

Hi-C guided assemblies reveal conserved regulatory topologies on X and autosomes despite extensive genome shuffling

Gina Renschler^{1,2}, Gautier Richard^{1,3}, Claudia Isabelle Keller Valsecchi¹, Sarah Toscano¹, Laura Arrigoni¹, Fidel Ramirez^{1*}, Asifa Akhtar^{1*}

¹ Max Planck Institute of Immunobiology and Epigenetics, Stübeweg 51, 79108 Freiburg im Breisgau, Germany

² Faculty of Biology, University of Freiburg, Schänzlestr. 1, 79104 Freiburg, Germany

³ IGEPP, INRA, Agrocampus Ouest, Univ Rennes, 35600 Le Rheu, France

*Correspondence: akhtar@ie-freiburg.mpg.de or ramirez@ie-freiburg.mpg.de

Supplemental Information Table of Contents

Supplemental Fig. S1, related to Figure 1. Generated PacBio and Hi-C datasets.

Supplemental Fig. S2, related to Figure 1. Hi-C assembly quality control.

Supplemental Fig. S3, related to Figure 1. HiCAssembler.

Supplemental Fig. S4, related to Figure 2. TAD calling validation.

Supplemental Fig. S5, related to Figure 2. Conserved SBs between *D. melanogaster* and *D. virilis* or *D. busckii*.

Supplemental Fig. S6, related to Figure 4. Characterization of conserved TADs (pink) compared to unconserved TADs (grey) and random unconserved regions (dark grey).

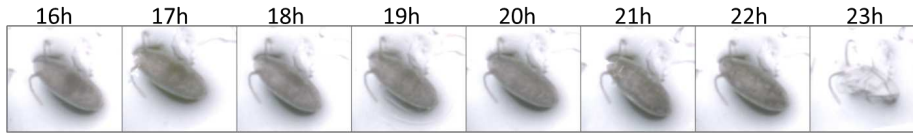
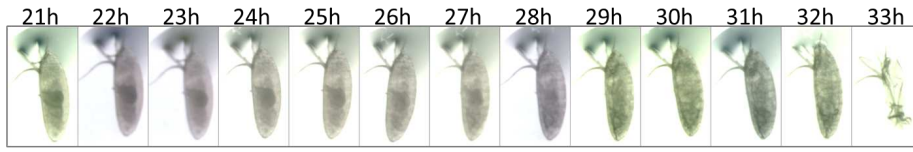
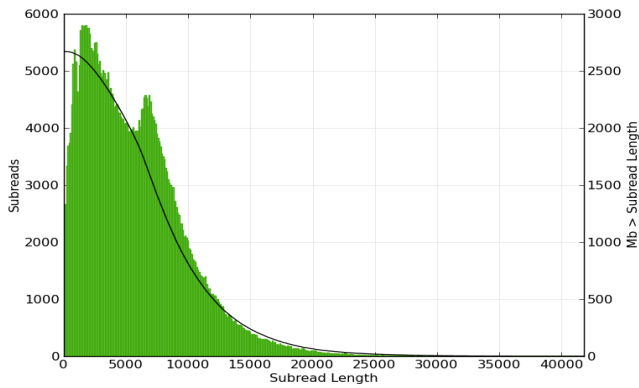
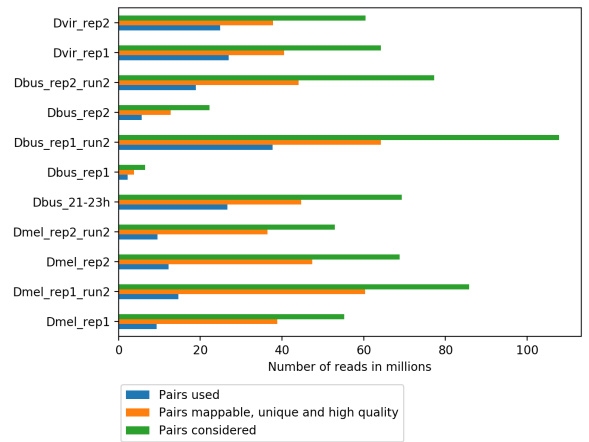
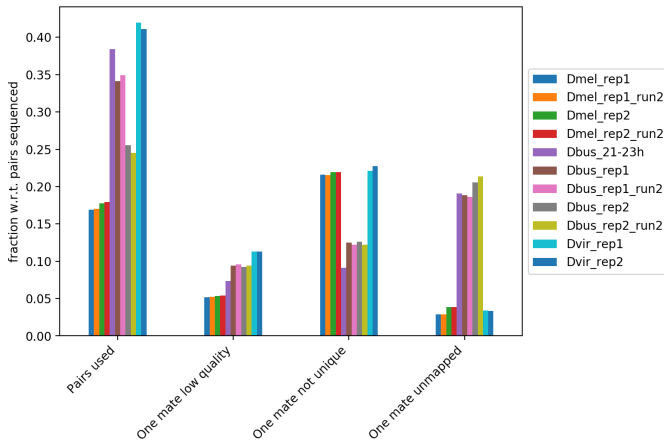
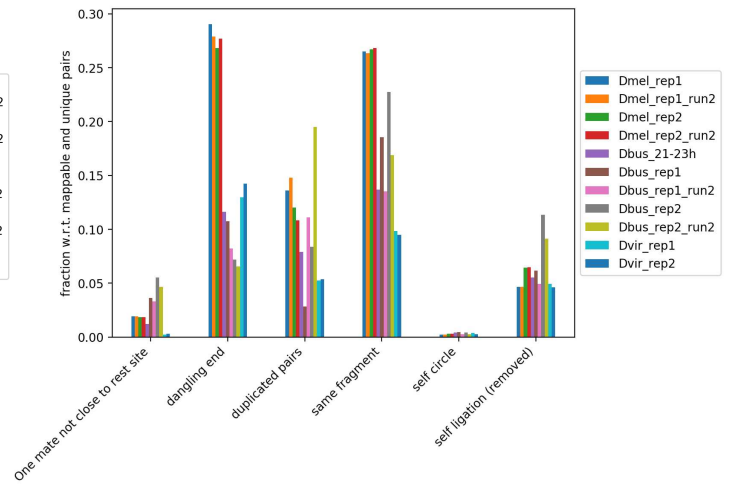
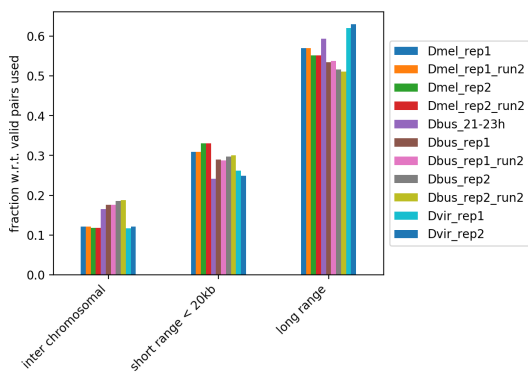
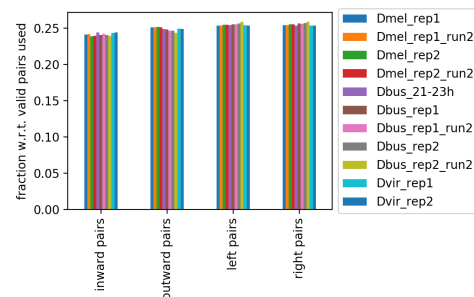
Supplemental Fig. S7, related to Figure 5 and 6. Immunostaining of polytene chromosomes and HAS characterization.

Supplemental Table S1, related to Figure 1. Qualitative comparison of HiCAssembler with other Hi-C scaffolding tools.

Supplemental Table S2, related to Figure 1. Sequenced and filtered valid reads of Hi-C samples.

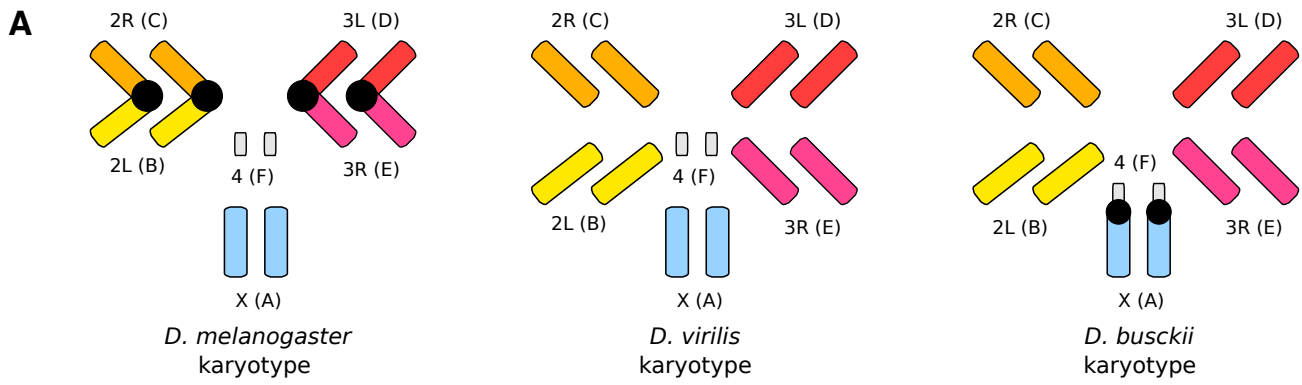
Supplemental Table S3, related to Figure 3A. Fischer two-tail test results for the overlap of TADs with SBs.

Supplemental References

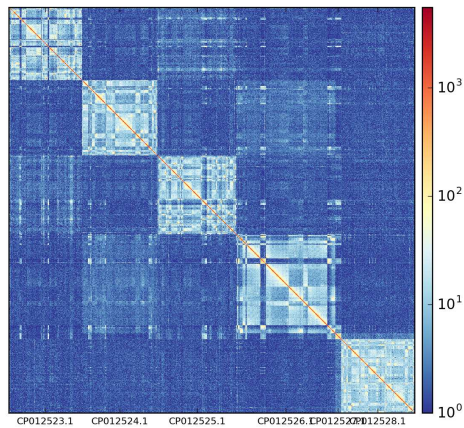
A *D. melanogaster***B** *D. virilis***C** *D. busckii***D****E****F****G****H****I****Figure S1**

Supplemental Fig. S1, related to Figure 1. Generated PacBio and Hi-C datasets.

(A,B,C) Developing *D. melanogaster* **(A)**, *D. virilis* **(B)** and *D. busckii* **(C)** embryo from stage 15-16 until hatching. The embryos were kept at 25°C and imaged every hour. The midgut is characteristically visible at stage 15 as a dark spot in the center of the embryo. The tracheal tree becomes visible at stage 17, just before hatching. **(D)** Histogram of the PacBio subread length. The mean subread length is 5.7 kb. **(E)** Number of sequenced (green) mappable, unique and high quality (orange) as well as valid used (blue) Hi-C reads per sequenced sample. If sequencing of one sample was split into two runs, the second run is indicated with 'run2'. Exact numbers of sequenced and valid reads are available in Table S2. **(F)** Fraction of used, unmappable and non-unique read pairs with respect to the sequenced read pairs. **(G)** Fraction of read pairs (with respect to mappable and unique reads) that were discarded when building the Hi-C matrix. **(H)** Fraction of inter-chromosomal, short-range (< 20 kb) and long-range Hi-C contacts with respect to the valid read pairs used. **(I)** Fraction of inward, outward, left and right read pairs with respect to the valid read pairs used.



B *D. busckii* ASM127793v1 assembly



C *D. busckii* neo-X

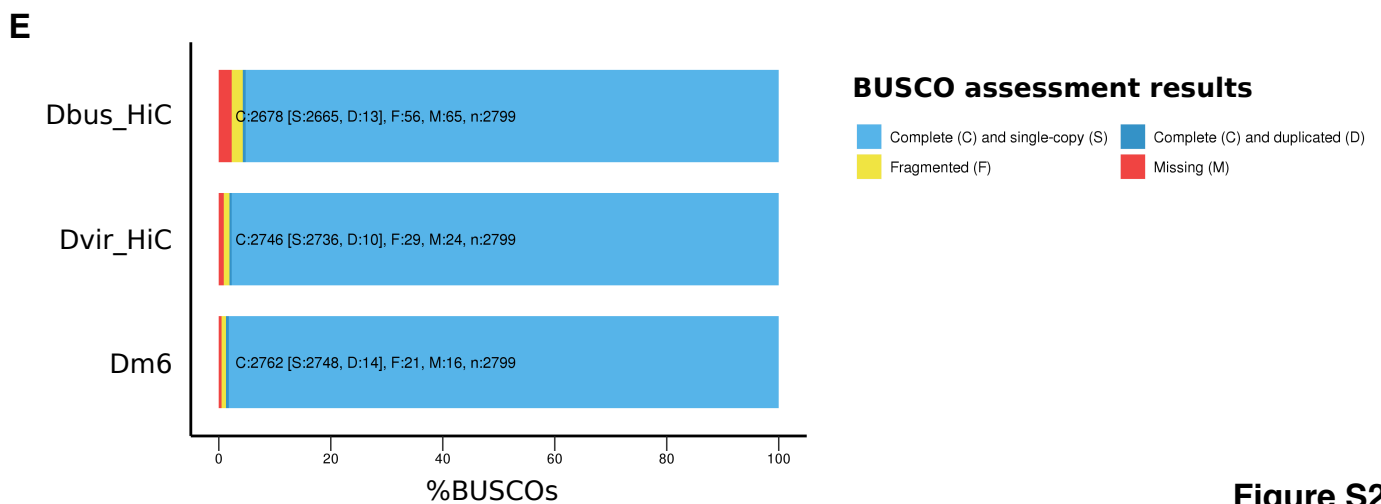
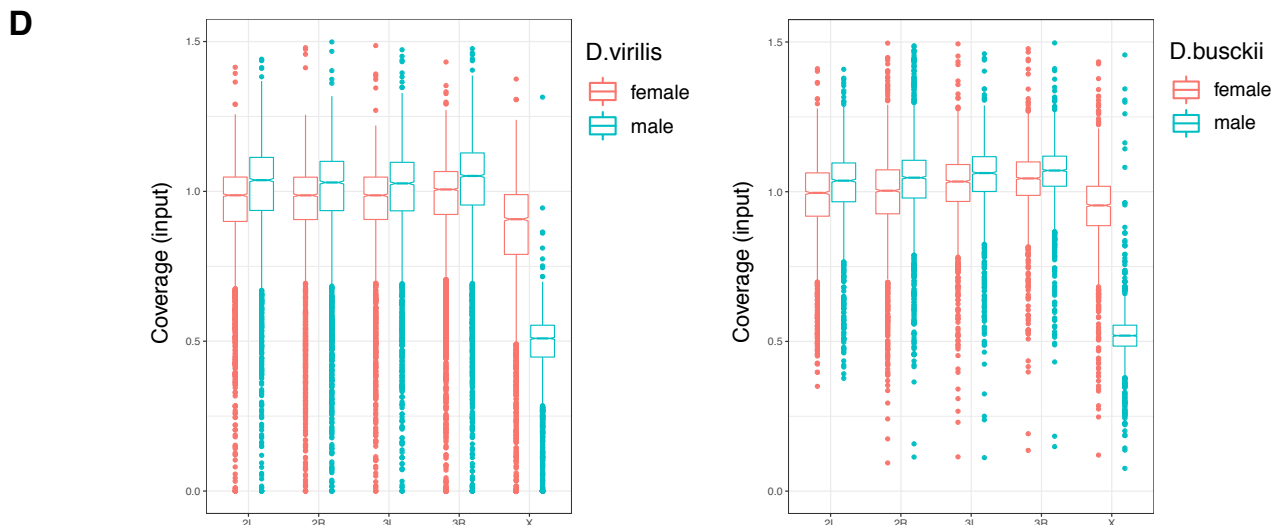
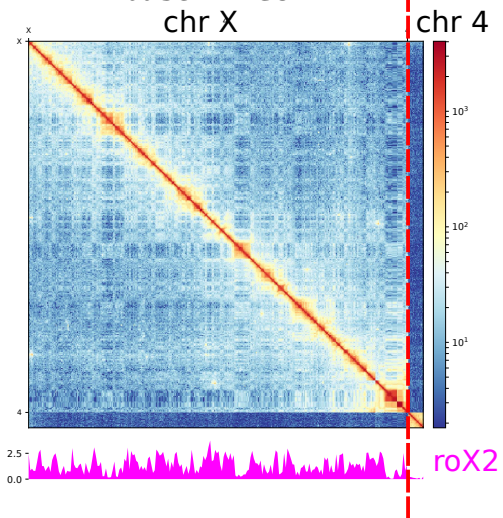


Figure S2

Supplemental Fig. S2, related to Figure 1. Hi-C assembly quality control.

(A) Female karyotypes of *D. melanogaster*, *D. virilis* and *D. busckii*. Chromosomal arms X, 2L, 2R, 3L, 3R and 4 corresponding to Muller elements A, B, C, D, E and F are shown as colored boxes. Fusions of chromosome arms are indicated with black circles. **(B)** The *D. busckii* ASM127793v1 assembly (current RefSeq representative genome) was generated by aligning assembled scaffolds to the *D. melanogaster* genome to generate pseudo-chromosomal sequences (Zhou and Bachtrog 2015). The equivocated order of scaffolds is visible as abruptly changing Hi-C contacts within and outside of chromosome arms. **(C)** roX2 ChIRP-seq data on the neo-X chromosome of *D. busckii*. In *D. busckii*, chromosome X (chr X) and small heterochromatic chromosome 4 (chr 4) are fused into one neo-X chromosome (Krivshenko 1959, 1955). Zhou and Bachtrog showed that Muller element F (chr 4) is not enriched for H4K16ac despite its fusion to the X chromosome (Zhou and Bachtrog 2015). Analysis of ChIRP-seq data of roX2 (Quinn et al. 2016) does not show enrichment of roX2 on chromosome 4 compared to the X chromosome as expected. The border between the X chromosome and chromosome 4 is highlighted with a red dashed line. **(D)** Coverage of mapped reads of male and female ChIP-seq input sample which was normalized to 1x (see methods) per 10 kb bin. The coverage of the male X chromosome is approximately 0.5 compared to 1 in females which shows that it is present in a single copy in males and two copies in females. **(E)** The *Diptera* (odb9) dataset containing 2799 Benchmarking Universal Single-Copy Orthologs (BUSCOs) (Waterhouse et al. 2017) was used to measure the completeness of our new genome assemblies (D_busckii_HiC and D_virilis_HiC) compared to the Dm6 genome. Reported are complete (C) and single-copy (S, light blue) or duplicated (D, dark blue) as well as fragmented (F, yellow) and missing (M, red) BUSCOs.

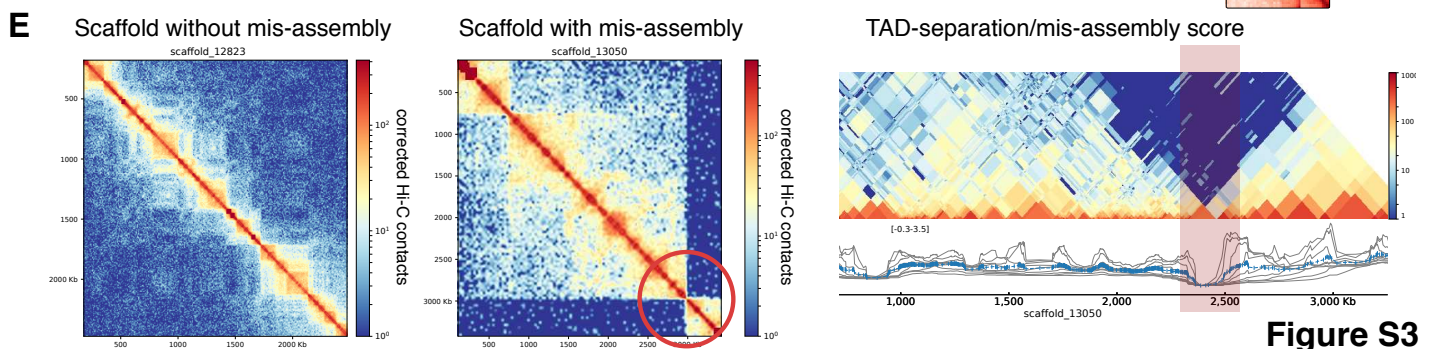
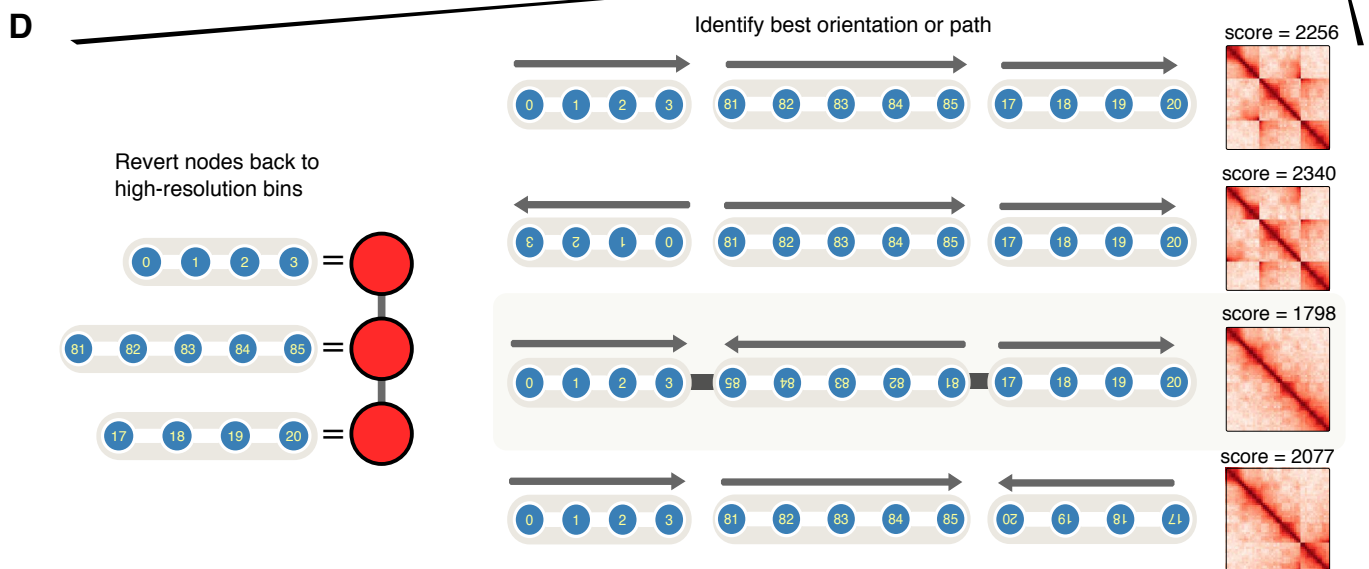
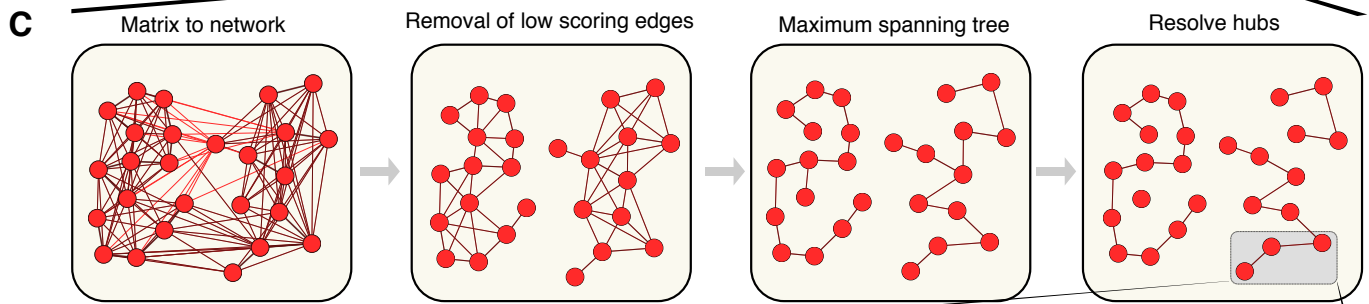
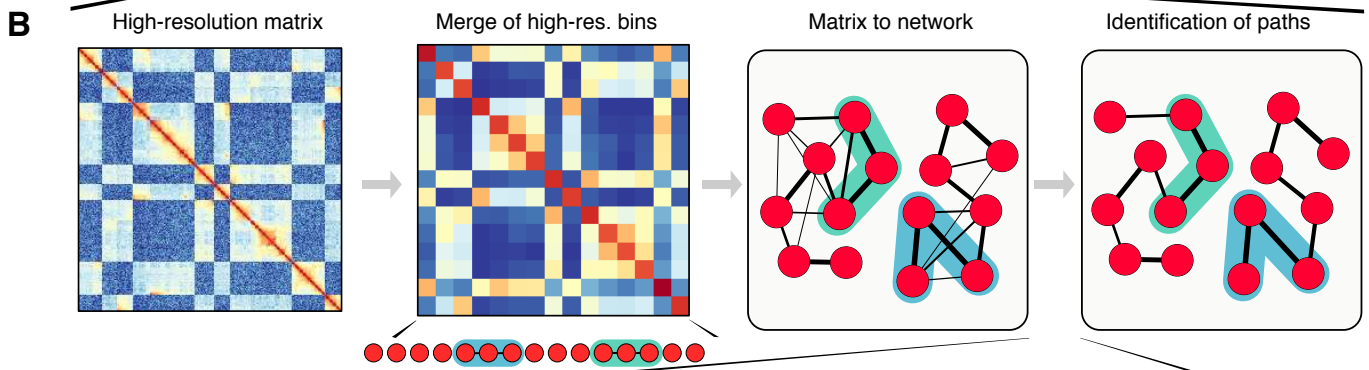
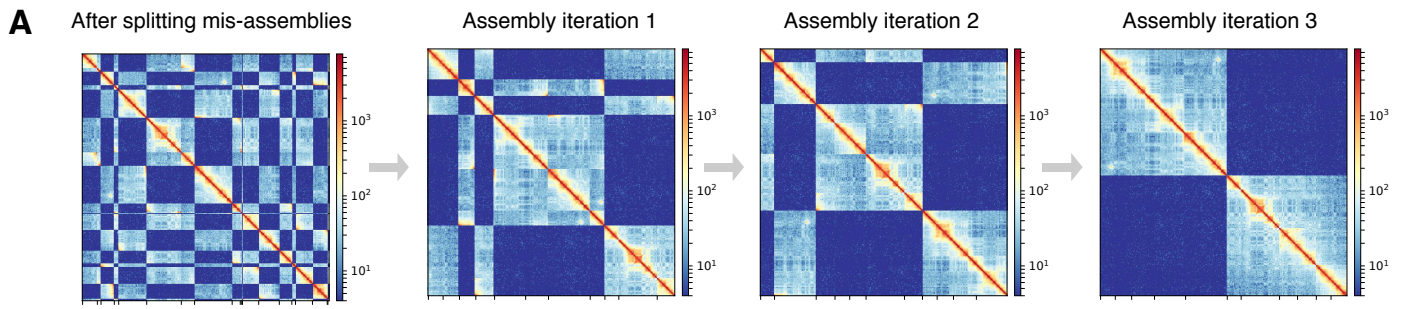


Figure S3

Supplemental Fig. S3, related to Figure 1. HiCAssembler.

(A) Example Hi-C matrix depicting the iterative progression of the Hi-C assembly strategy as in (Dudchenko et al. 2017). First, the Hi-C data is mapped to scaffolds, a Hi-C matrix of restriction fragment bins is created and corrected. Then, the original scaffolds are split if they contain mis-assemblies and small scaffolds are removed (first panel). In each iteration of the Hi-C assembly algorithm, scaffolds are joined and oriented to form what we call Hi-C scaffolds. After each iteration the Hi-C scaffolds become bigger until chromosome-length assemblies are obtained as shown in the last panel where two separated blocks remain. Afterwards, the small scaffolds that were initially removed are inserted into the Hi-C scaffolds. **(B)** Merging of high-resolution bins and determination of paths. In each iteration, all high resolution bins of the Hi-C matrix belonging to one of the Hi-C scaffolds are merged together (during the first iteration, Hi-C scaffolds are equal to the initial scaffolds). In the second panel (from left to right), each of the circles represent one Hi-C-scaffold whose high-resolution bins were merged. The new matrix is corrected to eliminate differences caused by merging Hi-C scaffolds of different length. Also, larger scaffolds are divided before merging them into equally long parts whose size is about the size of the smaller Hi-C scaffold. In the example shown, two large Hi-C scaffolds are divided into three parts; this is represented by the three connected circles with blue and green backgrounds. Using the divided scaffolds, the average number of corrected contacts between each part is computed to identify a cut-off threshold of the average number of counts between parts separated by one part. Next, the matrix is converted into a weighted graph to identify all Hi-C scaffolds that should be joined. In this graph, each node is a Hi-C scaffold (or a part of a Hi-C scaffold) and each edge's weight is the corrected number of contacts shared by any pair of Hi-C scaffolds. In the two rightmost panels, the Hi-C scaffolds that were divided are shown with the blue and green background and the edge weight is proportional to the line thickness. **(C)** Example usage of a larger network. For simplicity, subdivided Hi-C scaffolds are not shown. After the merged Hi-C matrix is converted into a weighted graph, all edges having a weight below the cut-off threshold are removed. This divides the network into connected components

representing groups of Hi-C scaffolds that are at most the length of one Hi-C scaffold apart in genomic distance. Then, the maximum spanning tree is computed and hubs are resolved by either pruning single nodes or by removing the weakest edges. The resulting graph contains the order in which the Hi-C scaffolds should be joined. **(D)** Once the ordering has been determined, the orientation of all parts is decided. For this, we use the high-resolution bins that form each Hi-C scaffold and compute their hic-score (see methods) in different orientations. The set of orientations that minimizes the hic-score is used to join the Hi-C scaffolds into a larger Hi-C scaffolds. In the figure, an example of a path of three Hi-C scaffolds is shown together with its respective high resolution bins (left). Four possible combinations of orientations are shown. For each case, the high-resolution submatrix (in log scale) containing only the bins of the three hic-scaffolds is shown on the right together with the hic-score. The third combination of orientations is the correct one as this produces the typical Hi-C matrix. **(E)** Depicted are two scaffolds of the Dvir_caf1 assembly, one without a mis-assembly (left) and one with a mis-assembly (middle). Mis-assemblies (red circles) are readily spotted as discontinuous regions in the Hi-C contact matrix that do not follow the power law decay with respect to genomic distance. Mis-assemblies can automatically be identified as positions that share significantly less contacts compared to the global average. This can be detected by HiCAssembler as minima in the TAD-separation/mis-assembly score (right).

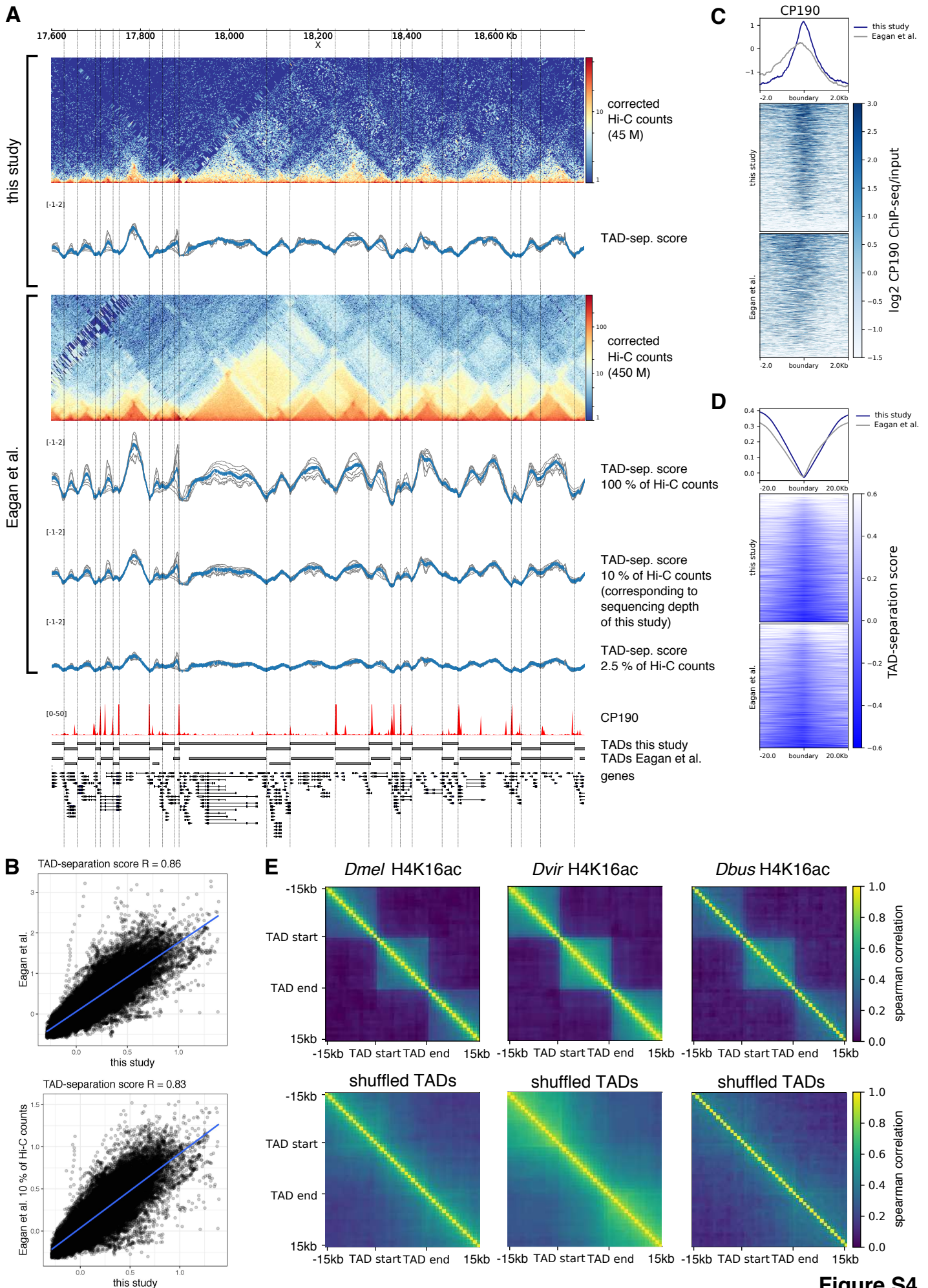


Figure S4

Supplemental Fig. S4, related to Figure 2. TAD calling validation.

(A) Example region on chromosome X showing a Hi-C contact matrix and TAD-separation score (TAD-sep. score) in *D. melanogaster* embryos (this study) and in a 10-fold deeper sequenced dataset from Kc167 cells (Eagen et al. 2017). The TAD-sep. score calculated using HiCExplorer is based on a z-score transformation of the Hi-C matrix as described in (Ramírez et al. 2018). TAD boundaries correspond to local minima of this score. TAD boundaries called using the TAD-sep. score in this study are shown as vertical lines. The TAD-sep. score calculated using HiCExplorer is robust to sequencing depth differences as shown by downsampling to 10 % (approximately corresponding to the sequencing depth used in this study) and 2.5 % of the Eagen et al. dataset. The following tracks show normalized ChIP-seq coverage for the common insulator protein cofactor CP190 (Li et al. 2015) (GSM1535980) as well as TADs called in this study using HiCExplorer (n= 2209) and TADs called by Eagen et al. using the Arrowhead algorithm (n=2127). **(B)** Spearman correlation of the TAD-sep. score from this study (45M Hi-C contacts) and Eagen et al. (450M Hi-C contacts) as well as from this study and downsampled dataset from Eagen et al. **(C)** Comparison of TAD boundaries called in this study and TAD boundaries called by (Eagen et al. 2017) with respect to CP190 (Li et al. 2015). Our boundaries overlap more frequently with CP190. **(D)** Similar to (C), but comparing TAD-sep. score. **(E)** Correlation of the histone mark H4K16ac (female 3rd instar larvae, this study) within and outside TADs (+15 kb flanking regions) as well as shuffled TADs as a control in *D. melanogaster*, *D. virilis* and *D. busckii*. Each pixel in the matrix represent the Spearman correlation of the histone mark in and outside of all TADs.

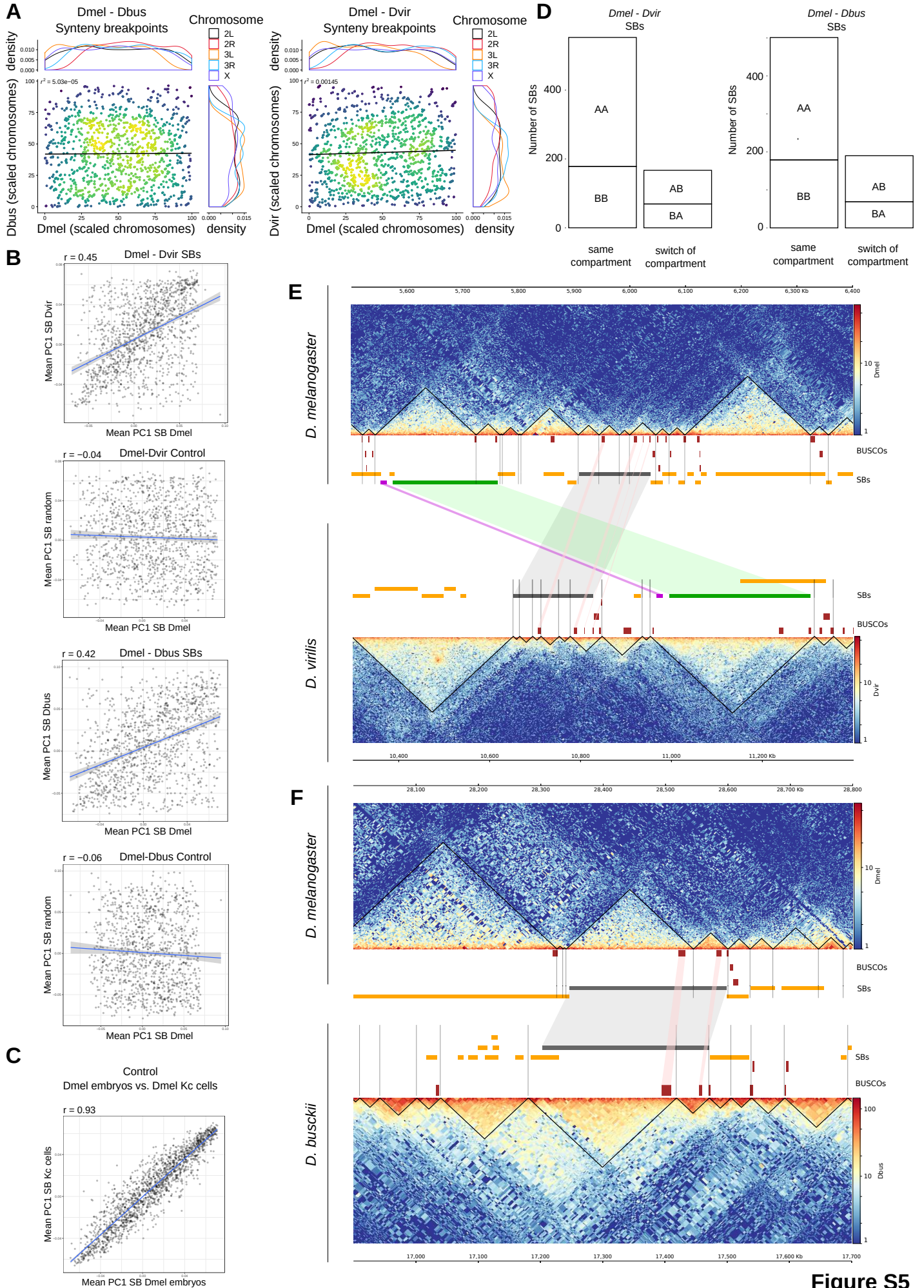


Figure S5

Supplemental Fig. S5, related to Figure 2. Conserved SBs between *D. melanogaster* and *D. virilis* or *D. busckii*.

(A) Density of syntenic block starts (syntenic breakpoints) for all chromosomes at once, scaled to the same size (*D. melanogaster* compared to *D. busckii* (left panel) or to *D. virilis* (right panel)). The black line corresponds to the linear regression calculated from the scatter plot (and its associated r^2 in the top left corner). The density of syntenic breakpoints along each chromosome in each species (scaled to the same size), is shown in the top (for the x axis) and right (for the y axis) part of each plot. **(B)** From top to bottom. Pearson correlation of mean PC1 (A/B compartment) of corresponding SB in *D. melanogaster* (Dmel) and *D. virilis* (Dvir), of random SB in Dvir and actual SB in Dmel, of corresponding SB in Dmel and *D. busckii* (Dbus) and of random SB in Dbus and actual SB in Dmel after removal of outliers ($-0.1 < PC1 < 0.1$). **(C)** Pearson correlation of A/B compartments (PC1) of syntenic blocks between the *D. melanogaster* embryo (our data) and *D. melanogaster* Kc cell dataset (Eagen et al. 2017). **(D)** Number of *Dmel-Dvir* or *Dmel-Dbus* SBs that lie completely in the A or B compartment in both species and stay in the same compartment (AA and BB) or switch compartment (AB and BA) between both species. **(E)** Hi-C contact heatmaps including called TADs (black triangles), BUSCOs (red boxes) and SBs (orange boxes) in *D. melanogaster* and *D. virilis*. Correspondence between conserved SBs (gray, green and magenta boxes) is indicated with connections between the SBs in both species. Corresponding BUSCOs are indicated with connections between them. **(F)** Comparison of a 800 kb region in between *D. melanogaster* and *D. busckii*. One corresponding SB (gray box) is indicated with a gray connection between both species and corresponding BUSCOs are indicated with light red connections. The Hi-C profile of the corresponding SB is similar in both species.

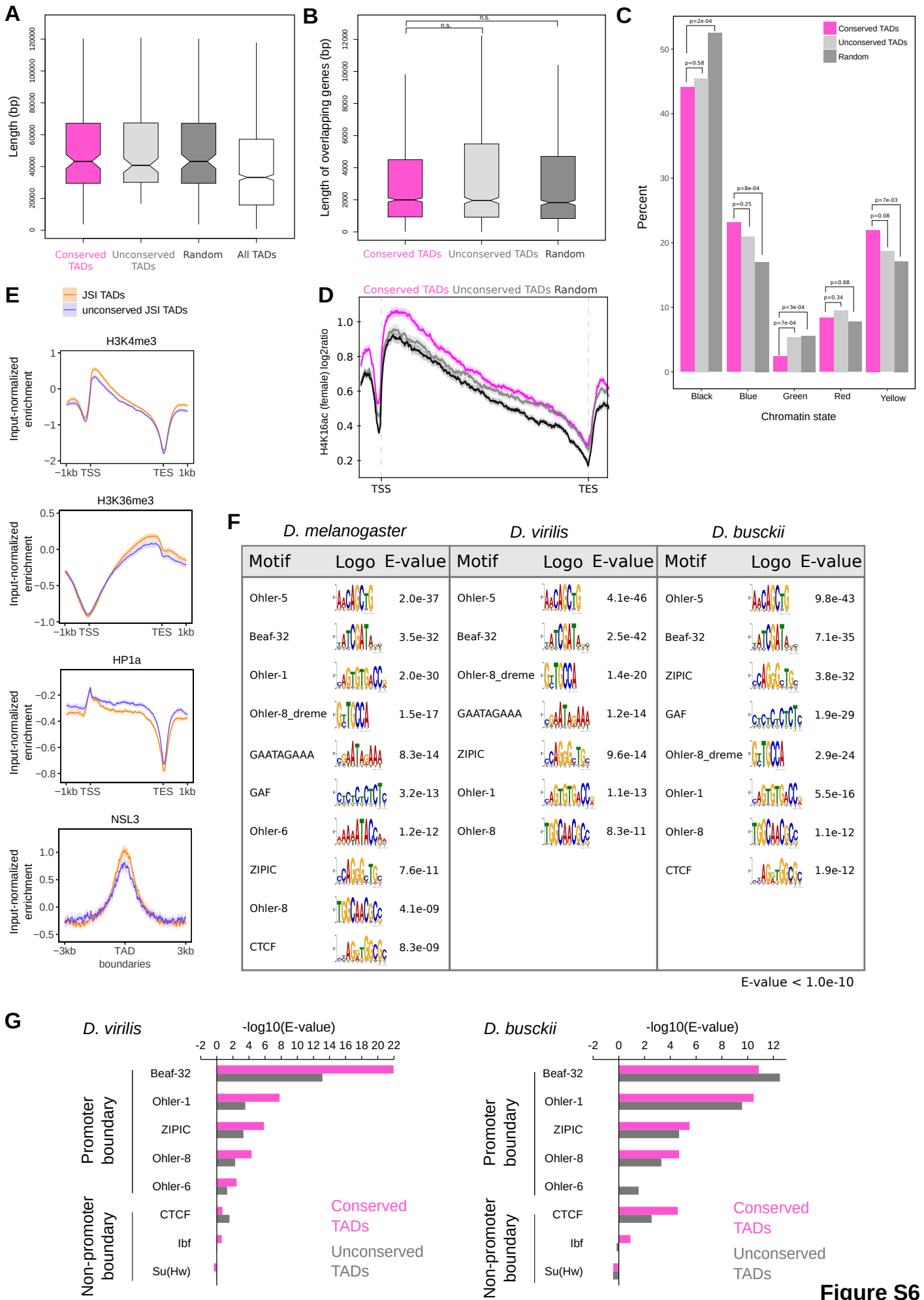
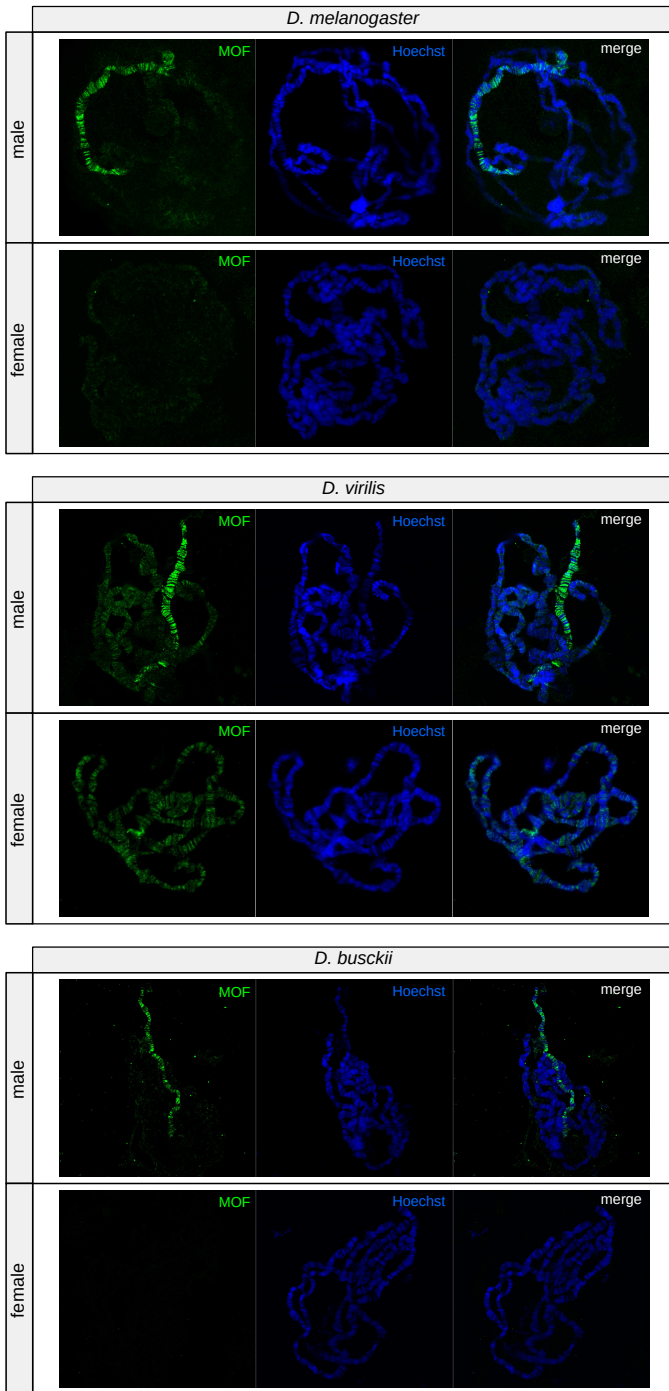


Figure S6

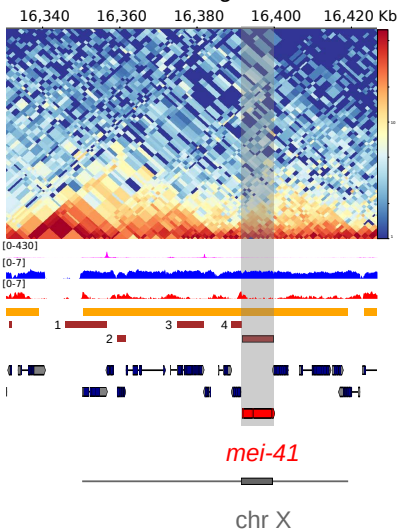
Supplemental Fig. S6, related to Figure 4. Characterization of conserved TADs (pink) compared to unconserved TADs (grey) and random unconserved regions (dark grey).

(A) Size distribution of the three test sets compared to all TADs (white). **(B)** Length of genes overlapping with all test sets. This is a control for Figure 4B. **(C)** Percentage of overlap of conserved TADs, unconserved TADs and random genomic regions with black, blue, green, red and yellow chromatin states defined in (Filion et al. 2010). *P*-values were calculated using Wilcoxon rank-sum tests on the chromatin states coverage (expressed in kb). **(D)** Female H4K16ac signal over input of genes (transcription start site (TSS) to transcription end site (TES)) in conserved TADs (pink), unconserved TADs (grey) and random regions (black). ChIP-seq profiles show mean (thick line) and standard error (shadowed area) of input-normalized ChIP-seq enrichment along scaled genes and unscaled 1 kb before the TSS and after the TES. **(E)** Log₂ratio of H3K4me₃, H3K36me₃, HP1 alpha (Celniker et al. 2009) and NSL3 (Lam et al. 2012) ChIP-seq reads over input reads along genes (transcription start site (TSS) to transcription end site (TES)) or at TAD boundaries of conserved and unconserved TADs by Jaccard similarity index (JSI). ChIP-seq profiles show mean (thick line) and 95 % CI (shadowed area) of input-normalized ChIP-seq enrichment, computed using deepStats (Richard 2019). **(F)** Enrichment of TAD boundary motifs in *D. melanogaster*, *D. virilis* and *D. busckii*. Enrichment analysis using AME from the MEME suite (McLeay and Bailey 2010) for motif enrichment in 500 bp around TAD boundaries in *D. melanogaster*, *D. virilis* and *D. busckii*. We used motifs that were previously described in *D. melanogaster* (Ramírez et al. 2018) and analyzed their enrichment against shuffled input sequences as control, average odds score as the sequence scoring method and one-tailed Wilcoxon rank-sum test as the motif enrichment test. **(G)** Analysis of motif enrichment of promoter and non-promoter boundary motifs at conserved TAD boundaries in *D. virilis* and *D. busckii* compared to unconserved TAD boundaries using AME from the MEME suite (McLeay and Bailey 2010).

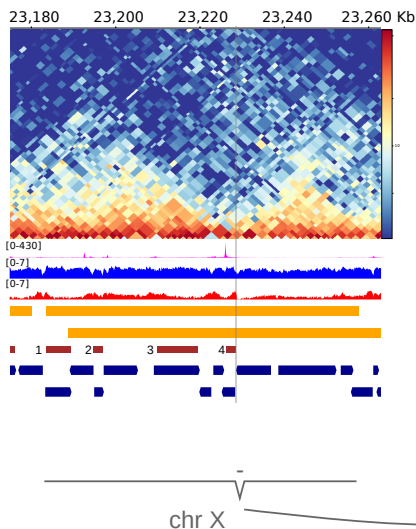
A Immunostaining of polytene chromosomes



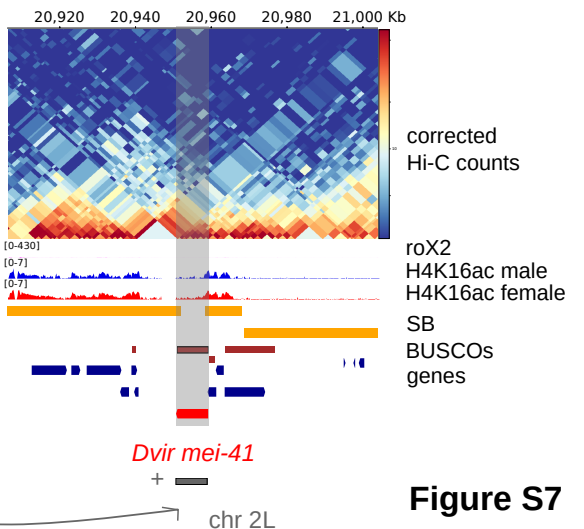
B *D. melanogaster*



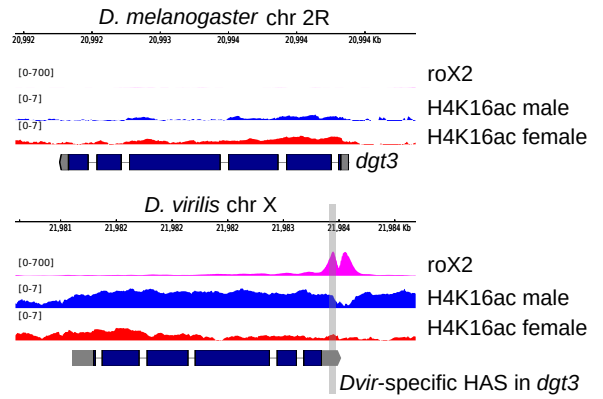
D. virilis



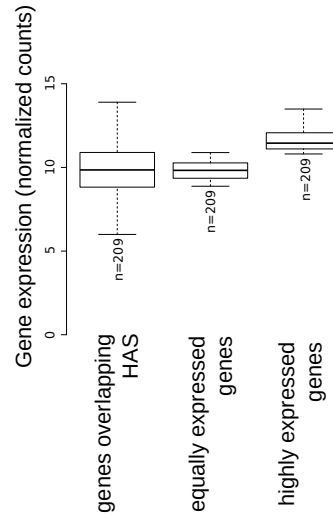
D. virilis



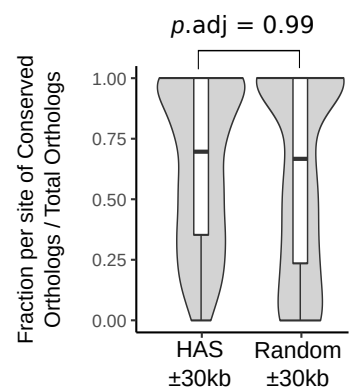
C



D



E



F

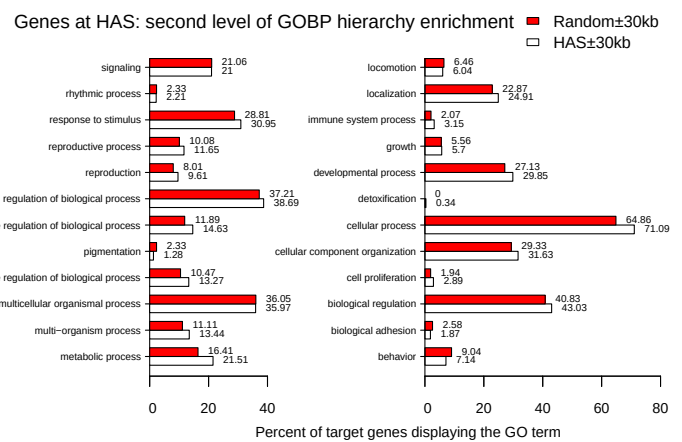


Figure S7

Supplemental Fig. S7, related to Figure 5 and 6. Immunostaining of polytene chromosomes and HAS characterization.

(A) Male and female polytene chromosomes were stained with MOF antibody (green) in *D. melanogaster*, *D. virilis* and *D. busckii*. DNA is counterstained with Hoechst (blue). MOF shows enrichment on the X chromosome compared to all other chromosomes in males but not females in all three species. **(B)** Example gene that moved between the X chromosome and an autosome when comparing *D. melanogaster* and *D. virilis*. *mei-41* is localized on chromosome X in *D. melanogaster* but on chromosome 2L in *D. virilis*. The surrounding genes on this SB on chromosome X maintained the same order (see corresponding BUSCOs numbered from 1-4 and gene track). *mei-41* in *D. virilis* (Dvir GJ22445) is localized on chromosome 2L in between two surrounding SBs and only has a H4K16ac peak at the promoter (male and female) but the gene body is not covered with H4K16ac. **(C)** The gene *dim gamma-tubulin 3* (Dmel\dgt3; Dvir\GJ18725) moved from chromosome 2L in *D. melanogaster* to the X chromosome in *D. virilis* and there gained roX2 binding in its 3' UTR. Thus, this HAS is a species-specific one. **(D)** Gene expression (normalized counts) of genes overlapping with HAS compared to equally expressed genes and highly expressed genes (n = 209), referring to Figure 6E. Comparison of gene expression was performed using library size normalized RNA-seq counts from 14-20 h aged embryos from modENCODE datasets obtained from (Ramírez et al. 2018) and also available on the chorogenome web server (<http://chorogenome.ie-freiburg.mpg.de/>). **(E)** Fraction of genes that are conserved compared to all genes per corresponding HAS pair between between *D. melanogaster* and *D. virilis* compared to random regions, both extended by 30 kb in 5' and 3' direction. Only genes with *D. melanogaster* to *D. virilis* orthologs were considered. The plotted distributions correspond to the fraction of orthologs in the *D. melanogaster* genome computed for each HAS or random site. The best matching pair of HAS or random region between both species was defined based on the highest (conserved orthologs / total orthologs) ratio. The displayed random regions boxplot and violin plot correspond to one of the 100 random sets analysed. The adjusted *p*-value is computed on all 100 random sets and was obtained by comparing

the HAS distribution to every random set distribution using the Wilcoxon rank-sum test and Benjamini-Hochberg correction for multiple comparisons. **(F)** Gene Ontology Biological Process (GOBP) enrichment at level 2 of the GO hierarchy for genes located within 60 kb around HAS (HAS \pm 30kb, white) and within 60 kb around random active sites (Random \pm 30kb, red) using goProfiles (<https://doi.org/doi:10.18129/B9.bioc.goProfiles>). No enrichment is statistically significant between the HAS \pm 30kb and Random \pm 30kb gene lists using Holm corrected Fisher exact test (as suggested in the goProfiles vignette). The values are expressed in percent of the genes displaying a given GO term out of the total number of genes contained in each gene list.

Supplemental Tables

Supplemental Table S1, related to Figure 1. Qualitative comparison of HiCAssembler with other Hi-C scaffolding tools.

Hi-C scaffolding tool	HiCAssembler	3D-DNA	SALSA
Reference	this study https://github.com/mexplanck-ie/HiCAssembler	Dudchenko et al. 2017 https://github.com/theaidenlab/3d-dna	Ghurye et al. 2017 https://github.com/machinegun/SALSA
Documentation			
Aim (Proof-of-concept/Hi-C scaffolding tool)	Tool	Tool	Tool (latest version SALSA2)
Requirements	Python dependencies (pip)	Java version >=1.7 Bash >=4 GNU Awk >=4.0.2 GNU coreutils sort >=8.11 LastZ version 1.03.73 (for diploid mode only) Python >=2.7 (for chromosome number-aware splitter module only) scipy numpy matplotlib (for chromosome number-aware splitter module only)	Python 2.7 BOOST libraries networkx
Input	pre-assembled contigs or scaffolds Hi-C interaction matrix in .h5 format (from HiCExplorer Hi-C analysis software)	pre-assembled contigs or scaffolds Hi-C interaction matrix in .hic format (from Juicebox Hi-C analysis software)	pre-assembled contigs or scaffolds aligned Hi-C data (.bed format)
Predicts number of chromosomes	yes	yes	yes
Orients contigs	yes	yes	yes
Automatic misassembly detection	yes	yes	yes
Manual misassembly detection	yes	yes	no
Accounts for duplications	no	under development	no
Output	genome.fasta file .h5 file liftover.chain file (allows liftover of data from initial contigs to assembly) files for visualization of assembly path graph (.graphml files)	genome.fasta file .hic file scaffold boundary files tracking file of modifications to the input contigs (.assembly file)	genome.fasta file
Visualization	automatic visualization at each assembly iteration	Juicebox Assembly Tools (JBAT) can be used for visualization	no
Tested genome(s)	<i>Drosophila busckii</i> , <i>Drosophila virilis</i>	<i>Aedes aegypti</i> , <i>Culex quinquefasciatus</i>	human, goat

Hi-C scaffolding tool	LACHESIS	DNA Triangulation	GRAAL
Reference	Burton et al. 2013 https://github.com/shendurelab/LACHESIS	Kaplan and Dekker 2013 https://github.com/NoamKaplan/dna-triangulation	Marie-Nelly et al. 2014 https://github.com/koszulab/instaGRAAL
Documentation			
Aim (Proof-of-concept/Hi-C scaffolding tool)	Tool (not developed anymore)	Proof-of-concept	Tool (latest version instaGRAAL)
Requirements	unix environment gcc zlib compression library boost C++ libraries SAMtools toolkit	Python modules (numpy, scipy, scikit-learn)	NVIDIA graphic card Cuda Toolkit PyCuda with OpenGL support Python dependencies (pip)
Input	pre-assembled contigs or scaffolds aligned Hi-C data (.bam format) ini file specifying several parameters	pre-assembled contigs or scaffolds of equally-sized bins Hi-C contact matrix	contigs or scaffolds split into bins of several restriction fragments Hi-C contact matrix at several resolutions (pyramid file)
Predicts number of chromosomes	no	yes	yes
Orients contigs	yes	no	yes
Automatic misassembly detection	no	no	yes
Manual misassembly detection	no	no	no
Accounts for duplications	no	no	yes
Output	genome.fasta file assembly metrics no	position of each bin (in an arbitrary units)	fasta file of genome assembly with the highest likelihood info about initial fragments assembly movie
Visualization		no	
Tested genome(s)	human, mouse and <i>Drosophila</i>	human	<i>Saccharomyces cerevisiae</i> , <i>Trichoderma reesei</i> , human

Supplemental Table S2, related to Figure 1. Sequenced and filtered valid reads of Hi-C samples. Upper panel contains numbers of sequenced reads for each sample. Some samples were sequenced in two runs to get higher sequencing depth which is indicated as 'Run_1' and 'Run_2'. Lower panel contains valid used reads for each sample after creation of the Hi-C contact matrix using `hicBuildMatrix` from HiCExplorer (see methods: Creation of Hi-C contact matrix or (Ramírez et al. 2018)), as well reported in percent with respect to pairs sequenced.

	Sample	<i>Dmel</i> Rep1	<i>Dmel</i> Rep2	<i>Dvir</i> Rep1	<i>Dvir</i> Rep2	<i>Dbus</i> Rep1	<i>Dbus</i> Rep1	<i>Dbus</i> 21-23h
Sequenced reads	Run 1	55,222,595	68,801,187	64,177,938	60,426,728	6,491,637	22,258,838	69,290,579
	Run 2	85,848,394	52,938,683	0	0	107,801,339	77,256,684	0
	Total	141,070,989	121,739,870	64,177,938	60,426,728	114,292,976	99,515,522	69,290,579
	Total Rep1&2	262,810,859		124,604,666		213,808,498		69,290,579

	Sample	<i>Dmel</i> Rep1	<i>Dmel</i> Rep2	<i>Dvir</i> Rep1	<i>Dvir</i> Rep2	<i>Dbus</i> Rep1	<i>Dbus</i> Rep1	<i>Dbus</i> 21-23h
Useful reads HiCExplorer 1.8.1	Run 1	9,349,695	12,242,090	26,929,877	24,857,426	2,216,509	5,687,685	26,811,828
	Run 2	14,611,364	9,501,536	0	0	37,669,762	18,934,139	0
	Total	23,961,059	21,743,626	51,787,303		39,886,271	24,621,824	26,811,828
	Total Rep1&2	45,704,685		51,787,303		64,508,095		26,811,828

	Sample	<i>Dmel</i> Rep1	<i>Dmel</i> Rep2	<i>Dvir</i> Rep1	<i>Dvir</i> Rep2	<i>Dbus</i> Rep1	<i>Dbus</i> Rep1	<i>Dbus</i> 21-23h
Useful / Sequenced reads	Run 1 (%)	16.9	17.8	42.0	41.1	34.1	25.6	38.7
	Run 2 (%)	17.0	18.0	NA	NA	34.9	24.5	NA
	Average (%)	17.4		41.6		29.8		38.7

Supplemental Table S3, related to Figure 3A. Fischer two-tail test results for the overlap of TADs with SBs.

Comparison	Species	Condition	Fisher two tail <i>p</i> -value	Overlapping proportion
<i>D. busckii</i> vs <i>D. melanogaster</i>	<i>D. busckii</i>	Called TADs	2.29E-62	0.17
		Shuffled SBs	5.55E-01	0.07
		Shuffled TADs	6.76E-01	0.07
		Shuffled TADs & SBs	3.00E-01	0.07
	<i>D. melanogaster</i>	Called TADs	4.84E-193	0.22
		Shuffled SBs	5.88E-01	0.06
		Shuffled TADs	5.86E-02	0.07
		Shuffled TADs & SBs	9.70E-01	0.06
<i>D. busckii</i> vs <i>D. melanogaster</i>	<i>D. virilis</i>	Called TADs	1.41E-65	0.13
		Shuffled SBs	5.85E-02	0.06
		Shuffled TADs	6.81E-01	0.05
		Shuffled TADs & SBs	7.71E-01	0.05
	<i>D. melanogaster</i>	Called TADs	6.33E-212	0.22
		Shuffled SBs	8.36E-01	0.06
		Shuffled TADs	5.04E-01	0.06
		Shuffled TADs & SBs	7.75E-01	0.06

References

- Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat Biotechnol* **31**: 1119–1125.
- Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, et al. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**: 92–95.
- Eagen KP, Aiden EL, Kornberg RD. 2017. Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. *Proc Natl Acad Sci U S A* **114**: 8764–8769.
- Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, et al. 2010. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* **143**: 212–224.
- Ghurye J, Pop M, Koren S, Bickhart D, Chin C-S. 2017. Scaffolding of long read assemblies using long range contact information. *BMC Genomics* **18**: 1–11.
- Kaplan N, Dekker J. 2013. High-throughput genome scaffolding from *in vivo* DNA interaction frequency. *Nat Biotechnol* **31**: 1143–1147.
- Krivshenko J. 1959. New Evidence for the Homology of the Short Euchromatic Elements of the X and Y Chromosomes of *Drosophila Busckii* with the Microchromosome of *Drosophila Melanogaster*. *Genetics* **44**: 1027–1040.
- Krivshenko JD. 1955. A CYTOGENETIC STUDY OF THE X CHROMOSOME OF *Drosophila Busckii* AND ITS RELATION TO PHYLOGENY. *Proc Natl Acad Sci U S A* **41**: 1071–1079.
- Li L, Lyu X, Hou C, Takenaka N, Nguyen HQ, Ong C-T, Cubeñas-Potts C, Hu M, Lei EP, Bosco G, et al. 2015. Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. *Mol Cell* **58**: 216–231.
- Marie-Nelly H, Marbouty M, Cournac A, Flot J-F, Liti G, Parodi DP, Syan S, Guillén N, Margeot A, Zimmer C, et al. 2014. High-quality genome (re)assembly using chromosomal contact data. *Nat Commun* **5**: 5695.
- McLeay RC, Bailey TL. 2010. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* **11**: 165.
- Quinn JJ, Zhang QC, Georgiev P, Ilik IA, Akhtar A, Chang HY. 2016. Rapid evolutionary turnover underlies conserved lncRNA-genome interactions. *Genes Dev* **30**: 191–207.
- Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Habermann B, Akhtar A, Manke T. 2018. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun* **9**: 189.
- Waterhouse RM, Seppely M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2017. BUSCO applications from quality assessments to gene

prediction and phylogenomics. *Mol Biol Evol.*

Zhou Q, Bachtrog D. 2015. Ancestral Chromatin Configuration Constrains Chromatin Evolution on Differentiating Sex Chromosomes in *Drosophila*. *PLoS Genet* **11**: e1005331.