Leveraging evolutionary relationships to improve *Anopheles* genome assemblies

Robert M. Waterhouse[1,*], Sergey Aganezov[2,3], Yoann Anselmetti[4], Jiyoung Lee[5], Livio Ruzzante[1], Maarten J.M.F. Reijnders[1], Sèverine Bérard[4], Romain Feron[1], Phillip George[6], Matthew W. Hahn[7], Paul I. Howell[8], Maryam Kamali[6,9], Sergey Koren[10], Daniel Lawson[11], Gareth Maslen[11], Ashley Peery[6], Adam M. Phillippy[10], Maria V. Sharakhova[6,12], Eric Tannier[13,14], Maria F. Unger[15], Simo V. Zhang[7], Max A. Alekseyev[16], Nora J. Besansky[15], Cedric Chauve[17], Scott J. Emrich[18], Igor V. Sharakhov[5,6,12,*]

[1] Department of Ecology and Evolution, University of Lausanne, and Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland.
[2] Department of Computer Science, Princeton University, Princeton, New Jersey 08450, USA.
[3] Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, 21218, USA.
[4] Institut des Sciences de l'Evolution de Montpellier, Université de Montpellier, Centre National de la Recherche Scientifique, Institut de Recherche pour le Développement, École Pratique des Hautes Études, 34090 Montpellier, France.
[5] The Interdisciplinary PhD Program in Genetics, Bioinformatics, and Computational Biology, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, USA.
[6] Department of Entomology, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, USA.
[7] Departments of Biology and Computer Science, Indiana University, Bloomington, Indiana 47405, USA.
[8] Centers for Disease Control and Prevention, Atlanta, Georgia 30329, USA.
[9] Department of Medical Entomology and Parasitology, School of Medical Sciences, Tarbiat Modares University, Tehran, Iran.
[10] Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.
[11] European Molecular Biology Laboratory - European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD, UK.
[12] Laboratory of Ecology, Genetics and Environmental Protection, Tomsk State University, Tomsk 634050, Russia.
[13] Laboratoire de Biométrie et Biologie Evolutive, Université Lyon 1, Unité Mixte de Recherche 5558 Centre National de la Recherche Scientifique, 69622 Villeurbanne, France.
[14] Institut national de recherche en informatique et en automatique, Grenoble, Rhône-Alpes, 38334 Montbonnot, France.
[15] Eck Institute for Global Health and Department of Biological Sciences, University of Notre Dame, Galvin Life Sciences Building, Notre Dame, Indiana 46556, USA.
[16] Department of Mathematics and Computational Biology Institute, George Washington University, Ashburn, Virginia 20147, USA.
[17] Department of Mathematics, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada.
[18] Department of Computer Science and Engineering, Eck Institute for Global Health, Cushing Hall, University of Notre Dame, Notre Dame, Indiana 46556, USA.
[*] Correspondence should be addressed to: robert.waterhouse@unil.ch (RMW), igor@vt.edu (IVS)

Running title: Evolutionary scaffolding of mosquito genomes

Abstract

While new sequencing technologies have lowered financial barriers to whole genome sequencing, resulting assemblies are often fragmented and far from 'finished'. Subsequent improvements towards chromosomal-level status can be achieved by both experimental and computational approaches. Requiring only annotated assemblies and gene orthology data, comparative genomics approaches that aim to capture evolutionary signals to predict scaffold neighbours (adjacencies) offer potentially substantive improvements without the costs associated with experimental scaffolding or re-sequencing. We leverage the combined detection power of three such gene synteny-based methods applied to 21 *Anopheles* mosquito assemblies with variable contiguity levels to produce consensus sets of scaffold adjacency predictions. Three complementary validations were performed on subsets of assemblies with additional supporting data: six with physical mapping data; 13 with paired-end RNA sequencing (RNAseq) data; and three with new assemblies based on re-scaffolding or incorporating Pacific Biosciences (PacBio) sequencing data. Improved assemblies were built by integrating the consensus adjacency predictions with supporting experimental data, resulting in 20 new reference assemblies with improved contiguities. Combined with physical mapping data for six anophelines, chromosomal positioning of scaffolds improved assembly anchoring by 47% for *A. funestus* and 38% *A. stephensi*. Reconciling an *A. funestus* PacBio assembly with synteny-based and RNAseq-based adjacencies and physical mapping data resulted in a new 81.5% chromosomally mapped reference assembly and cytogenetic photomap. While complementary experimental data are clearly key to achieving high-quality chromosomal-level assemblies, our assessments and validations of gene synteny-based computational methods highlight the utility of applying comparative genomics approaches to improve community genomic resources.

Keywords: genome assembly, gene synteny, comparative genomics, mosquito genomes, orthology, bioinformatics, computational evolutionary biology, chromosomes, physical mapping

Introduction

Reduced costs of new sequencing technologies have facilitated the rapid growth of draft genome assemblies from all kingdoms of life. Nevertheless, the often painstaking process of progressing from draft status to that of a 'finished' reference genome—a near-complete and near-contiguous chromosomal-level assembly—remains the exclusive accomplishment of relatively few species. Chromosomal ordering and orienting of contigs or scaffolds may be achieved by experimental approaches including fluorescence *in situ* hybridization (FISH) (Bauman et al. 1980), genetic linkage mapping (Hahn et al. 2014; Fierst 2015), optical (restriction site) mapping (Levy-Sakin and Ebenstein 2013), or analysis of chromatin interaction frequency data (Kaplan and Dekker 2013; Burton et al. 2013). When resources allow, combined approaches can produce excellent results, e.g. for Brassicaceae plants (Jiao et al. 2017), the three-spined stickleback (Peichel et al. 2017), and the mosquitoes, *Aedes aegypti* and *Culex quinquefasciatus* (Dudchenko et al. 2017; Matthews et al. 2017).

While many research applications do not strictly require such high-quality assemblies, improvements in completeness, contiguity, and chromosomal anchoring can substantially add to the power and breadth of biological and evolutionary inferences from comparative genomics or population genetics analyses. For example, extensive contiguity and chromosomal-level anchoring are clearly important when addressing questions concerning karyotype evolution or smaller-scale inversions and translocations, re-sequencing analyses of population-level samples, reconstructing rearrangement-based phylogenies, identifying and characterising genes that localise within quantitative trait loci (QTLs), examining genomic sexual conflicts, or tracing drivers of speciation. In many such studies, assembly improvements were critical to enable more robust analyses, e.g. QTL analysis with rape mustard flowering-time phenotypes (Markelz et al. 2017); contrasting genomic patterns of diversity between barley cultivars (Mascher et al. 2017); defining rearrangements of the typical avian karyotype (Damas et al. 2017); detecting chromosome fusion events during butterfly evolution (Davey et al. 2016); characterising the ancestral lepidopteran karyotype (Ahola et al. 2014); identifying the chromosomal position and structure of the male determining locus in *Ae. aegypti* (Matthews et al. 2017); and characterising a melon fly genetic sexing strain as well as localising the sexing trait (Sim and Geib 2017).

Available genome assemblies for anopheline mosquitoes vary considerably in contiguity and levels of chromosomal anchoring. Sequencing the first mosquito genome produced an assembly for the *A. gambiae* PEST strain with 8'987 scaffolds spanning 278 megabasepairs (Mbp), where 303 scaffolds spanned 91% of the assembly and physical mapping assigned 84% of the genome to chromosomal arms (Holt et al. 2002). Additional FISH mapping and orienting of 28 scaffolds and bioinformatics analyses later facilitated

an assembly update by removing haplotype scaffolds and bacterial sequences and anchoring a third of previously unmapped scaffolds to chromosomes (Sharakhova et al. 2007). Since then, more than 20 new assemblies have been built for the anophelines, several with mapping efforts that enabled at least partial chromosomal anchoring. Sequencing of the *A. gambiae* Pimperena S form and *A. coluzzii* (formerly *A. gambiae* M form) produced assemblies with 13'050 and 10'525 scaffolds, respectively, with 89% of each of these assemblies alignable to the closely related PEST assembly (Lawniczak et al. 2010). The much smaller 174 Mbp assembly of the more distantly related neotropical vector, *A. darlingi*, comprised 8'233 scaffolds, but they remained unanchored (Marinotti et al. 2013). Physical mapping assigned 62% of the 221 Mbp *A. stephensi* Indian strain assembly (23'371 scaffolds) (Jiang et al. 2014) and 36% of the *A. sinensis* Chinese strain assembly (9'597 scaffolds) (Zhou et al. 2014; Wei et al. 2017) to polytene chromosomes. The *Anopheles* 16 Genomes Project (Neafsey et al. 2013) produced assemblies ranging from a few hundred to several thousand scaffolds and used mapping data from four species to anchor *A. funestus* (35%), *A. atroparvus* (40%), *A. stephensi* SDA-500 strain (41%), and *A. albimanus* (76%) genome assemblies to chromosomal arms (Neafsey et al. 2015). Additional physical mapping data for *A. atroparvus* subsequently improved this initial assembly to 90% chromosomal anchoring (Artemov et al. 2018), and for *A. albimanus* to 98% (Artemov et al. 2017), higher levels of assignment than *A. gambiae* PEST.

For a genus such as *Anopheles* with already more than 20 genome assemblies available (Ruzzante et al. 2018), contiguity can be improved by leveraging information from cross-species comparisons to exploit patterns of conservation and identify potential scaffold adjacencies. While genome rearrangements can and do occur, multiple homologous genomic regions with conserved orders and orientations, i.e. regions with maintained synteny, offer an evolutionarily guided approach for assembly improvement. Specifically, employing orthologous genes as conserved markers allows for the delineation of maintained syntenic blocks that provide support for putative scaffold adjacencies. Here we present results from applying three computational approaches, ADSEQ (Anselmetti et al. 2018), GOS-ASM (Aganezov and Alekseyev 2016), and ORTHOSTITCH (this study), to assess the performance of evolutionarily guided assembly improvements of multiple anopheline genomes. Consensus predictions offer well-supported sets of scaffold adjacencies that lead to the improved contiguity of draft assemblies without the associated costs or time-investments required for experimental support. Validations of these predictions exploiting experimental data for subsets of the anophelines supported many adjacencies and highlighted the complementarity of experimental and computational approaches. Thus, whether employed as supporting data for experimentally based assembly improvement approaches, as complementary data to further enhance such improvements, or as stand-alone evidence as part of an assembly building pipeline, these evolutionarily guided methods offer a handy new set of utensils in any genome assembly toolbox. These comparative genomics approaches will help to propel the draft assemblies from similar species-clusters along the journey towards becoming 'finished' reference genomes.

Results

Synteny-improved contiguities of *Anopheles* genome assemblies

The paucity of supporting experimental data for many species means that improving the contiguity of current draft assemblies must often rely solely on comparative genomics approaches. The anophelines, as a genus where numerous assemblies are available and a few are already highly contiguous, present an ideal opportunity to apply and assess the performance of such approaches. Using orthologues delineated across 21 anopheline gene sets (**Supplemental Table S1**) and combining the results from three synteny-based approaches, ADSEQ (Anselmetti et al. 2018), GOS-ASM (Aganezov and Alekseyev 2016), and ORTHOSTITCH (**see Methods; Supplementary Online Material; Supplemental Fig. S1; Supplemental Tables S2, S3**), two-way consensus sets of well-supported predicted scaffold adjacencies resulted in substantial improvements for several assemblies (**Fig. 1**). The two-way consensus adjacencies were required to be predicted by at least two of the approaches with no third-method conflicts (**see Methods**). Improvements were quantified in terms of the absolute (**Fig. 1A**) and relative (**Fig. 1B**) increases in scaffold N50 values (a median-like metric where half the genome is assembled into scaffolds of length N50 or longer) and decreases in scaffold counts, considering only scaffolds with annotated orthologous genes used as input data for the scaffold adjacency predictions.

The greatest absolute increases in scaffold N50 values were achieved for *A. dirus* and *A. minimus*, while the greatest absolute reductions in scaffold counts were achieved for *A. christyi*, *A. culicifacies*, *A. maculatus*, and *A. melas* (**Fig. 1A**), reflecting the different levels of contiguity of their input assemblies. Reductions in the numbers of scaffolds that comprise each assembly varied from 1'890 fewer for the rather fragmented *A. melas* assembly to just one fewer for the already relatively contiguous *A. albimanus* assembly. Even without large reductions in the numbers of scaffolds, when a few adjacencies bring together relatively long scaffolds then they can lead to marked improvements in N50 values. For example, *A. dirus* and *A. minimus* improved with N50 increases of 5.1 Mbp and 4.8 Mbp and only 36 and 12 fewer scaffolds, respectively.

The general trend indicates that reducing the number of scaffolds by about a third leads to a doubling of the N50 value (**Fig. 1B**). Exemplifying this trend, *A. epiroticus* showed the greatest relative reduction in the number of scaffolds (40%) and achieved a 2.1-fold N50 increase. Notable exceptions include *A. farauti*, which showed a 1.4-fold N50 increase with a 30% reduction in the number of scaffolds, while *A. dirus* and *A. stephensi* (Indian) achieved 1.66-fold and 2.08-fold N50 increases with only 14% and 19% reductions in the number of scaffolds, respectively. Using only three-way consensus adjacencies led to

more conservative improvements, while employing a liberal union of all non-conflicting adjacencies resulted in a trend of a ~30% scaffold reduction to double the N50 value (**Supplemental Figs. S2, S3**). The enhanced contiguities of these anopheline assemblies based on predicted scaffold adjacencies demonstrate that while the results clearly depend on the quality of the input assemblies, applying synteny-based approaches can achieve substantial improvements.

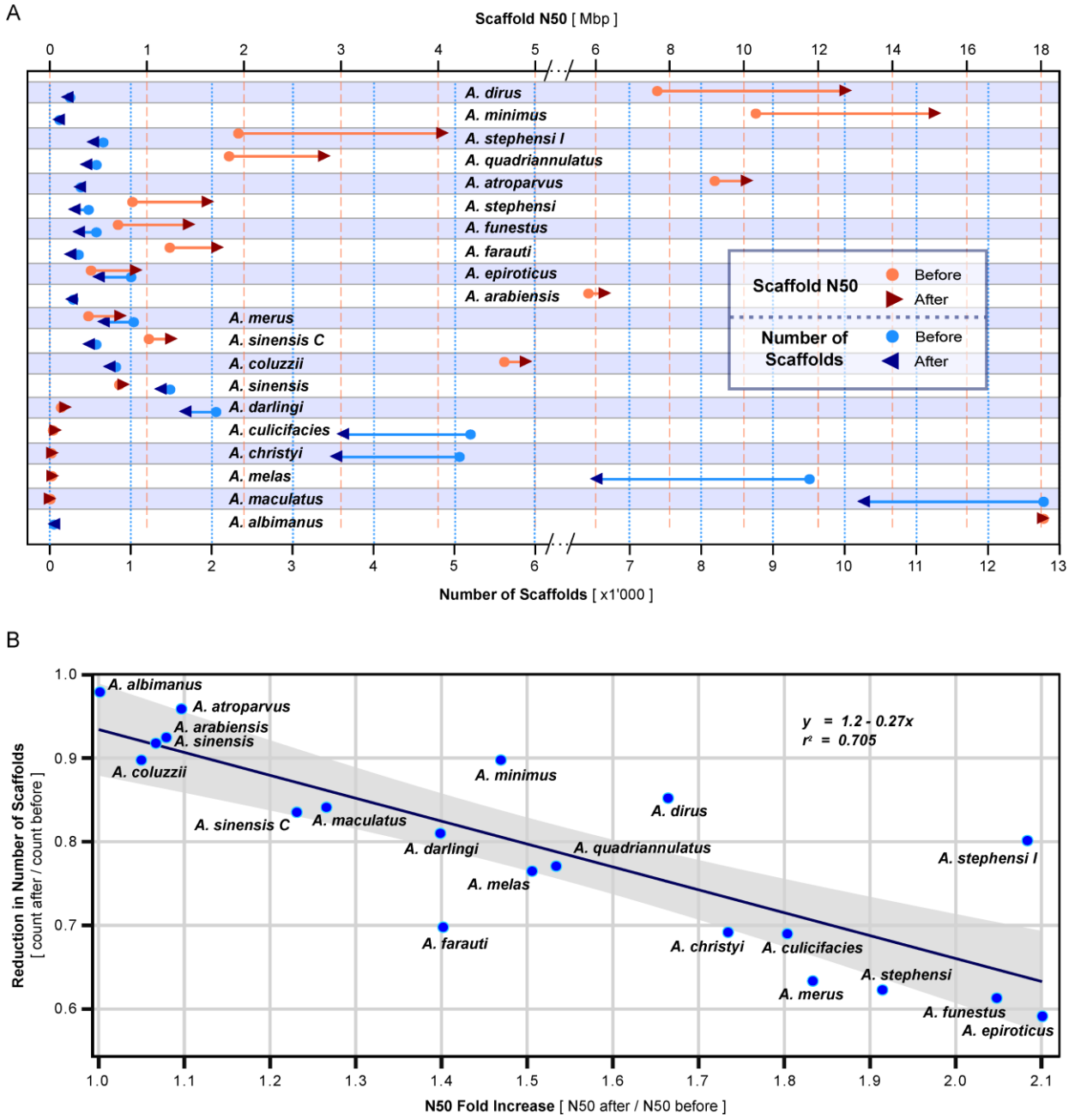


Figure 1. Improved genome assemblies for 20 anophelines from synteny-based scaffold adjacency predictions. Results from ADSEQ, GOS-ASM, and ORTHOSTITCH predictions were compared to define two-way consensus adjacencies predicted by at least two of the three approaches, where the third approach did not conflict. These adjacencies were used to build new assemblies with improved contiguities, quantified by comparing before and after scaffold counts and N50 values (half the total assembly length comprises scaffolds of length N50 or longer). The counts, values, and ratios represent only scaffolds with annotated orthologous genes used as the input dataset for the scaffold adjacency predictions. (***A***) Scaffold counts (blues, bottom axis) and N50 values (red/orange, top axis) are shown before (dots) and after (arrowheads) synteny-based improvements were applied. The 20 anopheline assemblies are ordered from the greatest N50 improvement at the top for *Anopheles dirus* to the smallest at the bottom for *Anopheles albimanus*. Note axis scale changes for improved visibility after N50 of 5 Mbp and scaffold count of 6'000. (***B***) Plotting before to after ratios of scaffold counts versus N50 values (counts or N50 after / counts or N50 before superscaffolding of the adjacencies) reveals a general trend of a ~33% reduction in scaffold numbers resulting in a ~2-fold increase of N50 values. The line shows the linear regression with a 95% confidence interval in grey. Results for two strains are shown for *Anopheles sinensis*, SINENSIS and Chinese (C), and *Anopheles stephensi*, SDA-500 and Indian (I).

Consensus adjacencies from complementary synteny-based methods

Although each of the computational methods aims to predict scaffold adjacencies based on gene collinearity, they differ in some of their underlying assumptions and in their implementations that identify, score, and infer the most likely scaffold neighbours. ADSEQ uses reconciled gene trees to reconstruct ancestral genomes and to delineate extant gene adjacencies in a duplication-aware parsimonious evolutionary scenario of adjacency gains and breaks that also identifies extant adjacencies between genes at scaffold extremities (Anselmetti et al. 2018). GOS-ASM employs the concept of the breakpoint graph and utilizes the topology of the species phylogeny to perform evolutionary rearrangement analysis of orthologous genes from multiple genomes from which scaffold adjacencies can then be inferred (Aganezov and Alekseyev 2016). ORTHOSTITCH, a new approach developed as part of this study, interrogates gene orthology data from cross-species comparisons to evaluate synteny evidence that supports putative scaffold adjacencies (**see Methods**). Similar to traditional meta-assembly-like methods that leverage such differences to identify well-supported consensus predictions, we compared the results of all scaffold adjacencies predicted by each method using the Comparative Analysis and Merging of Scaffold Assemblies (CAMSA) tool (Aganezov and Alekseyev 2017) (**see Methods; Supplementary Online Material; Supplemental Table S3**).

For the full set of 20 assemblies, GOS-ASM and ORTHOSTITCH predicted about 10'000 oriented adjacencies each, with just over twice as many predictions from ADSEQ. Comparing all predictions identified almost 30'000 distinct scaffold adjacencies, 36% of which were supported by at least two methods; this fraction comprised 10% that were in three-way agreement and a further 20% that were in two-way agreement with no conflicts with the third method (**Fig. 2; Supplemental Fig. S4**). The larger total number of predictions from ADSEQ resulted in much higher proportions of unique adjacencies (**Fig. 2**). Adjacencies in three-way agreement constituted 30% of GOS-ASM and 27% of ORTHOSTITCH predictions, and just 13% of the much more numerous ADSEQ predictions. In pairwise comparisons, ADSEQ supported almost two-thirds of each of the other prediction sets, while about a third of ADSEQ and GOS-ASM adjacencies agreed with ORTHOSTITCH, and slightly less than a third of ADSEQ and ORTHOSTITCH predictions were supported by GOS-ASM.

From the liberal union sets of all non-conflicting adjacencies for all assemblies, the adjacencies in three-way agreement made up 17% of the total, 46% of GOS-ASM, 39% of ORTHOSTITCH, and 19% of ADSEQ predictions (**Fig. 2B**). Considering only the supported predictions that were used to build the two-way consensus sets of adjacencies for the synteny-based assembly improvements presented in **Fig. 1**, i.e. excluding adjacencies predicted by only one method, the three-way consensus adjacencies made up 33% of the total, 54% of GOS-ASM, 44% of ORTHOSTITCH, and 33% of ADSEQ predictions (**Fig. 2B**). A third of these two-way supported consensus adjacencies that were employed to build the new superscaffolded assemblies were predicted by all three methods, with 98% supported by ADSEQ, 74% by ORTHOSTITCH, and 61% by GOS-ASM. Thus, comparing the results from the three methods and employing a two-way agreement with no third-method conflict filter resulted in an improved level of three-way adjacency agreements from a tenth to a third.

For the individual assemblies, more than half of the distinct scaffold adjacencies were in agreement for *A. epiroticus*, *A. merus*, and both the *A. stephensi* assemblies, with *A. funestus* achieving the highest consistency at 58% (**Fig. 2C; Supplemental Fig. S5**). Some of the most fragmented input assemblies produced some of the largest sets of distinct adjacency predictions but the agreement amongst these predictions was generally lower than the other assemblies. For example, *A. maculatus* was the least contiguous input assembly and produced more than 8'000 distinct predictions, of which only 18% showed at least two-way agreement with no conflicts (**Fig. 2C; Supplemental Fig. S5**).

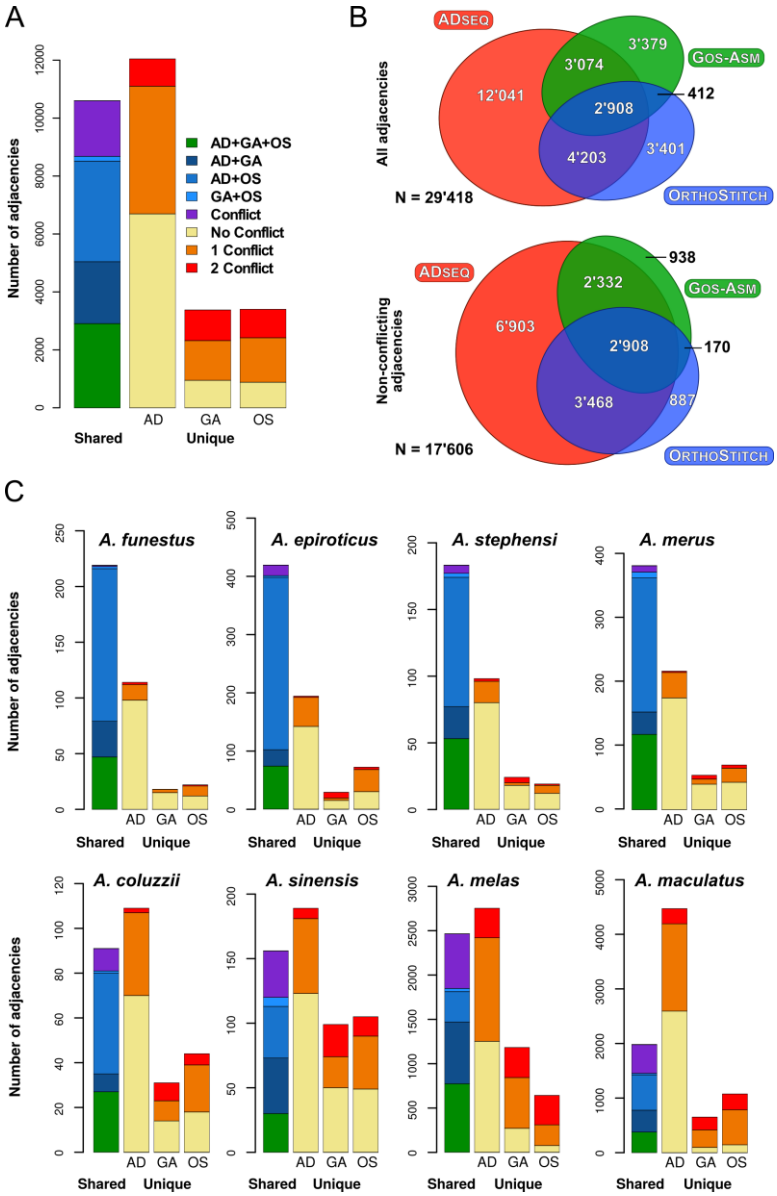


Figure 2. Comparisons of synteny-based scaffold adjacency predictions from ADSEQ (AD), GOS-ASM (GA), and ORTHOSTITCH (OS). Bar charts show counts of predicted adjacencies (pairs of neighbouring scaffolds) that are shared amongst all three methods (green), or two methods without (blues) and with (purple) third method conflicts, or that are unique to a single method and do not conflict (yellow) or do conflict with predictions from one (orange) or both (red) of the other methods. (***A***) Results of all adjacencies summed across all 20 anopheline assemblies. (***B***) Area-proportional Euler diagrams showing (top) the extent of the agreements amongst the three methods for all 29'418 distinct scaffold adjacencies, and (bottom) the extent of the agreements amongst the three methods for the 17'606 distinct and non-conflicting scaffold adjacencies (the liberal union sets), both summed over all 20 assemblies. (***C***) Individual results of adjacencies for representative anopheline assemblies, four with more than 50% agreement (top row), and four with lower levels of agreement (bottom row). Colours for each fraction are the same as in panel A, y-axes vary for each assembly with maxima of 120 for *Anopheles coluzzii* to 5'000 for *Anopheles maculatus*. Results for *Anopheles stephensi* are for the SDA-500 strain.

Validated adjacencies with physical mapping and RNA sequencing data

Physical mapping data generated from a subset of the anophelines considered here allowed for independent, quantitative validations of the synteny-based predictions and their consensus sets. Building cytogenetic photomaps and conducting extensive FISH experiments mapped 31 *A. albimanus* scaffolds (Artemov et al. 2017), 46 *A. atroparvus* scaffolds (Artemov et al. 2015; Neafsey et al. 2015; Artemov et al. 2018), 204 *A. funestus* scaffolds (Sharakhov et al. 2002, 2004; Xia et al. 2010; Neafsey et al. 2015) (including additional mapping for this study), 52 *A. sinensis* scaffolds (Chinese) (Wei et al. 2017), 99 *A. stephensi* (SDA-500) scaffolds (Neafsey et al. 2015), and 118 *A. stephensi* (Indian) scaffolds (Jiang et al. 2014) (including additional mapping for this study) (**see Methods; Supplementary Online Material; Supplemental Tables S4, S5; Supplemental Fig. S6**). The scaffold adjacencies identified from these physical mapping data, i.e. pairs of neighbouring mapped scaffolds, were compared with adjacencies predicted by each of the three methods and the CAMSA-generated two-way consensus sets, as well as the conservative three-way consensus sets and the liberal union sets of all non-conflicting adjacencies (**Supplemental Table S6**). With 85 physically mapped scaffold adjacencies, *A. funestus* validations confirmed 12-17% of the different sets of synteny-based adjacencies and highlighted conflicts with just 4-8% (**Fig. 3A**). Five of the 15 two-way consensus synteny-based predictions were confirmed by physical mapping of *A. atroparvus* scaffolds and only one conflict was identified (**Fig. 3A**). Examining the identified conflicts in detail revealed that most were resolvable. As not all scaffolds were targeted for physical mapping, neighbouring scaffolds on the physical maps could have shorter unmapped scaffolds between them that were identified by the synteny-based approaches. For *A. funestus*, five conflicts were resolved because the synteny-based neighbour was short and not used for physical mapping and an additional four conflicts were resolved by switching the orientation of physically mapped scaffolds, which were anchored by only a single FISH probe and therefore their orientations had not been confidently determined.

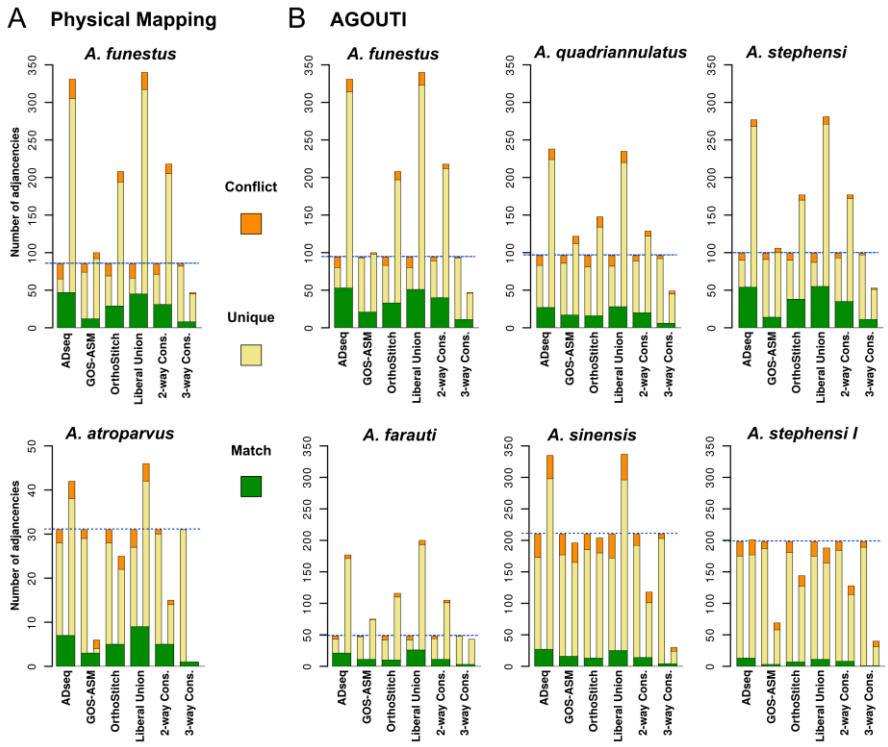


Figure 3. Scaffold adjacency validations with physical mapping and RNA sequencing data. The bar charts show counts from each set of synteny-based scaffold adjacency predictions compared with the adjacencies from the physical mapping (***A***) or AGOUTI (***B***) sets. The synteny-based sets comprise predictions from three different methods, ADSEQ, GOS-ASM, and ORTHOSTITCH, as well as their Liberal Union (all non-conflicting predictions), their two-way consensus (2-way Cons. predicted by two methods and not conflicting with the third method), and their three-way consensus (3-way Cons. predicted by all three methods). Adjacencies that are exactly matching form the green base common to both sets in each comparison, from which extend bars showing physical mapping or AGOUTI adjacency counts (left) and synteny-based adjacency counts (right) that are unique (yellow) or conflicting (orange) in each comparison. Blue dashed lines highlight the total adjacencies for the physical mapping or AGOUTI sets. For comparison all y-axes are fixed at a maximum of 350 adjacencies, except for *Anopheles atroparvus*. Results for two strains are shown for *Anopheles stephensi*, SDA-500 and Indian (I).

Transcriptome data from RNAseq experiments can provide additional information about putative scaffold adjacencies when individual transcripts (or paired-end reads) reliably map to scaffold extremities. The Annotated Genome Optimization Using Transcriptome Information (AGOUTI) tool (Zhang et al. 2016) employs RNAseq data to identify such adjacencies as well as correcting any fragmented gene models at the ends of scaffolds. Using available paired-end RNAseq mapping data from VECTORBASE (Giraldo-Calderón et al. 2015), scaffold adjacencies predicted for 13 anophelines ranged from just two for *A. albimanus* to 210 for *A. sinensis* (SINENSIS) (**see Methods; Supplementary Online Material;**

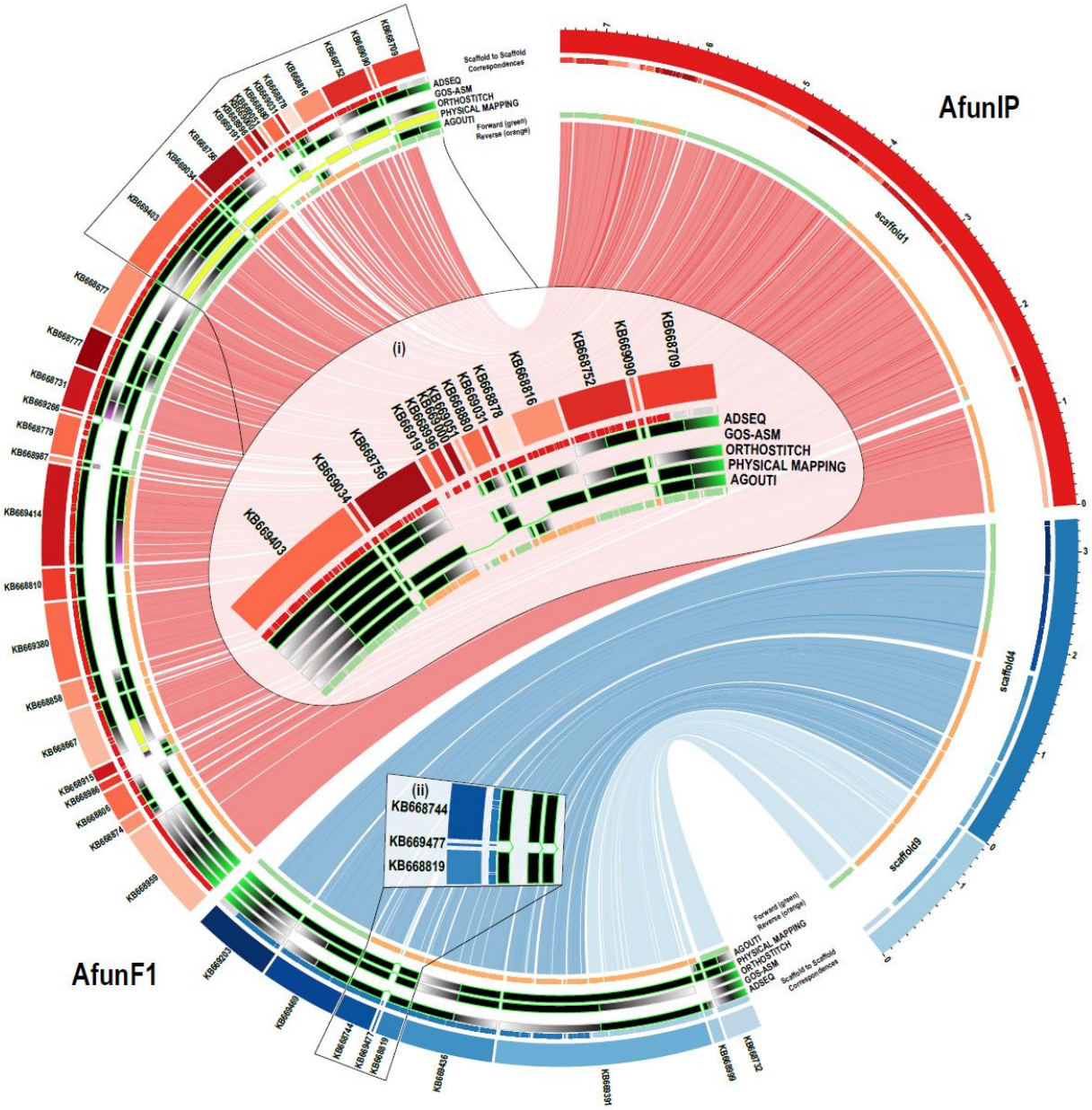
Supplemental Table S7). These AGOUTI-based scaffold adjacencies were compared with the adjacencies predicted by each of the three methods and the CAMSA-generated two-way consensus sets, as well as the conservative three-way consensus sets and the liberal union sets of all non-conflicting adjacencies (**Fig. 3B; Supplemental Table S8**). Across all 13 assemblies, 18% of AGOUTI-based scaffold adjacencies supported the two-way consensus synteny-based adjacencies, 75% were unique to the AGOUTI sets, and only 7% were in conflict. Nearly 200 AGOUTI-based scaffold adjacencies for *A. stephensi* (Indian) confirmed only eight and conflicted with 14 of the two-way consensus set adjacencies (**Fig. 3B**). In contrast, about half as many AGOUTI-based scaffold adjacencies each for *A. stephensi* (SDA-500) and *A. funestus* confirmed four to five times as many two-way consensus set adjacencies and conflicted with only five and six, respectively. Notably, 68% of the AGOUTI-based scaffold adjacencies that produced conflicts with the two-way consensus set adjacencies comprised scaffolds with no annotated orthologues. These cases can be resolved by noting that only scaffolds with orthologous genes were used for synteny-based predictions: therefore, the inferred neighbouring scaffolds could have shorter un-annotated scaffolds between them that were identified by AGOUTI. Such un-annotated scaffolds were also numerous amongst the adjacencies that were unique to AGOUTI where for 66% either one or both scaffolds had no annotated orthologues.

Validated adjacencies with new genome assemblies

A new *A. funestus* assembly, designated AfunIP, was generated as part of this study by merging approximately 70X of PacBio sequencing data with the reference assembly (AfunF1), with subsequent scaffolding using the original Illumina sequencing data (**see Methods; Supplementary Online Material; Supplemental Fig. S7; Supplemental Table S9**). This AfunIP genome assembly for *A. funestus* enabled the validation of the scaffold adjacency predictions for the AfunF1 assembly by examining collinearity between the two assemblies. AfunF1 scaffolds were ordered and oriented based on their alignments to AfunIP scaffolds and the resulting 321 alignment-based scaffold adjacencies were then compared with the synteny-based and AGOUTI predictions as well as with the physical mapping adjacencies to identify supported, unique, and conflicting adjacencies (**Fig. 4; Supplemental Fig. S8; Supplemental Table S10**). Each of the three synteny method prediction sets, as well as the two-way consensus and liberal union sets, had 14-17.5% in common with the alignment-based scaffold adjacencies, fewer than a quarter in conflict, and almost two thirds that were neither supported nor in conflict (**Supplemental Table S10**). The physical mapping adjacencies had generally more support, but also more conflicts as about half disagreed with the alignment-based adjacencies. Several disagreements were easily resolved by comparing these conflicts with those identified from the synteny-based

adjacencies (those shown in **Fig. 3A**) and confirming that switching the orientation of physically mapped scaffolds corrected the relative placements of these scaffolds, e.g. **Fig. 4 inset (i)**. Similarly to the validations with the physical mapping and RNAseq data presented above, apparent conflicts with the alignment-based adjacencies can also arise because using genome alignment data considered all alignable scaffolds while physical mapping targeted only large scaffolds and synteny methods did not consider scaffolds with no annotated orthologues (i.e. short scaffolds). This is exemplified in **Fig. 4 inset (ii)** where the alignment data placed a short scaffold between two scaffolds predicted to be neighbours by ADSEQ, ORTHOSTITCH, and physical mapping data. Skipping such short scaffolds (<5 Kbp) to define a smaller set of alignment-based adjacencies considering only the longer scaffolds resulted in increased support for the synteny-based sets of 19-23%, and most notably up to 39% for the physical mapping adjacencies, while only marginally increasing support for AGOUTI predictions from 15% to 17% (**Supplemental Table S10**).

Figure 4. Whole genome alignment comparisons of selected *Anopheles funestus* AfunF1 and AfunIP scaffolds. The plot shows correspondences of three AfunIP scaffolds (right) with AfunF1 (left) scaffolds based on whole genome alignments, with links coloured according to their AfunIP scaffold. Putative adjacencies between AfunF1 scaffolds are highlighted with tracks showing confirmed neighbours (black with bright green borders), supported neighbours with conflicting orientations (yellow), scaffolds with putative adjacencies that conflict with the alignments (purple gradient), scaffolds without putative adjacencies and thus no conflicts with the alignments (grey gradient) for: from outer to inner tracks, ADSEQ, GOS-ASM, ORTHOSTITCH, physical mapping, and AGOUTI. The innermost track shows alignments in forward (green) and reverse (orange) orientations. The outermost track shows alignments coloured according to the corresponding scaffold in the other assembly (light grey if aligned to scaffolds not shown). Inset (i) shows how corrected orientations of physically mapped scaffolds agree with the other methods. Inset (ii) shows how the alignments identified a short scaffold that was placed between two scaffolds identified by three other methods.


AfunIP
AfunF1
scaffold1
scaffold4
scaffold9
(i)
(ii)
KB668744
KB669477
KB668819
KB668709
KB669090
KB668752
KB668816
KB668878
KB669031
KB668880
KB669191
KB668756
KB669034
KB669403
ADSEQ
GOS-ASM
ORTHOSTITCH
PHYSICAL MAPPING
AGOUTI
Scaffold to Scaffold Correspondences
Forward (green)
Reverse (orange)

Re-scaffolding of the initial *A. farauti* (AfarF1) and *A. merus* (AmerM1) assemblies employed large-insert 'Fosill' sequencing libraries and reduced the numbers of scaffolds from 550 to 310 and 2'753 to 2'027 and increased N50 values from 1'197 Kbp to 12'895 Kbp and 342 Kbp to 1'490 Kbp, respectively (Neafsey et al. 2015). The availability of these re-scaffolded assemblies enabled the validation of the synteny-based and AGOUTI-based scaffold adjacency predictions for the AfarF1 and AmerM1 assemblies by examining corresponding scaffolds from the AfarF2 and AmerM2 assemblies (**see Methods; Supplementary Online Material; Supplemental Fig. S9**). The comparisons identified full support for the majority (87% and 82%) of the two-way synteny consensus set adjacencies and unresolvable conflicts for just 5% and 10%, while the AGOUTI-based adjacencies achieved similarly high levels of full support (81% and 67%), but with slightly greater proportions of conflicts (**Supplemental Table S11**).

New *Anopheles funestus* cytogenetic photomap and physical genome map

The collated data for *A. funestus* allowed for a comprehensive update of the previously published chromosomal photomap from ovarian nurse cells (Sharakhov et al. 2002). The existing images of polytene chromosomes of the five arms common to all anophelines (X, 2R, 2L, 3R, and 3L) were further straightened to facilitate linear placements of the genomic scaffolds on the photomap. Major structural updates for the cytogenetic map included reversal of the order of divisions and subdivisions within the 3La inversion to follow the standard 3L+[a] arrangement, and merging of two small subdivisions with larger neighbouring subdivisions: 5D to 6 and 34D to 34C (**Fig. 5**). The extensive additional physical mapping performed for *A. funestus*, together with the new AfunIP assembly and sequence alignment-based comparisons with the AfunF1 assembly, enabled a physical genome map to be built (**Fig. 5**). The 126 previously FISH-mapped (Sharakhov et al. 2002, 2004; Xia et al. 2010) and 66 newly FISH-mapped DNA markers (**Supplemental Fig. S6**) were located with BLAST searches to 139 AfunF1 scaffolds and then compared with AfunIP scaffolds using whole genome pairwise alignments (**see Methods; Supplementary Online Material**). The placement of scaffolds along the photomap took advantage of comparisons with the synteny-based scaffold adjacency predictions and with the AfunF1-AfunIP whole genome pairwise alignments. Synteny- or alignment-based scaffold neighbours were added to the genome map when they were short and thus had not been used for physical mapping. Additionally, scaffolds which were anchored with only a single FISH probe (i.e. with undetermined orientations) were reoriented when synteny- or alignment-based scaffold adjacencies provided supporting evidence to correct their relative placements on the map. The resulting physical genome map for *A. funestus* includes 204 AfunF1 scaffolds (**Supplemental Table S5**), with a further 99 neighbouring scaffolds after incorporating the synteny-based and AGOUTI-based adjacencies.

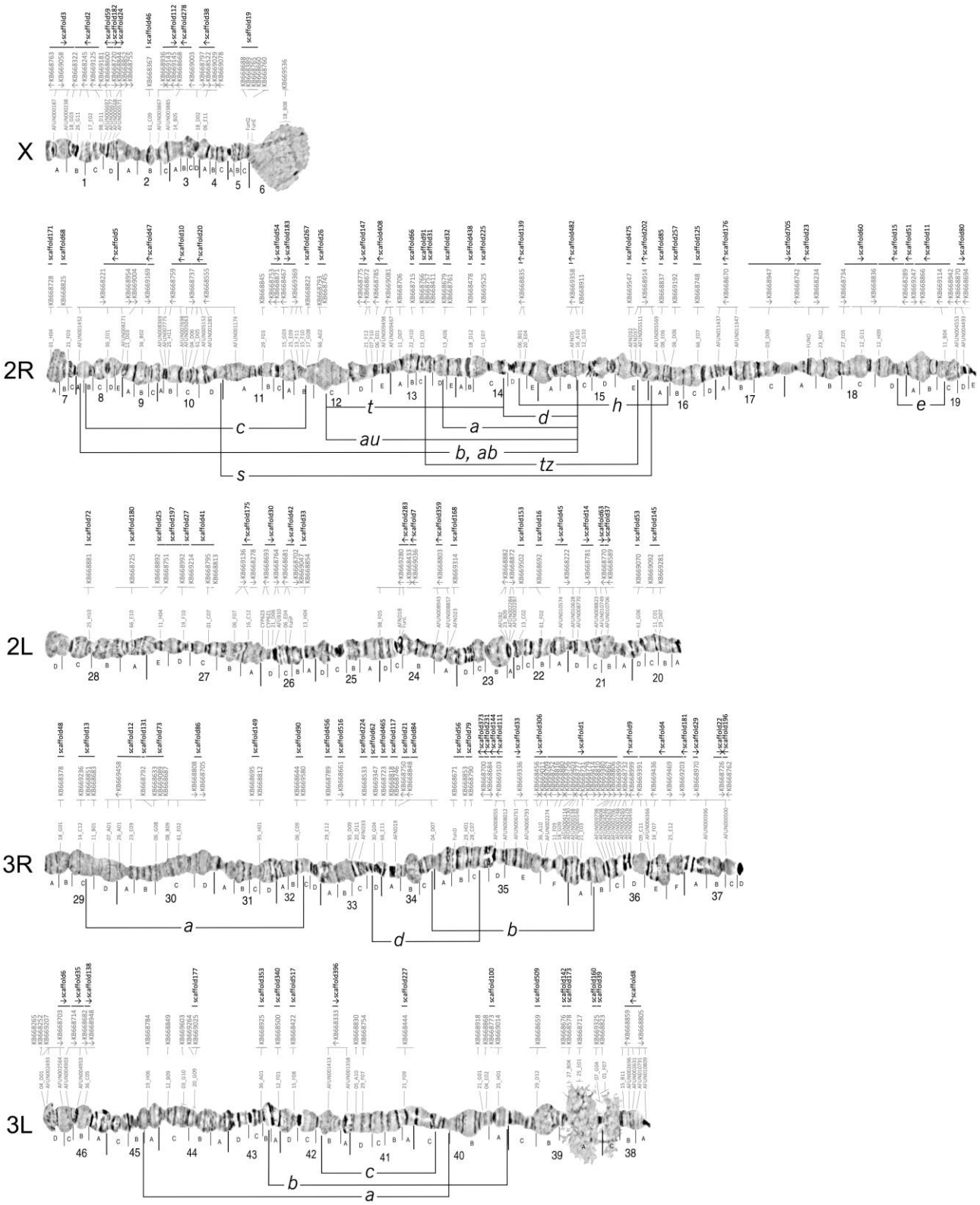

X
2R
2L
3R
3L

Figure 5. The *Anopheles funestus* cytogenetic photomap of straightened polytene chromosomes with anchored scaffolds from the AfunF1 and AfunIP assemblies. FISH-mapped DNA markers (grey probe identifiers directly above each chromosome) show the density of physical mapping along the chromosomal arm subdivisions (labelled with letters A, B, C, etc. directly below each chromosome) and divisions (labelled with numbers 1-46 below the subdivision labels). Scaffolds from the AfunF1 (KB66XXXX identifiers, grey font and thin horizontal lines) and AfunIP (scaffoldXX identifiers, black font and thick horizontal lines) assemblies are ordered along the photomap above each chromosome. Orientation of the scaffolds in the genome, if known, is shown by the arrows below each of the scaffold identifiers. Known polymorphic inversions are shown for chromosomal arms 2R, 3R, and 3L.

New reference genome assemblies for 20 anophelines

The consensus sets of synteny-based adjacencies for all species, validated with and complemented by physical mapping and/or RNAseq data for subsets of the anophelines, enabled improved genome assemblies to be generated by building superscaffolds from all scaffold neighbours (**Fig. 6**). This involved the design and implementation of a reconciliation workflow to integrate the different sets of scaffold adjacencies from synteny, physical mapping, AGOUTI, or alignment data for each assembly (**see Methods; Supplementary Online Material; Supplemental Fig. S10**). The total span of scaffolds that now form part of superscaffolds is more than 150 Mbp for seven assemblies and between 80 and 150 Mbps for another seven assemblies, although their actual contiguity levels remain variable (e.g. *A. atroparvus* in 34 superscaffolds and *A. funestus* in 114 superscaffolds). The greatest reductions in the total numbers of scaffolds were achieved for *A. christyi*, *A, culicifacies*, *A. maculatus*, and *A. melas*, with the largest improvements in scaffold N50 values observed for *A. atroparvus*, *A, dirus*, and *A. minimus* (**Table 1**). Given the heterogeneity of the input assemblies the relative changes highlight some of the most dramatic improvements, e.g. the *A. funestus* and *A. stephensi* (SDA-500) scaffold counts both dropped by almost 22% and N50s increased 4.0-fold for *A. atroparvus*, 3.1-fold for *A. funestus*, and 2.4-fold for *A. stephensi* (Indian) (**Table 1**). The largest relative improvements seen for *An. farauti* and *An. merus* are mainly from the additional "Fosill"-based scaffolded version 2 assemblies.

For the six anophelines with physical mapping data, the contributions of the synteny-based and/or AGOUTI-based adjacencies to the numbers and genomic spans of anchored scaffolds are largest for *A. stephensi* (SDA-500) and *A. funestus*, but negligible or low for the recently updated *A. albimanus* (Artemov et al. 2017), *A. atroparvus* (Artemov et al. 2018), and *A. sinensis* (Chinese) (Wei et al. 2017) assemblies (**Table 2**). For *A. albimanus*, reconciling the AalbS2 assembly (Artemov et al. 2017) with the AalbS1

adjacencies showed that the single neighbouring pair from the two-way consensus (and two of the three pairs unique to ORTHOSTITICH) agreed with the physical mapping data and were thus already incorporated into the new assembly, so only two short scaffolds from AGOUTI were added. Reconciling the *A. atroparvus* AatrE3 assembly (Artemov et al. 2018) with the AatrE1 adjacencies showed that three AGOUTI-based adjacencies were already incorporated, as were five two-way consensus synteny-based adjacencies, as well as two, one, and one, unique to the ADSEQ, GOS-ASM, and ORTHOSTITCH sets, respectively. These, and the 41 other adjacencies that did not involve any of the 46 physically mapped scaffolds, together led to a reduction of the initial AatrE1 scaffold number by 5.3% and a 4.0-fold increase in scaffold N50 for the new superscaffolded assembly. For the other anophelines with physical mapping data, the two *A. stephensi* assemblies achieved total assembly anchoring of 79% (improvements of 17% and 38%) and *A. funestus* more than doubled to reach 82% (**Table 2**). These chromosomally-anchored scaffolds and superscaffolds, together with all of the new improved genome assemblies have been submitted to VECTORBASE for processing and incorporation as new reference assemblies for the benefit of the entire research community.

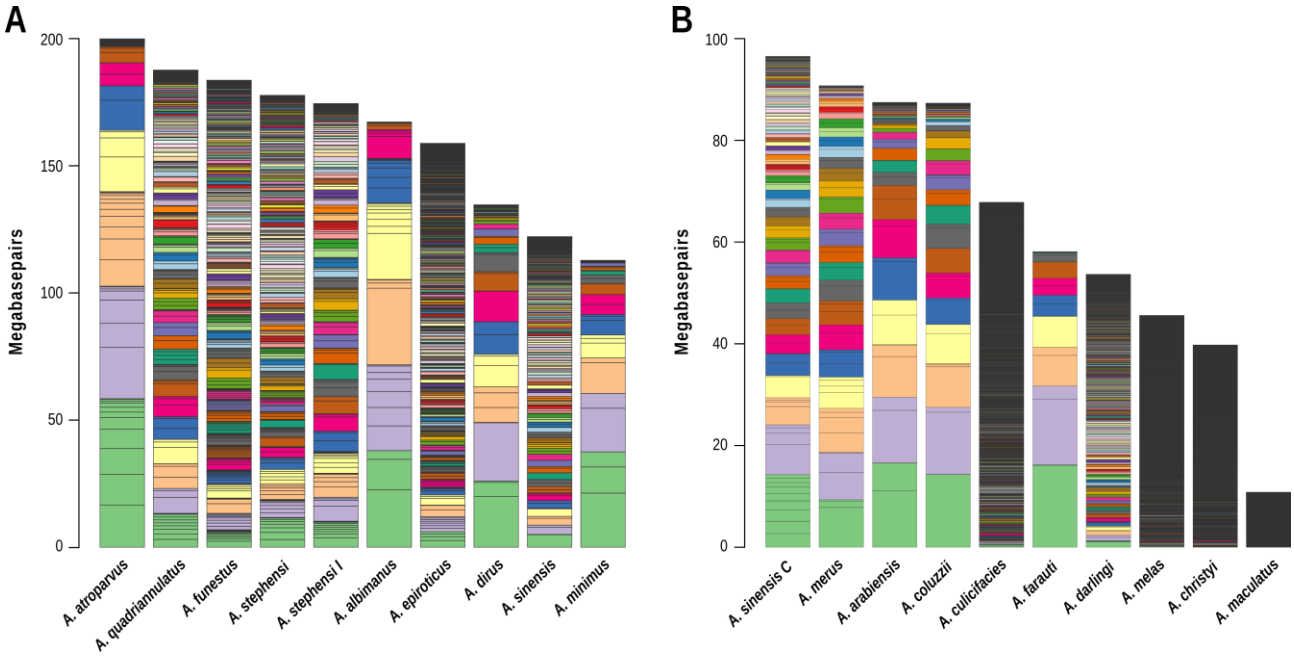


Figure 6. Resulting superscaffolds and their total genomic spans for each of the 20 *Anopheles* assemblies. Assemblies are ordered from the largest total genomic span (left) to the smallest (right), and sets of 10 assemblies are plotted separately in (***A***) and (***B***) with y-axis maxima of 200 and 100 megabasepairs, respectively. Superscaffolds are shown as stacked bars of the same colour with grey lines indicating the sizes of their constituent scaffolds, and with superscaffolds and scaffolds ordered from the largest (bottom) to the smallest (top). Stacked superscaffolds are distinguished by colours delineated from a palette with an ordered set of 60 contrasting hues, which are repeated for assemblies with more than 60 superscaffolds.

Table 1. Summary statistics of scaffold counts and N50 values of the 20 input and improved *Anopheles* assemblies after applying synteny (SYN), or AGOUTI (AGO), or physical mapping (PHY), or PacBio sequencing (PB) approaches.

Species	Input Assemblies			Approaches Applied	New Assemblies				
	Assembly Version	Number of Scaffolds	Scaffold N50 (Kbp)		Assembly Version	Number of Scaffolds [% reduced]		Scaffold N50 (Kbp) [Fold increase]	
A. albimanus	AalbS1	204	18'068	SYN+AGO+PHY	AalbS3§	203	[0.0]	33'601	[1.9]
A. arabiensis	AaraD1	1'214	5'604	SYN+AGO	AaraD2	1'160	[4.4]	6'693	[1.2]
A. atroparvus	AatrE1	1'371	9'207	SYN+AGO+PHY	AatrE4§	1'297	[5.4]	37'151	[4.0]
A. christyi	AchrA1	30'369	9	SYN	AchrA2	28'853	[5.0]	10	[1.1]
A. coluzzii	AcolM1	10'521	4'437	SYN	AcolM2	10'440	[0.8]	4'778	[1.1]
A. culicifacies	AculA1	16'162	22	SYN	AculA2	14'593	[9.7]	29	[1.3]
A. darlingi	AdarC3	2'221	115	SYN	AdarC4	1'838	[17.2]	159	[1.4]
A. dirus	AdirW1	1'266	6'906	SYN+AGO	AdirW2	1'211	[4.3]	12'741	[1.8]
A. epiroticus	AepiE1	2'673	367	SYN+AGO	AepiE2	2'254	[15.7]	814	[2.2]
A. farauti	AfarF1	550	12'895	SYN+AGO	AfarF3§	299	[45.5]	15'480	[12.9]
A. funestus	AfunF1	1'392	672	SYN+AGO+PHY+PB	AfunF2	1'090	[21.7]	2'051	[3.1]
A. maculatus	AmacM1	47'797	4	SYN	AmacM2	46'342	[3.0]	4	[1.0]
A. melas	AmelC1	20'281	18	SYN	AmelC3§	18'604	[8.0]	21	[1.2]
A. merus	AmerM1	2'753	342	SYN+AGO	AmerM3§	1'976	[28.2]	1'896	[5.5]
A. minimus	AminM1	678	10'313	SYN+AGO	AminM2	652	[3.8]	15'145	[1.5]
A. quadriannulatus	AquaS1	2'823	1'641	SYN+AGO	AquaS2	2'617	[7.3]	2'675	[1.6]
A. sinensis	AsinS2	10'448	579	SYN+AGO	AsinS3	10'136	[3.0]	638	[1.1]
A. sinensis (Chinese)	AsinC2	9'592	814	SYN+PHY	AsinC3	9'482	[1.1]	1'025	[1.3]
A. stephensi	AsteS1	1'110	837	SYN+AGO+PHY	AsteS2	870	[21.6]	1'788	[2.1]
A. stephensi (Indian)	AsteI2	23'371	1'591	SYN+AGO+PHY	AsteI3	23'050	[1.4]	3'775	[2.4]

§ New assemblies built from adjacencies of input assembly versions via reconciliation with updated assembly versions: physical mapping improvements for AalbS2, AatrE2, & AatrE3, additional "Fosill"-based scaffolding for AfarF2 & AmerM2, and haplotype removal for AmelC2.

Table 2. Summary of scaffold counts and genomic spans added to the initial physical maps from synteny-based (SYN) and AGOUTI-based (AGO) adjacencies, and counts of physically mapped (PHY) scaffolds that gained oriented neighbours after incorporating the SYN and AGO adjacencies.

Assembly	Physically mapped scaffolds	Scaffolds added to physical map by:			Total scaffolds added	PHY scaffolds now with oriented neighbours	Total basepairs added	% of assembly added	Total % of assembly anchored
		Synteny	AGOUTI	SYN+AGO					
A. albimanus	31	0	2	0	2	0	2'160	0.00	98.28
A. atroparvus	46	5	7	3	9	0	870'748	0.39	89.12
A. funestus	204	89	44	34	99	85	26'131'556	11.60	81.53
A. sinensis (Chinese)	52	18	NA	NA	18	14	5'791'225	2.62	43.72
A. stephensi	99	102	52	45	109	81	47'397'794	21.03	78.85
A. stephensi (Indian)	118	76	47	33	90	93	10'975'818	4.96	78.82

Discussion

Applying synteny-based scaffold adjacency predictors across 21 *Anopheles* genome assemblies resulted in almost 43'000 predicted adjacencies. The identification of consensus predictions produced supported subsets that were used to build improved assemblies for which the general trend showed that a reduction in the total number of orthologue-bearing scaffolds of about a third could double the scaffold N50 (**Fig. 1**). Notably, when the scaffolds involved were long, even a handful of adjacencies could greatly increase the N50 value, however, the numerous adjacencies for the rather fragmented input assemblies improved their contiguity but led to only minor improvements in N50 values. For the six assemblies with starting N50 values of between 340 Kbp and 840 Kbp (considering all scaffolds, not only those with orthologues), the average improvement was just less than 400 Kbp, demonstrating what can be achieved using only synteny-based approaches. By way of comparison, the honeybee genome assembly upgrade relied on millions of reads from ~20X SOLiD and ~5X Roche 454 sequencing to improve the scaffold N50 by 638 Kbp from 359 Kbp to 997 Kbp (Elsik et al. 2014). Thus, while the *Anopheles* results varied considerably depending on the input assemblies, using only synteny-based adjacencies from a combined analysis of the results from three methods achieved substantial contiguity improvements for many assemblies.

ADSEQ predicted about twice as many adjacencies as GOS-ASM and ORTHOSTITCH, providing support for about two thirds of each, while ORTHOSTITCH supported about a third and GOS-ASM supported slightly less than a third of each of the other prediction sets (**Fig. 2**). Thus, the more conservative GOS-ASM and ORTHOSTITCH results do not show a substantially greater overlap than either of them do with ADSEQ. Only 10% of all distinct scaffold adjacencies were predicted by all three methods, but building the two-way consensus adjacency sets increased this agreement to 33%, where almost all were supported by ADSEQ, nearly three-quarters by ORTHOSTITCH, and three-fifths by GOS-ASM. Consensus building can therefore achieve the goal of identifying a subset of well-supported adjacencies. Considering each of the assemblies individually, these two-way consensus adjacencies made up at least half of the distinct predictions for eight assemblies, with a maximum of 58% for *A. funestus*. While levels of agreement were generally lower for fragmented assemblies with many predicted adjacencies, some others with many fewer predicted adjacencies also showed below-average agreements. Thus the number of predicted adjacencies is not necessarily a reliable indicator of the level of agreement that can be achieved.

These results highlight the challenge of inferring accurate adjacencies as well as the importance of employing multiple approaches. Synteny block delineation, which then allows for scaffold adjacencies to be predicted, is itself a complex task where results from different anchor-based approaches can vary

considerably (Liu et al. 2018). Several key differences distinguish the three methods applied to the *Anopheles* assemblies, for example, GOS-ASM employs only single-copy orthologues as anchors so any gene duplications are excluded from the ancestral genome reconstructions, whereas the other two methods do take paralogues into account. Furthermore, both GOS-ASM and ADSEQ are 'phylogeny-aware' algorithms as they use the species tree topology, and ADSEQ additionally employs individual gene trees for each orthologous group. In contrast, ORTHOSTITCH does not take the species tree or gene phylogenies into account and instead relies on enumerating levels of support across the dataset to score putative adjacencies. These differences affect the sensitivity and specificity of each method, reflected by the more numerous predictions from ADSEQ that can explore complex gene evolutionary histories within the species tree topology (e.g. identifying adjacencies supported only within sub-clades of the phylogeny), versus the smaller sets of adjacencies from GOS-ASM, which excludes complexities introduced by gene duplications, and ORTHOSTITCH that simplifies the search by not imposing any evolutionary model. So while the consensus approach applied to the predictions across all the anophelines results in reduced sensitivities, it takes advantage of the different underlying assumptions and algorithmic implementations of each method to identify well-supported sets of scaffold adjacencies.

Another factor that may influence the number of predicted adjacencies, the level of agreement amongst different methods, and the resulting contiguity improvements, is the number of scaffolds with annotated orthologues, i.e. scaffolds used as input rather than the total number of scaffolds in an assembly (e.g. for *A. stephensi* (Indian) there were 23'371 scaffolds but only 660 with orthologues). The quality of the gene annotations themselves also likely plays a role, as fragmented gene models make orthology delineation difficult and hence reduce the ability to identify synteny blocks. Nevertheless, substantial improvements can be achieved even for relatively fragmented assemblies so long as gene models are nonetheless mostly complete. The evolutionary divergence of the set of species, as well as the total number of species, to which these methods are applied would also impact their ability to recover reliable adjacencies, because the complexity of the task of inferring synteny blocks is greatly reduced if the input orthology dataset consists mainly of near-universal single-copy orthologues. As gene duplications and losses accumulate over time the proportion of near-universal single-copy orthologues will shrink, and even amongst those that are maintained translocations and genomic shuffling events will add to the steady erosion of the evolutionary signals on which these methods rely. Shuffling rates may also vary in different lineages—e.g. lepidopteran genomes appear to have reduced levels of gene rearrangements (Kanost et al. 2016)—so seemingly equally divergent (in terms of time to last common ancestor) sets of species may be differentially amenable to synteny delineation.

The availability of alternative datasets with scaffold adjacency information for subsets of the anophelines provided the opportunity to validate the predictions based solely on synteny inferences. These included

physical mapping data, RNAseq data, a re-assembly incorporating PacBio sequencing data, and two re-scaffolded assemblies using extra 'Fosill' sequencing libraries. Although generally few scaffold adjacencies were obtained from the physical mapping data (because only larger scaffolds were selected for mapping they may not be direct neighbours) the comparisons were able to identify support for many synteny-based adjacencies (**Fig. 3A**). Several conflicts were also identified; however, most of these were due to the fact that the synteny-based neighbour was a short scaffold that had not been targeted for physical mapping and could be positioned between the two much larger physically mapped scaffolds, thus, they are not truly conflicts. Importantly, other conflicts involved only the relative orientation of neighbouring scaffolds and occurred with scaffolds that were anchored with only a single FISH probe and whose orientations had thus not been confidently determined. In these cases the synteny-based adjacencies therefore provided key complementary information and helped to correct the orientations of the physically mapped scaffolds.

Validations with RNAseq-based AGOUTI-predicted adjacencies also provided support for many synteny-based predictions (**Fig. 3B**). Two-thirds of the adjacencies unique to AGOUTI were between scaffolds where one or both scaffolds had no annotated orthologues. As AGOUTI is not restricted to large scaffolds preferred for physical mapping or scaffolds with annotated orthologues required for synteny-based approaches, it can provide complementary predictions that capture shorter unannotated scaffolds that would otherwise not be recovered. While this would not substantially improve N50 values it is nonetheless important for improving gene annotations as correcting such assembly breaks could allow for more complete gene models to be correctly identified.

The *A. funestus* PacBio-based AfunIP assembly scaffolds allowed for the alignment-based ordering and orientation of AfunF1 scaffolds for comparisons with the adjacency predictions and physical mapping data (**Fig. 4, Fig. 5**). These supported up to almost a quarter of *A. funestus* two-way consensus synteny adjacencies and about 40% of the physical mapping adjacencies. Importantly, most were neither supported nor in conflict, and conflicts generally occurred when the alignment-based adjacencies included short scaffolds that were not considered by the synteny-based or physical mapping approaches, and thus could be resolved. For *A. farauti* and *A. merus*, the genome-alignment-based comparisons of their initial assemblies with the re-scaffolded AfarF2 and AmerM2 assemblies provided much higher levels of support for the two-way consensus synteny adjacencies, with very few conflicts. This reflects the radically different approaches between re-scaffolding, where the additional 'Fosill' library data served to build longer scaffolds from the initial scaffolds, versus the re-assembly of *A. funestus* where PacBio sequencing data were first merged with the initial assembly before re-scaffolding with the original Illumina data to produce the hybrid AfunIP assembly. These comparisons therefore validate many of the

synteny-based adjacency predictions while conceding that short intervening scaffolds may be overlooked due to the limitations of having to rely on scaffolds with annotated orthologues.

As modern long-read sequencing technologies are capable of producing more contiguous assemblies than those based mainly on assembling short-reads, it is conceivable that many fragmented draft genomes will be completely superseded by new independently built high-quality reference assemblies. For example, single-molecule sequencing technologies were recently employed to produce assemblies of 15 *Drosophila* species, 14 of which already had previously reported sequenced genomes (Miller et al. 2018). Alternatively, re-sequencing to obtain proximity data to use in conjunction with contigs from draft assemblies can also achieve high-quality references to replace the fragmented initial versions, e.g. (Putnam et al. 2016; Dudchenko et al. 2017). Furthermore, although reference-assisted assembly approaches may mask true genomic rearrangements (Liu et al. 2018), high-quality chromosomal-level genomes of very close relatives can be used to improve draft assemblies, often employing alignment-based comparisons such as assisted assembly tools (Gnerre et al. 2009), reference-assisted chromosome assembly (Kim et al. 2013), CHROMOSOMER (Tamazian et al. 2016), or the RAGOUT 2 reference-assisted assembly tool (Kolmogorov et al. 2018). What role then is there for comparative genomics approaches that use evolutionary signals to predict scaffold adjacencies in draft assemblies?

Firstly, while recognising that downward trending costs of many new technologies are making sequencing-based approaches more accessible to even the smallest of research communities, the costs associated with experimental finishing or re-sequencing efforts remain non-trivial and acquired expertise is required for high-quality sample preparation and library building. Furthermore, the disappointing reality is that re-sequencing and re-scaffolding does not always lead to vastly improved assemblies, albeit an anecdotal reality because failures are not reported in the published literature. Secondly, hybrid assembly approaches benefit from the complementarity of the different types of input data that they employ, and our validations and comparisons show that synteny-based adjacencies can further complement the experimental data. In this regard, even if synteny-based results are not directly included in such hybrid approaches, they can nevertheless serve as a benchmark against which to quantify the effectiveness of different combinations of approaches (or different parameters used) and help guide re-assembly procedures towards producing the best possible improved assemblies. Thirdly, our results show how physical mapping datasets can be augmented or even corrected through comparisons with synteny-based scaffold adjacency predictions. Where subsets of scaffolds have already been mapped to chromosomes, adding neighbouring scaffolds from synteny-based predictions can add to the overall total proportion anchored without more labour-intensive experimental work. Additionally, as physical and genetic mapping of a scaffold requires the presence of at least one marker, and confirming the correct orientation of a localised scaffold requires at least two markers, draft assemblies with many short scaffolds require

much higher densities of markers in order to achieve near-complete chromosomal-level anchoring. Thus synteny-based predictions of scaffold adjacencies which reduce the total numbers of scaffolds to be mapped will allow for greater proportions of fragmented genome assemblies to be anchored using fewer markers.

Our three-method synteny-based scaffold adjacency prediction workflow is relatively easily implemented and may flexibly include results from additional adjacency predictors or, as evidenced with our various types of validation datasets, alternative sources of adjacency information (**Supplemental Fig. S10**). Rather than prescribing a panacea to cure all assembly ailments, we conclude that the components of this workflow may be adapted, substituted, extended, or simplified according to the needs and resources of any given draft genome assembly improvement project. Assessing the performance of three comparative genomics approaches and comparing their results with available experimental data demonstrates their utility as part of assembly improvement initiatives, as well as highlighting their complementarity to experimental approaches. The consensus predicted scaffold adjacencies can lead to substantial improvements of draft assemblies (**Fig. 1; Fig. 6; Table 1**) without requiring additional sequencing-based support, and they can add to and improve physical mapping efforts (**Table 2**). These evolutionarily guided methods therefore augment the capabilities of any genome assembly toolbox with approaches to assembly improvements or validations that will help draft assemblies mature into high-quality reference genomes.

Methods

Synteny-based scaffold adjacency predictions

The synteny-based prediction tools require as input both delineated orthology and genomic location data for the annotated genes from each assembly. All gene annotations were retrieved from VECTORBASE (Giraldo-Calderón et al. 2015) and orthology data was retrieved from ORTHODB V9 (Zdobnov et al. 2017): versions of the genome assemblies and their annotated gene sets are detailed in **Supplemental Table S1**, along with counts of scaffolds, genes, and orthologues. The complete 'frozen' input datasets of orthology relationships and genomic locations of the annotated genes for each of the 21 assemblies are presented in **Supplementary Dataset SD1**. ADSEQ analysis first builds reconciled gene trees for each orthologous group (gene family), then for pairs of gene families for which extant genomic adjacencies are observed, or suggested by sequencing data, a duplication-aware parsimonious evolutionary scenario is computed, via Dynamic Programming (DP), that also predicts extant adjacencies between genes at the extremities of contigs or scaffolds. This DP algorithm also accounts for scaffolding scores obtained from paired-end reads mapped onto contigs and provides a probabilistic score for each predicted extant adjacency, based on sampling optimal solutions (Anselmetti et al. 2018). ADSEQ was applied across the full anopheline input dataset to predict scaffold adjacencies (**Supplemental Table S2**). GOS-ASM employs an evolutionary rearrangement analysis strategy on multiple genomes utilizing the topology of the species phylogenetic tree and the concept of the breakpoint graph (Aganezov and Alekseyev 2016). Fragmented genomes with missing assembly 'links' between assembled regions are modelled as resulting from artificial 'fissions' caused by technological fragmentation that breaks longer contiguous genomic regions (chromosomes) into scaffolds. Assembling these scaffolds is therefore reduced to a search for technological 'fusions' that revert non-evolutionary 'fissions' and glue scaffolds back into chromosomes. GOS-ASM was applied to the full anopheline input dataset to predict such scaffold 'fusions' (**Supplemental Table S2**). The ORTHOSTITCH approach was first prototyped as part of the investigation of greater synteny conservation in lepidopteran genomes (Kanost et al. 2016), and subsequently further developed as part of this study to include a scoring system and additional consistency checks. Searches are performed to identify orthologues (both single-copy and multi-copy orthologues are considered) at contig or scaffold extremities in a given assembly that form neighbouring pairs in the other compared assemblies, thereby supporting the hypothesis that these scaffolds should themselves be neighbours. ORTHOSTITCH was applied to the full anopheline input dataset to predict scaffold adjacencies (**Supplemental Table S2; Supplemental Fig. S1**). The CAMSA tool (Aganezov and Alekseyev 2017) was used to compare and merge scaffold assemblies produced by the three methods by identifying adjacencies in three-way and two-way agreement (with no third method conflict) (**Supplemental Table S3**). CAMSA was also used to build merged assemblies using only conservative

three-way consensus adjacencies, and using liberal unions of all non-conflicting adjacencies. Quantifications of assembly improvements considered only scaffolds with annotated orthologous genes (because the synteny-based methods rely on orthology data) to count the numbers of scaffolds and compute scaffold N50 values before and after merging (**Fig. 1; Supplemental Figs. S2, S3**). The results of the CAMSA merging procedure were used to quantify all agreements and conflicts amongst the different sets of predicted adjacencies (**Fig. 2; Supplemental Table S3; Supplemental Figs. S4, S5**). A DOCKER container is provided that packages ADSEQ, GOS-ASM, ORTHOSTITCH, and CAMSA, as well as their dependencies, in a virtual environment that can run on a Linux server. See **Supplementary Online Material** for additional details for all synteny-based predictions and their comparisons, and the DOCKER container.

Validations with physical mapping and RNA sequencing data

Methods for chromosomal mapping of scaffolds are detailed for *A. albimanus* (Artemov et al. 2017), *A. atroparvus* (Artemov et al. 2015; Neafsey et al. 2015; Artemov et al. 2018), *A. stephensi* (SDA-500) (Neafsey et al. 2015), *A. stephensi* (Indian) (Jiang et al. 2014), and *A. sinensis* (Chinese) (Wei et al. 2017). *A. funestus* mapping built on previous results (Sharakhov et al. 2002, 2004; Xia et al. 2010) with additional FISH mapping (**Supplemental Fig. S6**) to further develop the physical map by considering several different types of mapping results. The complete 'frozen' input datasets of the physically mapped scaffolds for each of the six assemblies are presented in **Supplementary Dataset SD2**, with usable scaffold pair adjacencies in **Supplemental Table S4**, and the final mapped *A. funestus* scaffolds in **Supplemental Table S5**. These adjacencies were compared with the CAMSA-generated two-way consensus assemblies, as well as the predictions from each method and the conservative and liberal consensus assemblies (**Fig. 3A; Supplemental Table S6**). The RNAseq validations used genome-mapped paired-end sequencing data for 13 of the anophelines available from VECTORBASE (Giraldo-Calderón et al. 2015) (Release VB-2017-02), including those from the *Anopheles* 16 Genomes Project (Neafsey et al. 2015) and an *A. stephensi* (Indian) male/female study (Jiang et al. 2015). AGOUTI (Zhang et al. 2016) analyses were performed to identify transcript-supported scaffold adjacencies for these 13 anophelines (**Supplemental Table S7**). These adjacencies were compared with the CAMSA-generated two-way consensus assemblies, as well as the predictions from each method and the conservative and liberal consensus assemblies (**Fig. 3B; Supplemental Table S8**). See **Supplementary Online Material** for additional details for physical mapping and AGOUTI adjacencies and their comparisons.

Building the new assemblies

The new *A. funestus* assembly generated as part of this study was based on approximately 70X of PacBio sequencing data polished with QUIVER (from PacBio's SMRT Analysis software suite). This was

combined with the reference assembly (AfunF1) using METASSEMBLER (Wences and Schatz 2015) to generate a merged assembly, and this merged assembly was then scaffolded with SSPACE (Boetzer et al. 2011) using the original Illumina sequencing data, and designated the *A. funestus* AfunIP assembly. The AfunIP assembly improves on the reference AfunF1 assembly at contig level but not at scaffold level (**Supplemental Fig. S7; Supplemental Table S9**). Where AfunIP scaffolds span the ends of AfunF1 scaffolds they provide support for AfunF1 scaffold adjacencies. Thus, whole genome alignments of the two assemblies were performed using LASTZ (Harris 2007) and used to identify corresponding genomic regions that enabled the alignment-based ordering and orientation of AfunF1 scaffolds, which were then compared with the synteny-based, physical mapping based, and AGOUTI-based, adjacencies (**Fig. 4, Supplemental Fig. S8; Supplemental Table S10**). Using the AfunF1 assembly as the basis, and incorporating evidence from the AfunIP assembly through scaffold correspondences established from the whole genome alignments, the physical mapping data and the synteny-based and AGOUTI-based adjacency predictions were integrated to build the new reference assembly for *A. funestus*. The comprehensive update to the photomap employed BLAST searches to identify positions of the physically mapped DNA markers within the AfunF1 and AfunIP assemblies, and whole genome pairwise alignments to reconcile these two assemblies with the new photomap. Whole genome alignments of versions 1 and 2 assemblies for *A. farauti*, and *A. merus* were used to delineate corresponding scaffolds and identify supported, unsupported, and conflicting adjacencies (**Supplemental Fig. S9; Supplemental Table S11**). The new assemblies were built using the different datasets available for each of the anophelines (**Supplemental Fig. S10**): synteny data only for six, *A. christyi*, *A. coluzzii*, *A. culicifacies*, *A. darlingi*, *A. maculatus*, & *A. melas*; synteny and AGOUTI data for eight, *A. arabiensis*, *A. dirus*, *A. epiroticus*, *A. farauti*, *A. merus*, *A. minimus*, *A. quadriannulatus*, and *A. sinsensis* (SINENSIS); synteny and physical mapping data for *A. sinensis* (Chinese); synteny, AGOUTI, and physical mapping data for four, *A. albimanus*, *A. atroparvus*, *A. stephensi* (SDA-500), *A. stephensi* (Indian); and synteny, AGOUTI, physical mapping data, and the new PacBio-based assembly for *A. funestus*. See **Supplementary Online Material** for additional details on the PacBio assembly generation, the genome alignment based comparisons of the AfunF1 and AfunIP assemblies, and the workflow to integrate different adjacency predictions and build the new assemblies.

Data Access

The updated assemblies of 20 anophelines have been submitted to VECTORBASE (www.vectorbase.org) and will become available as the new reference assemblies. The input data for the synteny analyses of orthology relationships and genomic locations of the annotated genes are presented in **Supplementary Dataset SD1**. The complete input datasets of the physically mapped scaffolds for each of the six

assemblies are presented in **Supplementary Dataset SD2**. The final sets of scaffold adjacencies and superscaffolds for all assemblies are presented in **Supplementary Dataset SD3**.

Acknowledgements

Physical mapping and PacBio sequencing of *A. funestus* were supported by the United States (US) National Institutes of Health (NIH) National Institute of Allergy and Infectious Diseases (NIAID) grant R21 AI112734 to NJB, with SJE and IVS as co-investigators. Physical mapping for *A. stephensi* and *A. albimanus* was supported by the US NIH NIAID grant R21 AI099528 and the US Department of Agriculture National Institute of Food and Agriculture Hatch project 223822 to IVS, and for *A. atroparvus* by grant number 15-14-20011 from the Russian Science Foundation to IVS. The authors acknowledge Marcia Kern for technical assistance with the physical mapping, and Vasily Sitnik for assistance with data submission to VectorBase. SA and MAA were supported by the US National Science Foundation (NSF) grant IIS-1462107. SA was supported by the US NSF grants CCF-1053753 and DBI-1350041 and by US NIH grants U24CA211000 and R01-HG006677. YA, SB and ET were supported by the French *Agence Nationale pour la Recherche* Ancestrome project ANR-10-BINF-01-01. ADSEQ data computation and analyses benefited from the Montpellier Bioinformatics Biodiversity platform service. SK and AMP were supported by the Intramural Research Program of the NIH National Human Genome Research Institute 1ZIAHG200398. CC was supported by a Mitacs Globalink grant, the Natural Sciences and Engineering Research Council of Canada Discovery Grant RGPIN-249834, and a resource allocation from Compute Canada. MWH and SVZ were supported by US NSF grant DEB-1249633. RMW, LR, MJMFR, and RF were supported by Swiss National Science Foundation grant PP00P3_170664.

Author Contributions

RMW and IVS conceived the study. SA and MAA developed and implemented GOS-ASM and CAMSA. YA, SB, ET and CC developed and implemented ADSEQ. RMW developed and implemented ORTHOSTITCH. SJE contributed to synteny-based analyses. JL, PG, MK, AP, MVS, MFU and IVS carried out physical mapping experiments. RMW, MWH and SVZ performed AGOUTI analyses. RMW, SA, LR and MJMFR compared synteny-based with physical mapping and AGOUTI adjacencies. PacBio *funestus* sequencing and data production: PIH, SJE and NJB; assembly: SK and AMP; and assembly comparisons: RMW, JL, MVS and IVS. Reconciliation to produce final assemblies: RMW, JL, LR, MJMFR, RF, DL, GM, VS and IVS. Docker container: MJMFR. The manuscript was written by RMW with input from all authors.

Disclosure declaration

All authors declare no conflicts of interest.

References

Aganezov S, Alekseyev MA. 2016. Multi-genome scaffold co-assembly based on the analysis of gene orders and genomic repeats. In *Lecture Notes in Computer Science*, Vol. 9683 of, pp. 237–249, Springer, Cham.

Aganezov SS, Alekseyev MA. 2017. CAMSA: a tool for comparative analysis and merging of scaffold assemblies. *BMC Bioinformatics* **18**: 496.

Ahola V, Lehtonen R, Somervuo P, Salmela L, Koskinen P, Rastas P, Välimäki N, Paulin L, Kvist J, Wahlberg N, et al. 2014. The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nat Commun* **5**: 1–9.

Anselmetti Y, Duchemin W, Tannier E, Chauve C, Bérard S. 2018. Phylogenetic signal from rearrangements in 18 Anopheles species by joint scaffolding extant and ancestral genomes. *BMC Genomics* **19**: 96.

Artemov GN, Bondarenko SM, Naumenko AN, Stegniy VN, Sharakhova M V., Sharakhov I V. 2018. Partial-arm translocations in evolution of malaria mosquitoes revealed by high-coverage physical mapping of the Anopheles atroparvus genome. *BMC Genomics* **19**: 278.

Artemov GN, Peery AN, Jiang X, Tu Z, Stegniy VN, Sharakhova M V, Sharakhov I V. 2017. The physical genome mapping of Anopheles albimanus corrected scaffold misassemblies and identified interarm rearrangements in genus Anopheles. *G3 Genes|Genomes|Genetics* **7**: 155–164.

Artemov GN, Sharakhova M V, Naumenko AN, Karagodin DA, Baricheva EM, Stegniy VN, Sharakhov I V. 2015. A standard photomap of ovarian nurse cell chromosomes in the European malaria vector Anopheles atroparvus. *Med Vet Entomol* **29**: 230–237.

Bauman JGJ, Wiegant J, Borst P, van Duijn P. 1980. A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochrome-labelled RNA. *Exp Cell Res* **128**: 485–490.

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.

Damas J, O'Connor R, Farré M, Lenis VPE, Martell HJ, Mandawala A, Fowler K, Joseph S, Swain MT, Griffin DK, et al. 2017. Upgrading short-read animal genome assemblies to chromosome level using comparative genomics and a universal probe set. *Genome Res* **27**: 875–884.

Davey JW, Chouteau M, Barker SL, Maroja L, Baxter SW, Simpson F, Merrill RM, Joron M, Mallet J, Dasmahapatra KK, et al. 2016. Major improvements to the Heliconius melpomene genome assembly used to confirm 10 chromosome fusion events in 6 million years of butterfly evolution. *G3 Genes|Genomes|Genetics* **6**: 695–708.

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 (80-)* **356**: 92–95.

Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, de Graaf DC, Debyser G, Deng J, Devreese B, et al. 2014. Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC Genomics* **15**: 1–29.

Fierst JL. 2015. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. *Front Genet* **6**: 220.

Giraldo-Calderón GI, Emrich SJ, MacCallum RM, Maslen G, Dialynas E, Topalis P, Ho N, Gesing S, Madey G, Collins FH, et al. 2015. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res* **43**: D707-13.

Gnerre S, Lander ES, Lindblad-Toh K, Jaffe DB. 2009. Assisted assembly: how to improve a de novo genome assembly by using related species. *Genome Biol* **10**: R88.

Hahn MW, Zhang S V., Moyle LC. 2014. Sequencing, assembling, and correcting draft genomes using recombinant populations. *G3 Genes|Genomes|Genetics* **4**: 669–679.

Harris RS. 2007. Improved pairwise alignment of genomic DNA. The Pennsylvania State University.

Holt RA, Mani Subramanian G, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JMC, Wides R, et al. 2002. The genome sequence of the malaria mosquito Anopheles gambiae. *Science (80-)* **298**: 129–149.

Jiang X, Biedler JK, Qi Y, Hall AB, Tu Z. 2015. Complete dosage compensation in Anopheles stephensi and the evolution of sex-biased genes in mosquitoes. *Genome Biol Evol* **7**: 1914–1924.

Jiang X, Peery A, Hall AB, Sharma A, Chen X-G, Waterhouse RM, Komissarov A, Riehle MM, Shouche Y, Sharakhova M V, et al. 2014. Genome analysis of a major urban malaria vector mosquito, Anopheles stephensi. *Genome Biol* **15**: 459.

Jiao WB, Accinelli GG, Hartwig B, Kiefer C, Baker D, Severing E, Willing EM, Piednoel M, Woetzel S, Madrid-Herrero E, et al. 2017. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res* **27**: 778–786.

Kanost MR, Arrese EL, Cao X, Chen Y-RR, Chellapilla S, Goldsmith MR, Grosse-Wilde E, Heckel DG, Herndon N, Jiang HHH, et al. 2016. Multifaceted biological insights from a draft genome sequence of the tobacco hornworm moth, Manduca sexta. *Insect Biochem Mol Biol* **76**: 118–147.

Kaplan N, Dekker J. 2013. High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat Biotechnol* **31**: 1143–1147.

Kim J, Larkin DM, Cai Q, Asan, Zhang Y, Ge R-L, Auvil L, Capitanu B, Zhang G, Lewin HA, et al. 2013. Reference-

assisted chromosome assembly. *Proc Natl Acad Sci* **110**: 1785–1790.
Kolmogorov M, Armstrong J, Raney BJ, Streeter I, Dunn M, Yang F, Odom D, Flicek P, Keane TM, Thybert D, et al. 2018. Chromosome assembly of large and complex genomes using multiple references. *Genome Res*.
Lawniczak MK, Emrich SJ, Holloway AK, Regier AP, Olson M, White B, Redmond S, Fulton L, Appelbaum E, Godfrey J, et al. 2010. Widespread divergence between incipient Anopheles gambiae species revealed by whole genome sequences. *Science (80-)* **330**: 512–514.
Levy-Sakin M, Ebenstein Y. 2013. Beyond sequencing: optical mapping of DNA in the age of nanotechnology and nanoscopy. *Curr Opin Biotechnol* **24**: 690–698.
Liu D, Hunt M, Tsai IJ. 2018. Inferring synteny between genome assemblies: a systematic evaluation. *BMC Bioinformatics* **19**: 26.
Marinotti O, Cerqueira GC, de Almeida LG, Ferro MI, Loreto EL, Zaha A, Teixeira SM, Wespiser AR, Almeida E Silva A, Schlindwein AD, et al. 2013. The genome of Anopheles darlingi, the main neotropical malaria vector. *Nucleic Acids Res* **41**: 7387–7400.
Markelz RJC, Covington MF, Brock MT, Devisetty UK, Kliebenstein DJ, Weinig C, Maloof JN. 2017. Using RNA-Seq for genomic scaffold placement, correcting assemblies, and genetic map creation in a common Brassica rapa mapping population. *G3 Genes|Genomes|Genetics* **7**: 2259–2270.
Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, et al. 2017. A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**: 427–433.
Matthews BJ, Dudchenko O, Kingan S, Koren S, Antoshechkin I, Crawford JE, Glassford WJ, Herre M, Redmond SN, Rose NH, et al. 2017. Improved Aedes aegypti mosquito reference genome assembly enables biological discovery and vector control. *bioRxiv* 240747.
Miller DE, Staber C, Zeitlinger J, Hawley RS. 2018. High-quality genome assemblies of 15 Drosophila species generated using Nanopore sequencing. *G3 Genes|Genomes|Genetics* g3.118.200160.
Neafsey DE, Christophides GK, Collins FH, Emrich SJ, Fontaine MC, Gelbart W, Hahn MW, Howell PI, Kafatos FC, Lawson D, et al. 2013. The evolution of the Anopheles 16 genomes project. *G3 Genes|Genomes|Genetics* **3**: 1191–4.
Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, Amon J, Arcà B, Arensburger P, Artemov G, et al. 2015. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science (80-)* **347**: 1258522–1258522.
Peichel CL, Sullivan ST, Liachko I, White MA. 2017. Improvement of the threespine stickleback genome using a Hi-C-based proximity-guided assembly. *J Hered* **108**: 693–700.
Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, et al. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res* **26**: 342–350.
Ruzzante L, Reijnders MJMF, Waterhouse RM. 2018. Of genes and genomes: mosquito evolution and diversity. *Trends Parasitol*.
Sharakhov I, Braginets O, Grushko O, Cohuet A, Guelbeogo WM, Boccolini D, Weill M, Costantini C, Sagnon N, Fontenille D, et al. 2004. A microsatellite map of the African human malaria vector Anopheles funestus. *J Hered* **95**: 29–34.
Sharakhov I V, Serazin AC, Grushko OG, Dana A, Lobo N, Hillenmeyer ME, Westerman R, Romero-Severson J, Costantini C, Sagnon N, et al. 2002. Inversions and gene order shuffling in Anopheles gambiae and A. funestus. *Science (80-)* **298**: 182–185.
Sharakhova M V, Hammond MP, Lobo NF, Krzywinski J, Unger MF, Hillenmeyer ME, Bruggner R V, Birney E, Collins FH. 2007. Update of the Anopheles gambiae PEST genome assembly. *Genome Biol* **8**: R5.
Sim SB, Geib SM. 2017. A chromosome-scale assembly of the Bactrocera cucurbitae genome provides insight to the genetic basis of white pupae. *G3 Genes|Genomes|Genetics* **7**: 1927–1940.
Tamazian G, Dobrynin P, Krasheninnikova K, Komissarov A, Koepfli KP, O'Brien SJ. 2016. Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences. *Gigascience* **5**: 38.
Wei Y, Cheng B, Zhu G, Shen D, Liang J, Wang C, Wang J, Tang J, Cao J, Sharakhov I V., et al. 2017. Comparative physical genome mapping of malaria vectors Anopheles sinensis and Anopheles gambiae. *Malar J* **16**: 235.
Xia A, Sharakhova M V., Leman SC, Tu Z, Bailey JA, Smith CD, Sharakhov I V. 2010. Genome landscape and evolutionary plasticity of chromosomes in malaria mosquitoes ed. W.J. Murphy. *PLoS One* **5**: e10592.
Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simao FA, Ioannidis P, Seppey M, Loetscher A, Kriventseva E V. 2017. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res* **45**: D744–D749.
Zhang S V., Zhuo L, Hahn MW. 2016. AGOUTI: improving genome assembly and annotation using transcriptome data. *Gigascience* **5**: 31.
Zhou D, Zhang D, Ding G, Shi L, Hou Q, Ye Y, Xu Y, Zhou H, Xiong C, Li S, et al. 2014. Genome sequence of Anopheles sinensis provides insight into genetics basis of mosquito competence for malaria parasites. *BMC Genomics* **15**: 42.