



CHESS enables quantitative comparison of chromatin contact data and automatic feature extraction

Silvia Galan^{1,2,8}, Nick Machnik^{1,3,8}, Kai Kruse¹, Noelia Díaz¹, Marc A. Marti-Renom^{2,4,5,6} and Juan M. Vaquerizas^{1,7}✉

Dynamic changes in the three-dimensional (3D) organization of chromatin are associated with central biological processes, such as transcription, replication and development. Therefore, the comprehensive identification and quantification of these changes is fundamental to understanding of evolutionary and regulatory mechanisms. Here, we present Comparison of Hi-C Experiments using Structural Similarity (CHESS), an algorithm for the comparison of chromatin contact maps and automatic differential feature extraction. We demonstrate the robustness of CHESS to experimental variability and showcase its biological applications on (1) interspecies comparisons of syntenic regions in human and mouse models; (2) intraspecies identification of conformational changes in Zelda-depleted *Drosophila* embryos; (3) patient-specific aberrant chromatin conformation in a diffuse large B-cell lymphoma sample; and (4) the systematic identification of chromatin contact differences in high-resolution Capture-C data. In summary, CHESS is a computationally efficient method for the comparison and classification of changes in chromatin contact data.

Eukaryotic genomes follow similar global organizational principles: a multilayer, hierarchical organization into domains, with specific 3D interactions between individual genomic regions¹. Local chromatin conformation, however, can be variable across species^{2–6}, developmental stages^{7–9} and cell types^{10,11}, and can change dynamically with transcription¹², during replication¹³ and cell division¹⁴, among other contexts. Mutations affecting nuclear architecture have been shown to cause misregulation of gene expression leading to developmental disorders and disease (reviewed in Spielmann et al.¹⁵ and Krijger et al.¹⁶). Therefore, it is important to elucidate the relationships among nuclear architecture, evolution and fundamental biological processes.

Existing approaches to identify changes in the 3D conformation of genomic regions have relied partially on visual analysis of differences, such as a side-by-side evaluation of Hi-C matrices^{17,18} or fold change maps¹⁹. While visual comparisons can highlight specific changes in Hi-C matrices, results are often difficult to quantify and, by nature, cannot be automated to compare large numbers of matrices. More quantitative approaches have been developed. One class of tools focuses on the assessment of the degree of similarity or reproducibility between full chromatin contact matrices/datasets and does not allow for the identification of regions with particularly strong similarities or differences^{20–24}. Another focuses on the comparison of specific features, such as topologically associating domains (TADs)^{25,26}, or loops^{10,27}, which limits the discovery of differences to the specific feature analyzed. A third class of tools aims to find single pairs of bins with significantly differential interactions^{28–32} without providing any information about the specific type of structural feature that changes.

Therefore, there is a need for methods that allow a systematic comparison of the 3D conformation of genomic regions, that is, at the same time quantitative, able to identify and classify a range of structural variations and corresponding well to the visual perception of differences.

In this technical report, we describe CHESS, an algorithm that can be used to robustly identify and classify specific similarities or differences and features in chromatin contact data using a feature-free approach. CHESS applies the concept of the structural similarity index (SSIM) widely used in image analysis^{33,34} to chromatin contact matrices, assigning a structural similarity score and an associated *P* value to pairs of genomic regions. Next, CHESS uses image processing approaches to automatically extract 3D chromatin conformation features, such as TADs, stripes or loops. We first demonstrate the robustness of CHESS scores by evaluating the method on artificially generated and real Hi-C matrices of different sizes, sequencing depths and varying levels of noise. Then, we highlight the usefulness of CHESS in different real-world applications: (1) genome-wide comparisons of syntenic regions between human and mouse models; (2) the detection of conformational changes in *Drosophila melanogaster* on knockdown of the transcription factor Zelda during early embryonic development; (3) the detection of 3D chromatin conformation changes in the B cells of a patient with diffuse large B-cell lymphoma (DLBCL); and (4) the automatic detection and classification of subtle changes in chromatin conformation from genome editing experiments. Overall, our results demonstrate that CHESS can be successfully applied to diverse chromatin contact datasets to quantitatively determine structural differences between them.

¹Max Planck Institute for Molecular Biomedicine, Münster, Germany. ²National Centre for Genomic Analysis, Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain. ³Institute of Science and Technology Austria, Klosterneuburg, Austria. ⁴Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain. ⁵Pompeu Fabra University, Barcelona, Spain. ⁶Catalan Institution for Research and Advanced Studies, Barcelona, Spain. ⁷Medical Research Council London Institute of Medical Sciences, Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London, UK. ⁸These authors contributed equally: Silvia Galan, Nick Machnik. ✉e-mail: jmv@mpi-muenster.mpg.de

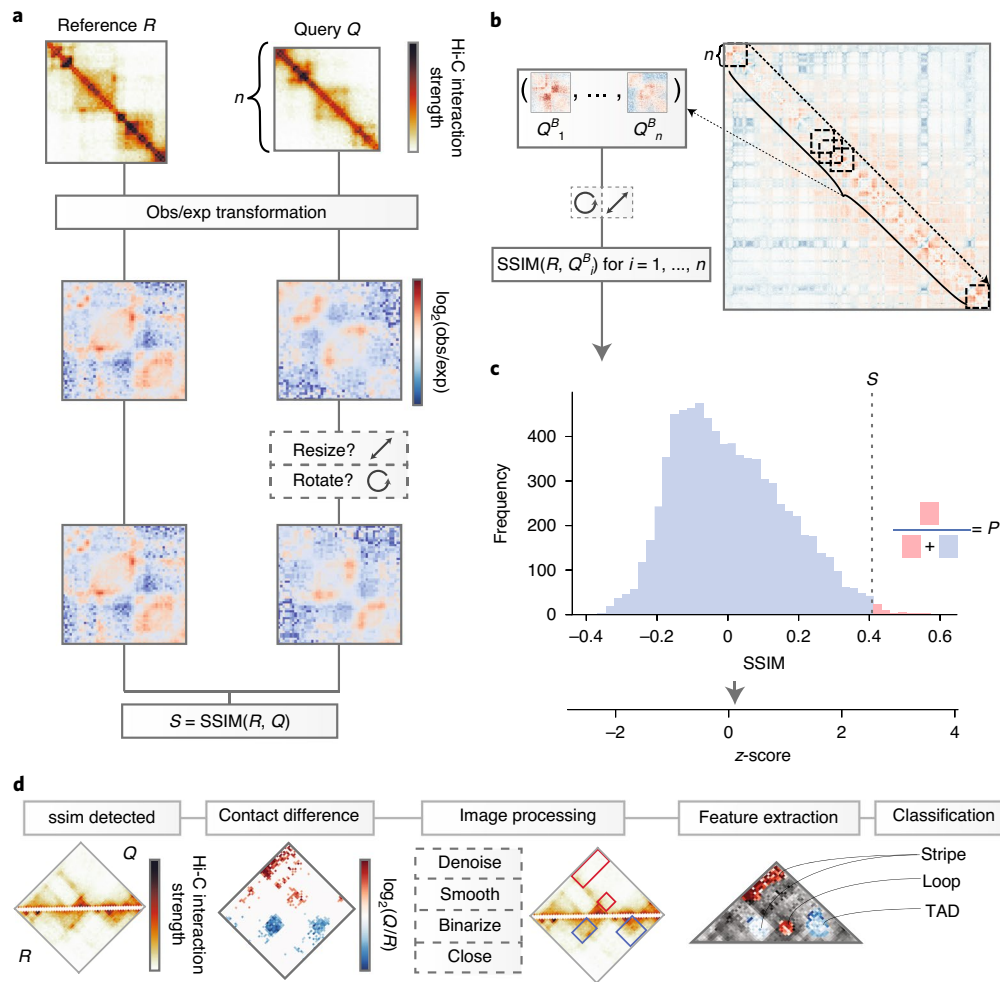


Fig. 1 | CHES overview and examples. **a**, CHES workflow, showing the obs/exp transformation, size/orientation adjustments and structural similarity score S calculation on two example matrices R and Q . **b**, Example of a background model for empirical calculation of z-scores and P values. Specifically, similarity scores are calculated for each $n \times n$ matrix Q^B_i at every position i along the diagonal of the whole chromosome matrix. **c**, The distribution of similarity scores for Q^B_i can be used to calculate P values and z-scores for S (Methods). **d**, Feature extraction workflow, showing the contact difference map between R and Q matrices identified by CHES and the list of image filters applied. The specific differential gained and lost structures are highlighted in red and blue boxes, respectively. Finally, all the features are classified according to their structural pattern, such as TADs, loops and stripes.

Results

Overview of the CHES algorithm. CHES assesses the degree of similarity between any pair of normalized chromatin contact matrices, such as those produced from Hi-C, or tiled Capture-C experiments (Fig. 1a). It provides a measure for quantifying matrix similarity, which allows the identification of particularly similar or dissimilar pairs. There are very few limitations on the origin of input matrices. These can be different regions in the same genome; the same genomic region across two experimental conditions; developmental time points; and even regions from different species. CHES uses the SSIM, a widely used metric for matrix similarity (Methods).

To calculate the similarity between a reference (R) and query (Q) matrix, their entries (that is, contact pairs or pixels in the maps) are first divided by the expected contact intensity at the respective distance. This observed/expected (obs/exp) transformation is necessary to remove the distance dependency of pairwise contact probabilities characteristic of Hi-C matrices³⁵, which is otherwise the dominant topological characteristic (Extended Data Fig. 1 and Methods). CHES then scales the matrices to equal size and calculates the SSIM between R and Q . The SSIM score (S) is a single

value, combining brightness, contrast and structure differences between two matrices (Extended Data Fig. 2). Brightness is calculated as the mean of the signal intensity. Contrast is calculated as the variance in signal. The structure term is calculated as the correlation between signal values of two matrices. S is then defined as a weighted product of these three components, which are scaled such that S ranges between -1 (inversion) and 1 (identity), where $S=0$ indicates no similarity. We performed an in-depth evaluation of the three components of S (Extended Data Fig. 2), which showed that the main contributor to similarity when applied to Hi-C matrices is the product of contrast and structure terms.

One application of CHES is to identify regions with strong changes in chromatin conformation between two conditions genome-wide. For this, S can be used directly to quantify changes in chromatin contacts within windows of a given size across the genome: for identical matrices $S=1$ and the lower the score the larger the change (Methods). Therefore, CHES can be used to rank genomic regions by the amount of chromatin changes within them.

An additional application of CHES is to assess whether contact matrices originating from different genomic regions, or different genomes, are similar. An appropriate null model can be used to

test whether the similarity measured by S is statistically significant. For example, a region R containing just a single TAD might obtain a high score when compared to a particular query Q , which also contains a single TAD. However, the same is true for any region with a single TAD, which is why the similarity of R and Q is not particularly informative in this instance. The score for the comparison of R versus Q should then be assigned a low significance. Conversely, when two highly complex regions with many structural features are assigned a strong similarity score S , it is unlikely to find an equally similar region in the genome by chance and the comparison is given high statistical significance. To compute a suitable null model, CHESS compares the reference matrix R to all other regions of the same size across the genome (referred to as Q_i^B in Fig. 1b). The distribution of scores from the null model is then used to calculate a z -score, corresponding to a normalized effect size, and a P value denoting the frequency of scores equal to or higher than S in the null model (Fig. 1c and Methods). Therefore, CHESS enables a quantitative comparison and assessment of statistical significance of contact matrix similarities.

Automatic feature selection and classification of structural changes. In addition to the identification of changes in chromosome conformation data, a major analytical task consists of the recognition of changes in specific 3D chromatin organization features, such as TADs among others. To specifically determine which features change between a pair of samples, CHESS implements a simple and fast workflow of image filters that allow their automatic identification and classification (Fig. 1d). First, a differential contact matrix is computed for each CHESS comparison and gained and lost contacts in each matrix are separated for further analysis. Second, the matrices are denoised, smoothed and binarized to apply a close morphology filter to extract the individual areas changing in each comparison. Subsequently, the two-dimensional (2D) cross-correlation between all the extracted areas in a dataset is computed. Finally, k -means clustering is used to detect the main structural features identified in these areas (Fig. 1d). Overall, this strategy allows one to automatically identify the precise 3D structural features, such as TADs, loops and stripes, which are changing in a particular region.

CHESS requires only low sequencing depth and tolerates a high level of noise. To robustly estimate the performance of CHESS with regard to different experimental conditions (for example, noise, sequencing depth) and matrix parameters (for example, size, size difference), we generated a set of synthetic Hi-C data designed to reflect features commonly observed in real-world Hi-C matrices, including an exponential decay of contact frequency with genomic distance, TADs and loops (Methods). To test the sensitivity of CHESS to noise and sequencing depth, we compared an artificial matrix R to a copy Q of itself, while adjusting the sequencing depth and adding noise to both of the matrices independently. As a background to calculate P values and z -scores, similarity scores were additionally calculated in comparisons of R to 1,000 randomly generated artificial matrices at the same sequencing depth and noise level (Fig. 2a). These simulations show that Q is correctly assigned the best CHESS score for ‘deeply sequenced’ matrices up to a noise level of 90% (Fig. 2b). Beyond that, ranking quickly becomes random, which is reflected in a uniform distribution of P values (Extended Data Fig. 3). For artificial matrices with fewer contacts, CHESS still tolerates noise levels of 60–80%. Correspondingly, z -scores are consistently high for the same levels of noise depending on sequencing depth (Fig. 2c). Interestingly, z -scores do not peak at 0% noise. This can be explained by the changing s.d. of the background scores; with increasing noise, the similarity of random matrices increases, leading to progressively narrower distributions of CHESS scores. This leads to a slight increase in z -scores up until the point that CHESS is

no longer able to identify Q as the top-ranking hit in these comparisons (Extended Data Fig. 3).

To verify these results in a real-world setting, we repeated the above analyses on a deeply sequenced mouse embryonic stem cell (mESC) Hi-C dataset¹². Interestingly, the results on real data indicated an even higher robustness of CHESS to high noise and low sequencing depth than observed on synthetic datasets; while high robustness requires sufficiently large region sizes (>2.5 Mb), which is probably due to the increased amount of distinctive features in larger regions, CHESS tolerates sequencing depths as low as 0.06 M Mb⁻¹ and 80% noise (Extended Data Fig. 4), demonstrating the applicability of this approach in shallow-sequenced datasets. The increased robustness to noise is probably due to a stronger signal enrichment in structural features, compared to the artificial data since structural features remain visible by eye even at 95% noise (Extended Data Fig. 4). Additionally, CHESS results are highly robust to parameter changes, including comparison window span, step size, matrix resolution and sequencing depth (Extended Data Fig. 5 and Methods). Overall these results demonstrate the ability of CHESS to reliably detect similarities between Hi-C matrices even at very low sequencing depths and with high amounts of noise.

Finally, to benchmark CHESS, we compared it to three widely used differential interaction detection packages: HOMER³², diffHic³⁰ and ACCOST³¹. All methods were run using default parameters on Hi-C interaction matrices at a 5-kilobase (kb) resolution for chromosome 19 from mESCs and neural progenitor cells (NPCs)¹². Because a standard criterion for assessing accuracy of differential chromatin interactions does not exist, we performed the analysis by examining the degree of overlap in the differential interacting regions identified by the three methods. Overall, we found a high level of overlap (Extended Data Fig. 6). However, it is important to note that CHESS identifies entire regions with differences while diffHic, HOMER and ACCOST identify specific pairs of bins with significant differences in contact counts. A small proportion of differences were reported only by HOMER, diffHic and ACCOST. Importantly, many of these differential interactions were filtered out by CHESS due to low signal-to-noise ratios.

Together, these results demonstrate that CHESS can robustly identify changes in chromatin conformation features over a wide range of experimental conditions.

CHESS similarity scores are consistent for matrices of different sizes. An immediate advantage of the analytical strategy behind CHESS is that it allows the calculation of S for matrices of different sizes. This is needed to measure, for example, the similarity between Hi-C maps of different species or for paralog-containing regions within the same genome. To do so, we implemented an upscaling transformation of the smaller of the two matrices (in case these were of different sizes) using nearest neighbor interpolation (Methods). To test the performance of CHESS on matrices of different sizes, we calculated S using an artificial matrix R and a matrix Q that maintains the relative positions, sizes and intensities of all features in R (that is, TADs and loops) but differs in size by a certain scaling factor (Fig. 2d, Supplementary Fig. 1 and Methods). Randomly generated matrices of the same size as Q serve as the background to calculate statistical significance. In a ‘deeply sequenced’ Hi-C matrix of 1.5 M Mb⁻¹, divided into equally sized regions of at least 60 bins, CHESS consistently ranked Q as the matrix most similar to R even if Q was less than half the size of R (Fig. 2e,f). Small matrices Q (smaller than 30 bins) do not rank higher than random matrices since they do not provide enough space to fit the features (that is, TADs and loops) of the reference matrix. A test with a simulated sequencing depth of 100 k Mb⁻¹ and 25% noise led to similar results, demonstrating that the method’s ability to detect similarities between matrices of different sizes is robust to experimental noise and different levels of sequencing depth (Extended Data Fig. 7).

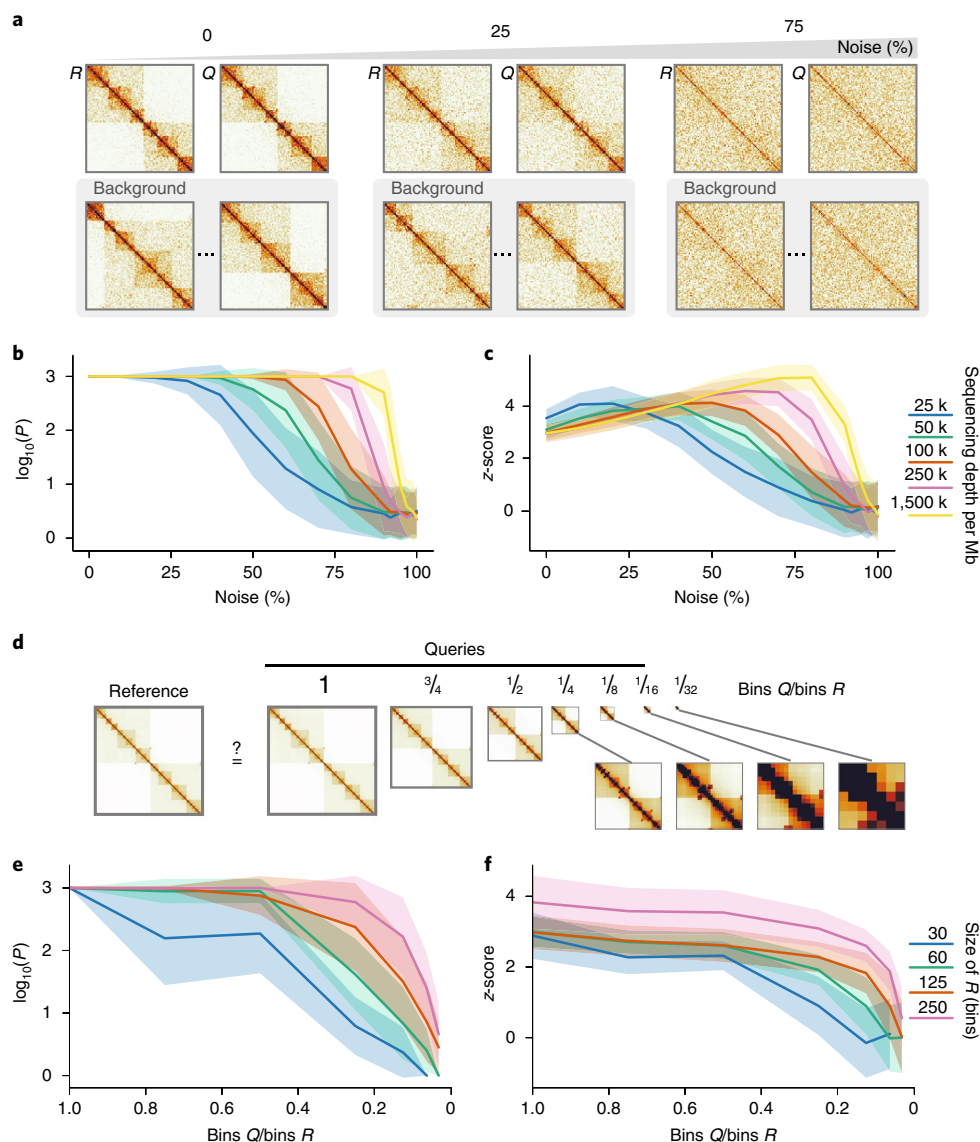


Fig. 2 | CHES evaluation on synthetic Hi-C matrices. **a–c**, Tests for the tolerance of CHES to increasing levels of simulated experimental noise. **a**, Schematic representation of the tests for the sensitivity of CHES to noise. **b**, Empirically determined CHES P values on synthetic matrices with increasing levels of noise (Methods). **c**, CHES z -scores on synthetic matrices with increasing levels of noise. **d,e**, Tests for the performance of CHES when comparing matrices of different sizes. **d**, Schematic representation of different scaling factors used to generate a query matrix Q from a reference R . **e**, Dependence of CHES P values on the scaling factor of the synthetic matrices Q and R . **f**, Dependence of CHES z -scores on the scaling factor. **b,c,e,f**, The solid lines indicate the mean, the shaded areas the s.d. over 100 simulations per parameter combination. Lowest possible P value in all tests = 0.001.

Comprehensive ranking of syntenic regions by structural similarity. Having validated the ability of CHES to reliably and robustly detect similarities and differences between synthetic and real Hi-C matrices, we next showcase its use in a real research scenario. Previous studies have examined the level of chromatin conformation similarity for regions of synteny (highly conserved sequences across species) finding a high degree of structural conservation between them^{2,10,18}. These comparisons have mainly focused on visual examination of individual examples^{10,18}, correlation analyses of specific 3D genome features, such as the binding of architectural proteins², measures of insulation³⁶ or the contact strength correlation within syntenic regions². However, a genome-wide quantification of the degree of similarity at the contact matrix level is lacking.

We used CHES to determine the level of chromatin conformation conservation for 175 regions of synteny between human and

mouse obtained from Synteny Portal³⁷. To calculate the statistical significance for the degree of conservation, we computed S scores for 100 random permutations of syntenic region pairs. CHES z -scores for true syntenic region pairs were strongly and consistently higher than those of random pairs (Fig. 3a, $P=0.01$; permutation test). Therefore, in agreement with previous observations^{2,10,18}, these results demonstrate genome-wide that overall regions of synteny between human and mouse share a similar 3D chromatin organization. However, our results highlight that not all regions of synteny have the same degree of structural similarity (Fig. 3b), suggesting that the evolutionary constraints on 3D chromatin structure are not uniform across the genome, resulting in different rates of evolution. In summary, our results demonstrate that CHES can be used to automatically quantify and rank 3D structural similarity genome-wide between species.

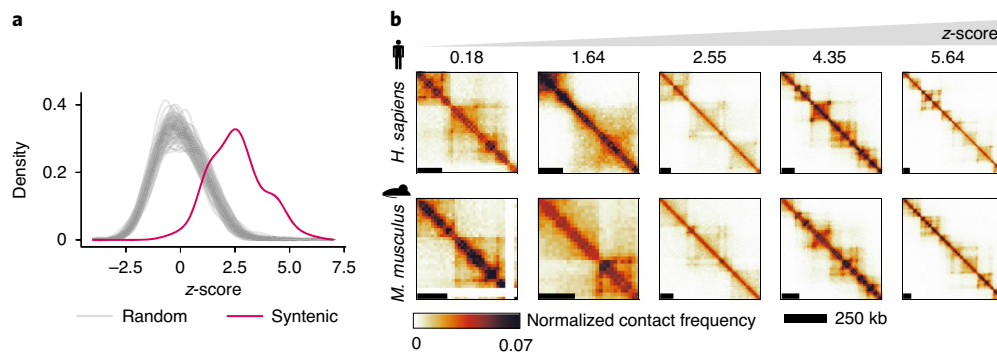


Fig. 3 | Global comparison of syntenic region similarity between human and mouse using CHES. **a**, Distributions of empirically determined CHES z-scores for 175 syntenic region pairs in human and mouse (red) and 100 random permutations of region pairs (gray) (Methods). A one-sided randomization test $P=0.01$, was used to compare the mean scores of randomly permuted pairs to the mean score of real syntenic regions. **b**, Examples of syntenic regions with increasing CHES z-scores from left to right.

Detection of structural variation on genetic perturbation. A common problem in genomics is the detection of emerging changes in a system with different experimental conditions, including targeted introduction of disturbances into the system. When applied to chromatin conformation, this approach has been fundamental to determine the contribution to 3D chromatin organization of different factors, such as CTCF-binding factor, cohesin or WAPL cohesin release factor, among others^{38–43}. However, these studies mostly relied on the detection of visual differences between Hi-C maps or the comparison of measurements derived from these maps, such as directionality or insulation indices.

Using the insulation score¹⁹, a metric that is low at TAD borders and high within TADs, we have previously shown that depletion of the pioneer transcription factor Zelda during early embryonic development in *Drosophila* leads to a weakening of insulation at TAD boundaries in loci strongly bound by Zelda in wild-type (WT) embryos⁹. Therefore, we sought to evaluate the sensitivity of CHES in detecting these changes, as well as its ability to detect further modifications in chromatin conformation that would have escaped detection by a simple comparison of insulation scores.

Running CHES in a comparison between WT nuclear cycle 14 and Zelda-depleted embryos resulted in the detection of 65 regions in the genome with changes in chromatin conformation. Out of the 62 differential boundaries identified previously⁹, 29 were contained in regions marked as changing by CHES. Visual inspection of the structural changes at the remaining 33 differential boundaries revealed that the differences in contact intensities were typically small and primarily caused by a decline in short distance contacts around the boundary (Supplementary Fig. 2). We reasoned that a smaller matrix size should increase the sensitivity of CHES with regard to these types of changes since they would correspond to a larger fraction of the input matrix pixels. Indeed, after reducing the size of the input matrices from 250 kb to 125 kb, we detected 51 out of 63 differential boundaries as changing, along with 163 additional regions. It is important to note that this approach would probably miss changes occurring far away from the diagonal, such as differences in the contact probability decay, long-range loops or large TADs. Therefore, we conclude that, by altering the size of the compared matrices, it is possible to fine-tune CHES to the scale of changes it can detect.

Visual examination of the regions captured in the first run and of control regions confirmed the detected differences. Notably, besides the already reported loss of insulation at a subset of Zelda-bound TAD boundaries (Fig. 4a)⁹, the newly identified regions highlighted a range of structural changes, such as differing signal intensity away from the main diagonal of the Hi-C matrix, suggesting changing

levels of chromatin compaction, and varying contact intensities inside TADs and at long distances (Fig. 4b–e). These results demonstrate that CHES can systematically identify regions that undergo structural changes on genetic perturbation, covering a broad spectrum of structural features, which correspond well to visual perception of differences.

CHES identifies structural abnormalities in DLBCL. We next sought to determine whether CHES could detect structural differences in clinically relevant samples. The characterization of these differences is of prime importance since changes in the 3D structure of chromatin in mammals can have a strong impact on genomic regulation and thereby give rise to disease phenotypes^{44,45} and the activation of oncogenes^{46,47}. We reasoned that by comprehensively scanning a genome for structural abnormalities compared to a healthy control (Ctrl), CHES could greatly aid our understanding of the relationship between nuclear architecture and disease. To test this, we performed a CHES comparison using a recent Hi-C dataset from primary DLBCL and healthy B cells⁴⁸. Across the whole genome, CHES identified 810 regions of 2 Mb with prominent structural variations in DLBCL (Fig. 5a,b). After filtering these for regions with high experimental noise (Methods), we obtained a high-confidence set of 112 regions exhibiting clear changes between healthy and diseased B cells (Fig. 5c–e). One of the most striking examples displayed the emergence of well-defined TAD structures in a region that seemed devoid of structural features in healthy cells (Fig. 5e). Despite these differences, our analysis also revealed that the majority of structures remained unchanged (Fig. 5f). To gain further insight into the nature of the changes, we applied the feature extraction component of CHES to the 112 selected regions. This resulted in the identification of 144 gained features (104 stripes and 40 TADs) and 53 lost loops in the DLBCL sample compared to the Ctrl (Fig. 5g). This illustrates the application of CHES to examine disease-related processes by systematically identifying genomic regions and characterizing the specific features whose 3D structure differs between healthy and diseased cells.

Identification and automatic classification of structural features in Capture-C data. Finally, we investigated the ability of CHES to automatically extract features in additional types of chromatin conformation capture datasets, such as tiled Capture-C experiments⁴⁹. To do so, we analyzed previously published Capture-C experiments for CRISPR–Cas9-mediated genome edits of architectural features, such as deletion of CTCF-binding factor binding sites, modifications of TAD boundaries and a TAD inversion at the *Sox9/Kcnj2* locus⁵⁰. CHES identified all previously described 3D rearrangements in

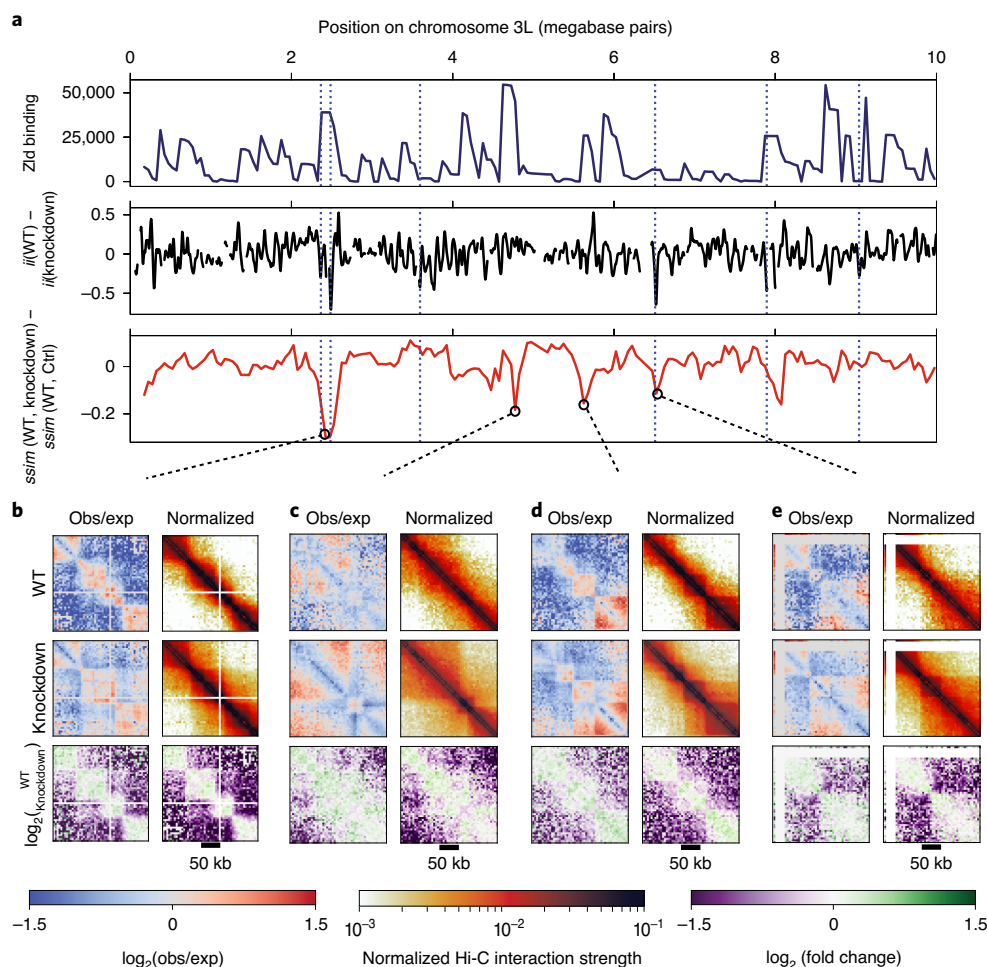


Fig. 4 | Identification of chromatin conformational changes in fly embryos after Zelda (*zld*) knockdown. **a**, Zelda binding signal in WT fly embryos (top), insulation score difference between WT and *zld* knockdown embryos (middle, smoothed) and difference between similarity scores calculated on WT to knockdown and WT to water injection Ctrl (bottom) for regions on a subset of chromosome 3L. The dotted blue lines indicate differential boundaries as identified by Hug et al.⁹ **b–e**, Four examples of regions with the strongest conformational changes between WT and knockdown, showing obs/exp and normalized Hi-C matrices, and \log_2 fold change matrices for WT/knockdown. The white lines on the Hi-C plots correspond to regions of the genome masked from the analysis due to low mappability.

the different mutants compared to WT mice (Fig. 6 and Extended Data Fig. 8). In addition, CHES identified marked differences that had not been reported in the original study. For example, besides the previously reported TAD fusion resulting from the *Sox9* regulatory domain inversion (InvC), CHES captured the loss of chromatin loops between the two TADs (Fig. 6a). A similar inversion not including the TAD boundary (Inv-Intra) did not result in a TAD fusion. However, CHES captured an increase in contact frequencies across the boundary in the form of a stripe (Fig. 6b). Applying CHES to all generated mutants systematically characterized the set of very subtle differences across these samples (Extended Data Fig. 8). This demonstrates that CHES can automatically identify and classify chromatin contact differences.

Discussion

The increasing wealth of available Hi-C datasets calls for fast quantitative algorithms that enable a systematic comparison of local chromatin structure. However, currently there are no algorithmic approaches for Hi-C data analysis that allow automated comparisons and classification of the identified 3D genome changes directly at the matrix level. This results in a lack of identification and characterization of a broad spectrum of differences in chromatin

conformation maps that can be visually recognized but that may be missed by more specialized approaches relying on preprocessed features. We have developed CHES to fill this gap by providing automated, systematic Hi-C matrix comparisons and feature classification that correspond well to the visual perception of structural differences (Fig. 1). A major feature of CHES is that it is not limited to comparing regions within a single dataset, but comparisons can be made among samples, cell types, developmental stages and even across different species, which makes it widely applicable. We demonstrated that CHES is robust to experimental noise and usable on shallow-sequenced datasets (Fig. 2). Furthermore, we showed that CHES can be used to perform cross-species comparisons (Fig. 3) and that it can detect 3D genome changes in genomes of different sizes (Figs. 4 and 5). Finally, we demonstrated that CHES can be used to analyze chromatin conformation capture datasets generated using different experimental approaches, such as Capture-C⁵⁰ (Fig. 6). Therefore, we expect CHES to be immediately applicable to other datasets, including tethered chromatin conformation capture⁵¹, digestion-ligation-only Hi-C⁵², genome architecture mapping⁵³ and microscopy-based methods, such as Hi-M⁵⁴.

An additional advantage of CHES is the fast and highly efficient implementation of the structural similarity algorithm that has a very

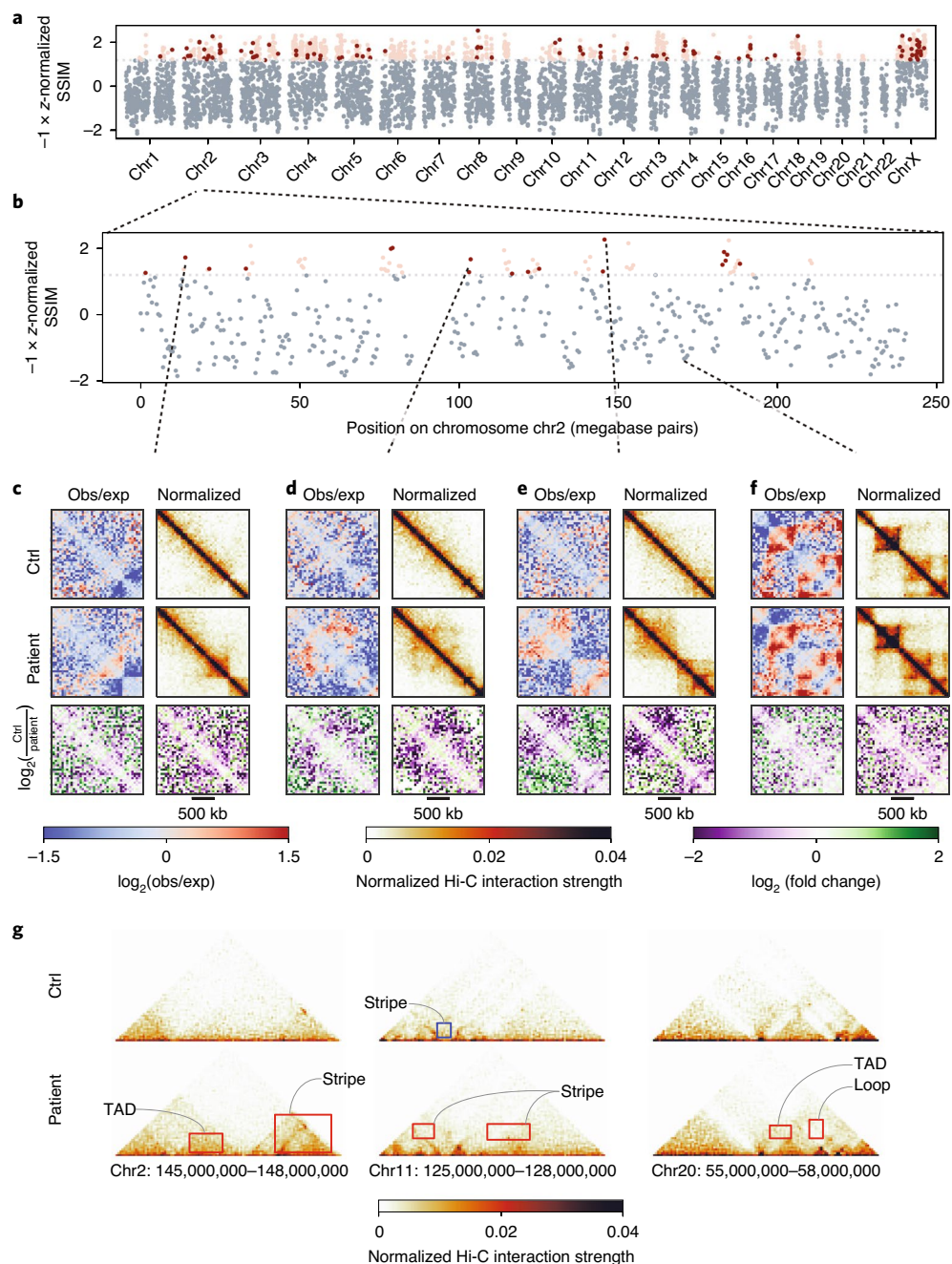


Fig. 5 | Identification of structural changes in a DLBCL. a, b, Similarity (z-normalized similarity score) of Hi-C data generated from healthy B cells (Ctrl) and a patient with DLBCL as assessed by CHES for 2-Mb regions. Highly dissimilar regions (z-normalized similarity score ≤ -1.2) are colored in red, whereas noisy regions (signal-to-noise ratio < 0.6) are in light red. **c–f**, Examples of regions with conformational changes (**c–e**) and conservation (**f**) between healthy and diseased B cells, showing obs/exp and normalized Hi-C matrices and \log_2 fold change matrices for Ctrl/patient. **g**, Three examples of highly dissimilar regions identified by CHES, with the gained and lost features highlighted in red and blue, respectively. The features are annotated according to their structural category.

small memory footprint since only the two matrices that are being compared need to be loaded. As a comparison, when scanning a whole chromosome for structural differences among conditions, CHES achieves a 4–320 times speedup at 3 times lower memory consumption compared to HOMER, diffHic and ACCOST^{30–32} (Extended Data Fig. 6). This makes the approach usable without requiring an advanced computational infrastructure. In addition, the nature of CHES comparisons makes them trivially parallelizable so that the algorithm can be efficiently sped up by dedicating

more computational resources to it. This allows CHES to make the myriad of comparisons necessary for more complex biological questions, including the background computation for comparing regions of different origin.

Within this context, a promising outlook for CHES applications is the de novo discovery of structurally similar regions between two genomes using an all-against-all comparison approach. These ‘structurally syntenic’ region pairs could provide fundamental insights on the evolution of nuclear architecture and its 3D

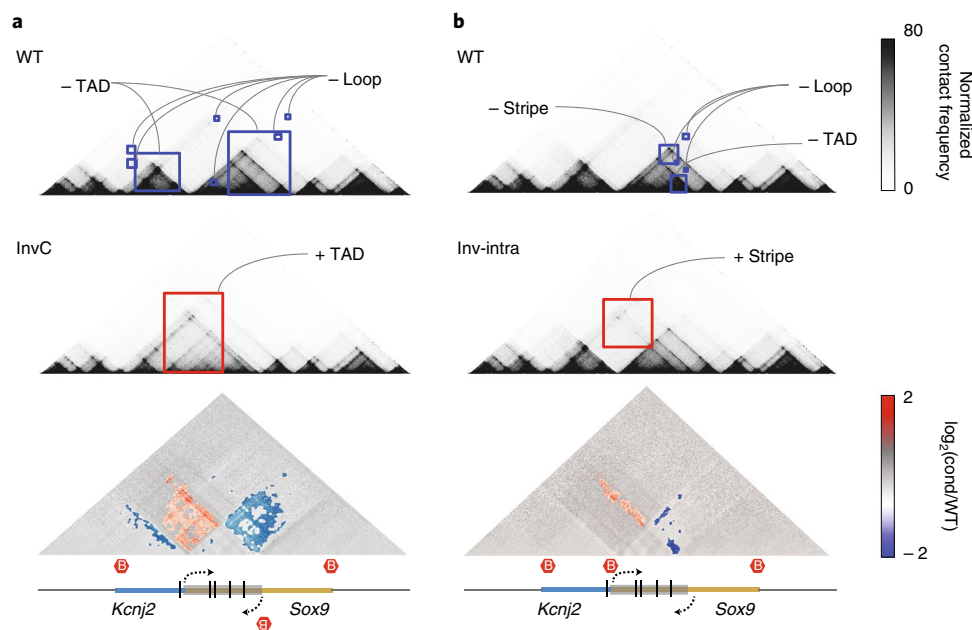


Fig. 6 | Feature extraction from Capture-C data. Data from Despang et al.⁵⁰. **a**, Feature extraction of WT against InvC mutant maps, which present an inversion in the *Sox9* sequence represented by a gray box (bottom). Lost and gained structures in the mutant are highlighted in blue and red squares, respectively. The bottom plot shows a \log_2 fold change map with identified features colored according to the directionality of the change. Red hexagons demarcate the positions of TAD boundaries. **b**, Same as **a** for feature extraction of WT against Inv-intra mutant maps, where the same sequence is inverted, not including the border between the two TADs.

constraints, including the effects of processes affecting 3D chromatin organization, such as rearrangements or changes in the binding of architectural proteins. Despite the efficiency of CHES, further work and heuristics are necessary to make this computationally tractable. Highlighting the importance of considering the 3D genome in evolutionary analyses, we find different degrees of structural conservation across mammalian evolution (Fig. 4), suggesting different rates of evolutionary change in these regions. This demonstrates how CHES can already facilitate the study of evolutionary genomics in the context of 3D structure.

Similarly, the identification of structural variation and its association with abnormalities in 3D genome organization and gene expression misregulation is central to evaluate the contribution of chromatin organization to disease-generating processes. As a proof of principle, in this study we demonstrate how CHES can be used to detect a number of chromatin conformational alterations genome-wide in B cells from a patient with DLBCL without the need of previous knowledge regarding the nature of the aberrations. Interestingly, our analysis identified regions in the genome gaining structural features, such as TADs and loops, despite the lack of protein-coding genes in these regions. Instead, these regions frequently contained long noncoding RNAs and pseudogenes. Future work integrating other patient-matched genome-wide datasets, such as chromatin accessibility or RNA sequencing, are necessary to determine the cause and consequence of these changes in relation to disease.

Future improvements of CHES might benefit from a further dissection of the SSIM, which may allow us to pinpoint the contributions of individual regions to overall matrix similarity. In general, structural similarity of images is an active field of research. Modifications improving the robustness of the SSIM to small shifts in position⁵⁵ or its power to identify similar subimages⁵⁶ are promising, but it remains to be determined which of the algorithms developed for assessing image similarity are compatible with the specific requirements of Hi-C matrix comparisons.

In conclusion, CHES is an algorithm that can be used to quantitatively assess and classify the structural similarity of two genomic regions from chromosome conformation capture data, without the need for feature selection before comparison. CHES is highly tolerant of differences in chromatin conformation capture library size and the noise level of datasets. Its applications include the ranking of known region pairs by similarity, such as syntenic regions in different species, and the discovery of structural changes, such as chromatin conformational changes of the same genomic region in two different conditions. CHES has great utility in the field of chromatin conformation and can simplify the identification of disease-associated structural variation in clinical applications.

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/s41588-020-00712-y>.

Received: 23 July 2018; Accepted: 4 September 2020;
Published online: 19 October 2020

References

1. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat. Rev. Genet.* **17**, 661–678 (2016).
2. Vietri Rudan, M. et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.* **10**, 1297–1309 (2015).
3. Acemel, R. D., Maeso, I. & Gómez-Skarmeta, J. L. Topologically associated domains: a successful scaffold for the evolution of gene regulation in animals. *Wiley Interdiscip. Rev. Dev. Biol.* **6**, e265 (2017).
4. Lazar, N. H. et al. Epigenetic maintenance of topological domains in the highly rearranged gibbon genome. *Genome Res.* **28**, 983–997 (2018).
5. Eres, I. E., Luo, K., Hsiao, C. J., Blake, L. E. & Gilad, Y. Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. *PLoS Genet.* **15**, e1008278 (2019).

6. Yang, Y., Zhang, Y., Ren, B., Dixon, J. R. & Ma, J. Comparing 3D genome organization in multiple species using phylo-HMRF. *Cell Syst.* **8**, 494–505.e14 (2019).
7. Ke, Y. et al. 3D chromatin structures of mature gametes and structural reprogramming during mammalian embryogenesis. *Cell* **170**, 367–381.e20 (2017).
8. Du, Z. et al. Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature* **547**, 232–235 (2017).
9. Hug, C. B., Grimaldi, A. G., Kruse, K. & Vaquerizas, J. M. Chromatin architecture emerges during zygotic genome activation independent of transcription. *Cell* **169**, 216–228.e19 (2017).
10. 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).
11. Dixon, J. R. et al. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
12. Bonev, B. et al. Multiscale 3D genome rewiring during mouse neural development. *Cell* **171**, 557–572.e24 (2017).
13. Nagano, T. et al. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61–67 (2017).
14. Gibcus, J. H. et al. A pathway for mitotic chromosome formation. *Science* **359**, eaao6135 (2018).
15. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018).
16. Krijger, P. H. L. & de Laat, W. Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. Cell Biol.* **17**, 771–782 (2016).
17. Darrow, E. M. et al. Deletion of *DXZ4* on the human inactive X chromosome alters higher-order genome architecture. *Proc. Natl Acad. Sci. USA* **113**, E4504–E4512 (2016).
18. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
19. Crane, E. et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240–244 (2015).
20. Yang, T. et al. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.* **27**, 1939–1949 (2017).
21. Sauria, M. E. G. & Taylor, J. QuASAR: quality assessment of spatial arrangement reproducibility in Hi-C data. Preprint at *bioRxiv* <https://doi.org/10.1101/204438> (2017).
22. Ursu, O. et al. GenomeDISCO: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics* **34**, 2701–2707 (2018).
23. Yan, K.-K., Yardımcı, G. G., Yan, C., Noble, W. S. & Gerstein, M. HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps. *Bioinformatics* **33**, 2199–2201 (2017).
24. Shavit, Y. & Lio, P. Combining a wavelet change point and the Bayes factor for analysing chromosomal interaction data. *Mol. Biosyst.* **10**, 1576–1585 (2014).
25. Huynh, L. & Hormozdiari, F. TAD fusion score: discovery and ranking the contribution of deletions to genome structure. *Genome Biol.* **20**, 60 (2019).
26. Paulsen, J. et al. HiBrowse: multi-purpose statistical analysis of genome-wide chromatin 3D organization. *Bioinformatics* **30**, 1620–1622 (2014).
27. Lareau, C. A. & Aryee, M. J. diffluoop: a computational framework for identifying and analyzing differential DNA loops from sequencing data. *Bioinformatics* **34**, 672–674 (2018).
28. Djekidel, M. N., Chen, Y. & Zhang, M. Q. FIND: differeNtial chromatin INteractions Detection using a spatial Poisson process. *Genome Res.* **28**, 412–422 (2018).
29. Stansfield, J. C., Cresswell, K. G., Vladimirov, V. I., Dozmorov, M. G. HiCcompare: an R-package for joint normalization and comparison of Hi-C datasets. *BMC Bioinformatics* **19**, 279 (2018).
30. Lun, A. T. L. & Smyth, G. K. diffHiC: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* **16**, 258 (2015).
31. Cook, K. B., Hristov, B. H., Le Roch, K. G., Vert, J. P. & Noble, W. S. Measuring significant changes in chromatin conformation with ACCOST. *Nucleic Acids Res.* **48**, 2303–2311 (2020).
32. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
33. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).
34. Wang, Z. & Bovik, A. C. A universal image quality index. *IEEE Signal Process. Lett.* **9**, 81–84 (2002).
35. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
36. Harmston, N. et al. Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *Nat. Commun.* **8**, 441 (2017).
37. Lee, J. et al. Synteny Portal: a web-based application portal for synteny block analysis. *Nucleic Acids Res.* **44**, W35–W40 (2016).
38. Schwarzer, W. et al. Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51–56 (2017).
39. Nora, E. P. et al. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930–944.e22 (2017).
40. Haarhuis, J. H. I. et al. The cohesin release factor WAPL restricts chromatin loop extension. *Cell* **169**, 693–707.e14 (2017).
41. Rao, S. S. P. et al. Cohesin loss eliminates all loop domains. *Cell* **171**, 305–320.e24 (2017).
42. Wutz, G. et al. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.* **36**, 3573–3599 (2017).
43. Gassler, J. et al. A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. *EMBO J.* **36**, 3600–3618 (2017).
44. Lupiáñez, D. G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
45. Franke, M. et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).
46. Flavahan, W. A. et al. Insulator dysfunction and oncogene activation in *IDH* mutant gliomas. *Nature* **529**, 110–114 (2016).
47. Hnisz, D. et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016).
48. Diaz, N. et al. Chromatin conformation analysis of primary patient tissue using a low input Hi-C method. *Nat. Commun.* **9**, 4938 (2018).
49. Hughes, J. R. et al. Analysis of hundreds of *cis*-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* **46**, 205–212 (2014).
50. Despang, A. et al. Functional dissection of the *Sox9-Kcnj2* locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.* **51**, 1263–1271 (2019).
51. Kalhor, R., Tjong, H., Jayatilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* **30**, 90–98 (2011).
52. Lin, D. et al. Digestion-ligation-only Hi-C is an efficient and cost-effective method for chromosome conformation capture. *Nat. Genet.* **50**, 754–763 (2018).
53. Beagrie, R. A. et al. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519–524 (2017).
54. Cardozo Gizzi, A. M. et al. Microscopy-based chromosome conformation capture enables simultaneous visualization of genome organization and transcription in intact organisms. *Mol. Cell* **74**, 212–222.e5 (2019).
55. Sampat, M. P., Wang, Z., Gupta, S., Bovik, A. C. & Markey, M. K. Complex wavelet structural similarity: a new image similarity index. *IEEE Trans. Image Process.* **18**, 2385–2401 (2009).
56. Homola, T., Dohnal, V. & Zezula, P. Searching for sub-images using sequence alignment. In *Proc. 2011 IEEE International Symposium on Multimedia* 61–68 (IEEE, 2011).

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

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2020

Methods

The CHES pipeline. The CHES pipeline is illustrated in Fig. 1. CHES takes two normalized^{57,58}, whole-genome Hi-C matrices as input. We recommend using matrices at least 100×100 bins in size (20×20 is the absolute minimum allowed by CHES) with no more than 10% of all bins unmappable (without signal). In a first step, these are transformed to obs/exp matrices⁵⁵ by dividing each matrix entry by the average of all entries at the same distance (see below). This transformation is necessary to remove the distance dependency of pairwise contact probabilities that is characteristic of Hi-C matrices. CHES comparisons of matrices that are not corrected for this distance dependency of contact probabilities are sensitive to varying experimental noise and relative region sizes (Extended Data Fig. 1). From the whole transformed matrices, the submatrices corresponding to the specified regions of interest are extracted, forming a comparison pair R (reference) and Q (query). Subsequently, R or Q are resized to the dimensions of the larger matrix using nearest neighbor interpolation (`skimage.transform.resize` in the `scikit-image` package⁵⁹ (v.0.13.1)). If regions are located on different strands, the matrices are rotated by 180 degrees (for example, in case a syntenic region is annotated on the reverse strand). All bins marked as unmappable in either of the matrices are removed from both matrices. The resulting processed versions of R and Q are handed to the structural similarity function, yielding a raw similarity score. We use an implementation of the original structural similarity algorithm^{33,34} available for the Python programming language in the `scikit-image` module⁵⁹ (v.0.13.1). While this function was initially developed for the evaluation of image quality^{33,34}, it does not make any assumptions about its input data other than that it comes in the form of two matrices of same dimensionality, with numerical entries, irrespective of how the data in these matrices have been generated. Outside of the computer vision field, it is, for example, also used in transportation research to compare matrix representations of origin-destination graphs, which differ from chromatin contact graphs conceptually only in that they are directed graphs^{60,61}, and as a similarity metric for acoustic pressure signals^{62,63}.

In some applications, R may be compared to a pool of matrices P forming the background model. The process described above for the pair R and Q is repeated for each pair R and $Q^p \in P$. We use the similarity scores B obtained from these background comparisons to calculate a P value and a z -score for the raw score $s = \text{SSIM}(R, Q)$:

$$P = \frac{|\{x \in B | x \geq s\}|}{|B|}$$

$$z = \frac{s - \mu_B}{\sigma_B}$$

where μ_B denotes the mean and σ_B the s.d. of the scores in B . We used two kinds of background models for this manuscript: (1) all submatrices of Q 's size located on the same chromosome as Q for the comparison of chromatin structures between syntenic regions; and (2) a pool of synthetic matrices built with the same parameters as Q but randomly generated features for the tests of CHES on synthetic Hi-C data. CHES P values are not automatically corrected for multiple testing since this is not necessary for all use cases. If CHES is used to identify significantly similar or different regions across the genome with a fixed acceptance threshold, then CHES P values need to be corrected for multiple testing.

Next, CHES extracts individual features that are different between two genomic regions (Fig. 1d). First, gained and lost contacts in the R matrix are computed and separated as increased/decreased interactions with respect to Q . Then, a set of image filters, with the parameters automatically adjusted according to the matrix size or user-defined, are applied to these two matrices that are from now on considered as images: (1) denoise the image using a bilateral filter⁶⁴. This is an edge-preserving filter that averages pixels based on their spatial closeness and their radiometric similarity; by default, they are computed using a window size of three. The Gaussian function of the Euclidean distance between two pixels and its s.d. are used to obtain the spatial closeness. The Euclidean distance between two color values is used for the radiometric similarity; CHES by default uses the mean value of the matrix. Notably, higher values of spatial closeness and radiometric similarity will average the bins with larger differences; (2) smooth the image using a median filter. This scans the image using a square-shaped array with an area computed automatically depending on the picture size. This array scans the image using a windows size and computes the median of the pixels, smoothing the signal. Higher values will smooth larger structures, while smaller values will consider more subtle signals; (3) image binarization using Otsu's method⁶⁵. It returns a threshold value that separates the pixels in two classes. This algorithm searches for the threshold that minimizes the intraclass variance. This threshold by default is calculated using whole matrix values to be more refined; (4) morphological closing of the image. This filter is used to remove small dark spots and connect small bright cracks. This helps to remove the remaining noise and enclose the individual structures. By default, CHES uses a square of eight bins, higher values will enclose larger structures while lower will consider smaller or more punctuated signals.

With the four filters, CHES extracts individual structures, which can be used to get the main structural clusters according to their pattern of interactions. First, the 2D cross-correlation between all the individual features is computed. Finally,

the k -means clustering algorithm is applied to obtain the main structural clusters. The optimal number of clusters is computed according to the elbow method by fitting the model within a range of 1–15 clusters, which may vary depending on the number of identified differential features. The robustness of the clustering was assessed by downsampling the identified structural features from the Hi-C data generated from healthy B cells (Ctrl) and patient with a DLBCL and computing the optimal number of clusters. This process was repeated 1,000 times. The clustering step proved to be highly robust to data sparsity (Supplementary Fig. 3).

Calculation of obs/exp matrices. We calculated the obs/exp form $M^{\text{obs/exp}}$ of a balanced matrix M by first computing the expected matrix M^{exp} by determining the average value of each diagonal in M :

$$M_{ij}^{\text{exp}} = \frac{\sum_{n=1}^{N-D} M_{D+n,n}}{N-D} \text{ with } D = i - j, i \geq j$$

Since M is symmetric around $M_{i=j}$, we computed this only for $i \geq j$ and then set $M_{ij}^{\text{exp}} = M_{ji}^{\text{exp}}$.

We then calculated $M^{\text{obs/exp}}$:

$$M_{ij}^{\text{obs/exp}} = \frac{M_{ij}}{M_{ij}^{\text{exp}}}$$

As for the matrix balancing, the obs/exp calculation was performed on a per-chromosome basis for real Hi-C data.

Generation of synthetic Hi-C matrices. To test the performance of CHES on datasets with different sequencing depths, we generated synthetic matrices within a range of numbers of simulated read pairs. Typical numbers of valid read pairs in current Hi-C studies range from 0.1 million per megabase (M Mb⁻¹) (refs. ^{10,18}) to 1 M Mb⁻¹ (ref. ¹²). As an example of a deeply sequenced dataset, we generated matrices with an equivalent depth of 1.5 M Mb⁻¹, corresponding to approximately 4.5 billion mapped reads across the whole genome in a Hi-C experiment on human cells. Datasets at lower sequencing depths were then generated by downsampling the number of read pairs in the original matrix by randomly removing pairs of contacts (Supplementary Fig. 4 and the 'Simulation of different sequencing depths and experimental noise' section below). This ensures that the overall structure of the dataset is maintained for the evaluation of sequencing depth-related effects. In addition, experimental noise was also simulated in the synthetic datasets by removing a number of contacts and adding them at random locations (Supplementary Fig. 4). This allowed us to model the effect of random ligations, a main contributor to noise in chromatin contact maps.

To generate a synthetic matrix, we performed the following steps. First, we produced an empty matrix M of dimensions n^2 . We then filled M with simulated pairs of reads, modeling the power-law decay of signal away from the main diagonal^{35,66} by:

$$x_i = \left(10^{-4} + \frac{1 - 10^{-4}}{1,000} i\right)^{-0.85}, i = 0, 1, \dots, n$$

where x_i denotes the read counts at the i th diagonal, counted as moving away from the main diagonal at $i=0$. At this point, the number of reads is uniform in each diagonal and the mean number of reads per bin is inversely proportional to the matrix size. We then added structural features resembling TADs and loops to M .

First, three layers of TADs were added. TAD size was randomly determined by drawing a size s from a truncated normal distribution, while TAD intensity was modeled by adding a constant read count to a square of area s^2 . For each consecutive layer TAD size decreased while TAD intensity increased (Supplementary Table 1). To start with, we placed a first TAD at a randomly chosen position on the main diagonal at least $0.1n$ away from each end of the diagonal. We then filled the main diagonal to both sides of this initial TAD with adjacent TADs. In cases where a space smaller than the lower bound of the truncated normal occurred at the ends of a diagonal, we covered it by adding a small TAD that can be thought of as being part of a bigger TAD reaching into the field of view from an adjacent genomic region. In the second and third round, smaller TADs were placed inside the TADs generated in the previous round.

Second, corner loops were added to TADs with a chance of 1/3. Loops were modeled as squares with an additional, randomly selected intensity of either 90, 140 or 220% of the intensity and side lengths of either 5, 7 or 10% of the side length of the corresponding TAD.

Simulation of different sequencing depths and experimental noise. Hi-C matrix quality and resolution are primarily affected by two properties: (1) random ligations, which are distinct from proximity ligations; and (2) sequencing depth. These properties have distinct effects: random ligations are an indicator of poor library quality and introduce 'noise' into the Hi-C matrix, while sequencing depth determines the achievable matrix resolution. However, they are not entirely independent. Increased sequencing depth can mitigate the effects of random ligations by enriching contacts in regions with 'true' proximity signal.

In all our tests of CHES, lower sequencing depths were simulated by subsampling, that is, the random removal of 'ligation fragment' pairs in a high-resolution matrix. Random ligations, in contrast, were simulated by the random replacement of pairs in a matrix. In particular, to model different sequencing depths, we lowered the density of read pairs in M by removing random read pairs until we reached the desired number of read pairs d . Subsequently, we simulated an experimental noise level ε by reassigning $r = \varepsilon \times d$ read pairs to randomly selected pairs of loci.

As stated above, our noise models random ligations in the Hi-C experiment that can occur after the genomic material has been digested. These random ligations are intramolecular ligations (not within a cross-linked pair). Therefore, the probability of a random ligation of two fragments is not related to the linear genomic distance between them. Consequently, the distance decay graph is expected to approach a flat line as the fraction of random ligations in the Hi-C library increases. Our noise procedure moves a fraction of the reads in each bin to randomly chosen bins, where each bin in the map has the same chance of receiving a read. This procedure results in the expected behavior of the distance decay graph, as shown in Supplementary Fig. 5.

The downsampling procedure, in contrast, removes randomly chosen reads without adding them anywhere. Since the fraction of reads removed is on average the same at all genomic distances, the distance decay graph does not change substantially due to sampling. Slight deviations occur only close to the maximum distance, where the numbers of reads and bins are small enough to allow for random fluctuations of the mean after sampling. We show the largely unchanged distance decay in Supplementary Fig. 5.

We simulated the size difference of regions with similar structural features by first generating a reference matrix of a certain size n_r and saving the relative positions and intensities of structural features in it. We then generated query matrices of smaller sizes n_q and placed the structural features at the same relative positions (rounded to the next full bin) with the same intensities. To ensure equal sequencing depth relative to the matrix size between the scaled matrices, we subsequently adjusted the depth of the scaled matrices to a scaled depth d_q in relation to the depth of the reference matrix d_r :

$$d_q = d_r \frac{n_q}{n_r}$$

The structural similarity algorithm in chromatin contact map comparisons.

The SSIM score for whole matrices can be calculated from the average of multiple 'subscores' obtained on smaller subsets of a matrix, the size of which can be controlled with dedicated window size parameters (Extended Data Fig. 2). Each subscore consists of three components: corrections for illuminance (differences in brightness); corrections for contrast; and the correlation coefficient between the two matrices (Extended Data Fig. 2) (refs. ^{33,34}). By default, CHES does not use subscores but computes a single SSIM value for the whole matrix comparison immediately.

We quantified the contribution of each component to the final score in comparisons of a random synthetic Hi-C reference matrix to an identical copy of itself and to a pool of 1,000 randomly generated matrices of the same size. We assessed the dependence of the final score on each component using multiple window sizes (Extended Data Fig. 2). For sufficiently large window sizes, SSIM subscores are perfectly reflected by the combination of the contrast and correlation components. Only for very small window sizes does illuminance play a minor role. While different window sizes affect the scores and relative rank of random matrices (Extended Data Fig. 2), the comparison of the reference matrix to its identical copy yields a perfect score in all comparisons, independently of window size.

Tests on synthetic Hi-C matrices. We tested CHES on synthetic matrices in two main test scenarios: noise/sequencing depth tests; and size/size difference tests. The test setup was similar in both scenarios. For each test run, we first defined the test conditions by setting the following parameters: the size of reference matrices; query matrices; the reference noise level; query noise level and sequencing depth (always the same for reference and query); the type of the input matrices (normalized or obs/exp); and the window size parameter of the structural similarity function. Using these parameters, we then generated a reference set of 100 synthetic Hi-C matrices. Corresponding to these references we then produced 100 query matrices, differing in a certain parameter (noise level or size) by a certain factor, but with the same structural features in the same positions (as described in 'Generation of synthetic Hi-C matrices'). We then generated a decoy pool of 1,000 synthetic matrices with all parameters equal to the query matrices but with randomly generated features. Each reference matrix was compared to its corresponding query using the structural similarity algorithm with the specified parameters and also to each matrix in the decoy pool, which we used as a simulation of the genomic background. The P values and z -scores were then calculated as described in 'The CHES pipeline'. The best possible P value in this test ($P \leq \frac{1}{1,000}$) was achieved when the comparison score of R versus Q was greater than all scores from comparisons to the 1,000 random matrices.

Processing of Hi-C matrices. We obtained Hi-C sequencing reads for human IMR90 (ref. ¹⁰), mouse CH12.LX¹⁰ (GSE63525), mouse ESCs (mESCs) and NPCs¹²

(GSE96107) and fly embryos at nuclear cycle 14, in WT Zelda knockdown cells (*zld* knockdown) and injected water control⁹ (E-MTAB-4918), as well as B-cell lymphoma and DLBCL⁴⁸ (E-MTAB-5875). The B-cell lymphoma and DLBCL data were processed as described previously⁴⁸. Fly data were processed as described in Hug et al.⁹

All human and the CH12.LX mouse paired-end FASTQ files were mapped independently to the reference genome (hg19 and mm10, respectively) in an iterative fashion using Bowtie v.2.2.4 with the -very-sensitive preset. Briefly, unmapped reads were truncated by 15 base pairs (bp) and realigned iteratively, until a valid alignment could be found or the truncated read was shorter than 25 bp. Only uniquely mapping reads with a mapping quality ≥ 30 were retained for downstream analysis. mESC and NPC FASTQ files were mapped using the Burrows-Wheeler Aligner MEM v.0.7.17-r1188 in a noniterative fashion with default parameters.

Restriction fragments were computationally predicted using the Biopython⁶⁷ v.1.71 Restriction module. Reads were assigned to fragments and fragments pairs were formed according to the read pairs. Pairs were then filtered for self-ligated fragments, PCR duplicates (both read pairs mapping within 1 bp of each other), read pairs mapping further than 5 kb from the nearest restriction sites and ligation products indicating uninformative ligation products⁴⁸. The Hi-C matrix was built by binning each genome at a given resolution of 10 kb and 25 kb and counting valid fragment pairs falling into each respective pair of bins. Finally, bins with less than 25% (human) or 10% (mouse) of the median number of fragments per bin were masked and the matrix was normalized using Knight-Ruiz (KR) matrix balancing⁵⁸ on each chromosome independently.

Tests on real Hi-C matrices. The robustness of CHES was also tested using real Hi-C data from experiments in Bonev et al.¹² and Diaz et al.⁴⁸. We show the results in the Supplementary Fig. 5 and Extended Data Fig. 5.

First, the data from Bonev et al.¹², binned at 25 kb, were used to repeat the analysis performed on synthetic matrices ('Tests on synthetic Hi-C matrices'). Different levels of noise (5, 20, 35, 50, 65, 80 and 95%) were added to the raw Hi-C matrix of chromosome 19. This was done twice independently to obtain versions *A* and *B* of the matrix to model matrices coming from independent experiments. Each of these was then downsampled (to 1, 5, 50 and 95% of the original number of reads), corrected and transformed to obs/exp matrices. For each combination of noise and sampling depth, CHES was run in default mode (using comparisons to the rest of the chromosome as the background model) to compare the same regions in the *A* and *B* matrices. These region pairs were obtained from a sliding window of sizes 1, 2.5, 5, 7.5 and 10 Mb, with a step size of 25 kb. The resulting mean P values and z -scores, as well as their variances were plotted as shown in Extended Data Fig. 4. The best possible P value, or perfect performance, was achieved when no other region got an equal or higher similarity score than the region with identical positions in *A* and *B*. We found the dependence on window size to be the main parameter governing the robustness of CHES; smaller bin size, that is, higher resolution of the maps, did not qualitatively change the results (Extended Data Fig. 4).

Second, Hi-C data from mESCs and NPCs at a resolution of 25 kb from the same dataset¹² were used for the reproducibility analysis of CHES when varying two data parameters. CHES was run using different combinations of window span (250 kb, 500 kb, 1 Mb, 2 Mb and 3 Mb) and step sizes (25 kb, 250 kb, 500 kb and 1 Mb). The Jaccard index was calculated to obtain the overlap between the identified genomic regions. To check the reproducibility of CHES results using different sequencing depths (percentage of reads: 80, 60, 40 and 20), we applied CHES using a 25-kb resolution, 3-Mb windows span and 500-kb step size. The data from Diaz et al.⁴⁸ were used to check how consistent CHES was when varying three data parameters, namely different values of windows span (250 kb, 500 kb, 1 Mb, 2 Mb and 3 Mb), step sizes (25 kb, 250 kb, 500 kb and 1 Mb) and resolutions (25 and 10 kb). One point in the plot (Extended Data Fig. 5) corresponds to the Jaccard index computed for a pair of CHES runs with different combinations of parameter values. All possible pairs were compared.

Benchmark analysis. Hi-C interaction matrices at a 5-kb resolution for chromosome 19 from mESCs and NPCs from Bonev et al.¹² were scanned for differences using different tools. HOMER³² (v.4.11), diffHic⁴⁰ (v.1.22.0) and ACCOST³¹ (<https://bitbucket.org/noblelab/accost/src/master;commit=b4a2f4b>) were run using default parameters. CHES was run using a windows span of 1 Mb and a step size of 500 kb. All tools were run using a single central processing unit computational machine of the following characteristics: Intel Xeon W, 3 GHz with 128 Gb of random access memory.

In particular, CHES was run using a windows span of 1 Mb and a step size of 500 kb. CHES ran approximately 7 times faster than HOMER, approximately 15 times faster than diffHic and approximately 320 times faster than ACCOST and had an approximately 4 times lower peak memory consumption than the two other tools (Extended Data Fig. 6).

To assess the similarities and differences among the three methods, we selected for each method all bins that were involved in a significant difference between mESC and NPC contact maps. Then, the selected bins were intersected to identify

those bins common to the three methods, any common bin by at least two of the three methods and, finally, any bin identified only by one of the methods (Extended Data Fig. 6). CHES and HOMER identified about 9,000 bins with differential interactions between mESCs and NPCs while diffHic identified about 4,000 and ACCOST about 6,000. Of the total identified differences, approximately 12% were identified by CHES, HOMER, diffHic and ACCOST. About 50% of differences were identified by CHES and HOMER alone.

Comparison of syntenic regions between *Homo sapiens* and *Mus musculus*.

We retrieved the annotations for syntenic blocks between hg19 (selected as reference) and mm10 with a resolution of 300 kb using SynBuilder³⁷. We used CHES to compare syntenic region pairs in Hi-C matrices at a 25-kb resolution for the human fibroblasts and mouse lymphoblasts. As control, we also carried out comparisons between region pairs with shuffled syntenic region IDs. This was repeated 100 times. To reduce the runtime of our method, we used a randomly chosen subset of 175 syntenic regions. For the same reason, we restricted the background calculation to the query chromosome the syntenic region was located on.

Detecting structural changes between WT and *zld* knockdown in *Drosophila melanogaster*. We obtained the locations of differential boundaries in *Drosophila melanogaster* and Hi-C data at a 5-kb resolution for the WT, *zld* knockdown and water control for nuclear cycle 14 from Hug et al.³. From these Hi-C data we computed insulation scores as described in the same publication. We smoothed the resulting index track with a Savitzky–Golay (implemented in Scipy⁶⁹ (v.1.0.0)) filter (window = 29, polyorder = 2, derivative = 0). We obtained data for Zld binding chromatin immunoprecipitation followed by sequencing experiments from Blythe et al.⁷⁰.

We partitioned the *D. melanogaster* genome into 250/125-kb regions with a step size of 50/25 kb. We ran CHES on the obs/exp-transformed Hi-C matrices corresponding to these regions, always comparing a region in WT to the same region in water control and knockdown. Inside the same windows we summed the $\log_{10}(q)$ values for all Zld peaks with a $\log_{10}(q)$ value >10 to generate the Zld binding tracks.

Using the CHES comparisons between WT–water control and WT–knockdown, we defined regions with structural changes as regions located at local minima of the track with values smaller or equal to −0.1.

Differential boundaries were defined as boundaries present in WT c14 cells (calls are available at <https://github.com/vaquerizaslab/Hug-et-al-Cell-2017-Supp-Site>) at which the difference in the \log_2 (insulation index) between WT c14 and *zld* knockdown was ≥ 0.3 .

We defined differential boundaries that were closer than 125/62.5 kb to the center of a structurally changing region as captured by CHES.

Detecting structural changes between healthy B cells and a patient with DLBCL.

We obtained Hi-C data from Díaz et al.⁴⁸ and processed them as described in the original publication. We partitioned the human hg19 genome into 2-Mb regions with a step size of 500 kb. We used CHES to compare the corresponding regions in the obs/exp-transformed Hi-C data from healthy B cells (Ctrl) and a patient with DLBCL. To distinguish between actual structural differences and those attributable to noise we calculated a signal-to-noise ratio r for the differential signal of each matrix pair:

$$r = \frac{\mu_M}{\sigma_M^2}, M = M_{\text{control}} - M_{\text{patient}}$$

This was done for a sliding window of 7×7 pixels on the matrix. The total signal-to-noise ratio was taken as the mean of all windows. Regions with a z -normalized similarity score ≤ -1.2 and a signal-to-noise ratio of $r \geq 0.6$ were labeled and accepted as changing.

Feature extraction from Capture-C data. The CHES feature extraction was applied to the Capture-C experiments from Despang et al.⁵⁰ (GSE125294). Interaction matrices normalized by the KR balancing method were downloaded. All mutants were compared to WT. All the differential features were extracted and clustered according to their interaction pattern (The CHES pipeline). Three structural clusters were obtained: TAD; loop; and stripe. Some examples are shown in Fig. 6 and Extended Data Fig. 8.

Statistics. The following statistical tests were used. In Fig. 3a we tested whether syntenic regions were structurally more similar than expected by chance using a one-sided randomization test with 100 permutations of pairwise syntenic region assignments. For the analyses in Figs. 2, 3 and Extended Data Figs. 1, 4 and 7 we tested whether a particular matrix A was more similar to another particular matrix B than to other 1,000 artificially generated matrices (Fig. 2 and Extended Data Figs. 1 and 7) or to all other matrices along the diagonal of B's whole chromosome matrix of the same size as A (Fig. 3 and Extended Data Fig. 4). The specific details of each of these tests are described in the Methods sections The CHES pipeline and Generation of synthetic Hi-C matrices.

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

Data availability

The datasets analyzed in this study have been obtained from the Gene Expression Omnibus (Rao et al.¹⁰, GSE63525; Bonev et al.¹², GSE96107; Despang et al.⁵⁰, GSE125294) and ArrayExpress (Hug et al.⁹, E-MTAB-4918; Díaz et al.⁴⁸, E-MTAB-5875).

Code availability

The CHES source code and the code for generating the synthetic Hi-C matrices and running tests on them is available on GitHub: (<https://github.com/vaquerizaslab/CHES>). The intervaltree and tqdm packages used internally in CHES can be found at <https://github.com/chaimleib/intervaltree> and <https://github.com/tqdm/tqdm>, respectively. In addition, CHES uses internally the following published packages: FAN-C⁷¹ (<https://github.com/vaquerizaslab/fanc>); Cython⁷²; SciPy⁶⁹; scikit-image⁶⁸; NumPy^{73,74}; Pandas⁷⁵; Pathos⁷⁶; Pybedtools⁷⁷; Knead⁷⁸.

References

- Imakaev, M. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
- Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* **33**, 1029–1047 (2013).
- van der Walt, S. et al. scikit-image: image processing in Python. *PeerJ*. **2**, e453 (2014).
- Behara, K. N. S., Bhaskar, A. & Chung, E. Geographical window based structural similarity index for OD matrices comparison. *J. Intell. Transp. Syst.*, <https://doi.org/10.1080/15472450.2020.1795651> (2020).
- Djukic, T., Hoogendoorn, S. & Van Lint, H. Reliability assessment of dynamic OD estimation methods based on structural similarity index. In *Proc. Transportation Research Board 92nd Annual Meeting* (Transportation Research Board, 2013).
- Breakey, D. & Meskell, C. Comparison of metrics for the evaluation of similarity in acoustic pressure signals. *J. Sound Vib.* **332**, 3605–3609 (2013).
- Hines, A. & Harte, N. Speech intelligibility prediction using a Neurogram Similarity Index Measure. *Speech Commun.* **54**, 306–320 (2012).
- Tomasi, C. & Manduchi, R. Bilateral filtering for gray and color images. In *Proc. Sixth International Conference on Computer Vision* 839–846 (IEEE, 1998).
- Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern. A Syst. Hum.* **9**, 62–66 (1979).
- Sexton, T. et al. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148**, 458–472 (2012).
- Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
- Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R. & Mozziconacci, J. Normalization of a chromosomal contact map. *BMC Genomics* **13**, 436 (2012).
- Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
- Blythe, S. A. & Wieschaus, E. F. Zygotic genome activation triggers the DNA replication checkpoint at the midblastula transition. *Cell* **160**, 1169–1181 (2015).
- Kruse, K., Hug, C. B. & Vaquerizas, J. M. FAN-C: a feature-rich framework for the analysis and visualisation of C data. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.02.03.932517> (2020).
- Behnel, S. et al. Cython: the best of both worlds. *Comput. Sci. Eng.* **13**, 31–39 (2011).
- Oliphant, T. E. A *Guide to NumPy* (Trelgol Publishing, 2006).
- 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).
- McKinney, W. Data structures for statistical computing in Python. In *Proc. Python in Science Conference* 56–61 (ScyPy.org, 2010).
- McKerns, M. M., Strand, L., Sullivan, T., Fang, A. & Aivazis, M. A. G. Building a framework for predictive science. Preprint at <https://arxiv.org/abs/1202.1056> (2012).
- Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424 (2011).
- Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a ‘Kneedle’ in a haystack: detecting knee points in system behavior. In *Proc. 31st International Conference on Distributed Computing Systems Workshops* 166–171 (IEEE Computer Society, 2011).

Acknowledgements

Work in the Vaquerizas laboratory is funded by the Max Planck Society, the Deutsche Forschungsgemeinschaft (DFG) Priority Programme SPP 2202 'Spatial Genome Architecture in Development and Disease' (project no. 422857230 to J.M.V.), the DFG Clinical Research Unit CRU326 'Male Germ Cells: from Genes to Function' (project no. 329621271 to J.M.V.), the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 643062—ZENCODE-ITN to J.M.V.) and the Medical Research Council in the UK. This research was partially funded by the European Union's H2020 Framework Programme through the European Research Council (grant no. 609989 to M.A.M.-R.). We thank the support of the Spanish Ministerio de Ciencia, Innovación y Universidades through grant no. BFU2017-85926-P to M.A.M.-R. The Centre for Genomic Regulation thanks the support of the Ministerio de Ciencia, Innovación y Universidades to the European Molecular Biology Laboratory partnership, the 'Centro de Excelencia Severo Ochoa 2013–2017', agreement no. SEV-2012-0208, the CERCA Programme/Generalitat de Catalunya, Spanish Ministerio de Ciencia, Innovación y Universidades through the Instituto de Salud Carlos III, the Generalitat de Catalunya through the Departament de Salut and Departament d'Empresa i Coneixement and cofinancing by the Spanish Ministerio de Ciencia, Innovación y Universidades with funds from the European Regional Development Fund corresponding to the 2014–2020 Smart Growth Operating Program. S.G. thanks the support from the Company of Biologists (grant

no. JCSTF181158) and the European Molecular Biology Organization Short-Term Fellowship programme.

Author contributions

N.M. and J.M.V. conceptualized the study. S.G., N.M. and K.K. devised the methodology. N.M. and J.M.V. carried out the investigation. S.G., K.K. and N.D. obtained the resources. S.G., N.M., K.K., M.A.M.-R. and J.M.V. prepared and wrote the original draft of the manuscript. S.G., N.M., K.K., N.D., M.A.M.-R. and J.M.V. wrote, reviewed and edited the draft. J.M.V. supervised the study. M.A.M.-R. and J.M.V. acquired the funding.

Competing interests

The authors declare no competing interests.

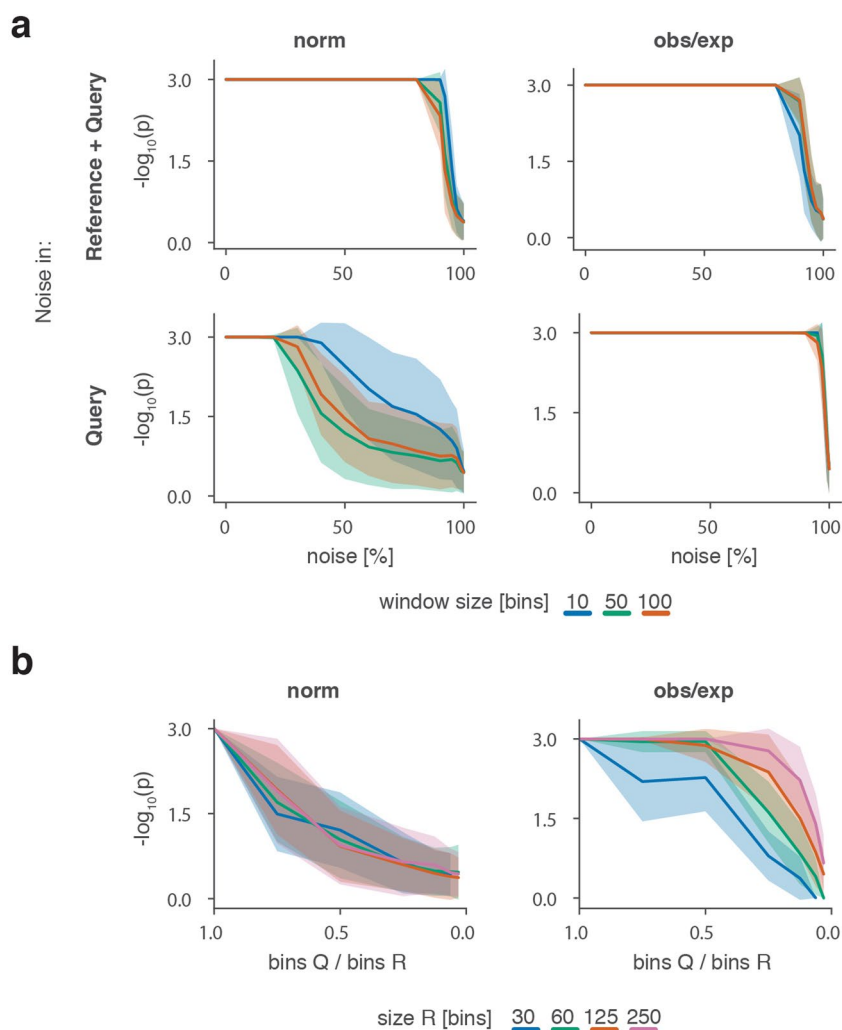
Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41588-020-00712-y>.

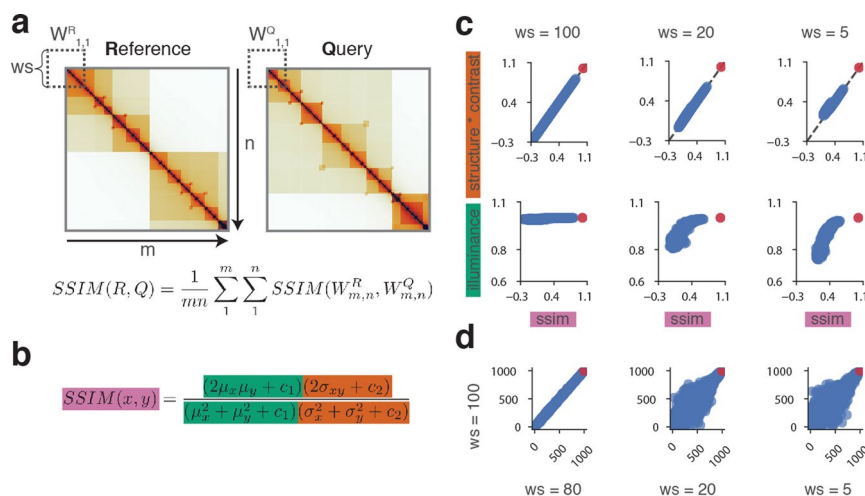
Supplementary information is available for this paper at <https://doi.org/10.1038/s41588-020-00712-y>.

Correspondence and requests for materials should be addressed to J.M.V.

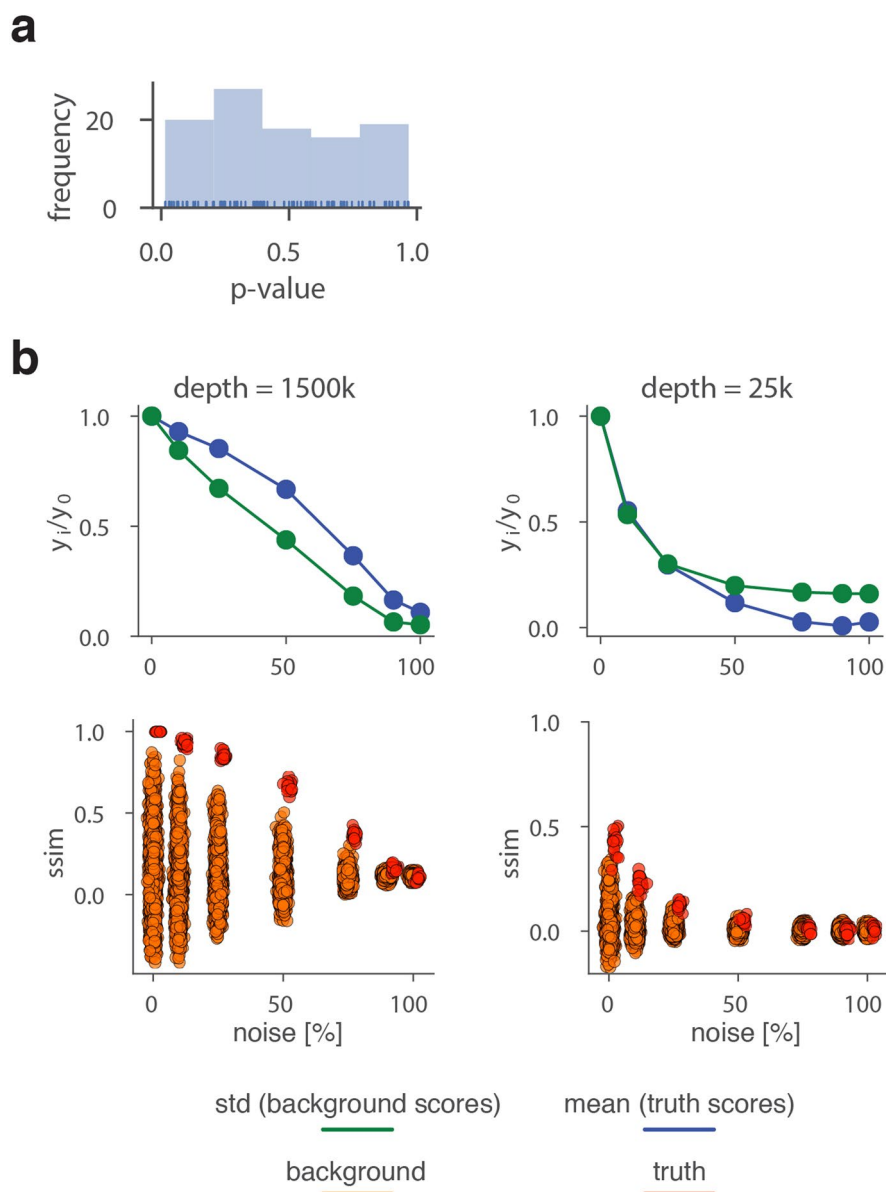
Reprints and permissions information is available at www.nature.com/reprints.



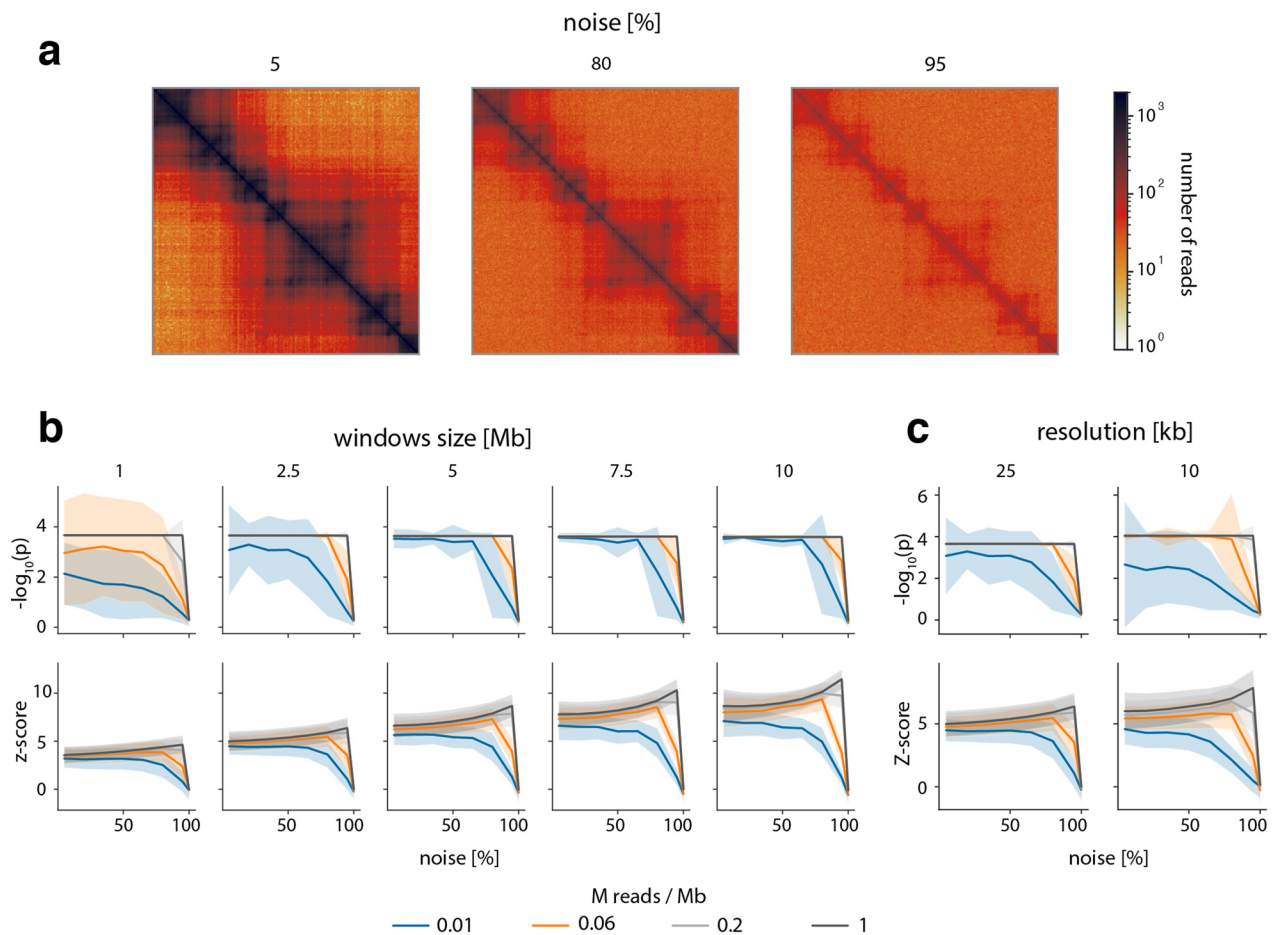
Extended Data Fig. 1 | Performance analysis of the CHES algorithm. a, CHES P values in dependence of the relative noise level in synthetic matrices. Shown are the cases of equal amounts of noise in reference R and query Q (top) and different amounts of noise (bottom, noise only added to Q). Each case is examined for normalised and observed/expected (obs/exp) matrices, and different window sizes in the SSIM algorithm. **b**, Empirically determined CHES P values in dependence of the size factor between R and Q for normalised (left) and observed/expected (obs/exp) matrices (right) (details in Methods). **a**, **b**, Solid lines indicate the mean, shaded areas the standard deviation over 100 simulations per parameter combination.



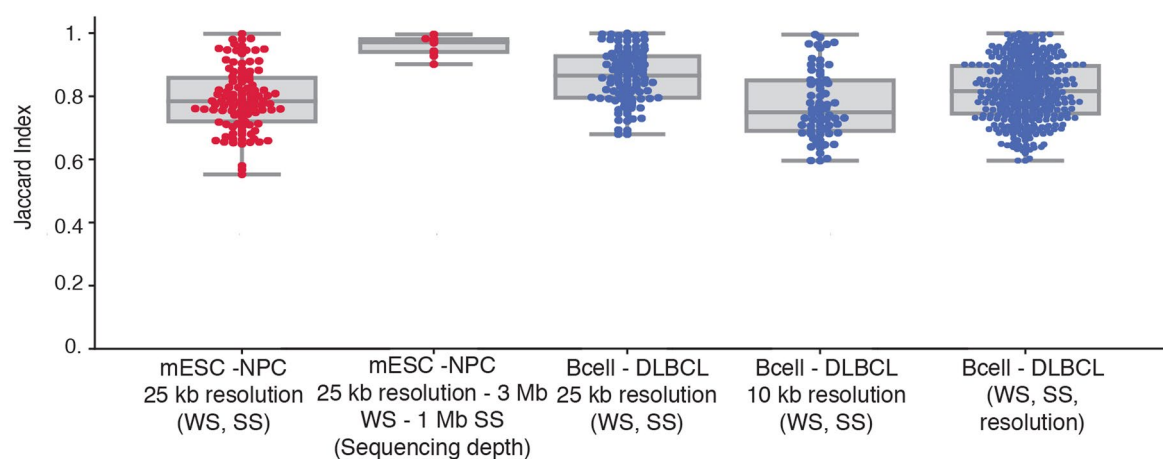
Extended Data Fig. 2 | Technical details of the SSIM algorithm applied to Hi-C matrices. **a**, Schematic overview of the structural similarity algorithm (SSIM). SSIM scores are calculated on all submatrices of R/Q at a given window size (WS). The final SSIM score is the mean of all SSIM submatrix scores. **b**, SSIM submatrix formula. Different components are coloured: illuminance (green), structure * contrast (red). x, y refer to submatrices (at the same positions) of the two full matrices for which the SSIM average is computed (see panel a). μ indicates the mean, σ the standard deviation, c_1 and c_2 are small constants that are introduced only for numerical reasons. **c** and **d**, SSIM comparisons of a matrix to itself (red dots) and 1,000 random matrices of the same size (blue dots). **c**, SSIM component values in dependence of SSIM score for different SSIM window sizes. **d**, Scatterplots of ranked SSIM scores at window size 100 vs. ranked scores at smaller window sizes.



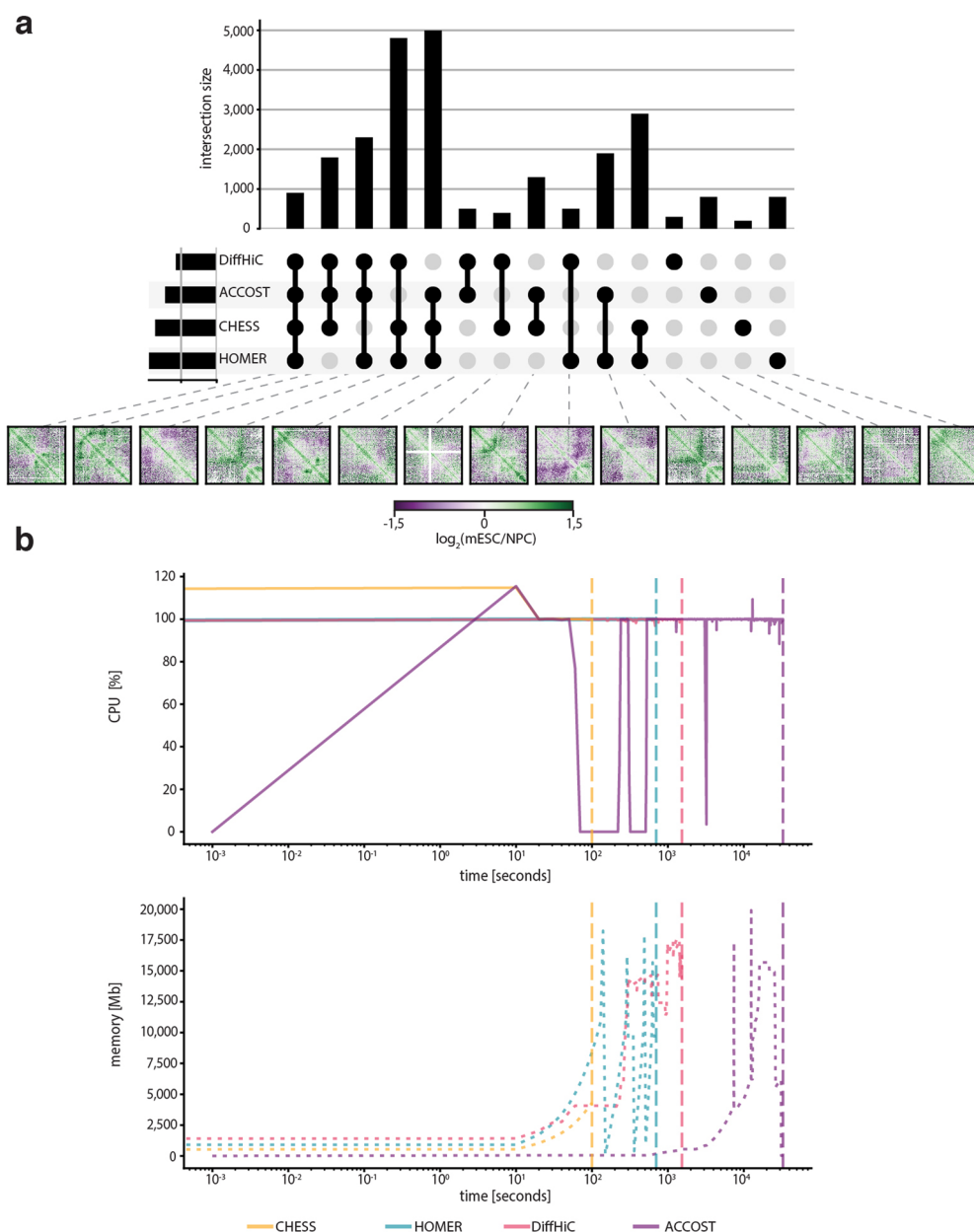
Extended Data Fig. 3 | Additional analysis of the CHES algorithm. **a**, Uniform distribution of empirically determined CHES P values for comparisons of matrices with 100 % noise added. **b**, Distribution of structural similarity scores (SSIM) for background and truth comparisons at 25 k/Mb and 1.5 M/Mb simulated sequencing depth. Above each: Fractional change (value at x % noise/value at 0 % noise) of the standard deviation (std) of background scores and mean of truth scores over 100 simulations per parameter combination.



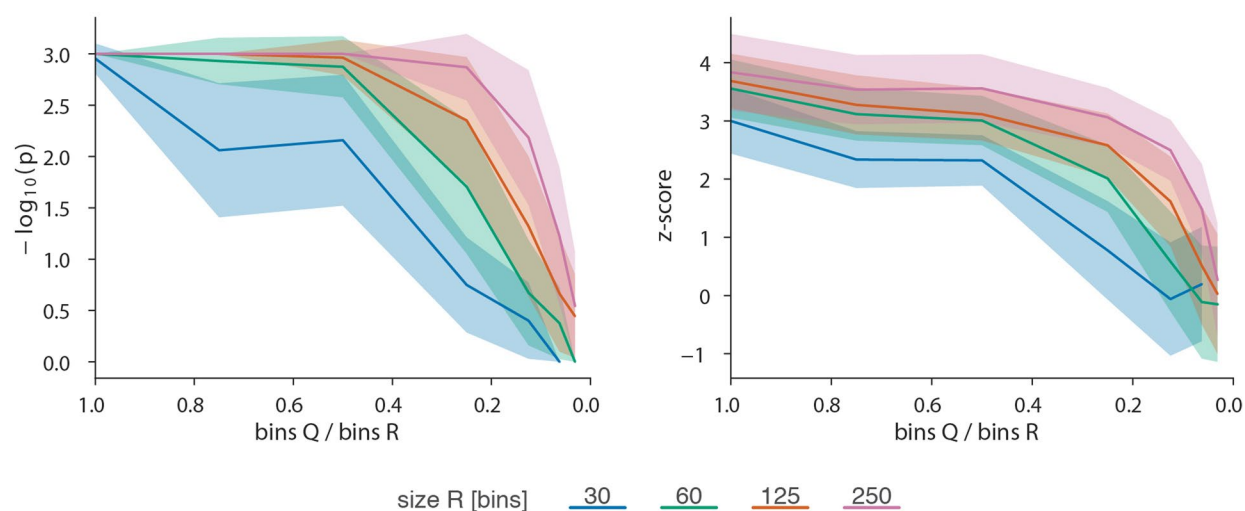
Extended Data Fig. 4 | CHES is robust to changes in noise due to random ligations and sequencing depth in real Hi-C data. a, Examples of 5 Mb matrices used in this analysis including a 5, 80 and 95 % of added noise (random ligations between pairs of loci). We tested to what extent CHES is able to identify two matrices as being identical, after noise and sequencing depth were adjusted independently in them. Matrices are based on chromosome 19 data from Bonev et al.¹². **a**, examples of the data with different amounts of noise. **b**, empirically determined P values and z -scores of CHES runs with different window sizes, noise levels and simulated sequencing depths (details in Methods). Step size and matrix resolution were both 25 kb. Lines for 2×10^5 and 1×10^6 overlap for runs with window sizes > 1 Mb. **c**, As in panel a, but comparing CHES runs with 2.5 Mb window size on matrices binned at 25 kb and 10 kb. **b** and **c**, solid lines indicate the mean, shaded areas the standard deviation over 1976, 2066, 2156, 2246, 2300 matrix pairs for window sizes 10 Mb, 7.5 Mb, 5 Mb, 2.5 Mb, 1 Mb, respectively.



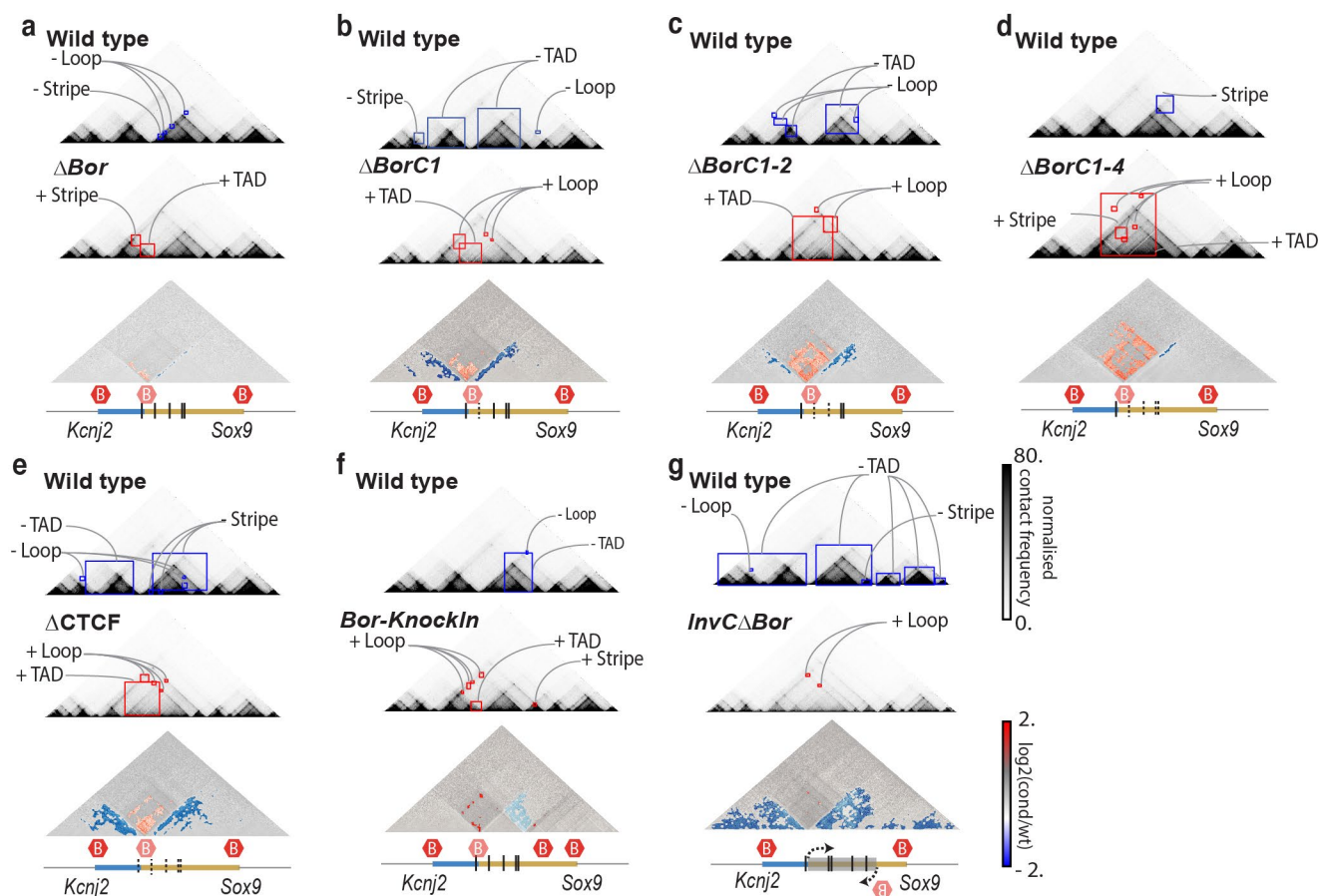
Extended Data Fig. 5 | Reproducibility of CHES using different window (WS) and step sizes (SS), sequencing depths and resolutions. For this analysis were tested the WS (250 kb - 3 Mb), SS (25 kb - 1 Mb), sequencing depths (percentage of reads between 20 and 80) and resolutions (10 kb and 25 kb) (details in Methods). X-axis labels: varied parameters in parentheses, fixed parameters before. The first two boxplots with red dots represent the Jaccard indices (JI) between CHES results in Bonev et al.¹² using different WS, SS and sequencing depths. The boxplots with blue dots correspond to the Díaz et al.⁴⁸ dataset; in this case using different WS, SS, and then between different WS, SS and resolutions. mESC mouse embryonic stem cells, NPC neural progenitor cells. Boxplot elements: centre line: median, whiskers: 1.5x interquartile range, box limits: upper-lower quartile.



Extended Data Fig. 6 | CHES benchmark against HOMER, diffHiC and ACCOST. **a**, Upset plot representing the intersection size between differential interactions of CHES, HOMER, diffHiC and ACCOST. Below, an example is shown for each intersected group. **b**, Computational requirements of CHES, HOMER, diffHiC and ACCOST. The first line plot shows the CPU usage, the second the memory consumption. The vertical dashed line represents the end of the run.



Extended Data Fig. 7 | CHES performance on differently sized simulated matrices with realistic noise and sequencing depth. Shown are empirically determined CHES p- and z-scores (details in Methods) for comparisons of R with a read depth of 100 read pairs / 100 bins and a resized copy Q . Scaling factor is indicated on the x-axis. A noise level of 25 % was added to both matrices independently. Sequencing depth was adjusted to 100 k/Mb. Solid lines indicate the mean, shaded areas the standard deviation over 100 simulations per parameter combination. Colours correspond to the different sizes of R .



Extended Data Fig. 8 | Feature extraction from Capture-C data. Examples of differential feature extraction with CHESS between the wt (top contact map) and different mutants (middle contact map) in the Despang et al.⁵⁰ dataset. Lost and gained structures in the mutants are highlighted in blue and red squares, respectively. \log_2 fold-change maps are depicted below (bottom contact map) with identified features coloured according to the directionality of the change. Below each comparison, the genomic annotation is represented, highlighting the modification of each mutant. The vertical lines define the CTCF binding motifs, dashed when deleted. Red hexagons demarcate TAD boundaries. Feature extraction between wt and **a**, ΔBor , in which the border was deleted. **b**, $\Delta BorC1$, in which the border and the first CTCF binding motif were deleted. **c**, $\Delta BorC1-2$, in which the border and the two first CTCF binding motifs were deleted. **d**, $\Delta BorC1-4$, in which the border and four CTCF binding motifs were deleted. **e**, $\Delta CTCF$, in which the border and all the CTCF binding motifs were removed. **f**, *Bor-KnockIn*, in which the border was moved to a new location within the *Sox9* locus. **g**, *InvC ΔBor* , in which the *Sox9* sequence was inverted and the border was removed.

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 |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> 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) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> 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	No software was used for data collection.
Data analysis	Python packages: python (3.6, 3.7), cython (0.29.16), scipy (1.0.0), numpy (1.14.0), scikit-image (0.13.1), pandas (0.22.0), pathos (0.2.1), kneed (0.6.0), tqdm (4.43.0), intervaltree (3.0.2), pybedtools (0.8.1) and future (0.16.0), fanc (0.8) Hi-C analysis: Biopython (1.71), Bowtie 2.2.4, BWA mem 0.7.17-r1188 CHES: https://github.com/vaquerizaslab/chess

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

The datasets analysed in this study have been obtained from Gene Expression Omnibus (GEO; Rao, et. al., 2014: GSE63525; Bonev, et. al., 2017: GSE96107; Deshpande, et. al., 2019: GSE125294) and ArrayExpress (Hug, et. al., 2017: E-MTAB-4918; Díaz, et. al., 2018: E-MTAB-5875).

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](https://www.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	Sample sizes were not predetermined with a statistical method, but chosen according to standards of the field, where appropriate.
Data exclusions	No data were excluded from the analyses.
Replication	In order to ensure proper performance of our method, we have tested it on multiple independent datasets, as well as synthetic datasets with 100 independent replications per parameter combination, as specified in detail in the Methods section.
Randomization	To calculate statistical significance for the degree of conservation of syntenic regions, we computed similarity scores for 100 random permutations of syntenic region pairs.
Blinding	The investigators were not blinded with respect to group allocation. Knowledge of group labels was necessary in order to carry out the research.

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

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging