Title: Hi-C approach in revealing the principles of 3D-genomic organization in Anopheles mosquitoes

Varvara Lukyanchikova[1,2,3,4,+], Miroslav Nuriddinov[3,+], Polina Belokopytova[3,4], Jiangtao Liang[1,2], Maarten J.M.F. Reijnders[5], Livio Ruzzante[5], Robert M. Waterhouse[5], Zhijian Tu[2,6], Igor V. Sharakhov[1,2,7,*], Veniamin Fishman[3,4,*]

1. Department of Entomology, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

2. Fralin Life Science Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

3. Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

4. Novosibirsk State University, Novosibirsk, Russia

5. Department of Ecology and Evolution, University of Lausanne, and Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

6. Department of Biochemistry, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

7. Department of Cytology and Genetics, Tomsk State University, Tomsk, Russia

[+] co-first authors

*** correspondence to I.V.S. (igor@vt.edu) and V.F. (minja-f@ya.ru)**

Abstract

Chromosomes are hierarchically folded within cell nuclei into territories, domains and subdomains, but their functional importance and evolutionary dynamics are not well defined. Here, we comprehensively profiled genome organization of five Anopheles species.

We showed that chromatin interactions are linked to genetic and epigenetic profiles are connected to each other and how epigenetic properties

Introduction

3-dimensional genome organization has recently gained attention as a complex and dynamic mechanism of gene regulation. Development of chromosome conformation

capture methods allowed studying chromatin contacts genome-wide at high resolution. In addition, data obtained using 3C-methods could be used to generate chromosome-level genome assemblies, allowing comprehensive analysis of genomes evolution.

Comparative studies performed on multiple vertebrate species revealed that genome architecture is evolutionarily conserved and could be explained by dynamic interplay between processes of cohesin-mediated loop extrusion and chromatin compartmentalization. In insects, comprehensive analysis and cross-species comparison of genome architecture were performed only for *Drosophila* species. These studies suggested that, in contrast to mammals, the process of loop extrusion does not define the structure of chromatin contacts. Instead, separation of active and repressed chromatin plays an essential role in formation and interaction of topologically associated domains (TADs), which are basic units of chromatin organization in *Drosophila*.

Recently Ghavi-Helm et al. suggested that breaking TADs does not influence gene expression based on analysis of *Drosophila* lines with highly rearranged genomes. In the same time, Renschler et al. used three distantly related *Drosophila* species to show that chromosomal rearrangements might shuffle the position of the entire TAD in the genome but preferentially maintain TADs as units. Thus, the role of 3-dimensional chromatin interactions in functioning and evolution of insect genomes is currently unclear.

To fill this gap, we have studied genome architecture of five mosquito species belonging to *Anopheles* genus using the Hi-C technique. Malaria has a devastating global impact on public health and welfare and *Anopheles* mosquitoes are exclusive vectors of malaria. In addition, sequencing and characterization of the genomes for 16 *Anopheles* mosquito species, including *An. atroparvus*, (Neafsey *et al.* 2015) has discovered a high rate of chromosomal rearrangements especially on the X chromosome, which makes them attractive as a model for studying interconnections between structural variations, chromosome evolution and genome architecture.

The obtained Hi-C data allowed us to improve existing genome assemblies of three mosquito species and generate new chromosome-length assemblies for two more species. Our analysis of TADs and compartments, supplemented by improved algorithms for compartments identification, demonstrated conservation of principles organizing chromatin in *Anopheles* species and other insects.

We found specific looping interactions, sometimes spanning over several dozens of megabases (Mb) in all studied *Anopheles* species, and showed that these interactions are evolutionary conserved for at least 80 million years. We generated RNA-seq and ChiP-seq data to show that these loops can not be explained using known molecular mechanisms and represent a specific type of chromatin interactions.

Aggregating long-range chromatin interactions, we found that there is a decrease of contact probabilities beyond a certain genomic distance. Performing broad evolutionary comparison between Anopheles species, other insects, and vertebrate data we showed that this limiting genomic distance is taxon-dependent and suggested a mechanistic explanation of this phenomenon.

Results

Hi-C-guided genome assembly of Anopheles mosquitoes.

In the Hi-C experiment we used 15-18h embryos of mosquito species from 3 different subgenera of *Anopheles* genus: Cellia (*An. coluzzii*, *An. merus*, *An. stephensi*), Nyssorhynchus (*An. albimanus*) and Anopheles (*An. atroparvus*) (Fig. 1, A-C). In addition, we sequenced Hi-C libraries from adult *An. merus* mosquito. Based on our analysis, phylogenetic relationships of the selected species represent a broad range of evolutionary distances, from 0.5 million years (MY) between closely related *An. coluzzii* and *An. merus* species (Thawornwattana et al, 2018) to 100 MY separating the most distant lineages such as *An. coluzzii* and *An. albimanus* (Neafsey et al 2015) (Fig. 1, A). Our RAxML molecular phylogeny shoes the following branching pattern (((((Anopheles_coluzzii:0.02132, Anopheles_merus:0.00671):0.06419, Anopheles_stephensi:0.08062):0.06959, Anopheles_atroparvus:0.12526):0.03922, Anopheles_albimanus:0.16921) :0.11740, Aedes_aegypti:0.29684). The RAxML time-calibrated phylogeny suggest the following divergence times in MY among the mosquito lineages ((((Anopheles_coluzzii:0.50000, Anopheles_merus:0.50000): 39.05308, Anopheles_stephensi:39.55308):37.17110, Anopheles_atroparvus:76.72418):23.27582, Anopheles_albimanus:100.00000):65.05236, Aedes_aegypti:165.05236).

To date, chromosome-length assemblies were already available for 2 species (*An. albimanus*, *An. atroparvus*), whereas for *An. coluzzii*, *An. merus* and *An. stephensi* there were only scaffolds/contigs published with N50 equal to 330-kb (Kingan et al., 2019), 37-kb and 793-bp correspondingly. While evolutionary superscaffolding and chromosomal

anchoring improved these assemblies, they did not reach full chromosomal level (Waterhouse et al., 2020).

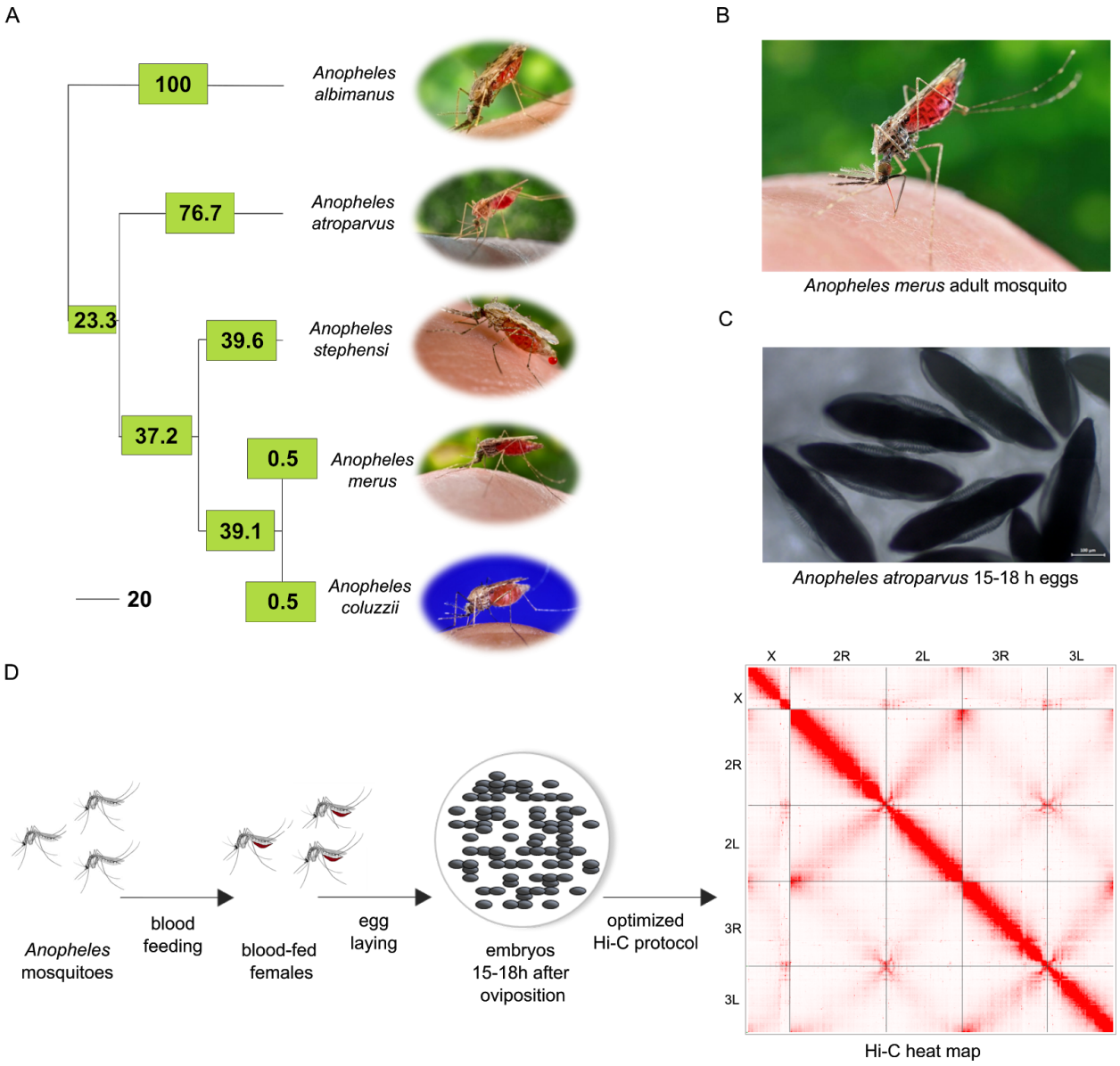

Fig. 1. ***Anopheles species taken into Hi-C experiment.*** *A) A time-calibrated phylogenetic tree shows evolutionary distance among the experimental Anopheles species; numbers in the box show time distances in MY for each branch; the scale bar = 20 MY; B) Anopheles merus adult female mosquito; C) Anopheles atroparvus 15-18h embryo developmental stage. Adult mosquito illustrations were taken from VectorBase repository; D) The scheme of Hi-C experiment.*

After sequencing the libraries and merging biological replicas we obtained 60-194 mln of unique alignable reads for each species (Supplementary Table 1). Library statistics show high quality of the obtained data (Supplementary Table 1).

Genomes for Anopheles species have been challenging to assemble to chromosomal levels; due to the presence of high repetitive DNA clusters, regular Illumina sequencing leaves a large part of mosquito genomes uncovered (ref). We employed 3D-DNA pipeline to assemble *An. coluzzii*, *An. merus* and *An. stephensi* genomes *de novo* using generated Hi-C data set. We reassembled chromosomes of *An. albimanus* and *An. atroparvus* using available chromosomal assemblies as drafts. For all species, five large scaffolds corresponding to the number of chromosomal arms (X, 2R, 2L, 3R, 3L) were identified (Table 1). Multiple misassemblies and several chromosome rearrangements were detected and fixed manually. Available physical maps of genomes were used to verify the corrections (George P., et al., 2010; Sharakhova M.V., et al., 2010; Kamali M., et al., 2011; Artemov, G.N., et al., 2017; Artemov, G.N., et al., 2018).

For *An. coluzzii Mopti* we used PEST assembly as a draft which represents a hybrid assembly between *An. gambiae* and *An. coluzzii*. We were able to identify numerous misassemblies taking place at Hi-C heat map (“break points” column in Table 1). Correcting misassemblies requires to split original scaffolds into smaller fragments. The region located near misassembly breakpoint cannot be unambiguously attributed to the split contigs, and thus should be treated as unplaced (debris) portion of the genome. Because of the high number of misassemblies visualized by Hi-C, a substantial portion of its genome was moved to the debris. To provide better assembly of this species, we decided to use recently published PacBio contigs from a single *An. coluzzii* Ngousso mosquito (ref), characterized by N50 equal to 3,47-Mb. This indeed results in more accurate assembly which was used in further analysis (Table 1).

Species	Scaffold length	Reference genome			*De novo* assembly			
		length of chromosomes	length of gaps	N50	length of chromosomes	length of gaps	N50	break points
An. albimanus	170,336,140	167,376,415	3,003,101	1,652,686	169,448,125	10,000	195,377	23
An. atroparvus	224,290,125	200,912,972	1,005,100	1,513,463	216,747,066	116,200	29,376	465
An. coluzzii	251,414,185	-	-	330,848	231,617,039	13,700	899,071	77
An. merus	300,704,392	-	-	36,858	234,366,716	27,100	176,838	17
An. stephensi	221,324,304	-	-	793	196,394,606	62,100	63,099	144

Table 1. *The column "break-points" displays numbers of misjoins corrected manually by redirection or reposition scaffolds.*

Pairwise alignment of the obtained assemblies showed that the length of alignment blocks and percentage of alignable nucleotides correlates with evolutionary distance between species, ranging from 19% to 93% (Fig. 2, A). We found that the vast majority of the rearrangements occur within the same arm, and even for most evolutionary distant species, inter-chromosomal translocations are extremely rare (Fig. 2, B). Comparing individual chromosomes, we found that for all species alignment blocks on X chromosome were smaller than on autosomes (Fig. 2, C), in agreement with previously shown elevated gene shuffling on the X chromosome (ref. Neafsey 2015).

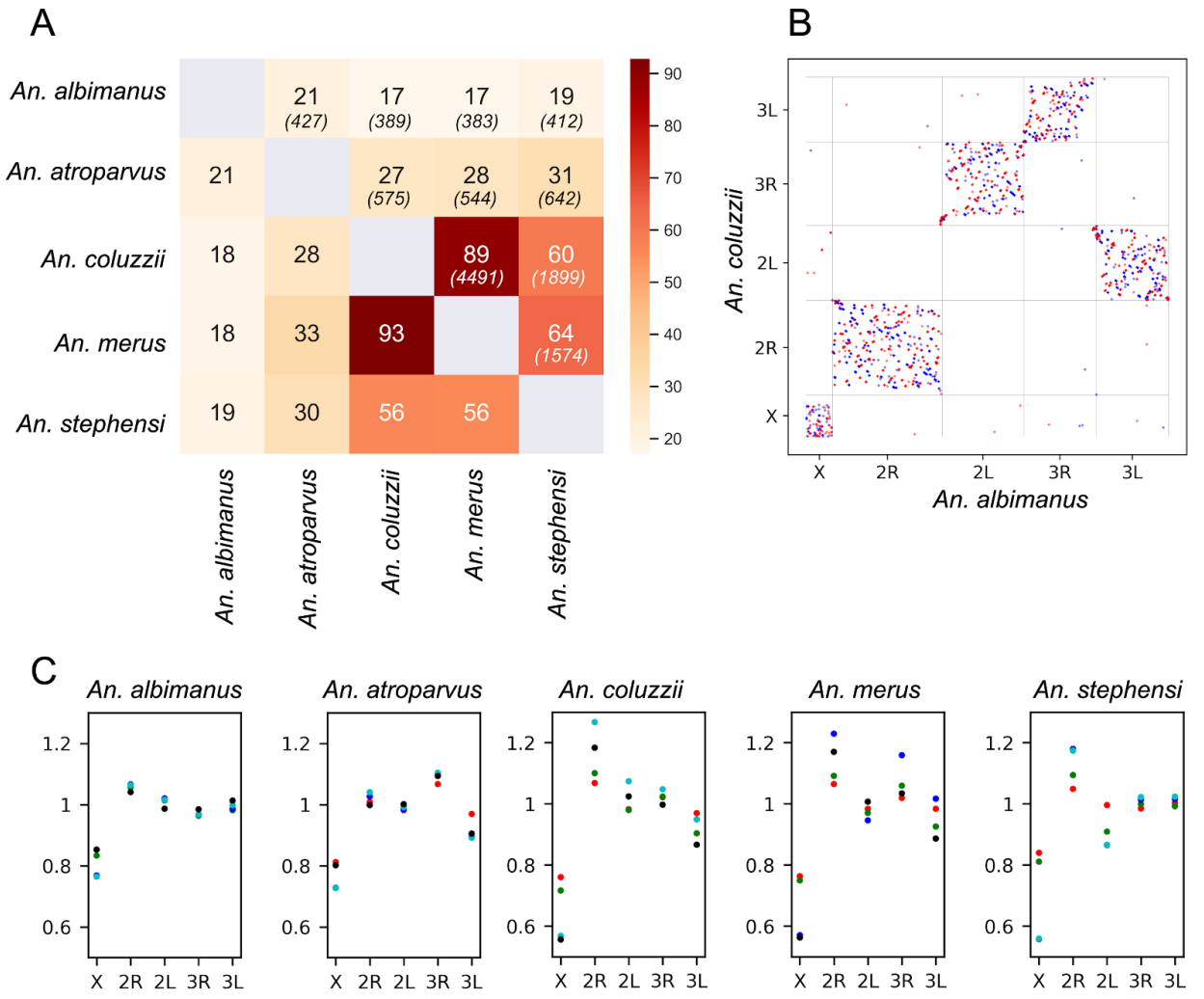

Fig. 2. Comparison of five Anopheles genomes. *A. Heatmap representing the percentage of alignable (remappable) nucleotides and length of synteny blocks (numbers shown in italic) for each pair of genomes. B. Synteny dot-plot showing results of pairwise alignment between An. coluzzii and An. albimanus. C. Comparison of alignment block lengths for each chromosome arm. Y-axes represent the ratio of alignment block length on the specific chromosome (depicted on X-axes) to the genome-wide average. Titles of plots indicate alignment reference and colors of dots correspond to query species. Note that we only used alignment blocks longer than 1000 bp to filter out blocks representing individual exons.*

Overall, Hi-C data allowed us to improve existing genomic assemblies for two *Anopheles* species and generate chromosome-length assemblies for three species *de novo.*

Identification of polymorphic inversions using Hi-C data.

Obtained chromosome-length Hi-C maps of five Anopheles species revealed several regions characterized by “butterfly” contacts pattern, associated previously (ref) with balanced inversions. The most prominent example of such contacts was found on chromosome 2R of *An. stephensi* (Fig. 3, A), where Hi-C map suggests an inversion of a large (~16 MB) chromosomal fragment.

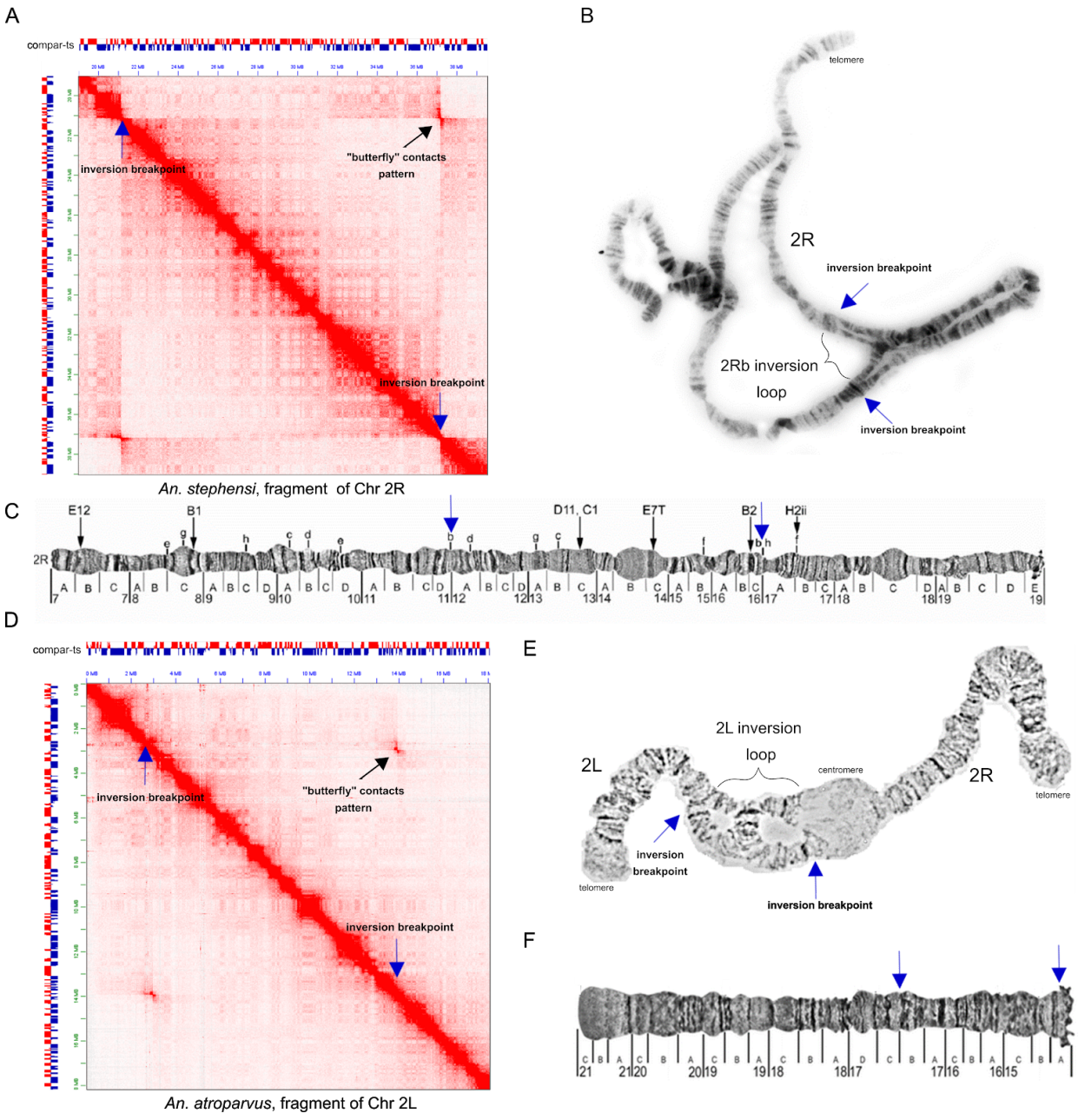

Fig 3. Polymorphic inversions identified using Hi-C data *(inversions break points marked with blue arrows at all images). A. Fragment of An. stephensi Hi-C heat map with 2Rb-polymorphic inversion manifested as "butterfly"* pattern*; B. Light microscope image of An. stephensi polytene chromosomes with 2Rb-inversion loop; C. Comparison with physical map of 2R chromosome of An. stephensi, published previously in Sharakhova M.V., et al., 2010, break points marked with blue arrows; D. Fragment of An. atroparvus Hi-C heat map with 2L-polymorphic inversion; E. Light microscope image of An. atroparvus polytene chromosome 2 with 2L-inversion loop; F. Physical map of 2L chromosome of An. atroparvus, published previously in Artemov G.N., et al., 2017.*

We observed both off-diagonal long-range interactions, as well as interactions near diagonal, suggesting that both variants with inverted and normally arranged sequences were present in the population. Demonstrated paracentric inversion was previously shown for Indian wild-type laboratory strain of An. stephensi (ref) and known as polymorphic 2Rb-inversion (Coluzzi et al. 1973a; Mahmood and Sakai 1984; Gayathri Devi and Shetty 1992; ref). The boundaries of inversion on Chromosome 2R were in agreement with previous cytological data (ref).
Further examination of Anopheles Hi-C heat maps allowed us to identify four inversions, ranging from 2.8 Mb to 16 Mb (the 2Rb inversion) (Supplementary Table 2, Supplementary Figure 1).

3D-chromatin structure of Anopheles genome revealed by Hi-C
Generated Hi-C heat maps visualized five chromosomal elements for all experimental genomes - X, 2R, 2L, 3R, 3L (Fig. 4, A). Strong centromere-centromere clustering was detected as well as inter- and intra-chromosomal telomere-telomere interactions. That evidence represents Rabl-like configuration (ref) in Anopheles genomes where centromeres and telomeres form clusters within a nuclear space (Fig. 4, B). Additionally, we have observed another manifestation of Rabl-like configuration, represented by interactions between chromosomal arms as perpendicular to the main diagonal "wings" pattern on Hi-C heat map (ref -

crops) (Fig. 4, A). Quantitative analysis of contact frequencies confirmed increase of interactions between loci from different chromosome arms located equidistantly from centromere (Supplementary Fig. 2). This increase of interactions was more pronounced in embryonic tissues (1.49-2.45 times) than in adult mosquito data (1.19-1.25 times), suggesting that Rabl-like configuration might be a feature of some but not all cell types. 3D-FISH experiments on *Anopheles stephensi* ovarian tissue confirmed the existence of Rabl-like configuration (centromeric clustering) in follicular cells and the lack of centromeric interactions in nurse cells within polytene chromosomes (Fig. 4, C).

Magnified view on the contact maps revealed large insulated blocks of chromatin located near centromeres (Fig. 4, G-H, Supplementary Figure 3, A-F). The size and location of these blocks were in accordance with the position of pre-centromeric and intercalary heterochromatic blocks on standard cytological maps (Supplementary Figure 4). Outside of the heterochromatic blocks we observed a checkerboard-like pattern of long-range interactions, which was previously shown to represent spatial compartments of the chromatin (Fig. 4, D). Some long-range chromatin contacts were extremely pronounced, corresponding to looping interactions between loci located several Mb away from each other (Fig. 4, D). At the highest resolution (1-25 kb) we observed triangles formed above diagonal, corresponding to the TADs (Fig. 4, E), and chromatin loops (Fig. 4, F).

We next aimed to quantitatively characterize TADs, compartments and loops identified in *Anopheles* genomes.

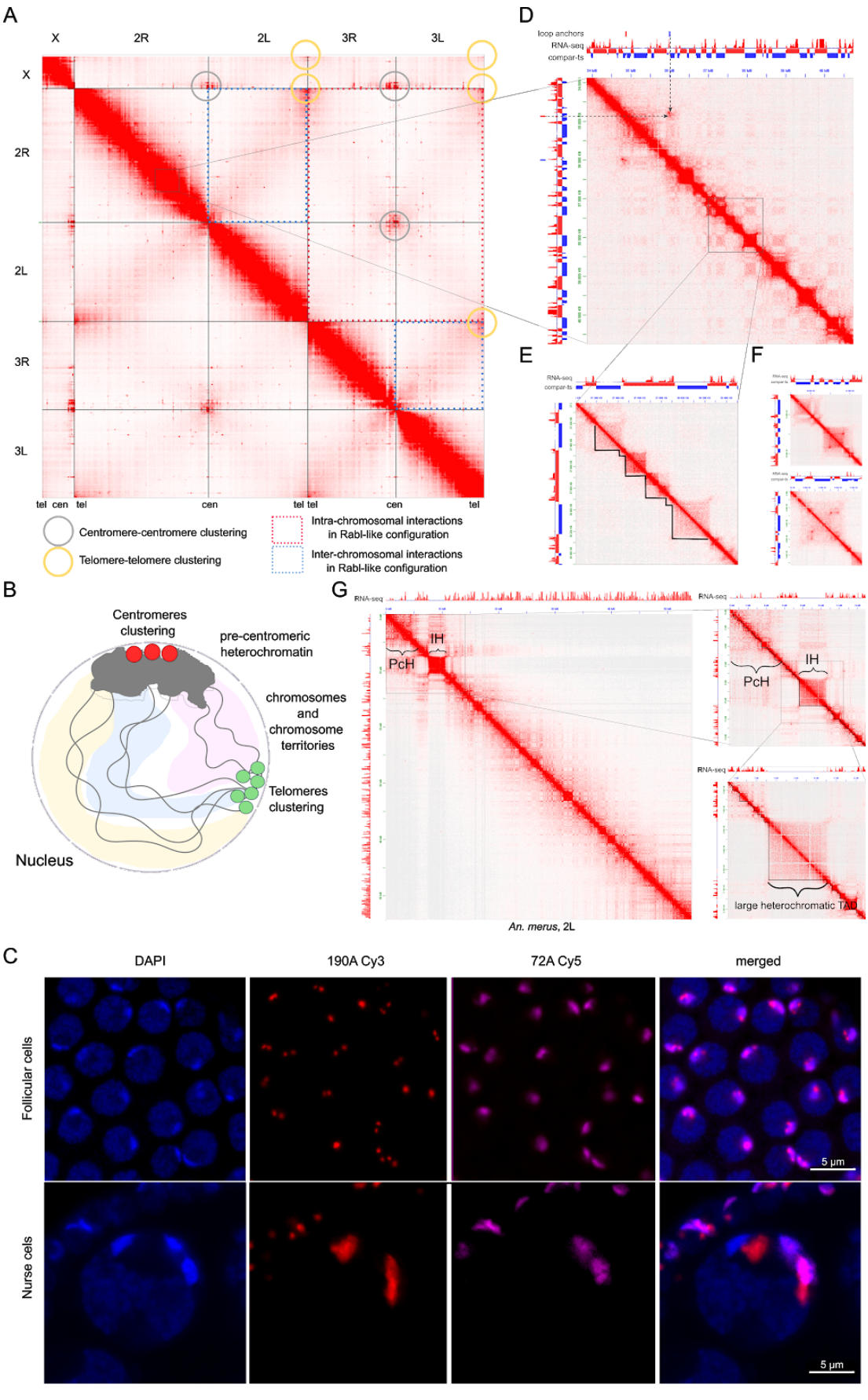
A
X
2R
2L
3R
3L
tel cen tel
cen
tel tel
cen
tel
Centromere-centromere clustering
Telomere-telomere clustering
Intra-chromosomal interactions in Rabl-like configuration
Inter-chromosomal interactions in Rabl-like configuration
D
loop anchors
RNA-seq
compar-ts
E
F
B
Centromeres clustering
pre-centromeric heterochromatin
chromosomes and chromosome territories
Telomeres clustering
Nucleus
G
RNA-seq
IH
PcH
large heterochromatic TAD
An. merus, 2L
C
DAPI
190A Cy3
72A Cy5
merged
Follicular cells
Nurse cels
5 µm
5 µm

Fig. 4. 3D-chromatin structure of Anopheles mosquitoes genome. *A) Hi-C heat map for Anopheles albimanus; B) Chr 2R: 33,900,000-40,700,000; long-distance loop pointed by arrows; compartments C) TADs: Chr 2R: 37,000,000-38,500,000; D) loops; E) Rabl-like configuration; F) FISH: 190A Cy3 probe - autosomal centromere, 72A Cy5 probe - Chr X centromere*

Anopheles genomes are partitioned into compartments

Inspecting genomic interactions on various resolutions, we observed prominent signs of genome compartmentalization, manifested as plaid-pattern of Hi-C contacts. However, when we employed principal component analysis, which is widely used to identify spatial chromatin compartments (ref-s), we found that for the majority of chromosomes the first principal component (PC1) of the Hi-C matrix does not reflect observed plaid-pattern (Fig. 5, A-B; Supplementary Fig. 5. A-I). In most cases, chromosomal arms were split into two or three large, contiguous fragments characterized by similar PC1 values. Thus, PC1 values reflect the position of locus along the centromere-telomere axis rather than the plaid-pattern of contacts.

This discordance between compartmental interactions and PC1-values could be explained by Rabl-configuration of the chromosomes and/or clustering of large heterochromatic blocks in centromeric and/or telomeric regions. First, clustering of centromeric and/or telomeric chromatin may dominate the clustering of compartments. This explains those cases when centromeric and telomeric regions display similar PC1 values, whereas the rest of the chromosomal arms display opposite PC1 values (Supplementary Fig. 5, A-I). Second, it was shown in *Drosophila* cells that Rabl-like configuration results in more elongated shapes of the chromosomal territories (ref), preventing clustering of actively expressed genes located far from each other on the centromere-telomere axis. Supporting this hypothesis, distributions of PC1 values observed in Anopheles mosquitoes data were very similar to the distributions obtained from the data from barley cells (Fig. 5, C), which also displays Rabl-configuration of chromosomes (ref). Moreover, for adult *An. merus* samples, where signs of

Rabl-like configuration of chromosomes were much less pronounced, standard PC1 algorithm was able to define compartments that agree with plaid-pattern on all chromosomes (Supplementary Fig. 5, K-M). Thus, we suggested that Rabl-configuration of chromosomes attenuates long-range interactions between compartments.

We developed an approach for robust identification of compartments on chromosomes with Rabl-configuration. Briefly, our approach relies on normalization of Pearson's correlation values within small blocks of the data matrix before PC1 calculation (see Supplementary Note I for details and comparison of different approaches). We called this new algorithm "contrast enhancement" because the scaling of Pearson's correlation within small blocks resembles enhancement of image contrast (Supplementary Fig. 6). We refer to the obtained PC1 values as cePC1 (contrast-enhanced PC1 values). As could be seen from Fig. 5 B, using contrast enhancement we were able to identify compartments that correspond well to the plaid-pattern observed on the Hi-C maps.

To understand epigenetic mechanisms underlying *Anopheles* chromatin compartments, we performed RNA-seq on the same embryonic stages as used for Hi-C libraries preparation. We found that compartments show moderate correlation with expression levels, slightly lower correlation with the gene density and only weakly correlate with GC-content (Fig. 5, D; Supplementary Table 4). Correlations were significantly higher for cePC1, obtained using contrast enhancement, whereas for original PC1 values we found almost no correlation with the aforementioned epigenetic features (Supplementary Table 4). Based on cePC1 values, we split genome into two non-overlapping A- and B-compartments so that loci belonging to A-compartment show higher gene density, gene expression, and GC-content.

Overall, we have shown that Rabl-configuration of chromosomes attenuates compartmental interactions. We developed a new computational approach to detect compartments and showed that spatial compartmentalization distinguishes active (gene dense, expressed GC-rich) and inactive (gene-poor, silent, GC-pour) chromatin.

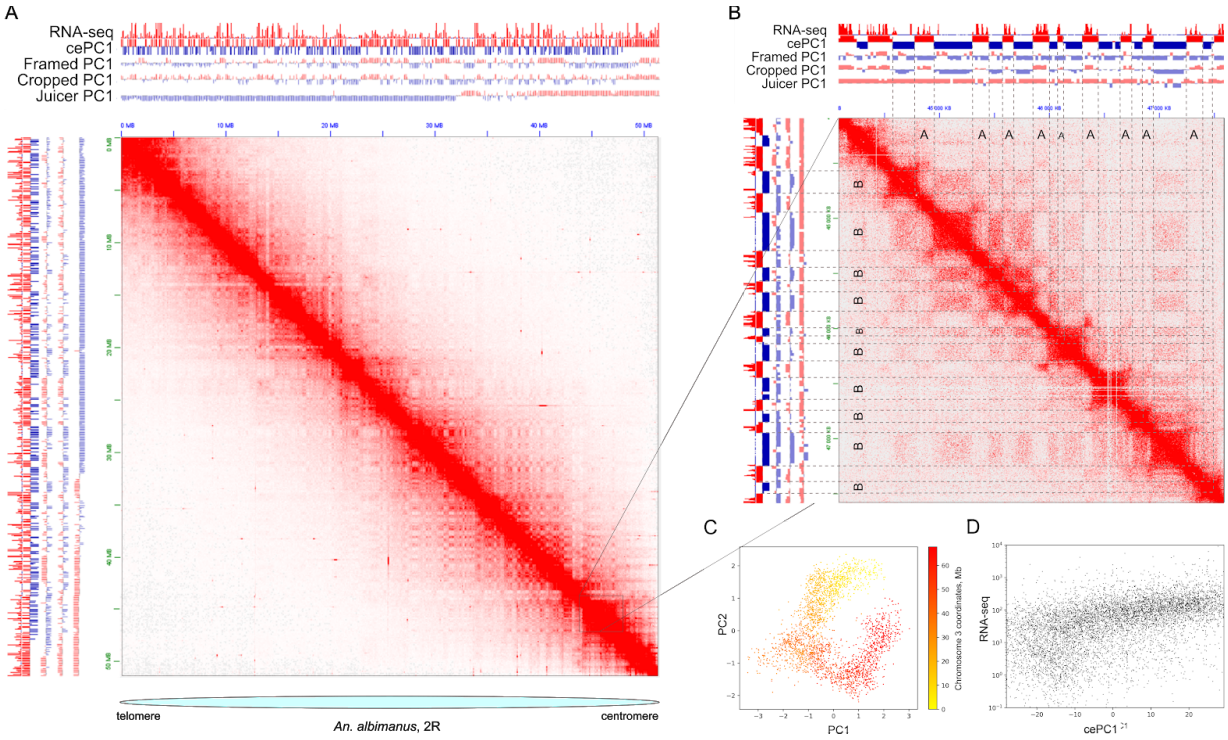

Fig.5. Anopheles genomes partitioning into compartments. *A. Compartmentalization demonstrated for Chr 2R, An. albimanus; B. Region zoomed-in from Fig. 5,A: correspondence between RNA-seq data and PC1 values, obtained with different algorithms; C. distributions of PC1 values within the length of Chr 3 observed in An. albimanus; D. cePC1 correlation with RNA-seq data.*

Transitions between compartments correspond to TAD boundaries in *Anopheles* genomes.

In addition to the plaid pattern, Hi-C maps display triangles above the main diagonal (TADs). For each species, we called TADs at 5-kb resolution (Methods, Fig. 6, A, B). TADs range in size from 15 to 650-kb, with a median size of ~135-kb (Supplementary Table 5). Distribution of TAD sizes was similar in all chromosomes (Supplementary Fig. 7, A)(Fig. 6, C); however, within each chromosome, smaller TADs tend to co-localize with A-compartment (euchromatic regions), whereas longer TADs belong to B-compartment (heterochromatin) (Fig. 6, D; Supplementary Fig. 7, B). As evident from the analysis of cePC1-values, TADs longer than 400-kb are

almost exclusively located in gene-poor, heterochromatic regions with low level of gene expression. These TADs do not form long-range interactions, typical for smaller TADs. Comparing positions of large TADs with cytological maps of *Anopheles* polythene chromosomes, we found that these structures often correspond to intercalary heterochromatin blocks (Supplementary Figure 4, Fig. 6, E) demonstrating reduction in RNA-seq signal.

We next analyzed TAD boundaries and found that in the majority of cases the boundaries coincide with the transitions between compartments (Fig. 6, F, Supplementary Fig. 7, C), and magnitude of cePC1 changes was higher around TAD boundaries than within TADs (p-value < 10E-30 for all species, Supplementary Fig. 7). We additionally inspected all cases when the strong TAD boundary (top quartile of insulatory scores distribution) separates regions with similar cePC1 values (cePC1 difference below bottom quartile of the distribution). We confirmed that all these cases were due to the errors in TADs or compartment calling (Supplementary Fig. 7, D).

Thus, we concluded that both TADs and compartments represent similar chromatin features; local insulation from the neighboring genomic regions is captured by TAD-callers, whereas preferences between long-range interactions of these locally insulated genomic regions are captured by cePC1. In-chromosome puffs, representing intercalary heterochromatic regions, correspond to the large TADs depleted for long-range interactions.

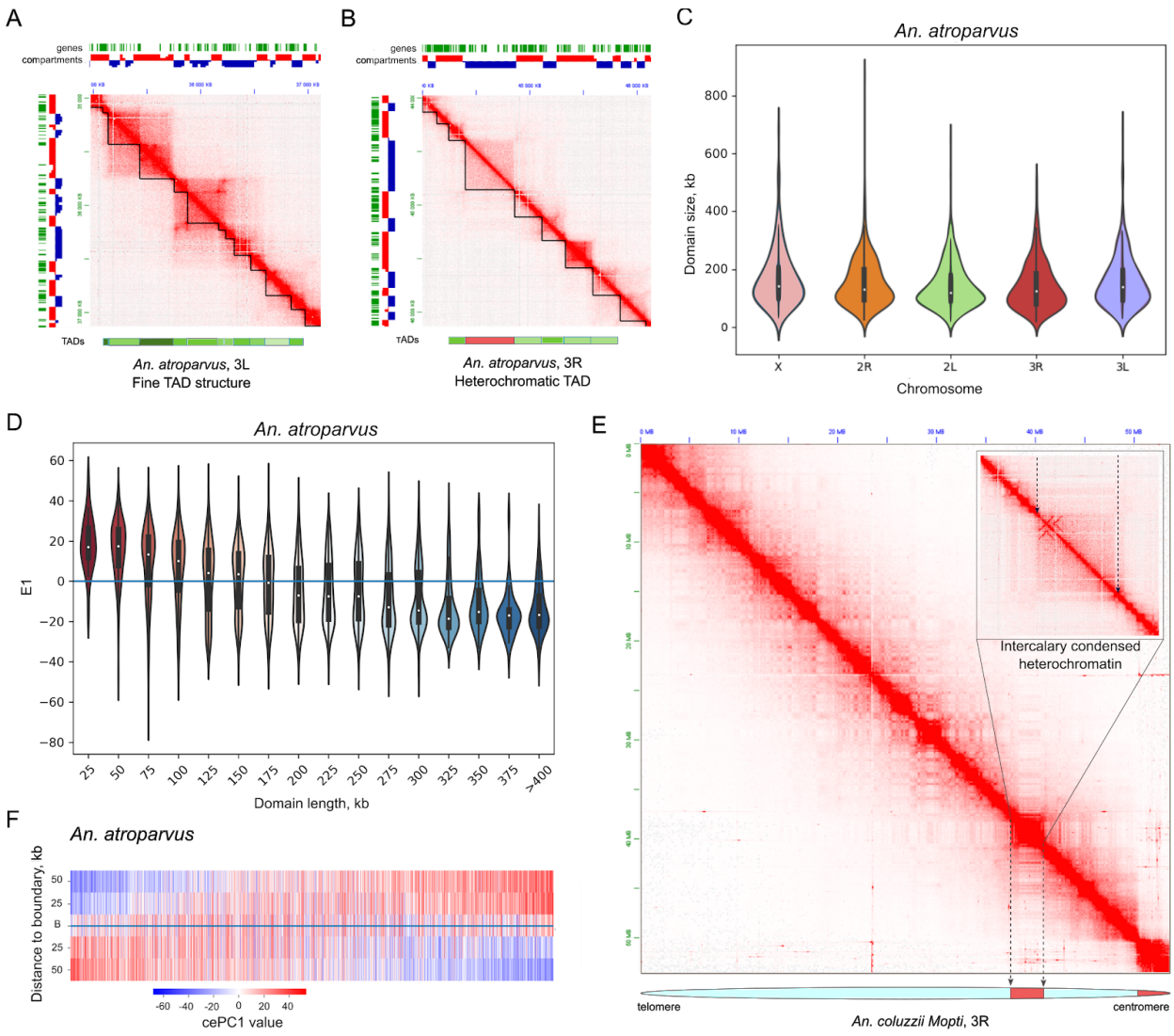

Fig. 6. TADs on Anopheles Hi-C heat maps. *A. Typical TAD structure; B. heterochromatic TAD; C. TADs size distribution within the chromosomes; D. Violin-plot illustrating distribution of* cePC1-values for different *TAD lengths; E. Intercalary heterochromatic block visualized on Hi-C map of Chr 3R Anopheles coluzzii Mopti; F. Heatmap of cePC1 values around TAD boundaries. Each line represents one TAD boundary.*

Long-range interactions identified on Hi-C maps and loop validation by FISH

Examination of Hi-C data revealed looping interactions between specific loci (Fig. 4, D, F). Many of these interactions occurred within the same

TAD, and some anchors formed networks of interactions. As chromatin loops were previously associated with Polycomb (ref), we performed H3K27me3 ChIP-seq on *An. atroparvus* and found that indeed anchors of some loops are located within H3K27me3-enriched regions (Supplementary Fig.8, A-H).

Whereas the vast majority of loops spanned genomic distance less than one Mb, we found notable examples of extremely long-distant loops, connecting loci separated by up to 31-Mb in the linear genome (Fig. 7, A-E). We found 2-6 of such giant loops in each species (Supplementary Table 3), and focused on two of them, one located on chromosome X (X-loop) (Fig. 7, A-E), and another on an autosome (A-loop) (Supplementary Figure 7), which were, according to synteny analysis provided below, present in all examined species. Anchors of these loops were represented by large (~200-300-kb) loci, which interact significantly more than expected taking their genomic distance (Fig. 7, B = **graph expected)**. Moreover, within these broad anchors, we often observed two pairs of smaller loci (~25-kb) showing peaking interaction frequencies (Fig. 7, A).

Interestingly, anchors of X-loop do not interact with anchors of A-loop above expected level (**graph expected between anchors)**, and both X- and A-anchors do not interact with anchors of smaller loops located nearby.

To understand whether loops identified in different *Anopheles* species are homologous to each other, we performed whole-genome alignment to identify conserved elements (CEs) in the *Anopheles* genomes, and used conserved elements located within loop anchors to remap anchor positions between species. Obtained results showed that loops are formed between homologous (synteny) regions (Fig. 7, A-G; Supplementary Fig. 9, A-F).

However, none of the conserved elements was located in the peak of the interaction in all species (Fig, 7, A; Supplementary Fig. 9), indicating that identified CEs could not by semself explain formation of loops.

We next analyzed gene expression and H3K27me3 profiles within loop anchors. Several genes were located within anchors of all examined species (Supplementary Table 6), and some of them were expressed,

although not all anchors contain actively expressed genes (Fig. 7, F-G, Supplementary Fig. 9, F) and expression level was moderate. Samely, H3K27me3 signal shows only weak enrichment within X- and A-loop anchors, much below the level observed in Polycomb loops (Fig. 7, E-G, Supplementary Table 7, Supplementary Fig. 10).

We found that A- and X-loops are present in both adult and embryonic data available for *An. merus*, indicating that these loops are not developmental-stage specific (Supplementary Fig. 11).

Finally, we performed 2D and 3D FISH to validate the interactions between the putative loop anchors in nuclei of follicle cells and ovarian nurse cells of *An. coluzzii*, *An. stephensi*, and *An. atroparvus* (Fig. 7, H). In all tested cases, we found no colocalization of genes at putative anchors in highly polytenized chromosomes of nurse cells by both 2D and 3D FISH. For the *An. coluzzii* 7.5 Mb X chromosome loop (7,380,000-7,610,000 – 14,950,000-15,400,000), we found 100% colocalization of genes at anchors in follicle cells by 2D FISH and variable colocalization by 3D FISH. For the *An. stephensi* 5.5-Mb 2R loop (37,355,000-37,525,000 – 42,920,000-43,025,000) no colocalization in follicle cells was identified by 3D FISH, but the anchors colocalized with high frequency (5 out of 7) in ovarian nurse cells with low-level polytene chromosomes.

Species with loop	Chr	Gen.coordinates of the 1st anchor	Gen. coordinates of the 2nd anchor	Chr	Gen.coordinates of the 1st anchor	Gen. coordinates of the 2nd anchor
An.coluzzii	X	7,370,000-7,610,000	14,950,000-15,400,000	2R	41,330,000-41,445,000	43,045,000-43,225,000
An.merus	X	7,150,000-7,385,000	14,610,000-14,900,00	2R	41,670,000-41,840,000	43,635,000-43,945,000
An.stephensi	X	9,440,000-9,730,000	14,670,000-15,080,000	2R	37,355,000-37,525,000	42,920,000-43,025,000
An.atroparvus	X	625,000-780,000	10,015,000-10,490,000	3R	7,670,000-8,100,000	38,625,000- 38,770,000
An.albimanus	X	4,600,000-4,665,000	6,615,000-6,670,000	2R	34,775,000-34,875,000	35,965,000-36,035,000

Table 2. Long-distance conservative loops coordinates in Anopheles species.

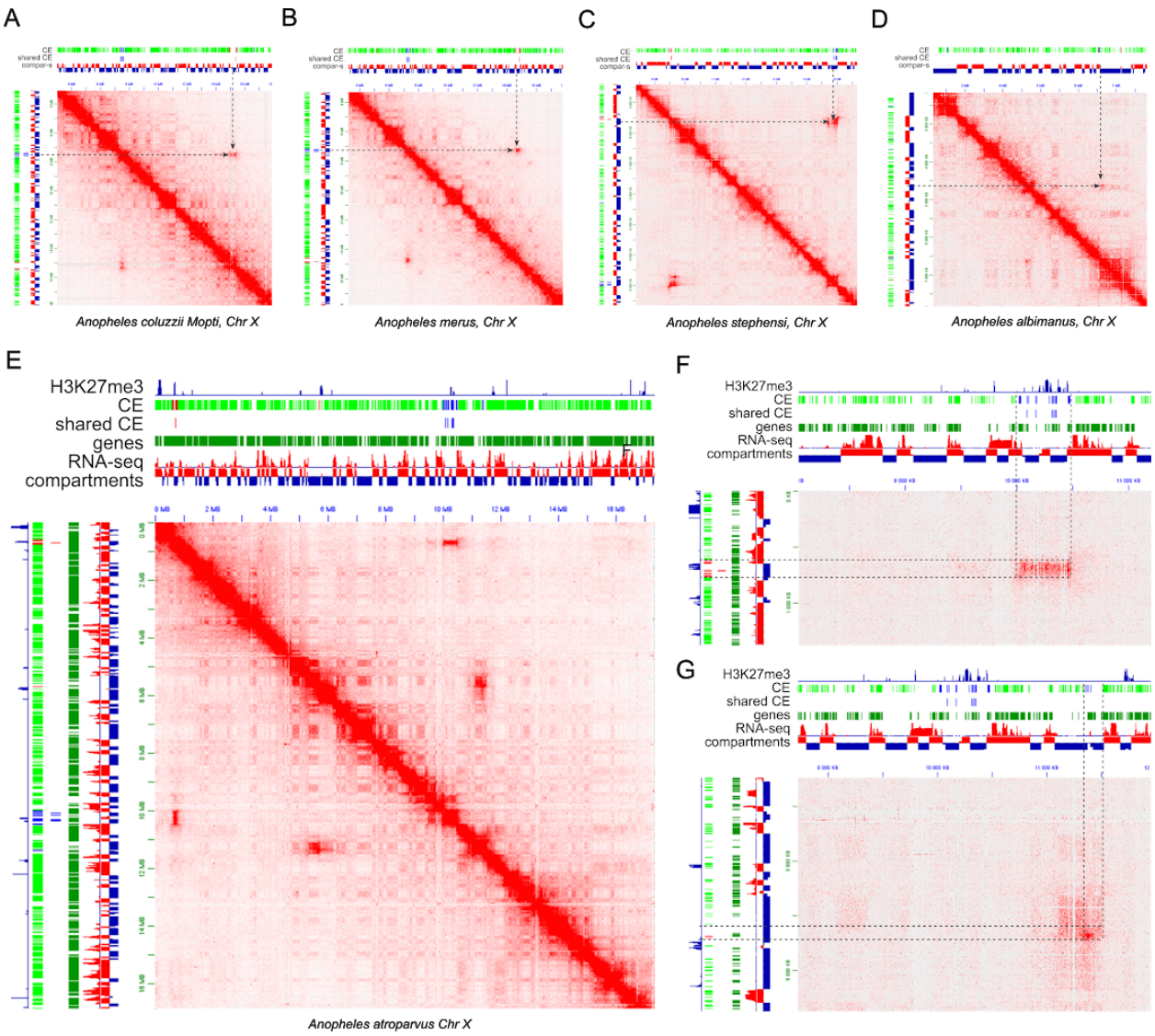
A
B
C
D
Anopheles coluzzii Mopti, Chr X
Anopheles merus, Chr X
Anopheles stephensi, Chr X
Anopheles albimanus, Chr X
E
H3K27me3
CE
shared CE
genes
RNA-seq
compartments
Anopheles atroparvus Chr X
F
G
H
DAPI
X-loop_anchor 1
X-loop_anchor 2
merged
Folicullar cells
Nurse cell
5 µm

Figure 7. Long-range interactions identified on Hi-C maps and loop validation by FISH. *A. X conservative loop in An. coluzzii; B. X conservative loop in An. merus; C. X conservative loop in An. atroparvus; D. X conservative loop in An. albimanus; E. X conservative loop in An. stephensi; F. X conservative loop in An. stephensi, zoomed-in; G. X conservative loop, verification by 3D FISH in follicular cells and nurse cells.*

We tested three loops of different sizes in *An. atroparvus*: 31-Mb A-loop on arm 3R: 7,670,000-8,100,000 – 38,625,000- 38,770,000, 6-Mb X-loop on chromosome X: 5,320,000-5,755,000 – 11,100,000-11,510,000 and 12-Mb loop on arm 2R: 11,210,000-11,700,000 – 23,055,000-23,700,000. The anchors are always colocalized in follicle cells in 2D FISH experiments. However, colocalization in 3D FISH experiments was variable in follicle cells: from no colocalization for the large 31-Mb 3R loop, to partial colocalization for the medium 12-Mb 2R loop, to 100% colocalization for the small 6-Mb X chromosome loop. Colocalization for the 6-Mb X chromosome loop was also confirmed by 3D FISH in low-level polytenized chromosomes of nurse cells.

Overall, we validated most of the long-range interactions detected by Hi-C using FISH, although colocalization of loop anchors was found only in a subset of examined cell types. Obtained data showed that *Anopheles* chromatin forms several extremely long-range, locus-specific contacts, which are evolutionarily conserved for ~80 millions of years and underlying by currently unknown molecular mechanisms independent of active transcription or Polycomb-group proteins.

Spatial contacts of the chromatin in insects and vertebrates are constrained in a genome-size-dependent manner.

To study general principles of genome organization in *Anopheles* we analyzed how chromatin contacts probability (P) scales with genomic separation s, P(s). As reported in previous studies, contact frequencies decay rapidly as genomic distance increases (Fig. 8, A). Since it was shown previously that P(s) follows a power law, we characterized the exponent by computing slope of the decay curve in log-log coordinates

(Fig. 8, B). We found that decay speed is not uniform, and could be described by two different decay phases.

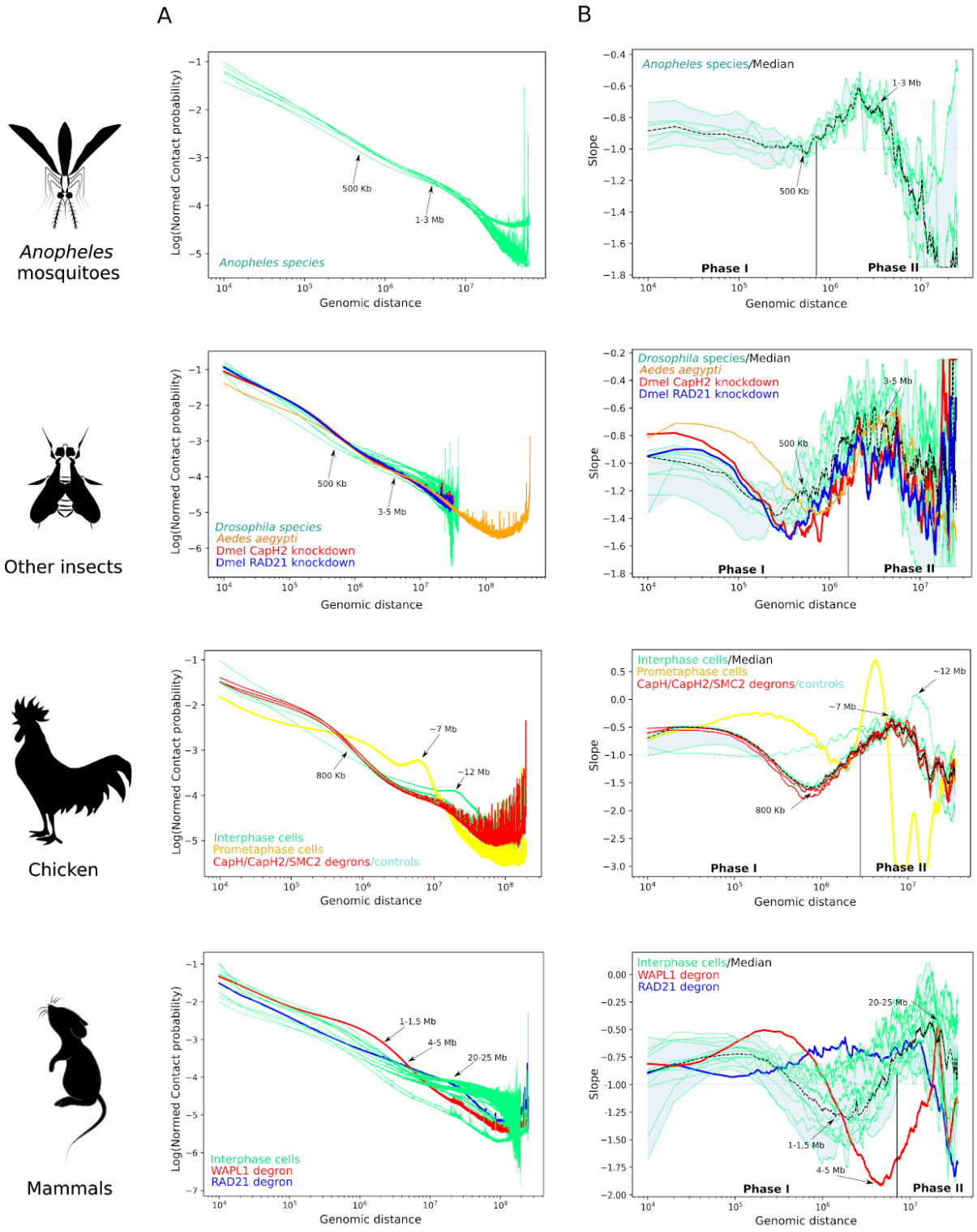

Fig. 8. Contact frequency decays non-uniformly with genomic distance. *A - dependence of contact probability on genomic distance for Anopheles, other insects, chicken and mammals shown in log-log coordinates. B - slope of the curve depicted on A as a function of genomic distance. Black dashed line shows the median, and the blue area represents minimal and maximal values for several species. Slope-plots for individual species are shown as thin green lines.*

The first phase characterized by a U-shaped slope curve occurs between 10-kb and 1-Mb. For distances between 10-kb and ~200-500-kb slope decreases modestly starting from ~-0.8 and reaching the lowest point slightly below -1, which is the characteristic value for the fractal globule, and then starts increasing. At ~500-kb starts phase II, where the slope rapidly increases reaching values around -0.6 for genomic distances ~1-3-Mb and then falls sharply to the values below -1.

These two phases were observed in all *Anopheles* species (Supplementary Fig. 12), and analysis of *Drosophila* and *Aedes* Hi-C datasets demonstrated very similar patterns (Fig. 8, B, Supplementary Fig. 12, A). We next analyzed vertebrates Hi-C data, and again found specific shape of the slope curve: the phase I, characterized by prominent decrease of the slope, was much more prolonged than in insects, and reaches minimum at genomic distances ~800-kb for chicken and ~1-1,5-Mb for mammals; the phase II display maximum at ~7 Mb for chicken and ~20-25 Mb for mammals, and after this point we observed sharp drop of the slope similarly to the insect's curve.

We used available Hi-C datasets obtained from cells lacking cohesin subunit RAD21, cohesin release factor WAPBL or cohesin loader NIPBL to confirm that U-shape of the slope curve observed in the phase I reflects formation of TADs, and that genomic distance characterized by the minimal slope value corresponds to the characteristic TAD length (see supplementary note II and associated supplementary figures). This allows us to provide estimation of TADs length without biases introduced by TAD-calling algorithms. The characteristic TAD length in *Anopheles* genome is around 200-400-kb, which is very close to *Drosophila* species, and slightly smaller than *Aedes* TADs (500-800-kb).

We next aimed to explain changes of the slope during the phase II, namely the abrupt drop of contact probabilities observed at genomic distance ~3-Mb for insects, ~7-Mb for chicken and ~20-25-Mb for mammals. We confirmed that this decrease of the slope was seen in multiple experiments on unsynchronized cells, G2-phase synchronized chicken DT40 cells, G1-phase quiescent chicken erythroblasts (Suppl. Fig. 12, B) and S-phase synchronized *Drosophila* S2 cells or embryonic cells (Suppl. Fig. 12, C) cells, indicating that there is no cell type or cell cycle specificity and that observed results couldn't be explained by fraction of mitotic cells. The only one cell type where we didn't observe the drop of the slope was mammalian post-mitotic neurons (Supplementary Fig. 12, D), suggesting that progression through mitosis might be essential. Moreover, cohesin- or condensin-mediated loop-extrusion process does not influence this slope drop, as we clearly observed it under RAD21-degron conditions, NIPBL-degron conditions, both CapH-, CapH2- and SMC2-degron conditions and in chicken erythrocytes which do not have extrusion-mediated TADs (Fig. 8 and Supplementary Fig. 12, A-D). However, we should note that cells lacking condensin subunits pass through mitosis before inducing degradation (ref). Thus, our analysis shows that condensins are not required to maintain constraints underlying slope shape, but we can not exclude that they are required to form these constraints.

We next speculate about possible constraints underlying observed drop of the slope. We suggested that the decrease of contact frequencies observed at large distances might be explained by the specific shape of chromosome territory. Indeed, it is known that chromosome territories are not spherical (DOI: 10.1007/s10577-007-1172-8, 10.1007/s00412-014-0480-y). If DNA fills an ellipsoidal or cylindrical shape, then the radius of the cylinder will determine the characteristic spherical volume inside which the drop of contacts should be in agreement with the fractal globule model. Upon exiting this volume, loci are located in different slices of the cylinder, which makes the likelihood of contacts between them smaller than expected for the fractal globule model.

Following this hypothesis, our data suggest that the interphase chromosomes of all animals fill elongated volumes, characterized by the specific minor radius, different for insects, chicken, and mammals. We estimated that the size of a spherical globule of DNA that fits in this volume is ~1-3-Mb for insects, ~7-Mb for chicken, and ~20-25-Mb for mammals.

Discussion

Here we used Hi-C data to comprehensively characterize 3-dimensional genome organization in five *Anopheles* species. We find that several properties of chromatin organization previously identified in *Drosophila* are shared by *Anopheles* species. In particular, *Anopheles* data suggests that the genome is divided into active and inactive compartments, which correlate with gene expression, and that loci belonging to different compartments are insulated from each other by TAD boundaries.

At the same time, we found some features of genome organization which were not previously described in *Drosophila* or other insect data.

Evolutionary conserved long-range chromatin interactions in Anopheles genomes.

We found that certain genomic regions form giant loops spanning dozens megabases. Loops described previously in *Drosophila* Hi-C data are typically smaller in size and mainly attributed to interactions between active genes or Polycomb proteins. Our data suggest that loops found in *Anopheles* are formed by other, yet unknown, mechanisms.

FISH analysis showed cell-by-cell variability of contacts between loop anchors, suggesting that 100% formation of the loops is not essential for the function of follicle cells. If such variation exists in embryos then the Hi-C method is sensitive enough to identify the interaction in the assembly of cells. In the case of ovarian nurse cells, we found that colocalization correlates with the low-levels of chromosome polyteny. We conclude that the high polyteny likely creates specific mechanical properties of chromosomes that hinder specific interaction between the loop anchors. An alternative explanation is that highly-polytenized chromosomes do not require such interactions. Interestingly, no specific long-range interactions

were found in polytene chromosomes of D. melanogaster by microscopic analysis (Hochstrasser, M.; Sedat, J.W., 1987) and Hi-C (Eagen et al 2015 Cell). However, specific distant chromosomal contacts have been detected in Hi-C heatmaps of Drosophila embryos (Sexton, T et al 2012 Cell).

Chromosome territories shape might constraint long-range contacts.
We have shown that long-range contacts are constrained in genomes of all examined animal species, and suggested that chromosome territory shape might underline these constraints. It is not known what determines the non-spherical shape of chromosomes. We confirmed that the drop of the slope is present after CapH2-knockdown in *Drosophila* cells, as well as in chicken cells with CapH, CapH2 or Smc2 proteins degraded, thus condensin proteins probably do not play an essential role in the maintenance cylindrical chromosome territories shapes during interphase. As was mentioned above, RAD21 degradation also does not show any effect on slope changes at large genomic distances, arguing against the role of cohesin in this process. Interestingly, the only cell type where we haven't observed a drop of the slope were quiescent mammalian neurons (Supplementary Fig. 12, D). This allows us to speculate, that chromosomes acquire elongated form during mitosis, and this elongated form might be preserved during the cell cycle. When cells exist cycle and stay quiescent for long enough time, diffusion processes intermix chromatin allowing contacts between more far loci. Similar hypotheses were recently proposed to explain the difference between slope changes in mammalian oocytes, which are arrested at the prophase stage for months and sperm cells, which are not subjected to such prolonged mitotic arrest.

Rabl-like chromosome configuration dominates compartmental interactions.
Elongated chromosome territory shape reduces the frequency of interactions between loci more than several Mb away from each other. At the same time, we observed the clustering of centromeric and telomeric heterochromatin. According to principal component analysis, these factors dominate genomic compartmentalization, although compartmental

interactions are present and well-pronounced at mid-range distances. We proposed new approach for the identification of compartments in genomes with Rabl-configuration and successfully employed it to call compartments in *Anopheles* genomes.

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its Additional files. The raw sequencing data for five Anopheles species have been deposited in the NCBI SRA database with accession numbers PRJNA615788 (RNA-seq raw reads), PRJNA623252 (ChIP-seq raw reads), and PRJNA615337 (Hi-C raw reads). Processed data, including genome assemblies, Hi-C heatmaps, RNA-seq and ChIP-seq tracks, finalized TADs and compartments are available at https://genedev.bionet.nsc.ru/Anopheles.html

Materials and methods

Mosquito colony maintenance:

Laboratory colonies of the following strains were used for the experiments: MOPTI strain for *Anopheles gambiae* (MRA-763); MAF strain for *Anopheles merus* (MRA-1156); STE2 strain for *Anopheles stephensi* (MRA-128); EBRO strain of *Anopheles atroparvus* (MRA-493); STECLA for *Anopheles albimanus* (MRA-126). All mosquito strains were initially obtained through Malaria Research and Reference Reagent Resource Center (MR4) stocks and BEI Resources, NIAID, NIH and maintained in the insectary of the Fralin Life Science Institute at Virginia Polytechnic Institute and State University. Mosquito specimens were hatched from eggs in unsalted water and incubated for 10-15 days undergoing four larvae and pupae developmental stages at 27C. Adult mosquitoes were maintained in the incubator at 27 °C, 75% humidity, with a 12h cycle of light and darkness. 5-7d adult mosquitoes were blood-fed on defibrinated sheep blood using artificial bloodfeeders. Approximately 48-72h post-blood feeding, the egg dishes (Suppl. Fig Y) were placed and after 15-18h embryos were collected for further experiments.

In situ Hi-C on mosquito embryos

The step-by-step mosquito Hi-C protocol can be found in Supplementary Protocol 1. In brief, procedure for mosquito embryos was modified based on the previously published high-resolution 3C protocol (Comet, I., et al., 2011), Drosophila Hi-C protocol (Sexton, T., et al., 2012), and in situ Hi-C protocol for mammalian cells (Rao, Suhas S.P., et al., 2014). Mosquito eggs were collected (optimized egg dish can be found in Supplementary Fig. 13) and Hi-C libraries were prepared using nuclei isolated from ~1000-3000 embryos of mixed sexes. Embryos were fixed with 3% formaldehyde at the developmental stage of 15-18 hours after oviposition. MboI restriction enzyme (NEB, #R0147) with average restriction fragment size ~250-300 bp) was used in the experiment. Two biological replicates of Hi-C libraries were generated, prepared with

NEBNext® Ultra™ II DNA Library Prep Kit for Illumina (NEB, #E7103), and sequenced using 150-bp pair-ended sequencing on Illumina platform. 3D-DNA pipeline was employed to assemble the genomes de novo using the generated Hi-C data set. Misassemblies were identified and fixed manually using assembling mode in Juicebox software (Durand, N.C., et al., 2016). The physical genome maps were used to assess the assemblies (George P., et al., 2010; Sharakhova M.V., et al., 2010; Kamali M., et al., 2011; Artemov, G.N., et al., 2017; Artemov, G.N., et al., 2018).

ChIP-seq

The anti-trimethyl-histone H3 (Lys27) (Millipore #07-449) antibody was used for ChIP seq experiment. ~1000 eggs were bleached and fixed according to … protocol. 8-10 cycles of 10/10sec ON/OFF in lysis buffer with SDS and NP40 were performed on Bioruptor Diagenode machine for chromatin sonication. The detailed protocol can be found in Supplementary Protocol 2.

RNA-seq

1500-2000 embryos of 15-18h Anopheles mosquitoes were bleached for 5 minutes. Total RNA was extracted following the Monarch Total RNA Miniprep Kit (NEB #T2010S) protocol with minor modifications. Mosquito embryos were homogenized in 800 µL of lysis buffer with 2 mL Dounce homogenizer. Samples were incubated at RT for 10 minutes, then proteinase K was added and samples were incubated for an additional 5 minutes at 55C. After that, the tubes were centrifuged at max speed for 2 minutes. Supernatant was transferred to fresh RNAse/DNAse-free tubes and proceeded with gDNA removal columns. The incubation time with DNAse was increased to 20 minutes in total. Total RNA was eluted with 50 µL H2O. Sample concentration was verified with Nanodrop and 1µg of total RNA was used for the next procedures. Samples were prepared for Illumina sequencing with NEBNext® Ultra™ II RNA Library Prep Kit for Illumina (NEB #E7775) accompanied by NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB #E7490) with RNA insert size of 200 bp.

Ovary preservation and polytene chromosome preparation

To prepare high-quality polytene chromosome slides we followed the protocol described in details in … with minor exceptions. Approximately 24-30h after the second or third blood feeding (the timeline for Christopher's III developmental stage varies for different species and should be estimated by visual inspection), ovaries were fixed in Carnoy's solution (3:1, ethanol: glacial acetic acid by volume), kept at RT for 24h, then stored at -20C for a long term.

At least one week after fixation, ovaries were dissected in Carnoy' solution. Cleaned and separated follicles were then placed on a slide in a drop of 50% propionic acid for ~5 min (5-10 follicles per slide), where they were macerated and squashed in fresh portion of 50% propionic acid. The quality was briefly checked with light. microscope and good-quality preparations were proceeded with flash-freezing in liquid nitrogen. After freezing the slides were immediately placed in pre-chilled 50% ethanol and kept at -20 overnight. Next day, after removing coverslips, preparations were dehydrated in ethanol series (50, 70, and 96%), air-dried, and the quality of polytene chromosomes was checked. High-quality slides were placed in a cardboard holder and stored at RT up to 3 months.

2D-FISH

Probes were prepared by Random Primer Labeling method described in Protocols for Cytogenetic Mapping of Arthropod Genomes (ref). gDNA was freshly extracted using Monach Genomic DNA purification kit, PCR product amplification was performed by regular PCR with DreamTaq/Q5 polymerase. PCR Product was purified with Qagen purification columns and ~200-250 ng were used for 25 µL labeling reaction. After overnight incubation in thermocycler at 37C, labeled probes were precipitated with 96% ethanol, dried and dissolved in 30 µL hybridization buffer. 10-15µL of one probe were applied to slide. Prepared slides were incubated for 25 minutes at 70C in humid chamber following the overnight incubation at 39C. Washing steps included 2 times wash in 1xSSC at 39 for 20 minutes, 1 time wash in 1xSSC at RT for 20 minutes, 1 time wash in 1xPBS at RT for 10 minutes. One drop of ProLong™ Gold Antifade Mountant with DAPI (ThermoFisher P36931) was added to the slide and signal detection was performed with fluorescent Olympus microscope. Set of primers for PCR product amplification can be found in Supplementary Table 8. All PCR primers were designed against unique exon regions.

3D-FISH

Fluorescent probes were prepared by the same method as for 2D-FISH. Probe pellets were dried and resuspended in 20-30 hybridization buffer. 20µL of one probe were used per one experiment where the total volume of hybridization probe solution contained at least 80µL. Probes were denatured in Thermomixer at 90C for 5 minutes, then transferred to 39C and pre-annealed for at least 30 minutes.

24-30h after blood feeding ovaries were dissected in 1xPBS and placed in 1mL 1xPBST. Fixation was performed in 4% paraformaldehyde solution for 20 minutes with rotation. Then, ovaries were washed in PBS and treated with 0.2µg/µL RNAse solution in PBS at 37C for 20 minutes. After rinsing in PBS, ovaries were placed in 1%

Triton-X-100/0.1M HCl solution and incubated at RT for 20 minutes with rotation. After brief washing, DNA denaturation was performed in 50% formamide/2xSSC solution at 75C for exactly 30 minutes. Then, ovaries were rinsed in 100µL of hybridization buffer which was replaced with hybridization mix. Tubes were incubated at 39C overnight with slow mixing. Next day, the ovaries were washed in a series of washing solutions at 39C with rotation: 3 washes in 50% formamide/2xSSC; 3 washes in 2xSSC. Drop of ProLong™ Gold Antifade Mountant with DAPI was added and ovaries were transferred to the 3D-FISH slide. Signal detection was performed with confocal microscope.

Computational methods

Whole-genome alignments and CE calling

For whole-genome alignments we used LastZ tool (**Harris, R.S. (2007) Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University.**) with parameters high-scoring segmantal pares (HSPs) threshold (--hspthresh) = 3000 interpolation threshold (--inner) = 2000, step size (--step) =20, alignmet processed with gap-free extension of seeds, gappes extension of HSPs and exluded chaining if HSPs (--gfextend --nochain --gapped).

To find alignment blocks used Mugsy (**Angiuoli SV and Salzberg SL. Mugsy: Fast multiple alignment of closely related whole genomes. Bioinformatics 2011 27(3):334-4**) with default option. To call CE, we used PhastCons. The target coverage is varied from 0.48 to 0.58 and expected lenth of CE is varied from 30 to 38 nucleotides in relation of reference species and chromosome.

Hi-C data processing

Raw reads were processed using Juicer protocol (ref). Contacts were normalized using KR-normalization. Expected contact counts were obtained by dumping expected vectors using juicer tools *dump* tool.

ChIP-seq data processing

All ChIP-seq data were processed using aquas pipeline (https://github.com/kundajelab/chipseq_pipeline) stopped at signal stage.

RNA-seq data processing

All data were processed using standard protocols with HISAT2, bedtools Genome Coverage, StringTie tools. The sequencing data were uploaded to the Galaxy web platform, and we used the public server at usegalaxy.org to analyze the data (Afgan et al. 2016).

Compartments calling

The default approach for PC1 values computation relies on using juicer tools *eigenvector* tool (ref) at 25-, 50 or 100-kb resolutions. For other approaches, explained in details in Supplementary Note I, we dumped observed/expected matrices and computed principal components using default correlation and principal component analysis functions available in R. To compute PC1 values for intrachromosomal submatrices we used window sizes equal to 10Mb.

TADs calling

To call TADs, we first used Armatus, Lavarbust, Dixon caller and hicExplorer *findTADs* utils (refs) with default parameters at 25- and 5-kb resolutions. Visual inspection of obtained TADs revealed that results were similar for Armatus, Lavarbust and hicExplorer algorithms, whereas Dixon caller results in large, megabase-scaled TADs, which do not correspond well with triangles visible on heatmaps. Although this difference is most probably due to default parameters of Dixon TAD caller, originally developed on mouse and human data, and results most probably could be improved by tweaking parameters, we decided to focus on identification of hicExplorer TADs at 5-kb resolution because visual assessment suggested that this caller provides the best results. By tweaking parameters we found that changing the delta value to 0.05 provides TADs most exactly matching triangles on heatmaps; changing other parameters does not lead to substantial improvements of TADs. Finally, resulting TADs were visually

inspected to correct boundaries positions in problematic regions (long heterochromatin blocks, gaps and etc.).

Slope plots analysis

To produce slope plots, we produced the best linear fit of the log-log scaled dependence of expected contact frequencies from genomic distance. For each distance, we used few points to obtain local fit. As contacts become noisier with distance, we used more points to fit regression at larger distances. In particular, to compute slope at distance X we used all expected values in the interval [X, 50 000 + 0.25*X]. We computed slope for each chromosome with a length of more than 20 Mb individually, and then used median of all obtained values. We cropped the resulting plot at 50 Mb because slope values become too noisy around this point.

Phylogenomic analysis and calculation of divergence times

Genome assemblies for the six mosquito species available from VectorBase (Giraldo-Calderón et al. 2015; PMID: 25510499) release VB-2019-08 were analysed with the Diptera dataset of the Benchmarking Universal Single-Copy Orthologue (BUSCO v3.0.2) assessment tool (Waterhouse et al. 2018; PMID: 29220515). From the results, 1'258 BUSCO genes present as single-copy orthologues in all six species were identified. The protein sequences for each BUSCO were aligned with MAFFT v7.450 (Katoh & Standley 2013; PMID: 23329690) and then filtered/trimmed with TrimAl v1.2 (Capella-Gutiérrez et al. 2009; PMID: 19505945) using automated1 parameters to produce a concatenated superalignment. AliStat v1.12 (https://www.biorxiv.org/content/10.1101/2020.01.15.907733v2) assessment of the superalignment: 6 sequences; 854'431 sites; completeness score: 0.96383. The phylogeny was then estimated using RAxML v8.0.0 (Stamatakis, 2014; PMID: 24451623) with the GAMMAPROTJTT model with 100 bootstrap samples. Rooted with *Aedes aegypti*, the molecular phylogeny was converted to an ultrametric time-calibrated phylogeny using the chronos function in R (Paradis 2013:

PMID: 23454091) using the discrete model and fixing *An. gambiae* complex age at 0.5 million years according to (Thawornwattana et al, 2018: PMID: 30102363) and the *Anopheles* genus age at 100 million years in line with (Neafsey et al, 2015) and the geological split of western Gondwana.

Funding: This work was supported by the NSF grant MCB-1715207, NIH NIAID grant R21AI135298, and the USDA National Institute of Food and Agriculture Hatch project 223822 to IVS. The reported study of *An. atroparvus* was partly funded by RFBR according to the research project №19-34-50051 to IVS and VL. PacBio sequencing of *An. merus* was funded by a grant from the University of Lausanne Department of Ecology and Evolution to RMW and NIH grant ….to ZT. MJMFR, LR, and RMW were supported by Novartis Foundation for medical-biological research grant #18B116 and Swiss National Science Foundation grant PP00P3_170664. VL was partly supported by the Fulbright Foreign Student Program, Grantee ID: 15161026. This work was supported by the Ministry of Education and Science of Russian Federation, grant #2019-0546 (FSUS-2020-0040).

Acknowledgments: The following reagents were obtained through BEI Resources, NIAID, NIH: *An. coluzzii*, Strain MOPTI, Eggs, MRA-763, contributed by Gregory C. Lanzaro; *An. merus*, Strain MAF, MRA-1156, contributed by Maureen Coetzee; *An. atroparvus*, Strain EBRO, Eggs, MRA-493, contributed by Carlos Aranda and Mark Q. Benedict; *An. albimanus*, Strain STECLA, Eggs, MRA-126, contributed by Mark Q. Benedict. All computations were performed using nodes of the high-throughput cluster of the Novosibirsk State University, and bioinformatics resource center of the Institute of Cytology and Genetics (Budget Project 0324-2019-0041-C-01)

Supplementary materials

Please see attached document