nature portfolio

Peer Review File

Inter-chromosomal contacts demarcate genome topology along a spatial gradient

Corresponding Author: Professor Philipp Maass

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
In this manuscript, the authors develop a new computational tool named Signature, which aims to detect inter-chromosomal structures from Hi-C data. Higher-order DNA structures formed by regions on different chromosomes and their function are certainly less explored in the field of genome organization; hence, the development of such a tool is a timely and important task. The authors validate the detection of multiple already described structures from inter-chromosomal interactions and demonstrate that these are indeed present in different cell types (e.g., speckle contacts, Rabl configuration). However, I have trouble pinpointing the main discovery of the manuscript, and thus I have certain reservations about the presented work.

Major:

- The authors provide the rationale that Hi-C does not detect inter-chromosomal contacts due to technical limitations. What is the rationale for developing a machine learning tool based on that data rather than data that captures inter-chromosomal interactions to a higher extent, such as GAM, SPRITE, or PoreC? This needs to be addressed.
- What are the technical limitations of Hi-C, and what type of conclusions cannot or should not be inferred from such data? Authors should discuss this clearly.
- "Topological anchoring community" regions sound simply like nuclear speckles. Authors suggest they are something different, but it is not presented clearly why that is the case and how they differ from speckles. This needs to be explained and addressed. Also, are the differences across different cell types regarding them simply related to different regions coming closer and further from speckles as their expression patterns change? That should be addressed.
- What are the novel inter-chromosomal interactions that authors identify? Are there any contacts that authors detected, which do not belong to any already known structures? Authors mention constitutive and specific "anchored loci" but do not provide their identity, nor explain if these are earlier mentioned speckle regions. Again, this needs to be clarified. If they are indeed new regions, these should be validated by FISH, and their clustering and change in localization should be presented. Moreover, any potential functions they might be playing would be of high interest for the manuscript.

Minor:

- Figures are extremely busy and not reader-friendly (in several going up to m or further); numbering often changes in the text, and what is presented is not well described in figure legends. It is very difficult to extract important information from Figures, and I highly recommend modifying them to ensure that readers can appreciate the presented work.
- Authors propose the term "non-homologous chromosomal contacts" to describe inter-chromosomal contacts. I think we should all use similar terms to describe certain observations and not come up with new acronyms when they are not needed. If there is a difference, it is not well described, as I did not understand it.

Few missing citations and works that should be discussed:

- https://www.nature.com/articles/s41467-022-32980-z)
- https://www.nature.com/articles/s41592-021-01135-1)

Reviewer #3

(Remarks to the Author)
In this study, the authors introduce Signature, a machine learning method for identifying NHCCs (non-homologous chromosomal contacts) in Hi-C data. They analyze a large dataset of Hi-C samples from various human cell types / samples. Their analysis reveals a varied landscape of NHCCs, with non-random distribution and association with active regions. Additionally, it identifies cell-type-specific and sex-specific NHCCs, potentially related to their functions and sex-based features. While this work represents a relevant contribution to understanding the role of NHCCs in genome organization and function, I have a few questions and comments.

Could the authors expand the discussion on how NHCCs are deterministic? NHCCs might be stochastic and transient, potentially influenced by factors like transcriptional activity and cell cycle stage, which may be the factors that lead to being more deterministic as well.
Could the authors expand the comments on the Rabl configuration in human genome organization? Based on ref. 50 (Hoencamp, C. et al.). This configuration is often considered more prevalent in organisms like yeast and file based on the presence of the complete condensin II subunits. In addition, the centromere and telomeres polarization or cluster in opposite directions may partially distort the territories.
One complementary study to ref 74 that could be mentioned by the authors (Contessoto, V et al. Nat. Comm, 2023) shows that the formation of helices in interphase and compaction are related to the Rabl configuration.

Reviewer #5

(Remarks to the Author)
Mokhtaridoost et al. introduce ‘Signature’, a pipeline for identifying non-homologous chromosomal contacts (NHCCs). Signature identifies NHCCs through LOESS regression to flag outliers and uses community detection methods to cluster and embed chromosomal bins in an interpretable manner. The authors conclude that NHCCs are strongly associated with gene expression, telomeric and centromeric clustering occurs pervasively in human cells, and that constitutive NHCCs ground cell type specific gene activity.

Signature would be the first pipeline for studying inter-chromosomal contacts, and to date, this area has remained relatively unexplored due to conflicting reports on the existence of such contacts. Their findings re-capitulate many well-known examples of chromosomal organization, such as previously identified NHCCs, Rabl’s configuration, and speckle associated contacts. While the paper has strong conceptual and biological appeal, the authors occasionally overstate their contributions, overinterpret their results, and fail to include clear methods for their results.

Nonetheless, Mokhtaridoost et al.’s paper is a promising first step into the exploration of NHCCs. This is an underappreciated and unexplored topic in the field of 3D genomics, and recent developments in this area indicate a broader interest in NHCCs. This work would augment existing analysis methodologies for Hi-C assays and stimulate the further development of approaches aimed at analyzing NHCCs and inter-chromosomal contacts in general.

Major Comments:
1. Line 62 – The authors cite multiple sources claiming that NHCCs are not readily detectable in Hi-C data. However, it is unclear how all of these sources support that claim.
a. Source 17 (HiC-DC+) is exclusively concerned with intra-chromosomal contact identification.
b. Source 18 (Higashi) is a sc-HiC method for imputing sc-HiC data with a hypergraph representation and is not focused on chromosomal contact identification, though the authors acknowledge the difficulty of using inter-chromosomal contacts.

2. Line 73 – Signature can assess intra- and inter-chromosomal interactions, but their assessment of this capability appears sparse. At the minimum, some comparisons against extant intra-chromosomal contact callers and statistics on their enrichment above the local background via APA would be helpful.

3. Line 94 – Weights are discussed, but at this point in the manuscript, what “weights” is referring to, is unclear. Throughout the remainder of the manuscript, the authors appear to flip between interaction weights/interaction frequencies; I am almost sure that these are referring to the same object. If not, please clarify.

4. Line 94 – The authors say they developed a non-parametric supervised learning approach called Locally Weighted Linear Regression. I have two minor objections to this.
a. LOESS – Their method calls the LOESS function in R (loess.sd in msir calls loess) to accomplish the non-parametric smoothing of the data.
b. Linear – This is not true. The default degree of LOESS yields local quadratic smoothing, and the authors have not specified d=1 in the calls to LOESS (or used LOWESS) and thus, they are performing a locally weighted polynomial (quadratic) regression.

5. Line 634 – Clarify if a one sided or two-sided z-test was used.

6. Line 107 – References Figure 1A, but lack of figure annotations makes it unclear what is being plotted in the right-hand side of the first facet. Is it the z-score of the interaction after removing the background?

7. Line 111 – Can the authors clarify the meaning of “how”.

8. Line 123 – Many non-interacting regions are proposed to exist, but this is briefly mentioned in passing. Have the authors explored the possible biological functions of these non-interacting regions?

9. Line 126 – There are 46 chromosomes, each corresponding to the paternal and maternal copies, but as far as I understand, an alignment would need to be performed with that goal in mind, whereas the authors have only performed an alignment against the reference genome.
a. Fig 1e is very suggestive, but the description of the methods for visualizing this are not clear.
b. After community detection into 46 communities, what does it mean to cluster interactive bins? Is the linking of consecutive bins what generates the chromosomal outlines in Fig 1e?
c. Based on the description of the ForceAtlas 2 usage and Gephi, it seems like since there are 46 communities, and between community interactions are excluded, it seems there should be quite a few distinct communities that number more than the 23 chromosomes present in Fig 1e? What are these?

10. Line 666 – Intercommunity interactions are excluded from the ForceAtlas2 layout used to visualize the CD results, so I am confused what that implies about the clustering of the centromeres and telomeres?
a. Lines 204-215 – This section claims that since inter-telomeric and inter-centromeric NHCCs occur together, this is evidence for Rabl's configuration, but the methods indicate that interactions (weights?) between communities (chromosomes?) are omitted. So telomeric interactions across chromosomes would presumably be omitted as well since it is demonstrated that the chromosomes appear to form distinct strands in the visualization. Could the authors clarify the methodology which justifies their conclusions in this section.

11. Line 166 – It is claimed that the NHCC frequencies do not depend on the interactions within chromosomes are not related to the ones for NHCCs, but prima facie I was expecting a negative correlation.
a. If a population of cells makes more NHCCs, then it does so at the expense of intra-chromosomal contacts.
b. Extended Data Fig. 3f is supposed to support this point, and while the coefficient of correlation appears small, wouldn't the overall negative slope suggest an inverse relationship between NHCC formation and intra-chromosomal contacts? Further, NHCC z-scores are plotted, and this includes z-scores of 0, which would not yield a significant p-value? Were all inter-chromosomal z-scores plotted because I thought that NHCCs were flagged as a positive z-score with $q < 0.05$.

12. Line 172 – To clarify, when it says a span of 1.84 Mb on average, that means that the total length of interacting NHCCs (consecutive 1 Mb) bins making up that domain averages out to 1.84 Mb?
a. For example: In case 2, when you have (consecutive?) bins with one bin in the middle interact with different single bins on chromosome B. This is only computed as 2 Mb, and not 3 Mb. Further, the bins on chromosome B do not factor into this.
b. Also, are there cases where bins are assigned to multiple domains? If so, how is this handled?
c. Could the domain definition be relegated to the methods instead of the extended data section?

13. Line 174 – Why are domains that are detectable at 1.84 Mb exclude the ones at 3.38 Mb?

14. Line 192 – Could the authors comment on how well mapped the centromeric and telomeric regions are as highly repetitive regions of the genome generally are difficult to map with short reads, and whether this affects the performance of Signature.

15. Line 223 – For GTEx analysis, are all the counts across the various tissues just being averaged after transformation?
a. Is the justification for this choice being that the cell types being analyzed are being pooled from various tissues?
b. If that is the case, I can appreciate why the authors sought out the GTEx for their analysis, but it would have been helpful to study NHCCs comprehensively in a selected cell type to see if their inferences across the entirety of their dataset held for their selected cell type.

16. Line 233 – This section explores the correlations between NHCCs and underlying genomic properties, but this does not imply the NHCCs support the gradient.

17. Line 282 – The randomization procedure discussed in the methods is unclear to me. Is it the following?
a. 40282 total NHCCs, 23351 (discrepancy between text and methods on this) unique
b. On each randomization, draw 40282 NHCCs (with replacement?)
c. What does it mean to check the percentage that is unique on each draw?

18. Line 340-341 – The asymmetric spatial genome gradient that is demonstrated in the paper and proposed is identified from the global embedding after pooling all the individual maps. In the absence of the individual embeddings and characterizations, this finding could be a fallacy of division.

19. Line 565 – HiC Pro was compared to the 4DN protocol for mapping.
a. I am surprised to see mapping rates for the RPE1 datasets being that low. HiC Pro uses a global alignment (bowtie2 end-to-end) rather than the local alignment (bwa mem) of the 4DN protocol. Were there issues with the raw data fastqs (adapter contaminants, low quality tiles) that prohibited proper end-to-end alignment?
b. Further, bwa mem allows for multi mappers to exist, but they are assigned a mapping quality of 0. A previous work from Johanson et al 2018 Plos Genetics indicated that when they identified NHCCs using a local enrichment measurement, blacklisted and hard to map regions made up many their NHCCs. Are the authors aware if this is also an issue for the NHCCs they have identified using Signature?

20. Line 759 – Related to earlier comments, please make obvious that you considered the fact that the repetitive sequences in the telomeric/centromeric regions are unmappable.

21. Extended Data Fig 3. A) – A few questions
a. Does each point correspond to one of the 62 HiC maps?
b. How is Cooler being used here for the NHCC step? It it their implementation of HiCCUPS but on the non-homologous chromosomal pairs? If that is the case, I am surprised that the called percentage of NHCCs is so high. Also, I am surprised that an interaction caller using a distance function yields a similar number of calls to Cooler; the methods section for the intra-chromosomal interaction caller for Signature is sparsely described.
c. In general, I think this figure needs a more extensive description of what is being shown.

22. Extended Data Fig 3. B) – Why is it titled NHCCs over seq. depth? Again, I presume that the percentage of NHCCs identified by Cooler and Signature are what is being plotted, and each of the dots corresponds to the different data sets used.

Minor Comments:
1. Line 58 – The authors reference limited software tools for detecting NHCCs, but to my knowledge, this would be the first tool explicitly designed for detecting NHCCs.

2. Line 63 – "Hi-C … captures NHCCs …", but in Line 61-62, "NHCCs have been considered as … not readily detectable in Hi-C data", I think there is a point of distinction to be made: Hi-C contains NHCC data, by virtue that there are counts mapping to trans contacts, but the issue of detection is another thing entirely, which is what I think 62 is discussing.

3. Line 75 – What are these orthogonal approaches?

4. Line 93 – "assumed a direct relationship of spatial proximity and inter-chromosomal …", I think the authors are pre-empting their justification of the background model they fit. Their background model is fitted against the genomic distance variable in Fig 1A, but it is simpler to just say use the term "genomic coordinate". In contrast to intra-chromosomal contacts, where distance can be clearly defined with respect to the distance between two interaction anchors, in inter-chromosomal contacts, distance as a concept is not the same.

5. Line 101 – Fig 1a, should be

6. Line 102 – "We cross-validated the span…".

7. Line 589 – 590 – To make the process computationally feasible, "LWLR retains data from local bins …, rather than storing information for all regions, as performed in linear regression." This is a well-known fact about LOESS.

8. Line 625 – A kind of weighted mean is being used, but it'd be more accurate to say just that the LOESS fitted values are being used?

9. Line 631 – The stated procedure for choosing the z-score can be justified as a minimum p-value pooling procedure.

10. Line 632 – There are obviously large deviations away from the normal distribution, but isn't that to be expected if there are true events that do not correspond to the null? So,

11. Line 135 – Please supply CDFs of the Merfish distances classified by NHCC vs non-interacting regions. It is difficult to appreciate the reported differences from the bar charts alone.

12. Line 137-138 – Reported overlaps are really the recall rate? (116/130~90%,72/164~40%)

13. Line 143-145 – Fig 1j, Is there a way to make the blocks called significant obvious?

14. Line 184 – For ease, could just mention in the figure legends that the expected contacts was based on chromosomal length.

15. Line 650 – Why can't interdependent data be clustered? This claim would imply that gene expression data should not be clustered because of dependencies between genes.

16. Line 221 – Can the authors assess whether these findings hold true when they have matched RNA-seq data for which they can assess if NHCCs harbor most of the active genes?

17. Line 227 – Could the authors define tissue specific NHCCs in line?

Reviewer #6

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
In the revised version of the manuscript authors addressed and clarified several of my previous comments which helped with the clarity and understanding of the manuscript. I still find it very cluttered and not easy to follow, but inter-chromosomal interactions are not much explored by the community and this work will provide useful tool. However, one of the points are still not clear to me.
Specifically, my questions regarding novel detected structures and general usage of the developed software is not addressed. I was asking about novel structures that might have potential functions and not contacts that were not detected by other methods (that is what revised version and rebuttal is referring to). For example, do authors see new higher-order structures that suggest existence with novel nuclear bodies not described previously? I think it is an important point that will help to appreciate potential power of this tool.

Reviewer #3

(Remarks to the Author)
The authors addressed all my questions.

Reviewer #5

(Remarks to the Author)
The authors have addressed all major points from the previous review. Major methodological questions have been sufficiently addressed, and the illustration and further descriptions regarding their clustering procedure and “domain” construction procedure are appreciated. Further investigations into NHCCs in the context of specific tissues and cell types were appreciated as well.

Minor:
For Figure 1A, instead of an epsilon symbol, a standard set membership symbol could be used.

Reviewer #6

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
In the revised versions authors aimed to address my previous comment regarding novel nuclear compartments detected by their method. I appreciate the efforts and addition of the new observations that I believe omprove the manuscripts.

REVIEWER COMMENTS

We would like to thank the reviewers for dedicating their time to review our study and for providing constructive feedback, which proved to be very valuable. Following the reviewer comments, we have revised the manuscript by adding new analysis, by rephrasing the text, and by implementing the required modifications, as elaborated below.

Reviewer #2 (Remarks to the Author):

In this manuscript, the authors develop a new computational tool named Signature, which aims to detect inter-chromosomal structures from Hi-C data. Higher-order DNA structures formed by regions on different chromosomes and their function are certainly less explored in the field of genome organization; hence, the development of such a tool is a timely and important task. The authors validate the detection of multiple already described structures from inter-chromosomal interactions and demonstrate that these are indeed present in different cell types (e.g., speckle contacts, Rabl configuration). However, I have trouble pinpointing the main discovery of the manuscript, and thus I have certain reservations about the presented work.

We would like to thank the reviewer for the evaluation of our work and for the helpful comments on our manuscript. We have addressed all raised concerns and have edited our current draft. Please find the details below.

Major:

- The authors provide the rationale that Hi-C does not detect inter-chromosomal contacts due to technical limitations. What is the rationale for developing a machine learning tool based on that data rather than data that captures inter-chromosomal interactions to a higher extent, such as GAM, SPRITE, or PoreC? This needs to be addressed.

We thank the reviewer for the comment. Hi-C is the most widely used technique to study 3D genome organization and it captures NHCCs. Orthogonal approaches, such as GAM, SPRITE and imaging techniques also detect NHCCs, but these methodologies harbor various technical intricacies, require further resources, their depth is not comparable to Hi-C, since they are of lower throughput than Hi-C. These were the main rationales to develop Signature for ‘C-based’ applications. We rephrased our rationale in the introduction. The text reads as follows:

‘Two commonly used approaches for investigating NHCCs are imaging and chromatin capture, both of which are limited in determining NHCCs. Specifically, imaging is not scalable to genome-wide approaches and chromosome conformation capture (i.e., proximity ligation-based Hi-C) as the most widely used technique to study 3D genome organization mainly focuses on analyzing intra -chrom osomal contacts (Lieberman-Aiden, van Berkum et al. 2009, Dekker, Alber et al. 2023). Moreover, both methodologies often

caused discordant results when studying NHCCs that do not complement one another (Dekker 2016, Maass, Barutcu et al. 2019, Payne, Chiang et al. 2021). Importantly, Hi-C datasets contain 'trans-reads', but current computational and statistical analysis has limitations in confidently determining true NHCCs. Hence, NHCCs have been considered as stochastic, singular events (Bashkirova and Lomvardas 2019, Maass, Barutcu et al. 2019), that are not readily detectable in Hi-C data (Maass, Barutcu et al. 2018, Zhang, Zhou et al. 2022). The non-ligation-based methodologies, such as SPRITE (Quinodoz, Ollikainen et al. 2018), imaging approaches (Chen, Zhang et al. 2018, Maass, Barutcu et al. 2018, Maass, Barutcu et al. 2018, Nguyen, Chattoraj et al. 2020, Su, Zheng et al. 2020, Takei, Yun et al. 2021, Park, Nguyen et al. 2023), and HiPore-C (Zhong, Niu et al. 2023) have assayed single cell types and although they determined NHCCs around the nuclear speckles and nucleoli (Quinodoz, Ollikainen et al. 2018, Zhong, Niu et al. 2023), their depth is not comparable to Hi-C. In summary, while some examples of NHCCs are well established and critical for cellular processes, we still lack a comprehensive view of the fundamental principles of NHCCs. This is owing to analytical Hi-C limitations where a robust statistical framework is required to confidently determine true NHCCs above background noise. Here, we developed a new machine learning method assessing the Spatially Interacting GeNomic ArchitecTURE (Signature) towards a comprehensive and systematic detection of NHCCs, their extent across cell types, and their putative impact on non-random chromosome positioning. Signature is the first tool explicitly designed to examine intra- and inter-chromosomal interactions in Hi-C datasets (including Omni-C, capture Hi-C, and micro-C (Krietenstein, Abraham et al. 2020)), without technical intricacy, further resources, and time to perform orthogonal approaches, which is advantageous for the field.'

- What are the technical limitations of Hi-C, and what type of conclusions cannot or should not be inferred from such data? Authors should discuss this clearly.

We would like to refer to the text section above where we describe the analytical hurdles to call NHCCs confidently in Hi-C data above genomic background noise.

- "Topological anchoring community" regions sound simply like nuclear speckles. Authors suggest they are something different, but it is not presented clearly why that is the case and how they differ from speckles. This needs to be explained and addressed. Also, are the differences across different cell types regarding them simply related to different regions coming closer and further from speckles as their expression patterns change? That should be addressed.

We distinguish the anchor community from the nuclear speckles because we investigated NHCCs that are in proximity to the speckles and not the subnuclear compartment of the speckles. These NHCCs are not part of the nuclear speckles, they rather overlap or are close to the speckles. Our approach determined many constitutive NHCCs in the periphery of the nuclear speckles, however, we do not claim that these NHCCs contribute to speckle formation. We have described in lines 248-251 that the constitutive NHCCs happen in genomic regions with extreme expression levels and many genes, and occurred independently of gene expression patterns in more than 50 % of the datasets. To clarify the text, we followed the reviewer's suggestion and have rephrased our wording.

The new text reads as follows:

'Remarkably, LWPR determined 61 constitutive NHCCs in ≧ 50 % of the 62 datasets (q < 0.05), of which 56 overlapped with the off-centered pattern identified by CD (overlap 91.8 % [56/61], Fig. 3f, permutation testing empirical p = 0, Extended Data Fig. 5a, Supplemental Table 7 and Video 2). This 'topological anchor community' was proximal to q-telomeric NHCCs (Fig. 2h), converged with patterns of genomic features (Fig. 2e, 3c-d), had higher gene density than remaining NHCCs (mean: 47.21/Mb vs. 23.04/Mb for genome), as well as gene expression (Mann-Whitney test p = 3.31x10-11, Extended Data Fig. 5b). Since nuclear speckles are sub-nuclear organelles formed by high gene density and expression, we determined that constitutive NHCCs are associated with speckle periphery via close spatial proximity but are not part of the speckles themselves.'

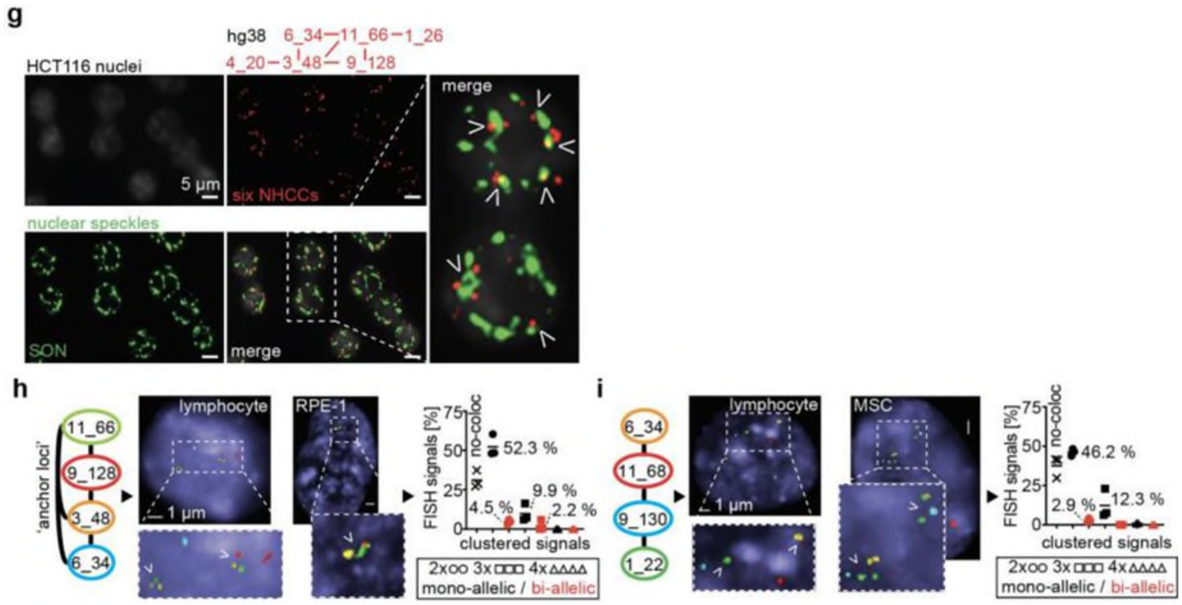
- What are the novel inter-chromosomal interactions that authors identify? Are there any contacts that authors detected, which do not belong to any already known structures?

The reviewer raises an important point about the novelty of the Signature-identified interactions. To address this question properly, we revisited our benchmarking experiments and excluded any overlapping bins of Signature NHCCs with SPRITE and HiCAN. Then, using *Signature*, we detected a total of 8343 unique and novel interactions using Signature compared to HiCAN from 10 analyzed datasets (2x GM12878, HMEC, HUVEC, 2x IMR90, 2x NHEK, 2x teloHAEC datasets) and 1034 new interactions in two GM12878 datasets when compared with SPRITE. The orthologous approaches were unable to significantly detect these interacting regions/bins. We added the described numbers to the figure caption:

Extended Data Figure 3. d. *Comparison of Signature NHCCs to HiCAN(Joo, Cho et al. 2023) (2x GM12878, HMEC, HUVEC, 2x IMR90, 2x NHEK, 2x teloHAEC datasets, q < 0.05) yielded in a total of 850 unique bins and 8343 novel interactions that were solely determined by Signature, and* ***e.*** *to SPRITE(Quinodoz, Ollikainen et al. 2018) (2x GM12878, q < 0.05). Signature identified 255 unique bins that determined 1034 interactions not detected by SPRITE.*

Authors mention constitutive and specific "anchored loci" but do not provide their identity, nor explain if these are earlier mentioned speckle regions. Again, this needs to be clarified. If they are indeed new regions, these should be validated by FISH, and their clustering and change in localization should be presented. Moreover, any potential functions they might be playing would be of high interest for the manuscript.

We would like to refer to table S7 where we list all constitutive NHCCs, and to table S8 that shows all unique NHCCs of the investigated cell and tissue types. Further, the "anchor loci" used in FISH are detailed in Fig. 3h-i. We have described (see also the last comment), that the constitutive NHCCs are associated with the speckles and either overlap or are in spatial proximity. To validate these loci, we had used oligopainting (Figure 3g) and multi-color FISH analysis (Figures 3h-i) and found that nine of the newly identified constitutive NHCCs are overlapping / co-localized to the speckles. For reference:



g. *Oligopainting of six NHCCs (red, hg38 positions: [chr_Mb]) with speckle marker SON (green) in HCT116 cells is shown (n = 2, each > 300 nuclei); arrows depict either co-localization or close spatial proximity; scale bars = 5 µm* ***h-i.*** *FISH of 'anchor loci' (shown with position [chr_Mb]) in lymphocytes, RPE-1 and MSCs [each ˜100 nuclei]). Examples show clustered proximal signals with white arrows; scale bars = 1 µm. Plots show quantification of either no-colocalizations, or mono-allelic (black) and bi-allelic (red) signal frequencies of double (2x), triple (3x), and quadruple (4x) clustered signals of NHCCs with depicted means (%). Means (horizontal line) and datapoints of three analyzed cell lines are shown.*

Minor:

- Figures are extremely busy and not reader-friendly (in several going up to m or further); numbering often changes in the text, and what is presented is not well described in figure legends. It is very difficult to extract important information from Figures, and I highly recommend modifying them to ensure that readers can appreciate the presented work.

We thank the reviewer for raising this important point. We agree that figures showed unrequired details. Following the reviewer's comment, we modified figure 1 and show panels a-h instead of panels a-m. Figure 2 harbors now panels a-i instead of a-l. We also modified the figure legends throughout the main manuscript, as well as of the extended data.

- Authors propose the term "non-homologous chromosomal contacts" to describe inter-chromosomal contacts. I think we should all use similar terms to describe certain observations and not come up with new acronyms when they are not needed. If there is a difference, it is not well described, as I did not understand it.

We acknowledge the suggestion to use the term *inter*-chromosomal contacts instead of non-homologous chromosomal contacts (NHCCs). We used the term in Maass et al, EMBO J 2018, Maass et al NSMB 2018, and Maass et al JBC 2019, because of the need to distinguish between transvection (homologous chromosomal contacts) and non-homologous chromosomal contacts. Therefore, we think that using the term NHCC(s) is appropriate to clearly define the type of the *inter*-chromosomal contact.

Few missing citations and works that should be discussed:

- https://www.nature.com/articles/s41467-022-32980-z)
- https://www.nature.com/articles/s41592-021-01135-1)

We thank the reviewer for suggesting these publications. Now, we have cited the paper from Dotson et al. when we introduce the idea of multi-way interactions. Although we have introduced and cited the orthogonal approaches to Hi-C, such as GAM and SPRITE, we do not compare between Hi-C, GAM, SPRITE and polymer modeling. Thus, we decided to omit the paper from Fiorillo et al.

Reviewer #3 (Remarks to the Author):

In this study, the authors introduce Signature, a machine learning method for identifying NHCCs (non-homologous chromosomal contacts) in Hi-C data. They analyze a large dataset of Hi-C samples from various human cell types / samples. Their analysis reveals a varied landscape of NHCCs, with non-random distribution and association with active regions. Additionally, it identifies cell-type-specific and sex-specific NHCCs, potentially related to their functions and sex-based features. While this work represents a relevant contribution to understanding the role of NHCCs in genome organization and function, I have a few questions and comments.

We would like to thank the reviewer for the evaluation of our work and for the helpful comments on our manuscript. By addressing all raised concerns, we have significantly improved our current draft. Please find the details below.

Could the authors expand the discussion on how NHCCs are deterministic? NHCCs might be stochastic and transient, potentially influenced by factors like transcriptional activity and cell cycle stage, which may be the factors that lead to being more deterministic as well.

We agree with the reviewer that mentioning reasons for NHCC formation and giving a perspective for follow-up studies could be useful.

Therefore we have added the following to the discussion:

‘Exploring the biological functions of non-interacting regions, such as cis gene-regulatory hubs and intra-chromosomal organization would be interesting, as well as the extent of transient factors, such as transcriptional activity and cell cycle stages, that influence deterministic NHCC formation.’

Could the authors expand the comments on the Rabl configuration in human genome organization? Based on ref. 50 (Hoencamp, C. et al.). This configuration is often considered more prevalent in organisms like yeast and file based on the presence of the complete condensin II subunits. In addition, the centromere and telomeres polarization or cluster in opposite directions may partially distort the territories.

Following the reviewer’s comment, we have elaborated on the Rabl configuration in the context of Hoencamp et al.

We have added:

‘Despite recent findings suggesting that condensin II prevents inter-centromeric clustering and negatively influences Rabl’s configuration in the human genome (Hoencamp, Dudchenko et al. 2021), Rabl’s configuration and chromosomal territories may not be mutually exclusive. Rather, they both may coexist to structure a flexible genome architecture that can react to external stimuli and undergo mitotic re-organization.’

One complementary study to ref 74 that could be mentioned by the authors (Contessoto, V et al. Nat. Comm, 2023) shows that the formation of helices in interphase and compaction are related to the Rabl configuration.

We thank the reviewer for suggesting the paper from Contessoto et al. as valuable complementary study. However, since we have already exceeded the volume of references and we cited different examples of Rabl’s configuration, we decided to not include this additional paper.

Reviewer #5 (Remarks to the Author):

Mokhtaridoost et al. introduce ‘Signature’, a pipeline for identifying non-homologous chromosomal contacts (NHCCs). Signature identifies NHCCs through LOESS regression to flag outliers and uses community detection methods to cluster and embed chromosomal bins in an interpretable manner. The authors conclude that NHCCs are strongly associated with gene expression, telomeric and centromeric clustering occurs pervasively in human cells, and that constitutive NHCCs ground cell type specific gene activity.

Signature would be the first pipeline for studying inter-chromosomal contacts, and to date, this area has remained relatively unexplored due to conflicting reports on the existence of such contacts. Their findings re-capitulate many well-known examples of chromosomal organization, such as previously identified NHCCs, Rabl's configuration, and speckle associated contacts. While the paper has strong conceptual and biological appeal, the authors occasionally overstate their contributions, overinterpret their results, and fail to include clear methods for their results.

Nonetheless, Mokhtaridoost et al.'s paper is a promising first step into the exploration of NHCCs. This is an underappreciated and unexplored topic in the field of 3D genomics, and recent developments in this area indicate a broader interest in NHCCs. This work would augment existing analysis methodologies for Hi-C assays and stimulate the further development of approaches aimed at analyzing NHCCs and inter-chromosomal contacts in general.

We would like to thank the reviewer for the detailed and precise evaluation of our work and for the thoughtful comments on our manuscript. We have addressed all raised concerns, comments, and ideas, which led to significant improvement of our current manuscript. Please find detailed point-by-point responses below.

Major Comments:
1. Line 62 – The authors cite multiple sources claiming that NHCCs are not readily detectable in Hi-C data. However, it is unclear how all of these sources support that claim.
a. Source 17 (HiC-DC+) is exclusively concerned with intra-chromosomal contact identification.
b. Source 18 (Higashi) is a sc-HiC method for imputing sc-HiC data with a hypergraph representation and is not focused on chromosomal contact identification, though the authors acknowledge the difficulty of using inter-chromosomal contacts.

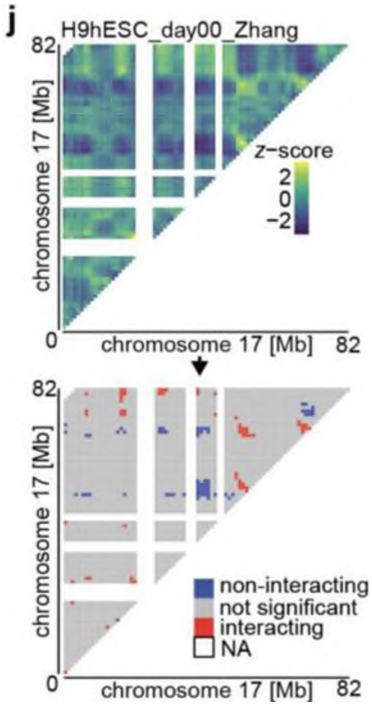
The reviewer is correct that the presented citations were unclear. To cite the different sources more precisely and to improve this part of our introduction as requested below, we have revised the entire section, reordered its citations, and removed the HiC-DC+ paper. The text reads as follows:

'Two commonly used approaches for investigating NHCCs are imaging and chromatin capture, both of which are limited in determining NHCCs. Specifically, imaging is not scalable to genome-wide approaches and chromosome conformation capture (i.e., proximity ligation-based Hi-C) as the most widely used technique to study 3D genome organization mainly focuses on analyzing intra -chromosomal contacts (Lieberman-Aiden, van Berkum et al. 2009, Dekker, Alber et al. 2023). Moreover, both methodologies often caused discordant results when studying NHCCs that do not complement one another (Dekker 2016, Maass, Barutcu et al. 2019, Payne, Chiang et al. 2021). Importantly, Hi-C datasets contain 'trans-reads', but current computational and statistical analysis has limitations in confidently determining true NHCCs. Hence, NHCCs have been considered as stochastic, singular events (Bashkirova and Lomvardas 2019, Maass, Barutcu et al.

2019), that are not readily detectable in Hi-C data (Maass, Barutcu et al. 2018, Zhang, Zhou et al. 2022)....'

2. Line 73 – Signature can assess intra- and inter-chromosomal interactions, but their assessment of this capability appears sparse. At the minimum, some comparisons against extant intra-chromosomal contact callers and statistics on their enrichment above the local background via APA would be helpful.

To address the reviewer's comment, we now mention the number of significant *intra*-chromosomal interactions that we determined according to the approach described in Sanyal et al. 2012 with our cross-validated span parameter. Specifically, we determined 120,106 ($q < 0.05$) significant *intra*-chromosomal interactions at 50 kb genomic resolution (31,604,799 [$p < 0.05$], submitted to CEO, methods). We have added this information to the text. Moreover, we show that our results can be plotted and visualized with classic interaction matrices (such as APA plots and other interaction matrix plots, see Extended Data Fig. 1j below). Finally, we addressed the enrichment of *cis* contacts over background noise. Specifically, we generated a classic Hi-C interaction matrix with *z*-scores and compared it to significant interactions. Using LOESS (Sanyal, Lajoie et al. 2012) and our span selection, *Signature* identifies long-range *cis* interactions, similarly to orthogonal approaches that analyze Hi-C datasets and *cis* contacts.



Extended Data Fig. 1j. *Signature determines intra-chromosomal interaction weights according to the method described by Sanyal et al. 2012 (Sanyal, Lajoie et al. 2012) with our cross-validated span selection. The example depicts (top) the z-scores and (bottom) p-values of the interaction matrix for chromosome 17 cis interactions, including long range interactions, in the H9-ESC Hi-C* *dataset (Zhang, Li et al. 2019). Significantly interacting regions are represented in red and significantly non-interacting regions in blue (p < 0.05), grey represents no significance, and white represents no available data.*

Moreover, we expanded on the direct comparison of NHCCs and *intra*-chromosomal contacts by comparing averaged z-scores genome-wide for cis contacts with NHCCs (all *z*-scores) across 62 Hi-C datasets. Please see our approach and response to point 11 below.

3. Line 94 – Weights are discussed, but at this point in the manuscript, what "weights" is referring to, is unclear. Throughout the remainder of the manuscript, the authors appear to flip between interaction weights/interaction frequencies; I am almost sure that these are referring to the same object. If not, please clarify.

We acknowledge that we used the terms 'frequencies' and 'weights' inconsistently throughout the manuscript. Using the term 'interaction weight' is more appropriate since this metric is derived from the normalization process. To clarify and to use the most precise terminology, we changed the text and now use *weight* throughout the manuscript and extended data.

4. Line 94 – The authors say they developed a non-parametric supervised learning approach called Locally Weighted Linear Regression. I have two minor objections to this.
a. LOESS – Their method calls the LOESS function in R (loess.sd in msir calls loess) to accomplish the non-parametric smoothing of the data.
b. Linear – This is not true. The default degree of LOESS yields local quadratic smoothing, and the authors have not specified d=1 in the calls to LOESS (or used LOWESS) and thus, they are performing a locally weighted polynomial (quadratic) regression.

We thank the reviewer for bringing this up. Indeed, our approach is rather a polynomial than local linear regression. We therefore changed the term throughout the entire manuscript and extended data to Locally Weighted Polynomial Regression (LWPR).

5. Line 634 – Clarify if a one sided or two-sided z-test was used.

We have used two-sided tests only, which we describe under our statistical analysis section of the methods. For reference:

"All statistical tests conducted were two-sided, unless stated otherwise."

6. Line 107 – References Figure 1A, but lack of figure annotations makes it unclear what is being plotted in the right-hand side of the first facet. Is it the z-score of the interaction after removing the background?

We acknowledge that additional annotation is required to improve the figure and its caption. Following the reviewer's comment, we added *z-scores* as Y axis label and modified the figure 1a caption to:

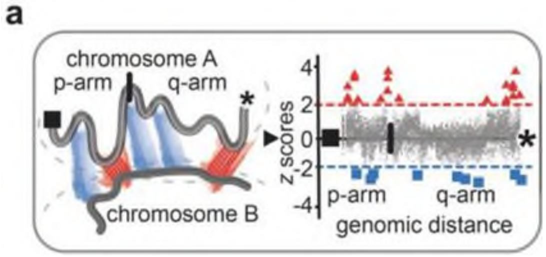


Figure 1a. '*Supervised learning in Signature. Scheme depicts genomic regions between different chromosomes that are evaluated for their interacting (red = positive z scores) and non-interacting bins (blue = negative z scores). All genomic regions are queried against all other regions in an 'All vs. All' approach. Right: example of interaction weights with significance cutoffs* (dashed lines) *between two chromosomes.*'

7. Line 111 – Can the authors clarify the meaning of "how".

We agree with the reviewer that this sentence needed clarification.

It now reads as follows:

'This feature complements supervised learning (LWPR) and visualizes spatial peculiarities of where NHCCs impact genome topology.'

8. Line 123 – Many non-interacting regions are proposed to exist, but this is briefly mentioned in passing. Have the authors explored the possible biological functions of these non-interacting regions?

The reviewer raises a very interesting point. Our *Signature* pipeline provides the advantage to determine both interacting and non-interacting regions. The latter may harbor important biological function, such as *intra*-chromosomal interactions clustering in distinct sub-nuclear compartments to support chromosomal stability, *cis* gene-regulatory hubs for maintaining basic cell functions, etc. Here, we primarily focused on addressing NHCC characteristics to start illuminating this unexplored topic of 3D genome organization. However, we reanalyzed the non-interacting regions and filtered for common and tissue-specific unique contacts in 13 datasets to give readers a perspective for follow-up studies. We have extracted annotated genes and performed GO enrichment analysis. Interestingly, we also found that non-interacting regions related to biological function of the Hi-C dataset's origin. We have this result to Extended Data Fig. 6c and modified the text to:

‘Notably, genes at these unique NHCCs and also at significantly non-interacting regions related to meaningful biological functions by GO-term analysis(Zhou, Zhou et al. 2019) (Fig. 4f, Extended Data Fig. 6c).’

Since a more detailed exploration of non-interacting regions would be beyond the scope of this study, we added a perspective to the discussion:

‘*Moreover, further exploring the biological functions of non-interacting regions, such as cis gene-regulatory hubs and intra-chromosomal organization would be interesting.’*

9. Line 126 – There are 46 chromosomes, each corresponding to the paternal and maternal copies, but as far as I understand, an alignment would need to be performed with that goal in mind, whereas the authors have only performed an alignment against the reference genome.

The reviewer is referring to our strategy to establish the Community Detection approach to study and to visualize genome topology. We acknowledge that our process requires a much more detailed explanation which we have added to the manuscript.
Specifically, we reasoned that we need at least 24 communities because the Hi-C data is produced of gonosomes and autosomes of diploid human cells. However, even without separating parental alleles during the mapping against the reference genome, 46 chromosomes in the human genome may interact with one another. To allow communities and NHCCs to form while keeping the structures of 46 chromosomes intact, we decided to set the number of communities to 46. In this case, chromosomes and their interactions that belong, based on Hi-C interaction weights, to more than one community, can be visualized in our genome topology maps.

a. Fig 1e is very suggestive, but the description of the methods for visualizing this are not clear.
b. After community detection into 46 communities, what does it mean to cluster interactive bins? Is the linking of consecutive bins what generates the chromosomal outlines in Fig 1e?

We acknowledge that the figure caption 1e, as well as the steps describing how we generated genome topology maps could be improved. We used the clustered bins to visualize communities, and by considering the physical connection of consecutive bins, we generated the chromosomal outlines as an estimation of genome topology.
Following the reviewer’s suggestion, we have modified Figure 1d caption to introduce the visualization concept for genome topology maps. The legend reads as follows:

‘Fig. 1d. *Consecutive bins of each chromosome are strung together to* generate the chromosomal outlines *and to visualize CD-approximated genome topology across 62 Hi-C datasets. Large chromosomes 1-7 (red & pink) and small, gene-dense chromosomes 16-22 (blue & black) are highlighted.’*

Moreover, to make the process to generate topology maps more transparent, we enlarged a region of figure 1e and added a zoomed-in window. This shows that consecutive bins generate the chromosomal outlines in the visualization.

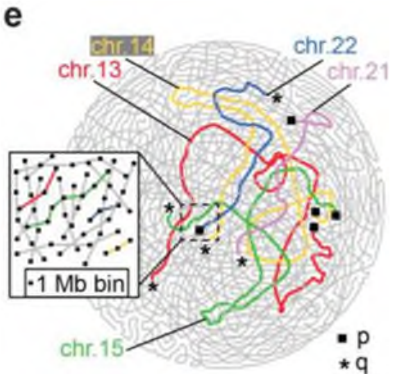
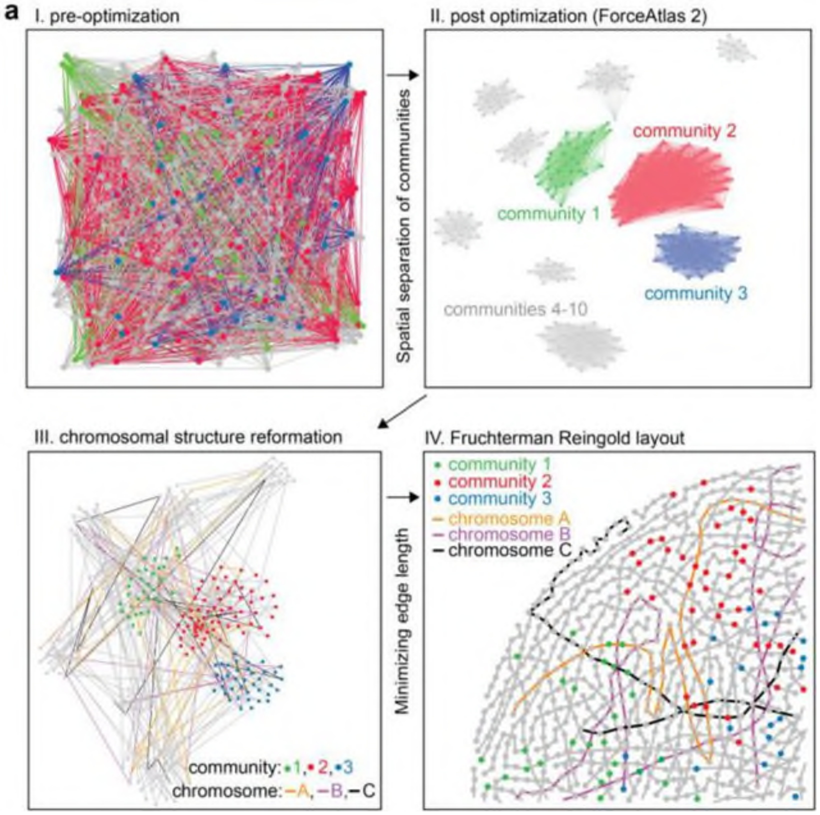


Figure 1e. *Acrocentric chromosomes 13-15, 21, and 22 are colored in genome topology map. Telomeric p- arms and q-arms are shown as black squares or asterisks, respectively. Enlargement depicts how CD strung bins together to generate chromosomal outlines.*

Finally, we changed the method section and added a step-by-step procedure to Extended Data Fig. 2a.



Extended Data Fig. 2a. *Schematic overview of visualization process of CD results in genome topology maps. Figure contains toy data for simplicity. I: Weighted network with CD results of 10 communities shows only intra-community Hi-C interactions. Inter-community interactions were removed. Each node represents a bin, and each weighted edge between two nodes depicts the Hi-C interaction weight between them. Three random communities colored in green, red, and blue are highlighted. II. Application of Force Atlas 2 optimizes the network and positions 10 communities separately. Node positions within each community are optimized independently from other communities based on their network interactions. III. Integration of chromosomal structure by removal of edges from step II and connecting consecutive bins. For example, the first bin of chromosome 6 (node 6_0) is connected to the second bin of chromosome 6 (6_1). IV. Final genome topology estimation optimized after step III using the Fruchterman-Reingold method in Gephi. Edge colors represent chromosomes, and node colors indicate the same communities from the previous step. This 4-step visualization approach incorporates the results of the CD algorithm and simulates chromosomal structure by connecting consecutive bins while recapitulating the proximity of nodes within each community as much as the chromosomal structure allows.*

The methods read as follows:

'To visualize a maximum of 46 chromosomes (two gonosomes [female XX / male XY] and 22 pairs of autosomes = 24 possible chromosomes) in human diploid cells, we set the number of possible communities to 46. This resembles the human genome and allows each community to include only one chromosome's bin if there is an intra-chromosomal domain structure isolated from the rest of the genome. For the visualization of the genome topology, we used the Gephi (Bastian, Heymann et al. 2009) software. To reflect the results of CD and to ensure clear separation between bins from different communities, we first optimized the visualization process. We excluded inter-community interactions and plotted all bins as nodes using the ForceAtlas-2 (Jacomy, Venturini et al. 2014) visualization layout, based on intra-community Hi-C interaction weights. This step ensured that bins within each community are visualized close together and separated from other communities, in turn facilitating the visualization of each community in the genome topology map. The distribution across the topology map as a 'mock nucleus' resulted in 46 distinct communities, where bins with higher interaction weights in each community were placed closer together to better visualize their interactions. Next, we added the physical connections between consecutive bins as edges in the network and optimized the network layout using the Fruchterman-Reingold layout algorithm (parameters: area = 5000, gravity = 5, speed = 10). This ensured that bins within the same community remained close together and the structural connections between consecutive bins across chromosomes were maintained. In the final genome topology estimation, consecutive bins were positioned next to each other to outline of the chromosomal structures. Moreover, bins within the same community were as close to each other as the physical constraints allowed.'

c. Based on the description of the ForceAtlas 2 usage and Gephi, it seems like since there are 46 communities, and between community interactions are excluded, it seems there should be quite a few distinct communities that number more than the 23 chromosomes present in Fig 1e? What are these?

In each of our genome topology maps, there are exactly 46 distinct communities, plotted separately due to the removal of *inter*-community interactions. However, the final figure shown in Fig 1e is after adding chromosomal structures and optimizing the network as described above. In Extended Data Fig. 2a, we now show how communities (node colors) in the genome topology network represent each bin's community, and where edges represent 24 chromosomes of human diploid cells. Since distinguishing colors of 46 communities and chromosomes is infeasible, we performed the described optimization steps (see stepwise CD visualization in Extended Data Fig. 2a and Methods) and presented simplified genome topology maps throughout the manuscript.

10. Line 666 – Intercommunity interactions are excluded from the ForceAtlas2 layout used to visualize the CD results, so I am confused what that implies about the clustering of the centromeres and telomeres?

We want to clarify our intentions behind excluding *inter*-community interactions. Excluding them ensured that bins within each community are plotted close together and separated from other communities which enhances the visualization of each community. Our aim was to distribute chromosomes and their interactions across a 'mock nucleus' (genome topology map) to facilitate the most accurate visualization of their interactions and the communities that CD determined. To improve the understanding of the CD visualization, we added a step-wise visualization of toy data in Extended Data Fig. 2a and modified the methods (see above under point 9).

a. Lines 204-215 – This section claims that since inter-telomeric and inter-centromeric NHCCs occur together, this is evidence for Rabl's configuration, but the methods indicate that interactions (weights?) between communities (chromosomes?) are omitted. So telomeric interactions across chromosomes would presumably be omitted as well since it is demonstrated that the chromosomes appear to form distinct strands in the visualization. Could the authors clarify the methodology which justifies their conclusions in this section.

The reviewer refers to our visualization approach of CD results. We explained in our response to point 9 that we excluded inter-community interactions to facilitate the visualization of communities and to accurately reflect the CD results. However, this does not contradict the clustering of bins from different chromosomes (such as inter-centromeric or inter-telomeric bins) within the same community.

The support for Rabl's configuration is based on the following observations: centromeric bins cluster together, and telomeric bins cluster together in different communities, positioned on opposite sides in the genome topology estimation. This arrangement is due to strong interactions between telomeres of different chromosomes and between centromeres of different chromosomes, as indicated by the Hi-C data. Conversely, telomeric bins and centromeric bins do not interact with each other, resulting in their positioning on opposite sides of the genome topology map.

11. Line 166 – It is claimed that the NHCC frequencies do not depend on the interactions within chromosomes are not related to the ones for NHCCs, but prima facie I was expecting a negative correlation.

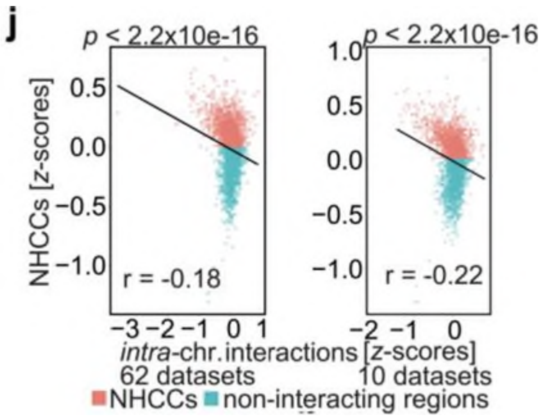
a. If a population of cells makes more NHCCs, then it does so at the expense of intra-chromosomal contacts.

b. Extended Data Fig. 3f is supposed to support this point, and while the coefficient of correlation appears small, wouldn't the overall negative slope suggest an inverse relationship between NHCC formation and intra-chromosomal contacts? Further, NHCC z-scores are plotted, and this includes z-scores of 0, which would not yield a significant p-value? Were all inter-chromosomal z-scores plotted because I thought that NHCCs were flagged as a positive z-score with q < 0.05.

We agree with the reviewer that a negative correlation could be expected, and we acknowledge that our first comparison of *intra*-chromosomal contacts with NHCCs was not appropriate because we did not apply the same criteria for *cis* & *trans*. Therefore, we redesigned this experiment. We considered all bins of the genome in the 62 Hi-C datasets and averaged their *z*-scores to avoid any bias by ignoring a subset of bins. If we had used only significant values, a given bin may represent only single Hi-C datasets and prevent investigating the overall genome-wide relationship of *cis/trans* contacts.
Moreover, we selected the top five and bottom five datasets with the highest and lowest number of NHCCs to test if more NHCCs occur at the expense of *intra*-chromosomal contacts. Remarkably, both approaches led to negative correlations (R = -0.18 and -0.22, p < 2.2E-16, respectively).

We incorporated the new results and have modified the text:

'To investigate if NHCCs depend on intra-chromosomal interactions, we performed genome-wide correlations of averaged cis and trans interaction weights per bin for all 62 datasets. Moreover, using 10 Hi-C datasets with the highest and lowest number of NHCCs, we tested if more NHCCs occur at the expense of intra-chromosomal contacts. We found weak negative correlations for both experiments (Pearson's R = -0.18 and -0.22, p < 2.2×10^{-16}, Extended Data Fig. 3j), that suggest that bins involved in NHCCs are mostly spatially separated from intra-chromosomal contacts.'



Extended Data Fig. 3j. *left: Pearson correlation of bins involved in intra-chromosomal interactions (averaged z-scores, 1Mb bins) and NHCCs (averaged z-scores, 1Mb bins) genome-wide in 62 Hi-C datasets (r = -0.18, p < 2.2x10^{-16}). NHCCs are shown in red and non-interacting regions in cyan. Right: Pearson correlation of averaged intra-chromosomal interactions and NHCCs in datasets with highest and lowest number of NHCCs (each category five datasets, r = -0.22, p < 2.2x10^{-16}).*

12. Line 172 – To clarify, when it says a span of 1.84 Mb on average, that means that the total length of interacting NHCCs (consecutive 1 Mb) bins making up that domain averages out to 1.84 Mb?

Yes, this is correct.
We rephrased the sentence to ‘*Specifically, NHCCs spread across 1.84 Mb on average, whilst significantly non-interacting regions comprise 3.38 Mb.*’

a. For example: In case 2, when you have (consecutive?) bins with one bin in the middle interact with different single bins on chromosome B. This is only computed as 2 Mb, and not 3 Mb. Further, the bins on chromosome B do not factor into this.

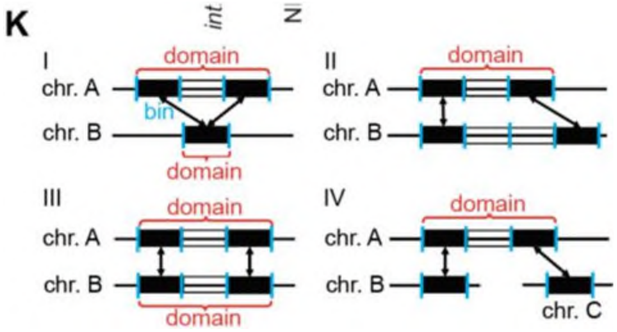
Yes, this is correct. Each bin’s position was solely used once for measuring NHCC size. A gap of one bin between consecutive bins was taken into consideration when we determined domain sizes. Gaps with two or more bins were handled separately. We have modified the figure caption accordingly to clarify. Please see also below.

b. Also, are there cases where bins are assigned to multiple domains? If so, how is this handled?

No. We have just used the bin information to quantify the mean NHCC size, independently of the number of significant interactions it was involved in. If a bin fits into one domain on its right side and another domain on its left flank, the two domains were integrated into a single domain.

c. Could the domain definition be relegated to the methods instead of the extended data section?

Yes, of course. We have clarified our criteria and the scheme in Extended Data Figure 3k to define domains, and added the description to the main methods section.



'Domain definition for NHCCs and non-interacting regions.

We considered four different scenarios (I-IV) to determine the average size of NHCC and non-interacting domains (Extended Data Fig. 3k). I: consecutive bins of chromosome A, with a maximum of one bin as gap in between, interact with the same bin on chromosome B. This builds domains on chromosomes A and B. Individual interacting bins were considered once to measure average NHCC sizes. II: consecutive bins of chromosome A with one bin as gap in between interact with different single bins on chromosome B. This builds a domain on chromosome A only. III: consecutive bins of chromosome A, with one bin as gap in between, interact with similar setup on chromosome B. This builds domains on chromosomes A and B. IV: consecutive bins of chromosome A, with one bin as gap in between, interact with different bins on two different chromosomes B and C. This builds a domain on chromosome A only.'

13. Line 174 – Why are domains that are detectable at 1.84 Mb exclude the ones at 3.38 Mb?

We are unclear about the reviewer's comment. We found that NHCCs spread across 1.84 Mb on average due to their positive significant *z*-scores, whilst significantly non-interacting regions with negative *z*-scores comprise 3.38 Mb. We hope that our modifications of the text, our scheme in Extended Data Fig. 3k to call domains, and the methods describing our approach (see point 12) clarify the reviewer's comment.

14. Line 192 – Could the authors comment on how well mapped the centromeric and telomeric regions are as highly repetitive regions of the genome generally are difficult to map with short reads, and whether this affects the performance of Signature.

We have filtered repetitive regions out according to BWA and the standards of the 4D Nucleome Hi-C analysis pipeline. Unmapped regions in sub-telomeric regions, around centromeres, and across the acrocentric p arms have not been considered, and thus, they are not influencing the performance of LOESS and/or Signature. Since we analyzed only continuous sequences of mapped regions, we cannot predict the interactive behavior of genomic regions in highly repetitive sequences, such as centromeres. To clarify which elements we considered for our conclusions, we have changed the text to:

'*When analyzing mapped regions flanking the centromeres and sub-telomeric regions, Signature identified thousands of significant contacts of p-arms and particularly of q-arms across all cell types.'*

15. Line 223 – For GTEx analysis, are all the counts across the various tissues just being averaged after transformation?

We recognize that our description regarding the GTEx analysis could improve from clarification. We have defined four different biotypes of genes (protein coding, small non-coding, long non-coding and pseudogenes). Since the common NHCCs were composed from a total of 61/62 Hi-C datasets, we averaged TPM values (per gene) across every tissue provided by GTEx (17382 samples from 948 donors and 54 tissues), then transformation was applied. This is already plotted in Figure 3b.

a. Is the justification for this choice being that the cell types being analyzed are being pooled from various tissues?

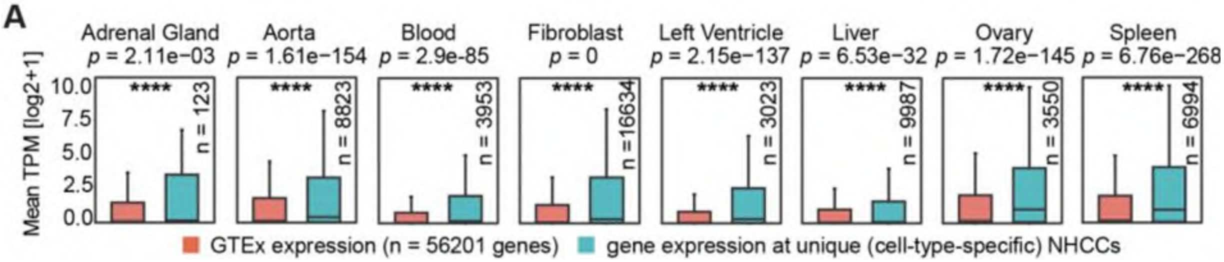
Yes, the reviewer's assumption is correct.

b. If that is the case, I can appreciate why the authors sought out the GTEx for their analysis, but it would have been helpful to study NHCCs comprehensively in a selected cell type to see if their inferences across the entirety of their dataset held for their selected cell type.

We have addressed the question if matched pairs of RNAseq and Hi-C data show that NHCCs harbor most of the active genes. In line with the reviewer comments under point 15, we have revised our text to explain our approach and findings better.

The text reads as follows:

'Gene expression across all 62 datasets was significantly higher for common NHCCs when compared to all of GTEx (Consortium 2020) (Mann-Whitney test, range: $p = 2.79x10^{-15}$ - $p = 6.11x10^{-52}$, Fig. 3b). The same was true when we compared tissue-specific gene expression profiles from GTEx matching our Hi-C samples by showing that NHCC regions harbor the most active genes (Extended Data Fig. 4a).'



Extended Data Figure 4a. *Comparison of mean gene expression shown as mean TPM (log2 +1) at unique (cell-type-specific) NHCCs (q < 0.05) with entire GTEx catalogue. n represents number of genes found in unique NHCCs. Box limits represent upper and lower quartiles. Central boxplot line represents the median and whiskers represent 1.5x IQR. GTEx samples: 258 GTEx samples from adrenal gland, 432 GTEx samples from aorta, 755 GTEx samples from whole blood, 504 GTEx samples from cultured fibroblasts, 432 GTEx samples from left ventricle, 226 GTEx samples from liver, 180 GTEx samples from ovary, 241 GTEx samples from spleen.*

16. Line 233 – This section explores the correlations between NHCCs and underlying genomic properties, but this does not imply the NHCCs support the gradient.

We agree with the reviewer. We have revised the text and toned down our conclusion. The

text now reads as follows:

'This spatial genome gradient along an axis of activity may either form because of TF accumulation, transcription, and RNA or these features are a consequence of chromatin flexibility and diffusibility of DNA and subnuclear organization.'

17. Line 282 – The randomization procedure discussed in the methods is unclear to me. Is it the following?

a. 40282 total NHCCs, 23351 (discrepancy between text and methods on this) unique

We apologize for the typo in the main text. We have corrected the number to 23251.

b. On each randomization, draw 40282 NHCCs (with replacement?)
c. What does it mean to check the percentage that is unique on each draw?

We have revisited our description of the applied methodology and have modified the text to:

'We applied random selection to a simulation of NHCC numbers. For each dataset, we randomly selected the same number of NHCCs as presented in the dataset. For example, dataset aorta_Leung showed 2210 NHCCs. Thus, we randomly chose 2210 out of the total number of NHCCs without replacement. This process was repeated for all datasets, and then iterated 10000 times. In each iteration, we calculated the ratio of uniqueness (the

number of NHCCs that was selected in only one dataset). Our actual unique NHCC ratio was 0.577, whereas the highest ratio observed in the 10,000 iterations was 0.162.'

18. Line 340-341 – The asymmetric spatial genome gradient that is demonstrated in the paper and proposed is identified from the global embedding after pooling all the individual maps. In the absence of the individual embeddings and characterizations, this finding could be a fallacy of division.

The reviewer is raising an important point that we addressed by two different approaches. As presented in figure 5 where we analyzed approximately 50% of the 62 Hi-C datasets to illuminate sex-specific NHCCs, we have found the asymmetric spatial genome gradient for the different groups of Hi-C datasets. However, to further prove that the gradient is even true for a subset of data, we re-analyzed a smaller subset of Hi-C datasets. Specifically, we selected 10 cardiovascular-related datasets (16.1% of entire compendium; left and right ventricle, stages of cardiomyocyte differentiation, aorta, smooth muscle cells, etc.) and redid the described analysis steps. Remarkably, we still find the asymmetric genome gradient and 96.7 % of constitutive NHCC loci. Moreover, when we map the expression profiles of aorta and ventricle (GTEx samples) to the subset of cardiovascular Hi-C datasets, we recapitulate the asymmetric off-centered genome gradient that we find across all GTEx samples. Our new results indicate that individual maps of much smaller subsets of our compendium still hold true for the claimed findings. We decided to add these results to the manuscript:

'Moreover, a subset of 10 cardiovascular Hi-C datasets recapitulated the spatial genome gradient with 96.7% of the constitutive NHCCs and showed asymmetric gene expression (Extended Data Fig. 7a, b), revealing that the genome gradient holds true in smaller subsamples.'

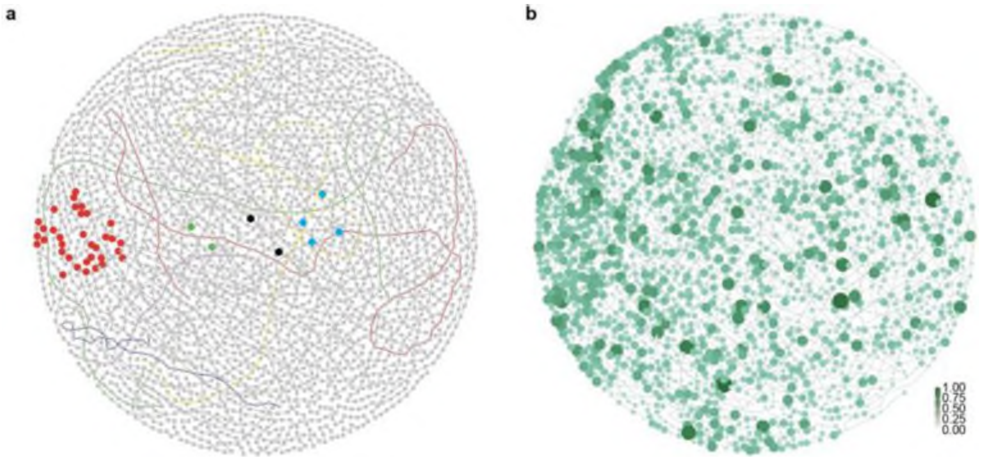


Figure S7a. *Genome topology map of 10 cardiovascular Hi-C datasets with constitutive NHCCs (red dots) and acrocentric chromosomes 13-15, 21, and 22 (colored chromosomal outlines).* ***b****. Scaled average gene expression of aorta and ventricle samples from GTEx (Consortium 2020)*

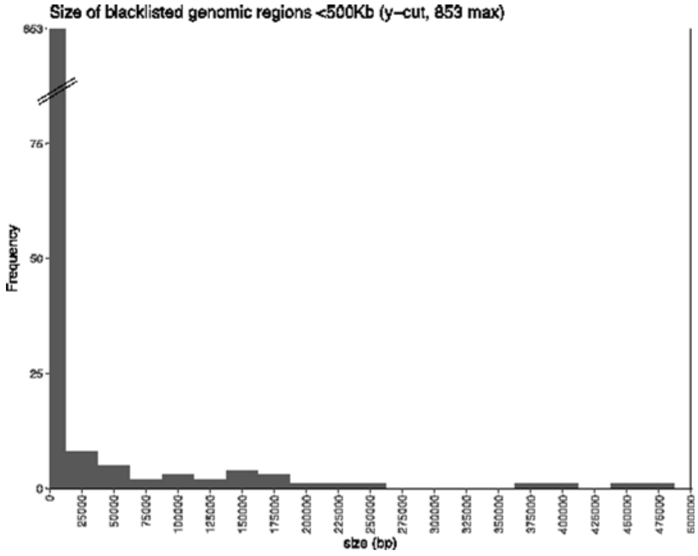
per 1 Mb bin across genome topology of 10 cardiovascular Hi-C datasets shown in panel a (high [green] vs. low expression [white]).

19. Line 565 – HiC Pro was compared to the 4DN protocol for mapping.
a. I am surprised to see mapping rates for the RPE1 datasets being that low. HiC Pro uses a global alignment (bowtie2 end-to-end) rather than the local alignment (bwa mem) of the 4DN protocol. Were there issues with the raw data fastqs (adapter contaminants, low quality tiles) that prohibited proper end-to-end alignment?

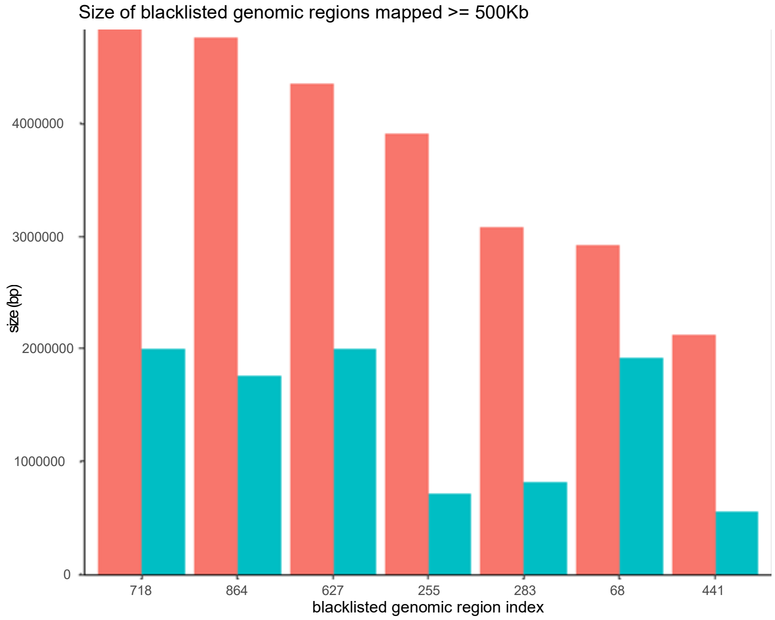
The reviewer is referring to mapping rates of the RPE-1 Hi-C datasets with Hi-C Pro. We did not look into details for single datasets, because the BWA mapping strategy was overall more efficient than bowtie2 (see figure S1l), and yielded with the 4DN protocol to more aligned and valid read pairs. Since our first Hi-C Pro mapping attempts go back to pre-pandemic times, we have to acknowledge that we cannot recall all these details. Since we do not describe results of Hi-C Pro-mapped datasets, we therefore think that further evaluation would be out of the scope of our study. We can exclude the supplemental figure to avoid confusion.

b. Further, bwa mem allows for multi mappers to exist, but they are assigned a mapping quality of 0. A previous work from Johanson et al 2018 Plos Genetics indicated that when they identified NHCCs using a local enrichment measurement, blacklisted and hard to map regions made up many their NHCCs. Are the authors aware if this is also an issue for the NHCCs they have identified using Signature?

We devised the computational steps to map and to analyze the Hi-C reads in line with the 4DN recommendation for processing Hi-C datasets, which does not include pairtools pruning (considering ENCODE's blacklist). This was surprising to us. We are aware that hard-to-map regions may confound results. In pairtools, we filtered out duplicate reads and selected for uniquely mapped read pairs, which removed repetitive sequences, reads with mismatches, and parts of the blacklisted regions.
However, to further address the reviewer's concern, we have used the current blacklist and compared the regions to our NHCCs. Most of the blacklisted regions (853) are up to 12.5 kilobases (see below) and if not filtered out in the pre-processing steps, these regions would likely not impact LOESS' ability to determine interaction weights when processing all *vs* all with our span parameter of ˜4-7 Mb.



Of the ENCODE Blacklist, 2.5% (23/910) are >= 500 kb and all of them were either fully or partially unmapped by *Signature*. Since our NHCC analysis is based on 1 Mb resolution, we considered blacklisted elements that are mapped in more than 50% of our bins (>500 kb, total: 7 blacklisted regions = 11x 1Mb mapped bins, see below [cyan: mapped region, red: entire region]). We checked the overlap with of our significant 40,282 NHCCs and found no overlap. However, we found that 424 interactions (1.05% of all NHCCs) contained a blacklisted region on one side of its interaction. Thus, we are confident that the identified NHCCs are true NHCCs.



20. Line 759 – Related to earlier comments, please make obvious that you considered the fact that the repetitive sequences in the telomeric/centromeric regions are unmappable.

We thank the reviewer for pointing this out again. We have modified the methods according to the reviewer comments under point 14 and 20. The text reads as follows:

'To explore potential patterns in chromosomal interactions and positioning of telomeric and centromeric regions, we focused on specific segments of mapped sequences, excluding repetitive sequences.'

21. Extended Data Fig 3. A) – A few questions
a. Does each point correspond to one of the 62 HiC maps?

Yes, the reviewer's assumption is correct. We have added *'Each datapoint represents one Hi-C dataset'* to the figure caption.

b. How is Cooler being used here for the NHCC step? It it their implementation of HiCCUPS but on the non-homologous chromosomal pairs? If that is the case, I am surprised that the called percentage of NHCCs is so high. Also, I am surprised that an interaction caller using a distance function yields a similar number of calls to Cooler; the methods section for the intra-chromosomal interaction caller for Signature is sparsely described.
c. In general, I think this figure needs a more extensive description of what is being shown.

We acknowledge that expanding the description of figure S3a is required. To determine detection rates of either intra-chromosomal contacts or NHCCs, we calculated the theoretical possible number of interactions dependent on the selected genomic resolution and plotted detection rates for each dataset in figure S3a. We have not used HICCUPS.
We think that the number of detected NHCCs in comparison to the theoretical number of NHCCs has not been described before. We were also surprised about the called percentage of NHCCs. However, we want to point out that these identified NHCCs represent all interactions and have not been tested for significance. Following the reviewer's comment, we modified the figure caption to the following:

*'**a.** left: detection rates based on theoretical number of possible interactions. Possible number of intra-chromosomal interactions is $1.9x10^8$ at 50 kb resolution, whilst $4.52x10^6$ NHCCs can theoretically occur at 1 Mb genomic resolution. Comparison of sequencing depth (number of mapped reads) and detection rate of either intra-chromosomal interactions (black) or NHCCs (red) is shown post-Cooler. right: Pearson's rho (r) and p values are denoted. Each datapoint represents one Hi-C dataset. Signature determined a median of 19.1 % intra-chromosomal interactions, where detection rates highly correlated with the number of mapped reads (Pearson correlation r = 0.77, p < 0.0001). Unexpectedly, Signature detected a median detection rate of 73.7 % for all NHCCs ([range 68.4 – 79.8 %], even in datasets with low sequencing depth*

(Pearson correlation r = 0.45, p = 0.0003), indicating recurrent NHCCs. *Ultra-deep sequencing with ˜8 - 12 billion (B) reads per sample (i.e., cardiomyogenesis and H1-ESCs) (Zhang, Li et al. 2019, Akgol Oksuz, Yang et al. 2021) only led to a marginal increase of the NHCC detection rate (79.8 %).'*

22. Extended Data Fig 3. B) – Why is it titled NHCCs over seq. depth? Again, I presume that the percentage of NHCCs identified by Cooler and Signature are what is being plotted, and each of the dots corresponds to the different data sets used.

The reviewer is correct. We have modified figure S3b and its caption to: *'Comparison of NHCCs (%) in Cooler (Abdennur and Mirny 2020) and Signature (mean Δ0.07 %) outputs by Pearson correlation (r = 0.999, p < 0.0001). Each datapoint represents one Hi-C dataset.*

Minor Comments:

1. Line 58 – The authors reference limited software tools for detecting NHCCs, but to my knowledge, this would be the first tool explicitly designed for detecting NHCCs.

We thank the reviewer for highlighting *Signature*'s uniqueness. Accordingly, we have revised the introduction and say that '*Signature is the first tool explicitly designed to detect intra-and inter-chromosomal interactions in Hi-C datasets (including Omni-C, capture Hi-C, and micro-C (Krietenstein, Abraham et al. 2020)), without technical intricacy, further resources, and time to perform orthogonal approaches, which is advantageous for the field.'*

2. Line 63 – "Hi-C ... captures NHCCs ...", but in Line 61-62, "NHCCs have been considered as ... not readily detectable in Hi-C data", I think there is a point of distinction to be made: Hi-C contains NHCC data, by virtue that there are counts mapping to trans contacts, but the issue of detection is another thing entirely, which is what I think 62 is discussing.

The reviewer's assumption is correct and we agree that we need to clearly point out the technical (analytical) limitations of current Hi-C analysis. In line with the reviewer's comment, we have therefore modified this part of the introduction. The text now reads as follows:

'Specifically, imaging is not scalable to genome-wide approaches and chromosome conformation capture (i.e., proximity ligation-based Hi-C) as the most widely used technique to study 3D genome organization mainly focuses on analyzing intra-chromosomal contacts (Lieberman-Aiden, van Berkum et al. 2009, Dekker, Alber et al. 2023). Moreover, both methodologies often caused discordant results when studying NHCCs that do not complement one another (Dekker 2016, Maass, Barutcu et al. 2019, Payne, Chiang et al. 2021). Importantly, Hi-C datasets contain 'trans-reads', but current computational and statistical analysis has limitations in confidently determining true NHCCs. Hence, NHCCs have been considered as stochastic, singular events (Bashkirova and Lomvardas 2019, Maass, Barutcu et al. 2019), that are not readily detectable in Hi-C data (Maass, Barutcu et al. 2018, Zhang, Zhou et al. 2022).'

3. Line 75 – What are these orthogonal approaches?

We have added SPRITE, GAM, and HiPore-C to describe the orthogonal approaches to Hi-C.

4. Line 93 – "assumed a direct relationship of spatial proximity and inter-chromosomal ...", I think the authors are pre-empting their justification of the background model they fit. Their background model is fitted against the genomic distance variable in Fig 1A, but it is simpler to just say use the term "genomic coordinate". In contrast to intra -chromosomal contacts, where distance can be clearly defined with respect to the distance between two interaction anchors, in inter-chromosomal contacts, distance as a concept is not the same.

We fully agree with the reviewer. Therefore, we have re-ordered the section and rephrased our description. The part now reads as follows:

'Classic Hi-C analysis identifies significant intra-chromosomal interactions by taking the linear genomic distance between two interaction anchors into account (Sanyal, Lajoie et al. 2012). In contrast, distance as a concept for NHCCs is not the same. Thus, to define true NHCCs where loci of different chromosomes are in spatial proximity, our model is fitted against all genomic coordinates and evaluates inter-chromosomal Hi-C interaction weights between chromosomes. Specifically, we developed a non-parametric supervised learning approach (Local Weighted Polynomial Regression [LWPR]) that systematically assesses relationships between all loci on all chromosomes. LWPR queries each chromosomal position against all other genomic regions (1 megabase [Mb] bins) in an 'All bins vs. All bins' approach, which has not been accomplished for Hi-C analysis before (Fig. 1a, Extended Data Fig. 1a-e).'

5. Line 101 – Fig 1a, should be

We are unclear about the reviewer's question and/or concern, because the comment under point 5 seems to miss information [shown as '?']. Therefore, we kindly ask the editor to clarify this point with the reviewer.

6. Line 102 – "We cross-validated the span...".

Thank you for suggesting this text edit. The sentence now reads: '*To determine the best local regression fit, we cross-validated the span parameter.'*

7. Line 589 – 590 – To make the process computationally feasible, "LWLR retains data from local bins ..., rather than storing information for all regions, as performed in linear regression." This is a well-known fact about LOESS.

The reviewer is correct that this is a well-known fact about LOESS. Thus, we deleted this unnecessary detail of our method description.

8. Line 625 – A kind of weighted mean is being used, but it'd be more accurate to say just that the LOESS fitted values are being used?

We agree with the reviewer and now use '*LOESS fitted values'.*

9. Line 631 – The stated procedure for choosing the z-score can be justified as a minimum p-value pooling procedure.

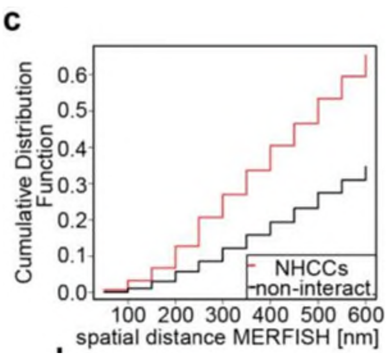
To avoid false positive results, we chose the more conservative *z*-score (i.e., closer to zero), which corresponds to a higher *p*-value. Therefore, we think that this is actually the opposite of a minimum *p*-value pooling procedure. Our approach is motivated by the high dimensionality of our data and the need to filter it as rigorously as possible.

10. Line 632 – There are obviously large deviations away from the normal distribution, but isn't that to be expected if there are true events that do not correspond to the null? So,

We have added the goodness-of-fit analysis in Extended Data Fig. 1d, which indicates that considering only one pair of chromosomes (chromosomes 12 and 17 from H9hESC_day00_Zhang) demonstrates a normal distribution, with both the theoretical and empirical density functions and cumulative distribution functions closely aligned. We also performed the Cramér-von Mises test to quantitatively assess the normality of the data, which confirmed its significant normality. However, we did not include the Cramér-von Mises test in the paper as it was outside the primary scope of the project. Additionally, the analysis in Extended Data Fig. 1d was limited to one chromosome pair, but our compendium comprises all-vs-all interactions. According to the Central Limit Theorem, with a sufficiently large sample size, the sampling distribution of the sample mean will approximate a normal distribution, regardless of the population distribution's shape. Therefore, the complete dataset used in our analysis is even closer to a normal distribution.

11. Line 135 – Please supply CDFs of the Merfish distances classified by NHCC vs non-interacting regions. It is difficult to appreciate the reported differences from the bar charts alone.

We thank the reviewer for suggesting the CDF plot, which indeed depicts the differences between NHCCs and non-interacting regions in MERFISH data better. We have edited Extended Data Fig. 3c and the corresponding legend.



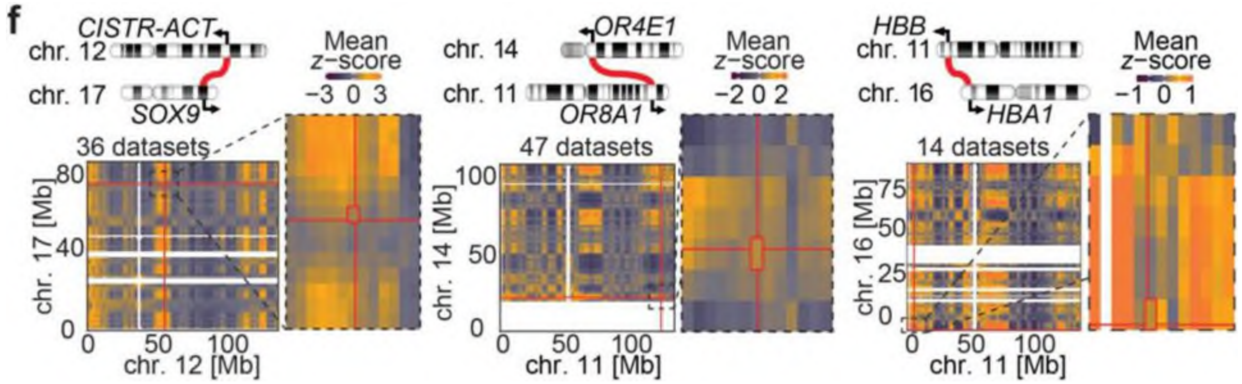
Extended Data Fig. 3c. *Cumulative Distribution Function (CDF) plot of number of Signature NHCCs and non-interacting regions (q < 0.05) over spatial proximities from MERFISH (Su, Zheng et al. 2020) (2x IMR90 datasets). Mann-Whitney test determined significance in raw data (p = 0.0083).*

12. Line 137-138 – Reported overlaps are really the recall rate? (116/130"90%,72/164"40%)

The reviewer is correct. Instead of overlap, it is more accurate to mention '*recall rates'*, instead of overlap (line 137-138). We have changed the text accordingly.

13. Line 143-145 – Fig 1j, Is there a way to make the blocks called significant obvious?

We appreciate the reviewer's comment on the visualization of the interacting genomic regions. Initially, we intended to present validated NHCC as two stacked heatmaps, one per chromosome of the interaction. However, we had made the decision to average out the datasets to represent both chromosomes of the NHCC in one plot (Figure 1f), which is more appropriate and correct.
To address the reviewer's comment, we think that enlarged magnifications in Figure 1f to show the bin(s) of the validated would make the most sense.



'Figure 1f. *Ideograms depict reported NHCCs tested by Signature. Each heatmap represents a pair of interacting chromosomes. Mean z-scores are shown and red lines indicate genomic positions of reported loci (shown in Mb = megabases). Enlargements highlight region of* *interest, each cell is a 1 Mb bin. Unmapped regions such as acrocentric p arm of chromosome 14 are shown in white.'*

14. Line 184 – For ease, could just mention in the figure legends that the expected contacts was based on chromosomal length.

We thank the reviewer for the valuable comment. We have added to figure 2e caption that contacts were tested '*based on chromosomal length'.*

15. Line 650 – Why can't interdependent data be clustered? This claim would imply that gene expression data should not be clustered because of dependencies between genes.

We think that our description led to a misunderstanding. We intended to say that interdependent data cannot be clustered by vector clustering, but not other approaches. We have rephrased the sentence to:

'Chromosomal interactions are interdependent data that cannot be clustered using vector clustering, justifying our use of CD for network clustering.'

16. Line 221 – Can the authors assess whether these findings hold true when they have matched RNA-seq data for which they can assess if NHCCs harbor most of the active genes?

We have addressed the question if matched pairs of RNAseq and Hi-C data show that NHCCs harbor most of the active genes. In line with the reviewer's comment, we have revised our text to explain our approach and findings better.

The text reads now as follows:

'Gene expression across all 62 datasets was significantly higher for common NHCCs when compared to all of GTEx(Consortium 2020) (Mann-Whitney test, range: $p = 2.79x10^{-15}$ - $p = 6.11x10^{-52}$, Fig. 3b). The same was true when we compared tissue-specific gene expression profiles from GTEx samples matching our Hi-C data by showing that NHCC regions harbor the most active genes (Extended Data Fig. 4a).'

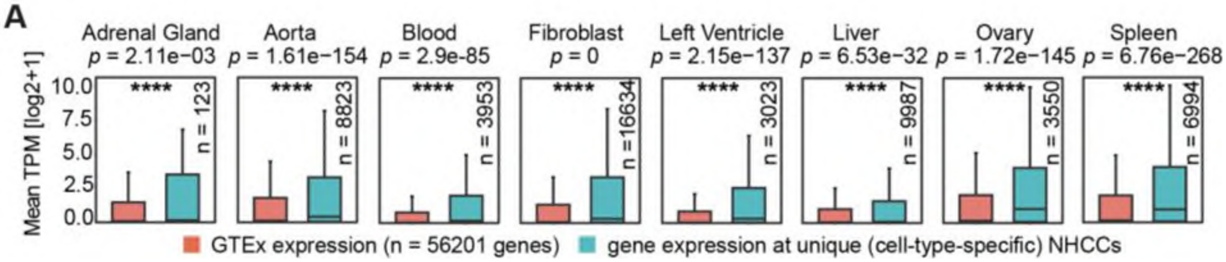


Figure S4a. *Comparison of mean gene expression shown as mean TPM (log2 +1) at unique (cell-type-specific) NHCCs (q < 0.05) with entire GTEx catalogue. n represents number of genes found*

in unique NHCCs. Box limits represent upper and lower quartiles. Central boxplot line represents the median and whiskers represent 1.5x IQR.

17. Line 227 – Could the authors define tissue specific NHCCs in line?

We acknowledge that clarification is required to describe our experiment properly. Specifically, we selected transcription factor (TF) binding profiles from ChIPseq experiments that matched the cell type of our Hi-C dataset to address the extent of TF binding at NHCCs.

We have rephrased the text to:

'Similarly, we compared transcription factor (TF) binding from ChIPseq experiments (Zou, Ohta et al. 2022) in cell types that matched our tissue-specific unique NHCCs. We found more TF binding at NHCCs than globally (Mann-Whitney test range: $p = 2.39x10^{-6}$ - $p = 1.15x10^{-71}$, Extended Data Fig. 4b).'

Reviewer #6 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We thank the reviewer for the thoughtful comments and ideas which have significantly improved the quality of our study. Please see our point-by-point responses above.

** See Nature Portfolio's author and referees' website at www.nature.com/authors for information about policies, services and author benefits.

References.

Abdennur, N. and L. A. Mirny (2020). "Cooler: scalable storage for Hi-C data and other genomically labeled arrays." Bioinformatics **36**(1): 311-316.

Akgol Oksuz, B., L. Yang, S. Abraham, S. V. Venev, N. Krietenstein, K. M. Parsi, H. Ozadam, M. E. Oomen, A. Nand, H. Mao, R. M. J. Genga, R. Maehr, O. J. Rando, L. A. Mirny, J. H. Gibcus and J. Dekker (2021). "Systematic evaluation of chromosome conformation capture assays." Nat Methods **18**(9): 1046-1055.

Bashkirova, E. and S. Lomvardas (2019). "Olfactory receptor genes make the case for inter-chromosomal interactions." Curr Opin Genet Dev **55**: 106-113.

Bastian, M., S. Heymann and M. Jacomy (2009). "Gephi: An Open Source Software for Exploring and Manipulating Networks." Proceedings of the International AAAI Conference on Web and Social Media **3**(1): 361-362.

Chen, Y., Y. Zhang, Y. Wang, L. Zhang, E. K. Brinkman, S. A. Adam, R. Goldman, B. v. Steensel, J. Ma and A. S. Belmont (2018). "TSA-Seq Mapping of Nuclear Genome Organization." bioRxiv **Preprint Posted April 25, 2018**.

Consortium, G. T. (2020). "The GTEx Consortium atlas of genetic regulatory effects across human tissues." Science **369**(6509): 1318-1330.

Dekker, J. (2016). "Mapping the 3D genome: Aiming for consilience." Nat Rev Mol Cell Biol **17**(12): 741-742.

Dekker, J., F. Alber, S. Aufmkolk, B. J. Beliveau, B. G. Bruneau, A. S. Belmont, L. Bintu, A. Boettiger, R. Calandrelli, C. M. Disteche, D. M. Gilbert, T. Gregor, A. S. Hansen, B. Huang, D. Huangfu, R. Kalhor, C. S. Leslie, W. Li, Y. Li, J. Ma, W. S. Noble, P. J. Park, J. E. Phillips-Cremins, K. S. Pollard, S. M. Rafelski, B. Ren, Y. Ruan, Y. Shav-Tal, Y. Shen, J. Shendure, X. Shu, C. Strambio-De-Castillia, A. Vertii, H. Zhang and S. Zhong (2023). "Spatial and temporal organization of the genome: Current state and future aims of the 4D nucleome project." Mol Cell **83**(15): 2624-2640.

Hoencamp, C., O. Dudchenko, A. M. O. Elbatsh, S. Brahmachari, J. A. Raaijmakers, T. van Schaik, A. Sedeno Cacciatore, V. G. Contessoto, R. van Heesbeen, B. van den Broek, A. N. Mhaskar, H. Teunissen, B. G. St Hilaire, D. Weisz, A. D. Omer, M. Pham, Z. Cola ric, Z. Yang, S. S. P. Rao, N. Mitra, C. Lui, W. Yao, R. Khan, L. L. Moroz, A. Kohn, J. St Leger, A. Mena, K. Holcroft, M. C. Gambetta, F. Lim, E. Farley, N. Stein, A. Haddad, D. Chauss, A. S. Mutlu, M. C. Wang, N. D. Young, E. Hildebrandt, H. H. Cheng, C. J. Knight, T. L. U. Burnham, K. A. Hovel, A. J. Beel, P. J. Mattei, R. D. Kornberg, W. C. Warren, G. Cary, J. L. Gomez-Skarmeta, V. Hinman, K. Lindblad-Toh, F. Di Palma, K. Maeshima, A. S. Multani, S. Pathak, L. Nel-Themaat, R. R. Behringer, P. Kaur, R. H. Medema, B. van Steensel, E. de Wit, J. N. Onuchic, M. Di Pierro, E. Lieberman Aiden and B. D. Rowland (2021). "3D genomics across the tree of life reveals condensin II as a determinant of architecture type." Science **372**(6545): 984989.

Jacomy, M., T. Venturini, S. Heymann and M. Bastian (2014). "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software." PLoS One **9**(6): e98679.

Joo, J., S. Cho, S. Hong, S. Min, K. Kim, R. Kumar, J. M. Choi, Y. Shin and I. Jung (2023). "Probabilistic establishment of speckle-associated inter-chromosomal interactions." Nucleic Acids Res **51**(11): 5377-5395.

Krietenstein, N., S. Abraham, S. V. Venev, N. Abdennur, J. Gibcus, T. S. Hsieh, K. M. Parsi, L. Yang, R. Maehr, L. A. Mirny, J. Dekker and O. J. Rando (2020). "Ultrastructural Details of Mammalian Chromosome Architecture." Mol Cell **78**(3): 554-565 e557.

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander and J. Dekker (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." Science **326**(5950): 289-293.

Maass, P. G., A. R. Barutcu and J. L. Rinn (2019). "Interchromosomal interactions: A genomic love story of kissing chromosomes." J Cell Biol **218**(1): 27-38.
Maass, P. G., A. R. Barutcu, D. M. Shechner, C. L. Weiner, M. Mele and J. L. Rinn (2018). "Spatiotemporal allele organization by allele-specific CRISPR live-cell imaging (SNP-CLING)." Nat Struct Mol Biol **25**(2): 176-184.
Maass, P. G., A. R. Barutcu, C. L. Weiner and J. L. Rinn (2018). "Inter-chromosomal Contact Properties in Live-Cell Imaging and in Hi-C." Mol Cell **69**(6): 1039-1045 e1033.
Nguyen, H. Q., S. Chattoraj, D. Castillo, S. C. Nguyen, G. Nir, A. Lioutas, E. A. Hershberg, N. M. C. Martins, P. L. Reginato, M. Hannan, B. J. Beliveau, G. M. Church, E. R. Daugharthy, M. A. Marti-Renom and C. T. Wu (2020). "3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing." Nat Methods **17**(8): 822-832. Park, D. S., S. C. Nguyen, R. Isenhart, P. P. Shah, W. Kim, R. J. Barnett, A. Chandra, J. M. Luppino, J. Harke, M. Wai, P. J. Walsh, R. J. Abdill, R. Yang, Y. Lan, S. Yoon, R. Yunker, M. T. Kanemaki, G. Vahedi, J. E. Phillips-Cremins, R. Jain and E. F. Joyce (2023). "High-throughput Oligopaint screen identifies druggable 3D genome regulators." Nature **620**(7972): 209-217. Payne, A. C., Z. D. Chiang, P. L. Reginato, S. M. Mangiameli, E. M. Murray, C. C. Yao, S. Markoulaki, A. S. Earl, A. S. Labade, R. Jaenisch, G. M. Church, E. S. Boyden, J. D. Buenrostro and F. Chen (2021). "In situ genome sequencing resolves DNA sequence and structure in intact biological samples." Science **371**(6532).
Quinodoz, S. A., N. Ollikainen, B. Tabak, A. Palla, J. M. Schmidt, E. Detmar, M. M. Lai, A. A. Shishkin, P. Bhat, Y. Takei, V. Trinh, E. Aznauryan, P. Russell, C. Cheng, M. Jovanovic, A. Chow, L. Cai, P. McDonel, M. Garber and M. Guttman (2018). "Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus." Cell.
Sanyal, A., B. R. Lajoie, G. Jain and J. Dekker (2012). "The long-range interaction landscape of gene promoters." Nature **489**(7414): 109-113.
Su, J. H., P. Zheng, S. S. Kinrot, B. Bintu and X. Zhuang (2020). "Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin." Cell **182**(6): 1641-1659 e1626.
Takei, Y., J. Yun, S. Zheng, N. Ollikainen, N. Pierson, J. White, S. Shah, J. Thomassie, S. Suo, C. L. Eng, M. Guttman, G. C. Yuan and L. Cai (2021). "Integrated spatial genomics reveals global architecture of single nuclei." Nature **590**(7845): 344-350.
Zhang, R., T. Zhou and J. Ma (2022). "Multiscale and integrative single-cell Hi-C analysis with Higashi." Nat Biotechnol **40**(2): 254-261.
Zhang, Y., T. Li, S. Preissl, M. L. Amaral, J. D. Grinstein, E. N. Farah, E. Destici, Y. Qiu, R. Hu, A. Y. Lee, S. Chee, K. Ma, Z. Ye, Q. Zhu, H. Huang, R. Fang, L. Yu, J. C. Izpisua Belmonte, J. Wu, S. M. Evans, N. C. Chi and B. Ren (2019). "Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells." Nat Genet **51**(9): 1380-1388.
Zhong, J. Y., L. Niu, Z. B. Lin, X. Bai, Y. Chen, F. Luo, C. Hou and C. L. Xiao (2023). "High-throughput Pore-C reveals the single-allele topology and cell type-specificity of 3D genome folding." Nat Commun **14**(1): 1250.
Zhou, Y., B. Zhou, L. Pache, M. Chang, A. H. Khodabakhshi, O. Tanaseichuk, C. Benner and S. K. Chanda (2019). "Metascape provides a biologist-oriented resource for the analysis of systems-level datasets." Nat Commun **10**(1): 1523.

Zou, Z., T. Ohta, F. Miura and S. Oki (2022). "ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data." Nucleic Acids Res.

REVIEWER COMMENTS

Reviewer #2 (Remarks to the Author):

In the revised version of the manuscript authors addressed and clarified several of my previous comments which helped with the clarity and understanding of the manuscript. I still find it very cluttered and not easy to follow, but inter-chromosomal interactions are not much explored by the community and this work will provide useful tool. However, one of the points are still not clear to me.
Specifically, my questions regarding novel detected structures and general usage of the developed software is not addressed. I was asking about novel structures that might have potential functions and not contacts that were not detected by other methods (that is what revised version and rebuttal is referring to). For example, do authors see new higher-order structures that suggest existence with novel nuclear bodies not described previously? I think it is an important point that will help to appreciate potential power of this tool.

We would like to thank the reviewer for the additional comments on our manuscript. We acknowledge the idea of determining novel subnuclear structures with our approach. To accomplish this and to exemplify Signature's power to yield new insight into genome structure, we first calculated the correlation between gene expression and transcription factor binding across genome topology (62 Hi-C datasets). The overall correlation was 0.6945, although the pattern varied across the genome topology map.
Then, we focused on regions exhibiting either correlated or anticorrelated patterns between gene expression and transcription factor binding. Indeed, we identified specific local hubs that suggest novel nuclear compartments. For example, one region of NHCCs with anticorrelated gene expression and TF binding harbors many TFs whilst genes are lowly expressed, indicating a reservoir of TFs. The annotated genes of this region are highly significantly enriched in deacetylated histones (log10[*q*] = -67.13). We have added our findings to the manuscript and we have highlighted the potential of combining *Signature* data with additional genomic features to determine novel subnuclear structures. We have modified the genome topology maps in figures 3c and 3d:

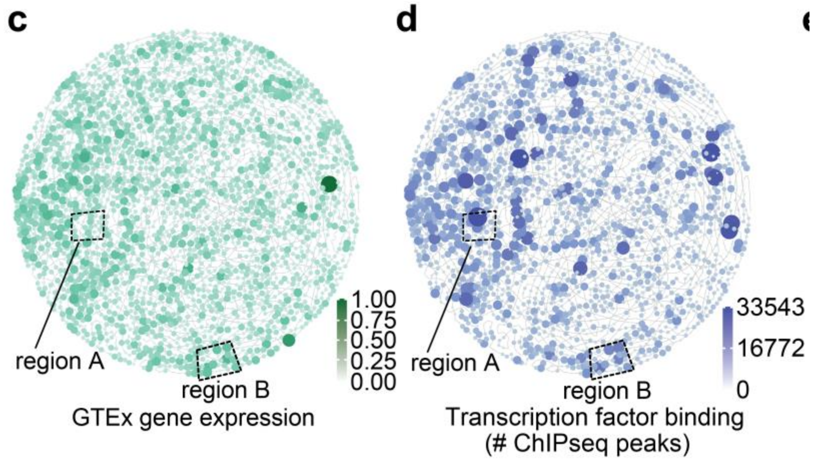


Figure 3c legend: Scaled average gene expression of GTex[52] per 1 Mb bin across genome topology (high [green] *vs.* low expression [white]). Dashed boxes indicate regions with either an anticorrelated pattern of expression and transcription factor (TF) binding (region A) or a correlated pattern (region B).

The text reads as follows:
'Remarkably, expression and TF binding were highly correlated (r = 0.6945), although not in a consistent pattern across the genome. Locally, we found regions with anticorrelated expression and TF binding pattern in the genome topology map. For example, region A (r = -0.161) harbors many bound TFs, whilst its genes are lowly expressed and relate to HDAC deacetylated histones (log10[q] = -67.13, Fig. 3c-d), indicating a compartment where TFs accumulate. In contrast, othe*r regions showed high correlation between expression and TF binding, such as region B where genes related to sensory perception function (r = 0.849, Fig. 3c-d). The combination of Signature results with other genomic features in genome topology maps, such as expression and ChIPseq data may help to reveal novel higher-order subnuclear structures and their function.'*

Discussion:

'Combining Signature outputs with additional genomic features (i.e., expression and ChIPseq data, methylation and acetylation pattern, recombination frequencies, etc.) in genome topology maps may complement Signature-derived data interpretation and identify novel subnuclear structures.'

Reviewer #3 (Remarks to the Author):

The authors addressed all my questions.

Thank you!

Reviewer #5 (Remarks to the Author):

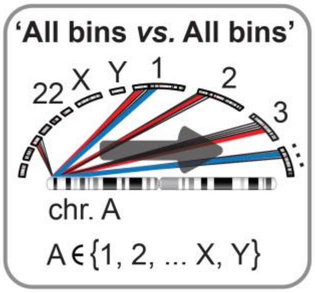
The authors have addressed all major points from the previous review. Major methodological questions have been sufficiently addressed, and the illustration and further descriptions regarding their clustering procedure and "domain" construction procedure are appreciated. Further investigations into NHCCs in the context of specific tissues and cell types were appreciated as well.

Thank you!

Minor:

For Figure 1A, instead of an epsilon symbol, a standard set membership symbol could be used.

We have modified figure 1A and added a membership symbol.



Reviewer #6 (Remarks to the Author):
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Thank you!