Inter-chromosomal contacts demarcate genome topology along a spatial gradient

Supplementary information.

Milad Mokhtaridoost[1], Jordan J. Chalmers[1,2], Marzieh Soleimanpoor[1], Brandon J. McMurray[1], Daniella F. Lato[1], Son C. Nguyen[3,4], Viktoria Musienko[5], Joshua O Nash[1,6], Sergio Espeso-Gil[1], Sameen Ahmed[1,2], Kate Delfosse[1], Jared W.L. Browning[1,2], A. Rasim Barutcu[7], Michael D. Wilson[1,2], Thomas Liehr[5], Adam Shlien[1,6], Samin Aref[8], Eric F. Joyce[3,4], Anja Weise[5], Philipp G. Maass[1,2,9]

[1] Genetics and Genome Biology Program, SickKids Research Institute, Toronto, ON, M5G 0A4, Canada

[2] Department of Molecular Genetics, University of Toronto, Toronto, ON, M5S 1A8, Canada

[3] Penn Epigenetics Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

[4] Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

[5] Jena University Hospital, Friedrich Schiller University, Institute of Human Genetics, Am Klinikum 1, 07747, Jena, Germany

[6] Laboratory of Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada

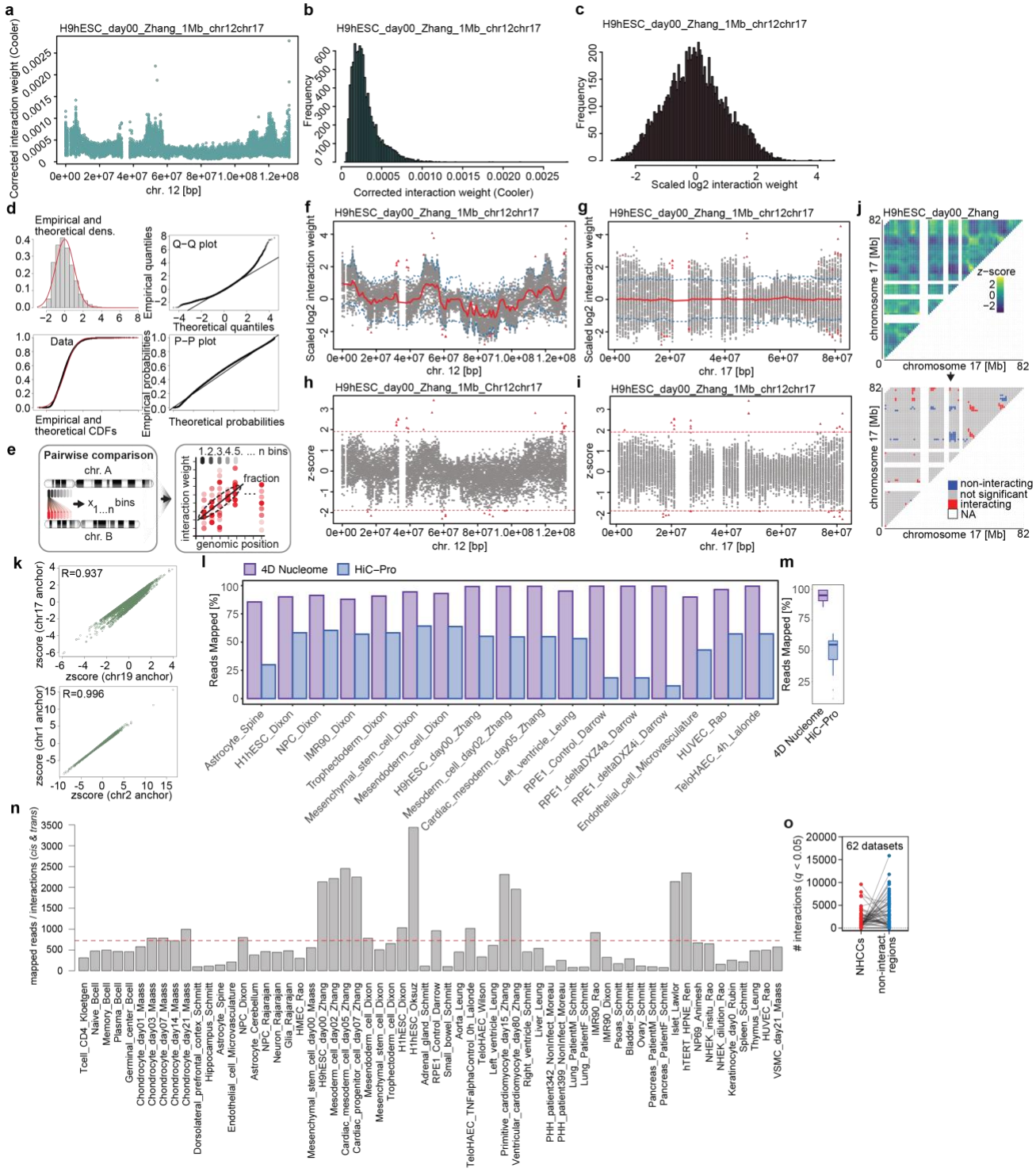
[7] Donnelly Centre, University of Toronto, Toronto, ON, M5S 3E1, Canada

[8] Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, M5S3G8, Canada

[9] corresponding author: philipp.maass@sickkids.ca

Inter-chromosomal contacts demarcate genome topology along a spatial gradient

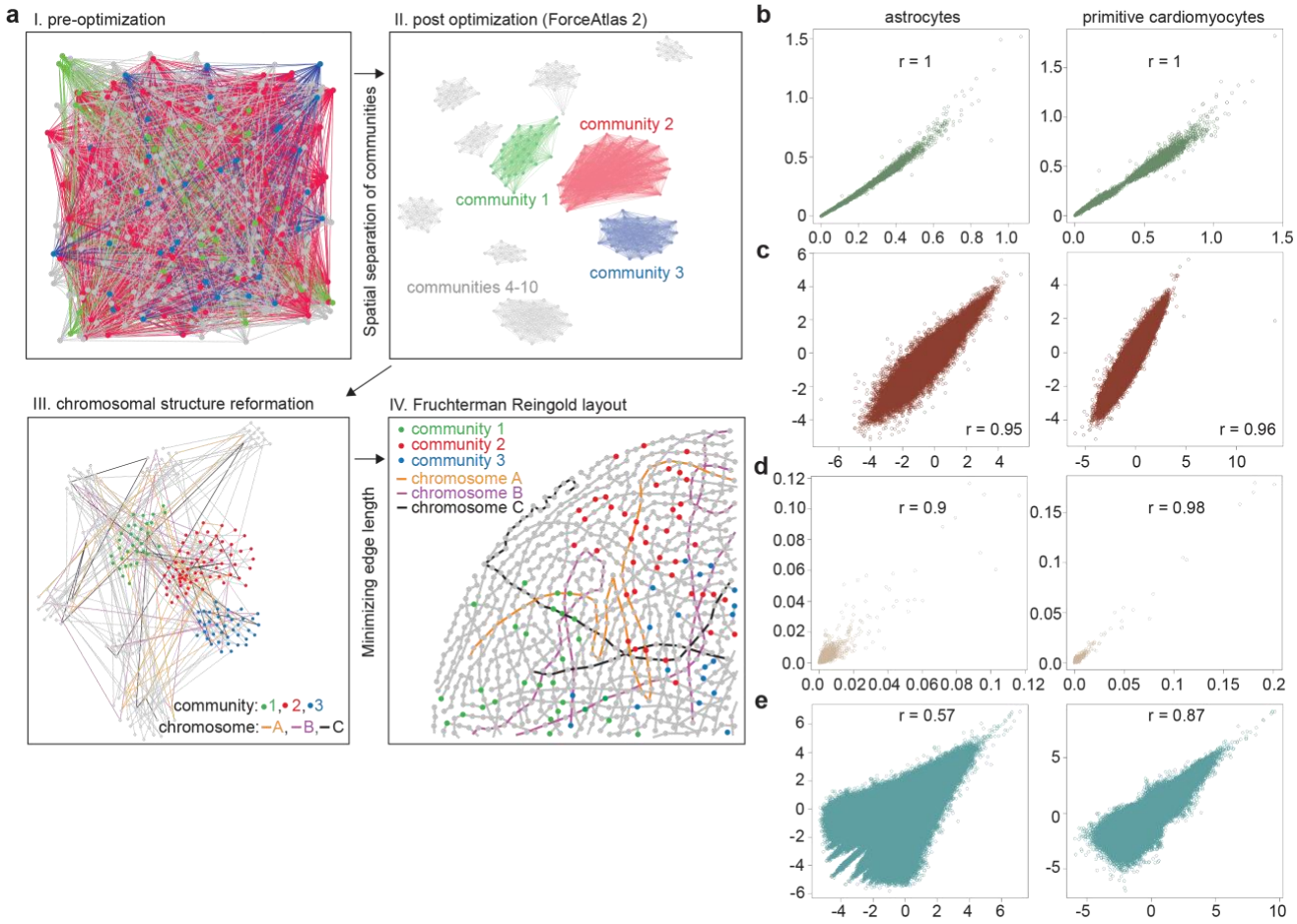
Supplementary Figure 1.



a. Examples of interaction weights between chromosomes (chr.) 12 and 17 of H9hESC_day00_Zhang dataset[1] are visualized along the linear genomic position of chromosome 12. **b.** Representation of the distribution of interaction weights of data shown in panel a. **c.** Scaled log2-transformed interaction weights of data shown in panel b. **d.** Goodness-of-fit graphs confirm the normal distribution of scaled log2-transformed interaction weights. **e.** In addition to '*All bins vs. All bins*', *Signature* can perform pairwise chromosomal comparisons as well (excluding the entire genomic background). **f.** Scaled log2-

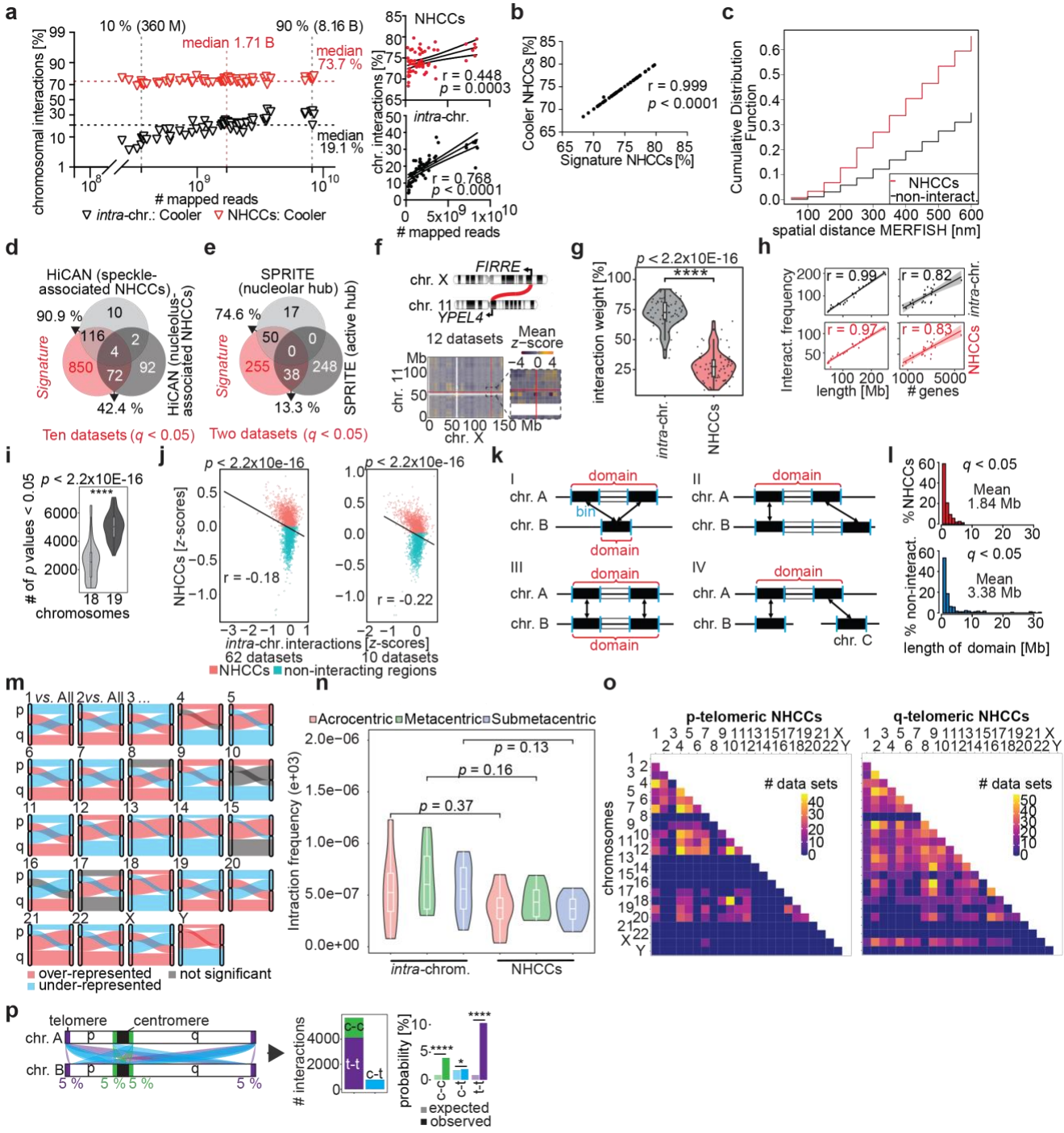
transformed pairwise interaction weights for chromosome 12-17 interactions, plotted along the linear genomic positions of chromosome 12 and (**g**) chromosome 17. The red line depicts the estimated weighted mean interaction across the linear genomic coordinates of the anchor chromosome. Blue dashed lines indicate the lower and upper bounds of the significance level. Red triangles highlight interactions that remain significant when considering both chromosomes 12 and 17 as the anchor chromosomes. **h.** Depiction of *z*-scores for the chromosome 12-17 interactions, presented along the linear genomic positions of chromosome 12 and (**i**) chromosome 17. Dashed red lines indicate the significance cutoff, while red triangles denote interactions identified as significant. Positive *z*-scores are considered as NHCCs, whilst negative *z*-scores describe significantly non-interacting regions. **j.** *Signature* determines *intra*-chromosomal interaction weights according to the method described by Sanyal et al. 2012[2] 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[1]. 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. **k.** Pearson correlations of *z*-scores of either 'anchor' chromosome in pairwise comparisons are depicted with correlation coefficients. **l.** The 4D Nucleome approach mapped more reads than Hi-C Pro[3] across various datasets. (solely 'RPE1_control' from Darrow *et al.* was used for the body map). **m.** Data of panel k summarized as boxplot. Central boxplot line represents the median and whiskers represent 1.5x IQR. **n.** Average number (dashed red line) of 724.3 mapped reads per interaction (*cis & trans*) shown for 1 Mb bins across all datasets. **o.** Significant NHCCs and non-interacting regions ($q < 0.05$) across 62 Hi-C datasets are shown.

Supplementary Figure 2.



a. 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[4] 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[5]. 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. **b.** Pearson correlation of *intra*-chromosomal (*in cis*) interaction weights between biological replicates of two previously published Hi-C datasets.[1,6] Pearson's rho (r) is denoted. **c.** same as panel b, but *z*-scores are shown. **d.** same datasets as shown in panel b but correlation of *inter*-chromosomal contact weights is depicted. **e.** same as panel d. Correlations of *z*-scores for NHCCs ('All *vs.* All') are depicted. *Signature* effectively captures trends within neighboring genomic positions, indicating that a higher interaction weight does not necessarily translate to higher *z*-scores, as demonstrated in panel e.

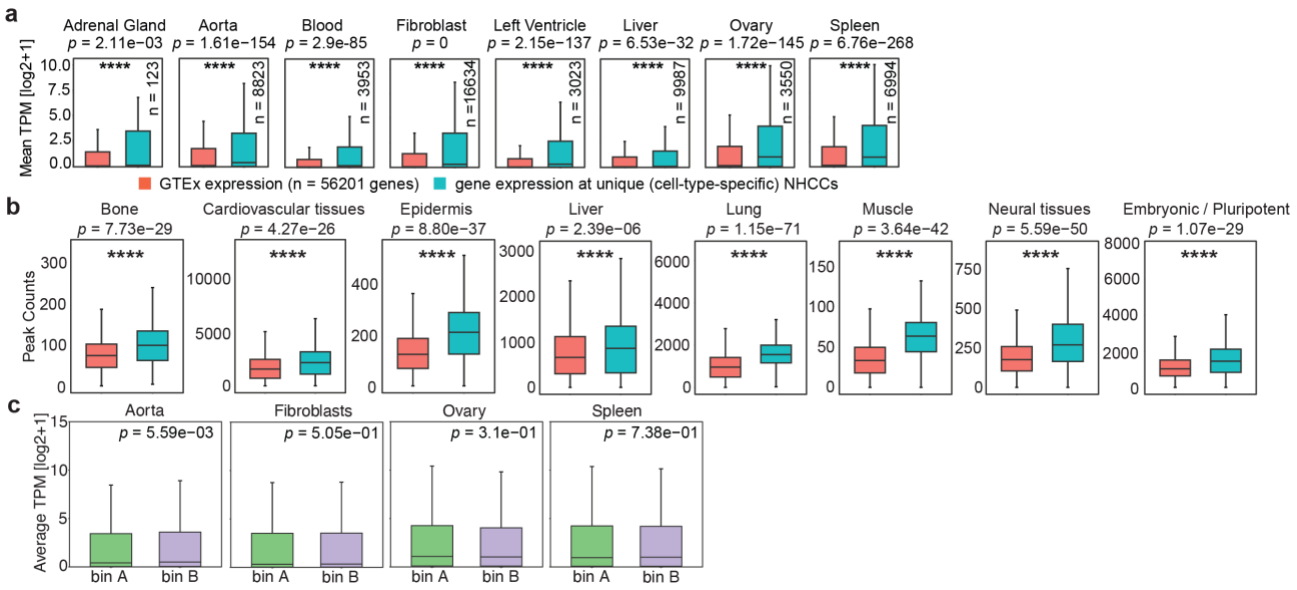
Supplementary Figure 3.



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*

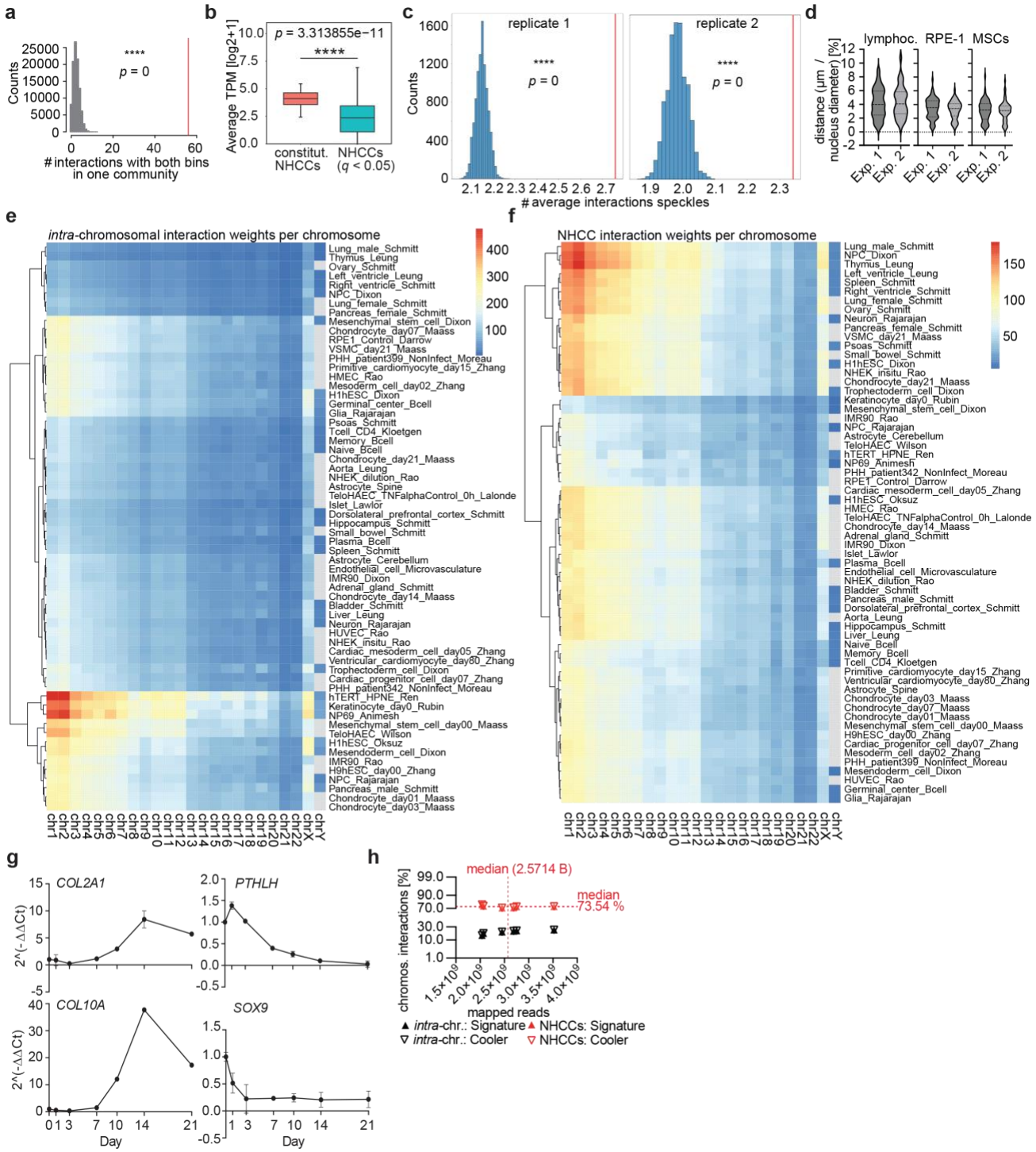
detected a median detection rate of 73.7 % for NHCCs ([range 68.4 – 79.8 %], even in datasets with low sequencing depth (Pearson correlation r = 0.45, p = 0.0003), indicating recurrent NHCCs. Ultra-deep sequencing with ˜8 - 12 billion (B) reads per sample (i.e., cardiomyogenesis and H1-ESCs)[1,7] only led to a marginal increase of the NHCC detection rate (79.8 %). **b.** Comparison of NHCCs (%) in Cooler[8] and *Signature* (mean Δ0.07 %) outputs by Pearson correlation (r = 0.999, p < 0.0001). Each datapoint represents one Hi-C dataset. **c.** Cumulative Distribution Function (CDF) plot of number of *Signature* NHCCs and non-interacting regions (q < 0.05) over spatial proximities from MERFISH[9] (2x IMR90 datasets). Mann-Whitney test determined significance in raw data (p = 0.0083). **d.** Comparison of Signature NHCCs to HiCAN[10] (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[11] (2x GM12878, q < 0.05). Signature identified 255 unique bins that determined 1034 interactions not detected by SPRITE. **f.** Ideograms depict reported NHCCs between *FIRRE* and *YPEL4.* The heatmap represents interacting chromosomes X and 14. Red line indicates genomic positions of reported loci along chromosome length (shown in Mb = megabases). Each cell is a 1 Mb bin. Unmapped regions of acrocentric p arm of chromosome 14 are shown in white. Mean z-scores are shown **g.** Interaction weights across 62 Hi-C datasets for *intra*-chromosomal interactions (mean 72 %, range: 34.7 – 92 %) in comparison to NHCCs (mean 28 %, range: 8 – 65.3 %). Mann-Whitney test determined significance ($p < 2.22 \times 10^{-16}$). Box limits represent upper and lower quartiles. Central boxplot line represents the median and whiskers represent 1.5x IQR. **h.** Pearson correlations with correlation coefficients and p-values of (left) *intra*-chromosomal contacts ([grey], r = 0.99, $p < 2.2 \times 10^{-16}$) and NHCCs ([red], r = 0.97, $p = 1.2 \times 10^{-15}$) with chromosome length and with (right) number of genes per chromosome (*intra*-chromosomal interactions: r = 0.82, $p = 1 \times 10^{-6}$; NHCCs: r = 0.83, $p = 4.5 \times 10^{-7}$). **i.** Total number of NHCCs (p < 0.05) of chromosomes 18 and 19. Mann Whitney testing determined significance $p < 2.2 \times 10^{-16}$. Box limits represent upper and lower quartiles. Central boxplot line represents the median and whiskers represent 1.5x IQR. **j.** 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, $p < 2.2 \times 10^{-16}$). 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, $p < 2.2 \times 10^{-16}$). **k.** Schematic overview of how NHCC domains were determined (Methods). **l.** Average NHCC domain size in megabases (Mb) is shown for significantly interacting domains (left, q < 0.05) and for non-interacting domains (right, q < 0.05). **m.** Binomial testing of p-p, p-q, and q-q-arm interactions of single chromosomes *vs.* all other chromosomes. Over-represented (red), under-represented (blue), and not significant (grey) interactions are shown. **n.** Interaction weights (Cooler) for *intra*-chromosomal interactions of acrocentric (red), metacentric (green), and submetacentric (blue) chromosomes in comparison to NHCCs across 62 Hi-C datasets. Mann Whitney testing tested for significance (p values are shown above violin plots). Box limits represent upper and lower quartiles. Central boxplot line represents the median and whiskers represent 1.5x IQR. **o.** Heatmaps show distribution of significant p- and q-telomeric NHCCs for each chromosomal pair across 62 Hi-C datasets. **p.** Scheme for testing telomeric (t-t, purple), centromeric (c-c, green), and centromeric-telomeric (c-t, blue) NHCCs in up to 5 % of mapped sequence per region. Bar plots show (left) number of interactions and (right) expected and observed probabilities (%) across 62 Hi-C datasets, n = 24 conditions per group (22 autosomes and two gonosomes). Binomial testing determined statistical significance (**** p = 0, * p = 0.017).

Supplementary Figure 4.



a. 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. **b.** Comparison of transcription factor binding shown as summed ChIPseq peaks[12] of tissues that match with Hi-C datasets (NHCCs = blue; genome = red). Box limits represent upper and lower quartiles. Central boxplot line represents the median and whiskers represent 1.5x IQR. **c.** In the context of *Signature's* design, each NHCC generated two *z*-scores, with each bin of a pairwise interaction serving as the anchor ('*predictor for regression*'). The final *z*-score for each interaction was determined by considering the lower *z*-score, chosen as the more conservative score (depicted as bin B in the figure, see also Methods). Utilizing GTex[13], we find that the gene expression levels in the affiliated genomic regions of both bins did not exhibit a significant difference. This finding also certifies that the selection of the more conservative *z*-score in *Signature* does not impact NHCCs. Box limits represent upper and lower quartiles. Central boxplot line represents the median and whiskers represent 1.5x IQR.

Supplementary Figure 5.

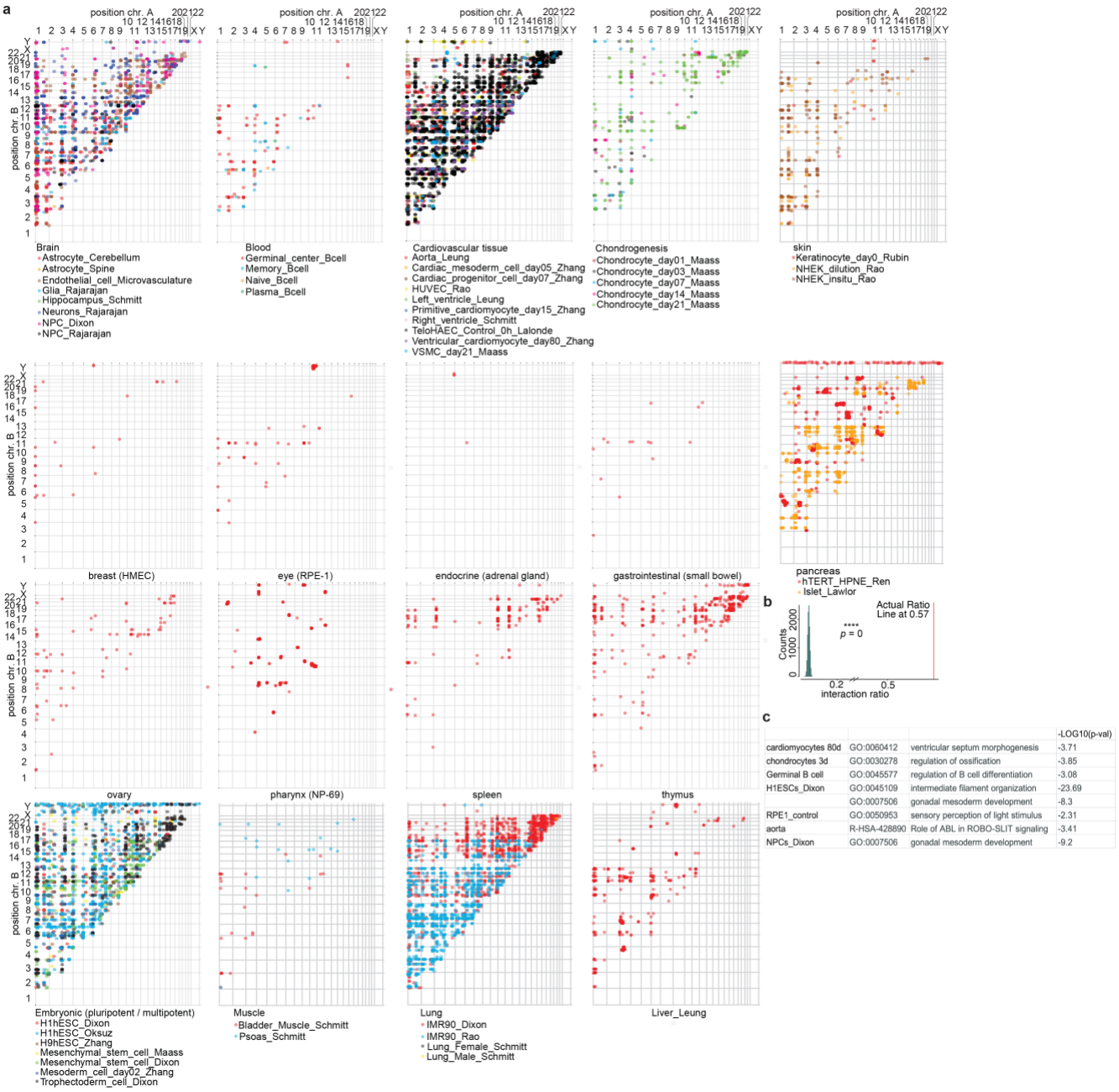


a. Permutation test of overlap between LWPR and community detection in comparison to random overlap (empirical $p = 0$). **b.** Comparison of mean gene expression shown as mean TPM (log2 +1) of largest community of 61 constitutive NHCCs with gene expression of all significant NHCCs ($q < 0.05$). Asterisk depicts significance determined by Mann Whitney testing. Outliers are excluded. Box limits represent upper and lower quartiles. Central boxplot line represents the median and whiskers represent 1.5x IQR. **c.** Permutation tests of associations between NHCCs and speckles. Two independent replicates are shown. Asterisks depict significance (empirical $p = 0$). **d.** Distances between proximal FISH signals of

two different experiments (Exp. 1 & 2, each ˜100 nuclei) are shown in relation to the corresponding nucleus sizes (lymphocytes: mean Exp.1 = 4.18 % = 438 nm, mean Exp.2 = 4.37 % = 559 nm; RPE-1: mean Exp.1 = 3.44 % = 485 nm, mean Exp.2 = 3.23 % = 400 nm; MSCs: mean Exp.1 = 3.15 % = 352 nm, mean Exp.2 = 3.1 % = 372 nm). Lines in violins represent median and quartiles. **e.** Hierarchical clustering of *intra*-chromosomal interactions of 62 Hi-C datasets across chromosomes. Grey cells in chromosome Y column depict female datasets. **f.** same as panel e but for NHCCs. **g.** RT-qPCRs of chondrogenic markers (*COL2A1, COL10A1, PTHLH,* and *SOX9*) across chondrogenic differentiation (two biological replicates). **h.** Detection rates of *intra*-chromosomal interactions (black) and NHCCs (red) of chondrogenic differentiation (two biological replicates) either after Cooler (light triangles) or after Cooler and *Signature* (filled triangles) over mapped reads. Median of mapped reads and median of detected NHCCs are shown.

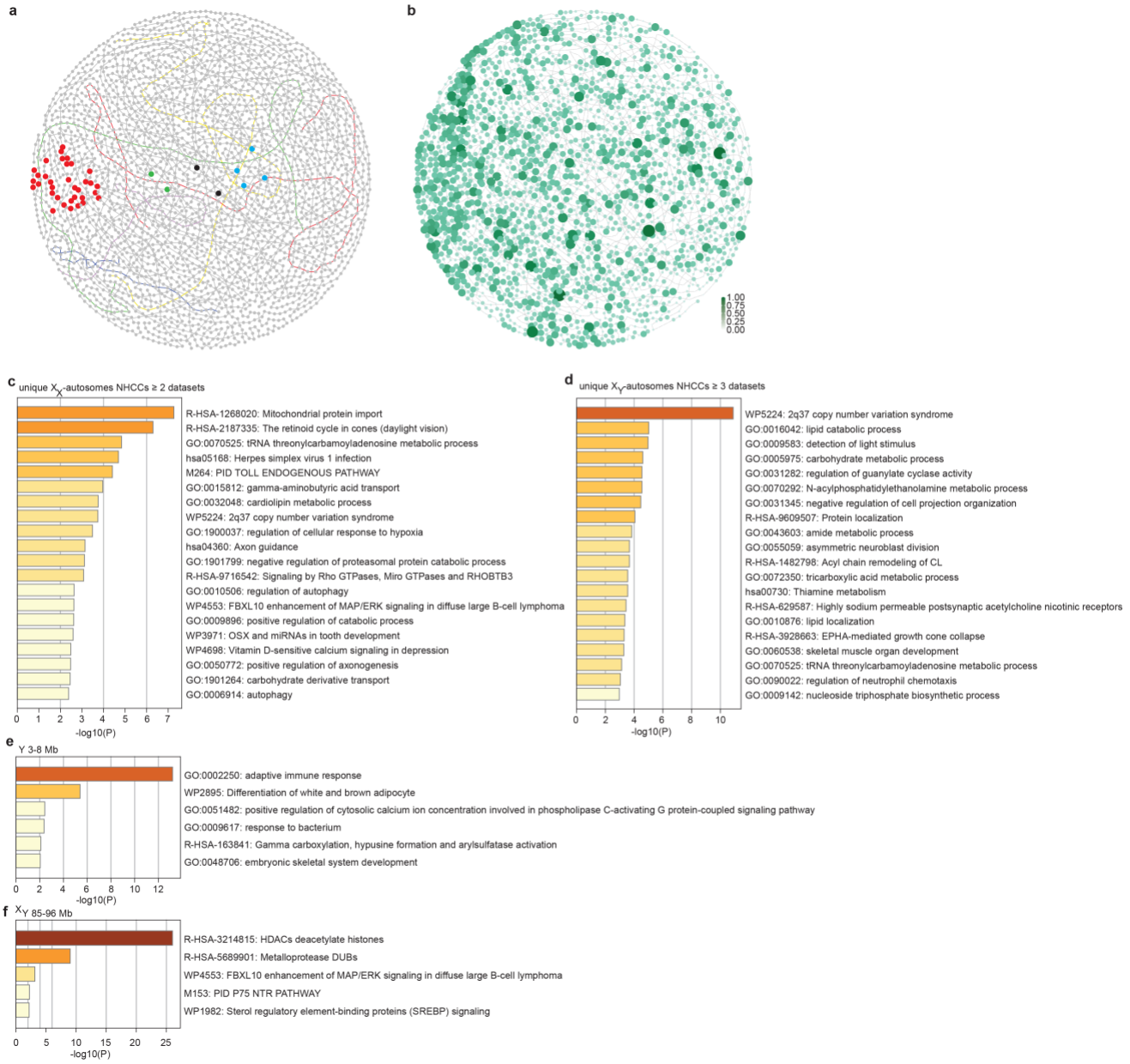
Inter-chromosomal contacts demarcate genome topology along a spatial gradient

Supplementary Figure 6.



a. Unique (cell-type-specific and/or tissue-type-specific, respectively) NHCCs (q < 0.05) are shown (opacity = 50 %). **b.** Randomization test of unique (dataset-specific) NHCCs (q < 0.05) in comparison to randomly selected bins. Asterisks depict significance (empirical p = 0). **c.** Examples of GO terms of genes of significantly non-interacting unique regions.

Supplementary Figure 7.



a. 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[13] per 1 Mb bin across genome topology of 10 cardiovascular Hi-C datasets shown in panel a (high [green] vs. low expression [white]). **c.** Functional gene ontology (GO) analysis of genes annotated at unique X_X-autosome interactions determined in ≧ 2 datasets. **d.** GO analysis for unique X_Y-autosome interactions determined in ≧ 3 datasets. **e.** GO analysis for genes involved in Y-X interactions in male cells within chromosome Y 3-8 Mb range (hg38). **f.** GO analysis for genes involved in Y-X interactions in male cells within chromosome X 85-96 Mb range (hg38).

References.

1 Zhang, Y. *et al.* Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat Genet* **51**, 1380-1388 (2019). https://doi.org:10.1038/s41588-019-0479-7

2 Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-113 (2012). https://doi.org:10.1038/nature11279

3 Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16**, 259 (2015). https://doi.org:10.1186/s13059-015-0831-x

4 Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One* **9**, e98679 (2014). https://doi.org:10.1371/journal.pone.0098679

5 Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the International AAAI Conference on Web and Social Media* **3**, 361-362 (2009). https://doi.org:10.1609/icwsm.v3i1.13937

6 Consortium, E. P. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710 (2020). https://doi.org:10.1038/s41586-020-2493-4

7 Akgol Oksuz, B. *et al.* Systematic evaluation of chromosome conformation capture assays. *Nat Methods* **18**, 1046-1055 (2021). https://doi.org:10.1038/s41592-021-01248-7

8 Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311-316 (2020). https://doi.org:10.1093/bioinformatics/btz540

9 Su, J. H., Zheng, P., Kinrot, S. S., Bintu, B. & Zhuang, X. Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. *Cell* **182**, 1641-1659 e1626 (2020). https://doi.org:10.1016/j.cell.2020.07.032

10 Joo, J. *et al.* Probabilistic establishment of speckle-associated inter-chromosomal interactions. *Nucleic Acids Res* **51**, 5377-5395 (2023). https://doi.org:10.1093/nar/gkad211

11 Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* (2018). https://doi.org:10.1016/j.cell.2018.05.024

12 Zou, Z., Ohta, T., Miura, F. & Oki, S. 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* (2022). https://doi.org:10.1093/nar/gkac199

13 Consortium, G. T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020). https://doi.org:10.1126/science.aaz1776