Hi-C sequencing data from cortex laboratory rats

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

1 Department of Genetics, Genomics and Informatics

2 Department of Pharmacology, Addiction Science, and Toxicology

University of Tennessee Health Science Center, Memphis, TN 38103, USA

Correspondence:

Hao Chen

71. S. Manassas Street

Room 209 Translational Science Building

University of Tennessee Health Science Center

Memphis, TN 38103

hchen@uthsc.edu

Abstract

The three-dimensional conformation and packaging of chromosomes modulates the spatial organization of the nucleus, orchestrating DNA replication and repair, maintaining genome stability and integrity, and regulating gene expression. Hi-C methods provide high-resolution data on chromatin-to-chromatin interactions both within and among chromosomes at a genome-wide scale. Hi-C resolves topologically-associated domains (TADs) and chromatin loops that are linked to cell-specific transcriptional control. We present a comprehensive Hi-C dataset generated from the frontal cortex of laboratory rats, encompassing a diverse group 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 x BN/NHsdMcwi). This dataset serves as a valuable resource for studying the mechanisms by which three-dimensional chromatin architecture governs gene expression in the brain. Strain-specific variations in genome organization can illuminate the influence of chromatin structure on gene expression, neuronal functionality, and the predisposition to neurological and behavioral disorders.

Background & Summary

To understand 3D genome architecture within the nucleus, imaging and Chromosome Conformation Capture (3C) led the research for 3D genome structure[1]. Hi-C sequencing, derived from the combination of the 3C assay combined with high-throughput Next-Generation Sequencing (NGS) technology, has become widely utilized due to their ability to map chromatin interactions across the genome at a high resolution[2]. The chromatin contact frequencies obtained through Hi-C revealed the hierarchical 3D structures of the genome, including chromosomal compartments, topologically associated domains (TADs), and chromatin loops. Notably, such structures have been recognized for their significant roles in various biological and cellular processes, including the regulation of gene expression[3], modulation of chromatin accessibility[3], and their influence on disease mechanisms[4]. Also, these data on long-range interactions between chromatin segments are valuable for a range of applications in genomics and biomedical research.

A major application of the HI-C method is in genome assembly, where Hi-C data are used to scaffold contigs into chromosome-scale assemblies, enabling the construction of highly contiguous and accurate genome references. This is possible because Hi-C contact frequencies provide spatial proximity data, indicating which genomic fragments are likely to be physically adjacent. By analyzing the interaction data, researchers can accurately order and orient contigs, resolving ambiguities in assembly derived from linear sequencing data. This approach enhances the completeness and accuracy of genome assemblies. This application is particularly beneficial for resolving complex genomes, including those of polyploid species[5], and for haplotype phasing, which separates homologous chromosomes into distinct assemblies[6]. Additionally, Hi-C data can be exploited to identify structural variants (SVs) providing valuable insights into genomic rearrangements, including deletions, duplications, insertions, inversions, and translocations. These SVs typically involve alterations spanning 50 base pairs or more[7], making them challenging to detect using short-read sequencing technologies due to their limited capacity for their read length and repetitive region[8]. SVs, such as translocations or inversions, can significantly alter the spatial distance between two genomic regions. These SVs change the contact frequency of reads mapped to these regions. Compared to the background interaction frequencies of individuals without these variants, these altered interactions serve as indicators of SVs[9].

Chromatin loops identified by Hi-C, in general, are formed by enhancer-promoter interactions, which are essential for understanding transcriptional regulation[10]. In addition, information on promoter-enhancer interactions and structural variations obtained from Hi-C data have become an integral part of multi-omics integration, enabling researchers to better understand regulatory mechanisms in transcriptomic, epigenomic, and proteomic data[11]. This integrative approach provides a comprehensive view of how genome organization influences cellular processes, uncovering relationships between chromatin architecture and functional genomic elements in both health and disease contexts.

Methods

Tissue processing, Hi-C library preparation, and sequencing. Frozen rat brains were removed from long-term storage at -80℃, stored at -20℃ for ≥ 30 minutes, and cryosectioned at -15 ℃ into 120 µm slices. Using a -10℃ cold block, 30-200 mg of cortex tissue was microdissected using an 18-gauge needle with a bent tip, and a pair of forceps as needed. Optimal tissue input weight was found to be ~50 mg after extensive troubleshooting. Bilateral cortex tissue was immediately transferred to pre-chilled 1.5 mL DNA Lobind tubes (Eppendorf) kept on dry ice until all tissue was collected. Tissue was pulverized in small quantities using a Cellcrusher mini (Cellcrusher) pre-cooled in liquid nitrogen until all tissue had been pulverized and collected. 1X PBS was used to collect samples from the Cellcrusher. Tissue samples were then processed according to the Arima Hi-C+ Kit (Arima). Samples that passed QC1 in the Arima Hi-C protocol continued to the Arima-provided KAPA Hyper Prep Kit (Roche) protocol. Samples were fragmented using a Covaris S2 instrument (Table 1), generally with the following settings: Duty Cycle 10%, Intensity 5, Cycles/Burst 200, Time 55 s, # cycles 8-15, using Covaris microTUBEs (Covaris). Fragmented samples were analyzed by Agilent Bioanalyzer to confirm a median fragment size of 200-600. Size selection methods varied throughout the samples due to extensive troubleshooting; optimal size selection was performed with fresh Ampure XP beads (Beckman Coulter) in a 2-sided manner, with the goal of a narrow peak and 400 bp median fragment. Any steps that involve Ampure XP beads were performed with either a DynaMag (Thermo Fisher) compatible with 1.5 mL tubes, or a magnetic rack with high/low settings (Parse Biosciences) compatible with PCR-sized tubes; magnetic 96-well plates compatible with PCR-sized tubes were found to result in bead loss. Library preparation was performed using the KAPA HyperPrep kit and Illumina TruSeq DNA Single Indexes (Illumina; discontinued) or Illumina TruSeq UDI v2 (Illumina; current version). Pre-amplification libraries were quantified using the KAPA Library Quantification Kit (Roche) and Lightcycler 480. Amplified libraries were analyzed using Qubit HS DNA Assay (Invitrogen) and Agilent Bioanalyzer HS DNA chip. All other quantification steps in the Arima and KAPA protocols were also performed with the Qubit HS DNA Assay and Agilent Bioanalyzer HS DNA chips. Available part numbers for all listed reagents have been provided in **Table1**.

Table 1. Kits and reagents for library preparation.

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 6x16mm
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 Lib (E1AF)
Kapa Library Quant Kit	Roche	KK4824	Full name: KAPA Library Quantification Kit (Illumina) Lightcycler 480qPCR Mix

Hi-C sequencing data processing. Sequencing data were obtained in the FASTQ format. These data were then analyzed using the Juicer pipeline. mRatBN7.2/rn7 was used as the reference genome in this analysis. The analysis workflow is provided in our github repository.

Data Records

The Hi-C sequencing data comprise three groups: (1) four individual rats representing the parental strains of the two recombinant inbred (RI) strains of the hybrid rat diversity panel 2) five RI strains and (3) one F1 hybrid offspring (SHR/Olalpcv x BN/NHsdMcwi). These datasets, containing paired-end raw sequencing reads in fastq.gz format, have been deposited in the Sequence Read Archive (SRA) under accession numbers shown in Table 2.

Technical Validation

We analyzed the Hi-C data using the Juicer pipeline[12]. The mRatBN7.2/rn7 was used as the reference genome. The QC values for each sample were provided in Table 2. We aimed to obtain about 500-600 M paired-end reads per sample. The average reads per sample was 623.39 M reads in our data collection. Two samples, LE/Stm and F344/Stm, fell short of this goal. On average, only 1.31% was not mappable to the reference genome. Among the mapped reads, 41.30% were normal reads that did not contain segments from two different genomic locations (i.e., chimeric reads), 47.34% were chimeric and were mapped to a unique location in the reference genome. The remaining 10.06% were chimeric and matched to multiple locations in the reference genome. For the alignable reads (i.e., normal+chimeric paired), 0.83% were

optical duplicates, and 23.83% were PCR duplications, the remaining 63.98% were unique reads. On average, 22.32% of the unique reads mapped to the reference genome with a low quality (MAPQ score less than 30), the remaining 77.68% (297.38 M reads) had sufficient map quality and were used to generate information on Hi-C contacts. Among these reads, on average, 52.28 M were mapped to between chromosomal interactions, and the remaining 245.10 M reads per sample were mapped to within chromosomal interactions. These within chromosomal reads were further divided by the distance between the two interacting chromosomal locations. About 119.65 M reads per sample were from short range (<20kB) interactions, and the remaining 125.41 M reads were from long range (>20 Kb) interactions. Together, these analyses indicate that all our samples pass the quality recommendations put forth by Arima Genomics. These metrics include less than 6% unmapped reads, less than 20% chimeric ambiguous reads, greater than 80% alignable reads and approximately 20% interchromosomal reads. We anticipate this high quality dataset will contribute to the understanding of 3D genomic structure and functional genomics of the rats.

Table 2. Statistics of Hi-C sequencing after running the Juicer pipeline.

Strain	Accession	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	SRR 3185 8090	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	SRR 3185 8086	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	SRR 3185 8085	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/lpcv	SRR 3185 8082	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/lpcv	SRR 3185 8089	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/lpcv	SRR 3185 8084	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/lpcv	SRR 3185 8081	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	SRR 3185 8088	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	SRR 3185 8087	236,75 6,631	83,46 5,301 (35.2 5%)	117,3 39,37 0 (49.5 6%)	31,32 7,006 (13.2 3%)	4,62 4,95 4 (1.95 %)	200,804,67 1 (84.81%)	180, 520, 409 (76.2 5%)	19,40 8,381 (8.20 %)	875, 881 (0.37 %)	38,443,7 69 (21.30%)	142,076, 640 (78.70%)	28,198,9 93 (15.62%)	113,877,647 (63.08%)	45,033,988 (24.95%)	68,830,975 (38.13%)
SHR/OlalpcvxBN/NHsdMcwi (F1)	SRR 3185 8083	692,93 2,477	267,6 13,99 9 (38.6 2%)	347,3 57,78 7 (50.1 3%)	73,09 5,088 (10.5 5%)	4,86 5,60 3 (0.70 %)	614,971,78 6 (88.75%)	486, 090, 705 (70.1 5%)	125,4 05,98 7 (18.1 0%)	3,47 5,09 4 (0.50 %)	102,505, 023 (21.09%)	383,585, 682 (78.91%)	**80,822,3 69 (16.63%)**	302,763,313 (62.29%)	111,873,195 (23.01%)	190,858,836 (39.26%)

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/blob/main/README.md

Author Contributions

PK performed data analysis and drafted the manuscript. RRW generated the Hi-C data and contributed to manuscript preparation. HC oversaw data generation and analysis. HC, RWW, and BMS designed the experiments and revised the manuscript.

Competing Interests

None

Acknowledgments

The authors thank Dr. Melinda R. Dwinell (Medical College of Wisconsin) for providing the breeders for the Hybrid Rat Diversity Panel (HRDP) and acknowledge the Center for Integrative and Translational Genomics at the University of Tennessee Health Science Center for its support in maintaining the HRDP. The authors also acknowledge Mr. Angel Garcia Martinez and Ms. Caroline Jones for their assistance with HRDP breeding. The Molecular Resource Center and the Genomics Core at the University of Tennessee Health Science Center are acknowledged for their contributions to sequencing data generation. The majority of the computation for this work was performed on the University of Tennessee Infrastructure for Scientific Applications and Advanced Computing (ISAAC) computational resources. This work was supported by NIH/NIDA grants U01 DA-053672 (to BMS, RWW, and HC).

References

1. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.* **21**, 207–226 (2020).
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. Liu, R. *et al.* Hi-C, a chromatin 3D structure technique advancing the functional genomics of immune cells. *Front. Genet.* **15**, 1377238 (2024).
4. Krumm, A. & Duan, Z. Understanding the 3D genome: Emerging impacts on human disease. *Semin Cell Dev Biol* **90**, 62–77 (2019).
5. 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).
6. 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).
7. Mahmoud, M. *et al.* IMPORTANT: Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).
8. Dixon, J. R. *et al.* Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* **50**, 1388–1398 (10/2018).
9. 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).
10. Siersbæk, R. *et al.* Dynamic rewiring of promoter-anchored chromatin loops during adipocyte differentiation. *Mol. Cell* **66**, 420–435.e5 (2017).
11. Gong, H., Yang, Y., Zhang, S., Li, M. & Zhang, X. Application of Hi-C and other omics data analysis in human cancer and cell differentiation research. *Comput. Struct. Biotechnol. J.*

19, 2070–2083 (2021).

12. 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).