

RESEARCH

Open Access



Phylogenetic signal from rearrangements in 18 *Anopheles* species by joint scaffolding extant and ancestral genomes

Yoann Anselmetti^{1,2}, Wandrille Duchemin^{2,3}, Eric Tannier^{2,3}, Cedric Chauve⁴ and Sèverine Bérard^{1*}

From The Sixteenth Asia Pacific Bioinformatics Conference 2018
Yokohama, Japan. 15-17 January 2018

Abstract

Background: Genomes rearrangements carry valuable information for phylogenetic inference or the elucidation of molecular mechanisms of adaptation. However, the detection of genome rearrangements is often hampered by current deficiencies in data and methods: Genomes obtained from short sequence reads have generally very fragmented assemblies, and comparing multiple gene orders generally leads to computationally intractable algorithmic questions.

Results: We present a computational method, ADSEQ, which, by combining ancestral gene order reconstruction, comparative scaffolding and *de novo* scaffolding methods, overcomes these two caveats. ADSEQ provides simultaneously improved assemblies and ancestral genomes, with statistical supports on all local features. Compared to previous comparative methods, it runs in polynomial time, it samples solutions in a probabilistic space, and it can handle a significantly larger gene complement from the considered extant genomes, with complex histories including gene duplications and losses. We use ADSEQ to provide improved assemblies and a genome history made of duplications, losses, gene translocations, rearrangements, of 18 complete *Anopheles* genomes, including several important malaria vectors. We also provide additional support for a differentiated mode of evolution of the sex chromosome and of the autosomes in these mosquito genomes.

Conclusions: We demonstrate the method's ability to improve extant assemblies accurately through a procedure simulating realistic assembly fragmentation. We study a debated issue regarding the phylogeny of the *Gambiae complex* group of *Anopheles* genomes in the light of the evolution of chromosomal rearrangements, suggesting that the phylogenetic signal they carry can differ from the phylogenetic signal carried by gene sequences, more prone to introgression.

Keywords: Scaffolding, Comparative genomics, Mosquito genomics

Background

The promises of using genes as evolutionary markers for phylogeny, introduced half a century ago by Zuckerkandl and Pauling [1], have been unfulfilled so far [2, 3]. At every scale of evolution, gene histories can differ from organisms tree-like phylogeny due to non-tree like evolutionary mechanisms such as incomplete lineage sorting,

horizontal gene transfer, hybridization, symbiosis, among others. This happens for example in the *Gambiae complex*, composed of several African *Anopheles* mosquito species, whose phylogeny is important to shed light on the origin of malaria transmission to humans [4], but is difficult to trace because of apparent extensive gene introgression within this complex [5, 6]. Chromosomal rearrangements have been recognized, for an even longer time [7], as valuable phylogenetic markers, due to several reasons, including their lower occurrence rate

*Correspondence: Severine.Berard@umontpellier.fr

¹ISEM, Université de Montpellier, CNRS, IRD, EPHE, Montpellier, France
Full list of author information is available at the end of the article

compared to sequence evolution. They have proved to be of great interest for understanding mosquito evolution for example [8, 9], due to the fact introgression of whole chromosomes is much less frequent than introgression of genes [10]. Moreover, in terms of functional and ecological implications, chromosomal rearrangements have also been shown to be involved in important adaptation processes [11–13].

However chromosome evolution is still challenging to study, especially from short reads sequence data, and current methods have severe limitations that we outline now. First, some methods are limited to consider only chromosomal regions which are highly similar and whose differences are detected by genetic mapping, polytene chromosome banding, *in situ* hybridization or targeted sequencing [8]. Among methods which can handle whole genome sequence data [14–19], some consider only a small number of markers (often genes) with simple evolutionary histories (typically duplication-free histories and one-to-one orthologous gene families), and most of them require fully assembled extant genomes, aside of the recently published method DESCHRAMBLER [20]. As a consequence existing methods are hardly applicable to currently available genomic data, characterized by very short sequencing reads that can not resolve genomic repeats [21, 22], leading to highly fragmented genome assemblies, often in the form of hundreds or thousands of contigs or scaffolds, where evolutionary breakpoints can not be distinguished from assembly artifacts.

Various types of data can help to improve the contiguity of genome assemblies obtained from short reads. For example Third-Generation Sequencing (TGS) technologies generate long, albeit noisy, sequencing reads that can resolve ambiguities due to repeats [23]; alternatively, chromosome conformation capture technologies [24, 25] or genome maps [26–28] have also been used successfully for scaffolding large genomes. However in the absence of long range sequence data or genome maps, the most widely used approach to scaffold contigs is the comparative approach, using one or several related genomes to guide the scaffolding. The principle of comparative scaffolding is to align contigs of a fragmented genome assembly onto one, or a set of, assembled reference genome(s) and to deduce contig adjacencies from the contiguity of the corresponding alignments along the reference(s). Most comparative scaffolding methods rely on a single reference genome, assumed to be closely related enough that contiguity along the reference can confidently imply contiguity in the newly assembled genome [29–35]. Such methods have mostly been used to assemble pathogen genomes using closely related assembled references, but have also been shown to be applicable in wider contexts, such as scaffolding an antelope genome using a cow genome as reference [34]. There exists few

methods that can handle several reference genomes at once, that can be distinguished between methods that do require that the phylogenetic relation between the considered genomes are provided [36–39] and methods that do not need such information [40]. Moreover, only two methods make use of sequencing data that might not appear in the contigs to be scaffolded but can provide a valuable scaffolding signal that complements the comparative signal [35, 37]. All such methods are also limited to handle contigs containing repeats and discard repeated contigs.

An important feature of most of these methods is that they assume a hypothesis of genome rearrangement parsimony or near-parsimony to transfer contiguity information from the reference(s) to the genome of interest, this hypothesis being either explicit [30, 31, 33, 38] or implicit [36, 39]. This points at the fact that the comparative approach is a kind of conundrum: to scaffold genomes, comparative methods rely, at least implicitly, on a framework to compare genomes and detect conserved synteny and chromosomal rearrangements, while whole genome evolution methods do not fare well when provided with fragmented genome assemblies.

We introduce a new computational method, ADSEQ, that addresses the issues raised above, regarding both genome evolution by rearrangements and comparative genome scaffolding; we apply it to simultaneously study the chromosome evolution and improve the scaffolding of 18 *Anopheles* genomes, 16 of them recently sequenced by Neafsey et al. [4], including several important malaria vector species. The method ADSEQ does not need a fully assembled reference genome, as is required by most comparative scaffolding methods, but takes as input a set, that can be arbitrarily large, of fragmented genome assemblies, together with a species phylogeny. It also takes advantage of sequencing data such as paired-end reads, for species for which it is available. From this input, ADSEQ computes ancestral genome segments, as well as extant scaffolding adjacencies. Additionally, ADSEQ allows the user to infer evolutionary scenarios in terms of gene duplication, gene loss, gene displacement and genome rearrangement along each branch of the species phylogeny. A Gibbs-Boltzmann probabilistic framework based on the cost of adjacencies gain and breaks in evolutionary scenarios provides a statistical support on all ancestral and extant inferred adjacencies, with sequencing data used to define a prior on extant gene adjacencies. To handle genes whose history involves duplication and loss, ADSEQ relies on the use of reconciled gene trees, in terms of gene duplication and losses, which allows to use a much larger gene set than existing comparative scaffolding methods relying on one-to-one orthologous genes or gene families with simple duplication/loss histories. We present, together with

the ADSEQ method, a validation procedure for the extant genome scaffolding aspect of ADSEQ relying on a realistic framework to generate artificially fragmented genome assemblies.

Using ADSEQ, we provide an analysis of whole genome evolution in a large set of *Anopheles* species with an unprecedented precision, being able to quantify duplications, losses, gene displacements between chromosomes, and chromosomal rearrangements. We work at a much finer scale than cytogenetics methods [41–47], rely on a larger gene complement than traditional genome rearrangement methods based on orthologous genes, and we provide a refined evolutionary analysis compared to [4] due to the improvement of extant genome scaffolding. In particular we find that gene displacements between chromosomes are much more frequent for genes belonging to families with duplication/loss histories, that previous studies handling only one-to-one orthologous genes had mostly overlooked. We also use our method to compare two alternative *Anopheles* phylogenies. We find that *Anopheles* genomes are compartmentalized between autosomes and sex chromosome according to duplications and chromosomal rearrangements, just as they were found to be according to gene sequences. We provide an alternative hypothesis to the conclusions of Fontaine et al. [5] about introgression of the major part of the genome. Indeed our measures of rearrangements and duplications are in favour of the phylogeny supported by most genes.

Methods

We first describe the main methodological contribution of our work, the ADSEQ tool, followed by its application to the specific *Anopheles* genomes data set we analyze in detail. We begin by introducing some simple but important terminology. A gene, extant or ancestral, is seen as a directed DNA segment with two *extremities*. Genes are parts of larger segments containing one or several genes, which are contig, super-contigs, scaffolds or even chromosomes for well assembled genomes. Two genes that are contiguous along such a segment define an *adjacency* between one extremity of each gene, that we call a *gene adjacency*. Thus *segments of genes*, either extant or ancestral, are encoded by linearly organized sets of gene adjacencies. For extant genomes, such segments, corresponding to contigs, scaffolds or chromosomes, are observed, while for ancestral genomes, segments are reconstructed and are to be considered as hypothetical as they are not directly supported by sequence data. Ancestral segments have previously been termed Contiguous Ancestral Regions (CARs) [19, 48, 49]. However, to stress the similarity between scaffold in extant genomes and CAR in ancestral genome, we use the generic terminology “segment” for both throughout this paper.

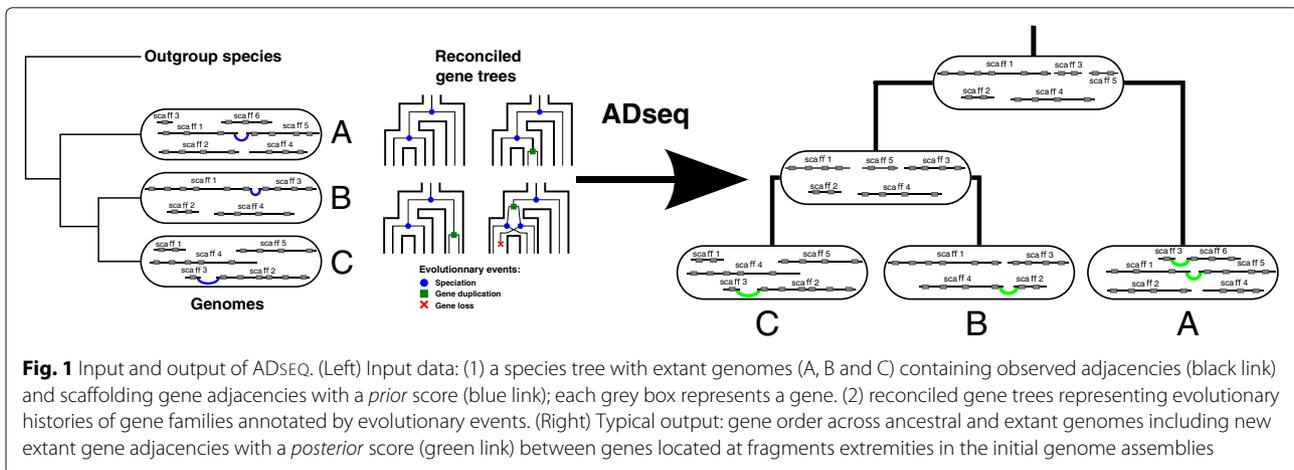
Assembly recovery through detection of Coevolution with sequencing data (ADSEQ)

ADSEQ builds upon a family of methods aimed at reconstructing the evolutionary history of gene adjacencies introduced with the DECO algorithm [50] and extended along several lines in [51, 52]. It is implemented within the package DECOSTAR [53]. The aim of ADSEQ is to jointly scaffold extant genomes and reconstruct ancestral gene orders, through the joint analysis of phylogenomics and sequencing data.

ADSEQ takes as minimal input a species tree and a set of extant genomes data: gene adjacencies, homologous gene families and their associated reconciled gene trees (see for example [54] for background on reconciled gene trees). Reconciled gene trees implicitly yield the gene content of ancestral genomes. A key feature of ADSEQ is that the extant gene adjacencies can originate from genomes in various state of assembly, from fully assembled genomes —where gene adjacencies encode the gene order along the chromosomes—, to ambiguously assembled genomes represented as scaffolding graphs, through fragmented genomes assembled into contigs or scaffolds. Each extant gene adjacency is assigned a *prior* score in $[0, 1]$, expected to represent the confidence that the adjacency actually occurs in the genome of interest. This prior can be obtained from sequencing data or genome maps for example; so adjacencies between genes in fully assembled genomes will have a high prior score, while a potential gene adjacency in a poorly assembled genome and that is not supported by a strong signal in terms of sequencing data will likely be assigned a low prior score.

ADSEQ processes independently all pairs of gene families for which at least one extant adjacency is observed between genes from these families. A *solution* of ADSEQ on such an instance is a set of extant and ancestral gene adjacencies between extant and ancestral genes of the two considered families, that are consistent with the given reconciled gene trees. Taking the obtained solutions over all pairs of gene families defining ADSEQ instances, ancestral adjacencies link ancestral genes extremities into ancestral genome segments, while extant adjacencies improve the scaffolding of the provided extant genomes and reduce their fragmentation. This, together with the reconciled gene trees, in turn provides an important input material for studying whole-genome evolution through mechanisms such as gene duplication, loss and transfer, introgression, or genome rearrangement. Figure 1 provides a high-level overview of the ADSEQ algorithm.

An important feature of ADSEQ is that, for each considered instance, it does not compute a single solution, but samples solutions from the, often large, associated search space. In order to sample solutions, ADSEQ relies on a *score*, function of the number of evolutionary events and the prior score of adjacencies. Any solution indeed yields



a number of gains and breakages of gene adjacencies, that model genome rearrangement events consistent with the provided reconciled gene trees (see a description of the propagation rules that allow to infer gains/breakages from reconciled gene trees in the DECO algorithm [50]). The *score* of a solution S is defined as $n(S) = g(S) + b(S) + c(S)$ where $g(S)$ is the number of gene adjacency gains scaled according to a user-defined unit cost of a gain; $b(S)$ is the number of gene adjacency breakages again scaled according to a user-defined unit cost; $c(S)$ is the cost of including or discarding extant adjacencies, based on their prior score: for an adjacency of *prior* score p , including the adjacency in a solution costs $-kT_0 \log(p)$ while discarding it from the solution costs $-kT_0 \log(1 - p)$, where kT_0 is a pseudo-temperature that we discuss in details in Additional file 1.

A polynomial time and space Dynamic Programming (DP) algorithm samples solutions for a given instance with a probability proportional to their score. More precisely, ADSEQ can sample solutions under a Gibbs-Boltzmann probability distribution defined as follows: the Gibbs-Boltzmann score of a solution S is equal to $\exp^{-n(S)/kT}$, where kT is a user-defined pseudo-temperature, and this score defines implicitly a probability distribution over the set of all solutions. Tuning the pseudo-temperature kT provides a control over the probability to sample parsimonious solutions: a low pseudo-temperature tends to increase the probability to sample most parsimonious solutions while a large pseudo-temperature skews the Gibbs-Boltzmann distribution toward the uniform distribution over the set of all solutions (we refer to [51] for more details on the Gibbs-Boltzmann framework applied to the DECO algorithm).

In the present work, the *prior* scores of extant adjacencies are either 1.0 for adjacencies that are observed in a contig or scaffold, or a scaffolding score obtained from sequencing data using the scaffolding software BESST.

Pairs of genes located at the extremities of contigs and for which sequencing data do not provide any evidence for a scaffolding adjacency receive a small prior score as described in [52] (see also Additional file 1). The *posterior* scores are defined as the frequency out of a sample of 100 solutions with temperatures $kT = kT_0 = 0.1$ that skews the Gibbs-Boltzmann distribution toward optimal solutions.

Linearization of extant and ancestral components

After ADSEQ is applied to all considered pairs of gene families, the obtained result is a set of ancestral and extant gene adjacencies, each adjacency being assigned a *posterior* score. False positives – i.e. pairs of genes predicted inaccurately to be contiguous in an extant or ancestral genome –, can be due to errors in the data (for example errors in gene families or reconciled gene trees) or to errors in the inference process (for example the parsimony assumption might be wrong for some gene adjacencies). As a result, it is possible that a given gene (or gene extremity) belongs to more than two (more than one) adjacencies, which is incompatible with the expected linear structure of chromosomes.

To address this issue, we process the set of ADSEQ ancestral and extant adjacencies in such a way that they define linear ancestral and extant segments. To do so, we apply, independently for each species, a method used both in ancestral genome reconstruction [55] and scaffolding algorithms [56]. It consists in extracting, for each species, a Maximum-Weight Matching (MWM) in the graph whose vertices are gene extremities and edges are gene adjacencies, weighted by their posterior score. This MWM can still include circular segments, that are linearized by removing the least-weight edge of each such circular segment. Moreover, prior to this linearization step, we discard adjacencies whose posterior score is below a user-defined threshold, that was set to 0.1

after simulations aimed at measuring the accuracy of the ADSEQ algorithm (see Fig. 2). This linearization step is done independently for each species, ancestral or extant.

Application to the *Anopheles* data set

We now describe how we generated the data for the 18 *Anopheles* genomes.

Species trees

The main species phylogeny we considered was taken from [5]. We call it the “X phylogeny” as it is based on the X chromosome genes. It is the species tree used by default unless another is specified. We also considered an alternative “Whole Genome (WG) phylogeny”, that was introduced in [5] as the most frequently observed among trees built using sequences from the autosomes (Fig. 3).

Genomic data

Most genomic data we used were produced by the *Anopheles* 16 Genomes project described in [4] and retrieved from VectorBase (<https://www.vectorbase.org>): genome assemblies (contigs, scaffolds, or chromosome arms), gene annotations, and gene sequences (CDS). For 16 out of the 18 considered *Anopheles* species, we retrieved from the NCBI Sequence Read Archive (SRA) paired-end Illumina libraries with an insert size of 180bp (‘fragment’ libraries) and mate-pair libraries with an insert size of 1.5kbp (‘jump’ libraries), both obtained from a single female individual. Additional low-coverage long-range sequencing libraries (‘Fosill’ libraries) were obtained from pools of individuals. Details are available in Additional file 17: Table S1.

Gene families and trees

We retrieved homologous gene families from orthologous gene groups of 21 *Culicidae* species recovered from the OrthoDBmoz2 database (<http://cegg.unige.ch/orthodbmoz2>) generated using OrthoDB [57]. We generated a multiple sequence alignment for each family, used RAXML [58] to compute draft gene trees with bootstrap supports, and then corrected these draft gene trees using PROFILENJ [59]. PROFILENJ contracts branches with low bootstrap support and, using the species tree, resolves the polytomies in a way that minimizes the number of duplications and losses in a reconciliation. This resulted in 14,940 gene trees containing 183,680 genes. We refer to Additional file 1, Additional file 2: Figure S1, Additional file 3: Figure S2, Additional file 4: Figure S3, and Additional file 5: Figure S4, and Additional file 18: Table S2 for a description of our preprocessing pipeline (including preprocessing of these gene families), a comparison of the newly inferred gene trees with the original ones computed by the *Anopheles* consortium and for statistics on the inferred gene duplications and losses in these gene trees. As a species phylogeny is required to reconcile gene trees, we repeated the process described above for both the X phylogeny and the WG phylogeny.

Extant adjacencies and prior scores from sequencing data
Sequencing reads from all libraries were filtered to discard low quality reads and trimmed to 75bp length using TRIMMOMATIC [60], then mapped onto the contigs or scaffolds of the considered species using BOWTIE2 [61]; for reads with multiple mappings, all of them were conserved. The scaffolding software BESST [62, 63] was then used to detect potential scaffolding adjacencies between pairs of contigs containing at least one annotated gene. Scaffolding adjacencies that were not supported by at least four pairs of reads were discarded. For remaining scaffolding adjacencies, we assigned a score defined as the arithmetic mean of the two scores computed by BESST, the link variation score and the link dispersity score. Detailed statistics on the scaffolding adjacencies so obtained are available in Additional file 6: Figure S5 and Additional file 7: Figure S6.

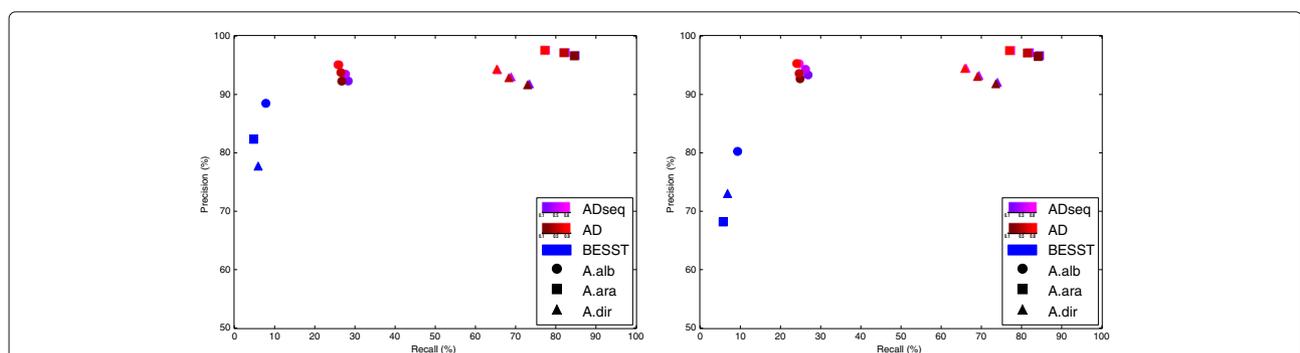
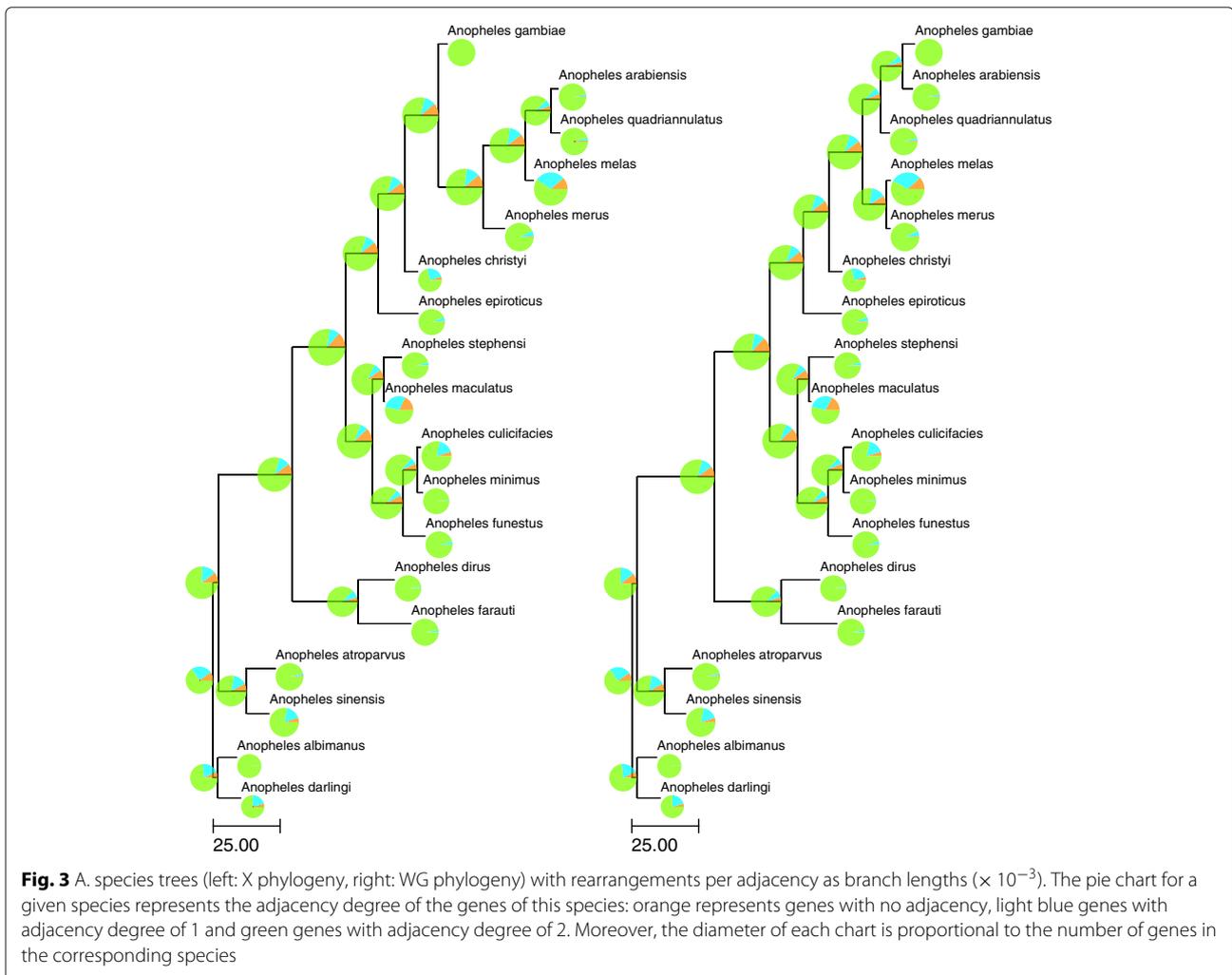


Fig. 2 Precision and recall statistics for scaffolding adjacencies on three artificially fragmented genomes (A.alb: *A. albimanus*, A.ara: *A. arabiensis* and A.dir: *A. dirus*). **Left graph:** results with 50% of reads. **Right graph:** results with all reads. The different methods results are plotted with the precision on the Y axis and the recall on the X axis. For ADSEQ and AD, results for three different adjacency support thresholds (0.1, 0.5 and 0.8) before genome linearization are plotted and represented with a color gradient. Note: A True Positive (TP) adjacency requires the proper orientation of both genes



Genome fragmentation simulations for measuring the accuracy of ADSEQ for extant scaffolding

We developed a validation protocol of ADSEQ to measure its ability to propose reliable extant scaffolding adjacencies (see Additional file 8: Figure S7 for an illustration of the protocol and Additional file 19: Table S3 for assembly statistics at different steps of the protocol). The key element is to provide to ADSEQ a genome whose assembly is more fragmented than the reference assembly, in order to verify that ADSEQ can retrieve the lost adjacencies. Moreover our simulation framework aims at generating a realistic fragmentation, as relying on a random fragmentation, as used in other validation protocols [16, 52], generates data that are in general easy to scaffold using comparative methods.

To avoid this pitfall we simulated a fragmented assembly by re-assembling the considered genomes using KMERGENIE [64] and MINIA [65]. MINIA was chosen due to its stringency in handling repeats, that leads to more conservative and fragmented assembly compared

to other contig assemblers. We applied this protocol to a randomly chosen either all or half of the raw sequence reads, independently three times, with the species *A. albimanus*, *A. arabiensis* and *A. dirus*, whose positions in the species tree allow to consider various evolutionary contexts.

We ran ADSEQ as described above on these new assemblies and compared its results (scaffolding adjacencies) with the reference assemblies. We call a True Positive (TP) adjacency an adjacency inferred by ADSEQ and present in the initial genome assembly. A False Positive (FP) adjacency corresponds to an adjacency inferred by ADSEQ and not present in the reference genome assembly. A FP can however be a true adjacency not found by the reference assembly (e.g., connecting two scaffolds), so we call Certain False Positive (CFP) a FP adjacency which extremities are not scaffold or contig extremities in initial genome assembly. Finally a False Negative (FN) is a pair of gene extremities that are contiguous in the initial assembly but are not inferred as an adjacency by ADSEQ. From these

values we compute the usual *precision* and *recall* statistics, but using CFP for the false positives count.

Gene order evolution analysis

Assignment of chromosome segments

To compare the evolution of *Anopheles* chromosomes, especially the apparent differences between the X chromosome and the autosomes described in [4], it is necessary to assign extant and ancestral chromosome segments to either the X chromosome or the autosomes. As *A. gambiae* is the only fully assembled genome in our data set, this is also the only genome for which such information is readily available; in all other species the genomes are assembled into scaffolds with no indication of whether this scaffold belongs to the X chromosome or an autosome, unless additional data is available, such as genome maps. We assigned extant and ancestral genes and segments to the X chromosome or autosomes using the following probabilistic method. For each gene g (ancestral or extant), a set of *An. gambiae* orthologs is defined as all *An. gambiae* genes from the same gene family than g whose last common ancestor with g in the reconciled gene tree is a speciation node. Note that this set might be empty, and that this definition includes the case where g is an ancestor of a *An. gambiae* gene. The probability of g being on the X chromosome is then defined as the frequency of orthologs located onto *An. gambiae* X chromosome, or, if no ortholog is present on this X chromosome, by a background probability, defined as the global frequency of *An. gambiae* genes on the X chromosome. Then each segment is given a probability to be located on the X chromosome as the mean of probabilities for all genes it contains. Each gene then inherits the probability of being on the X chromosome from the segment it belongs to.

Very recently an assignment of *An. albimanus* genes to chromosomes was published together with a new assembly [66] that we could use to verify that our assignment method is accurate: out of 8840 genes assigned to a chromosome in the novel assembly, we correctly predicted the autosomal/X placement of 8837 genes (comparing the assignment of higher probability and the assignment in the new assembly).

Gene movements (translocations)

For every couple of genes for which one is a direct descendant of the other, we inferred a gene movement (between the X chromosome and an autosome) if the probability of the ancestral gene being on the X chromosome is ≤ 0.2 while the probability of the descendant gene being on the X chromosome is ≥ 0.8 , or conversely.

Detecting chromosomal rearrangements

For every branch of the species tree, genes with exactly one exemplar in their family both in the ancestor and

descendant species were selected. Then conserved adjacencies were computed, which are adjacencies present between ancestral genes and descendant homologous genes. In order to discard gene displacements, which are counted elsewhere, we filtered also genes which are not involved in any common adjacency. A *rearrangement* (gain or breakage of a gene adjacency) is counted every time two gene extremities are contiguous on a segment (with respect to the reduced selected set of genes) of the ancestor, but not on the descendant, or conversely. When they are not contiguous, in order to detect a certain genome rearrangement, we require also that they are not both the extremities of their segments, to avoid counting as a rearrangement a potentially undetected scaffolding adjacency. Rearrangements were not directly counted as gains and breakages output from ADSEQ because this count can be blurred by adjacencies gained or broken by gene duplications and gene losses. As a consequence, gene duplications and losses are not counted as rearrangements, that are limited to synteny breakages due to balanced rearrangements (inversions, transpositions, large translocations), that do not change the gene content.

We stress that this method to detect genome rearrangements is conservative and underestimates rearrangement counts, as it does not detect the rearrangements hidden by the assembly fragmentation of the considered genomes. Moreover, this underestimation can be biased by the degree of fragmentation of the compared species, so two numbers of rearrangements are not necessarily comparable in biological terms, even for species closely located in the species phylogeny. However, given the same genomes in the input, the numbers of rearrangements for two different phylogenies are comparable, as are the number of rearrangements in sex chromosomes and autosomes.

Results

The results are organized in three parts. First, we describe the results of the simulation-based evaluation of the accuracy of ADSEQ to recover extant scaffolding adjacencies. Then we describe extant and ancestral genomes obtained with ADSEQ, and analyze important aspects of their evolutionary history. Finally we use ADSEQ to evaluate different species trees and re-examine the conclusions of [5] regarding species evolution.

Validation of the ADSEQ algorithm for extant scaffolding

We compared the scaffolding performance of ADSEQ with two other methods, on the same data set of realistically fragmented genomes (see [Methods](#)). One is using sequence information only, BESST [63], while the other one is using the comparative approach only, on the same phylogenomic data, AD (ADSEQ, where the possibility of

using sequence information is turned off). Figure 2 shows the precision statistic in function of the recall statistic for the two data sets (sampling 50% of the reads or using them all). For ADSEQ and AD, different values of the threshold for filtering adjacencies prior to linearization (see Methods) were tested.

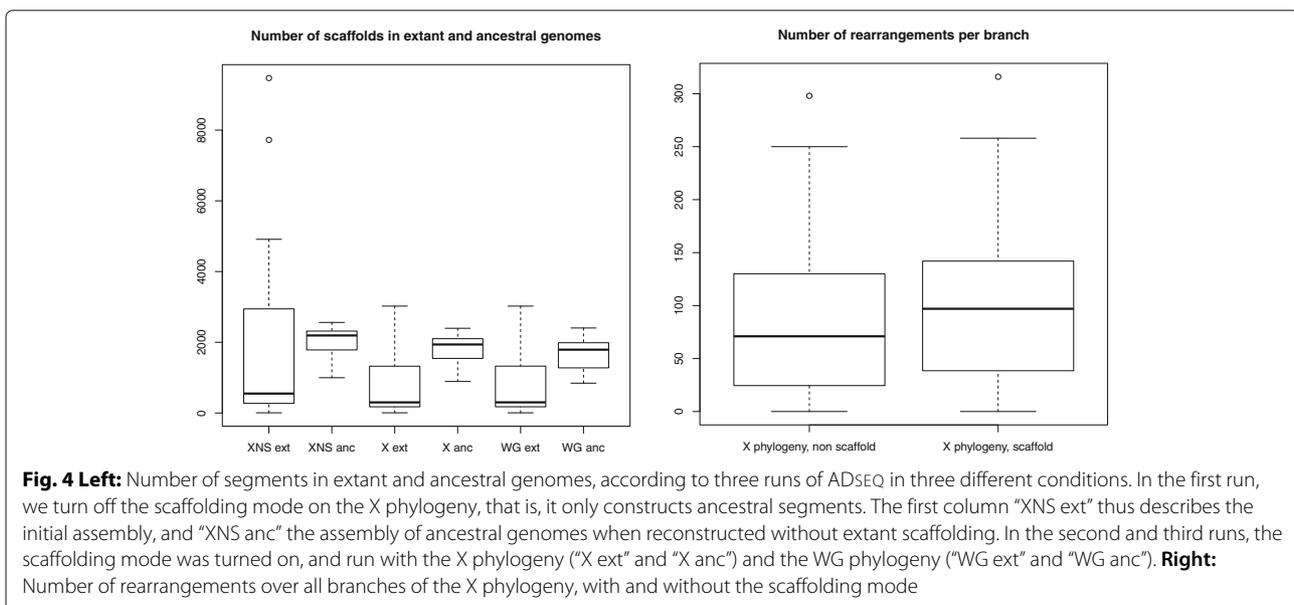
ADSEQ outperforms BESST in precision and recall independently of the threshold. This shows that in the majority of cases, adding phylogenomic information improves the recall without affecting the precision compared to a method purely based on sequence data. Additional results, where the gene orientation is not considered to determine an inferred adjacency as TP, are provided in Additional file 9: Figure S8 et Additional file 10: Figure S9. These results show that BESST has an equivalent precision and recall statistics compared to results of Fig. 2 where gene orientation is considered. This comparison shows that for most of inferred adjacencies the three methods inferred the right orientation for both genes involved in adjacency. In summary the combined approach using comparative signal and sequence data (ADSEQ) is giving significantly better results than a method based on sequence alone (BESST).

We now compare ADSEQ and AD, for a better illustration of the Fig. 2 to compare precision and recall statistics between ADSEQ and AD see subfigures zooming in Additional file 11: Figure S10. Recall statistic is slightly better with ADSEQ than with AD for all considered threshold values. For most of the condition the precision statistic is slightly better ADSEQ than AD (except in some conditions for *A. albimanus* and *A. dirus*, see Additional file 11: Figure S10). For *A. albimanus*, with all reads considered, ADSEQ outperforms AD for recall statistic for a

threshold fixed to 0.1 and 0.5. So from a quantitative point of view, adding sequence data seems to have a smaller impact on the recall and precision statistics compared to using synteny evolution. Note however that the combination of both supports for extant scaffolding adjacencies (sequence data and synteny evolution) is an important by-product of ADSEQ. A phylogenetic method alone is more difficult to trust in the absence of sequence data. So even if the general statistics are comparable, the additional support brought by the sequence data is an important feature. Moreover, additional results in Additional file 12: Figure S11 and Additional file 13: Figure S12 strongly support that a joint combination of phylogenetic and sequence signals (ADSEQ) overpasses an *a posteriori* combination of phylogenetic signal and sequence data (AD + BESST) for scaffolding improvement. These results show indeed that combining AD + BESST slightly overpasses ADSEQ in term of recall statistic (stronger TP adjacencies) but at the expense of a strong decrease of the precision.

Improved scaffolding of *Anopheles* extant and ancestral genomes

Properties of the improved assemblies for *Anopheles* extant genomes and of the reconstructed *Anopheles* ancestral genomes segments are summarized in Fig. 4. We describe three runs of ADSEQ: one without proposing extant scaffolding adjacencies (which amounts to use the DECLONE algorithm [51] to reconstruct ancestral genomes without improving extant genomes) and two with ADSEQ using the X and WG species phylogenies. The first observation that can be made is that the ability to create extant scaffolding adjacencies has a very significant impact on the ability to reconstruct ancestral segments,



that define ancestral genomes at a similar level of fragmentation than the improved extant genomes. This effect is important toward refined genome evolution analysis that rely on the ancestral segments as input material, especially to detect chromosomal rearrangements.

Additional file 20: Table S4 and Additional file 14: Figure S13, Additional file 15: Figure S14, and Additional file 16: Figure S15, provide more detailed illustration and statistics on the improved scaffolding. We observe that from 36,634 initial extant segments (contigs, supercontigs and scaffolds after the various filterings steps described in Methods and SI text), we scaffold the extant genomes into 13,525 segments, with an average number of 94 genes per segment up from 37 before running ADSEQ. Very similar results are obtained for all genomes independently of the chosen phylogeny, confirming the overall picture described in Fig. 4-Left.

On the right side of Fig. 4, we can observe that we retrieve a significantly (Wilcoxon paired test, p -value $< 10^{-4}$) higher number of rearrangements by the joint scaffolding technique than by just constructing ancestral genomes without scaffolding extant genomes. So the joint scaffolding of extant and ancestral genomes is beneficial to both. In particular scaffolding extant genomes while reconstructing ancestral genomes gives access to more information regarding the evolutionary history.

An interesting feature of ADSEQ is the possibility that the linearization step does delete an observed gene adjacency in an extant genome. This is unlikely as observed adjacencies have the highest score in the linearization procedure, but it can happen if it is in conflict with other adjacencies with high posterior probability. It happened only once in our dataset, for an adjacency between two *A. culicifacies* genes. The two genes were predicted in the reverse order, or equivalently in the reverse orientation, because all identified homologs were arranged similarly. This can be explained either by two inversions, one encompassing each gene, or by an assembly or annotation error. This shows that our approach can also detect questionable adjacencies in the given extant assemblies.

Evolution and phylogeny

ADSEQ is not a phylogenetic method *per se*, as it requires a given species phylogeny and does not include an extension to search an optimal phylogeny according to some evolutionary criterion. However as a method which infers ancestral gene orders and evolutionary events, and is computationally efficient (all steps that require a species phylogeny, including the correction of gene trees with ProfileNJ), the reconciliation of the gene trees with the species trees, and the joint scaffolding / ancestral genome reconstruction takes five hours on a laptop), it can be used to compare a few selected competing phylogenies. To this aim we compared several measures obtained by

the same methods using the two phylogenies, WG and X. They are the two main examined topologies in [5], where X is preferred as depicting the species history.

Duplications

Our pipeline using PROFILENJ to correct gene trees allows to record gene duplications. We counted a total of 6461 duplications for the X phylogeny, against 6159 duplications for the WG phylogeny (see Table 1). This means that for many gene families, a duplication was identified in the X phylogeny and not in the WG phylogeny. For these families, a well supported branch (100% bootstrap with RaxML) was compatible with the WG phylogeny but not with the X phylogeny, indicating that well supported branches are more often compatible with the WG phylogeny. This supports the result of [5] that most genes follow the WG phylogeny. The fact that this is observed on the autosomes and not on the X phylogeny also supports that the genomes evolve with two compartments.

Scaffolding and ancestral genome reconstruction

On the left side of Fig. 4, we can observe a first difference between the results obtained with the X and the WG phylogenies: the extant scaffolding is slightly better (in terms of fragmentation level of the extant genomes) with the X phylogeny (mean segment number is 835 genes for the X phylogeny, versus 840 for the WG phylogeny), while the ancestral scaffolding is better with the WG phylogeny (mean segment number is 1860 for the X phylogeny, versus 1756 for the WG phylogeny). The better extant scaffolding with the X phylogeny can be attributed to the basal position of the genome with best assembly (*A. gambiae*) in the *Gambiae* complex. Indeed in ADSEQ the sister species can be assembled according to *A. gambiae*, but outgroup species cannot, so the assembly is necessarily better if a fully assembled genome has more sisters species and less outgroups as it is the case with *A. gambiae* in the X phylogeny. Interestingly ancestral genomes are better scaffolded with the WG phylogeny, even with this sister-branch artifact that concerns extant genomes. This better ancestral genome reconstruction obtained with the WG phylogeny could be considered as a first signal contradicting the hypothesis that the X phylogeny is the true species phylogeny, although it does not allow to draw definitive conclusions.

Table 1 Numbers of inferred rearrangements and duplications in the X chromosome and in the autosomes, according to the phylogeny (X or WG) used as a parameter of ADSEQ

Event	X phylogeny		WG phylogeny	
	X chr.	Autosomes	X chr.	Autosomes
Duplications	604	5857	606	5553
Rearrangements	415	2949	416	2760

Conflict

With both phylogenies we also measured the level of syntenic conflicts, defined as the sum of the posterior scores of the adjacencies discarded during the linearization phase (data shown in SI text). We observe a higher level of syntenic conflicts in the X phylogeny (7665) than in the WG phylogeny (6319). According to simulations (described in SI text), the level of conflict is higher with a wrong phylogeny, even if it is not with the same order of magnitude than what we observe on our data. This could be seen as a second element contradicting the sequence-based hypothesis that the X phylogeny is the true species phylogeny, although the high level of conflict observed with both phylogenies here again does not allow to draw reliable conclusions.

Gene movements (translocations)

As it is believed there have been no large-scale rearrangement between the X chromosome and the autosomes in the *Anopheles* history [4, 67], we could assign most extant and ancestral segments (at least almost all that contain more than one gene) either to the X chromosome or to an autosome, with high accuracy (see [Methods](#)). Then we identified which genes moved from the X chromosome to an autosome, or conversely, by screening all couples of direct ancestor/descendant genes, one being in a segment assigned to the X chromosome and the other to an autosome. We found 429 genes having moved from the X chromosome to an autosome, and 469 from an autosome to the X chromosome, which confirms the trend found by Neafsey et al. [4] (59 over 132 gene movements originated from the X chromosome), although we draw our conclusion from experiments using now many more genes than in [4].

Genome rearrangements

We now turn to detecting genome rearrangements, as defined in the Methods section. In particular, we stress that we look for breaks and gains of gene adjacencies due to genome rearrangements such as inversions, transpositions and translocations, excluding duplications and losses, as well as adjacency breaks and gains due to duplications and losses. Moreover, due to the fragmented nature of many ancestral genomes, we expect to underestimate the true number of synteny breaks and gains.

We detect 3364 gains or breakages of adjacencies (860 in the *Gambiae complex*) using the X phylogeny, and 3176 using the WG phylogeny (590 in the *Gambiae complex*). The difference is illustrated in Fig. 3. Between the two competing phylogenies, one can observe a 30% decrease in the number of rearrangements within the *Gambiae complex* with the WG phylogeny compared to the X phylogeny. These gains/breaks of adjacencies can be combined along each branch to detect inversions, defined by

pairs of breaks in the ancestor and pairs of gains in the descendant involving the same four gene extremities. This lead to the identification of 242 inversions in the X phylogeny (including 16 inversions in the *Gambiae complex*, including 4 on the single lineage to *A. gambiae*) and 240 inversions with the WG phylogeny, with only 4 in the *Gambiae complex*, two of them on the branch leading to *A. gambiae*.

Comparison of sex chromosomes and autosomes evolution

Sex chromosome and autosomes have different evolutionary modes, according to duplications and rearrangements. Table 1 summarizes the number of inferred events of gene duplication and genome rearrangement in the sex chromosomes and autosomes, depending on the chosen phylogeny.

A striking observation is the different behavior of the X chromosome and of the autosomes regarding duplications and genome rearrangements. We do not count loss events to compare phylogenies because the absence of genes can be due to the fragmented assembly and not necessarily to actual gene losses during evolution. This compartmentalization was observed by [5] for genes and attributed to introgression in the autosomes; it was also noticed in [4] that the genome rearrangement rate was much higher in the X chromosome than in autosomes. We observe here a similar trend. We computed genome rearrangement rates by normalizing the number of observed gains and breaks of gene adjacencies by the number of gene adjacencies in the whole set of extant and ancestral genomes; with the X phylogeny we could observe that the X chromosome has a rearrangement rate equal to 1.46 times the rate observed in the autosomes, a figure that is similar (1.57) using the WG phylogeny. The observed higher rate of rearrangement in the X chromosome is in fact likely higher, as the relative fragmentation of the chromosome X is higher compared to the autosomes in most species both extant and ancestral (data not shown).

Moreover, we can observe interesting differences between the X and WG phylogenies. Constantly less events are found on the X chromosome with the X phylogeny, while less events are found on the autosomes with the WG phylogeny. It seems indeed that not only do genes follow different histories because of introgression [5], but also entire chromosomes do. However, the observed compartmentalization alone does not allow us to specify which part of the genome has followed the species diversification. As the *Gambiae complex* is estimated to be 2.2 million years old, it is reasonable to use parsimony arguments concerning rearrangements (see argument in the next paragraph). If we do so, we find less rearrangements in total in the WG phylogeny: even normalized by the number of adjacencies (because an increase in the number of rearrangements might be the effect of a higher

number of adjacencies): $9.15 \cdot 10^{-3}$ for the WG phylogeny versus $9.68 \cdot 10^{-3}$ for the X phylogeny. This means that rearrangements do not yield the same phylogenetic signal than the one suggested in [5]), which puzzles the evolutionary scenario in the *Gambiae complex* [68].

Assessing the relevance of rearrangements parsimony

The fact that parsimony can give a good account on the phylogeny can be questioned. Indeed, rearrangements in *Anopheles* are not uniformly distributed [69], they can show some degree of convergence, and rearrangements can show inter-species polymorphism. To test whether in the *Gambiae complex* we are in the domain of validity of parsimony, we compared the gene order of *A. gambiae* with *A. albimanus*, which, following the recent improved assembly of *A. albimanus*, are the two genomes which have their genes assigned to chromosomes. We selected all genes with an assignment to a chromosome, and applied the EM2 distance estimator [70]. It is based on a non uniform model of genome rearrangements which has proved to give the most reliable results on mammalian genomes, whose evolution spans a similar amount of time than the *Anopheles* genomes. We found an estimation of 1313 inversions with the statistical estimator, while the parsimony solution was 1300 (data not shown). So the parsimony result is within in the 1% interval of the statistical method, far from saturation. As *A. gambiae* and *A. albimanus* are separated by approximately 79 million years of evolution, we may suppose that, in the 2 or 3 million years that have shaped the *Gambiae complex*, rearrangements were not numerous enough to contradict parsimony.

Discussion

An important contribution of our work is the unification of two domains of research, namely genome scaffolding, and the evolution of gene order and ancestral genome reconstruction. They are usually separated despite the similarity of their objectives (reconstructing ancestral gene order is akin to a scaffolding procedure if ancestral genes are considered as contigs). Our work improves on previous works (especially [37, 39, 71]) in several aspects. In particular, we integrate elements coming from more traditional phylogenetic methods, such as gene trees and reconciliations, in order to be able to handle a large gene complement that includes gene families with complex evolutionary histories. Another important aspect of our work is the validation procedure of the scaffolding method. We propose a novel simulation procedure which takes real sequence data but lowers the coverage and rely on a conservative contigs assembler to obtain a realistic fragmentation. Using this validation method, we show that combining both sequence data and comparison with related genomes in a phylogenetic context produce better scaffolds, at least in the context of *Anopheles* genomes.

The benefit of the joint approach that considers in the same framework scaffolding extant genomes and reconstructing ancestral genomes is evident from both the improved extant genomes assemblies, where we reduce the fragmentation from roughly 36,000 segments to below 14,000 segments, and the detection of genome rearrangements, where we observe again a much better resolution of ancestral genomes. This allows us to detect a statistically significantly larger number of genome rearrangements that can not be confused with assembly artifacts. To the best of our knowledge, ADSEQ is currently the only method that can process such a data set with many genomes, most of them provided with fragmented assemblies, while using a large complement of gene families without being limited by the nature of the evolution of these families in terms of duplications and losses, and using also sequencing data. Regarding extant genomes scaffolding, the quality of our results depends of a set of factors, such as the quality of the initial extant assemblies and the position in the species phylogeny; we do not gain much for well assembled species such as *A. albimanus* which is almost an outgroup, while we refine very well the assembly of the genomes of species such as *A. minimus* or *A. dirus*. It is important to note that while we rely on sequencing data in the present work, other sources of data such as genome maps for example could be used to define a prior score for scaffolding assemblies. In terms of genome rearrangements, we likely underestimate their actual number due to the fragmentation of the reconstructed ancestral genomes. This is a consequence the very conservative approach we follow that detects only rearrangements for which there is a clear support. It remains to see if more realistic models of genome rearrangements that do not rely on reconstructed ancestral gene orders would be able to cope, in terms of computational complexity and of robustness of the detected rearrangements, with both the large number of species considered here and the level of fragmentation of the extant genomes assemblies. Nevertheless, the results we obtain support strongly the observation of [4] that the X chromosome evolves by genome rearrangements at a much higher rate than the autosomes.

Conclusions

Genes sequences and nucleotide substitutions are questionable markers for phylogeny because of introgression, due to hybridization or horizontal transfer, as it has been several times demonstrated recently, and in particular on *Anopheles* species. Genome rearrangements can help decipher phylogenetic relationships. However, they are difficult to infer from genomes with fragmented assemblies, a common feature for genomes obtained from short sequencing reads. We introduce a computational and statistical method for the inference

of genome rearrangement events in the presence of fragmented assemblies, proposing improved scaffolding based on sequencing data and an evolutionary model of gain/breakage of genes synteny along a given species phylogeny. We provide an evolutionary study of rearrangements, providing alternative hypotheses on the introgression process in the *Gambiae complex*.

Finally, our work opens the way to several research avenues. Generally, our general approach that relies on the joint analysis of sequencing data and the comparative approach to improve the quality of extant genome data could be extended to correct other types of errors that assembly breakpoints. To cite a specific example, it could be extended to account for the well known problem of unassembled genes [72], that create apparent gene loss and rearrangement breakpoints. Other avenues could include the development of metrics to compare alternative species phylogenies or the introduction in the evolutionary model of introgression events.

Additional files

Additional file 1: Supplementary text. (PDF 181 kb)

Additional file 2: Figure S1. Extant and ancestral genome gene content (left) and ancestral gene degree (right). Left: Number of genes of extant species (left), ancestral species using the reconciled VectorBase gene trees (middle), and ancestral species using the reconciled PROFILENJ gene trees (right). Right: Gene degree distribution of ancestral genes after applying ADSEQ with the RAW gene trees (blue graph) and the PROFILENJ gene trees (red graph), compared to the expected gene degree distribution for theoretical perfectly assembled genomes (black graph). The degree of a gene is defined as the sum of the ADSEQ *posterior* scores of adjacencies involving this gene. Here the value at coordinate x is the sums of all degrees in the interval $[x, x + 1]$. (PDF 92 kb)

Additional file 3: Figure S2. The pipeline is split into two parts: The first part (Blue one) processes genome content data to obtain extant genome adjacencies. **Step 1** detects genes that are included in other genes. In **Step 2**, gene families containing these genes are filtered out to avoid ambiguity in defining observed extant gene adjacencies. In **Step 3**, gene trees are inferred from the gene sequences, one gene tree per gene family (see Additional file 5: Figure S4 for more information on gene trees inference pipeline). Finally, in **Step 4** genes contained in gene families for which gene trees have not been inferred are discarded from the analysis (41 gene families containing representing 1,039 genes). The second part of the pipeline (Green one) processes sequencing data to obtain scaffolding adjacencies that will be used to improve extant genome assembly with the ADSEQ algorithm. **Step A** trims reads with TRIMMOMATIC to remove low qualities reads and remaining adapters. In **Step B**, trimmed reads are mapped onto their respective genome with BOWTIE2 considering all multiple mappings. In **Step C**, pairs of contigs for which paired-end reads suggest a possible contiguity along their chromosome are linked with the scaffolding software BESST, and the resulting potential scaffolding adjacencies are scored according to the BESST model. Then, in **Step 5** scaffolding gene adjacencies are determined from contigs adjacencies obtained from sequencing data processing part with genes present in gene trees. This results in scaffolds with observed scored scaffolding adjacencies that are used as input of DECoSTAR (**Step 6**). See Additional file 18: Table S2 for a description per species on dataset used for DECoSTAR. Pipeline to produce input data for ADSEQ on 18 *Anopheles* dataset. The pipeline takes as input a species tree for the 18 *Anopheles* species, the whole set of gene families and gene trees for these species and genomic data (contigs, scaffolds and chromosomes). The goal of the pipeline is to produce input

data for the ADSEQ algorithm to reconstruct ancestral genome structure and evolution and to improve extant genome scaffolding. (PDF 103 kb)

Additional file 4: Figure S3. Distribution of gene families number (red bars) and number of genes (blue bars) per families containing x species. **Left graph:** distribution of the 17,780 raw input gene trees corresponding to 212,800 genes. **Middle graph:** distribution of the 14,981 gene families, containing 184,719 genes, after discarding families containing included genes (after step 2 of Additional file 3: Figure S2). **Right graph:** distribution of the 14,940 gene trees, composed of 183,680 genes, after gene trees inference pipeline (after steps 3 and 4 of Additional file 5: Figure S4). (PDF 83 kb)

Additional file 5: Figure S4. Pipeline to improve gene trees inference from homologous gene family. CDS sequences of genes in gene trees have been obtained from VectorBase database and homologous gene families deduced from the 14,981 gene trees resulting of step 2 of Additional file 5: Figure S4. **Step A** consists to multiple align homologous genes with MUSCLE with parameter “-maxiters 2” if a gene sequence have a size upper than 32,000 bp. In **Step B**, GBLOCKS was applied on alignments to select high confidence alignment sites. At this step, 41 gene families have been discarded due to sequences that were not present in a selected blocks. For **Step C**, RAxML have been used to infer maximum likelihood gene trees with the GTR-GAMMA model and 100 bootstrap iterations. Finally in **Step D**, the maximum likelihood gene trees are processed with PROFILENJ to potentially changing branches with bootstrap support lower than 100% in a DL reconciliation model (min(Duplication, Loss)) with the species tree [73]. (PDF 35 kb)

Additional file 6: Figure S5. Distributions of scaffolding adjacencies scores computed by BESST for scaffolding adjacencies supported by at least 3 paired reads. **Left graph:** adjacency scores distribution between all contigs or scaffolds, over 405,939 scaffolding adjacencies. **Right graph:** adjacency scores distribution for contigs and scaffolds with gene corresponding to the 68,876 scaffolding gene adjacencies considered by DECoSTAR. Blue bars represent the link variation score, red bars the link dispersity score and purple bars the mean of the two link scores. For more information on the link scores see SI text and [62]. (PDF 37 kb)

Additional file 7: Figure S6. Distributions of scaffolding adjacencies link scores computed by BESST for scaffolding adjacencies supported by at least 3 paired reads, for each of the 18 *Anopheles* species. **Upper graphs:** distribution of scores all 405,939 potential scaffolding adjacencies. **Lower graphs:** distribution of scores for all 68,876 scaffolding gene adjacencies used as input by DECoSTAR. **Left graphs:** distribution of link dispersity scores. **Middle graphs:** distribution of link variation scores. **Right graphs:** distribution of the mean of link variation and dispersity scores. Each color corresponds to one species and the number between parenthesis in the legend indicates the number of scaffolding adjacencies inferred by BESST for each species. (PDF 64 kb)

Additional file 8: Figure S7. The ADSEQ validation protocol. (PDF 61 kb)

Additional file 9: Figure S8. Precision and recall statistics for scaffolding adjacencies on three artificially fragmented genomes (A.alb: *Anopheles albimanus*, A.ara: *Anopheles arabiensis* and A.dir: *Anopheles dirus*), when gene orientations are not accounted for. **Left graph:** results with 50% of reads. **Right graph:** results with all reads. The different methods results are plotted with the precision on the Y axis and the recall on the X axis. For ADSEQ and AD, results for three different adjacency support threshold (0.1, 0.5 and 0.8) before genome linearization are plotted and represented with a color gradient. These results show similar results to Fig. 2 showing that for most of the predicted adjacencies the three methods infer the correct gene orientation. (PDF 106 kb)

Additional file 10: Figure S9. Subfigures zooming of Additional file 9: Figure S8 to compare precision and recall statistics between ADSEQ and AD. **Upper graphs:** zoom of results with 50% of reads. **Lower graphs:** results with all reads. (PDF 154 kb)

Additional file 11: Figure S10. Subfigures zooming of Additional file 3: Figure S2 to compare precision and recall statistics between ADSEQ and AD. **Upper graphs:** zoom of results with 50% of reads. **Lower graphs:** results with all reads. (PDF 136 kb)

Additional file 12: Figure S11. Venn diagrams showing adjacencies shared by the three scaffolding methods ADSEQ, AD and BESST with a

sample of 50% of the reads. **Upper Venn diagrams:** results for *Anopheles albimanus*. **Middle Venn diagrams:** results for *Anopheles arabiensis*. **Lower Venn diagrams:** results for *Anopheles dirus*. **Left diagrams:** False Negative (FN) adjacencies, corresponding to adjacencies created by the fragmentation process and that have not been recovered. **Center diagrams:** results for True Positive (TP) adjacencies. Here, an adjacency is considered TP if the pair of genes is adjacent in the reference assembly and the orientation of genes involved in the adjacency is properly recovered. **Right diagrams:** results for Certain False Positive (CFP) adjacencies. An adjacency is determined as CFP when the pair of gene does not belong to the reference assemblies and one of the two genes is not located at a contig extremity in reference genome, or if the recovered orientation of genes is incorrect. If we consider method individually, these results show that ADSEQ outperforms AD and BESST with the lowest number of FN adjacencies, the largest number of TP adjacencies and the lowest number of CFP adjacencies (except for *An. albimanus* where AD has the lowest number of CFP (224 vs. 231 for ADSEQ). However, if we combine *a posteriori* AD and BESST, this performs better than ADSEQ in terms of recall (higher number of TP adjacencies) but at the expense of a strong decreases of precision (much higher number of CFP adjacencies). (PDF 20 kb)

Additional file 13: Figure S12. Similar to Additional file 12: Figure S11 with all reads included. (PDF 20 kb)

Additional file 14: Figure S13. Scatter plot exhibiting scaffolding improvement of the 18 *Anopheles* genomes by ADSEQ with X species tree phylogeny. Right plot is a zoom of a small part of the left graph. Each color corresponds to one species. For each species, upper part of vertical line corresponds to number of segments in initial genome assembly and lower part the number of segments after scaffolding improvement by ADSEQ. Circle diameter is proportional to the % of scaffolding improvement of the genome where scale is displayed in lower right part of the graphs. The X axis represent the sum of *a posteriori* scores of discarded adjacencies during linearization step representing the degree of syntenic conflicts in adjacencies prediction of ADSEQ (see paragraph "Conflict" in section "Results"). (PDF 56 kb)

Additional file 15: Figure S14. Similar to Additional file 14: Figure S13 with WG species tree phylogeny. (PDF 56 kb)

Additional file 16: Figure S15. Similar to Additional file 14: Figure S13 with RAW gene trees instead of PROFILENJ gene trees. (PDF 48 kb)

Additional file 17: Table S1. Summary of genome assemblies and sequencing data information. 16 on the 18 *Anopheles* species have been sequenced in [4] and data are available on the SRA database of the NCBI (see column 4 and 6 for BioProject and SRA ID). FASTQ files of paired sequencing data have been obtained with SRA-TOOLKIT. After mapping of paired reads on reference genome assemblies (column 2), median insert size of libraries have been determined with package "CollectInsertSizeMetrics" of PICARD TOOLS (v1.61) (column 7). Column "Library name" give information on the sequencing strategies employed in [4]. Where 'fragment' library corresponds to a Paired-End library with an expected insert size of 180bp and FR orientation (→←). 'jump' library corresponds to a Mate-Pair library with an insert size of 1.5kbp and RF orientation (←→). And 'fossil' corresponds to a library generated from a pool of hundred mosquitoes to improve the scaffolding with an expected insert size around 38kbp and FR orientation (→←). Column 3 gives the ID of gene set used in this study. (PDF 38 kb)

Additional file 18: Table S2. Assembly statistics on the 18 *Anopheles* genomes. Statistics before and after processing are displayed in columns 2-7 and after the pipeline to produce input data for the DeCoSTAR algorithm in columns 8-11 (see Additional file 3: Figure S2 for illustration of the data preprocessing step). For initial dataset assembly statistics, columns 2 and 3 present contigs number and N50 statistic in bp for all contigs in genome assemblies. In columns 4-7, only contigs with at least one gene are considered. Column 4 corresponds to contigs number with gene in reference assemblies. Columns 5 & 6 represent N50 statistics respectively in bp and in gene number. Column 7 represent the number of gene in reference genome assemblies. For genome assemblies used as input of DeCoSTAR, all contigs contains at least one gene. Column 8 gives the number of contigs after step 4 of Additional file 3: Figure S2. Columns 9 & 10 represent N50 statistics respectively in bp and in gene number. And column 11 represents the number of gene in genomes taken as input of

DeCoSTAR. The input dataset of DeCoSTAR is composed of 14,940 gene trees (see Additional file 4: Figure S3 and Additional file 5: Figure S4 for more information on gene trees) and 68,876 gene adjacencies with sequence support (scaffolding gene adjacencies) (see Additional file 6: Figure S5 and Additional file 7: Figure S6 for more information on scaffolding adjacencies). (PDF 33 kb)

Additional file 19: Table S3. Assembly statistics at various stages of the gene annotation step of validation protocol of ADSEQ (step 4 of Additional file 8: Figure S7). For each species and annotation step, table gives different assembly statistics (column2): the number of contigs in the assembly, the size of the assembly in bp and in gene number, the N50 statistics of the assembly in bp and in gene number (if available) and gene trees number corresponding to gene present in the assembly (for the two last columns). Column 3 (Initial assembly) corresponds to the assembly statistics of reference genomes. Columns 4-8 of upper and lower table corresponds to MINIA assembly statistics at different filtering step respectively with 50% reads sampling and without reads sampling. Column "initial" corresponds to assembly in output of MINIA algorithm assembly. MINIA contigs are then mapped on reference genome to annotate gene of reference assembly on MINIA contigs. Assembly statistics after filter1 corresponds to MINIA contigs that have been mapped on reference assembly with an identity and a coverage $\geq 90\%$. Filter2 consists to keep only contig with a unique optimal alignment (to avoid uncertainty in gene annotation). Column 7 corresponds to MINIA assembly statistics after merging of MINIA contigs overlapping a same gene (simulating RNA-seq scaffolding). Then last column corresponds to statistics after filter3 that consists to discard gene families of genes that have not been mapped on MINIA contigs. (PDF 51 kb)

Additional file 20: Table S4. Scaffolding statistics on the 18 *Anopheles* genomes before and after ADSEQ (with the X (upper table) and WG (lower table) species phylogenies). The columns 2-5 correspond to assemblies statistics before running the ADSEQ algorithm. Column 2 corresponds to the number of contigs in reference assemblies. The N50 statistic corresponding to the contig size where 50% of the total assembly length is comprised in contigs with size superior or equal to this value. This metric is computed with size considered both in bp (in column 3) and in gene number (in column 4). Column 5 gives the number of genes in genome assemblies give as input to ADSEQ. Columns 6-9 and 10-13 represent scaffolding statistics of ADSEQ respectively for X chromosome species tree topology and Whole-Genome topology. Columns 6 & 10 represent scaffolds number after ADSEQ. Columns 7 & 11, and 8 & 12 represent N50 statistics respectively for size in bp and size in gene number. Columns 9 & 13 represent new adjacencies inferred by ADSEQ (#scaff adj) represent the number of new adjacencies that are scaffolding adjacencies (i.e. adjacencies with sequence signal proposed by BESST and inferred by ADSEQ). (PDF 56 kb)

Abbreviations

CARs: Contiguous ancestral regions; CDS: Coding DNA sequence; CFP: Certain false positive; DP: Dynamic programming; FN: False negative; FP: False positive; MWM: Maximum-weight matching; SRA: Sequence read archive; TGS: Third-generation sequencing; TP: True positive; WG: Whole genome

Acknowledgements

Data computation and analyses benefited from the Compute Canada Western Canada Research Grid (Westgrid) and Montpellier Bioinformatics Biodiversity (MBB) platform services.

Funding

This work is funded by the Agence Nationale pour la Recherche, Ancestrrome project ANR-10-BINF-01-01, the MITACS Globalink program and the Natural Sciences and Engineering Research Council of Canada (NSERC), Discovery Grant RGPIN-249834 and a resource allocation from Compute Canada. Funding for the publication of this article was provided by the Agence Nationale pour la Recherche, Ancestrrome project ANR-10-BINF-01-01.

Availability of data and materials

Data and scripts used in this study are available on the GitHub repository: <https://github.com/YoannAnselmetti/ADseq-Anopheles-APBC2018>.

About this supplement

This article has been published as part of *BMC Genomics* Volume 19 Supplement 2, 2018: Selected articles from the 16th Asia Pacific Bioinformatics Conference (APBC 2018): genomics. The full contents of the supplement are available online at <https://bmcbgenomics.biomedcentral.com/articles/supplements/volume-19-supplement-2>.

Authors' contributions

ET, CC and SB designed the study. YA and WD implemented the method. YA applied the method to data. YA, ET, CC and SB analyzed the results and wrote the manuscript. All authors read and approved the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹ISEM, Université de Montpellier, CNRS, IRD, EPHE, Montpellier, France. ²Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR5558, 43 Boulevard du 11 novembre 1918, 69622 Villeurbanne cedex, France. ³INRIA Grenoble - Rhône-Alpes, 655 Avenue de l'Europe, 38330 Montbonnot-Saint-Martin, France. ⁴Department of Mathematics, Simon Fraser University, 8888 University Drive, V5A1S6 Burnaby, BC, Canada.

Published: 9 May 2018

References

- Zuckerkind E, Pauling L. Molecules as documents of evolutionary history. *J Theor Biol.* 1965;8(2):357–66.
- Springer MS, Gatesy J. The gene tree delusion. *Mol Phylogenet Evol.* 2016;94:1–33.
- Doolittle WF, Brunet TDP. What Is the Tree of Life? *PLOS Genet.* 2016;12(4):e1005912.
- Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, et al. Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. *Science.* 2015;347(6217):1258522.
- Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, et al. Mosquito genomics. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science.* 2015;347(6217):1258524.
- Wen D, Yu Y, Hahn MW, Nakhleh L. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Mol Ecol.* 2016;25:2361–72.
- Dobzhansky T, Sturtevant AH. Inversions in the Chromosomes of *Drosophila Pseudoobscura*. *Genetics.* 1938;23(1):28–64.
- Kamali M, Xia A, Tu Z, Sharakhov IV. A new chromosomal phylogeny supports the repeated origin of vectorial capacity in malaria mosquitoes of the *Anopheles gambiae* complex. *PLoS Pathog.* 2012;8:e1002960.
- Sharakhov IV. Chromosome phylogenies of malaria mosquitoes. *Tsitologiya.* 2013;55:238–40.
- della Torre A, Merzagora L, Powell JR, Coluzzi M. Selective introgression of paracentric inversions between two sibling species of the *Anopheles gambiae* complex. *Genetics.* 1997;146:239–44.
- Ayala D, Ullastres A, González J. Adaptation through chromosomal inversions in *Anopheles*. *Front Genet.* 2014;5:129.
- Ayala D, Acevedo P, Pombi M, Dia I, Boccolini D, Costantini C, et al. Chromosome inversions and ecological plasticity in the main African malaria mosquitoes. *Evolution.* 2017;71(3):686–701.
- Love RR, Steele AM, Coulibaly MB, Traore SF, Emrich SJ, Fontaine MC, et al. Chromosomal inversions and ecotypic differentiation in *Anopheles gambiae*: the perspective from whole-genome sequencing. *Mol Ecol.* 2016;25(23):5889–906.
- Tang J, Moret BME. Scaling up accurate phylogenetic reconstruction from gene-order data. *Bioinformatics.* 2003;19(Suppl 1):i305–i312.
- Large B, Simon D, Kadane J. Bayesian phylogenetic inference from animal mitochondrial genome arrangements. *J R Stat Soc Ser B.* 2002;64(4):681–693.
- Avdeyev P, Jiang S, Aganezov S, Hu F, Alekseyev MA. Reconstruction of Ancestral Genomes in Presence of Gene Gain and Loss. *J Comput Biol.* 2016;23(3):150–64.
- Hu F, Lin Y, Tang J. MLGO: phylogeny reconstruction and ancestral inference from gene-order data. *BMC Bioinformatics.* 2014;15:354.
- Zheng C, Sankoff D. On the PATHGROUPS approach to rapid small phylogeny. *BMC Bioinformatics.* 2011;12(S-1):S4.
- Bradley R, Jones ET, Rajaraman A, Chauve C. ANGES: reconstructing ANcestral GENomeS maps. *Bioinformatics.* 2012;28(18):2388–90.
- Kim J, Farré M, Auvil L, Capitanu B, Larkin DM, Ma J, et al. Reconstruction and evolutionary history of eutherian chromosomes. *Proc Natl Acad Sci.* 2017;114(27):E5379–E5388.
- Nagarajan N, Pop M. Sequence assembly demystified. *Nat Rev Genet.* 2013;14(3):157–67.
- Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet.* 2011;13(1):36–46.
- Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol.* 2015;33(6):623–30.
- Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science.* 2017. Epub ahead of print. <https://doi.org/10.1126/science.aal3327>.
- Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* 2016;26(3):342–50.
- Lewin HA, Larkin DM, Pontius J, O'Brien SJ. Every genome sequence needs a good map. *Genome Res.* 2009;19(11):1925–8.
- Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, et al. ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* 2015;16(1):3+.
- Fierst JL. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. *Front Genet.* 2015;6:220.
- Assefa SA, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics.* 2009;25(15):1968–9.
- Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD, Perna NT. Reordering contigs of draft genomes using the Mauve Aligner. *Bioinformatics.* 2009;25(16):2071–3.
- Dias Z, Dias U, Setubal JC. SIS: a program to generate draft genome sequence scaffolds for prokaryotes. *BMC Bioinformatics.* 2012;13:96.
- Bao E, Jiang T, Girke T. AlignGraph: algorithm for secondary de novo genome assembly guided by closely related references. *Bioinformatics.* 2014;30(12):319–28.
- Lu CL, Chen K, Huang S, Chiu H. CAR: contig assembly of prokaryotic draft genomes using rearrangements. *BMC Bioinformatics.* 2014;15:381.
- Tamazian G, Dobrynin P, Krasheninnikova K, Komissarov A, Koepfli KP, O'Brien SJ. Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences. *GigaScience.* 2016;5(11):38.
- Shaik S, Kumar N, Lankapalli AK, Tiwari SK, Baddam R, Ahmed N. Contig-Layout-Authenticator (CLA): A Combinatorial Approach to Ordering and Scaffolding of Bacterial Contigs for Comparative Genomics and Molecular Epidemiology. *PLoS ONE.* 2016;11(6):1–19.
- Husemann P, Stoye J. Phylogenetic comparative assembly. *Algorithm Mol Biol.* 2010;5:3.
- Kim J, Larkin DM, Asan CQ, Zhang Y, Ge RL, et al. Reference-assisted chromosome assembly. *Proc Natl Acad Sci.* 2013;110(5):1785–90.
- Kolmogorov M, Raney BJ, Paten B, Pham SK. Ragout - a reference-assisted assembly tool for bacterial genomes. *Bioinformatics.* 2014;30(12):302–9.
- Aganezov S, Alekseyev MA. In: Bourgeois A, Skums P, Wan X, Zelikovsky A, editors. Multigenome scaffold co-assembly based on the analysis of gene orders and genomic repeats. vol. 9683. Cham: Springer International Publishing; 2016. p. 237–249.
- Bosi E, Donati B, Galardini M, Brunetti S, Sagot M, Liò P, et al. MeDuSa: a multi-draft based scaffold. *Bioinformatics.* 2015;31(15):2443–51.

41. Sharakhov IV, Serazin AC, Grushko OG, Dana A, Lobo N, Hillenmeyer ME, et al. Inversions and gene order shuffling in *Anopheles gambiae* and *A. funestus*. *Science*. 2002;298:182–5.
42. Wang-Sattler R, Blandin S, Ning Y, Blass C, Dolo G, Touré YT, et al. Mosaic genome architecture of the *Anopheles gambiae* species complex. *PLoS ONE*. 2007;2(11):e1249.
43. Coluzzi M, Sabatini A, della Torre A, Di Deco MA, Petrarca V. A polytene chromosome analysis of the *Anopheles gambiae* species complex. *Science*. 2002;298:1415–8.
44. Caccone A, Min GS, Powell JR. Multiple origins of cytologically identical chromosome inversions in the *Anopheles gambiae* complex. *Genetics*. 1998;150:807–14.
45. Gentile G, Slotman M, Ketmaier V, Powell JR, Caccone A. Attempts to molecularly distinguish cryptic taxa in *Anopheles gambiae* s.s. *Insect Mol Biol*. 2001;10:25–32.
46. Mathiopoulos KD, della Torre A, Santolamazza F, Predazzi V, Petrarca V, Coluzzi M. Are chromosomal inversions induced by transposable elements? A paradigm from the malaria mosquito *Anopheles gambiae*. *Parassitologia*. 1999;41:119–23.
47. Appawu MA, Baffoe-Wilmot A, Afari EA, Nkrumah FK, Petrarca V. Species composition and inversion polymorphism of the *Anopheles gambiae* complex in some sites of Ghana, west Africa. *Acta Trop*. 1994;56:15–23.
48. Ma J, Zhang L, Suh BB, Raney BJ, Burhans RC, Kent WJ, et al. Reconstructing contiguous regions of an ancestral genome. *Genome Res*. 2006;16(12):1557–65.
49. Chauve C, Tannier E. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PLoS Comput Biol*. 2008;4(11):e1000234.
50. Bérard S, Gallien C, Boussau B, Szöllosi GJ, Daubin V, Tannier E. Evolution of gene neighborhoods within reconciled phylogenies. *Bioinformatics*. 2012;28(18):i382–i388.
51. Chauve C, Ponty Y, Zanetti JPP. Evolution of genes neighborhood within reconciled phylogenies: an ensemble approach. *BMC Bioinformatics*. 2015;16(Suppl 19):S6.
52. Anselmetti Y, Berry V, Chauve C, Chateau A, Tannier E, Bérard S. Ancestral gene synteny reconstruction improves extant species scaffolding. *BMC Genomics*. 2015;16(Suppl 10):S11.
53. Duchemin W, Anselmetti Y, Patterson M, Ponty Y, Bérard S, Chauve C, et al. DeCoSTAR: Reconstructing the ancestral organization of genes or genomes using reconciled phylogenies. *Genome Biol Evol*. 2017;9(5):1312–1319.
54. Jacox E, Chauve C, Szöllosi GJ, Ponty Y, Scornavacca C. ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*. 2016;32(13):2056–2058.
55. Manuch J, Patterson M, Wittler R, Chauve C, Tannier E. Linearization of ancestral multichromosomal genomes. *BMC Bioinformatics*. 2012;13(Suppl 19):S11.
56. Mandric I, Zelikovsky A. ScaffoldMatch: scaffolding algorithm based on maximum weight matching. *Bioinformatics*. 2015;31(16):2632–8.
57. Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Pozdnyakov IA, Ioannidis P, et al. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res*. 2015;43(Database issue):250–6.
58. Stamatakis A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analysis with thousands of taxa and mixed models. *Bioinformatics*. 2006;22:2688–90.
59. Noutahi E, Semeria M, Lafond M, Seguin J, Boussau B, Guéguen L, et al. Efficient gene tree correction guided by genome evolution. *PLoS ONE*. 2016;11(8):e0159559.
60. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 30;2014:2114.
61. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357–9.
62. Sahlin K, Vezzi F, Nystedt B, Lundeberg J, Arvestad L. BESST - Efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics*. 2014;15(1):281.
63. Sahlin K, Chikhi R, Arvestad L, Science C. Genome scaffolding with PE-contaminated mate-pair libraries. *bioRxiv preprint*. 2015;1–13.
64. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics*. 2014;30(1):31–7.
65. Chikhi R, Rizk G. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *Algorithm Mol Biol*. 2013;8(1):22+.
66. Artemov GN, Peery AN, Jiang X, Tu Z, Stegnyy VN, Sharakhova MV, et al. The Physical Genome Mapping of *Anopheles albimanus* Corrected Scaffold Misassemblies and Identified Interarm Rearrangements in Genus *Anopheles*. *G3 Genes Genomes Genet*. 2017;7(1):155–64.
67. Sharakhov IV, Artemov GN, Sharakhova MV. Chromosome evolution in malaria mosquitoes inferred from physically mapped genome assemblies. *J Bioinforma Comput Biol*. 2016;14(2):1630003.
68. Clark AG, Messer PW. Conundrum of jumbled mosquito genomes. *Science*. 2015;6217(347):27–8.
69. Pombi M, Caputo B, Simard F, Di Deco MA, Coluzzi M, della Torre A, et al. Chromosomal plasticity and evolutionary potential in the malaria vector *Anopheles gambiae* sensu stricto: insights from three decades of rare paracentric inversions. *BMC Evol Biol*. 2008;8:309.
70. Biller P, Guéguen L, Knibbe C, Tannier E. Breaking good: accounting for the diversity of fragile regions for estimating rearrangement distances. *Genome Biol Evol*. 2016;8:1427–39.
71. Lin Y, Nurk S, Pevzner PA. What is the difference between the breakpoint graph and the de Bruijn graph? *BMC Genomics*. 2014;15(Suppl 6):S6.
72. Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. *PLoS Comput Biol*. 2014;10(12):e1003998+.
73. Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst Zool*. 1979;28(2):132–63.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

