An Efficient Statistical Method of Detecting Introgressive Events from Big Genomic Data

by

Jingxue Feng

B.Sc., Simon Fraser University, 2017

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Spring 2019

Approval

Name:	**Jingxue Feng**
Degree:	**Master of Science (Statistics)**
Title:	**An Efficient Statistical Method of Detecting Introgressive Events from Big Genomic Data**
Examining Committee:	**Chair:** Tim Swartz Professor
	Liangliang Wang Senior Supervisor Assistant Professor
	Cedric Chauve Supervisor Professor Department of Mathematics
	Lloyd T. Elliott Internal Examiner Assistant Professor
Date Defended:	**April 9th, 2019**

Abstract

Introgressive hybridization, also called introgression, is the gene flow from one species to another due to mating between species. The genetic signals of introgression are not always obviously observed. Current methods of detecting introgressive events rely on the analysis of orthologous markers, and therefore do not consider gene duplication and gene loss. Since introgression leaves a phylogenetic signal similar to horizontal gene transfer, introgression events can be detected under a gene tree-species tree reconciliation framework, which simultaneously accounts for evolutionary mechanisms including gene duplication, gene loss, and gene transfer. In this work, the reconciliation-based method has been applied to a large dataset of *Anopheles* mosquito genomes. We recover extensive introgression that occurs in gambiae complex, a group of African mosquitoes, although with some variations compared to previous reports. Our analysis results also imply a possible ancient introgression between the Asian and African mosquitoes.

Keywords: introgression; horizontal gene transfer; Bayesian phylogenetic inference; gene tree-species tree reconciliation; *Anopheles* mosquito genomes

Acknowledgements

This thesis is based on a research conducted to investigate introgression among different mosquito species, which is a joint work with Liangliang Wang and Cedric Chauve. I am grateful for a number of people who encourage and support me along the way of research.

Firstly, I would like to express sincere thanks to my senior supervisor Liangliang Wang. Thank her for offering me opportunities to learn further in statistics. Not only did she teach me advanced statistical knowledge, but she also encouraged me to build up soft skills. I truly appreciate her guidance when I felt lost on my way during research. Without her aid and support, it will be impossible for me to complete this thesis.

I would also like to thank Professor Cedric Chauve from Department of Mathematics, who jointly worked with me and Liangliang Wang for this research. Thanks for his brilliant ideas and computational skills contributed during the exploratory data analysis. He is very patient whenever I have any questions. It is my great honor to work with such an expert in evolutionary computation.

Furthermore, I really appreciate all graduate students in our department who have ever offered me help, including Chuyuan (Cherlane) Lin, Anqi (Angela) Chen, Haoyao Ruan, Jiarui (Erin) Zhang, Ying (Daisy) Yu, Mengyang (Chris) Li, and Yifan (Lucas) Wu. Special thanks go to Shijia Wang and Yuping Yang because of their selfless support and careful proofreading on this thesis.

Finally, I would like to give a big thank to my mother, my father and my fiance for their extensive support during my graduate study. Thanks for their listening and understanding when I shared my joys and sorrows with them along the journey.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Hybridization is a process involving the mating between two different groups of species [13]. One outcome of interspecific hybridization is introgression, or introgressive hybridization, which is the transfer of genetic material from a donor species to a receptor species. The process of introgression can be neutral without an effect on evolution, but it can also be adaptive to the changing environment and spread in the recipient pool. The latter phenomenon is referred to as "adaptive introgression" [28]. Detecting signals of introgression in genomic data becomes a very important question in the evolution of eukaryotic genomes [26]. Some relevant studies of adaptive introgression include mimicry pattern in butterflies [7] and poison resistance of house mice [38]. Recently, the evolution of a group of African *Anopheles* mosquitoes, known as the *gambiae complex*, is of interest. This species complex is composed of the most important malaria vectors in sub-Saharan Africa, although their vectorial capacities may vary between different species. In 2015, Fontaine *et al.* demonstrated that there is extensive introgression within the gambiae complex, which probably implies rapid acquisition of enhanced vectorial capacities [11]. They detected introgression by comparing mean of divergence times since introgression would reduce the species divergence times. The extent of introgression within the gambia complex was later confirmed by Wen *et al.* that employs phylogenetic network methods [48], although the suggested introgression events were not in full agreement with Fontaine *et al.* This thesis follows this line of work, aiming at detecting signals of introgression within a larger group of *Anopheles* mosquito genomes, covering both African and Asian mosquitoes.

There exists several methods that have been designed specifically to detect signals of introgression from genomic data, and they can be classified into two major groups: methods based on summary statistics, and methods based on evolutionary models.

In the first group, summary statistics-based methods aim at detecting introgression between two closely related sister lineages. These methods use population genomics data to

detect haplotype blocks [1] at a genetic distance lower than the expected distance if no introgression was involved. We refer to [35] for a recent discussion on these methods. When four species are considered, the most common summary statistic method is *D statistics* [36], also called the *ABBA BABA statistics*. This method records the frequency of evolutionary trees (gene tree) that are incongruent with a given species phylogeny (species tree) over several *loci* [2], and tests if the disagreement between the observed incongruent topologies is significant against a null hypothesis assuming that phylogenetic incongruence is only due to *incomplete lineage sorting* (ILS) [3]. There exists relevant methods that extend this method to handle more than four species [32, 10], although at a significant computational cost. A common feature of these methods is that they aim at disentangling two evolutionary processes, ILS and introgression, because both of them can result in gene trees that are incongruent with the species tree.

Another group of methods detect introgression using phylogenetic networks by modeling ILS and hybridization simultaneously. This model-based approach has been implemented in combinatorial approaches [16, 52] and probabilistic frameworks [24, 47, 49, 53]. We refer the reader to [8] for a recent discussion 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. The model-based methods 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 [37, 51].

An important drawback of the methods outlined above is that they are limited to the analysis of orthologous genes [4], 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 [29], introgression through hybridization leaves a phylogenetic signal similar to *horizontal gene transfer* (HGT), although they are different from the biological point of view: introgression occurs in sexual species, whereas HGT occurs in asexual species. HGT is an evolutionary mechanism well handled by several efficient species tree-gene tree reconciliation algorithms [19, 40, 43, 41] that scale well to large data sets. It suggests that the framework of reconciling gene trees with a known species tree could

[1] A set of genes that tend to be inherited together through the evolution.

[2] A locus (plural loci) is a fixed location on the chromosome where the gene is found.

[3] Incomplete lineage sorting occurs when the gene copies fail to coalesce at the time of speciation but coalesce in an ancestral species if looking backwords in time.

[4] Orthologous genes are genes in different species that originate from a common ancestral DNA sequence. Their functions are similar to the functions of the ancestral genes.

be used for detecting introgression without filtering out paralogous genes [5]. We refer readers to Section 3.2 for details of the species tree-gene tree reconciliation. In the present work, we explore this idea, and apply a reconciliation-based method to detect signals of introgression over a large data set of 14 *Anopheles* genomes covering both African and Asian mosquitoes, including the gambiae complex.

The outline of this dissertation is organized as follows. Some background on phylogenetics and the *Anopheles* data set are discussed in Chapter 2. The data we started with is composed of the full genome sequences of 14 *Anopheles* species assembled at various contiguous levels. In Chapter 3, we first rely on MrBayes, which is a software program for Bayesian inference of phylogeny, to perform Bayesian inference of phylogenetic parameters on the DNA sequence alignments in order to sample gene trees. Then, we implement a probabilistic approach, called *amalgamated likelihood estimation* (ALE), to reconstruct gene trees that can be pieced together as a combination of clades in an evolutionary model accounting for gene duplication, gene loss and HGT. In order to distinguish introgression from ILS, we propose the hypothesis that introgression concerns blocks of contiguous genes, while ILS acts randomly along the chromosome, as discussed in [39]. A statistical test is then applied to detect genome regions with significantly more genes whose evolution involves a signal of HGT along chromosomes. In Chapter 4, we discuss the experimental results step by step, and display the signals of introgression along the chromosomes of several *Anopheles* species. Our results further confirm the extensive level of introgression within gambiae complex discussed in [11], although with some differences related to certain introgressed genome segments. In addition, we also find evidence for a potential ancient introgression event involving an African mosquito lineage and the most common ancestor of the clade of Asian *Anopheles*, which has not been discovered before. Ultimately, several concerns of current work and necessary future studies are discussed in Chapter 5.

[5]Unlike orthougous genes, paralogous genes are new genes and have new functions. They are diverged within one species.

Chapter 2

Materials

This chapter reviews some background on phylogenetics and introduces *Anopheles* data we used to detect introgression, including a species tree and full genome sequences for 14 *Anopheles* species. The species tree is taken from [1] without given branch lengths because the branching pattern remains highly debated [11, 45]. The 14 genomes assembled at different contiguous levels are obtained from the results of genome scaffolding illustrated in [1].

2.1 Background on Phylogenetics

Phylogenetic tree The objective of phylogenetics is to study the evolutionary relationships among different species or individuals, called *taxa*, based on biological sequence alignments [18]. The evolutionary relationships are usually represented by a *phylogenetic tree*, which is a binary tree structure composed of nodes and branches. An example is given in Figure 2.1. We call the black dots in the tree *nodes*. The external nodes or leaves of the phylogenetic tree — Taxon 1, Taxon 2, Taxon 3, Taxon 4, and Taxon 5 — represent extant or present-day organisms, and the internal nodes represent hypothetical ancestors splitting lineages where a *speciation* event occurs. The edge that connects a descendent and an ancestor is called a *branch*.

The phylogenetic trees are divided into *rooted trees* and *unrooted trees* on the basis of the presence or absence of the *ancestral root*. Figure 2.1 displays a rooted tree. The node at the top of the tree represents the most recent common ancestor of 5 taxa, called the root. The branches of the rooted tree contain the direction of time: we move forward in time when we move from the root to the leaves. By contrast, for an unrooted tree shown in Figure 2.2, there does not exist an ancestral root, and the branches are undirected in time. However, many phylogenetic analyses include an *outgroup* to regain a root in the unrooted tree. An outgroup is a taxon (or a set of taxa) that are less related to other taxa used in the study, and the root will therefore be placed on the branch between the other taxa and

the outgroup. For instance, in Figure 2.2, if taxon 5 is an outgroup, we can place the root on the branch leading to taxon 5 to obtain the rooted tree in Figure 2.1.

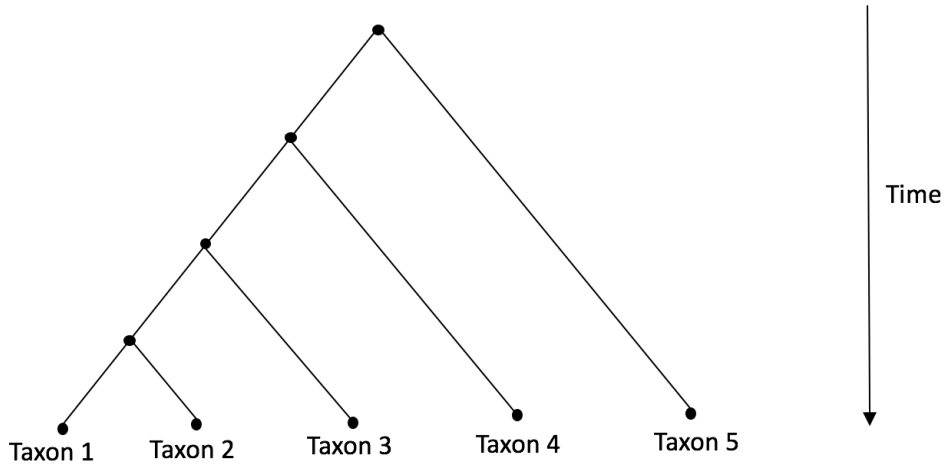

Figure 2.1: Rooted phylogenetic tree of 5 taxa. External nodes or leaves represent extant or present-day organisms. Internal nodes represent hypothetical ancestors splitting lineages (so-called speciation events). The branches are related to the direction of time: as we move from the root to the leaves, we move forward in time.

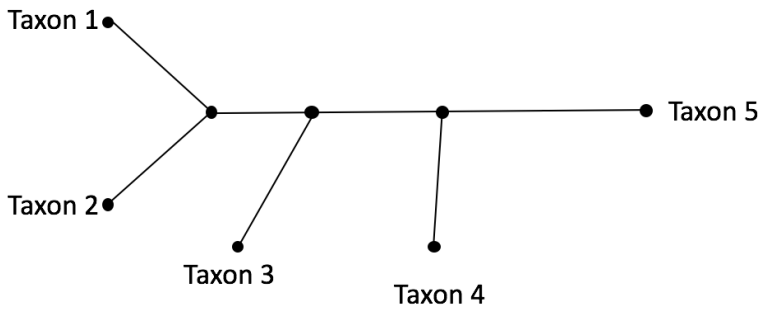

Figure 2.2: Unrooted phylogenetic tree of 5 taxa. External nodes or leaves represent contemporary organisms. Internal nodes represent hypothetical ancestors splitting lineages (so-called speciation events). The branches are not related to the direction of time.

Typically, the shape of a phylogenetic tree is conveyed by two major types of information. One is the *topology* that defines the branching pattern of the tree and the distribution of extant species among the leaves. The other is the *branch lengths* that represent evolutionary time, measured by the average amount of mutational change. These two components are considered critically when we use likelihood models to analyze a topology.

Data used in phylogenetic analysis Although the data used earlier for phylogenetic analysis was based on physical features of the organisms under consideration (i.e. morphological data), contemporary studies pay more attention to building phylogenetic trees using

molecular sequence data, mainly *DNA sequences*. DNA, or deoxyribonucleic acid, is a two-stranded twisted molecule that contains unique genetic code for each organism. DNA is made up of 4 nucleotides or bases: *Adenine* (A), *Guanine* (G), *Cytosine* (C), and *Thymine* (T). The order of these nucleotides determines the genetic code of DNA. The process of determining the order of these bases in a piece of DNA is called *DNA sequencing*.

When we obtain the DNA sequences of some taxa, we need to compare homologous nucleotides that have been acquired directly from the common ancestor of the taxa of interest. The process of finding out which regions of a set of DNA sequences are homologous and should be compared is called *DNA sequence alignment*. An example of the DNA sequence alignment for five taxa is given in Table 2.1. Different rows represent different species, and different positions in each DNA sequence represent different *sites*. All nucleotides in Table 2.1 are identical at most of the sites, which reflects the fact that the sequences compared are homologous. The differences at certain sites result from mutational changes during evolution. For instance, in the fourth column in Table 2.1, taxa 3, 4 and 5 have base C, whereas taxon 1 has a base A and taxon 2 has a base G. This reflects a fact that taxa 3, 4 and 5 are more closely related to each other than the remaining taxa. The objective of phylogenetic analysis is to recognize significant genomic similarities between distant species and create phylogenetic trees to explain these relationships. Since the amount of molecular data sets is usually very large, and more and more data is gathered nowadays, it is therefore necessary to use computational statistics in phylogenetic tree inference. Here, we focus on a relatively new probabilistic approach to infer phylogenetic trees — Bayesian inference, demonstrated in Section 3.1.

Taxon 1	...TGTATCGCTC...
Taxon 2	...TGTGTCGCTC...
Taxon 3	...AGTCTCGTTC...
Taxon 4	...TGTCTCGTTT...
Taxon 5	...AGTCTCATTC...

Table 2.1: A small section of the hypothetical DNA sequence alignment for five taxa. Rows represent taxa, and columns represent sequence positions.

Gene tree-species tree incongruence A phylogenetic tree (gene tree) constructed from DNA sequences for a genetic locus sometimes does not agree with the tree that represents the actual population split history (species tree). The incongruence between gene trees and a species tree might be due to gene duplication and loss, ILS, and HGT. Lineage sorting occurs because each individual in one population randomly contributes genetic material to the next generation [30]. ILS occurs when an ancestral species experiences several speciation

events in a short period of time. As shown in Figure 2.3, gene tree GT_1 is identical to the species tree (shaded tube), while gene tree GT_2 is not identical to the species tree due to ILS. The evolutionary scenarios involving gene duplication, loss and HGT that result in phylogenetic incongruence is demonstrated in Section 3.2.

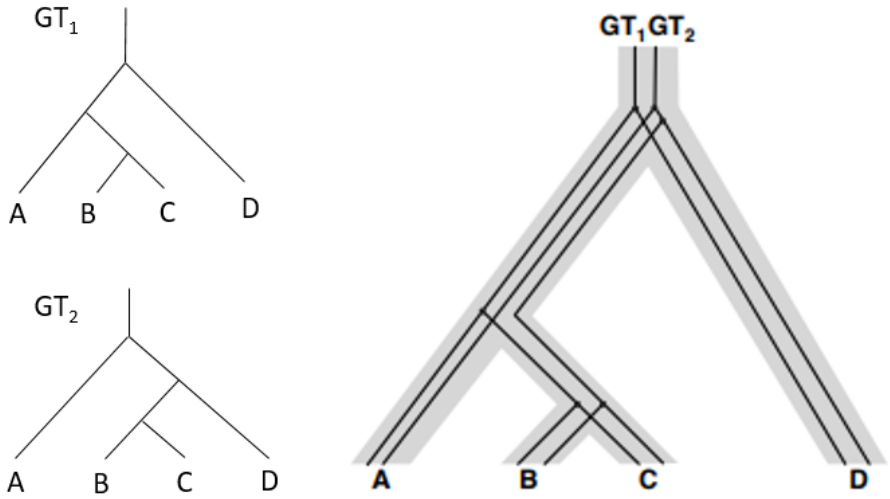

Figure 2.3: Incomplete lineage sorting that results in gene tree-species tree incongruence, adopted from Figure 1.1 in [30]. The topology of gene tree GT_1 is identical to that of the species phylogeny (shaded tube), whereas gene tree GT_2 is incongruent with the species phylogeny due to ILS.

2.2 *Anopheles* Data

Anopheles is a genus of mosquito transmits parasites between people resulting the spread of malaria in humans. Various *Anopheles* species are found to be the malaria vectors in different parts of the world [11]. The gambiae complex contains several closely related Afrotropical mosquito sibling species, and it has become the most important malaria vector [31]. At the same time, detecting introgression among those malaria vectors is of particular interest.

Genome sequences A genome contains all DNA of an organism. To sequence an entire genome, the DNA of the genome is broken into smaller pieces, the pieces are sequenced, and then the sequences are assembled into a single long "consensus". Our data starts from the entire genome sequences of 14 *Anopheles* species that are geographically distributed in sub-Saharan Africa and Asia, including *Anopheles gambia* (AGAMB), *Anopheles coluzzii*(ACOLU), *Anopheles arabiensis* (AARAB), *Anopheles quadriannulatus* (AQUAD), *Anopheles melas* (AMELA), *Anopheles merus* (AMERU), *Anopheles christyi* (ACHRI), *Anopheles epiroticus* (AEPIR), *Anopheles stephensi India* (ASTEI), *Anopheles stephensisensu stricto* (ASTES), *Anopheles maculatus* (AMACU), *Anopheles culicifacies* (ACULI), *Anopheles minimus* (AMINI), and *Anopheles funestus* (AFUNE). Only the genome of

Anopheles gambia is fully assembled among the 14 genomes. The genomes of other *Anopheles* species are fragmented assemblies. We refer the readers to [1] for a more precise discussion on the assembly of these genomes.

A gene is a small part of the genome, and genes are made of DNA. Genes can determine the characteristics of an organism, and different species have different number of genes in the genome. The considered genomes of 14 *Anopheles* species contain from $10,000$ to $15,000$ genes, as given in Figure 2.4.

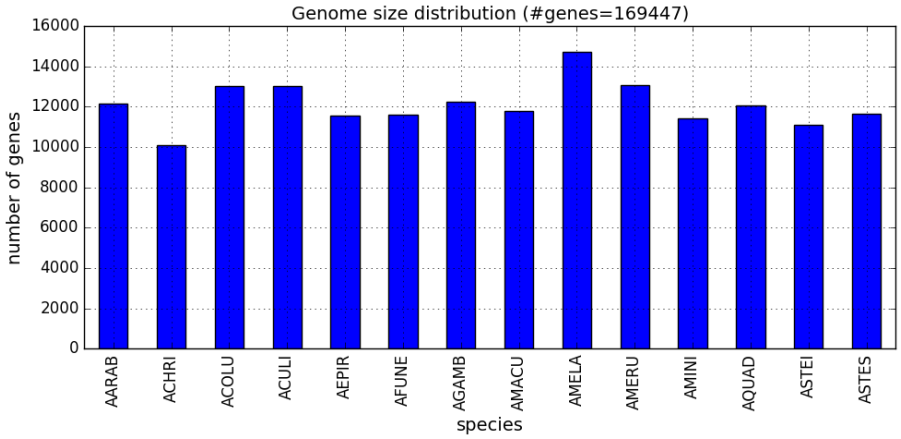

Figure 2.4: Number of genes per genome for each *Anopheles* species.

Gene Families A *gene family* is a group of related genes that share a common ancestor. The genes are similar in structure and in biochemical function within a gene family. Defining gene families is a necessary step for building gene trees. The genes from the 14 *Anopheles* species had been clustered into more than 17,000 homologous gene families based on their common ancestry using the OrthoDB algorithm demonstrated in [46].

Figure 2.5 shows the distribution of the sizes (number of genes) of gene families. It can be seen that there is a peak for gene families with 14 genes. An important observation is that there is a large number of small-size gene families, which is probably due to errors in assembling genes or clustering genes into homologous families [1]. Within each gene family, a *multiple sequence alignment* (MSA) was obtained using the method in [1], which had been converted to an aligned matrix in the standard NEXUS format prior to analysis. In Appendix B.2, we display an example of aligned matrix in NEXUS format taken from a gene family.

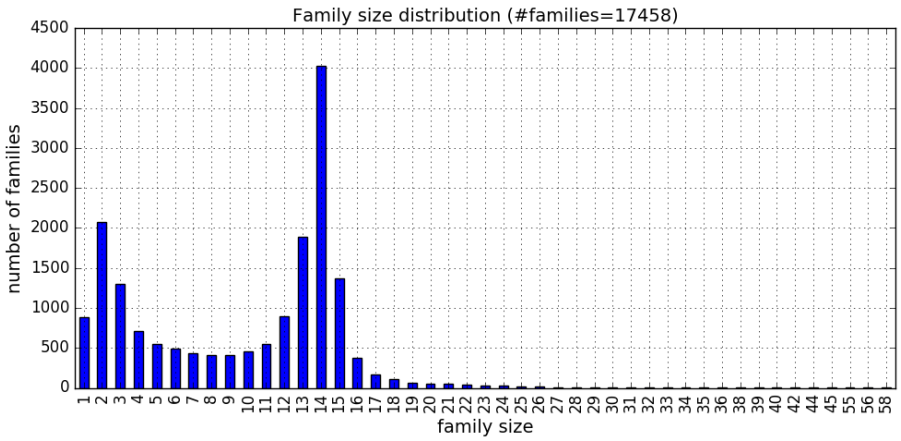

Figure 2.5: The sizes (number of genes) of gene families.

Species Tree A species tree demonstrates the evolutionary relationships among a group of taxa that are believed to have a common ancestor. The 14 considered *Anopheles* species can be classified into three groups:

- The gambiae complex composed of six sibling species, including *Anopheles gambia* (AGAMB), *Anopheles coluzzii*(ACOLU), *Anopheles arabiensis* (AARAB), *Anopheles quadriannulatus* (AQUAD), *Anopheles melas* (AMELA), and *Anopheles merus* (AMERU);
- Two outgroups to the gambiae complex: *Anopheles christyi* (ACHRI) in Africa and *Anopheles epiroticus* (AEPIR) in Asia. They serve as a reference group when we study the evolutionary relationships of the ingroup;
- A subtree composed of six Asian mosquitoes, including *Anopheles stephensi India* (ASTEI), *Anopheles stephensisensu stricto* (ASTES), *Anopheles maculatus* (AMACU), *Anopheles culicifacies* (ACULI), *Anopheles minimus* (AMINI), and *Anopheles funestus* (AFUNE), where the African *Anopheles funestus* and Oriental *Anopheles minimus* are genetically closely related [12]; from now this group is called *Asian clade*.

The rooted binary species tree of these 14 considered *Anopheles* species is shown in Figure 2.6. It is same as the tree in [1], namely "X-phylogeny", because the species phylogeny is constructed based on X chromosome genes. The species phylogeny within the gambiae complex is taken from [11]. In our experiment, we do not consider the dating information of the interior nodes of the species tree, i.e. without given information of branch lengths, except ensuring that the descendants appear later than their ancestors.

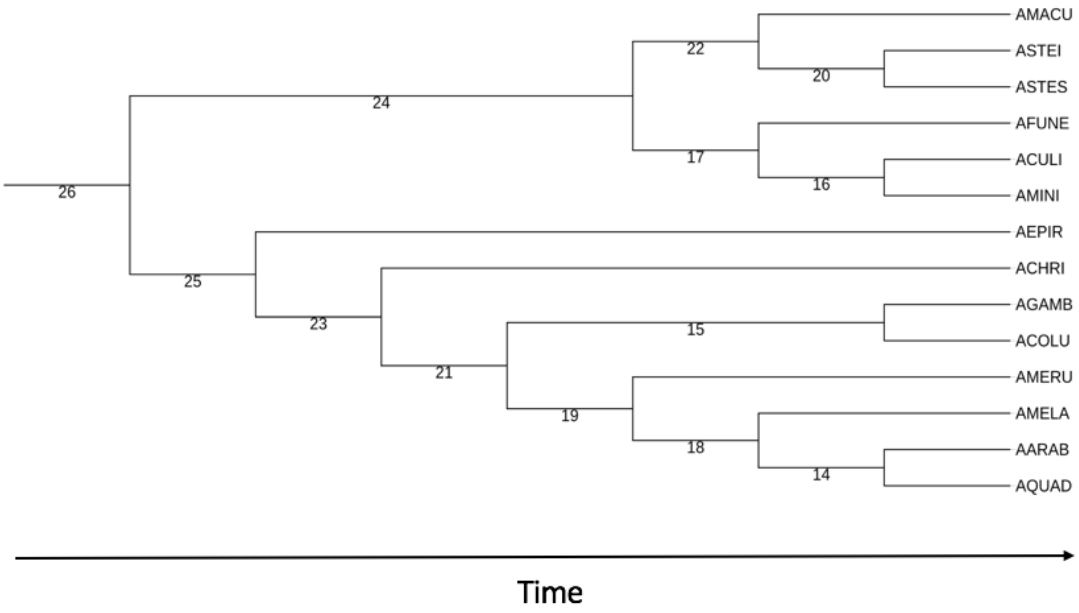

Figure 2.6: Species tree of the 14 *Anopheles* species considered in our study. Numbers on the branches of the tree represent ancestors.

Chapter 3

Methods

The methodologies we used to detect introgression are composed of three major steps. First, we use a MrBayes to obtain a sample of gene trees, and take the sampled gene trees as input in ALE to explore the space of sampled reconciled gene trees in a reconciliation model accounting for gene duplication, gene loss, and HGT. Then, we rely on an algorithm of time consistency to analyze the level of noises for the inferred HGTs. Eventually, based on the robust HGTs, we perform statistical hypothesis tests on the genome segments where a large number of HGT-like genes indicate potential signals of introgression.

3.1 MrBayes: Sampling Gene Trees

Given the observed MSA for each gene family, the first step is to estimate phylogenetic trees that best explain the genetic information. MrBayes is a popular program for doing Bayesian inference of phylogeny [34]. We first used MrBayes to approximate the posterior distribution of phylogenetic trees by sampling gene trees through *Markov Chain Monte Carlo (MCMC)* method. Other user-friendly software programs that can handle Bayesian phylogenies include BEAST [9], RevBayes [15], and BAMBE [22]. The Bayesian approach to phylogenetics is outlined in this section.

Bayesian inference of phylogeny The inferences of phylogeny in a Bayesian analysis is based on computing posterior probabilities of phylogenetic trees [17]. Let's consider τ_i as the topology of ith phylogenetic tree, $i = 1, 2, 3, ..., N(s)$, where $N(s)$ denotes the number of possible trees for s species. Let c denote the number of aligned position in DNA sequences (i.e. number of sites), and $\boldsymbol{D}$ is the matrix of s aligned DNA sequences of length c. To be specific, we can describe the matrix $\boldsymbol{D}_{s\times c}$ as

$$
\begin{aligned}
\boldsymbol{D} &= \{d_{ij}\} \\
&= \begin{Bmatrix} \boldsymbol{d}_1 & \boldsymbol{d}_2 & \boldsymbol{d}_3 & \dots & \boldsymbol{d}_{c-2} & \boldsymbol{d}_{c-1} & \boldsymbol{d}_c \end{Bmatrix} \\
&= \begin{array}{c} \begin{array}{ccccccc} \text{site } 1 & \text{site } 2 & \text{site } 3 & \dots & \text{site } c-2 & \text{site } c-1 & \text{site } c \end{array} \\ \left\{ \begin{array}{ccccccc} A & T & G & \dots & G & A & C \\ A & T & C & \dots & C & G & C \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ A & T & C & \dots & C & G & T \end{array} \right\} \begin{array}{c} \text{species } 1 \\ \text{species } 2 \\ \vdots \\ \text{species } s \end{array} \end{array},
\end{aligned}
$$

where $\boldsymbol{d}_i$ $(1 \leq i \leq c)$ denotes the observations at the ith site or ith column of the alignment. For example, we can observe $\boldsymbol{d}_1 = \begin{Bmatrix} A & A & \dots & A \end{Bmatrix}'$ at the 1st site. Each $\boldsymbol{d}_i$ is associated to the leaves of the tree. The sites are assumed to be independent in this dissertation.

For each gene family with the observed MSA, we first want to find the posterior distribution of phylogenetic tree topologies τ. According to Bayes theorem, the posterior probability of the ith phylogenetic tree topology τ_i conditional on the current data $\boldsymbol{D}$ is

$$
f(\tau_i|\boldsymbol{D}) = \frac{f(\boldsymbol{D}|\tau_i)f(\tau_i)}{\sum_{j=1}^{N(s)} f(\boldsymbol{D}|\tau_j)f(\tau_j)}, \tag{3.1}
$$

where $f(\boldsymbol{D}|\tau_i)$ is the likelihood of the ith tree topology, $f(\tau_i)$ is the prior probability of the ith tree topology. The denominator is the sum over all possible $N(s) = \frac{(2s-5)!}{2^{s-3}(s-3)!}$ unrooted trees for s species, which makes the posterior probability analytically impossible to calculate [17]. Thus, we used the *Metropolis-Hasting* (MH) algorithm to approximate the posterior probabilities, as described in Algorithm 1. The calculation of likelihood depends on some unknown parameters. A brief introduction of those parameters will be provided in the following subsections.

DNA sequence evolution model To calculate the likelihood function $f(\boldsymbol{D}|\tau_i)$, we assume the tree τ_i comes with a set of branch lengths in $\boldsymbol{v}_i$ and a stochastic model of DNA substitution. MrBayes was run using the *General Time Reversible* (GTR) model of sequence evolution with a proportion of *invariable sites* and a *Gamma-shaped distribution* of rates across sites, namely a GTR + I + Γ model.

The GTR model was first generally demonstrated by Simon Tavaré in 1986 [44]. The core of the GTR model is to specify the (unscaled) instantaneous rate of DNA substitution as a 4×4 stochastic matrix $\boldsymbol{Q}$:

$$\boldsymbol{Q} = \{q_{ij}\} = \begin{array}{c} \\ [A] \\ [C] \\ [G] \\ [T] \end{array} \begin{array}{c} \begin{array}{cccc} [A] & [C] & [G] & [T] \end{array} \\ \begin{pmatrix} . & \pi_C \cdot r_{A \rightleftharpoons C} & \pi_G \cdot r_{A \rightleftharpoons G} & \pi_T \cdot r_{A \rightleftharpoons T} \\ \pi_A \cdot r_{A \rightleftharpoons C} & . & \pi_G \cdot r_{C \rightleftharpoons G} & \pi_T \cdot r_{C \rightleftharpoons T} \\ \pi_A \cdot r_{A \rightleftharpoons G} & \pi_C \cdot r_{C \rightleftharpoons G} & . & \pi_T \cdot r_{G \rightleftharpoons T} \\ \pi_A \cdot r_{A \rightleftharpoons T} & \pi_C \cdot r_{C \rightleftharpoons T} & \pi_G \cdot r_{G \rightleftharpoons T} & . \end{pmatrix} \end{array}$$

where $q_{ij}(i \neq j)$ indicates the rate of change from nucleotide i to nucleotide j, and $i, j \in \{A, C, G, T\}$. The model is "time reversible" because the substitution from nucleotide i to nucleotide j have the same rate of substitution from nucleotide j to nucleotide i. The GTR model parameters are composed of the equilibrium base frequencies for four nucleotides, $\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$, indicating the frequency at which each base occurs at each site; and six substitution rate parameters, $\boldsymbol{r} = (r_{A \rightleftharpoons C}, r_{A \rightleftharpoons G}, r_{A \rightleftharpoons T}, r_{C \rightleftharpoons G}, r_{C \rightleftharpoons T}, r_{G \rightleftharpoons T})$, indicating the rate of replacement of one base by another. The diagonal elements are specified as the negative sum of the elements in each row, which means $q_{ii} = -\sum_{j \neq i} q_{ij}$ so that the sum of each row equals 0. For example, the dot on the first row represemts $-(\pi_C \cdot r_{A \rightleftharpoons C} + \pi_G \cdot r_{A \rightleftharpoons G} + \pi_T \cdot r_{A \rightleftharpoons T})$.

By default, MrBayes assumes equal rates of substitution among nucleotide sites [34]. It is unrealistic because different sites have different functional roles in the gene and therefore lead to substitution-rate variations. Instead, we run MrBayes using the $I + \Gamma$ model of sequence evolution with rate heterogeneity among sites. It is composed of two layers: the first layer assumes a proportion (λ) of sites are invariable while the other $(1 - \lambda)$ sites may evolve over time; the second layer assumes the substitution rate to be drawn from a gamma distribution, based on a discrete gamma model suggested by Yang [50]. In the discrete gamma model, they control the mean of gamma distribution to be 1, and set the scale and the shape parameters to be equal in order to avoid using too many parameters. In other words, the extent of rate variation is determined by a single shape parameter α, and a small value of α indicates a significant rate variation among sites. To sum up, the $I + \Gamma$ model of sequence evolution has two parameters, including the proportion of invariable sites (λ) and the shape parameter of the gamma distribution (α).

Prior probability distribution From Bayesian perspective, the parameters are viewed as random variables, therefore a prior probability density is required to be specified to cover the nature of the random variation. In summary, there are six types of parameters to set priors in the phylogenetic model : topology (τ_i), branch lengths ($\boldsymbol{v}_i$), equilibrium frequencies of the nucleotides ($\boldsymbol{\pi}$), nucleotide substitution rates ($\boldsymbol{r}$), the proportion of invariable sites

(λ), and the shape parameter of the gamma distribution that determines rate variation (α).

We used a default setting on those priors in MrBayes since they work well for most analyses. By default, MrBayes has a uniform setting on topologies by assigning equal prior probabilities to all different fully-resolved topologies τ_i. For the branch lengths $\boldsymbol{v}_i$, they are assumed to follow an exponential distribution with a rate parameter 10 (mean 0.1), i.e. $\boldsymbol{v}_i \sim \text{Exp}(10)$, allowing values of $\boldsymbol{v}_i$ to range from 0 to ∞. The stationary nucleotide frequencies $\boldsymbol{\pi}$ and the nucleotide substitution rates $\boldsymbol{r}$ of the GTR rate matrix are assumed to follow a "flat" Dirichlet distribution respectively, with all distribution parameters equal to 1 [23]:

$$\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T) \sim \text{Dir}(1, 1, 1, 1),$$

and

$$\boldsymbol{r} = (r_{A\rightleftharpoons C}, r_{A\rightleftharpoons G}, r_{A\rightleftharpoons T}, r_{C\rightleftharpoons G}, r_{C\rightleftharpoons T}, r_{G\rightleftharpoons T}) \sim \text{Dir}(1, 1, 1, 1, 1, 1),$$

which allows equal rates of change between nucleotides. The proportion of invariable site λ is assumed to be uniformly distributed between 0 and 1:

$$\lambda \sim \text{Uniform}(0,1).$$

For the shape parameter α from the gamma distribution of rate variation, we did not have good prior knowledge about the variance in site rates, thus we place an uninformative prior on α, which is an exponential distribution with rate parameter 1:

$$\alpha \sim \text{Exp}(1).$$

MCMC sampling method Now, the phylogenetic model gives us $f(\boldsymbol{D}|\tau_i, \boldsymbol{v}_i, \boldsymbol{\pi}, \boldsymbol{r}, \lambda, \alpha)$ instead of the likelihood $f(\boldsymbol{D}|\tau_i)$ from Bayes equation (3.1). Let $\boldsymbol{\theta}$ denote the parameters in GTR model, i.e. $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{r}\}$. Under the assumption of independent sites, the likelihood $f(\boldsymbol{D}|\tau_i)$ from Bayes equation (3.1) becomes

$$f(\boldsymbol{D}|\tau_i) = \iiiint f(\boldsymbol{D}|\tau_i, \boldsymbol{v}_i, \boldsymbol{\theta}, \lambda, \alpha) f(\boldsymbol{v}_i) f(\boldsymbol{\theta}) f(\lambda) f(\alpha) d\boldsymbol{v}_i \, d\boldsymbol{\theta} \, d\lambda \, d\alpha, \tag{3.2}$$

where $f(\boldsymbol{v}_i), f(\boldsymbol{\theta}), f(\lambda)$, and $f(\alpha)$ are our prior beliefs mentioned previously. Computing equation (3.2) is difficult because it involves multidimensional integrals, including discrete and continuous components, and therefore is too expensive to calculate.

Usually, the posterior distribution $f(\tau_i|\boldsymbol{D})$ is a complicated function over a large parameter space that cannot be calculated analytically. However, the posterior probability of phylogenies can be estimated through sampling trees from the posterior probability distribution [17]. We can simulate tree samples from the target distribution, $f(\tau_i|\boldsymbol{D})$, by constructing an

ergodic *Markov Chain* [1] such that this chain finally converges to the stationary distribution $f(\tau_i|\boldsymbol{D})$. *Markov Chain Monte Carlo* (MCMC) can be used to sample phylogenies according to the target distribution. To construct the Markov chain, the MCMC method we used here to approximate $f(\tau_i|\boldsymbol{D})$ is the *Metropolis-Hastings algorithm*.

Let $\Psi = \{\tau, \boldsymbol{v}, \boldsymbol{\theta}, \lambda, \alpha\}$ be a tree with branch lengths and model parameters, which denotes the current state of the chain. Let $Q(\Psi'|\Psi)$ denote the proposal distribution defined as the probability of proposing the new state Ψ' given the current state Ψ; while $Q(\Psi|\Psi')$ is defined reversely. The Metropolis-Hastings algorithm generates a collection of states through the following transitions:

Algorithm 1 Metropolis-Hastings Algorithm

1: Initialize the Markov chain at state $\Psi^{(0)}$.

2: Iteration for $q = 1, 2, ..., N$.

- Draw a new state $\Psi^{(q)}$ randomly from a proposal distribution $Q(\Psi^{(q)}|\Psi^{(q-1)})$.
- Calculate the acceptance probability $r(\Psi^{(q)}, \Psi^{(q-1)}) = \min(1, \frac{f(\boldsymbol{D}|\Psi^{(q)})}{f(\boldsymbol{D}|\Psi^{(q-1)})} \times \frac{f(\Psi^{(q)})}{f(\Psi^{(q-1)})} \times \frac{Q(\Psi^{(q-1)}|\Psi^{(q)})}{Q(\Psi^{(q)}|\Psi^{(q-1)})})$.
- Draw a random variable u from Uniform(0,1).
- If $u \leq r(\Psi^{(q)}, \Psi^{(q-1)})$, Move to the new state $\Psi^{(q)}$; otherwise remain in current state $\Psi^{(q-1)}$.

Eventually, the sequence of visited states forms an MCMC chain. The stationary distribution of the chain is the joint probability density of tree topologies, branch lengths, and substitution parameters. The inferences of phylogenetic trees can be saved during the course of MCMC analysis. The initial portion of an MCMC sample will be discarded to allow the Markov chain to reach stationarity, so-called *burn-in* period. We focus on those trees being generated after the burn-in period.

Convergence diagnostics For each gene family, two independent MCMC chains started from different random trees are computed with a burn-in period of 25% samples. The tree samples were saved (i.e., thinned) every 500 iterations, and we ended the Markov chain after $N = 10,000,000$ iterations, leading to a maximum of $20,000$ sampled gene trees. The gene families in which both MCMC chains generated less than $5,000$ samples were discarded from further analysis. The assessment of convergence in MrBayes is based on *average standard deviation in split frequencies* (ASDSF), which averages the standard deviations of the

[1] A Markov chain is a chain system that contains transitions from one state to another, where the probability of the future state of the system is dependent only on the current state of the system.

frequency of each clade across all runs. As the ASDSF approaches 0, the MCMC chain converges onto the stationary distribution [20, 34]. We used 0.01 as a threshold of ASDSF to determine if the MCMC chain has converged or not. These sampled gene trees with ASDSF below 0.01 are referred as *MrBayes trees* representing the evolution of the corresponding gene family. Finally, two sets of MrBayes trees were generated for each gene family.

3.2 ALE: Reconciling Sampled Gene Trees with the Species Tree to Infer HGTs

Next, the obtained MrBayes trees and the rooted species tree were provided as input to ALE for amalgamated likelihood estimation. ALE is a probabilistic approach that explores all reconciled gene trees that can be pieced together as a combination of clades observed in MrBayes trees [41]. It is implemented by accounting for gene duplication, gene loss and HGT [2]. According to a simulation study over 36 cyanobacteria genomes in [41], the gene trees reconstructed based on ALE approach are dramatically more accurate than those reconstructed based on molecular sequences alone. In this section, we provide a general idea of ALE approach. For the open source implementation of ALE, please refer to `https://github.com/ssolo/ALE`.

Gene tree-species tree reconciliation ALE is implemented in the context of a gene tree-species tree reconciliation framework, which allows for gene duplication, loss and transfer (DTL). A species tree is a phylogenetic tree that demonstrates the actual evolutionary history of the species. The internal nodes of a species tree correspond to species-level events, such as speciation events. A gene tree is a phylogenetic tree that represents the evolutionary pathway of the genes included in the study. Gene trees can provide information about both gene-level events and species-level events. Therefore, a gene tree and a species tree can be incongruent for many reasons including DTL events. Reconciling a gene tree with a species tree requires a mapping of each node of the gene tree to the nodes or branches of the species tree using a series of gene-level and species-level events [25]. An example of gene tree-species tree reconciliation invoking speciation and DTL events is given in Figure 3.1. The speciation events occur at the internal nodes of S. A gene transfer from species A to B results in a new gene copy b_1. And introgression in sexual species might leave a similar signal to this HGT. Gene duplication and loss events are also inferred during reconciliation. The gene tree that evolves inside a species tree is called a *reconciled gene tree*. Note that a

[2] As mentioned in Chapter 1, introgression and HGT leave similar genomic signatures although they are different evolutionary mechanisms. Since HGT is also well handled by several algorithms, one way to detect signals of introgression is to find out the HGT between two *Anopheles* species.

reconciliation may include one or more reconciled gene trees along with the species tree.

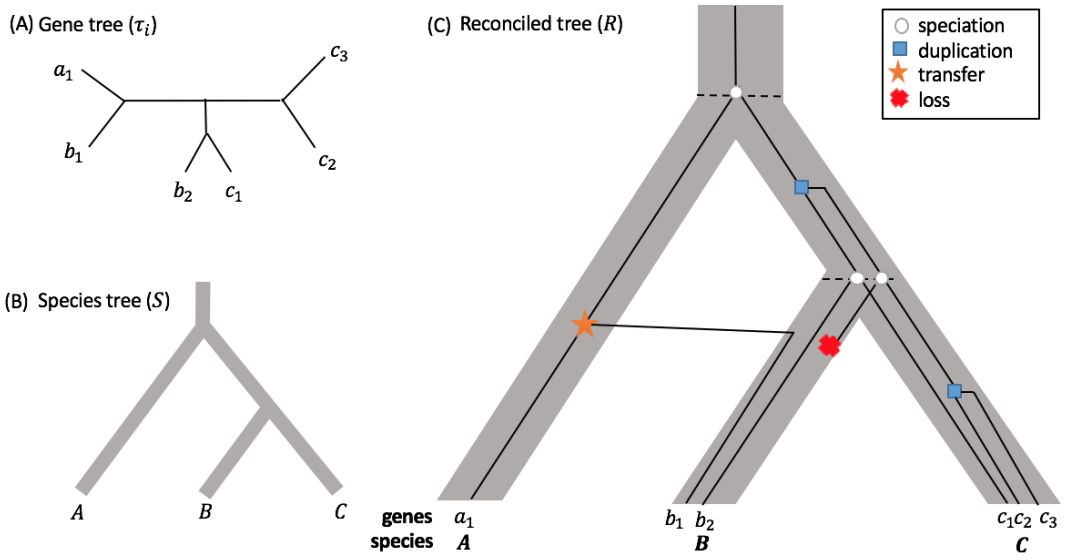

Figure 3.1: Reconciled tree R between an unrooted gene tree τ_i and a species tree S in a three-taxon scenario. (A) A gene tree topology. Each gene is named with the lowercase letter of the corresponding species. (B) A species phylogeny consists of species A, B and C. (C) Reconciled tree R is represented as a mapping of the nodes of the gene tree in (A) onto the nodes or branches of the species tree in (B) using three speciation events (white circle), two events of gene duplication (blue square), one event of gene loss (red cross), and one event of gene transfer (orange star).

Conditional Clade Probability Although MCMC sampling techniques are commonly used for estimating the posterior distributions of phylogenetic trees, a drawback of these methods is that the MrBayes trees sampled from the converged MCMC run only takes a small fraction of the total tree space, leading to inaccurate approximations of posterior probability distribution calculated with simple sample relative frequencies [21]. For example, given a gene family with 14 taxa, the total number of possible unrooted trees is $N(14) = 3.16234143 \times 10^{11}$, but the MCMC samples contain at most 20,000 of these trees. To reduce the uncertainty, the posterior probability distribution of trees recorded from MCMC can be accurately approximated by *conditional clade probability* (CCP) method introduced in [14].

A clade is a group of organisms that includes a common ancestor and all its descendents. The species within one clade are more closely related to each other than they are to those

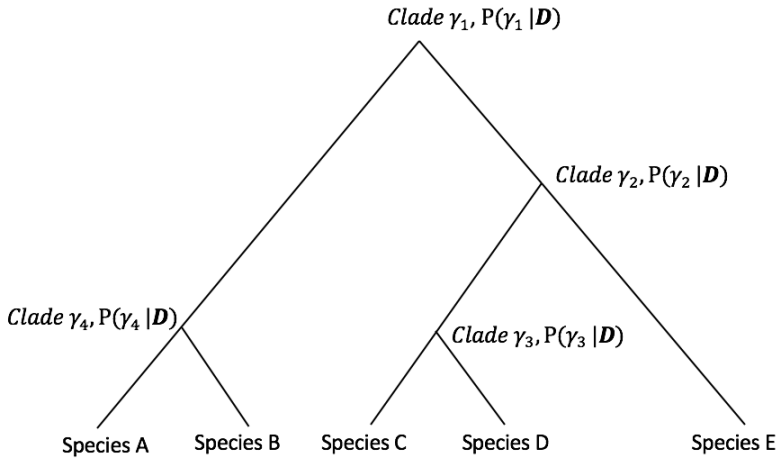

Figure 3.2: A consensus tree τ described by clades $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ and corresponding posterior probabilities $P(\gamma_1|\boldsymbol{D}), P(\gamma_2|\boldsymbol{D}), P(\gamma_3|\boldsymbol{D}), P(\gamma_4|\boldsymbol{D})$. Each clade is composed of several species: Clade γ_1 contains all the 5 species A, B, C, D and E. Clade γ_2 contains species C, D and E. Clade γ_3 contains species C and D. Clade γ_4 contains species A and B.

species outside the clade. Suppose Figure 3.2 gives an example of a consensus tree [3] and corresponding clade posterior probabilities from a sample of trees. The tree can be described by its clades $\gamma_1, \gamma_2, \gamma_3$, and γ_4, where γ_1 represents the set of species for the whole tree. We can write γ_1 as $\gamma_1 = (A, B, C, D, E)$. According to the rules of conditional probability, the event that τ being the true tree is equivalent to the intersection of events that the true tree contains clades γ_2, γ_3, and γ_4. In this case, the probability of τ being the true tree is

$$P(\tau) = P(\gamma_2 \cap \gamma_3 \cap \gamma_4) = P(\gamma_2 \cap \gamma_4)P(\gamma_3|\gamma_2 \cap \gamma_4). \tag{3.3}$$

Based on an assumption that the clades in different regions of the tree might be approximately independent, the conditional probabilities in Equation (3.3) can be further simplified. For instance, in Figure 3.2, given an edge that separates clade γ_4 from species C, D and E in the tree, the clade γ_2 that further develops the opposite side of this edge is approximately independent from γ_4. Under the principle of conditional independence of separated clades, the conditional probability $P(\gamma_3|\gamma_2 \cap \gamma_4)$ can be simplified as $P(\gamma_3|\gamma_2)$ because the edge separates clade γ_2 also separates its daughter clade γ_3 from clade γ_4 in the tree. Therefore, the probability in equation (3.3) leads to the expression as follows:

$$P(\tau) \approx P(\gamma_2 \cap \gamma_4)P(\gamma_3|\gamma_2). \tag{3.4}$$

[3] A consensus tree is a graphical way to display the collection of most probable clades from a sample of trees.

In a more general case, let γ_L and γ_R be the left subclade and right subclade splitting the mother clade γ. Each of the two probability terms in Equation (3.4) follows the form $P(\gamma_L \cap \gamma_R|\gamma)$. This is true even for subclades with only a single taxa, yielding $P(\gamma_2 \cap \gamma_4) = P(\gamma_2 \cap \gamma_4|\gamma_1)$ and $P(\gamma_3|\gamma_2) = P(\gamma_3 \cap \text{Species E}|\gamma_2)$. Consider an arbitrary binary rooted tree T where each clade γ contains at least 2 taxa including the clade Γ containing all taxa in the tree. The posterior probability of the tree T given the ubiquitous clade Γ representing all taxa can be approximated:

$$q_T(\Gamma) = \prod_{\gamma \in \text{ all clades of } T,\ |\gamma|>1} P(\gamma_L \cap \gamma_R|\gamma), \tag{3.5}$$

where

$$P(\gamma_L \cap \gamma_R|\gamma) = \frac{P(\gamma_L \cap \gamma_R \cap \gamma)}{P(\gamma)}.$$

$P(\gamma_L \cap \gamma_R|\gamma)$ can be estimated from the MCMC sample such that $P(\gamma_L \cap \gamma_R \cap \gamma)$ is the joint frequency of observing the split implying clades γ_L, γ_R associated with their parent clade γ, and $P(\gamma)$ is the frequency of observing the parent clade γ.

Amalgamated Likelihood Estimation ALE method combines the CCP-estimate sequence likelihood based on MrBayes sampled trees, with a reconciliation model that accounts for DTL events, to reconstruct gene trees and infer the HGTs.

Within a gene family, consider n samples of MrBayes trees denoted by $\boldsymbol{\tau} = \{\tau_1, \tau_2, \ldots, \tau_n\}$ with $\max(n) = 20,000$. As mentioned earlier, CCP method is used to estimate the posterior probability of a gene tree τ_i that can be pieced together from clades observed in the sample of trees $\boldsymbol{\tau} = \{\tau_1, \tau_2, \ldots, \tau_n\}$. This is a process called *gene tree amalgamation* [41]. The CCP-estimated posterior probability is nonzero only if the gene trees can be amalgamated, and zero otherwise.

The basic idea of ALE in our study is to exhaustively explore the space of reconciled gene trees that can be amalgamated from the clades present in MrBayes trees. To be specific, ALE was implemented to estimate the likelihood of DNA sequence alignment D conditional on the species tree S (given in Figure 2.6) and a reconciliation model M, where the model M allows for gene duplication (rate δ), gene loss (rate ϕ), gene transfer (rate ν), and speciation (rate σ), i.e. $M = \{\delta, \phi, \nu, \sigma\}$. According to Bayes' rule, the parameters of the model M can be estimated given the species tree S and alignment D, which is expressed as

$$f(M|S, D) \propto L_{joint}(D|S, M) f(S, M),$$

where the likelihood can be expressed as

$$L_{joint}(D|S,M) \approx \sum_{i=1}^{n} P(D|\tau_i)P(\tau_i|S,M) \approx \sum_{i=1}^{n} \left[q_{\tau_i}(\Gamma_{\tau_i}) \sum_{(e,t)} P_e(R_{\tau_i},t)\right]. \tag{3.6}$$

Based on the law of total probability, the likelihood can be first written as the sum of the product of $P(D|\tau_i)$ and $P(\tau_i|S,M)$ over all gene tree topologies. The value of $q_{\tau_i}(\Gamma_{\tau_i})$ is the CCP-estimate posterior probability of gene tree τ_i that can be amalgamated from clades present in MrBayes trees, where Γ_{τ_i} is a ubiquitous clade consisting of all taxa in τ_i. The value of $P_e(R_{\tau_i},t)$ is a reconciliation likelihood calculated using a series of speciation and DTL events that draw R_{τ_i}-rooted gene tree τ_i into the species tree S recursively in terms of the speciation time t and branch e along S. A recursive calculation of reconciliation likelihood is illustrated in Figure B.2. The sum $\sum_{(e,t)} P_e(R_{\tau_i},t)$ runs over all branch-time pairs in the species tree S [42]. Here ALE assumes the branch lengths of the species tree to be 1 by default. The amalgamated likelihood in Equation (3.6) is approximate because it is computed based on the approximate independence assumption for CCP estimates. The details of the joint likelihood in Equation (3.6) can be found in [41]. In general, given a set of MrBayes trees, ALE extracts the clades observed in these trees associated with their frequencies, and explore the space of reconciled gene trees that can be amalgamated from these clades while maximizing the likelihood of observing the reconciled gene tree. The result is a maximum likelihood amalgamated reconciled gene tree; when used with its Bayesian MCMC sampling mode, ALE can also determine the DTL rates and sample reconciled gene trees.

For each homologous gene family, the MrBayes trees and the species tree given in Figure 2.6 were provided as input to ALE. Then ALE was run independently on the two sets of MrBayes trees in a Bayesian MCMC sampling mode. For each ALE run, there were $100,000$ iterations of the MCMC chain in total, and a reconciled tree was sampled every 100 iterations, resulting in $1,000$ sampled reconciled gene trees. We call these two sets of sampled reconciled gene trees the *ALE trees*. In addition, a maximum likelihood amalgamated reconciled gene tree was computed for each ALE run. Gene families for which the two amalgamated reconciled gene trees were not identical were excluded from further analysis. In addition, ALE can produce inferred HGT events and their frequencies. An HGT event is defined by a donor species d and a receptor species r, which is denoted by the ordered pair (d,r). The frequency of an inferred HGT event is obtained by averaging the frequency of observing this HGT in the ALE trees over two independent ALE runs; note that d and r can both be either an extant or an ancestral species. The final output of this stage is a list of quadruples (donor d, receptor r, family g, frequency f). Each quadruple records that for a given family g, an HGT from species d to species r was observed in the sampled reconciled

gene trees with frequency f. We used the remaining inferred HGT events to detect signals of introgression in the rest of work.

3.3 MaxTic: Time Consistency of Inferred HGTs

It is known that accurate detection of HGT is challenging, especially when using an undated species tree. Thus, evaluating noises due to likely false HGTs is important. To do so, we rely on a method proposed by Chauve in 2017, called *Maximum Time Consistency*, or MaxTiC [5]. The implementation of MaxTiC written in Python is available at `https://github.com/ssolo/ALE/tree/master/maxtic`.

For each gene family, each HGT event inferred by ALE has a donor species d and a receptor species r. Let a denote the most recent common ancestor of the donor species d. In our case, a *time constraint* is then defined as $a > r$, which means a must be older than r. We assigned a *weight* to this time constraint $a > r$, which is the frequency that the constraint $a > r$ is found in $1,000$ reconciled gene trees, summed across all gene families. Based on the list of quadruples obtained from ALE results, consider the frequency of a HGT event (d, r) in a gene family g is $f_{(a>r|(d,r),g)}$. Let $\mathcal{G}$ represent the set of gene families. The weight of the time constraint associated with (d, r), i.e. $a > r$, can be written as

$$w_{(a>r|(d,r))} = \sum_{\text{all } g \in \mathcal{G}} f_{(a>r|(d,r),g)}. \tag{3.7}$$

A large number of time constraints can be used to order the internal nodes and produce a *ranked phylogeny*. However, there might exist conflicts in time constraints such that different constraints result in different ranked version of the phylogeny. An example of conflicting constraints is provided in Figure 3.3. On the basis of some HGT events, Y is found to be older than X, and Z is found to be older than T, but T is an ancestor of Y and X is an ancestor of Z. We cannot construct a phylogeny to satisfy both constraints. In this case, a conflict occurs.

Let C be the set of time constraints with weights inferred from l ALE gene transfers,

$$C = \{w_{(a_1>r_1|(d_1,r_1))}, w_{(a_2>r_2|(d_2,r_2))}, ..., w_{(a_l>r_l|(d_l,r_l))}\},$$

where a_i is the most recent ancestor of the donor species d_i, and r_i is the receptor species, $i = 1, 2, ..., l$. Given C derived from a set of HGTs, MaxTic searches for a subset $C_0 \subset C$ maximizing the total weights, where all constraints in C_0 are compatible with the total order of the species tree (so-called *time consistency*), and we denote this maximum weight

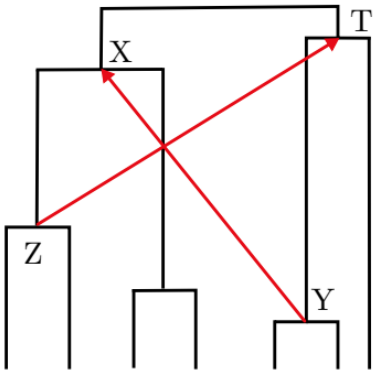

Figure 3.3: Two conflicting constraints: $Y > X$ and $Z > T$. Each of the constraints can be fulfilled by different ranked versions of the phylogeny, but they cannot occur simultaneously. This figure is retrieved from [5].

as $W_{\text{consistent}}$. 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. We applied MaxTiC with inputs composed of time constraints or ranking constraints derived from HGTs, obtained by filtering out inferred HGTs with frequencies below a threshold t. In our study, t ranges 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, where the weights are denoted by $W_{\text{consistent at } t}$; and the constraints conflict with the computed ranking of the internal nodes of the species tree, where the weights are denoted by $W_{\text{conflict at } t}$. We define the *consistency ratio* as the ratio between the weight of the time consistent constraints and the weight of all considered constraints at frequency threshold t,

$$\text{consistency ratio at } t = \frac{W_{\text{consistent at } t}}{W_{\text{consistent at } t} + W_{\text{conflict at } t}}. \tag{3.8}$$

Intuitively, a high consistency ratio points to a low proportion of erroneous HGTs inferred by ALE. We would like to observe the consistency ratio as high as possible to gain robust HGTs. Afterwards, the post-filtered HGTs are taken into next stage for statistical tests.

3.4 Multiple Hypothesis Testing: Detecting Potential Introgressed Segments

Gene duplication and HGT are two gene-level evolutionary mechanisms that can cause incongruence between a gene tree and a species tree, which have been taken into account

in ALE model. However, incomplete lineage sorting (ILS, as mentioned in Chapter 1) is another common mechanism that can cause phylogenetic incongruence, which is not considered in the ALE model. A crucial problem of detecting introgression in our study is to distinguish the effects of introgression from ILS because both of them can result in HGT-like patterns in reconciled gene trees, as shown in Figure 1 from Yu's work [52]. To resolve this problem, we rely on the assumption that unlike ILS, the introgression is more likely to impact blocks of contiguous genes [39]; while ILS is expected to impact genes randomly located along chromosomes. With this assumption, for a given pair of species (d, r), we aim at detecting genome regions where the concentration of genes and corresponding gene families whose evolutionary history involves post-filtered HGT from d to r is significantly higher. For example, the $(1, 1)$ subplot in Figure 4.5 shows a significant signal of introgression instead of ILS; whereas $(2, 1)$ shows oppositely. As *Anopheles gambia* is the only fully assembled genome in our data set, we used five chromosomal arms of *Anopheles gambia*, including chromosome 2L, 2R, 3L, 3R and X, to perform statistical tests within genome segments; we discuss the impact of this approximation in Chapter 5.

We designed our analysis as follows. Consider an *Anopheles gambiae* genome segment (called a *window* from now) containing k genes. Within a given window, let p denote the true probability of observing 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. Here, the HGTs we consider are the ones inferred from the ALE results. A statistical hypothesis test is conducted to test the null hypothesis (H_0), $p = p_0$, versus the alternative hypothesis (H_a), $p > p_0$. It is a one-sided test because we are interested in significant evidence of gene transfer events that suggest potential introgression. Let X_i be the number of observed HGTs from d to r in the m ALE sampled reconciled gene trees for the i-th gene in the window, where $i = 1, \ldots, k$. We assume the distribution of X_i to be Binomial(m, p). An unbiased estimator for p is

$$\hat{p} = \frac{\sum_{i=1}^{k} X_i}{mk}.$$

Under the assumption that X_i's are independent for simplicity, we have

$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{mk}.$$

Consequently, the test statistics

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/(mk)}}$$

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 $P(Z \geq z)$ conditional on the null hypothesis being true.

It is known that falsely rejecting the true hypothesis is a risk for a hypothesis test. This is known as the type I error and occurs with rate α. To reduce the risk of making a type I error, researchers always set a value for α in a hypothesis test. However, the problem becomes more serious if we perform many hypothesis tests because we are more likely to obtain "false discoveries". Therefore, for the multiple tests across all the windows on each chromosome, we used *False Discovery Rate* (FDR)-controlling procedures to control incorrect rejections, under the Benjamini-Yekutieli (BY) [3] procedure that controls the FDR for dependent tests. Suppose that there are h null hypotheses $H_i, i = 1, 2, ..., h$, with a p-value $p_{(i)}$ for each H_i. The FDR control under BY procedure generates as follows:

Algorithm 2 False Discovery Rate Control under Benjamini-Yekutieli procedure

1: We want to control FDR at level α.
2: Calculate p-values for the hypothesis tests and order the p-values as $p_{(1)} \leq ..... \leq p_{(i)} \leq .... \leq p_{(h)}$.
3: Multiply each $p_{(i)}$ by its adjustment factor $a_{(i)} = lh/i$, with $l = \sum_{k=1}^{h} 1/k$ for $i = 1, \ldots, h$.
4: If the multiplication in the last step does not follow the original ordering, apply a step-up method to decrease the highest p-values: $\tilde{p}_{(i)} = \min_{j=i,...,h} a_j p_{(j)}$.
5: Set $\tilde{p}_{(i)} = \min(\tilde{p}_{(i)}, 1)$ for all i.
6: For each $1 \leq i \leq h$, check if $\tilde{p}_{(i)} \leq \alpha$ is true; if true, then significant.

Step 2-5 in the BY controlling procedure is implemented in R using "`p.adjust`" function. The result of this analysis is a list of windows for which we detect a significantly higher concentration of genes supporting an HGT from d to r under an FDR at level 1%; we selected a window of size $n = 20$ genes, although results were similar with $n = 10$ or $n = 30$.

3.5 Flow Chart for the Research Methodology

The research methodology in our study of introgression is reviewed in Figure 3.4. First, the genes from 14 *Anopheles* genomes was clustered into more than 17,000 gene families and each family contains a multiple sequence alignment. Within each family, MrBayes was first applied on MSA to sample gene trees, which we call MrBayes trees. These MrBayes trees and the known species tree are provided as input to ALE to obtain bayesian sampling of 1,000 reconciled gene trees (ALE trees) built from clades of MrBayes trees, as well as a list of the inferred HGT events associated with frequencies. Later, the inferred HGTs are brought into MaxTiC to assess time consistency in order to remove probably false HGTs in further analysis. Finally, we use the filtered HGTs to detect genome regions with significantly more

genes whose evolution shows a signal of HGT than expected. The introgressed regions are demonstrated by large chromosomal regions that concentrate many genes indicating HGTs. Some chromoplots are presented in Section 4.4.

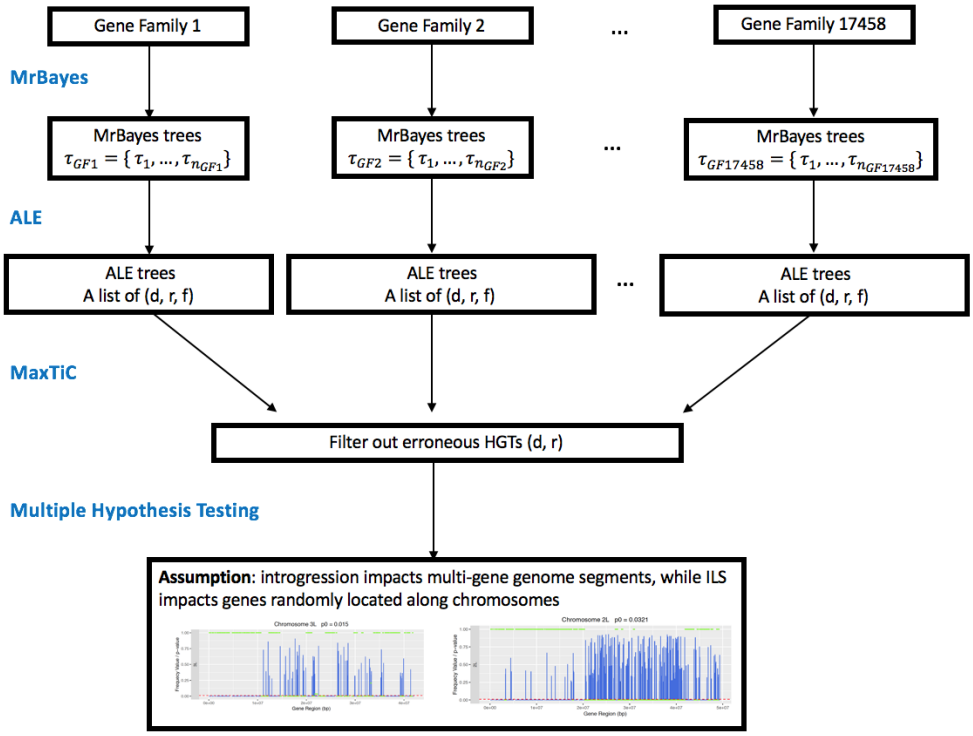

Figure 3.4: An overview of research methodology involved in our study.

Chapter 4

Experiment Results

4.1 Exploring the Space of Reconcilied Gene Trees

For each gene family, we first ran MrBayes using the family's MSA as input to produce 2 sets of up to 20,000 MrBayes trees. Those families with less than 4 genes, or at least one MCMC chain does not converge, or less than 5000 MrBayes trees, were discarded from further analysis. The rest of MrBayes trees and the fixed species tree were then provided to ALE to explore the space of reconciled gene trees by considering gene duplication, gene loss and HGT. Those gene families with two different amalgamated reconciled gene trees generated by ALE were excluded. After filtering out the trivial gene families using the MrBayes+ALE pipeline, the number of genes decreased from $169,447$ to $137,180$ with each species "losing" roughly $2,000$ genes, and the number of gene families dropped from $17,458$ to $11,589$. Figure 4.1 and Figure 4.2 illustrates the impact of this filtering on genome sizes and homologous families sizes respectively.

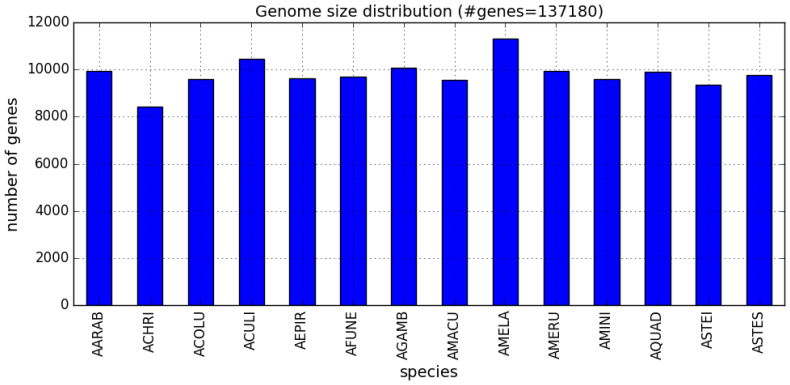

Figure 4.1: Number of genes for each *Anopheles* species after filtering out some gene families from the MrBayes+ALE pipeline.

Comparing with Figure 2.5, we can observe a significant decrease in the number of gene families with 12 or more genes, indicating that many of these families do not pass

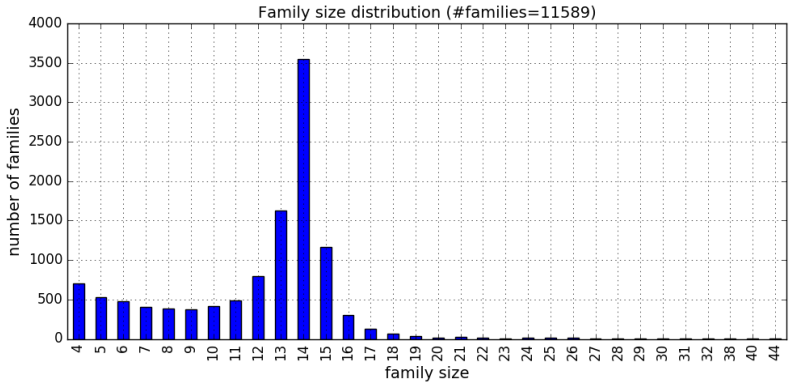

Figure 4.2: Number of genes per gene family after filtering out some gene families from the MrBayes+ALE pipeline.

our relatively stringent filtering criteria. But the post-filtering figures are very similar to the pre-filtering figures in general, indicating a limited impact of MrBayes+ALE filtering criteria.

Next, we consider the inferred HGTs that can suggest potential introgression events. For each gene family, after filtering out all HGTs that appear in less than 20% of both sets of ALE trees, the total number of conserved HGTs is $16,210$, leading to more than one inferred HGT events per gene family on average. Figure 4.3 shows that low-frequency HGTs dominate the landscape, although there are $4,771$ HGTs observed with frequency at least 50%, and $1,778$ HGTs observed with frequency at least 80%.

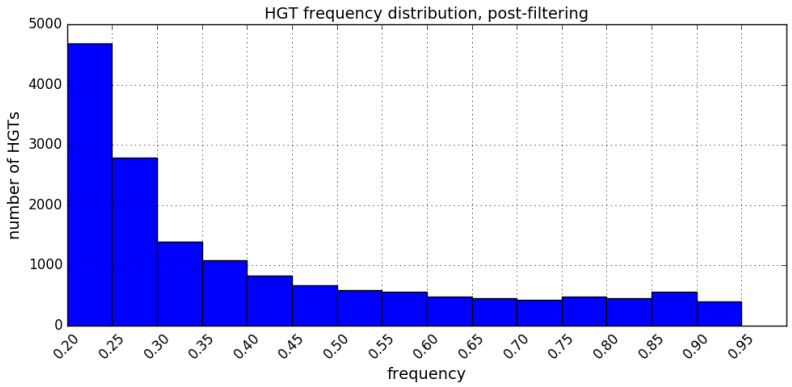

Figure 4.3: Distribution of the frequency of observed HGTs appearing with frequency at least 20% in ALE trees.

4.2 Discovering Noises in HGTs

After applying MaxTic over HGTs, the results suggest that the inferred HGTs do not show an apparent high level of noise, measured in terms of conflicting HGTs. Again, the consistency ratio represents the weight of the consistent constraints divided by the weight of all considered constraints. A higher consistency ratio means a lower level of noises. At threshold $t = 25\%$, we filtered out inferred HGTs with frequency $< 25\%$, the corresponding consistency ratio is $\frac{4172.58}{4172.68+377.05} \approx 0.9171$, as shown in Table 4.1. The ratio gradually increases from 0.9080 at $t = 20\%$, to 0.9382 at $t = 50\%$, and 0.9736 at $t = 80\%$, indicating a low level of time inconsistency for HGTs with frequency higher than 20%. The most interesting finding is that, at threshold $t = 85\%$, only two constraints with a significant weight are discarded, constraints $(18, 15)$ and $(14, 15)$. Here (x, y) means that node x should be ranked before node y, or x is older than y, while the reversed constraints $(15, 18)$ and $(15, 14)$ are conserved.

Threshold t(%)	**Kept Constraints (#)**	**Weight**	**Discarded Constraints (#)**	**Weight**	**Consistenty Ratio**
20	39	4464.98	35	447.84	0.9088
25	32	4172.68	31	377.05	0.9171
30	29	3877.52	23	315.33	0.9248
35	25	3626.85	19	285.85	0.9269
40	21	3359.89	17	245.53	0.9319
45	20	3156.61	16	220.20	0.9348
50	18	2983.73	14	196.64	0.9382
55	15	2823.46	13	164.98	0.9448
60	13	2658.03	10	138.58	0.9504
65	10	2473.69	10	117.98	0.9545
70	7	2288.02	6	90.64	0.9619
75	7	2083.75	3	67.66	0.9686
80	6	1806.85	3	48.96	0.9736
85	4	1561.26	2	22.77	0.9856
90	3	1186.73	1	1.80	0.9985
95	1	671.19	0	0.00	1.00

Table 4.1: HGTs time consistency ratio for different level of threshold t.

4.3 Criteria of Potential Introgression Events

Following the results of MaxTiC analysis, we used stringent criteria below to classify inferred HGTs as potential introgression events from a donor species d to a receptor species r:

- The inferred HGT must be observed in at least 50 gene families.
- Each inferred HGT must be observed at a frequency higher than 50%.
- The inferred HGT must be observed with an accumulated frequency greater than 50 across all gene families.

Figure 4.4 shows the potential introgression events detected using these criteria. As expected, most potential introgression events are recent and involve the gambiae complex,

which agrees well with the extensive amount of introgression seen in this group by Fontaine *et al* [11]; in particular, we retrieve the major introgression from *Anopheles arabiensis* to the common ancestor of *Anopheles gambiae* and *Anopheles coluzzi* (ancestral species 15) that was found in [11, 48]. We can also observe some bidirectional introgresion events at various levels of support: (*Anopheles arabiensis*, *Anopheles coluzzi*) and (*Anopheles coluzzi*, *Anopheles arabiensis*); (*Anopheles gambia*, *Anopheles arabiensis*) and (*Anopheles arabiensis*, *Anopheles gambia*); (15, *Anopheles arabiensis*) and (*Anopheles arabiensis*, 15); (*Anopheles quadriannulatus*, *Anopheles gambia*) and (*Anopheles gambia*, *Anopheles quadriannulatus*). The only other potential event within the gambiae complex found in our analysis is the HGT event (*Aopheles quadriannulatus*, *Anopheles merus*), agreeing with the direction proposed in [48] as opposed to [11], although with a limited support. We can also observe that the frequency of transfers is close to the number of transfers, suggesting most such transfers actually occur more often than 50%.

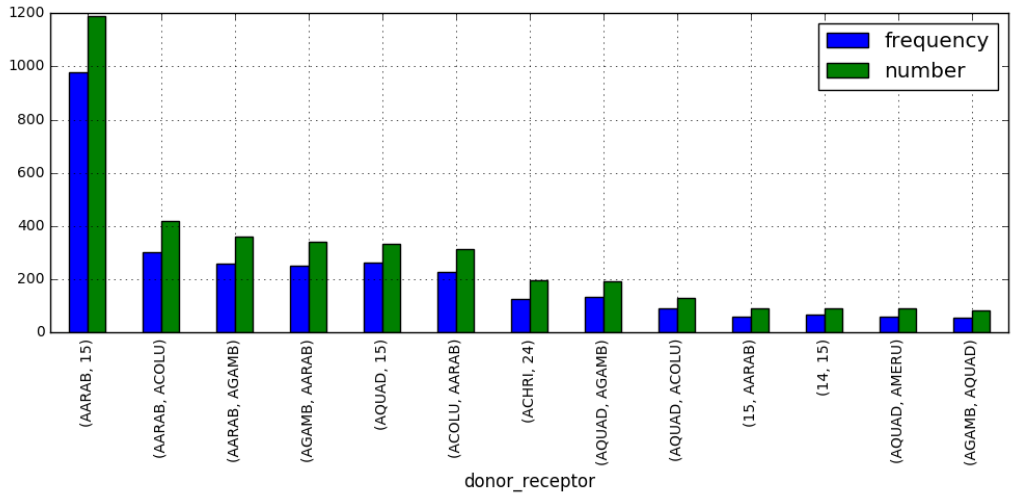

Figure 4.4: Potential introgression events based on sets of at least 50 inferred HGTs of frequency 0.5 or above and accumulated frequency at least 50. The x-axis shows 13 top potential introgression events, and y-axis indicates the total frequency or total count of each HGT. The green bars indicate the total number of corresponding HGTs, while the blue bars indicate the sum of frequencies for a particular HGT event across all gene families. For instance, the potential introgression event (ACHRI, 24) occurred 195 times across all families as shown by the green bar, and the sum of frequency is 126.75 as shown by the blue bar, leading to an average frequency of 0.65.

A new observation is the hypothesis of a potential introgression event from the lineage of *Anopheles christyi* to ancestral species 24 within Asian clade. To the best of our knowledge, such an ancient and 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, which is comparable to other possible introgression events, such as the one from *Anopheles quadriannulatus* to *Anopheles gambia* supported by 193 HGTs with an average

frequency 0.70.

In order to assess further the level of support for these various potential introgression events, we considered the taxon coverage (i.e. number of species covered) 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 (*Anophles arabiensis*, *Anopheles gambiae*) to 13.89 for (*Anopheles christyi*, species 24). The same analysis repeated after lowering the HGT frequency threshold to 0.2 leads to similar results, with a slight decrease of the average taxon coverage by gene families.

4.4 Evolutionary Histories of Gene Families

Following the method of detecting introgression described in section 3.4, we explored the concentration of genes belonging to gene families whose evolutionary history involves HGTs (d, r) to detect signals of introgression from d to r along the chromosomes of *Anopheles gambia*. Some chromoplots are given in Figure 4.4. All chromoplots are available at `https://github.com/cchauve/Anopheles_introgression_RECOMBCG_2018`. We discuss some interesting observations below.

First, the pattern of potentially introgressed genes for three HGT events (*Anopheles arabiensis*, 15), (*Anopheles arabiensis, Anopheles gambia*), and (*Anopheles arabiensis, Anopheles coluzzi*) along the chromosome arms 2L and 3L are displayed in Figure 4.5. In each plot, the horizontal axis represents the position along the chromosome ("bp" means base pair, one of the pairs A-T or C-G), and the vertical axis ranging from 0 to 1 represents the corresponding HGT frequency from ALE results and corrected p-values under BY controlling procedure. The average of the (d, r) HGT frequencies for all genes along the chromosome is given at the top of each plot. Those windows with p-values (green dots) $< 1\%$ (red dotted line) indicate significant evidence that we are more likely to observe the gene from a family whose evolution involves a (d, r) HGT, and a concentration of such genes indicate signatures of introgression. For the first HGT event (*Anopheles arabiensis*, 15), we observe a concentration of genes indicating introgression on chromosome 2L and 3L. For the second HGT event (*Anopheles arabiensis, Anopheles gambia*), it is interesting to observe that there is a fairly weak signal of introgression from *Anopheles arabiensis* to *Anopheles gambia* on chromosome arm 2L, although the signal tends to be a bit stronger on chromosome arm 3L. For the third HGT event (*Anopheles arabiensis*, *Anopheles coluzzi*), a signal of intrgression is strongly detected along chromosome arm 2L, and the region it centered around is called "2La inversion", a widespread polymorphism related to the level of malaria infection in the

gambiae complex [33]. There is no obvious introgression signals along chromosome 3L for the third HGT event. Similarly, limited regions of the X chromosome being introgressed can be observed, illustrated in Figure 4.6.

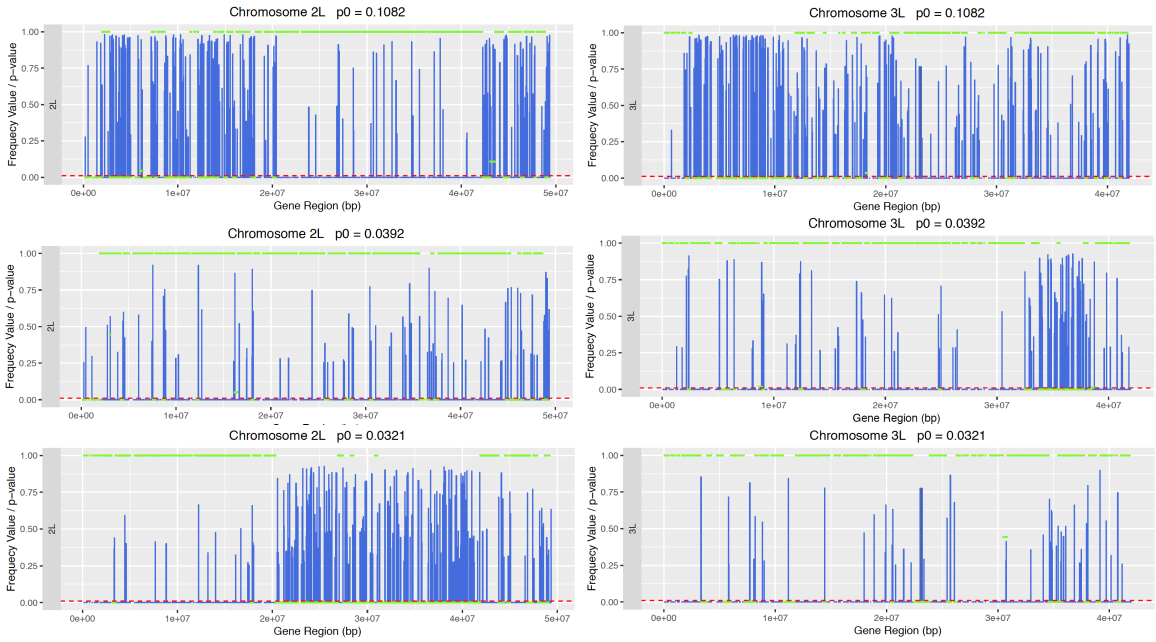

Figure 4.5: Chromoplots for HGT events (*Anopheles arabiensis*, 15) [Top], (*Anopheles arabiensis, Anopheles gambia*) [Middle], and (*Anopheles arabiensis, Anopheles coluzzi*) [Bottom] along chromosome arms 2L and 3L. Blue vertical bars indicate genes with their HGT frequency, the red dotted line is the FDR at level 1% and green dots represent the BY corrected p-values.

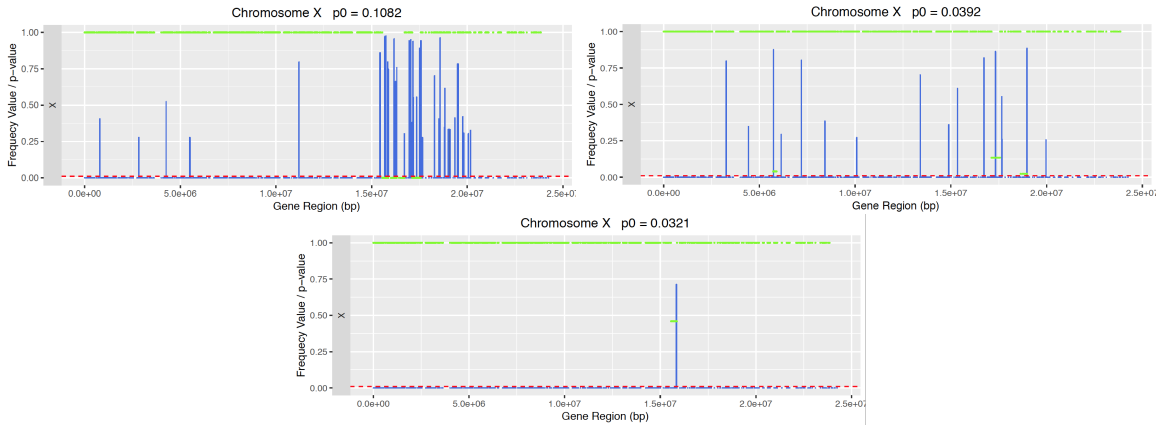

Figure 4.6: Chromoplots for the HGT events (*Anopheles arabiensis*, 15) [Top Left], (*Anopheles arabiensis, Anopheles gambia*) [Top Right], and (*Anopheles arabiensis, Anopheles coluzzi*) [Bottom] along chromosome X.

By looking at the chromoplots obtained from the HGTs observed between the lineage of *Anopheles christyi* and species 24 in Figure 4.7, we can see a level of support of introgression similar to the potential events located within the gambiae complex, although with a much stronger signal of introgression located on the X chromosome.

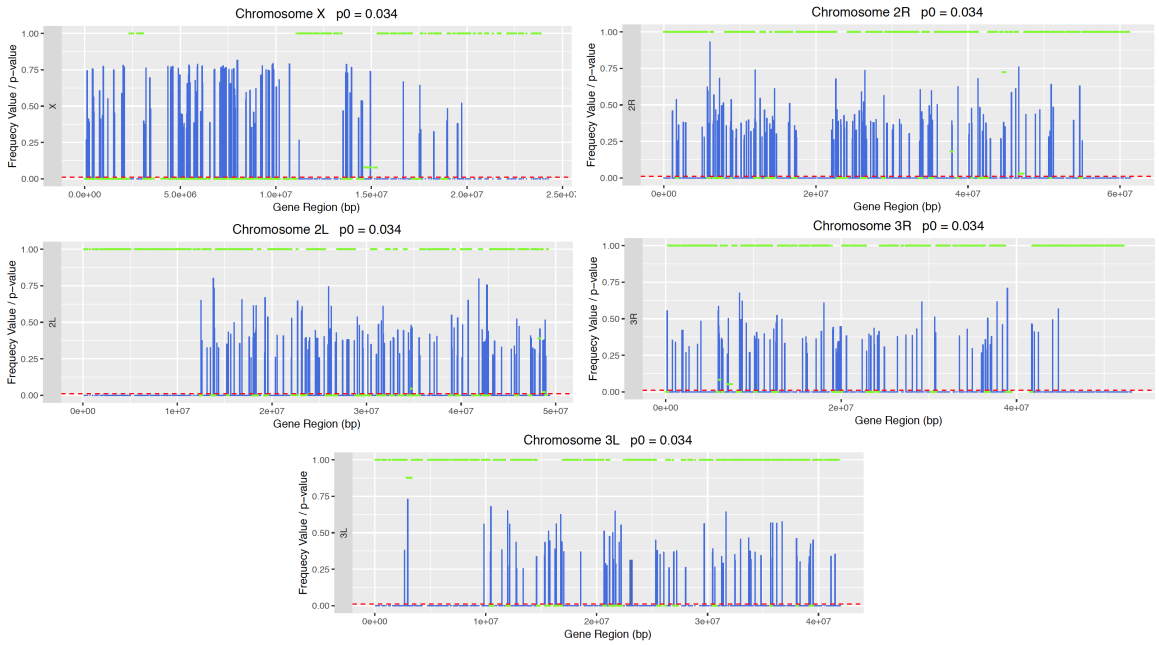

Figure 4.7: Chromoplots for the potential introgression event from *An. christyi* to species 24 on five chromosomes. A relatively strong signal of introgression is observed on chromosome X.

Within the gambiae complex, we retrieve patterns observed in other works. For example, Wen *et al.* mentioned that the introgression from *Anopheles quadriannulatus* to *Anopheles gambia* involves mostly 2La inversion [48], for which we obtained similar results as shown in Figure 4.8. We also detect the signal for an introgression event from *Anopheles quadriannulatus* to *Anopheles merus* on limited regions of chromosomal arms 3R and 3L, as given in Figure 4.9, which further confirms the findings discussed in [48].

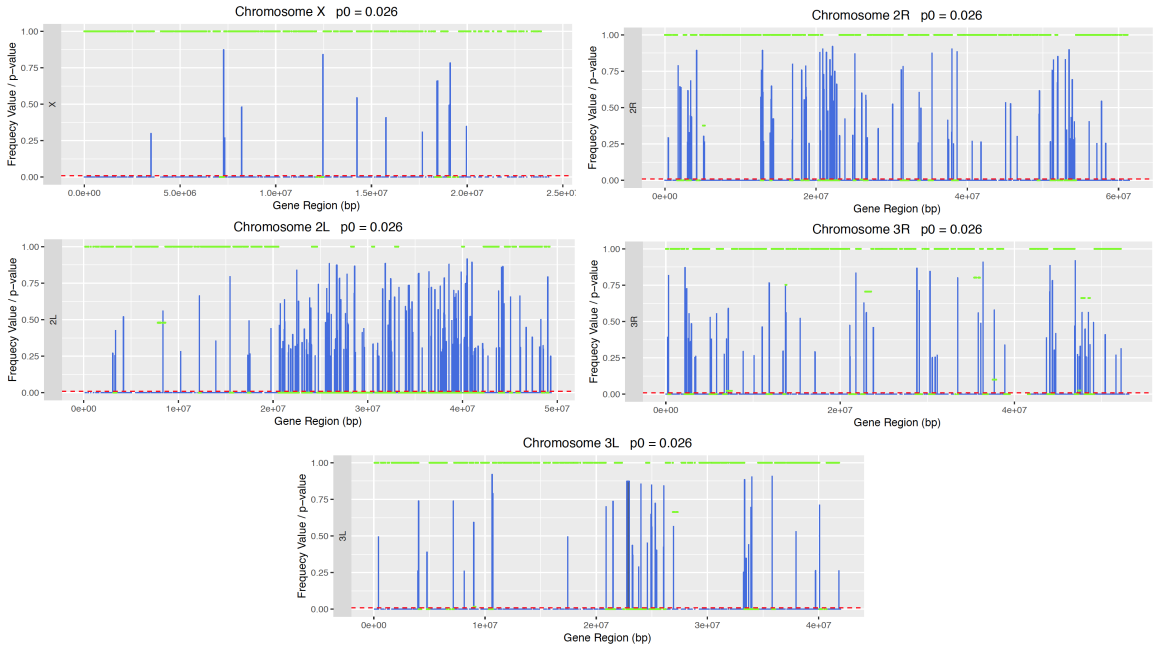

Figure 4.8: Chromoplots for introgression from *Anopheles quadriannulatus* to *Anopheles gambia* on five chromosomes. A relatively strong signal of introgression is observed on chromosome 2L, so-called 2La inversion.

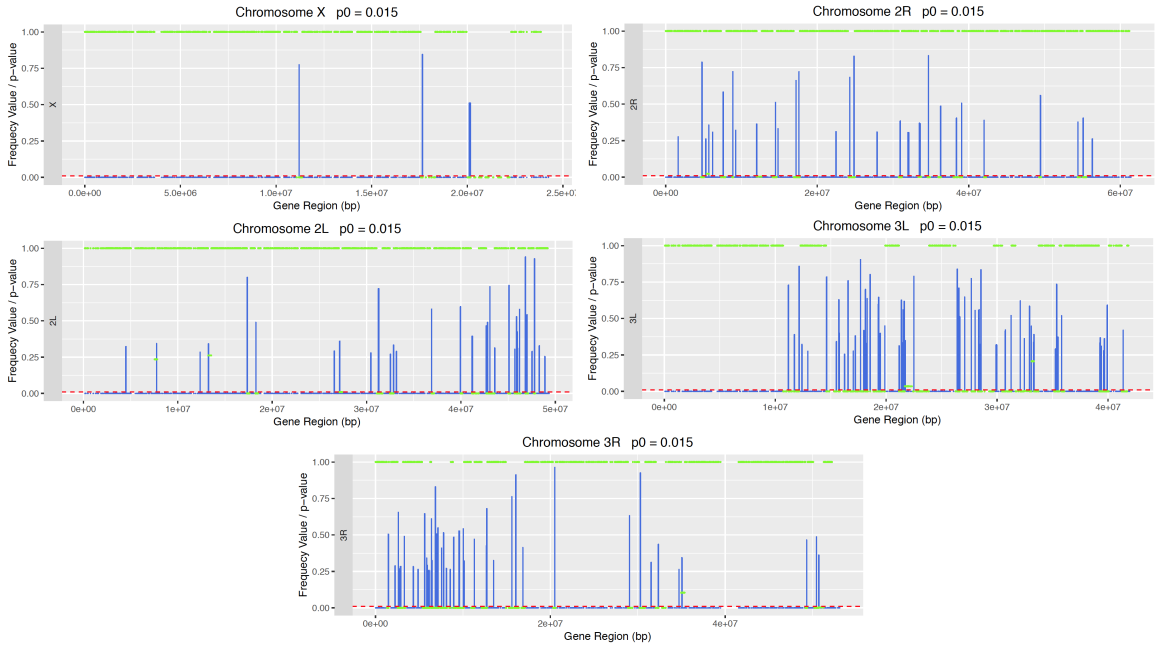

Figure 4.9: Chromoplots for introgression from *Anopheles quadriannulatus* to *Anopheles melus* on five chromosomes. Introgression is detected on limited regions of chromosome 3L and 3R.

Chapter 5

Discussion

We present an efficient statistical method that detects signals of introgression events by inferring HGTs through a gene tree-species tree framework, which accounts for gene duplication and gene loss simultaneously. This approach benefits from the recent development of phylogenetic analysis tools, including MrBayes, ALE, and MaxTiC. These tools are applied on a well-studied large dataset of 14 *Anopheles* genomes covering both African and Asian mosquitoes. Relying on the hypothesis that introgression concerns blocks of contiguous genes, we finally recover extensive introgression within the gambiae complex, as discussed in [11]; and propose a potential ancient introgression event between more distant species: from the lineage of *Anopheles christyi* to the common ancestor of Asian *Anopheles* mosquitoes.

The approach we proposed for detecting signals of introgression has several advantages. First, compared to existing methods, the reconciliation framework in our approach can handle gene families with gene duplication and gene loss events. Secondly, sampling reconciled gene trees provides a more nuanced view of gene family evolution, for which the sampling frequency of HGTs can be used to discard probably false HGTs, and therefore improves overall accuracy of detecting intrgression. More importantly, our approach is applied to a relatively large data sets containing 14 genomes, which makes a strong contrast with methods based on summary statistics.

However, this approach still needs to be studied in terms of accuracy and performance on simulated data sets. The impact of errors in gene families, gene trees and the considered species tree should also be assessed in these simulations. It would also allow for evaluating different reconciliation algorithms, including recently developed algorithms that account for ILS [4, 43]. SimPhy, a fast and flexible software package for phylogenetic simulation under ILS and gene DTL events, would be a possible way to conduct simulation in our case [27]. Also, we can make a comparison of introgression events between ALE and other different reconciliation methods, such as NOTUNG [6] and Ranger-DTL [2], in order to see if we

observe similar introgression results.

While disentangling introgression from ILS, we pay attention to HGTs in a block of contiguous genes to infer introgression events, comparing with genes randomly scattered on the chromosome in cases of ILS. Such spatial co-localization tests become less powerful when we only use the fully-assembled genome of *Anopheles gambia* to show all potential introgression events along five chromosomes. We could also test the impact of using fragmented assemblies of extant or ancestral species instead of *Anopheles gambia* on the spatial co-localization test [1]. But we expect to see less introgressed regions along fragmented assemblies of genome.

In conclusion, our work demonstrates a reconciliation-based approach to study introgression in a larger data set. The results on the *Anopheles* data set confirm previous reported results and infer an unreported introgression event between Asian and African *Anopheles*, which displays the power of our technique. Some more studies need to be conducted to check accuracy and performance, such as testing performance on simulated datasets, and comparing with other reconciliation methods. From an applied point of view, the hypothesis of the introgression event between *Anopheles christyi* and the Asian mosquitoes clade is an interesting direction to study further.

Bibliography

[1] Yoann Anselmetti, Wandrille Duchemin, Eric Tannier, Cedric Chauve, and Sèverine Bérard. Phylogenetic signal from rearrangements in 18 anopheles species by joint scaffolding extant and ancestral genomes. *BMC Genomics*, 19(2):96, 2018.

[2] Mukul S Bansal, Manolis Kellis, Misagh Kordi, and Soumya Kundu. Ranger-dtl 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics*, 34(18):3214–3216, 2018.

[3] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, pages 1165–1188, 2001.

[4] Yao-ban Chan, Vincent Ranwez, and Céline Scornavacca. Inferring incomplete lineage sorting, duplications, transfers and losses with reconciliations. *Journal of Theoretical Biology*, 432:1–13, 2017.

[5] Cédric Chauve, Akbar Rafiey, Adrian Davin, Celine Scornavacca, Philippe Veber, Bastien Boussau, Gergely Szollosi, Vincent Daubin, and Eric Tannier. Maxtic: Fast ranking of a phylogenetic tree by maximum time consistency with lateral gene transfers. 2017.

[6] Kevin Chen, Dannie Durand, and Martin Farach-Colton. Notung: a program for dating gene duplications and optimizing gene family trees. *Journal of Computational Biology*, 7(3-4):429–447, 2000.

[7] Kanchon K Dasmahapatra, James R Walters, Adriana D Briscoe, John W Davey, Annabel Whibley, Nicola J Nadeau, Aleksey V Zimin, Daniel ST Hughes, Laura C Ferguson, Simon H Martin, et al. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 487(7405):94, 2012.

[8] James H Degnan. Modeling hybridization under the network multispecies coalescent. *Systematic Biology*, 67(5):786–799, 2018.

[9] Alexei J Drummond and Andrew Rambaut. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1):214, 2007.

[10] Ryan Elworth, Chabrielle Allen, Travis Benedict, Peter Dulworth, and Luay Nakhleh. Dgen: A test statistic for detection of general introgression scenarios. *BioRxiv*, 2018. doi: 10.1101/348649. URL `https://www.biorxiv.org/content/10.1101/348649v1.abstract`.

[11] Michael C Fontaine, James B Pease, Aaron Steele, Robert M Waterhouse, Daniel E Neafsey, Igor V Sharakhov, Xiaofang Jiang, Andrew B Hall, Flaminia Catteruccia, Evdoxia Kakani, et al. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217):1258524, 2015.

[12] Claire Garros, Lizette L Koekemoer, Maureen Coetzee, Marc Coosemans, and Sylvie Manguin. A single multiplex assay to identify major malaria vectors within the african anopheles funestus and the oriental an. minimus groups. *The American Journal of Tropical Medicine and Hygiene*, 70(6):583–590, 2004.

[13] Richard G Harrison and Erica L Larson. Hybridization, introgression, and the nature of species boundaries. *Journal of Heredity*, 105(S1):795–809, 2014.

[14] Sebastian Höhna and Alexei J Drummond. Guided tree topology proposals for bayesian phylogenetic inference. *Systematic Biology*, 61(1):1–11, 2011.

[15] Sebastian Höhna, Michael J Landis, Tracy A Heath, Bastien Boussau, Nicolas Lartillot, Brian R Moore, John P Huelsenbeck, and Fredrik Ronquist. Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, 65(4):726–736, 2016.

[16] Barbara R Holland, Steffi Benthin, Peter J Lockhart, Vincent Moulton, and Katharina T Huber. Using supernetworks to distinguish hybridization from lineage-sorting. *BMC Evolutionary Biology*, 8(1):202, 2008.

[17] John P Huelsenbeck, F Ronquist, and B Hall. An introduction to bayesian inference of phylogeny, 2001.

[18] Dirk Husmeier. Introduction to statistical phylogenetics. In *Probabilistic Modeling in Bioinformatics and Medical Informatics*, pages 83–145. Springer, 2005.

[19] Edwin Jacox, Cedric Chauve, Gergely J Szöllősi, Yann Ponty, and Celine Scornavacca. eccetera: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, 32(13):2056–2058, 2016.

[20] Clemens Lakner, Paul Van Der Mark, John P Huelsenbeck, Bret Larget, and Fredrik Ronquist. Efficiency of markov chain monte carlo tree proposals in bayesian phylogenetics. *Systematic Biology*, 57(1):86–103, 2008.

[21] Bret Larget. The estimation of tree posterior probabilities using conditional clade probability distributions. *Systematic Biology*, 62(4):501–511, 2013.

[22] Bret Larget and Donald L Simon. Markov chain monte carlo algorithms for the bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, 16(6):750–759, 1999.

[23] Jiayu Lin. *On the Dirichlet Distribution*. PhD thesis, Queen's University, 2016.

[24] Kevin J Liu, Jingxuan Dai, Kathy Truong, Ying Song, Michael H Kohn, and Luay Nakhleh. An hmm-based comparative genomic framework for detecting introgression in eukaryotes. *PLoS Computational Biology*, 10(6):e1003649, 2014.

[25] Wayne P Maddison. Gene trees in species trees. *Systematic Biology*, 46(3):523–536, 1997.

[26] James Mallet, Nora Besansky, and Matthew W Hahn. How reticulated are species? *BioEssays*, 38(2):140–149, 2016.

[27] Diego Mallo, Leonardo de Oliveira Martins, and David Posada. Simphy: phylogenomic simulation of gene, locus, and species trees. *Systematic Biology*, 65(2):334–344, 2015.

[28] Simon H Martin and Chris D Jiggins. Interpreting the genomic landscape of introgression. *Current Opinion in Genetics & Development*, 47:69 – 74, 2017. doi: https://doi.org/10.1016/j.gde.2017.08.007. URL `http://www.sciencedirect.com/science/article/pii/S0959437X17300357`.

[29] Luay Nakhleh. Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in Ecology & Evolution*, 28(12):719–728, 2013.

[30] Luay Nakhleh, Derek Ruths, and Hideki Innan. Gene trees, species trees, and species networks. 2009.

[31] Daniel E Neafsey, Robert M Waterhouse, Mohammad R Abai, Sergey S Aganezov, Max A Alekseyev, James E Allen, James Amon, Bruno Arcà, Peter Arensburger, Gleb Artemov, et al. Highly evolvable malaria vectors: the genomes of 16 anopheles mosquitoes. *Science*, 347(6217):1258522, 2015.

[32] James B Pease and Matthew W Hahn. Detection and polarization of introgression in a five-taxon phylogeny. *Systematic Biology*, 64(4):651–662, 2015.

[33] Michelle M Riehle, Tullu Bukhari, Awa Gneme, Wamdaogo M Guelbeogo, Boubacar Coulibaly, Abdrahamane Fofana, Adrien Pain, Emmanuel Bischoff, Francois Renaud, Abdoul H Beavogui, et al. The anopheles gambiae 2la chromosome inversion is associated with susceptibility to plasmodium falciparum in africa. *Elife*, 6:e25813, 2017.

[34] Fredrik Ronquist, Maxim Teslenko, Paul Van Der Mark, Daniel L Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A Suchard, and John P Huelsenbeck. Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542, 2012.

[35] Benjamin K Rosenzweig, James B Pease, Nora J Besansky, and Matthew W Hahn. Powerful methods for detecting introgressed regions from population genomic data. *Molecular ecology*, 25(11):2387–2397, 2016.

[36] Camille Roux, Georgia Tsagkogeorga, Nicolas Bierne, and Nicolas Galtier. Crossing the species barrier: genomic hotspots of introgression between two highly divergent ciona intestinalis species. *Molecular Biology and Evolution*, 30(7):1574–1587, 2013.

[37] Claudia Solís-Lemus and Cécile Ané. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genetics*, 12(3):e1005896, 2016.

[38] Ying Song, Stefan Endepols, Nicole Klemann, Dania Richter, Franz-Rainer Matuschka, Ching-Hua Shih, Michael W Nachman, and Michael H Kohn. Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Current Biology*, 21(15):1296–1301, 2011.

[39] 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. doi: 10.1093/sysbio/syx035. URL `http://dx.doi.org/10.1093/sysbio/syx035`.

[40] Maureen Stolzer, Han Lai, Minli Xu, Deepa Sathaye, Benjamin Vernot, and Dannie Durand. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, 28(18):i409–i415, 2012.

[41] Gergely J Szöllősi, Wojciech Rosikiewicz, Bastien Boussau, Eric Tannier, and Vincent Daubin. Efficient exploration of the space of reconciled gene trees. *Systematic Biology*, 62(6):901–912, 2013.

[42] Gergely J Szöllősi, Eric Tannier, Nicolas Lartillot, and Vincent Daubin. Lateral gene transfer from the dead. *Systematic Biology*, 62(3):386–397, 2013.

[43] Gergely J Szöllősi, Adrián Arellano Davín, Eric Tannier, Vincent Daubin, and Bastien Boussau. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Phil. Trans. R. Soc. B*, 370(1678):20140335, 2015.

[44] Simon Tavaré. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on Mathematics in the Life Sciences*, 17(2):57–86, 1986.

[45] Yuyu Wang, Xiaofan Zhou, Ding Yang, and Antonis Rokas. A genome-scale investigation of incongruence in culicidae mosquitoes. *Genome Biology and Evolution*, 7(12): 3463–3471, 2015.

[46] Robert M Waterhouse, Fredrik Tegenfeldt, Jia Li, Evgeny M Zdobnov, and Evgenia V Kriventseva. Orthodb: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Research*, 41(D1):D358–D365, 2012.

[47] Dingqiao Wen and Luay Nakhleh. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Systematic Biology*, 67(3):439–457, 2017.

[48] Dingqiao Wen, Yun Yu, Matthew W Hahn, and Luay Nakhleh. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Molecular Ecology*, 25(11):2361–2372, 2016.

[49] Dingqiao Wen, Yun Yu, Jiafan Zhu, and Luay Nakhleh. Inferring phylogenetic networks using phylonet. *Systematic Biology*, 67(4):735–740, 2018.

[50] Ziheng Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10(6):1396–1401, 1993.

[51] Yun Yu and Luay Nakhleh. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, 16(10):S10, 2015.

[52] Yun Yu, R Matthew Barnett, and Luay Nakhleh. Parsimonious inference of hybridization in the presence of incomplete lineage sorting. *Systematic Biology*, 62(5):738–751, 2013.

[53] Chi Zhang, Huw A Ogilvie, Alexei J Drummond, and Tanja Stadler. Bayesian inference of species networks from multilocus sequence data. *Molecular Biology and Evolution*, 35(2):504–517, 2017.

Appendix A

Code scripts

A.1 MrBayes code

We ran MrBayes for the aligned DNA sequences within each gene family as following,

```
begin mrbayes;
set autoclose=yes nowarn=yes;
execute data.nex;
prset brlenspr=unconstrained:exp(10.0);
mcmcp ngen=10000000;
mcmcp Nchains=1;
lset nst=6 rates=invgamma ngammacat=4;
prset tratiopr = beta(1, 1);
lset nst=6;
mcmc;
sumt;
end;
```

Appendix B

Figures

B.1 NEXUS Data Block

```
BEGIN DATA;
 DIMENSIONS NTAX=2 NCHAR=492;
 FORMAT DATATYPE=DNA INTERLEAVE MISSING=-;
[Name: ACOM034036       Len:   492  Check:    0]
[Name: ACHR001523       Len:   492  Check:    0]

MATRIX
ACOM034036          ATGACGGATGTTTCGGCCAA AACTGAAAAGACCCCGGTCA CCACCTCCGAGAAAGAGGTA GAGTCGGACGAGGCAACGCC ACAAGTAGCGCCGGCCGACG
ACHR001523          ATGACGGATGTTTCGGCTAA AACTGAAAAGACCCCGGTCA CCACCTCCGAGAAAGAGGTA GAGTCGGACGAGGCGACGCC ACAAGTAGCGCCAGCCGATG

ACOM034036          CGACAGAGGAAACAGCAACC ACCGATAAAACTGCCGCCGT TGAAACCGACGGTGCCGCCG CCGACGCCGCAGAAGTAGCG GAGAAGAACGGCGAAGAAGA
ACHR001523          CGACAGAGGAATCTGCTACC ACCGAAAAAACT---GCCGT TGAAACCGAAGCTGCTGCCA CCGACGCCGCGGAAGTAGCG GAGAAGAACGGAGAAGAAGA

ACOM034036          GCCGGCCAAGGCG---GACA AAGACAACGGTGTCGCCACG GAGGAGGAGGCCACCACCAC CGCCGAGGGCGAACCGAAAG AATCCTCGACCGAGGACGGC
ACHR001523          GCCGGCTAAAGCGACAGATA AAGACAACGGTGTCGCTACG GAGGAGGAAGCCGCCACCAC CACCGAGGCCGAACCGAAAG AATCCTCCAACGAGGAAGGT

ACOM034036          GAGGAGGGCAAGGAGGAGGA CGGTGCTGCTGACAAGGCGG CCGAGGCCACCAAGAGGAAA GCCACCGAC---GTGAAGGC TGACGGTGCGGCGGCCGCGG
ACHR001523          GAGGAGAGCA---AGGAGGA CGGTGCTGCTGACAAGGCGG CCGAGGCCACCAAGAGGAAA GCCACCGAAGTGGTGAAGGC TGACGGTGCGGCGGTCGTCG

ACOM034036          AGGTTGACGGTGCCGAGCAT ACGACGCCCGAGAAGAAGGC CAAACTAGACGACAGCAGCG ACGCCAAAGCTGCCGAGGAA GTTTCCACTTAG
ACHR001523          AGGTTGACGGTGCCGAGCAT ACGACGCCCGAGAAGAAGGC CAAACTAGACGACAGCAGCG ACGCCAAAGCTGCCGAGGAA GTTTCCACTTAG
```

Figure B.1: An example of a NEXUS data block from one of the *Anopheles* gene families. As seen in the figure, the DNA sequence alignments has been translated to a 2 (number of taxa) $\times$ 492 (number of sites) aligned matrix, which will be taken as input for MrBayes.

B.2 Reconciliation Likelihood in ALE

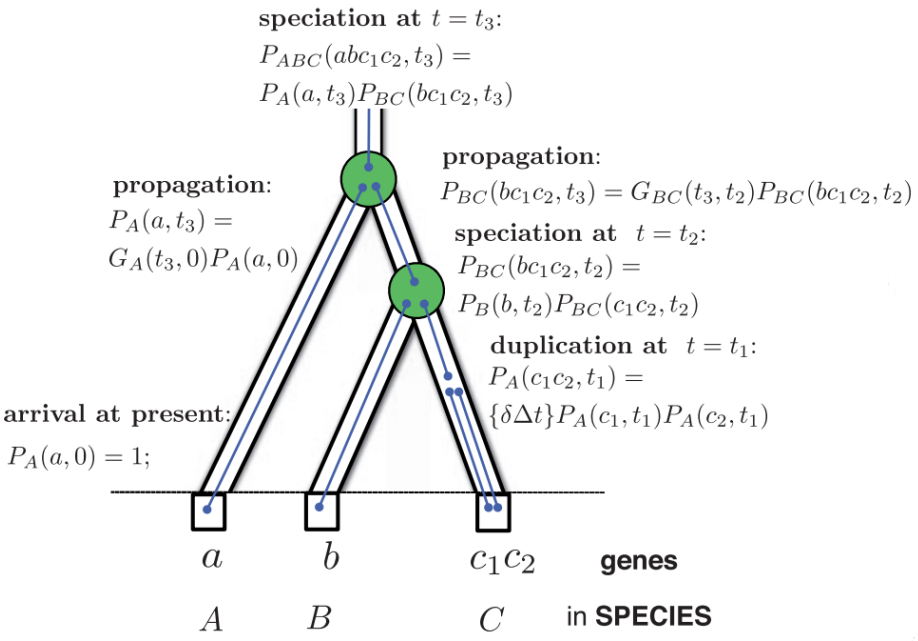

Figure B.2: Reconciling a gene tree τ (blue line) with the species tree S (outside tube) that involves a duplication and two speciations. This figure is adopted from Figure 1 created by Szöllősi *et al* [41]. The time along S from present day to history has been discretized into $[0, t_1)$, $[t_1, t_2)$, and $[t_2, t_3)$. Note that $P_A(c_1c_2, t_1)$, $P_A(c_1, t_1)$, and $P_A(c_2, t_1)$ should be $P_C(c_1c_2, t_1)$, $P_C(c_1, t_1)$, and $P_C(c_2, t_1)$. It calculates the probability $P_{ABC}(abc_1c_2, t_3)$ of seeing the root of τ at the root of S using reconciliation events that map τ into S (some terms are not shown). In general, the evolutionary scenario is unknown and we must sum over all possible ways to map τ into S. $G_{BC}(t_3, t_2)$ indicates the single-gene propagation probability between time t_2 and t_3 along branch splitting species B and C; similarly for $G_A(t_3, 0)$.