Supplemental Material

Analysis of Hi-C data using SIP effectively identifies loops in organisms from *C. elegans* to mammals

M. Jordan Rowley, Axel Poulet, Michael H. Nichols, Brianna J. Bixler, Adrian L. Sanborn, Elizabeth A. Brouhard, Karen Hermetz, Hannah Linsenbaum, Gyorgyi Csankovszki, Erez Lieberman Aiden, and Victor G. Corces

Table of Contents

SUPPLEMENTAL METHODS

Loop calling

Hi-C data obtained in human cells was mapped to GRCh38. This was the version of the human genome used in the original publication where HiCCUPS was first described (Rao et al. 2014) and we used the same version to be able to directly compare results obtained with SIP to those obtained with HiCCUPS. Although a newer version of the human genome is available, the use of GRCh38 should not affect the findings, since we are only performing a comparison between different loop calling tools. HiCCUPS (Rao et al. 2014) and HOMER (Heinz et al. 2018) loops were called at 5 kb resolution with default parameters, except when estimating the effect of varying FDR levels. Loops identified by cLoops were obtained using the parameters recommended for Hi-C data (Cao et al. 2019), however, due to memory costs, we were unable to run cLoops on the full *Drosophila* dataset. In order to perform comparisons with cLoops, we subsampled the *Drosophila* data on 2L to 50 million reads, an approach similar to what was done in the original publication (Cao et al. 2019). We used SIP in this subsampled data to ensure that the loops could still be identified. Fit-Hi-C interactions were identified at 5 kb resolution using default parameters and two passes (Ay et al. 2014). Example loci were visualized using Juicebox (Durand et al. 2016). Because cLoops does not place loops into bins, comparison of loop callers was done by placing loops into 25 kb bins before taking overlaps.

Loops called in *Drosophila* cells used Kc167 Hi-C maps combined from GSE80702 (Cubeñas-Potts et al. 2016) and GSE89112 (Eagen et al. 2017) genome build dm6. A. aegypti Hi-C was from GSE113256 (Matthews et al. 2018) genome build AaegL5.0.

Performance Testing

SIPs performance tests were run on a laptop with 2 cores and 16 GB of RAM running Windows 10. CPU performance, memory usage, and time were tracked using VisualVM (https://visualvm.github.io/index.html). SIP was also tested on several machines including Linux (Table S2).

Noise addition was done by first calculating the number of additional reads necessary to add to the matrix based upon the noise percentage desired. For example, to increase noise by 50%, an additional 50% of the total observed signal was added back in. These additional pseudo-reads were then distributed according to the distance decay and added to bins based on a random Poisson distribution.

Contribution of Transcription Factors

CTCF ChIA-PET data is from GSE72816 (Tang et al. 2015). CTCF peaks and motif orientations were obtained from previously published data (Rao et al. 2014), and ZNF143, YY1, and RNAPII ChIP-seq data were obtained from ENCODE ENCSR936XTK, ENCSR000BNP, and ENCSR000BGD, respectively. All datasets were obtained using GM12878 cells. ChIP-seq signal for CTCF and BORIS in K562 cells were obtained from GSE70764, remapped to hg19, and peaks were identified using MACS2 (Zhang et al. 2008) with bw 100 due to the reported insert size in the libraries (Pugacheva et al. 2015).

DPY-27 ChIP-seq signal was obtained from GSE67650(Kramer et al. 2015). Motifs were identified using MEME ChIP of loop anchors.

Hi-C and HiChIP in C. elegans

Late stage hermaphrodite *C. elegans* embryos were collected from the N2 Bristol strain, crosslinked in 1.1% formaldehyde, and then placed into isolation buffer (10 mM Tris-HCl pH 8.0,

10 mM NaCl, 0.2% Igepal CA-630, 5 mM DTT, 1 mM PMSF, 0.1% PU, and 0.5 mM EGTA). Embryos were ground on ice with pestle “A” from DWK Life Sciences to isolate nuclei. Nuclei were then crosslinked in 1% formaldehyde for 10 min, resuspended in 0.5% SDS buffer and incubated at 65° C for 5 min. HiChIP libraries were then prepared as described (Rowley et al. 2019) but using antibodies against the DPY-27 protein (Csankovszki et al. 2009). Hi-C with DpnII was performed using the *in-situ* protocol (Rao et al. 2014).

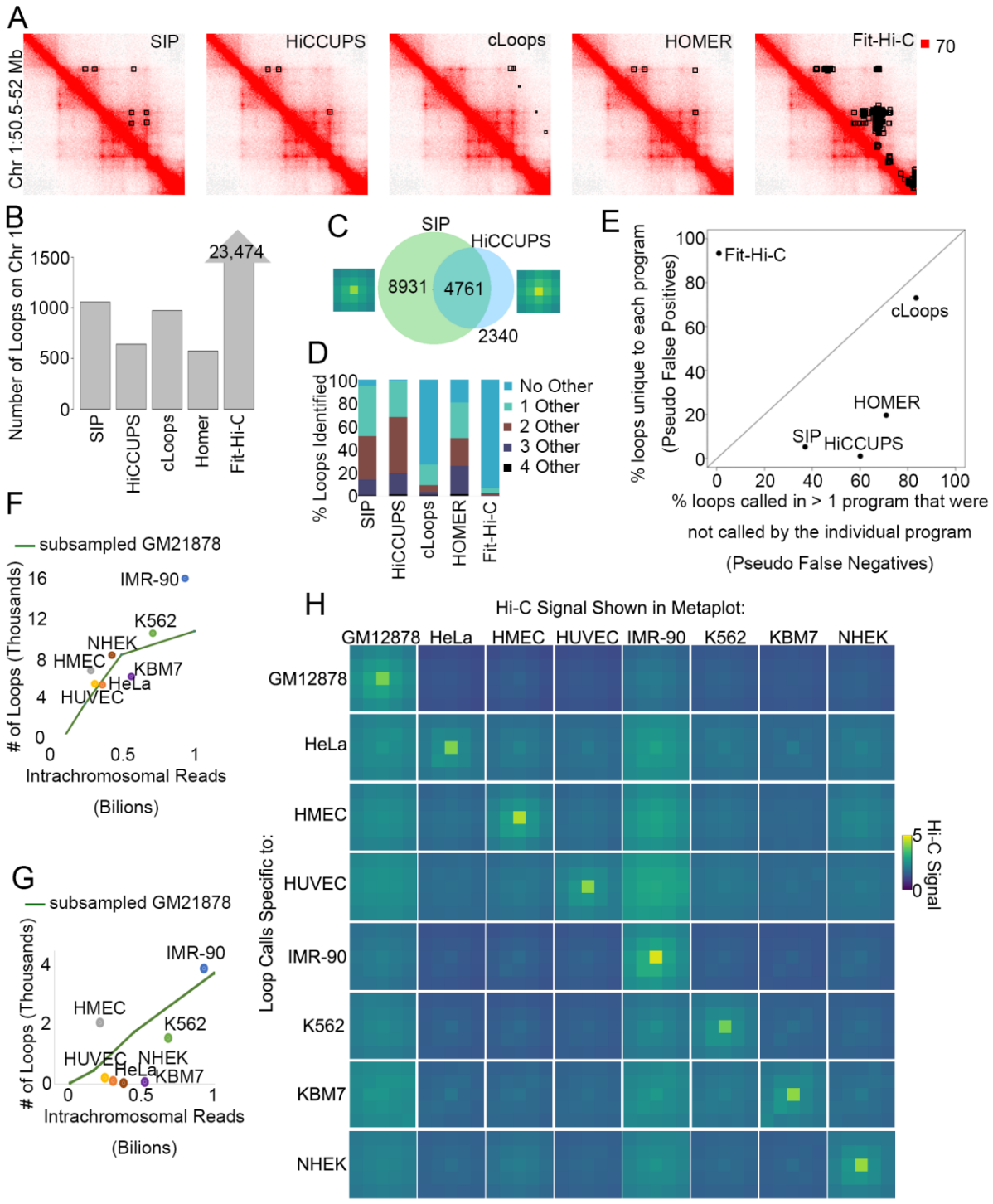

Figure S1. Comparison of loop calls obtained by different programs. A. Example locus showing loop calls in GM12878 Hi-C data by SIP, HiCCUPS, cLoops, and HOMER. Significant interaction calls by Fit-Hi-C are also shown for comparison. B. Number of loops or significant interactions on Chromosome 1 that were identified by each program. C. Venn diagram of SIP vs HiCCUPS loops as well as metaplots of signal on loops unique to each program. D. Percentage of loops identified by each program that were also identified by other loop callers. E. Percentage of loops that were unique to each program (y-axis) vs the percentage of loops that were called

in multiple programs, but not identified by the individual program (x-axis). F. Number of loops identified by SIP in each cell type at 5 kb resolution, compared to the number of loops identified in subsampled GM12878 cells. G. Number of loops identified by HiCCUPS in each cell type at 5 kb resolution, compared to the number of loops identified in subsampled GM12878 cells. H. Average Hi-C signal for loops categorized by the cell type in which they were uniquely called by SIP.

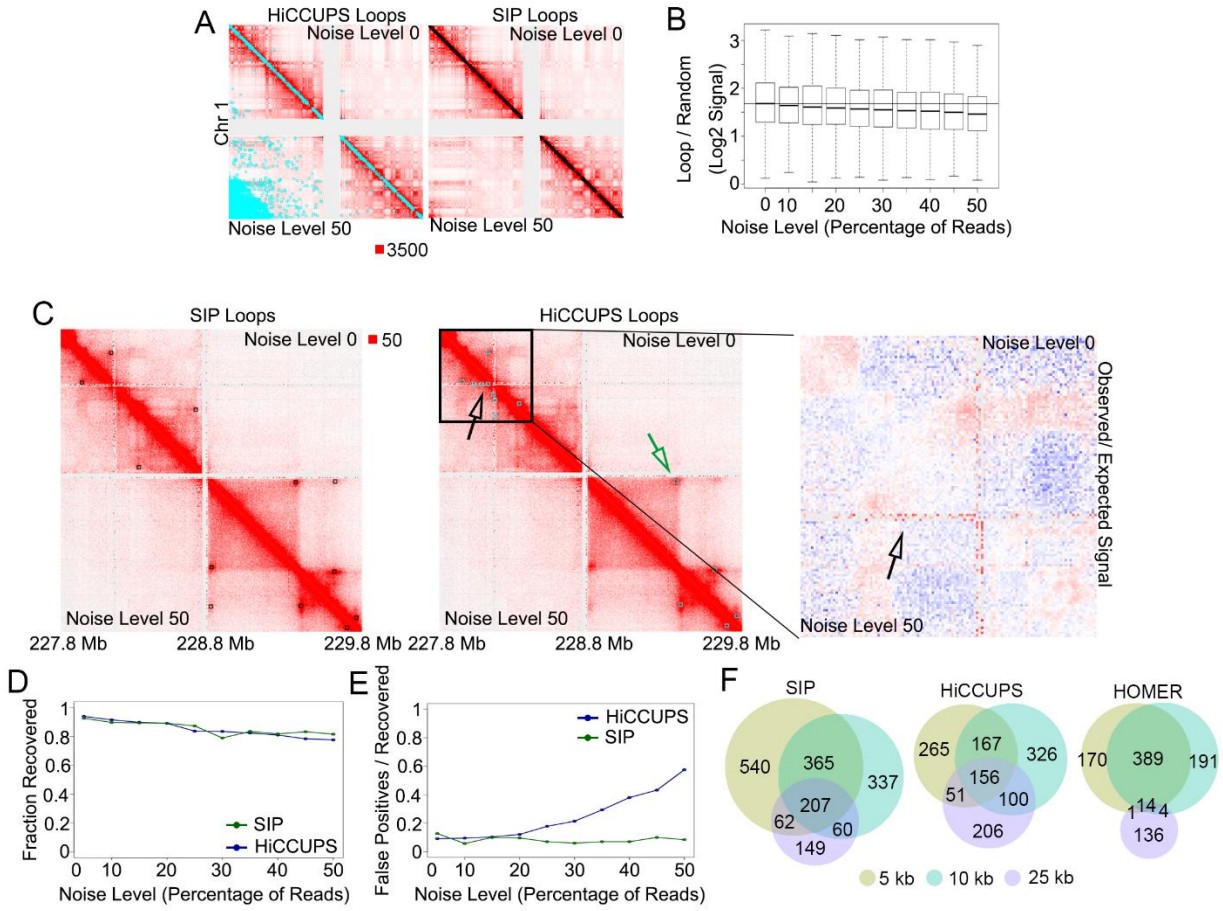

Figure S2. Effect of Hi-C data quality on loop calls by SIP versus HiCCUPS. A. Screenshot of Hi-C data for Chromosome 1 and the loops called in the original data (above diagonal) and in data with added noise. B. Boxplot of loop vs random signal after noise addition. Loop locations from the original dataset were used. C. Examples of loops missed (green arrow) or false positives gained (black arrow) after increasing noise. Right: Distance normalized signal highlighting the increased speckling after noise addition (black arrow). D. Fraction of loops called using the full dataset recovered by SIP (green) or by HiCCUPS (blue) in data with added noise. E. Ratio of false positives (loops not identified in the original dataset) vs loops recovered by SIP (green) or HiCCUPS (blue) in data with added noise. F. Overlap of loops called at 5 kb, 10 kb, or 25 kb resolution.

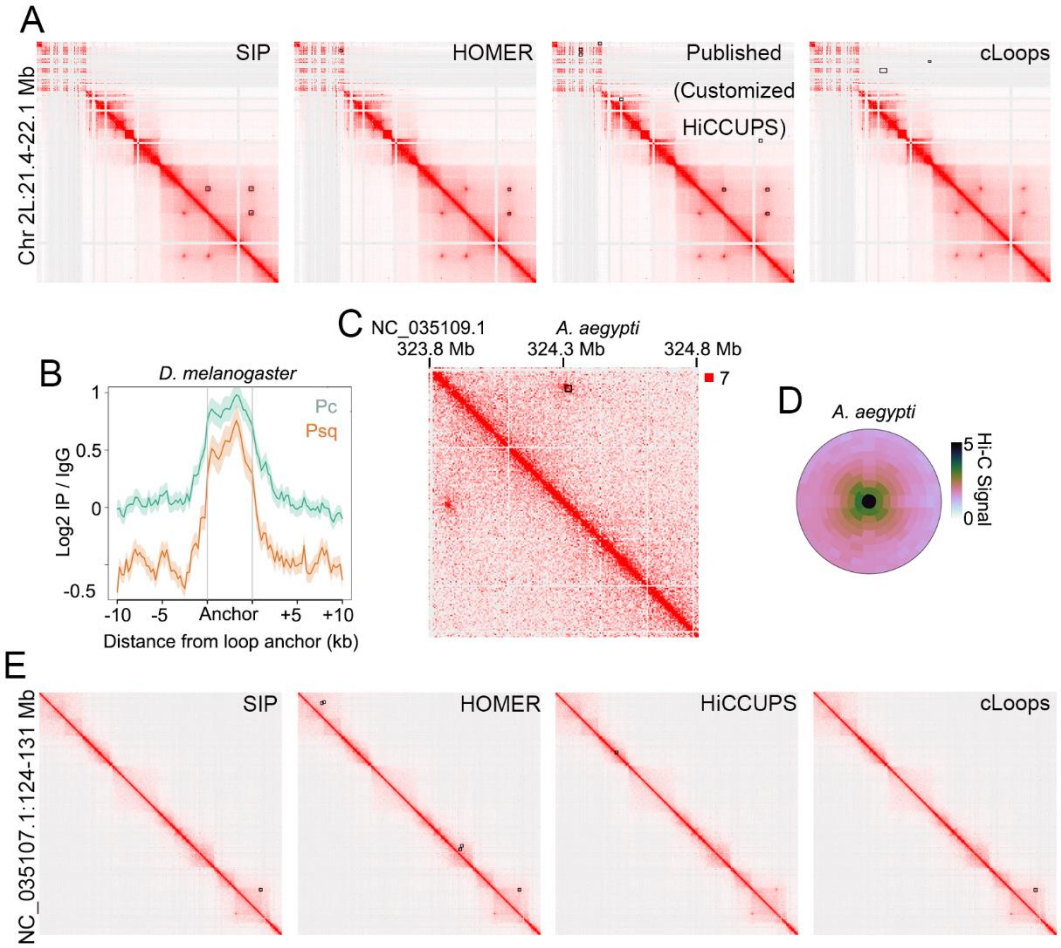

Figure S3. Performance of different loop calling tools on Hi-C data from *Drosophila* and *Aedes*. A. Examples of *D. melanogaster* loops called by SIP, Homer, published loops that use a method similar to HiCCUPS, and cLoops. B. Average profile of Pc (teal) and Psq (orange) ChIP-seq signal across *D. melanogaster* loop anchors. C. Zoomed in example of a SIP loop in *A. aegypti*. D. SIPMeta plot for loops detected in *A. aegypti*. E. Examples of *A. aegypti* loops called by SIP, Homer, HiCCUPS, and cLoops.

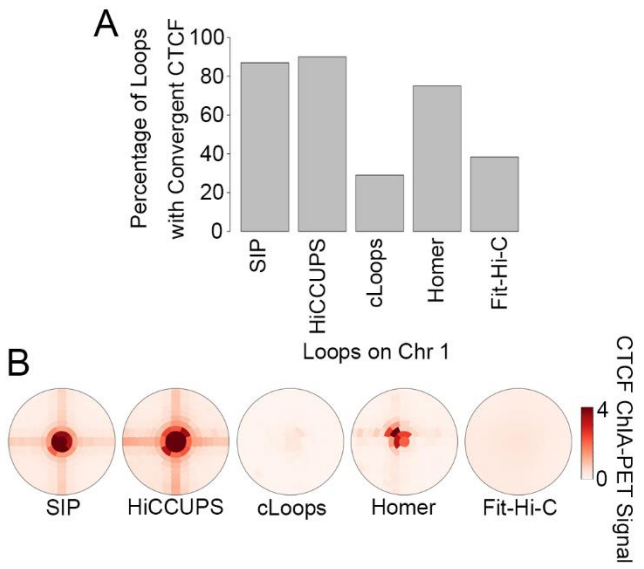

Figure S4. Performance of different programs in calling loops containing CTCF anchors. A. Percentage of loops identified by SIP, HiCCUPS, cLoops, Homer on Chromosome 1 that overlap with convergent CTCF. Also shown are the percentage of Fit-Hi-C interactions overlapping convergent CTCF. B. SIPMeta plots of CTCF ChIA-PET signal for loops from each program.

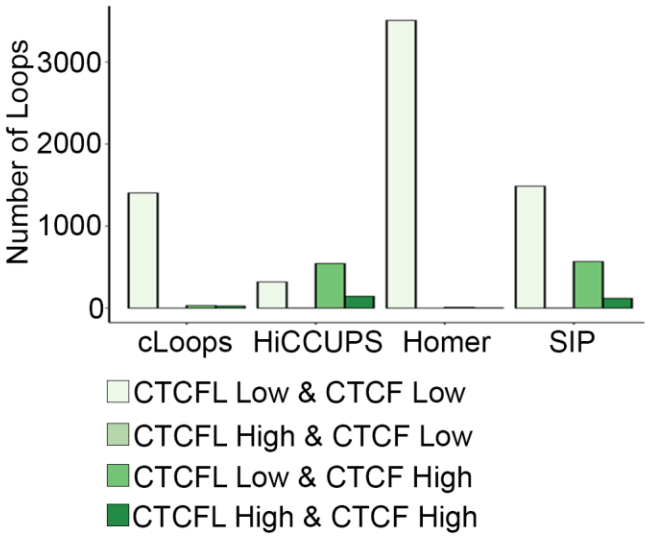

Figure S5. Performance of different programs in calling loops containing different architectural proteins. Number of loops identified by cLoops, HiCCUPS, Homer, and SIP in K562 cells for each category.

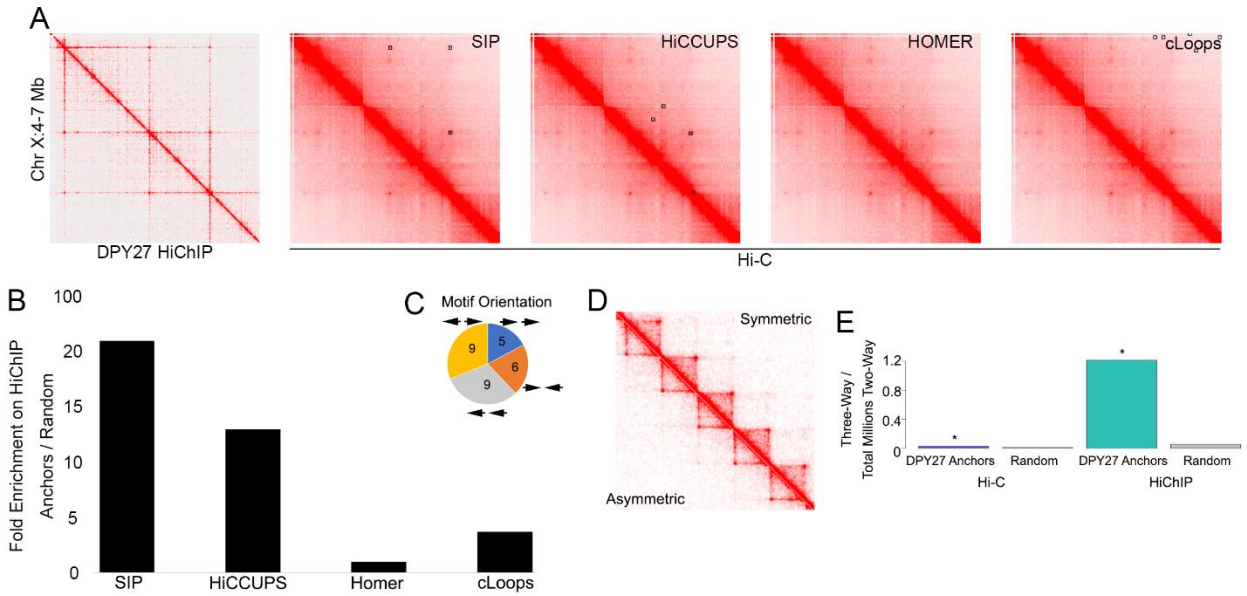

Figure S6. Performance of different programs in identifying loops in *C. elegans* HiChIP data. A. Example locus of loops identified by SIP, HiCCUPS, Homer, and cLoops. DPY27 HiChIP (left) is shown to provide reference to condensin I DCC looping. B. Enrichment of loops identified by each program for overlap of anchors with that of DPY27 HiChIP anchors compared to random regions. C. Number of MEX motifs at loops in each orientation that could be unambiguously identified. D. Polymer simulations of loops formed by symmetric extrusion starting from random sites (top right) or loops formed by asymmetric extrusion starting near loop anchors (bottom left). E. Relative three-way vs two-way interactions obtained by Hi-C or by DPY-27 HiChIP connecting DPY-27 loop anchors compared to the averages of permutations using an equal number of random regions on Chromosome X. * indicates p<.05 Monte-Carlo Permutation test.

Table S1. Runtime and memory usage of loop callers on Chromosome 1 of GM12878 Hi-C.

Computer specifications	
System	Ubuntu 18.04.3
CPU	Intel Xeon Gold 6140 CPU @ 2.30GHz
Available RAM	512G
Used threads	1

	Chromosome 1				
	SIP	HiCCUPS	Fit-Hi-C	Homer	cLoops
Time (Minutes)	2.3	4.67	75.87	32	2432
Memory (GB)	1.157496	1.140964	8.602388	62.05729	102.7938

Table S2. Runtime and memory usage of SIP on different systems

Computer Specifications						
System	Ubuntu 18.04.3	Windows 10	Linux Mint 19.2	Ubuntu 16.04.6	Windows 10	Windows 10
CPU	Intel Xeon Gold 6140 CPU @ 2.30GHz	Intel i7-7660U CPU @ 2.5GHz	Intel i7-5600U CPU @ 2.60GHz	AMD Fx-8320 @ 1.4GHz	Intel i9-9900X CPU @ 3.5GHz	intel i3-8130U @ 2.2GHz
Available RAM	512G	16G	16G	32G	64G	32G
Used threads	23	2	2	2	2	2
Time (minutes)	12	46	43	46	23	42
Memory (GB)	15.4	3.5	2.9	3.95	5.65	3.05

Table S3. Quality control of Hi-C libraries

	Hi-C Rep1	Hi-C Rep2
Sequenced Read Pairs	168,530,362	57,237,198
Normal Paired	138,491,810 (82.18%)	52,225,973 (91.24%)
Chimeric Paired	541 (0.00%)	101,731 (0.18%)
Chimeric Ambiguous	306 (0.00%)	17,799 (0.03%)
Unmapped	30,037,705 (17.82%)	4,891,695 (8.55%)
Ligation Motif Present	63,092,084 (37.44%)	10,404,176 (18.18%)
Alignable (Normal+Chimeric Paired)	138,492,351 (82.18%)	52,327,704 (91.42%)
Unique Reads	53,847,549 (31.95%)	43,726,559 (76.40%)
PCR Duplicates	84,642,981 (50.22%)	8,597,729 (15.02%)
Optical Duplicates	1,821 (0.00%)	3,416 (0.01%)
Library Complexity Estimate	59,723,465	141,252,108
Intra-fragment Reads	11,098,945 (6.59% / 20.61%)	14,960,699 (26.14% / 34.21%)
Below MAPQ Threshold	5,915,580 (3.51% / 10.99%)	5,282,707 (9.23% / 12.08%)
Hi-C Contacts	36,833,024 (21.86% / 68.40%)	23,483,153 (41.03% / 53.70%)
Ligation Motif Present	14,711,513 (8.73% / 27.32%)	5,434,745 (9.50% / 12.43%)
3' Bias (Long Range)	86% - 14%	79% - 21%
Pair Type %(L-I-O-R)	25% - 25% - 25% - 25%	25% - 25% - 25% - 25%
Inter-chromosomal	6,919,343 (4.11% / 12.85%)	11,967,710 (20.91% / 27.37%)
Intra-chromosomal	29,913,681 (17.75% / 55.55%)	11,515,443 (20.12% / 26.34%)
Short Range (<20Kb)	18,894,824 (11.21% / 35.09%)	6,066,369 (10.60% / 13.87%)
Long Range (>20Kb)	11,018,826 (6.54% / 20.46%)	5,448,944 (9.52% / 12.46%)

Table S4. Quality control of HiChIP libraries

	DPY27 HiChIP Rep1	DPY27 HiChIP Rep2	DPY27 HiChIP LongReads
Sequenced Read Pairs	40,649,544	51,962,635	1,067,363,468
Normal Paired	37,051,768 (91.15%)	45,895,018 (88.32%)	439,323,497 (41.16%)
Chimeric Paired	31,418 (0.08%)	108,196 (0.21%)	28,453,367 (2.67%)
Chimeric Ambiguous	2,890 (0.01%)	18,757 (0.04%)	182,607,181 (17.11%)
Unmapped	3,563,468 (8.77%)	5,940,664 (11.43%)	64,021,180 (6.00%)
Ligation Motif Present	4,022,726 (9.90%)	13,504,076 (25.99%)	185,342,026 (17.36%)
Alignable (Normal+Chimeric Paired)	37,083,186 (91.23%)	46,003,214 (88.53%)	467,776,864 (43.83%)
Unique Reads	5,857,090 (14.41%)	4,217,873 (8.12%)	12,750,497 (1.19%)
PCR Duplicates	31,225,432 (76.82%)	41,783,994 (80.41%)	454,930,912 (42.62%)
Optical Duplicates	664 (0.00%)	1,347 (0.00%)	95,455 (0.01%)
Library Complexity Estimate	5,867,653	4,217,950	12,750,497
Intra-fragment Reads	2,210,727 (5.44% / 37.74%)	1,489,962 (2.87% / 35.32%)	2,849,042 (0.27% / 22.34%)
Below MAPQ Threshold	723,722 (1.78% / 12.36%)	516,835 (0.99% / 12.25%)	4,504,473 (0.42% / 35.33%)
Hi-C Contacts	2,922,641 (7.19% / 49.90%)	2,211,076 (4.26% / 52.42%)	5,396,982 (0.51% / 42.33%)
Ligation Motif Present	314,414 (0.77% / 5.37%)	591,723 (1.14% / 14.03%)	2,482,298 (0.23% / 19.47%)
3' Bias (Long Range)	70% - 30%	79% - 21%	73% - 27%
Pair Type %(L-I-O-R)	25% - 25% - 25% - 25%	25% - 25% - 25% - 25%	25% - 25% - 25% - 25%
Inter-chromosomal	1,308,964 (3.22% / 22.35%)	856,584 (1.65% / 20.31%)	2,495,659 (0.23% / 19.57%)
Intra-chromosomal	1,613,677 (3.97% / 27.55%)	1,354,492 (2.61% / 32.11%)	2,901,323 (0.27% / 22.75%)
Short Range (<20Kb)	1,128,800 (2.78% / 19.27%)	766,413 (1.47% / 18.17%)	1,981,903 (0.19% / 15.54%)
Long Range (>20Kb)	484,861 (1.19% / 8.28%)	588,072 (1.13% / 13.94%)	919,388 (0.09% / 7.21%)

Supplemental References

Ay F, Bailey TL, Noble WS. 2014. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res* **24**: 999-1011.

Cao Y, Chen Z, Chen X, Ai D, Chen G, McDermott J, Huang Y, Xiaoxiao G, Han JJ. 2019. Accurate loop calling for 3D genomic data with cLoops. *Bioinformatics* doi:10.1093/bioinformatics/btz651.

Csankovszki G, Collette K, Spahl K, Carey J, Snyder M, Petty E, Patel U, Tabuchi T, Liu H, McLeod I et al. 2009. Three distinct condensin complexes control C. elegans chromosome dynamics. *Curr Biol* **19**: 9-19.

Cubeñas-Potts C, Rowley MJ, Lyu X, Li G, Lei EP, Corces VG. 2016. Different enhancer classes in Drosophila bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture. *Nucleic Acids Research* **45**: 1714-1730.

Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. 2016. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems* **3**: 99-101.

Eagen KP, Aiden EL, Kornberg RD. 2017. Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. *Proceedings of the National Academy of Sciences* **114**: 8764-8769.

Heinz S, Texari L, Hayes MGB, Urbanowski M, Chang MW, Givarkes N, Rialdi A, White KM, Albrecht RA, Pache L et al. 2018. Transcription Elongation Can Affect Genome 3D Structure. *Cell* **174**: 1522-1536 e1522.

Kramer M, Kranz AL, Su A, Winterkorn LH, Albritton SE, Ercan S. 2015. Developmental Dynamics of X-Chromosome Dosage Compensation by the DCC and H4K20me1 in C. elegans. *PLoS Genet* **11**: e1005698.

Matthews BJ, Dudchenko O, Kingan SB, Koren S, Antoshechkin I, Crawford JE, Glassford WJ, Herre M, Redmond SN, Rose NH et al. 2018. Improved reference genome of Aedes aegypti informs arbovirus vector control. *Nature* **563**: 501-507.

Pugacheva EM, Rivero-Hinojosa S, Espinoza CA, Méndez-Catalá CF, Kang S, Suzuki T, Kosaka-Suzuki N, Robinson S, Nagarajan V, Ye Z et al. 2015. Comparative analyses of CTCF and BORIS occupancies uncover two distinct classes of CTCF binding genomic regions. *Genome Biology* **16**.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**: 1665-1680.

Rowley MJ, Lyu X, Rana V, Ando-Kuri M, Karns R, Bosco G, Corces VG. 2019. Condensin II Counteracts Cohesin and RNA Polymerase II in the Establishment of 3D Chromatin Organization. *Cell Rep* **26**: 2890-2903 e2893.

Tang Z, Luo Oscar J, Li X, Zheng M, Zhu Jacqueline J, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J, Ruszczycki B et al. 2015. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163**: 1611-1627.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**: R137.