# Detecting introgression in Anopheles mosquito genomes using a reconciliation-based approach

Cedric Chauve[1], Jingxue Feng[2], Liangliang Wang[2]

[1] Department of Mathematics, Simon Fraser University,
8888 University Drive, Burnaby (BC), Canada
[2] Department of Statistics and Actuarial Sciences, Simon Fraser University,
8888 University Drive, Burnaby (BC), Canada

**Abstract.** Introgression is an important evolutionary mechanism in insects and animals evolution. Current methods for detecting introgression rely on the analysis of phylogenetic incongruence, using either statistical tests based on expected phylogenetic patterns in small phylogenies or probabilistic modeling in a phylogenetic network context. Introgression leaves a phylogenetic signal similar to horizontal gene transfer, and it has been suggested that its detection can also be approached through the gene tree / species tree reconciliation framework, which accounts jointly for other evolutionary mechanisms such as gene duplication and gene loss. However so far the use of a reconciliation-based approach to detect introgression has not been investigated in large datasets. In this work, we apply this principle to a large dataset of *Anopheles* mosquito genomes. Our reconciliation-based approach recovers the extensive introgression that occurs in the gambiae complex, although with some variations compared to previous reports. Our analysis also suggests a possible ancient introgression event involving the ancestor of *An. christyi*.

## 1 Introduction

Introgression is the transfer of genetic material between sympatric species, a donor and a receptor species, through hybridization between individuals of both species. It is an important evolutionary mechanism, that plays a key role in the evolution of eukaryotic genomes [15], especially toward the adaptation to a changing environment, a phenomenon known as adaptive introgression (reviewed in [16]). Among recent examples, the evolution of a group of African *Anopheles* mosquitoes, known as the *gambiae complex*, is of interest. This species complex includes most African vectors for the disease malaria, although not all species of the complex are malaria vectors. In 2015, Fontaine *et al.* demonstrated that there is extensive introgression within the gambiae complex, with possible implications related to the rapid acquisition of enhanced vectorial capacities [9]. The extent of introgression within the gambia complex was later confirmed by another work [29], using a different methodology, although the suggested introgression events were not in full agreement with Fontaine *et al.* The present paper follows this line of work, aiming at detecting traces of introgression within a larger group of *Anopheles* mosquito genomes, covering African and Asian mosquitoes.

There exist several methods that have been designed specifically for detecting footprints of introgression from genomic data, that can be classified into two main groups: methods based on summary statistics and methods based on evolutionary models. Among the first group, specific methods target the detection of introgression between two closely related sister lineages, relying on population genomics data for detecting haplotype blocks at a genetic distance lower than the expected distance if no introgression was involved; we refer to [20] for a recent discussion on these methods. When four species are considered, the most common summary statistic method is the *D statistics* [7], also called the *ABBA BABA statistics*. This method records, over several loci, the frequency of evolutionary trees that are incongruent with a given species phylogeny, and tests if the imbalance between the observed incongruent topologies is significant against a null hypothesis assuming that phylogenetic incongruence is solely due to Incomplete Lineage Sorting (ILS). There exist related methods that consider other invariants [3] or extend it to handle more than four taxa [18, 8], although at a significant computational cost. A common feature of these methods is that they aim at disentangling two evolutionary mechanisms that result in discordant gene trees compared to a given species tree: ILS and introgression. Another line of work is based on modeling introgression, that results from hybridization events, using phylogenetic networks, with evolutionary models that account for both ILS and hybridization. This model-based approach has been implemented in combinatorial [31, 11] and probabilistic frameworks [14, 30, 33, 28]. We refer the reader to [6] for a recent perspective on model-based approaches. These methods are highly parameterized, and generally their computational complexity grows exponentially with the number of reticulate edges considered in the phylogenetic network, and they have mostly been used with data sets of relatively moderate size so far, although recent pseudo-likelihood methods have shown promising improvements in computation time [32, 21].

An important drawback of the methods outlined above is that they rely on the analysis of orthologous loci, thus disregarding gene duplication and gene loss. While this can be a reasonable approach for small data sets, it does exclude many gene families for larger data sets. Moreover, as observed in [17], introgression through hybridization leaves a phylogenetic signal similar to Horizontal Gene Transfer (HGT), although both are very different from a biological point of view. HGT is an evolutionary mechanism that is well handled by several efficient gene tree / species tree reconciliation algorithms [25, 23, 24, 12] that scale well to large data sets. This suggests that the framework of reconciling gene trees with a known species tree could be used for detecting introgression without the need to filter out paralogous genes.

In the present work, we explore this idea, and apply a reconciliation-based method to detect signals of introgression in a large data set of 14 *Anopheles* genomes covering both African and Asian mosquitoes and including the gambiae complex. We use a combination of published methods to sample reconciled gene trees in an evolutionary model accounting for gene duplication, gene loss and HGT using almost the full complement of genes in our data set. In order to disentangle ILS and introgression, we rely on the hypothesis that introgression acts on larger genome segments, as discussed in [22], and we develop a statistical test to detect genome segments with significantly more genes whose evolution shows a signal of HGT than expected if such genes were located at random along chromosomes. Our approach recovers a strong signal for several introgression events within the gambiae complex, confirming the extensive level of introgression within this group of species, although with some differences related to specific introgressed segments. We also find support for a potential ancient introgression event involving the *An. christyi* lineage and the most common ancestor of the clade of Asian *Anopheles* mosquitoes.

## 2 Data and Methods

### 2.1 Data

Our starting data are the full genome sequences of 14 *Anopheles* species:

- the gambiae complex composed of *An. gambia* (AGAMB), *An. coluzzi*-(ACOLU), *An. arabiensis* (AARAB), *An. quadriannulatus* (AQUAD), *An. melas* (AMELA), *An. merus* (AMERU);
- two outgroups to the gambiae complex, *An. christyi* (ACHRI, an african mosquito) and *An. epiroticus* (AEPIR, an asian mosquito);
- a clade of asian mosquitoes, *An. stephensi India* (ASTEI), *An. stephensi sensu stricto* (ASTES), *An. maculatus* (AMACU), *An. culicifacies* (ACULI), *An. minimus* (AMINI), also including the african mosquito *An. funestus* (AFUNE) related to asian vectors [10]; from now we call this group the *Asian clade*.

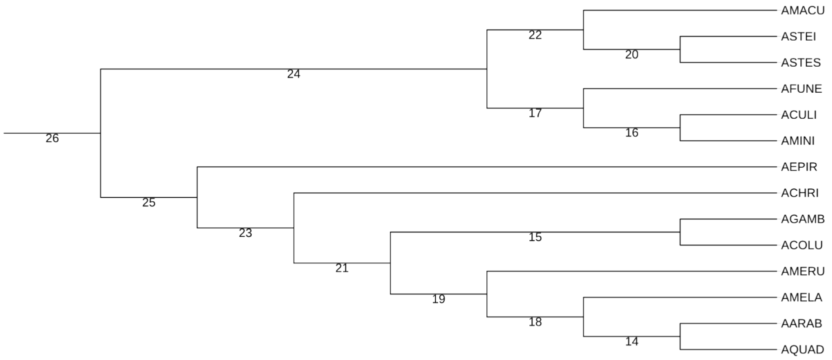

Fig. 1: Species tree of the 14 considered *Anopheles* species. Numbers on the internal branches identify ancestral species.

The species tree relating these species is given in Fig. 1; it is the so-called X-phylogeny used in [1]. In our experiments, we consider this tree as undated, i.e. with no given branch length. The branching pattern within the gambiae complex, a highly debated question, follows [9, 26].

The 14 genomes contain a single fully assembled genome, *An. gambia*, while some others are assembled at the contig level; we refer the reader to [1] for a precise discussion on the assembly of these genomes. The considered genomes contain from $10,000$ to above $14,000$ genes, that have been clustered prior to our study into homologous gene families using the OrthoDB algorithm [27] and represent an improvement of the set of gene families used in [1]. Fig. 2 below illustrates the distribution of the number of genes per genome and the sizes of the gene families.

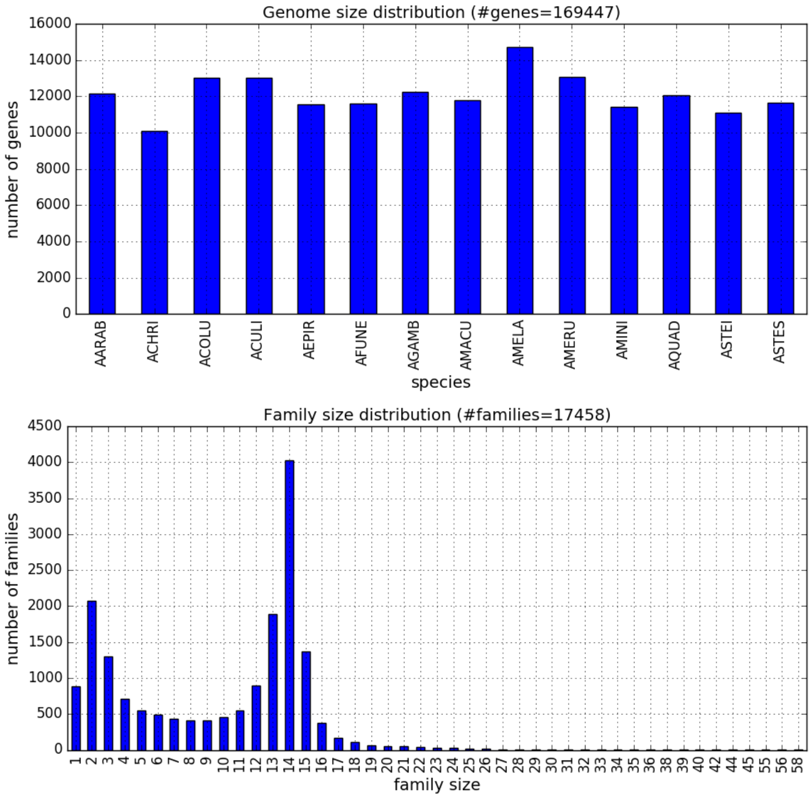

Fig. 2: (Top) Distribution of the number of genes per species. (Bottom) Distribution of the size (number of genes) of gene families.

An important observation is the large number of very small gene families, likely due to errors in assembling genes or in clustering genes into homologous families, an expected issue with large-scale multi-species genomic data sets. For each family, a multiple sequences alignment (MSA) of the coding sequences of the genes belonging to the family was obtained using the method described in [1].

### 2.2 Methods

Our analysis of this data set contains three main steps. In a first step, we sample reconciled gene trees for each homologous family, in a model including gene duplication, gene loss and HGT. Then we evaluate the consistency of these inferred HGTs to verify that they do not contain a high level of noise. Last we rely on robust HGT to detect potential introgression events and we apply a statistical test of the co-clustering of the involved genes to detect genome segments potentially involved in these introgression events.

*Reconciled genes trees.* For each gene family, we ran MrBayes, a software package for Bayesian phylogenetics inference [19], using the family MSA as input and computing two independent MCMC (Markov-Chain Monte-Carlo) chains per family. MrBayes was run using the General Time Reversible model of sequence evolution with a proportion of invariable sites and a $\Gamma$-shaped distribution of rates across sites. The MCMC ran for $10,000,000$ generations and both chains started from different random trees. Since the average standard deviation of split frequencies (ASDSF) approaching 0 indicates convergence [13, 19], we used 0.01 as the threshold of this statistic to determine if the MCMC chain has converged or not. The ASDSF was examined every $5,000$ iterations, and, after convergence, tree samples were saved every 500 MCMC iterations, leading to a maximum of $20,000$ sampled gene trees. Families for which at least one chain generated less than $5,000$ samples were discarded from further analysis. We refer to these sampled trees as the *MrBayes trees*, with two sets of MrBayes trees being generated for each gene family.

Next, for each selected family, the MrBayes trees were provided as input to ALE [25], a method for the exploration of the space of reconciled gene trees, accounting for gene duplication, gene loss and HGT[3]. Reconciled gene trees are gene trees augmented with a mapping of each internal node to a species of the species tree (extant or ancestral) and an annotation of the node as either a speciation or a duplication or a HGT; in the latter case the receptor species of the HGT is also indicated. Given a set of MrBayes trees, ALE extracts the clades observed in these trees and their frequencies, and explores the space of reconciled gene trees that can be assembled from these clades (a process called *gene tree amalgamation*) while maximizing the likelihood of observing the reconciled gene tree. The result is a maximum likelihood amalgamated reconciled gene tree; moreover, when used with its Bayesian MCMC mode, ALE determines the rates of gene duplication, loss and HGT and can sample reconciled gene trees.

For each homologous gene family, ALE was run independently on the two sets of MrBayes trees resulting of the two MrBayes MCMC chains, using its

[3] As mentioned previously, introgression and HGT are different evolutionary mechanisms; however, for expository reasons, we refer to the transfer of genetic material between two *Anopheles* species as an HGT.

Bayesian MCMC mode. For each run, an amalgamated reconciled gene tree was computed, together with a sample of $1,000$ reconciled gene trees, sampled every 100 iterations of the MCMC chain. We call these two sets of sampled reconciled gene trees the *ALE trees*. Gene families for which the two amalgamated reconciled gene trees were not identical were excluded from further analysis. For a given family, the frequency of an HGT, defined by a donor species $d$ and a receptor species $r$ and denoted by the ordered pair $(d, r)$, is obtained by averaging, over the two independent runs of ALE, the frequency of observing this HGT in the ALE trees; note that $d$ and $r$ can both be either an extant or an ancestral species. The final output of this step is a list of quadruples (donor $d$, receptor $r$, family $g$, frequency $f$): each such quadruple records that, for the given family $g$, an HGT from species $d$ to species $r$ was observed in the sampled reconciled gene trees with frequency $f$. In the rest of this work, we analyze the observed HGT to detect traces of introgression.

*Consistency of HGTs.* It is well known that the accurate detection of HGT is challenging, especially when using an undated species tree. It is then important to evaluate the noise due to likely erroneous HGTs. To do so, we rely on the recent method MaxTiC [5]. MaxTiC aims at ranking the internal nodes of a species tree provided with weighted ranking constraints derived from a set of HGTs, in order to maximize the total weight of the satisfied constraints. In our case, constraints are obtained from HGTs as described in [5]: a given HGT from a donor species $d$ to a receptor species $r$ defines a ranking constrain that the ancestor $a$ of $d$ should be older than $r$. Note that, by definition, a HGT whose donor is an extant species does not create a constraint that can conflict with a ranking of the internal nodes; as a consequence, we excluded such constraints from the input of MaxTiC. The weight of a ranking constraint is the sum of the frequencies of the HGTs defining it that are observed across all selected gene families. We applied MaxTiC with inputs composed of ranking constraints derived from several sets of HGTs, obtained by filtering out inferred HGTs whose frequency is below a threshold $t$, ranging from 0.20 to 0.95 by steps of 0.05.

The result of MaxTiC, for a given value of the threshold $t$, is composed of two sets of ranking constraints, the constraints consistent with the computed ranking of the internal nodes of the species tree, and the constraints in conflict with this ranking. We define the *consistency ratio* as the ratio between the weight of the consistent constraints divided by the weight of all considered constraints at frequency threshold $t$. Intuitively, a high consistency ratio points at a low proportion of erroneous HGTs.

*Detecting potentially introgressed segments.* Gene duplication and HGT are two mechanisms that can cause incongruence between a gene tree and a species tree, that are accounted for in ALE. However, ILS is a third common cause of phylogenetic incongruence, that is not considered in the ALE model. A crucial question toward detecting introgression is to distinguish inferred HGTs likely due to introgression to HGTs that could be due to ILS. To do so we rely on the hypothesis that, unlike ILS, introgression is more likely to impact blocks of contiguous genes [22]. Based on this hypothesis, for a given pair of species $(d, r)$, we aim at detecting genome regions where the concentration of genes belonging to families whose evolutionary history as given by sampled reconciled trees involves HGT from $d$ to $r$ is significantly higher compared to a null hypothesis that such families are scattered randomly along chromosomes. As *An. gambia* is the only fully assembled genome in our data set, we perform all tests using *An. gambia* chromosomes; we discuss the impact of this approximation in the Discussion section.

We designed our analysis as follows. Consider an *An. gambiae* genome segment (called *window* from now) containing $n$ genes. Let $p$ denote the probability of observing, within a given window, a gene from a family whose evolution involves a $(d, r)$ HGT, and $p_0$ be the average of the $(d, r)$ HGT frequencies for all the genes on the whole genome, where HGTs are the ones inferred from the ALE results. A statistical hypothesis testing is conducted to test the null hypothesis, $p = p_0$, versus the alternative hypothesis, $p > p_0$. Note that prior to this test, tandem arrays, i.e. segments of consecutive genes from the same family, were reduced to a single gene. Let $X_i$ be the number of observed HGTs from $d$ to $r$ in the $s$ ALE sampled reconciled gene trees for the $i$-th gene in the window, where $i = 1, \dots, n$. We assume the distribution of $X_i$ to be Binomial$(s, p)$. An unbiased estimator for $p$ is $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{sn}$. Under the assumption that $X_i$'s are independent for simplicity, we have $var(\hat{p}) = \frac{p(1-p)}{sn}$. Consequently, the test statistics $Z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/(sn)}}$ is approximately distributed as a standard Normal distribution. Let $z$ be the observed value of $Z$ given the ALE results. The $p$-value of the hypothesis testing can then be obtained by computing $P(Z \geq z)$.

For the multiple tests over all the windows on each chromosome, we used the Benjamini-Yekutieli (BY) [2] method to control the False Discovery Rate (FDR) in a multiple testing setting with dependencies between the tests, which is the case in our experiment as adjacent windows are not independent.

The result of this analysis is a list of windows for which we detected a significantly higher density of genes supporting an HGT from $d$ to $r$, under a FDR of 1%; we selected a window size of $n = 20$ genes, although results were similar with $n = 10$ or $n = 30$.

## 3 Results

*Reconciled gene trees.* After running MrBayes and ALE, and filtering out gene families for which both MrBayes chains did not generate at least $5,000$ sampled gene trees and gene families for which the two amalgamated gene trees generated by ALE were not identical, there are $11,589$ gene families containing a total of $137,180$ genes left, with each species "losing" roughly $2,000$ genes. Fig. 3 illustrates the impact of this filtering on homologous families sizes.

Comparing with Fig. 2, we observe a significant decrease in the number of gene families with 12 or more genes, indicating that many of these families do not generate consistent amalgamated gene trees under our relatively stringent

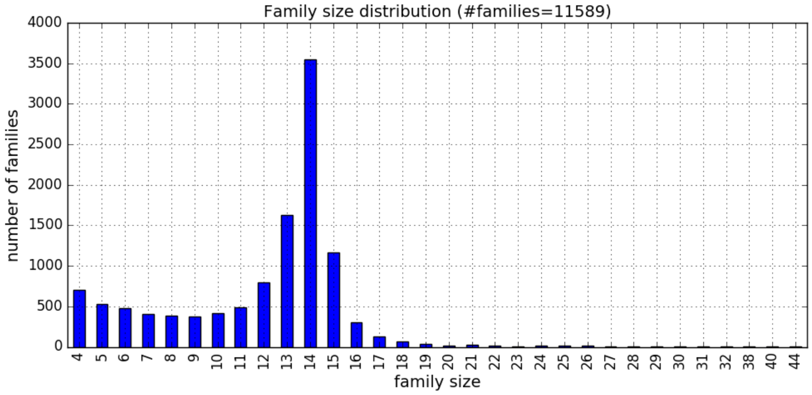

Fig. 3: Distribution of the number of genes per homologous family after filtering out families for consistency of the two runs of the MrBayes+ALE pipeline.

filtering criteria. We can also observe that the majority of the gene families passing our filtering step are not composed of one-to-one orthologous genes, motivating an approach based on an evolutionary model accounting for gene duplication and gene loss.

*Horizontal Gene Transfers.* Next we consider the inferred HGTs that can suggest potential introgression events. After filtering out, for each gene family, all HGT that do not appear in at least 20% of both sets of ALE trees, the total number of conserved HGTs is $16,210$, leading to an average number of inferred HGT per gene family slightly above 1. Fig. 4 shows that low-frequency HGTs dominate the landscape, although there are $4,771$ (resp. $1,778$) HGTs observed with frequency at least 50% (resp. 80%).

Next, the MaxTiC results suggest that the inferred HGTs do not show an apparent high level of noise, measured in terms of conflicting HGTs. The consistency ratio increases steadily from 0.908 at $t = 0.2$ to 0.938 at $t = 0.5$ and 0.973 at $t = 0.8$, indicating a low level of conflict among HGTs with frequency at least 0.2. The most interesting finding is that, at threshold $t = 0.7$, only two constraints with a significant weight are discarded, constraints $(18,15)$ and $(14,15)$ – where $(x,y)$ means that node $x$ should be ranked before node $y$ – while the reversed constraints $(15,18)$ and $(15,14)$ are among the conserved constraints, although with a weight an order of magnitude higher. The constraints $(15,14)$ and $(14,15)$ originate respectively from HGTs from *An. arabiensis* and *An. quadriannulatus* to ancestral species 15 and from *An. gambia* and *An. coluzzi* to ancestral species 14. This observed time inconsistent HGTs between these two groups of species is discussed in Section 4.

*Potential introgression events.* In order to classify inferred HGTs as potential introgression events from a donor species $d$ to a receptor species $r$, we used the following stringent criteria: the HGT must be observed in at least 50 gene

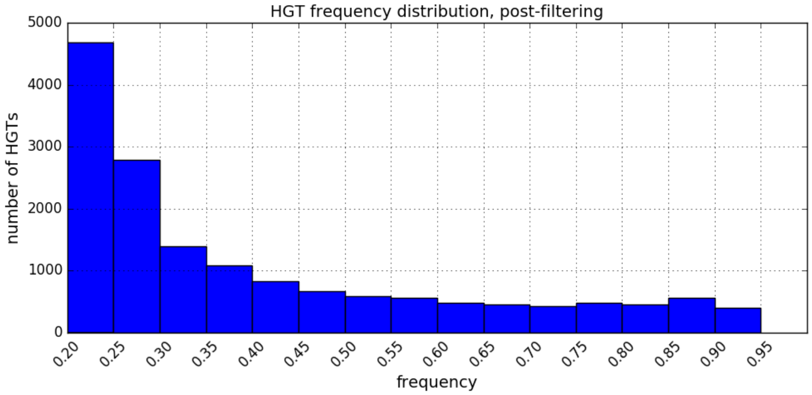

Fig. 4: Distribution of the frequency of observed HGTs appearing with frequency at least 20% in ALE trees.

families, at frequency at least 50%, with an accumulated frequency over all such gene families at least 50. These criteria are based on the results of the MaxTiC analysis. Fig. 5 shows the potential introgression events detected using these criteria.

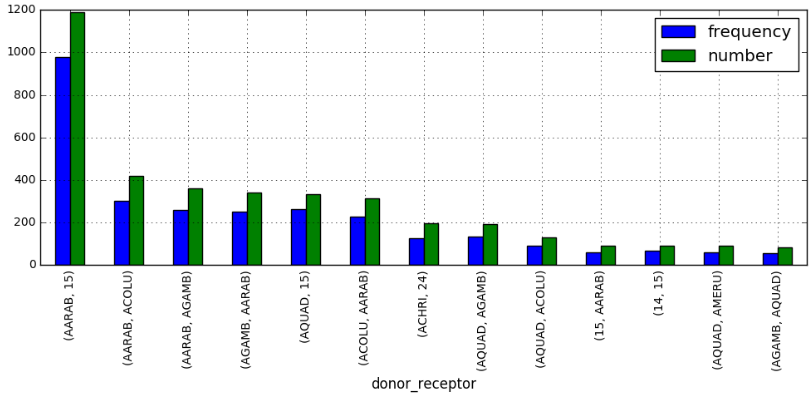

Fig. 5: Potential introgression events based on sets of at least 50 inferred HGTs of frequency 0.5 or above and accumulated frequency at least 50.

As expected, most potential introgression events are recent and concern the gambiae complex, in agreement with the extensive amount of introgression seen in this group [9]; in particular, we retrieve the major introgression from *An. arabiensis* to the common ancestor of *An. gambiae* and *An. coluzzi* (ancestral species 15) that was found in [9, 29]. We can also observe that almost all potential introgression between the two groups of *An. gambia*, *An. coluzzi* and their common ancestor on one side and *An. arabiensis*, *An. quadriannulatus* and species 14 on the other side seem to be bidirectional, although at various levels of support. The only other potential event within the gambiae complex found by our analysis is the event (*A. quadriannulatus*,*An. merus*), agreeing with the direction proposed in [29] as opposed to [9], although with a limited support.

As mentioned above, we do not find a strong support for any introgression between species of the Asian clade. In order to find such an event, one needs to relax significantly our criteria by considering HGTs observed with frequency as low as 20%; the only event found is then from *An. maculatus* to *An. culicifaces*.

However, the most striking observation is the hypothesis of a potential introgression event from the lineage of *An. christyi* to ancestral species 24, the last common ancestor of the Asian clade. To the best of our knowledge, such an ancient, potential, introgression event has not been discussed in the literature so far. This potential introgression is supported by 195 HGTs with an average frequency of 0.65, comparable to likely introgression events, such as the one from *An. quadriannulatus* to *An. gambia* (193 HGT, average frequency 0.70), discussed in [29].

In order to assess further the level of support for these various potential introgression events, we considered, for each such event, the taxon coverage of the gene families whose evolution involves an HGT supporting the event. The rationale is that for HGTs supported by gene families with low taxon coverage, the identification of the donor and receptor species could lack precision. Overall, we find that all potential introgression events are supported by gene families covering a large number of species, from an average of 12.51 for (*An. arabiensis*, *An. gambiae*) to 13.89 for (*An. christyi*, species 24). The same analysis repeated after lowering the HGT frequency threshold to 0.2 lead to similar results, with a slight decrease of the average taxon coverage by gene families; in this context the event from *An. maculatus* to *An. culicifaces* is supported by families covering on average 6.12 species, thus lowering further its support and confirming the absence of signal for introgression events within the Asian clade.

*Spatial distribution of gene families involved in HGTs.* Our analysis of the clustering of gene families involved in HGTs along the chromosomes of *An. gambia* showed that for all the potential introgression events shown on Fig. 5, some genome regions contain significant clusters of gene families supporting the event. We provide all the corresponding chromoplots images at `https://github.com/cchauve/Anopheles_introgression_RECOMBCG_2018` and discuss below some interesting observations.

First, considering the three events with *An. arabiensis* as donor species and the chromosome arm 2L, one of the 4 autosomal arms of all *Anopheles* species, the pattern of potentially introgressed genes is very different, as shown in Fig. 6. It is interesting to observe that there seems to be close to no introgression signal toward *An. gambia* on this arm, while the introgression to *An. coluzzi* is centered around the region of the so-called 2La polymorphic inversion. A similar pattern showing very specific regions of the X chromosome being introgressed can be observed, illustrated on Fig. 7.

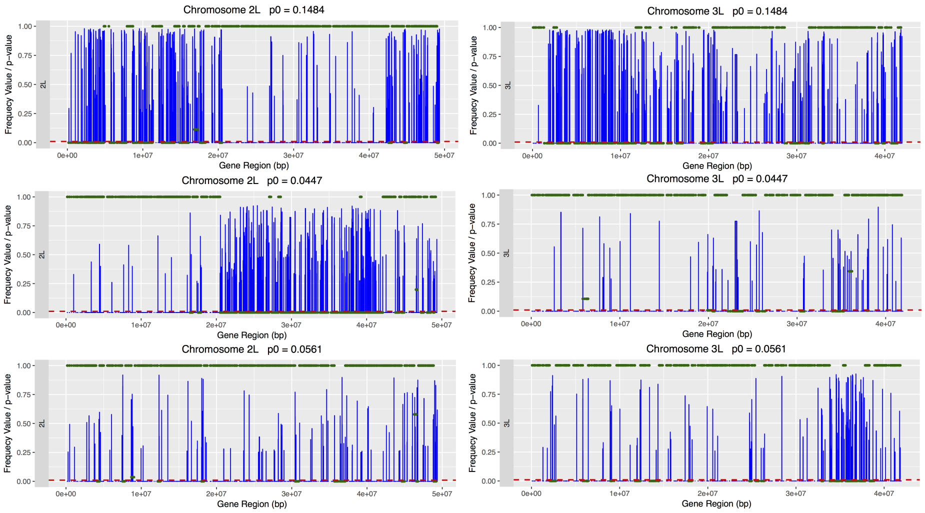

Fig. 6: Chromoplots for chromosome arms 2L and 3L for introgression from *An. arabiensis* to species 15 (Top), *An. gambia* (Middle) and *An. coluzzi* (Bottom). Blue vertical bars indicate genes with their HGT frequency, the red dotted line the FDR of 1% and green dots the BY corrected p-value.

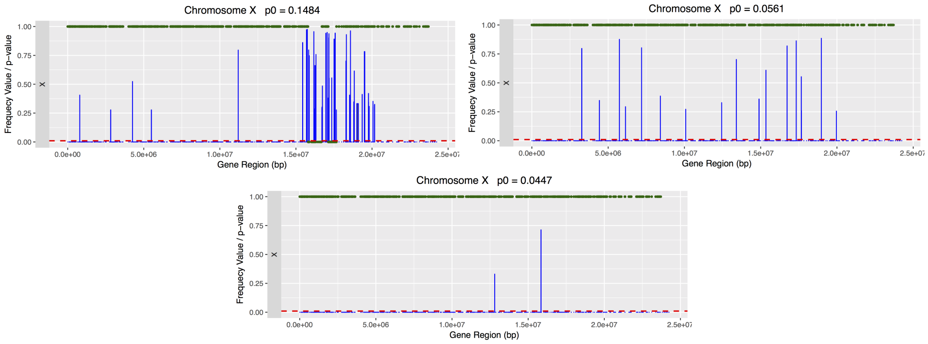

Fig. 7: Similar to Fig. 6 for chromosome X

Within the gambiae complex, we retrieve patterns observed in other works. We see for example that the introgression from *An. quadriannulatus* to *An. gambia* involves mostly the 2La inversion again, as was discussed in [29]. We can also see that the signal for an introgression event from *An. quadriannulatus* to *An. merus* involves limited regions of chromosomal arms 3R and 3L.

Last, looking at the chromoplots obtained from the HGTs observed between the lineage of *An. christyi* and species 24 (Fig. 8), we can see a level of support similar to the potential events located within the gambiae complex, although with a much stronger signal for introgression located on the X chromosome.

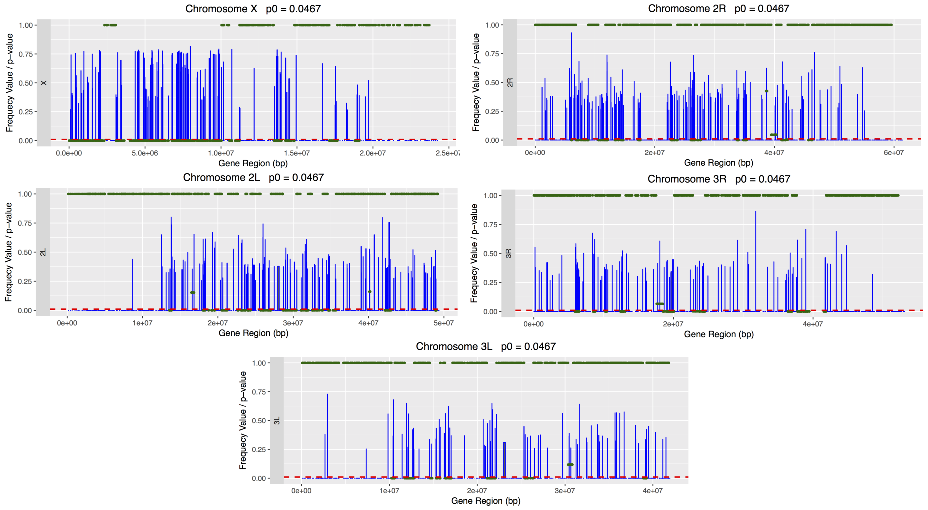

Fig. 8: Chromoplots for the potential introgression event from *An. christyi* to species 24.

## 4 Discussion

Our work builds upon the fact that reconciling gene trees with a given species tree, in an evolutionary model that accounts for HGTs, offers a natural framework to detect traces of introgression events. This approach benefits from the recent development of reconciliation algorithms that sample reconciled gene trees in evolutionary models accounting for HGT, both using parsimony [12] and probabilistic [25, 24] methods.

Using established phylogenetic and phylogenomics methods and stringent filtering criteria, this approach recovers the well accepted extensive introgression within the gambiae complex and leads to the hypothesis of an ancient introgression event from the lineage of *An. christyi* to the common ancestor of Asian *Anopheles* mosquitoes, involving predominantly the X chromosome. This hypothesis is interesting as the phylogenetic placement of *An. christyi* was questioned in [3] – that uses the same species tree as in the present work – where it was suggested that the significant level of phylogenetic incongruence is indeed due to biological causes (ILS or hybridization); nevertheless, the branching pattern of the gambiae complex, the Asian mosquitoes, *An. christyi* and *An. epiroticus*, deserves further investigation related to ILS and introgression.

Compared to methods based on summary statistics of gene trees, such as the D-statistics, the approach we suggest has several advantages, that are well illustrated by our work. First, the reconciliation framework can handle gene families with gene duplication and gene loss events. Therefore, we can use almost the full complement of genes in larger data sets, unlike methods based on the analysis of one-to-one orthologous loci. Moreover, for all considered gene families, evolutionary trees are computed using a more comprehensive evolutionary model that actually accounts for hybridization events; this contrasts with summary statistics methods that rely on the analysis of gene or loci trees computed without accounting for such evolutionary events. Finally, sampling reconciled gene trees is important toward providing a more nuanced view of evolutionary processes at play within a group of species; the impact of filtering out HGTs sampled with low frequency in our work illustrates this important feature of our approach. Finally, summary statistics methods are limited to data sets with few species, that are in general assumed to be closely related, unlike our approach, a feature which is a crucial toward raising the hypothesis of an ancient introgression along the lineage of *An. christyi*.

Despite the promising results we obtain on a well studied data set, the idea of using a reconciliation-based approach for detecting footprints of introgression requires to be evaluated very carefully. Indeed, our work can certainly not be considered sufficient to claim that HGT infrerred from reconciliations can capture accurately introgression events. Such an evaluation would require to assess its accuracy on simulated datasets, especially toward measuring the impact of using homologous gene families instead of orthologous gene families. The impact of errors in gene families, gene trees and the considered species tree, among other factors, should also be assessed in these simulations. It would also allow to evaluate different reconciliation algorithms, including recently developed algorithms that account for ILS [23, 4]. Regarding these two algorithms, it would be interesting to see if they could be extended to sample reconciled gene trees; also, both consider a parsimony framework and the ability of ALE to sample reconciled gene trees in a probabilistic framework was key in our decision to use it for our study.

Regarding the question of the species tree, it is very natural to argue that, given the level of introgression observed in the gambiae complex for example, an approach based on a starting species tree is questionable and phylogenetic networks could provide a better principled framework. However, current phylogenetic networks methods do not scale well to the number of species we consider in this work and the number of potential introgression events. Moreover, to maintain a reasonable computational complexity, they often require either prior potential reticulate edges or an upper bound on the number of reticulations to be given. It would be interesting to test more efficient pseudo-likelihood methods [21, 32] and methods jointly computing a species networks and gene trees [33, 28]. We nevertheless believe that an interesting feature of our approach is the ability to propose introgression events, that could be tested in a network framework. Last, the consistency analysis using MaxTiC also suggests that current models of phylogenetic networks might need further developments to account for extensive bidirectional and repeated hybridizations events that take place within a short amount of time, as is the case in the gambiae complex. For example, considering the significant level of introgression observed between the clades of *An. arabiensis* and *An. quadriannulatus* of *An. gambia* and *An. coluzzi* suggests that the speciations leading to these four species could have taken place over an extended period of time – which conflicts with the MaxTiC principle of ranking speciation events – during which extensive bidirectional introgression occurred.

The main issue with our approach concerns disentangling introgression from ILS. We followed an indirect approach based on synteny. This approach is further weakened by the fact that we use only the chromosomes of *An. gambia* for all potential introgression events, as it is the only fully assembled genome. This is especially questionable for the potential introgression involving *An. christyi*; but in this precise case, within the clade of Asian mosquitoes the best available assembled genome is *An. minimus*, which is fragmented in around one hundred scaffolds [1], which likely reduces its effectiveness as a support for a synteny analysis. Our synteny-based approach would then naturally benefit from better assembled genomes, including ancestral genomes [1].

Along the lines of making the most out of the reconciliation framework, the ability to model missing species, either because they are extinct or unsampled (called *ghost species*), is an intriguing avenue; the hypothesis that an ancient introgression event along the lineage of *An. christyi* would involve a ghost species is reasonable we believe. Two reconciliation methods, ALE and ecceTERA [12] can handle ghost species, although they require a dated species tree; this is another interesting future avenue to explore.

To conclude, we believe that our work demonstrates that a reconciliation-based approach to study introgression in larger data sets is worth exploring and several interesting methodological questions require further work such as integrating better ILS, networks and unresolved species phylogenies. From an applied point of view, the hypothesis of an introgression event between *An. christyi* and the Asian mosquitoes clade is an interesting case to study further.

**Acknowledgments.** CC is supported by Natural Science and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2017-03986. Most computations were done on the Cedar system of ComputeCanada through a resource allocation to CC. We thank Luay Nakhleh for useful feedback on an early draft of this work.

## References

1. Y. Anselmetti, W. Duchemin, E. Tannier, et al. Phylogenetic signal from rearrangements in 18 Anopheles species by joint scaffolding extant and ancestral genomes. *BMC Genomics*, 19(2):96, 2018.
2. Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.
3. P. D. Blischak, J. Chifman, A. D. Wolfe, and L. S. Kubatko. HyDe: A python package for genome-scale hybridization detection. *Systematic Biology*, 2018. Advance access, doi:10.1093/sysbio/syy023.
4. Y.-B. Chan, V. Ranwez, and C. Scornavacca. Inferring incomplete lineage sorting, duplications, transfers and losses with reconciliations. *Journal of Theoretical Biology*, 432:1 – 13, 2017.
5. C. Chauve, A. Rafiey, A. A. Davin, C. Scornavacca, et al. Maxtic: Fast ranking of a phylogenetic tree by maximum time consistency with lateral gene transfers, 2017. biorxiv 10.1101/127548. Reviewed at 10.24072/pci.evolbiol.100037.
6. J. H. Degnan. Modeling hybridization under the network multispecies coalescent. *Systematic Biology*, 2018. Advance access, doi:10.1093/sysbio/syy040.
7. E. Y. Durand, N. Patterson, D. Reich, and M. Slatkin. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28(8):2239–2252, 2011.
8. R. Elworth, C. Allen, T. Benedict, P. Dulworth, and L. Nakhleh. Dgen: A test statistic for detection of general introgression scenarios, 2018. biorxiv 10.1101/348649.
9. M. C. Fontaine, J. B. Pease, A. Steele, et al. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217):1258524, 2015.
10. C. Garros, L. Koekemoer, M. Coetzee, M. Coosemans, and S. Manguin. A single multiplex assay to identify major malaria vectors within the African Anopheles funestus and the Oriental An. minimus groups. *American Journal of Tropical Medicine and Hygiene*, 70:583–590, 2004.
11. B. R. Holland, S. Benthin, P. J. Lockhart, V. Moulton, and K. T. Huber. Using supernetworks to distinguish hybridization from lineage-sorting. *BMC Evolutionary Biology*, 8(1):202, 2008.
12. E. Jacox, C. Chauve, G. J. Szöllősi, Y. Ponty, and C. Scornavacca. eccetera: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, 32(13):2056–2058, 2016.
13. C. Lakner, P. Van Der Mark, J. P. Huelsenbeck, B. Larget, and F. Ronquist. Efficiency of Markov chain Monte Carlo tree proposals in bayesian phylogenetics. *Systematic biology*, 57(1):86–103, 2008.
14. K. J. Liu, J. Dai, K. Truong, Y. Song, M. H. Kohn, and L. Nakhleh. An hmm-based comparative genomic framework for detecting introgression in eukaryotes. *PLOS Computational Biology*, 10(6):1–13, 06 2014.
15. J. Mallet, N. Besansky, and M. W. Hahn. How reticulated are species? *BioEssays*, 38(2):140–149, 2015.
16. S. H. Martin and C. D. Jiggins. Interpreting the genomic landscape of introgression. *Current Opinion in Genetics & Development*, 47:69 – 74, 2017.
17. L. Nakhleh. Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in Ecology and Evolution*, 28(12):719–728, 2013.

18. J. B. Pease and M. W. Hahn. Detection and polarization of introgression in a five-taxon phylogeny. *Systematic Biology*, 64(4):651–662, 2015.
19. F. Ronquist, M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, et al. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542, 2012.
20. B. K. Rosenzweig, J. B. Pease, N. J. Besansky, and M. W. Hahn. Powerful methods for detecting introgressed regions from population genomic data. *Molecular Ecology*, 25(11):2387–2397, 2016.
21. C. Solìs-Lemus and C. Ané. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLOS Genetics*, 12(3):1–21, 03 2016.
22. F. Sousa, Y. J. K. Bertrand, J. J. Doyle, et al. Using genomic location and coalescent simulation to investigate gene tree discordance in *Medicago* l. *Systematic Biology*, 66(6):934–949, 2017.
23. M. Stolzer, H. Lai, M. Xu, D. Sathaye, B. Vernot, and D. Durand. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, 28(18):i409–i415, 2012.
24. G. J. Szöllosi, A. A. Davín, E. Tannier, V. Daubin, and B. Boussau. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678):20140335, 2015.
25. G. J. Szöllősi, W. Rosikiewicz, B. Boussau, E. Tannier, and V. Daubin. Efficient exploration of the space of reconciled gene trees. *Systematic Biology*, 62(6):901–912, 2013.
26. Y. Wang, X. Zhou, D. Yang, and A. Rokas. A genome-scale investigation of incongruence in culicidae mosquitoes. *Genome Biology and Evolution*, 7(12):3463–3471, 2015.
27. R. M. Waterhouse, F. Tegenfeldt, J. Li, et al. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Research*, 41(D1):D358–D365, 2012.
28. D. Wen and L. Nakhleh. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Systematic Biology*, 67(3):439–457, 2018.
29. D. Wen, Y. Yu, M. W. Hahn, and L. Nakhleh. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Molecular Ecology*, 25(11):2361–2372, 2016.
30. D. Wen, Y. Yu, J. Zhu, and L. Nakhleh. Inferring phylogenetic networks using PhyloNet. *Systematic Biology*, 67(4):35–40, 2018.
31. Y. Yu, R. M. Barnett, and L. Nakhleh. Parsimonious inference of hybridization in the presence of incomplete lineage sorting. *Systematic Biology*, 62(5):738–751, 2013.
32. Y. Yu and L. Nakhleh. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, 16(10):S10, 2015.
33. C. Zhang, H. A. Ogilvie, A. J. Drummond, and T. Stadler. Bayesian inference of species networks from multilocus sequence data. *Molecular Biology and Evolution*, 35(2):504–517, 2018.