg Center for Regenerative Therapies (BCRT), Charité – Universitätsmedizin Berlin, Berlin, Germany. ⁵Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany. ⁶Department Developmental Genetics, Max Planck Institute for Molecular Genetics, Berlin, Germany. ⁷Epigenetic Regulation and Chromatin Architecture Group, Berlin Institute for Medical Systems Biology, Max-Delbrück Center for Molecular Medicine, Berlin-Buch, Germany. ⁸Present address: Epigenetics and Sex Development Group, Berlin Institute for Medical Systems Biology, Max-Delbrück Center for Molecular Medicine, Berlin-Buch, Germany. ⁹These authors contributed equally: Simona Bianco, Darío G. Lupiáñez, Andrea M. Chiariello, Carlo Annunziatella. *e-mail: mundlos@molgen.mpg.de; mario.nicodemi@na.infn.it

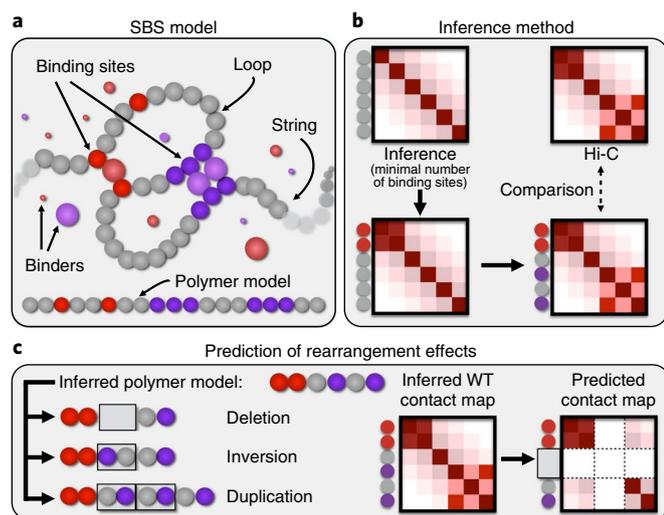


Fig. 1 | The PRISMR method: inference of molecular binders shaping chromatin folding. **a**, In the SBS model, chromatin is represented as a chain of beads interacting with molecular binders, in which the different types of binding sites (and their cognate bridging molecules) are visualized in different colors. **b**, PRISMR samples the thermodynamics ensemble of the conformations of a given SBS model to derive its contact matrix from polymer physics. By a simulated annealing Monte Carlo algorithm (SA), it iteratively finds the model that best describes the input contact matrix. **c**, By informing the model inferred from wild-type (WT) data with a given rearrangement, the effects of genomic mutations on folding can be predicted from only polymer physics without any fitting parameters.

3D configuration of the entire locus, we performed capture Hi-C (cHi-C)¹⁰ on embryonic day (E) 11.5 mouse limb buds and human skin fibroblasts. We also analyzed previously published Hi-C datasets from murine CH12-LX and human IMR90 cells²³. Regardless of the cell or tissue type or the species, we observed a subdivision of the locus in one large TAD, containing only *EPHA4*, in a smaller TAD, containing *PAX3* and *SGPP2*, and in a gene-dense region on the centromeric side, showing no clear TAD structure (Fig. 2a). Differences were apparent within the *EPHA4* TAD that likely reflect cell- and tissue-specific patterns of interaction and gene regulation⁵.

Toward developing predictive models of the architecture of the *EPHA4* locus across different cell types, we applied PRISMR to all four Hi-C datasets. The derived contact matrices were similar to the original Hi-C data, not only recapitulating the global TAD conformation of the locus, but also capturing cell-specific intra-TAD organization (Fig. 2a and Supplementary Note): the Pearson correlation, r , and distance-corrected correlation coefficient, r' , range up to $r = 0.95$ and $r' = 0.69$ (Fig. 2a, Methods, Supplementary Fig. 2, and Supplementary Table 1). PRISMR identifies the different binding domains of the locus, i.e., the sets of binding sites of the same type (i.e., color) that determine the folding patterns (Fig. 2b). In CH12-LX cells, for instance, the model predicted 21 different binding domains. They are all required to meet the accuracy cutoff in the description of Hi-C data, and six of them had main structural roles (see Methods). To check the robustness of our simulated annealing Monte Carlo procedure, we investigated whether there were differences between the best minima achieved by different runs of the PRISMR simulations. We found that independent runs gave correlations comparable with those of the experimental Hi-C data and similar corresponding polymer models ($P < 1 \times 10^{-250}$; Supplementary Note); additionally, the optimal model was robust to changes of the number of binding domain types (Supplementary Note). The model binding domains can be broadly divided into two categories (Fig. 2b

and see Methods): one category includes domains largely overlapping with annotated TADs (type-I), and the other includes domains extending over several TADs or even over the whole analyzed region (type-II). Comparing the genomic position of the model binding sites with ENCODE epigenetics data available for CH12-LX cells²⁴ showed that a single binding type (a ‘color’) did not correlate with a single molecular factor (see Methods and Supplementary Fig. 3).

To explore the roles of various factors to the folding patterns detected by PRISMR simulations of cHi-C data, we considered the architectural protein CTCF, a DNA-binding transcription factor thought to facilitate the formation of chromatin loops^{20,25,26}. Notably, some binding site types identified by PRISMR correlated with CTCF (Supplementary Fig. 3). As PRISMR does not exploit prior information on binding sites and factors, to test its reach we considered a variant of the model in which we included previous knowledge about the location of CTCF binding sites in the locus, which were added to interact with an additional type of binder that bridges opposed (forward or reverse) CTCF sites (see Methods). In limb tissue E11.5 cells, for instance, this variant (named ‘PRISMR + CTCF’) had correlations with Hi-C data similar to those of the initial model: it improved the visualization of the large *Epha4* TAD, mainly by strengthening the loop anchors characteristic for CTCF-associated loops, but it also resulted in additional contacts in the neighboring gene-dense region that were not present in the original cHi-C data (Fig. 2c and Supplementary Fig. 4). Conversely, a model with only CTCF (named ‘CTCF-only’) can describe some of the loops seen in the data²⁰, but poorly captured the global contact patterns of the *Epha4* locus (Supplementary Fig. 4), resulting in a lower correlation coefficient ($r' = 0.05$). These results indicate that other factors besides CTCF were important in chromatin folding and TAD configuration (see Methods) and that our approach can recapitulate most of the interactions of Hi-C data without a priori information on binding factors. Nevertheless, such information can be added to adapt and improve model predictions.

To test whether PRISMR can predict the effects of homozygous SVs on chromatin folding, we investigated three previously reported variants⁷ at the *Epha4* mouse locus: a deletion (*DelB*) encompassing a large part of the *Epha4* TAD and the telomeric TAD boundary (associated with brachydactyly due to misexpression of *Pax3*), a slightly smaller deletion (*DelBs*) that leaves the TAD boundary intact (no misexpression, no phenotype), and a balanced 1.1-Mb inversion (*InvF*) that causes misexpression of *Wnt6*. We implemented these mutations in polymer models of the wild-type CH12-LX and E11.5 limbs cells inferred by PRISMR and re-ran the ensemble of folding conformations to derive an average locus contact matrix. For E11.5 limb tissue, we tested both the PRISMR model and the PRISMR + CTCF version with the addition of CTCF sites (Fig. 3a and Supplementary Figs. 5a, 6a, and 7). To identify the regions of statistically significant ectopic interactions in each predicted rearrangement, we subtracted each mutant matrix from the wild-type matrix (Fig. 3b and Supplementary Figs. 5b, 6b, and 8). Although the studied locus is populated by more than 40 genes, our matrices predicted that only certain regions, containing a limited number of genes, would display changes in the interaction profiles. For example, in the larger deletion (*DelB*) including the *Epha4* TAD boundary, we identified new contacts that predicted fusion between the remaining *Epha4* and *Pax3* TADs, thus facilitating the association between *Epha4* enhancers and *Pax3* that results in ectopic gene activation and a pathogenic phenotype⁷. Ectopic contacts between the same regions were also predicted in the smaller deletion (*DelBs*), which leaves the *Epha4*–*Pax3* boundary intact. However, virtual 4C analysis derived from our predictions showed that the enhancers–*Pax3* ectopic interaction was diminished, consistent with the absence of *Pax3* activation in these mutants (Fig. 3c and Supplementary Figs. 5c and 6c). The inversion (*InvF*) was predicted to result in a rearrangement of the genomic content of the two

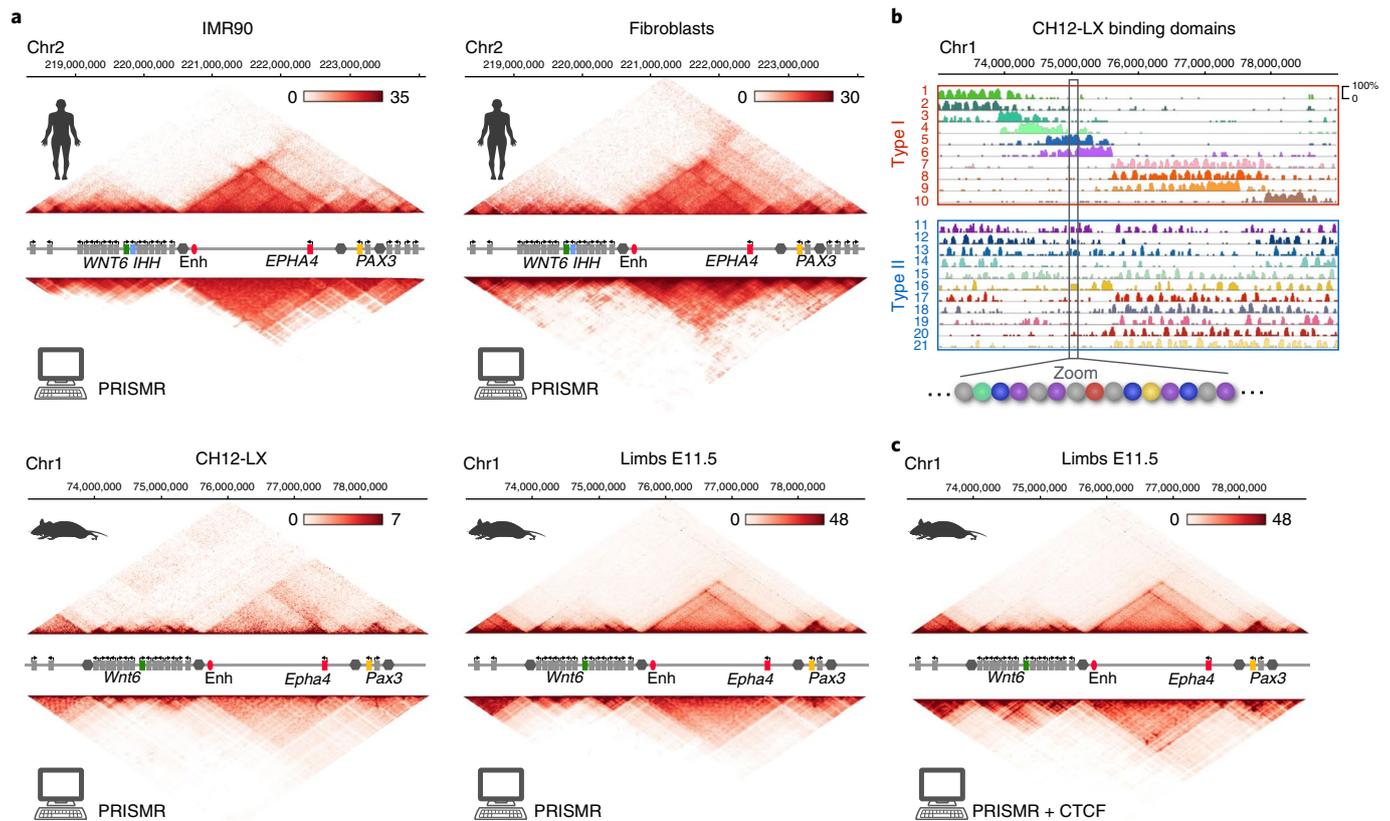


Fig. 2 | PRISMR recapitulates 3D conformation at the *EPHA4* locus. **a**, Published Hi-C data²³ (left) and our capture Hi-C data (right; $n = 1$ with an internal control comparing 4 different experiments; see Methods) compare well with the contact matrices derived by PRISMR. Their Pearson correlations, r , and distance-corrected Pearson correlation coefficients, r' , are comparatively high: $r = 0.92$, $r' = 0.64$ in IMR90; $r = 0.93$, $r' = 0.69$ in human fibroblasts; $r = 0.91$, $r' = 0.56$ in CH12-LX; and $r = 0.94$, $r' = 0.60$ in limb tissue E11.5. The genomic region, with genes (rectangles), TAD boundaries (hexagons), and enhancers (ovals), is shown schematically. Relevant genomic elements are highlighted with colors and names. Enh, enhancer; chr, chromosome. **b**, PRISMR also identifies the different binding sites (see Fig. 1a) along the locus that shape its 3D structure. The binding sites of the same type (color) form the shown different binding domains. In CH12-LX cells, for example, there are 21 statistically significant (one-sided Wilcoxon's rank-sum test; $P < 1.1 \times 10^{-7}$; see Supplementary Note) different binding domains. The plots represent their abundance (as percentages) along the genomic sequence. Type-I domains are spatially restricted, whereas type-II are ubiquitous, covering the whole region. Their overlapping genomic positions produce the observed complex interaction patterns. **c**, The contact matrix from a variant of the PRISMR model with addition of CTCF binding sites (PRISMR + CTCF model) has a similar Pearson correlation with experimental data ($n = 1$ with an internal control comparing 4 different experiments; see Methods); for example, in limb E11.5 tissue we find $r = 0.95$, $r' = 0.52$.

adjacent TADs with interaction hotspots between *Epha4* enhancers and a gene-dense region (three genes affected) that would be consistent with the ectopic *Wnt6* activation reported previously. We also observed ectopic interactions between a region near the centromeric breakpoint containing the *Wnt10a* gene and the remaining *Epha4* TAD. Therefore, PRISMR identified specific and localized regions of ectopic interactions across the entire locus as a consequence of genomic rearrangements, identifying a small number of genes whose regulation might be directly affected.

As a next step, we tested the accuracy of our predictions by comparison against a new experimental cHi-C dataset from mouse limb buds carrying homozygous mutations. Our dataset showed the same regions of ectopic interaction and displayed a noticeably high agreement with PRISMR predictions, not only across the entire locus but also when the regions of ectopic interaction were compared (Fig. 3, Supplementary Tables 1 and 2, and Supplementary Figs. 5, 6, and 8). Our results confirmed that the larger deletion in *DelB* mutant led to a fusion of the *Epha4* and *Pax3* TADs, not occurring in the smaller *DelBs* mutation, in which the TAD boundary remains intact (Supplementary Fig. 9). In the inversion, ectopic contacts were observed between *Wnt6* and the *Epha4* enhancer

region, which facilitated *Wnt6* activation as previously observed in vivo⁷, and between a region at the centromeric breakpoint and the entire *Epha4* TAD. Notably, the observed ectopic interaction was interrupted by the *Epha4* centromeric boundary, which, although inverted, appeared to retain its functionality (Supplementary Fig. 9). Hence, deletions and inversions that include boundary elements can result in fusions or reorganization of TADs, respectively.

Finally, we wanted to test the potential of PRISMR to predict the effects of heterozygous SVs on chromatin organization as they are commonly observed in human patient samples. A PRISMR polymer model of the *EPHA4* locus, inferred from healthy control human fibroblast cHi-C data (Fig. 2a), was employed to predict the effects of SVs on chromatin contact matrices (Fig. 4 and Supplementary Fig. 10). To test the model predictions, we used fibroblasts obtained from human patients to perform cHi-C (Fig. 4). We analyzed a 1.6-Mb deletion associated with brachydactyly (similar to mouse *DelB*), a 900-kb duplication (*DupP*) associated with polydactyly and *IHH* activation, and a 1.4-Mb duplication (*DupF*) associated with syndactyly and *WNT6* activation⁷. Subtraction maps identified the precise regions and intensity of significant ectopic interactions (Fig. 4b, Methods, and Supplementary Fig. 11). In the brachydactyly-associated deletion,

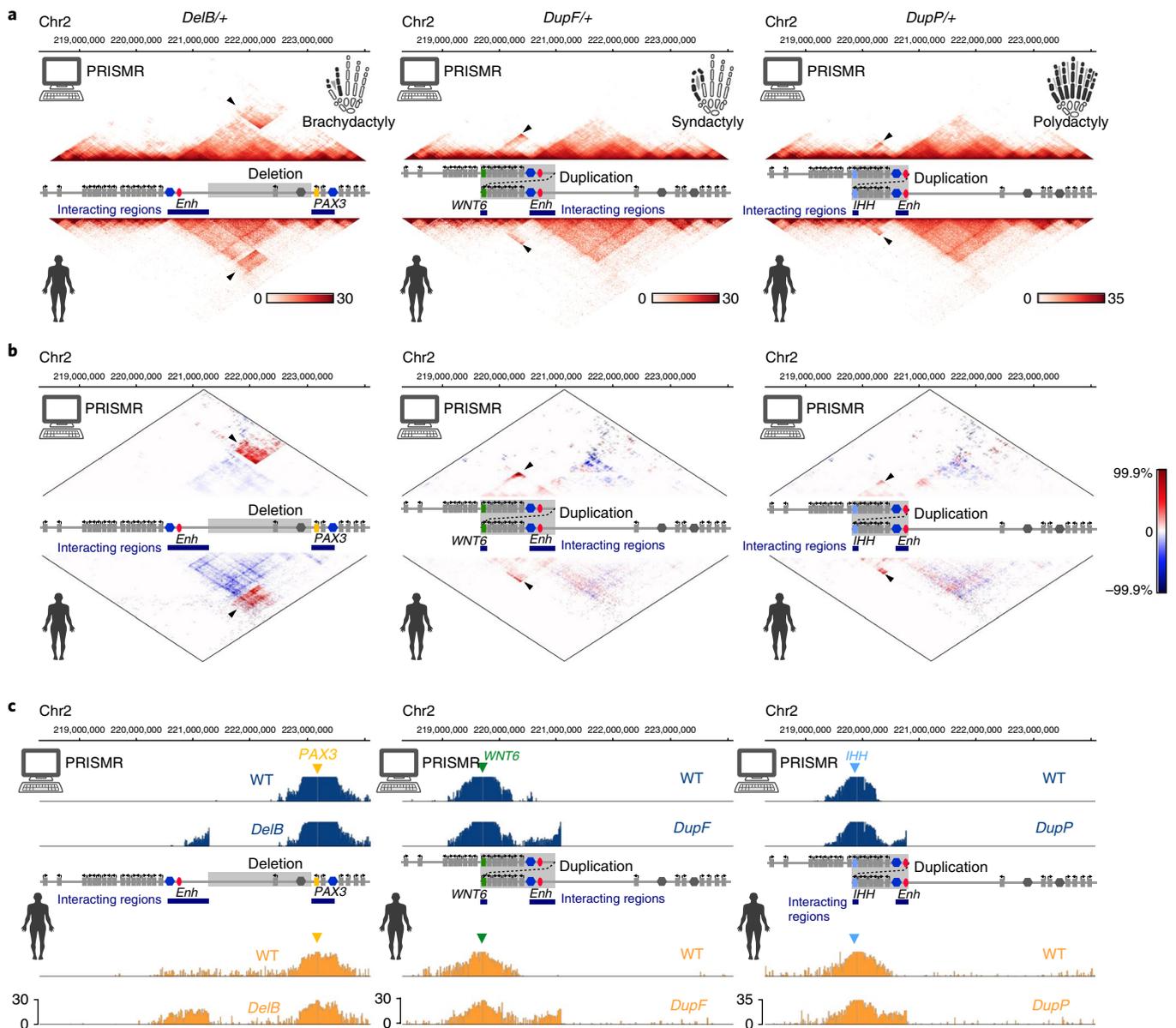


Fig. 4 | PRISMR predicts the effects of human heterozygous structural variants on chromatin architecture. a, Contact matrices from model predictions derived from WT data (top) and cHi-C experiments in mutation carrying cultured human skin fibroblasts (bottom; $n=1$ with an internal control comparing 4 different experiments; see Methods). Schematic genomic region displays genes (rectangles), TAD boundaries (hexagons), enhancers (ovals), and corresponding structural variant. Human phenotypes associated with the rearrangement are indicated on right. *DelB/+*: PRISMR predicts the chromatin effects of a 1.6-Mb heterozygous deletion (Pearson correlation $r=0.93$, distance-corrected Pearson correlation $r'=0.61$). Increased interaction is detected between the remaining *EPHA4* and *PAX3* TADs (arrowhead and blue bars), resulting in *PAX3* misexpression and brachydactyly. *DupF/+*: heterozygous 1.4-Mb duplication ($r=0.88$, $r'=0.52$). Increased interaction is detected between *EPHA4* enhancer cluster and *WNT6* regions. *DupP/+*: heterozygous 900-bp duplication ($r=0.90$, $r'=0.56$). Increased interaction is detected between *EPHA4* enhancer cluster and *IHH* regions. **b**, Subtraction maps produced (using a healthy control and patients) from predictions and cHi-C data ($n=1$ with an internal control comparing 4 different experiments; see Methods). Above, threshold gain of interaction is displayed in red and loss in blue (absolute differences > 2 s.d.; see Methods). Ectopic interactions between *EPHA4* TAD and genomic regions are indicated (arrowheads and blue bars). **c**, Virtual 4C plots derived from predictions and cHi-C data from the viewpoint on the respective phenotype-causing gene. *DelB/+*: note increased interaction of *PAX3* promoter with remaining *EPHA4* TAD, including *EPHA4* enhancer cluster in both, prediction and experimental data. *DupF/+*: note increased interaction of *WNT6* promoter with the *EPHA4* enhancer cluster. *DupP/+*: note increased interaction of *IHH* promoter with the *EPHA4* enhancer cluster.

and even in complex genomic regions with high gene density. PRISMR predictions can be used to identify regions of ectopic interaction that can then be scanned for their content, i.e., the presence of genes and enhancers that could interact. Furthermore, our results indicate that PRISMR can be used in cases where affected tissues or equivalent cell

types are not available. Recent advances in high-throughput sequencing have boosted the identification of SVs^{27–29}. In this scenario, polymer modeling by PRISMR emerges as a valid approach for predicting pathogenic effects, facilitating the interpretation and diagnosis of this type of genomic rearrangement.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at <https://doi.org/10.1038/s41588-018-0098-8>.

Received: 4 January 2017; Accepted: 27 February 2018;

Published online: 16 April 2018

References

- Fraser, J., Williamson, I., Bickmore, W. A. & Dostie, J. An overview of genome organization and how we got there: from FISH to Hi-C. *Microbiol. Mol. Biol. Rev.* **79**, 347–372 (2015).
- Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
- Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
- Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
- Phillips-Cremins, J. E. et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281–1295 (2013).
- Fraser, J. et al. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Syst. Biol.* **11**, 852 (2015).
- Lupiáñez, D. G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
- Hnisz, D. et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016).
- Lupiáñez, D. G., Spielmann, M. & Mundlos, S. Breaking TADs: how alterations of chromatin domains result in disease. *Trends Genet.* **32**, 225–237 (2016).
- Franke, M. et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).
- Duan, Z. et al. A three-dimensional model of the yeast genome. *Nature* **465**, 363–367 (2010).
- Lesne, A., Riposo, J., Roger, P., Cournac, A. & Mozziconacci, J. 3D genome reconstruction from chromosomal contacts. *Nat. Methods* **11**, 1141–1143 (2014).
- Serra, F. et al. Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS Lett.* **589**(20 Pt A), 2987–2995 (2015).
- Barbieri, M. et al. Complexity of chromatin folding is captured by the strings and binders switch model. *Proc. Natl. Acad. Sci. USA* **109**, 16173–16178 (2012).
- Bohn, M. & Heermann, D. W. Diffusion-driven looping provides a consistent framework for chromatin organization. *PLoS One* **5**, e12218 (2010).
- Brackley, C. A., Taylor, S., Papantonis, A., Cook, P. R. & Marenduzzo, D. Nonspecific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization. *Proc. Natl. Acad. Sci. USA* **110**, E3605–E3611 (2013).
- Chiariello, A. M., Annunziatella, C., Bianco, S., Esposito, A. & Nicodemi, M. Polymer physics of chromosome large-scale 3D organisation. *Sci. Rep.* **6**, 29775 (2016).
- Giorgetti, L. et al. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* **157**, 950–963 (2014).
- Nicodemi, M. & Prisco, A. Thermodynamic pathways to genome spatial organization in the cell nucleus. *Biophys. J.* **96**, 2168–2177 (2009).
- Sanborn, A. L. et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. USA* **112**, E6456–E6465 (2015).
- Scialdone, A., Cataudella, I., Barbieri, M., Prisco, A. & Nicodemi, M. Conformation regulation of the X chromosome inactivation center: a model. *PLoS Comput. Biol.* **7**, e1002229 (2011).
- Fudenberg, G. et al. Formation of chromosomal domains by loop extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
- Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
- ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Guo, Y. et al. CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* **162**, 900–910 (2015).
- Nora, E. P. et al. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930–944.e22 (2017).
- Gilissen, C. et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).
- Hehir-Kwa, J. Y. et al. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat. Commun.* **7**, 12989 (2016).
- Newman, S., Hermetz, K. E., Weckselblatt, B. & Rudd, M. K. Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. *Am. J. Hum. Genet.* **96**, 208–220 (2015).

Acknowledgements

We thank the sequencing core, transgenic unit, and animal facilities of the Max Planck Institute for Molecular Genetics for technical assistance. This work is supported by grants from the Deutsche Forschungsgemeinschaft (DFG) and the Max Planck Foundation (MPF) to S.M. and D.G.L., the Berlin Institute of Health (BIH) to S.M. and A.P.; by CINECA ISCRA Grant HP10CYFSP5 and HP10CRTY8P, computer resources at INFN and *Scope* at the University of Naples (M.N.), and the Einstein BIH Fellowship Award to M.N.

Author contributions

M.N. and S.M. designed the project. S.M., D.G.L., and M.V. devised the cHi-C experiments. S.B., A.M.C., C.A., and M.N. developed the modeling part; S.B., A.M.C., and C.A. ran the computer simulations and performed their analyses. L.W. derived mouse homozygous lines and performed tetraploid aggregations. D.G.L., K.K., and G.A. performed cHi-C experiments, and R.S. performed bioinformatic analyses. M.N., S.M., D.G.L., S.B., A.P., A.M.C., and C.A. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41588-018-0098-8>.

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

Correspondence and requests for materials should be addressed to S.M. or M.N.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

The studied loci. The murine *Epha4* locus discussed in this paper is a 6-Mb long region around the *Epha4* gene (coordinates mm9 chr1: 73,000,000–79,000,000). We employed in situ Hi-C data from CH12-LX cells²³ and our own cHi-C data in E11.5 mouse limb buds, at a 10-kb resolution. The studied human *EPHA4* locus in skin fibroblasts is 5.77 Mb long (coordinates hg19 chr2: 218,320,000–224,090,000); in that system we produced our own cHi-C data at 10-kb resolution. We also studied the *EPHA4* locus in human IMR90 cells, where we used previously published in situ Hi-C data at 10-kb resolution²³; the considered locus is 8 Mb long (coordinates hg19 chr2: 217,000,000–225,000,000).

The strings and binders switch (SBS) model. The string and binders switch (SBS) polymer model of chromatin folding quantifies the biological scenario where molecules, such as transcription factors, loop DNA by bridging distal cognate binding sites^{14,19}. In the SBS model, a chromatin filament is represented as a self-avoiding (SAW) polymer chain of N beads. The beads can be bound and bridged by molecular binders; each type of binder is defined by its specific, cognate type of binders (Fig. 1a). In our notation, n is the total number of different types of binding sites (represented with different colors in the figures). There are also inert sites along the chain, i.e., beads (represented in gray) that do not interact with any binder apart from steric hindrance. Each type of binder has a molar concentration c and their binding energy to their cognate beads is E_{int} . As well described in polymer physics and in previous studies^{14,17,19}, the above model exhibits a coil–globule phase transition, from an open conformation in the SAW universality class found at low c or E_{int} to a different conformational class corresponding to a globule, compact state³⁰. All details on the model and its simulations are given below and in the Supplementary Note.

The PRISMR algorithm. Our PRISMR algorithm aims to find the minimal number and types (colors) of binding sites in a SBS polymer chain, and their position along the chain, that best reproduces an input contact matrix of a given chromosomal locus, by folding solely according to the laws of physics (Fig. 1b). PRISMR is based on a standard simulated annealing Monte Carlo (SA) optimization procedure^{31,32} that minimizes the distance between the predicted polymer model and the input contact matrix, under a Bayesian weighting factor to avoid overfitting. The procedure involves five main iterative steps: (i) consider a polymer model with a given arrangement of binding sites; (ii) derive a thermodynamics ensemble of its 3D equilibrium conformations; (iii) compute its contact matrix; (iv) compare it with the input Hi-C data; and (v) change the polymer model accordingly and repeat until convergence. Those steps are described in full details below and in the Supplementary Note.

The sequence of the genomic region to be modeled is divided into L windows according to the genomic resolution of the considered Hi-C experiment. As a single DNA window could include many binding sites, we consider a polymer chain that is r times longer in order to include such details. The optimal value of r is returned as an intermediate output by PRISMR, as described below. An SBS polymer model is identified by the arrangement of binding sites of different types along the chain of beads, i.e., by the set $\{c_i\}$ of its color variables $c_i = 0, 1, \dots, n$ (0 corresponding to gray, inert sites), where the index $i = 1, \dots, N$ labels the i th bead. The output of PRISMR is then the best, minimal arrangement $\{c_i\}_m$ of beads along the chain to describe the input contact matrix.

Specifically, the iterative SA procedure of the algorithm minimizes a cost function, H , including two terms to fit well the data and to avoid overfitting. The first one, H_{in} , is the distance between the input, $C_{exp}(i,j)$, and model, $C(i,j)$, contact matrices; the second, H_{out} , is a Bayesian term (a chemical potential in statistical mechanics) that penalizes the addition of binding sites (see Supplementary Note for details). The weight of the Bayesian term is given by a positive factor λ : the larger λ is, the more the addition of binding sites is penalized (if $\lambda = 0$, there is no penalty).

Given n , r , and λ , PRISMR samples the huge space of the possible arrangements, $\{c_i\}$, of the binding sites along the polymer chain (having $(n+1)^N$ elements) to search for the minimum, $\{c_i\}_m$, of the cost function H by use of a standard SA iterative procedure^{31,32}. Starting from a random initial assignment of the binding sites along the chain, the color of a randomly chosen bead is changed at random, the average contact matrix of the new polymer is computed, and the cost function evaluated until convergence (see Supplementary Note). The SA procedure is repeated, from many initial conditions, with different values of n , r , and λ to find the minimal required number of different colors n^* , the minimal required value of r , r^* , and the minimal number of binding sites, i.e., the maximum allowed value of λ , λ^* , to explain the input data within a given accuracy (Supplementary Fig. 1 and Supplementary Note). Since r^* is systematically found to be smaller than n^* in all the cases studied here, for simplicity we consider the safe option where $r^* = n^*$. The optimal arrangement of binding sites, $\{c_i\}^*$, obtained with n^* and λ^* , is the final output of the PRISMR algorithm (as shown in Fig. 2b, in log scale).

To derive our polymer models of the *Epha4* locus, we applied the above procedure to four different input Hi-C datasets, all binned at 10-kb resolution and KR (Knights and Ruiz) normalized³³. In the studied murine *Epha4* locus, the algorithm returns $n^* = 21$ and $\lambda^* = 1.0$ in both the published in situ Hi-C data of CH12-LX cells²³ and in our limb tissue cHi-C data. In the considered human

EPHA4 locus, PRISMR finds $n^* = 16$ and $\lambda^* = 1.0$ in the published in situ Hi-C data in human IMR90 cells²³, and $n^* = 24$ and $\lambda^* = 1.0$ in the cHi-C data in human fibroblast produced in this study. To simplify the notation, n^* and r^* are renamed n and r in the Results and Discussions sections of this paper. All details about the PRISMR algorithm, its SA procedure, robustness and convergence, can be found below and in the Supplementary Note.

A computationally demanding step of PRISMR is the iterative calculation of the equilibrium thermodynamics average contact frequency, $C(i,j)$, for the sites of a given polymer model, during each iteration of the SA procedure. That can be achieved, for instance, by molecular dynamics (MD) computer simulations; however, these may require huge computational efforts. To speed up the computation of $C(i,j)$, we implemented a mean-field approximation, an approach typical of statistical mechanics³⁴ (see Supplementary Note). To test our approximation, we compare its results against full-scale MD simulations of the optimal model $\{c_i\}^*$ found by the SA procedure. We find that the contact matrices obtained by MD have a Pearson correlation with those derived under the mean-field scheme ranging from $r = 0.91$ to $r = 0.95$ across the studied cases (Supplementary Note and Supplementary Fig. 12). Unless otherwise stated, we always report the full MD results throughout the paper.

Analyses of the model binding domains. We ran different successful tests of the statistical significance and robustness of the optimal polymer models, $\{c_i\}^*$, identified by PRISMR (Supplementary Note). In particular, analyses were performed to quantify the relevance of each binding domain (color) to fold the structure of the *Epha4* locus, in all the studied cases. To this aim, we measured how the distance-corrected Pearson correlation, r' (see below), between model and experimental Hi-C data, was affected when individual colors were withdrawn from the model (Supplementary Note). For example, we found that in mouse CH12-LX cells, no color is redundant, as r' was reduced in a range from 8% up to 15% of its original value. In particular, there is a subgroup of 6 main binding domains having an impact higher than the average (12%). Analysis of the models in other murine and human cell types, in WT cells and mutants, gave similar results (Supplementary Note).

Each of the different n^* types of binding domains is specified by the coordinates (in bases) of their binding sites along the locus. To quantify the similarity between pairs of binding domains (colors), we measured their genomic overlap, which turns out to be far from random (Supplementary Note). Analogously, the assignment of the binding domains to class I–II (see Results) is based on their overlaps with the locus TADs⁵: specifically, a binding domain is of type I if it strongly overlaps (above median) only one or two consecutive TADs, else it is of type II (see Supplementary Note). Finally, to get some insights into the molecular nature of the inferred different types of binding sites, we correlated their genomic positions with epigenetic features available for CH12-LX cells in the ENCODE database²⁴ (Supplementary Table 3). Specifically, for each binding domain we computed the Pearson correlation coefficient between the locations of its binding sites and the peaks of a chromatin mark, in the corresponding 10-kb bins. Next, to check the statistical significance of such correlations, we computed the distribution of correlations with chromatin marks of a random control model obtained by bootstrapping our binding domains. The correlation with a chromatin feature is considered significant if it is above the 95th or below the 5th percentile of the distribution of correlations in the random model. The resulting matrix of correlations is shown in Supplementary Fig. 3. We found that our binding domains correlated with distinct combinations of chromatin marks, rather than matching each with a single molecular factor.

Molecular dynamics simulations of the *EPHA4* locus and its mutations. In our SBS model, the polymer beads and the binders are subject to a Brownian motion, described by the Langevin equation, numerically integrated with the Verlet algorithm in the LAMMPS package^{35,36}. We set the diameters of polymer beads and of binders to σ , and their masses to m . The interaction potential used in our MD simulations has been developed in classical computer studies of polymer physics³⁷. Two consecutive beads along the polymer are bound by a finitely extensible spring FENE potential³⁷. To model excluded volume effects, the Weeks–Chandler–Andersen potential is used³⁷. Finally, the binders interact with their cognate polymer beads via a short-ranged, attractive Lennard–Jones potential¹⁶. All details about the MD simulations are provided in the Supplementary Note.

The PRISMR polymer chain model of the WT *Epha4* locus in murine CH12-LX cells (Fig. 2a), derived from published in situ Hi-C data at 10-kb resolution²³, is made of $N = 12,600$ beads. To model structural variants, we start from the WT polymer, implement the mutation under consideration, and perform new MD simulations. For instance, we removed the part of the polymer corresponding to the DNA deletion in real cells, or we inverted the binding sites corresponding to the experimental inversion. The resulting polymer models for the *DelBs* and *DelB* deletions and the *InvF* inversion are made respectively of $N = 9,513$, $N = 9,093$, and $N = 12,600$ beads (Supplementary Fig. 5a). The models derived from our murine E11.5 cell cHi-C data, our CHiC data of human skin fibroblast cells, and published human IMR90 cell data²³ are analogous (Figs. 2a and 4a, Supplementary Fig. 6a, and Supplementary Note).

Polymer models including prior knowledge of CTCF binding sites. We considered a variant of our PRISMR model (PRISMR + CTCF) in which forward/reverse CTCF binding sites are included; they are supposed to interact with an additional type of binder that can only bridge oppositely directed CTCF sites. We use peak-called CTCF ChIP-seq data³⁸ and apply a standard motif finding analysis (using the FIMO tool in the MEME Suite online software version 4.12.0; see URLs³⁹) to obtain the best-matching peak in a 10-kb window and its orientation (Supplementary Note). The CTCF motif⁴⁰ was obtained from the JASPAR database (see URLs). Additionally, to try to dissect the specific effects of CTCF alone, we also considered a simpler polymer model (CTCF-only) including only the above described CTCF binding sites/binders, without the binding domains identified by the PRISMR model (Supplementary Note). The PRISMR + CTCF model does not significantly improve the comparison against cHi-C data with respect to PRISMR alone (Supplementary Table 1), whereas the CTCF-only model fits some of the loops well, but has a poor distance-corrected Pearson correlation with cHi-C data, $r' = 0.05$ (Supplementary Fig. 4).

Contact frequency matrices. The average pairwise contact frequency matrices of the polymer models are derived as discussed in the literature^{14,17}. The distance r_{ij} between any pair of beads i and j of the same type is measured in a given equilibrium polymer configuration; if r_{ij} is lower than a set threshold, i and j are considered in contact. Averages are then taken over independent, single-molecule 3D structures (Supplementary Note).

To identify significant ectopic interactions, the normalized subtraction matrices (Figs. 3b and 4b and Supplementary Figs. 5b, 6b, 8, and 11), are computed. First, the matrix corresponding to the mutation is multiplied by a factor to equalize, in the regions not affected by the mutation, the reads count to the WT case, and then the WT matrix is subtracted from it. Next, to account for the distance bias, such a matrix is normalized by dividing each subdiagonal by the average WT reads count at that genomic distance. Significant ectopic interactions correspond to absolute differences > 2 s.d. (Supplementary Note).

To better highlight ectopic interactions, we produced virtual 4C plots from the viewpoint of the phenotype causative genes in each mutant in mouse (Fig. 3c and Supplementary Figs. 5c and 6c) and in human (Fig. 4c). Virtual 4C plots are obtained by plotting the column in the contact matrix corresponding to the considered viewpoint. To have a fair comparison between WT and mutation, we first normalized the WT matrix by equalizing the number of its reads to the total reads in the mutation, as described above.

To better compare experimental and model predicted contact matrices, we also measured the distance corrected Pearson correlation, r' , i.e., the correlation after the effect of genomic distance is subtracted from a contact map (Supplementary Fig. 2). Specifically, we subtracted from each diagonal of the contact matrices (experimental and predicted) their average contact frequency at that genomic distance, and then calculated the Pearson correlation coefficient. The control case, matrices with bootstrapped diagonals, returns values of $r' < 0.05$ (Supplementary Note). Analogous comparisons of experimental replicates (see below and Supplementary Note) show that PRISMR versus cHiC correlations are comparable to those among replicates.

SureSelect design. We designed the SureSelect enrichment RNA probes over the genomic intervals chr1: 71,000,001–81,000,000 (mm9) and chr2: 218,314,001–224,093,000 (hg19) using SureDesign from Agilent. Probes were not specifically designed in proximity of DpnII sites but over the entire region, with a coverage of 85%.

Mouse samples. Mouse embryonic stem (ES) cell lines were established from *DelB* and *DelBs* homozygous blastocysts using N2B27 Medium supplemented with FGF/Erk and Gsk3 pathway inhibitors (2i) and LIF⁴¹. Both mouse strains (*DelB* and *DelBs*) were previously maintained by crossing them with C57BL/6J mice. The derived ES cells, in addition to the preexisting wild type G4 and *InvF* homozygous cell lines (129/Sv × C57BL/6 F1 hybrid)⁴², were tested for mycoplasma contamination using Mycoalert detection kit (Lonza, catalog number LT07-118) and Mycoalert assay control set (Lonza, catalog number LT07-518). These ES cells lines were used to generate homozygous embryonic litters using tetraploid complementation⁴². All animal procedures were conducted as approved by the local authorities (LAGeSo Berlin) under license numbers G0368/08 and G0247/13.

Human materials. Skin biopsies were obtained from the patients and controls by standard procedures. Fibroblasts were cultured in DMEM (Lonza) supplemented with 10% FCS (Gibco), 1% L-glutamine (Lonza), and 1% penicillin/streptomycin (Lonza). Cells were tested for mycoplasma contamination as described above. Written informed consent was obtained from all individuals studied to participate in this study. This study was approved by the Charité Universitätsmedizin Berlin ethics committee and complies with all relevant ethical regulations.

cHi-C. Libraries were prepared from mouse E11.5 mutant and wild-type distal fore- and hindlimbs and from human cultured skin fibroblasts from patients and healthy controls, as previously described⁴³, i.e. crosslinking, cell lysis, DpnII

digestion, ligation, and de-crosslinking. DNA was then sheared with a Covaris sonicator (duty cycle: 10%, intensity: 5, cycles per burst: 200, time: 6 cycles of 60s each, set mode: frequency sweeping, temperature: 4° to 7 °C). Adaptors were ligated to the sonicated DNA, which was then amplified according to Agilent instructions. The library was then hybridized to SureSelect RNA probes, indexed for sequencing following Agilent instructions and sequenced using a 100-bp paired-end mode.

Capture Hi-C experiments were performed as singletons. As an internal control, we compared the results from all four experiments for regions outside of the region of interest (for mouse: chr1: 71,000,001–74,820,000 and chr1: 78,080,001–81,000,000; for human: chr2: 218,320,001–219,730,000 and chr2: 223,030,001–224,090,000). The cHi-C maps of the internal control were highly correlated between the four samples (mouse samples Spearman r : WT/*DelB* = 0.99; WT/*DelBs* = 0.99; WT/*InvF* = 0.99; *DelB*/*DelBs* = 0.99; *DelB*/*InvF* = 0.99; *DelBs*/*InvF* = 0.99; human samples Spearman r : WT/*DelB* = 0.99; WT/*DupF* = 0.98; WT/*DupP* = 0.98; *DelB*/*DupF* = 0.99; *DelB*/*DupP* = 0.98; *DupF*/*DupP* = 0.98), confirming the high reproducibility of the methodology.

cHi-C data processing. DNA libraries were sequenced paired-end. Fastq files were processed with the HiCUP pipeline v0.5.8⁴⁴ (nofill: 1, format: Sanger, without di-tag length restriction) to perform the mapping as well as to filter for valid and unique di-tags. The pipeline was set up with Bowtie2 v2.2.6⁴⁵ and with reference genomes hg19 and mm9, respectively. Additionally, we created customized genome files containing the mutation for the SVs investigated in mouse (Supplementary Table 4). BAM files produced by the HiCUP pipeline were transferred into an intermediate text file format suited for Juicebox command line tools^{23,46}, which were used for binning and normalization (KR normalization³⁵). The genomic interval enriched by the cHi-C protocol covers 5.77Mb in hg19 and 10Mb in mm9, respectively, leading to three different regimes in the contact map: (i) enriched vs. enriched, (ii) enriched vs. nonenriched, and (iii) nonenriched vs. nonenriched. For binning and normalization, only regime (i) was considered. Therefore, di-tags were filtered for the enriched genomic interval and all coordinates were shifted by the start of the enriched interval (Supplementary Table 4). The chromosome size file for Juicebox command line tools was customized such that it only contained the size of the enriched region. The minimum MAPQ for the import was set to 30. After binning and normalization, all coordinates were shifted back to their original values. The applied KR (Knights and Ruiz) normalization³⁵ is a matrix-balancing algorithm that ensures equal sums for all rows and columns of the map. The underlying assumption for this type of normalization is that all loci should have an equal representation in the map. When SVs, such as duplications, are mapped to a WT reference genome, as well as at the map borders, the map deviates strongly from this assumption and matrix balancing corrects disproportionately. Therefore, only raw count maps were used to create subtraction maps between WT and SVs (see section “Contact frequency matrices”). In all cases, cHi-C data confirmed the presence of the corresponding SVs on each mouse and human studied sample.

Statistics. No statistical methods were used to predetermine sample size. There was no exclusion/inclusion of samples or animals in the analysis, no randomization of experiments, and investigators were not blinded during experiments and outcome assessment. No unique materials were used in this study.

All the statistical tests we used are described in the paper and all details provided in the Methods and Supplementary Note. We used Pearson correlation coefficients to compare experimental and model derived contact matrices. Pearson correlations were also used to match the model-inferred binding domains with epigenetic features from ENCODE²⁴. One-tailed Wilcoxon's rank-sum tests were applied to check the significance of the binding domains.

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

Code availability. We used publicly available computer codes (LAMMPS) for our molecular dynamics simulations. Custom codes used to generate results reported in the manuscript can be made available upon request. All details of the algorithms are illustrated above and in previous publications cited herein.

Data availability. Data have been deposited at GEO under accession code GSE92294.

URLs. FIMO: <http://meme-suite.org/tools/fimo>; JASPAR database, <http://jaspar.binf.ku.dk/>; Gene Expression Omnibus, <https://www.ncbi.nlm.nih.gov/geo/>.

References

- de Gennes, P. G. *Scaling Concepts in Polymer Physics* (Cornell Univ. Press, Ithaca, NY, 1979).
- Kirkpatrick, S., Gelatt, C. D. Jr & Vecchi, M. P. Optimization by simulated annealing. *Science* **220**, 671–680 (1983).
- Salamon, P., Sibani, P. & Frost, R. *Facts, Conjectures, and Improvements for Simulated Annealing* (SIAM, Philadelphia, 2002).

33. Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* **33**, 1029–1047 (2013).
34. Parisi, G. *Statistical Field Theory* (Westview Press, New York, 1998).
35. Rosa, A. & Everaers, R. Structure and dynamics of interphase chromosomes. *PLOS Comput. Biol.* **4**, e1000153 (2008).
36. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117**, 1–19 (1995).
37. Kremer, K. & Grest, G. S. Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. *J. Chem. Phys.* **92**, 5057 (1990).
38. Andrey, G. et al. Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding. *Genome Res.* **27**, 223–233 (2017).
39. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
40. Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
41. Nagy, K. N. J. in *Advanced Protocols for Animal Transgenesis* (eds. Pease, S. & Saunders, T. L.) 431–455 (Springer, Berlin, 2011).
42. Artus, J. & Hadjantonakis, A. K. Generation of chimeras by aggregation of embryonic stem cells with diploid or tetraploid mouse embryos. *Methods Mol. Biol.* **693**, 37–56 (2011).
43. Hagège, H. et al. Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nat. Protoc.* **2**, 1722–1733 (2007).
44. Wingett, S. et al. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res.* **4**, 1310 (2015).
45. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
46. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see [Reporting Life Sciences Research](#). For further information on Nature Research policies, including our [data availability policy](#), see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study. For [final submission](#): please carefully check your responses for accuracy; you will not be able to make changes later.

▶ Experimental design

1. Sample size

Describe how sample size was determined.

Capture Hi-C experiments are extremely reproducible among biological replicates when samples are processed appropriately. For this reason, experiments for each condition are performed only once (n=1). To control that experimental procedures are reproducible, samples from different experiments were compared. In this comparison, control genomic regions that are unaffected by structural variants (and therefore should show the same interaction profile) are compared. Such comparison shows that samples are highly reproducible from a technical point of view (see Methods, "CHiC" subsection, 2nd paragraph).

2. Data exclusions

Describe any data exclusions.

There was no exclusion/inclusion of samples or animals in the analysis.

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

To control that Capture Hi-C experimental procedures are reproducible, samples from different experiments were compared. In this comparison, control genomic regions that are unaffected by structural variants (and therefore should show the same interaction profile) are compared. Such comparison showed that samples are highly reproducible from a technical point of view (see Methods, "CHiC" subsection, 2nd paragraph). All experimental attempts were successful

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

There was no randomization of experiments.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Investigators were not blinded during experiments and outcome assessment.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
- A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- A statement indicating how many times each experiment was replicated
- 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 any assumptions or corrections, such as an adjustment for multiple comparisons
- Test values indicating whether an effect is present
Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.
- A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
- Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on [statistics for biologists](#) for further resources and guidance.

► Software

Policy information about [availability of computer code](#)

7. Software

Describe the software used to analyze the data in this study.

We used a standard motif finding analysis (using the FIMO tool in the MEME Suite online software version 4.12.0).

Fastq files were processed with the HiCUP pipeline v0.5.843 (Nofill:1, Format: Sanger, without di-tag length restriction) performing the mapping as well the filtering for valid and unique di-tags. The pipeline was set up with Bowtie2 v2.2.644 .

We used publicly available computer codes (LAMMPS) for our Molecular Dynamics simulations. Custom codes used to generate results reported in the manuscript will be made available upon request. All details of the algorithms are illustrated in the Methods section and in previous publications cited therein

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). [Nature Methods guidance for providing algorithms and software for publication](#) provides further information on this topic.

► Materials and reagents

Policy information about [availability of materials](#)

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

No unique materials were used in this study.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used in this study.

10. Eukaryotic cell lines

- a. State the source of each eukaryotic cell line used.

Mouse ES cell lines were established from DelB and DelBs homozygous blastocysts using N2B27 Medium supplemented with FGF/Erk and Gsk3 pathway inhibitors (2i) and LIF. Both mouse strains (DelB and DelBs) were previously maintained by crossing them with C57BL/6/J mice. The derived ES cells, in addition to the preexisting wild type G4 (129/Sv × C57BL/6 F1 hybrid; kindly provided by Andras Nagy's lab) and InvF homozygous cell lines (129/Sv × C57BL/6 F1 hybrid, generated through CRISPR genome editing from G4 cells; reported in Lupiáñez et al., 2015) were used to generate homozygous embryonic litters using tetraploid complementation.

Skin biopsies and were obtained from the patients and controls by standard procedures. Fibroblasts were cultured in DMEM (Lonza) supplemented with 10% fetal calf serum (Gibco), 1% l-glutamine (Lonza) and 1% penicillin/streptomycin (Lonza).

- b. Describe the method of cell line authentication used.

In all cases, cHi-C data confirmed the presence of the corresponding SVs on each mouse and human studied sample.

- c. Report whether the cell lines were tested for mycoplasma contamination.

All cell lines tested negative for mycoplasma contamination.

- d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No commonly misidentified cell lines were used.

▶ Animals and human research participants

Policy information about [studies involving animals](#); when reporting animal research, follow the [ARRIVE guidelines](#)

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.

Mouse ES cell lines (129/Sv × C57BL/6 F1 hybrid background and 129/Sv × C57BL/6 F1 hybrid background backcrossed with C57BL/6) were used to generate homozygous embryonic litters using tetraploid complementation. Resulting embryos of both sexes were analyzed at E 11.5

Policy information about [studies involving human research participants](#)

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Sex and gender from patients and controls are heterogeneous but not relevant for the results of the study as Capture Hi-C experiments are not influenced by those parameters