scientific **data**

OPEN

DATA DESCRIPTOR

Check for updates

Hi-C sequencing data from frontal cortex of laboratory rats

Panjun Kim[1], Rachel R. Ward[1], Burt M. Sharp[1], Robert W. Williams[1] & Hao Chen [2] ✉

The three-dimensional structure of chromosomes is integral to nuclear organization and influences processes such as DNA replication, repair, and gene expression. Hi-C methods generate high-resolution, genome-wide data that map physical interactions between different parts of the chromatin. These data have applications in *de novo* genome assembly, the detection of structural variants, and the analysis of gene regulatory mechanisms. This report describes a Hi-C dataset from the frontal cortex of laboratory rats. The dataset includes a diverse panel of inbred strains (SHR/OlaIpcv, BN-Lx/Cub, BXH6/Cub, HXB2/Ipcv, HXB10/Ipcv, HXB23/Ipcv, HXB31/Ipcv, LE/Stm, F344/Stm) and an F1 hybrid (SHR/OlaIpcv × BN/NHsdMcwi). This data set provides a valuable resource for advancing rat genomics and offers opportunities for integration with other omic datasets generated using this widely used model organism.

Background & Summary

High-throughput chromosome conformation capture (Hi-C) is a technique for studying the three-dimensional (3D) organization of the genome[1]. The method identifies regions of the genome that are physically close to one another inside the cell nucleus. This is achieved by first crosslinking the chromatin to fix its 3D structure. The DNA is then digested, and fragments that are in close proximity are ligated together. These newly joined DNA molecules, which represent interacting genomic regions, are identified through next-generation paired-end sequencing[2].

Hi-C data are frequently used for *de novo* genome assembly. The frequency of interactions between DNA fragments provides information about their spatial proximity, which allows smaller sequenced fragments (contigs) to be ordered and oriented into chromosome-length assemblies. This improves the contiguity and accuracy of the resulting genome. The approach is effective for assembling complex genomes, such as those of polyploid species[3], and can also assist in separating the contributions of homologous chromosomes in a process known as haplotype phasing[4].

Additionally, Hi-C data facilitate the detection of large-scale structural variants (SVs), such as deletions, duplications, insertions, inversions, and translocations. These variants, which span more than 50 base pairs[5], can be difficult to identify with standard short-read sequencing methods[6]. Since SVs alter the expected physical arrangement of the genome, they produce distinct changes in Hi-C interaction patterns. These changes serve as signatures for identifying the location and type of structural variation[7].

Hi-C is also a primary tool for mapping chromatin loops, which are 3D structures formed when distant genomic regions interact. These loops often bring gene promoters into contact with distal regulatory elements like enhancers or silencers, playing a key role in regulating gene expression[8]. Architectural proteins such as CTCF and cohesin frequently bind to specific DNA motifs to stabilize these interactions[9]. Hi-C therefore provides genome-wide maps of the physical interactions that are important for understanding long-range gene regulation.

We have generated Hi-C datasets from the frontal cortex tissue of ten distinct rat strains. This collection includes significant genetic diversity from the HXB/BXH recombinant inbred strains[10], which are a component of the Hybrid Rat Diversity Panel (HRDP)[11,12]. The HRDP is a resource for systems genetics that facilitates high-resolution mapping of complex traits and allows for integrative analyses of genomic, transcriptomic, and phenotypic data[11,13]. The HRDP contains both recombinant and genetically diverse inbred strains, which support reproducible studies of behavior, physiology, and disease. We expect these Hi-C datasets will aid in rat

[1]Department of Genetics, Genomics and Informatics. University of Tennessee Health Science Center, Memphis, TN, 38103, USA. [2]Department of Pharmacology, Addiction Science, and Toxicology. University of Tennessee Health Science Center, Memphis, TN, 38103, USA. ✉e-mail: hchen@uthsc.edu

genome assembly, the discovery of structural variation, and the analysis of regulatory architecture. They can also be integrated with other 'omic' datasets available for this model organism[14–16].

Methods

Animals. Rats were bred at the University of Tennessee Health Science Center from breeders supplied by the Medical College of Wisconsin. All rats were group-housed under standard laboratory conditions and had not been subjected to behavioral or drug treatments. Euthanasia was performed via isoflurane overdose, followed by rapid brain removal. The average age at euthanasia was 139 days. Brains were immediately frozen using freeze spray. All procedures were approved by the Institutional Animal Care and Use Committee of the University of Tennessee Health Science Center and followed the NIH Guidelines concerning the Care and Use of Laboratory Animals.

Tissue processing, Hi-C library preparation, and sequencing. Frozen rat brains were transferred from −80 °C storage to −20 °C for at least 30 minutes, then cryosectioned at −15 °C into 120 μm slices. On a −10 °C cold block, 30–200 mg of cortex tissue was microdissected using an 18-gauge needle with a bent tip and forceps as needed. Following extensive troubleshooting, the optimal tissue input weight was determined to be approximately 50 mg. The bilateral cortex tissue was immediately transferred to pre-chilled 1.5 mL DNA LoBind tubes (Eppendorf) and kept on dry ice until all tissue was collected. The tissue was then pulverized in small quantities using a Cellcrusher-mini pre-cooled in liquid nitrogen. After all tissue was pulverized, 1X PBS was used to collect the sample from the device.

Tissue samples were processed according to the Arima Hi-C + Kit protocol. Samples that passed the first quality control step (QC1) in the Arima protocol continued to the library preparation stage using the KAPA Hyper Prep Kit. Samples were fragmented using a Covaris S2 instrument, with settings for each library detailed in Table 1. The fragmented samples were analyzed with an Agilent Bioanalyzer to confirm a median fragment size between 200 and 600 bp.

Size selection methods varied across samples due to extensive troubleshooting. The optimal procedure was determined to be a two-sided selection using fresh Ampure XP beads, which aimed for a narrow peak with a 400 bp median fragment size. Steps involving Ampure XP beads were performed with either a DynaMag for 1.5 mL tubes or a magnetic rack from Parse Biosciences for PCR-sized tubes. It was noted that magnetic 96-well plates were unsuitable as they resulted in bead loss.

Library preparation was completed using the KAPA HyperPrep kit with either Illumina TruSeq DNA Single Indexes or Illumina TruSeq UDI v2. Pre-amplification libraries were quantified using the KAPA Library Quantification Kit on a Lightcycler 480. Amplified libraries and samples at all other quantification steps were analyzed using the Qubit HS DNA Assay and an Agilent Bioanalyzer HS DNA chip. A complete list of reagents and part numbers is available in Table 2.

Generating prerequisite files for the Juicer pipeline. The Juicer pipeline (v1.6)[17] requires three input files in addition to the FASTQ sequencing data: a reference genome, a chromosome size file, and a restriction site file. **a. Reference genome:** The mRatBN7.2/rn7 genome was used as the reference. The file ("rn7.fa.gz") was downloaded from the UCSC Genome Browser. To improve mapping specificity, only canonical chromosomes were retained, excluding unlocalized and unplaced sequences. **b. Chromosome size file:** The reference genome was indexed using the faidx subcommand of samtools (v1.16.1)[18]. The chromosome names and sizes were extracted from the first two columns of the resulting.fai file. **c. Restriction enzyme site file:** A file containing the genomic locations of restriction enzyme cleavage sites was generated using the generate_site_positions.py script from the Juicer pipeline. The four recognition sequences used in the Arima Genomics protocol (GATC, GANTC, CTNAG, and TTAA) were supplied as parameters along with the reference genome.

Processing Hi-C data using the Juicer pipeline. The Hi-C sequencing data and the three prerequisite files were processed using the Juicer pipeline (https://github.com/aidenlab/juicer/releases/tag/1.6) with a juicer_tools.jar file (https://github.com/aidenlab/juicer/wiki/Download). The pipeline automates read alignment, filtering of invalid reads, and generation of contact maps. **A. Alignment**. Paired-end FASTQ files were aligned to the reference genome using BWA-MEM (v.0.7.17)[19] with Juicer's default parameters (-SP5M). No pre-alignment trimming was performed. **B. Filtering**. After alignment, reads were classified as normal paired, chimeric paired, and chimeric ambiguous (mapped equally well to multiple locations), or unmapped. Chimeric ambiguous and unmapped reads are excluded from downstream analyses. **C. Duplicate removal**. PCR and optical duplicates were identified and removed using Juicer's dups.awk module. This step produced the final non-redundant dataset (merged_nodups.txt). **D. Generating contact matrices**. The alignments were then used to generate contact matrices, stored in .hic files, at two mapping quality (MapQ) thresholds. The inter.hic file includes reads with MapQ ≥ 1, while the high-stringency inter_30.hic file includes only reads with MapQ ≥ 30. The latter file, containing uniquely mapped reads, was used for downstream analyses.

Data Records

The dataset is publicly available through the NIH Short Read Archive[20–29]. Accession numbers are provided in Table 3. The Hi-C sequencing data are organized into three groups: (1) four individual rats from the parental strains of the hybrid rat diversity panel's recombinant inbred (RI) lines, (2) five RI strains, and (3) one F1 hybrid (SHR/Olalpcv × BN/NHsdMcwi). The deposited data consist of paired-end raw sequencing reads in fastq.gz format.

Library	Strain	Duty Cycle	Intensity	Cycles/Burst	Time (s)	# cycles	Med Fragment
592.A	SHR/OlaIpcv	10%	5	200	60	NA	200
592.B	SHR/OlaIpcv	10%	5	200	60	NA	300
607.B	HXB10/Ipcv	10%	5	200	60	NA	400
607.C	HXB10/Ipcv	10%	5	200	60	NA	400
A2D.B	LE/Stm	10%	5	200	60	8	350
74A.A	F344/Stm	10%	5	200	60	8	350
DE8B.A	BN-Lx/Cub	10%	5	200	60	12	700
D765.A	BXH6/Cub	10%	5	200	60	12	700
DBA9.A	HXB23/Ipcv	10%	5	200	60	15	600
DA21.A	SHR/OlaIpcvXBN/NHsdMcwi	10%	5	200	60	15	600
DA08.A	HXB2/Ipcv	10%	5	200	60	12	435
DA68.A	HXB31/Ipcv	10%	5	200	60	12	354

Table 1. Settings for Covaris instrument used for each library.

Kit/Reagent	Supplier	Part No.	Notes
Arima Hi-C + Kit	Arima	A410321	Component Box A
		A410322	Component Box B
		A410239	Component Box C
Kapa Hyper Prep Kit	Roche	KK8504	
Ampure XP Beads	Beckman Coulter	A63880	
Qubit HS DNA Assay	Invitrogen (Thermo)	Q32854	
Covaris microTUBE	Covaris	520045	Full name: Covaris microTUBE AFA Fiber Pre-Slit Snap-Cap 6 x 16 mm
Illumina TruSeq Adapters	Illumina	20015960	Full name: Illumina TruSeq DNA Single Indexes Set A; NOW DEFUNCT
Illumina TruSeq UDI v2	Illumina	20040870	Replacement for above; used in only most recent Hi-C library (E1AF)
Kapa Library Quant Kit	Roche	KK4824	Full name: KAPA Library Quantification Kit (Illumina) Lightcycler 480qPCR Mix

Table 2. Kits and reagents for library preparation.

Technical Validation

Hi-C sequencing data were processed with the Juicer pipeline, using the RatBN7.2/rn7 reference genome for alignment. Quality control metrics for each sample are detailed in Table 3. The experimental goal was to generate 500–600 million (M) paired-end reads per sample. The average yield was 623.39 M reads per sample. Two samples, LE/Stm and F344/Stm, had lower sequencing depth due to challenges in quantifying the library concentration, which led to imbalanced output in a pooled sequencing run. However, these two samples produced a high proportion of unique reads (76.51%) compared to the sample average of 60.85%.

Alignment to the reference genome found only 1.31% of the reads failed to map across all samples. Of the mapped reads, an average of 41.30% were normal (non-chimeric) pairs derived from single genomic fragments. The remaining 57.40% were chimeric, containing sequences from different genomic locations. Among these chimeric reads, 47.34% mapped to a unique genomic location, while 10.06% mapped ambiguously to multiple locations.

The alignable read pairs, which include both normal pairs and uniquely mapped chimeric pairs, constituted an average of 88.64% of the total reads. From this set, optical duplicates accounted for a small fraction (0.83% of total reads), while PCR duplicates were more common (23.83% of total reads). After removing duplicates, the final set of unique, alignable read pairs used for analysis represented an average of 63.98% of the initial total reads.

For high-confidence contact analysis, these unique read pairs were filtered based on mapping quality (MAPQ). Pairs with a MAPQ score below 30 were excluded, which removed an average of 22.32% of the unique pairs. The remaining 77.68% of unique pairs, averaging 297.38 M high-quality pairs per sample, were used to construct the final Hi-C contact maps.

Analysis of the high-quality pairs showed that inter-chromosomal contacts (interactions between different chromosomes) accounted for an average of 52.28 M pairs per sample (17.6%). The majority, 245.10 M pairs (82.4%), were intra-chromosomal contacts (interactions within the same chromosome). These intra-chromosomal interactions were further divided by the linear genomic distance between contacting sites.

Strain	Accession	File Size * (GB)	Sequenced Read Pairs	Normal Paired	Chimeric Paired	Chimeric Ambiguous	Unmapped	Alignable (Normal + Chimeric Paired)	Unique Reads	PCR Duplicates	Optical Duplicates	Below MAPQ Threshold	Hi-C Contacts	Inter-chromosomal	Intra-chromosomal	Short Range (<20Kb)	Long Range (>20Kb)
SHR/ Olalpcv	SRR34514260[26]	S1: 38.9/39.5 S2: 31.1/32.1	952,935,229	449,085,823 (47.13%)	402,527,810 (42.24%)	81,938,110 (8.60%)	19,383,486 (2.03%)	851,613,633 (89.37%)	421,930,987 (44.28%)	427,615,729 (44.87%)	2,066,917 (0.22%)	106,778,274 (25.31%)	315,152,713 (74.69%)	42,962,539 (10.18%)	272,190,174 (64.51%)	162,842,590 (38.59%)	109,277,669 (25.90%)
BN-Lx/Cub	SRR31858086[23]	R1/R2: 50.5/52.2	665,286,666	264,253,780 (39.72%)	330,656,732 (49.70%)	65,703,773 (9.88%)	4,672,381 (0.70%)	594,910,512 (89.42%)	391,501,702 (58.85%)	187,745,605 (28.22%)	15,663,205 (2.35%)	88,670,817 (22.65%)	302,830,885 (77.35%)	52,727,255 (13.47%)	250,103,630 (63.88%)	112,478,669 (28.73%)	137,617,246 (35.15%)
BXH6/Cub	SRR31858085[27]	R1/R2: 50.1/51.6	662,628,241	239,347,784 (36.12%)	352,448,200 (53.19%)	65,260,074 (9.85%)	5,572,183 (0.84%)	591,795,984 (89.31%)	366,754,424 (55.35%)	208,883,824 (31.52%)	16,157,736 (2.44%)	81,913,007 (22.33%)	284,841,417 (77.67%)	45,908,394 (12.52%)	238,933,023 (65.15%)	108,911,801 (29.70%)	130,011,848 (35.45%)
HXB2/Ipcv	SRR31858082[21]	R1/R2: 39.2/40.7	515,278,266	220,245,390 (42.74%)	241,424,605 (46.85%)	48,477,976 (9.41%)	5,130,295 (1.00%)	461,669,995 (89.60%)	281,137,924 (54.56%)	178,235,975 (34.59%)	2,296,096 (0.45%)	66,166,147 (23.54%)	214,971,777 (76.46%)	36,708,051 (13.06%)	178,263,726 (63.41%)	88,987,202 (31.65%)	89,242,367 (31.74%)
HXB10/Ipcv	SRR34514259[29]	S3: 25.0/26.2 S4: 24.6/25.3	671,477,361	354,240,739 (52.76%)	253,965,944 (37.82%)	50,055,079 (7.45%)	13,215,599 (1.97%)	608,206,683 (90.58%)	527,787,344 (78.60%)	77,514,001 (11.54%)	2,905,338 (0.43%)	107,234,103 (20.32%)	420,553,241 (79.68%)	62,958,554 (11.93%)	357,594,687 (67.75%)	212,515,598 (40.27%)	144,998,935 (27.47%)
HXB23/Ipcv	SRR31858084[28]	R1/R2: 55.6/56.5	729,389,632	310,749,674 (42.60%)	345,687,339 (47.39%)	67,223,826 (9.22%)	5,728,793 (0.79%)	656,437,013 (90.00%)	498,113,688 (68.29%)	154,531,847 (21.19%)	3,791,478 (0.52%)	116,191,104 (23.33%)	381,922,584 (76.67%)	80,489,692 (16.16%)	301,432,892 (60.51%)	126,805,703 (25.46%)	174,595,310 (35.05%)
HXB31/Ipcv	SRR31858081[24]	R1/R2: 65.0/67.3	866,086,396	357,846,036 (41.32%)	419,394,522 (48.42%)	78,756,301 (9.09%)	10,089,537 (1.16%)	777,240,558 (89.74%)	491,171,681 (56.71%)	281,278,878 (32.48%)	4,789,999 (0.55%)	109,418,687 (22.28%)	381,752,994 (77.72%)	61,258,667 (12.47%)	320,494,327 (65.25%)	178,745,611 (36.39%)	141,684,502 (28.85%)
LE/Stm	SRR34514258[25]	R1/R2: 18.8/17.8	241,132,264	88,512,979 (36.71%)	115,869,905 (48.05%)	32,117,729 (13.32%)	4,631,651 (1.92%)	204,382,884 (84.76%)	185,099,065 (76.76%)	18,224,944 (7.56%)	1,058,875 (0.44%)	38,966,122 (21.05%)	146,132,943 (78.95%)	30,786,881 (16.63%)	115,346,062 (62.32%)	48,297,170 (26.09%)	67,029,804 (36.21%)
F344/Stm	SRR34514257[22]	R1/R2: 17.5/18.5	236,756,631	83,465,301 (35.25%)	117,339,370 (49.56%)	31,327,006 (13.23%)	4,624,954 (1.95%)	200,804,671 (84.81%)	180,520,409 (76.25%)	19,408,381 (8.20%)	875,881 (0.37%)	38,443,769 (21.30%)	142,076,640 (78.70%)	28,198,993 (15.62%)	113,877,647 (63.08%)	45,033,988 (24.95%)	68,830,975 (38.13%)
SHR/ OlalpcvxBN/ NHsdMcwi (F1)	SRR31858083[20]	R1/R2: 53.8/52.9	692,932,477	267,613,999 (38.62%)	347,357,787 (50.13%)	73,095,088 (10.55%)	4,865,603 (0.70%)	614,971,786 (88.75%)	486,090,705 (70.15%)	125,405,987 (18.10%)	3,475,094 (0.50%)	102,505,023 (21.09%)	383,585,682 (78.91%)	80,822,369 (16.63%)	302,763,313 (62.29%)	111,873,195 (23.01%)	190,858,836 (39.26%)

Table 3. Statistics of Hi-C sequencing after running the Juicer pipeline. *File sizes for Read 1 (R1) and Read 2 (R2) available from SRA are provided in GB. SHR/Olalpcv has two sets of data, one marked as S1 and other marked as S2. Similarly, HXB10/Ipcv also has two sets of data, marked as S3 and S4. All other samples have one set of data files.

Short-range interactions (<20 kb) comprised an average of 119.65 M pairs, while long-range interactions (>20 kb) made up the remaining 125.41 M pairs.

Taken together, these processing steps and the resulting metrics indicate that the Hi-C dataset is of high quality, meeting or exceeding the benchmark recommendations provided by Arima Genomics. Key criteria met include a low percentage of unmapped reads (1.31% observed vs. <6% recommended), a limited number of ambiguous chimeric reads (10.06% observed vs. <20% recommended), a high proportion of alignable reads (with over 98% of reads initially mapped, compared to a recommendation of >80%), and an appropriate fraction of inter-chromosomal contacts (approximately 17.6% observed vs. ~20% recommended). These metrics suggest that the dataset provides a solid foundation for investigating the 3D genomic architecture and functional genomics of the rat.

Data availability

The Hi-C sequencing dataset supporting this Data Descriptor has been deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession number PRJNA1197090. The data are publicly accessible at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1197090/.

Code availability

Data processing was conducted using Juicer 1.6 (https://github.com/aidenlab/juicer/tree/main). The command line used and the input files are available on GitHub at https://github.com/distilledchild/rat-cortex-hic-descriptor/.

References

1. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat Rev Genet* **17**, 772 (2016).
2. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
3. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat Plants* **5**, 833–845 (2019).
4. Duan, H. *et al.* Physical separation of haplotypes in dikaryons allows benchmarking of phasing accuracy in Nanopore and HiFi assemblies with Hi-C data. *Genome Biol.* **23**, 84 (2022).
5. Mahmoud, M. *et al.* Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).
6. Dixon, J. R. *et al.* Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* **50**, 1388–1398 (10/2018).
7. Song, F., Xu, J., Dixon, J. & Yue, F. Analysis of Hi-C Data for Discovery of Structural Variations in Cancer. *Methods Mol Biol* **2301**, 143–161 (2022).
8. Siersbæk, R. *et al.* Dynamic rewiring of promoter-anchored chromatin loops during adipocyte differentiation. *Mol. Cell* **66**, 420–435. e5 (2017).
9. Davidson, I. F. *et al.* CTCF is a DNA-tension-dependent barrier to cohesin-mediated loop extrusion. *Nature* **616**, 822–827 (2023).
10. Printz, M. P., Jirout, M., Jaworski, R., Alemayehu, A. & Kren, V. Invited Review: HXB/BXH rat recombinant inbred strain platform: a newly enhanced tool for cardiovascular, behavioral, and developmental genetics and genomics. *J. Appl. Physiol.* **94**, 2510–2522 (2003).
11. Dwinell, M. R. *et al.* Establishing the hybrid rat diversity program: a resource for dissecting complex traits. *Mamm. Genome* https://doi.org/10.1007/s00335-024-10102-y (2025).
12. Tabakoff, B., Smith, H., Vanderlinden, L. A., Hoffman, P. L. & Saba, L. M. Networking in Biology: The Hybrid Rat Diversity Panel. *Methods Mol. Biol.* **2018**, 213–231 (2019).
13. Tabakoff, B., Hoffman, P. L. & Saba, L. The hybrid rat diversity panel for addiction research. 1.
14. de Jong, T. V. *et al.* A revamped rat reference genome improves the discovery of genetic diversity in laboratory rats. *Cell Genom* **4**, 100527 (2024).
15. Pattee, J. *et al.* Evaluation and characterization of expression quantitative trait analysis methods in the Hybrid Rat Diversity Panel. *Front. Genet.* **13**, 947423 (2022).
16. Duttke, S. H. *et al.* Glucocorticoid Receptor-Regulated Enhancers Play a Central Role in the Gene Regulatory Networks Underlying Drug Addiction. *Front. Neurosci.* **16**, 858427 (2022).
17. Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* **3**, 95–98 (2016-7).
18. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10** (2021).
19. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]* (2013).
20. https://identifiers.org/ncbi/insdc.sra:SRR31858083.
21. https://identifiers.org/ncbi/insdc.sra:SRR31858082.
22. https://identifiers.org/ncbi/insdc.sra:SRR34514257.
23. https://identifiers.org/ncbi/insdc.sra:SRR31858086.
24. https://identifiers.org/ncbi/insdc.sra:SRR31858081.
25. https://identifiers.org/ncbi/insdc.sra:SRR34514258.
26. https://identifiers.org/ncbi/insdc.sra:SRR34514260.
27. https://identifiers.org/ncbi/insdc.sra:SRR31858085.
28. https://identifiers.org/ncbi/insdc.sra:SRR31858084.
29. https://identifiers.org/ncbi/insdc.sra:SRR34514259.

Acknowledgements

Author contributions

P.K. performed data analysis and drafted the manuscript. R.R.W. generated the Hi-C data and contributed to manuscript preparation. H.C. oversaw data generation and analysis. H.C., R.W.W. and B.M.S. designed the experiments and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.C.

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

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