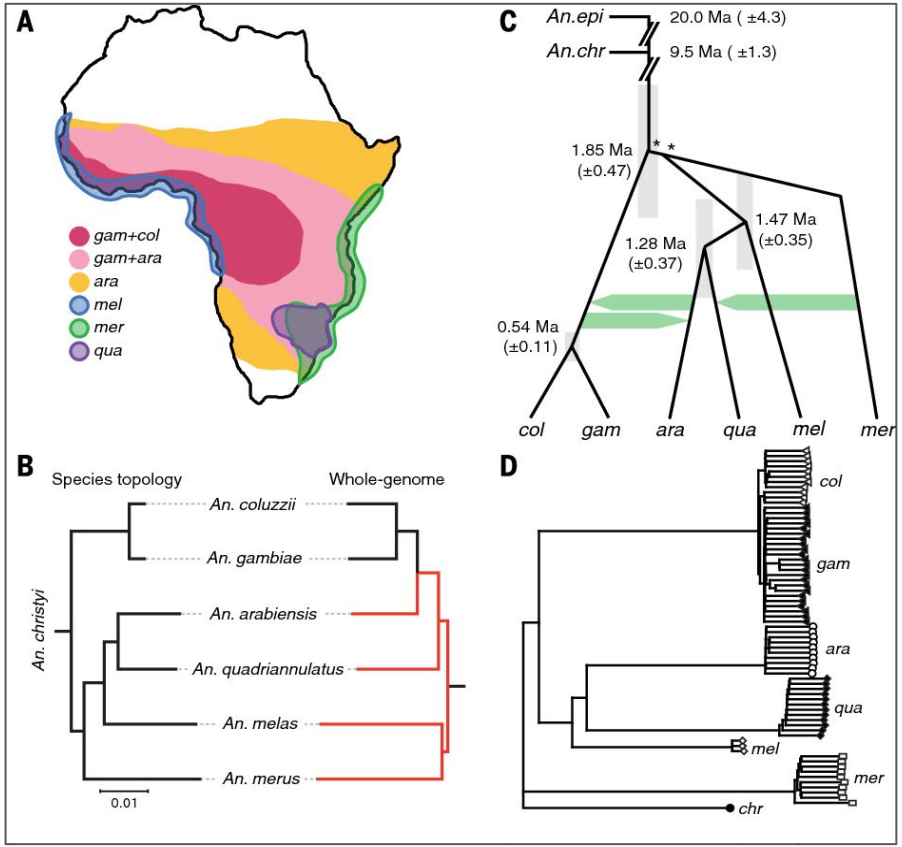
Phylogenetics for the *Gambia* complex and introgression

Cedric Chauve
Department of Mathematics
Simon Fraser University

The *Gambia* complex phylogeny controversy



[Fontaine 2015]

Extensive introgression in a malaria vector species complex revealed by phylogenomics.

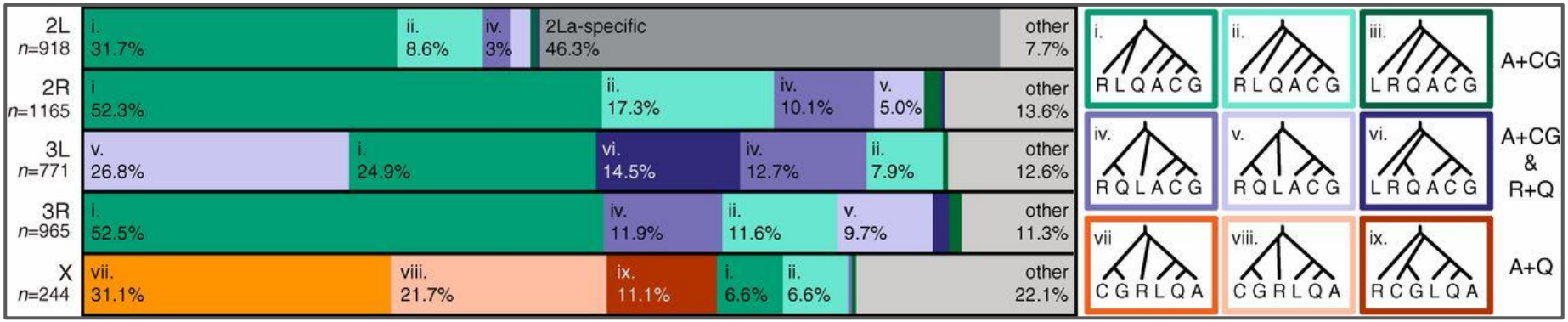
The phylogeny of the recently (6M years) diverged complex of African *Anopheles* mosquitoes (the *gambia complex*) is controversial.

The phylogenetic signal is likely difficult to capture due to extensive *introgressive hybridization* (introgression) between autosomes.

The true phylogeny for this group of species is given by the X chromosomes.

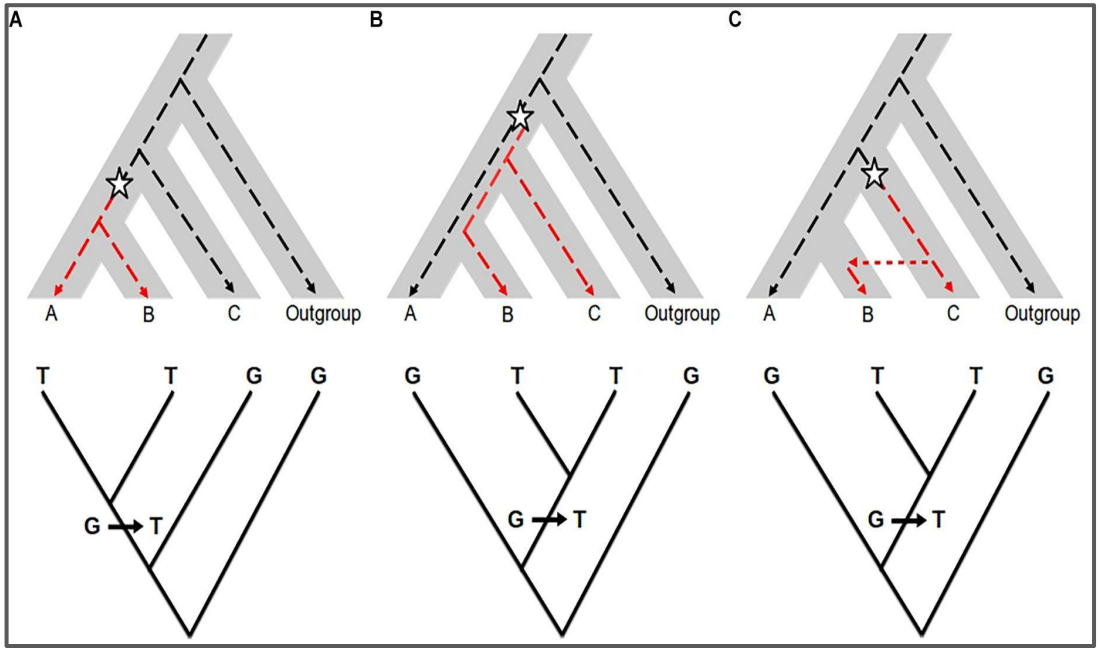
A very discordant phylogenetic signal

[Fontaine 2015]



ML-rooted phylogenies were inferred from *n* = 4063 50-kb genomic windows for *An. arabiensis* (A), *An. coluzzii* (C), *An. gambiae* (G), *An. melas* (L), *An. merus* (R), and *An. quadriannulatus* (Q), with *An. christyi*as out-group. The nine most commonly observed topologies across the genome (trees i to ix) are indicated on each of the five chromosome arms (if found) by correspondingly colored blocks whose length represents the proportion of all 50-kb windows on that arm that support the topology. The X chromosome most often indicates that A and Q are sister taxa (trees vii to ix), whereas the autosomes indicate that A and C+G are sister groups (trees i to vi). Large portions of the autosomes (particularly 3L and 3R) indicate that R and Q are sister taxa (trees iv to vi). **The most common phylogeny on the X chromosome (vii) represents the most likely species branching order.**

Two confounding factors: ILS and introgression



[Burgarella 2019]

ILS: Incomplete Lineage Sorting.
Genes diverge prior to speciation, or equivalently, gene lineages coalesce before the speciation.

Introgression.
Transfer (with replacement) of a genome segment due to hybridization between sympatric species.

Important point here.
The tree topologies are indeed the sames for ILS and introgression , but the divergence time for the confounding clade (B1) differs, especially compared to the speciation time.

Introgression as an important evolutionary factor

Hanemaaijer *et al. Malar J* (2019) 18:127
https://doi.org/10.1186/s12936-019-2759-1

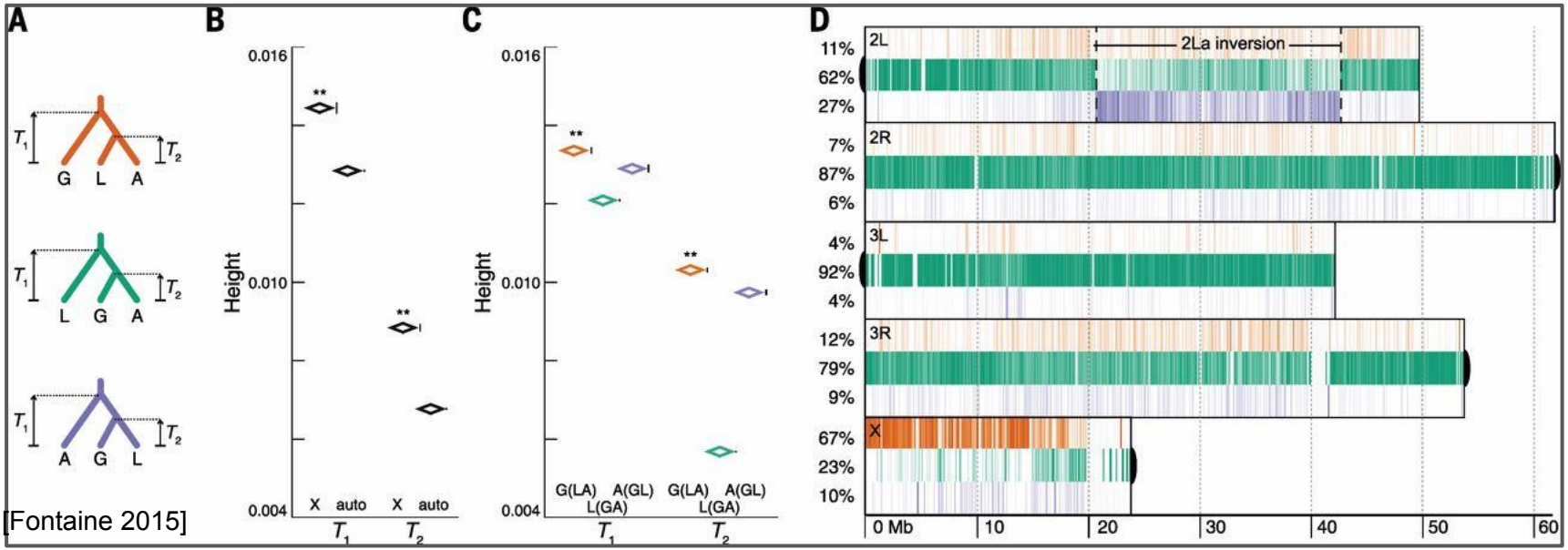
Malaria Journal

RESEARCH Open Access

Introgression between *Anopheles gambiae* and *Anopheles coluzzii* in Burkina Faso and its associations with *kdr* resistance and *Plasmodium* infection

Mark J. Hanemaaijer[1], Hannah Higgins[1], Ipek Eralp[1], Youki Yamasaki[1], Norbert Becker[2,3], Oscar D. Kirstein[1], Gregory C. Lanzaro[1*] and Yoosook Lee[1]

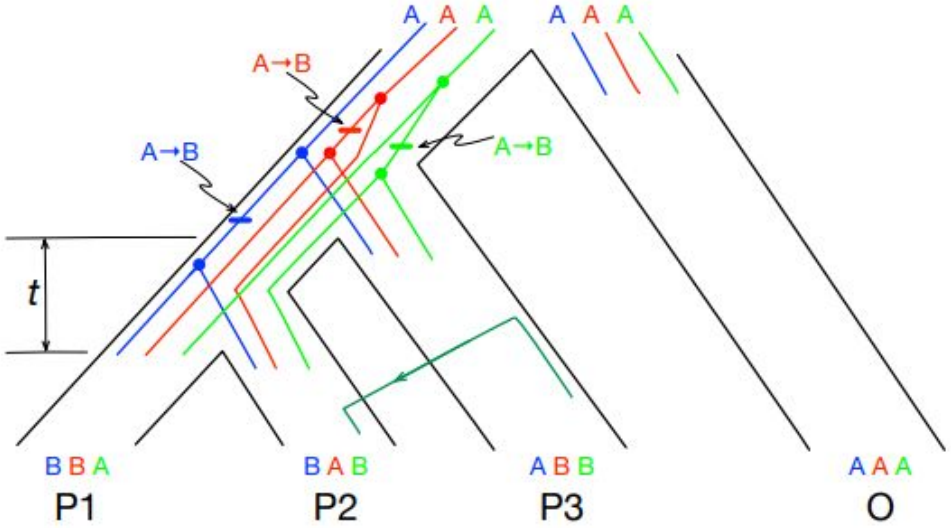
The signal for the “true” phylogeny

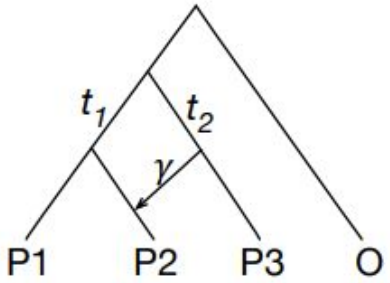


[Fontaine 2015]

The X chromosome shows significantly higher divergence times T1 and T2 than the autosomes (on trees inferred on windows od 10kb). Even among only autosomal loci, regions with the X majority relationship G(LA) (orange) have significantly higher T1 and T2, which indicates that the autosomal majority topology (green) is the result of widespread introgression between An. arabiensis and An. gambiae. The X majority relationship G(LA) therefore represents the true species branching relationships.

Disentangling ILS and introgression: the D-statistic

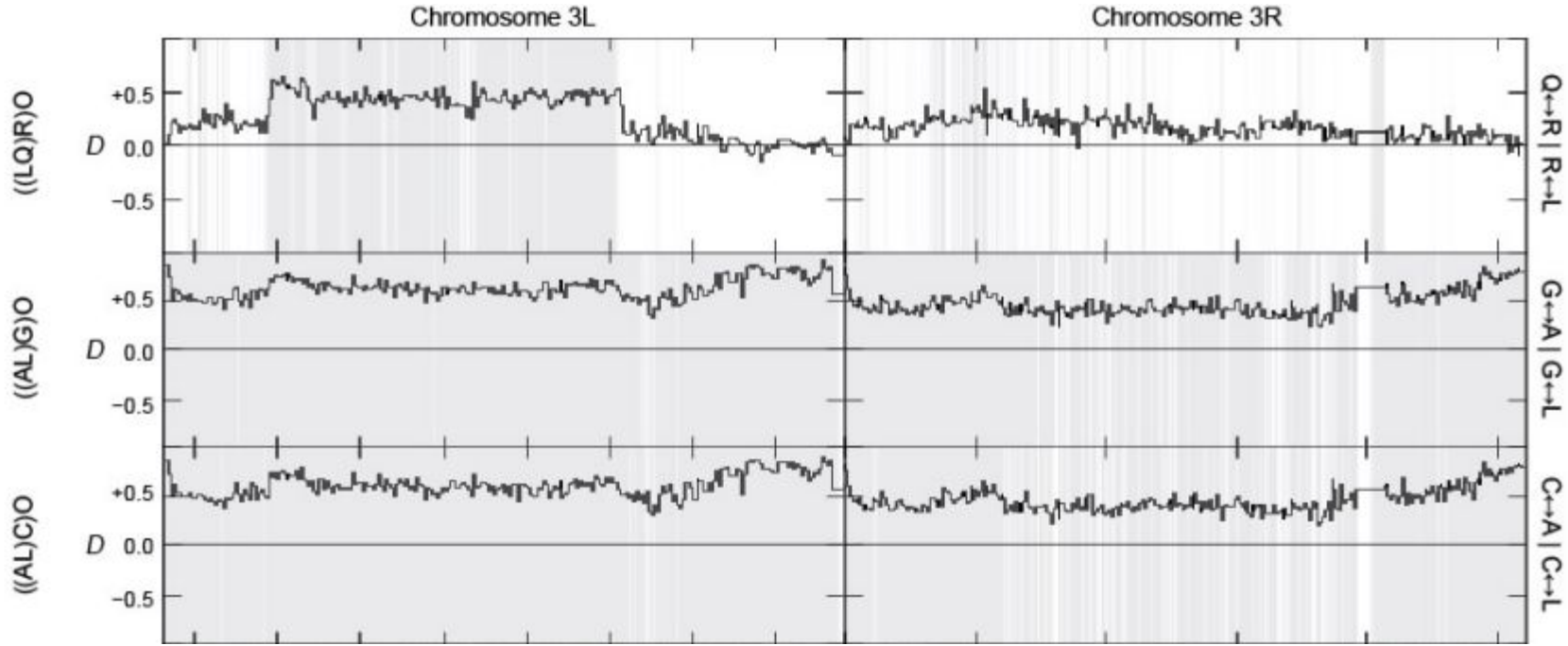




$$D = \frac{N_{ABBA} - N_{BABA}}{N_{ABBA} + N_{BABA}}$$

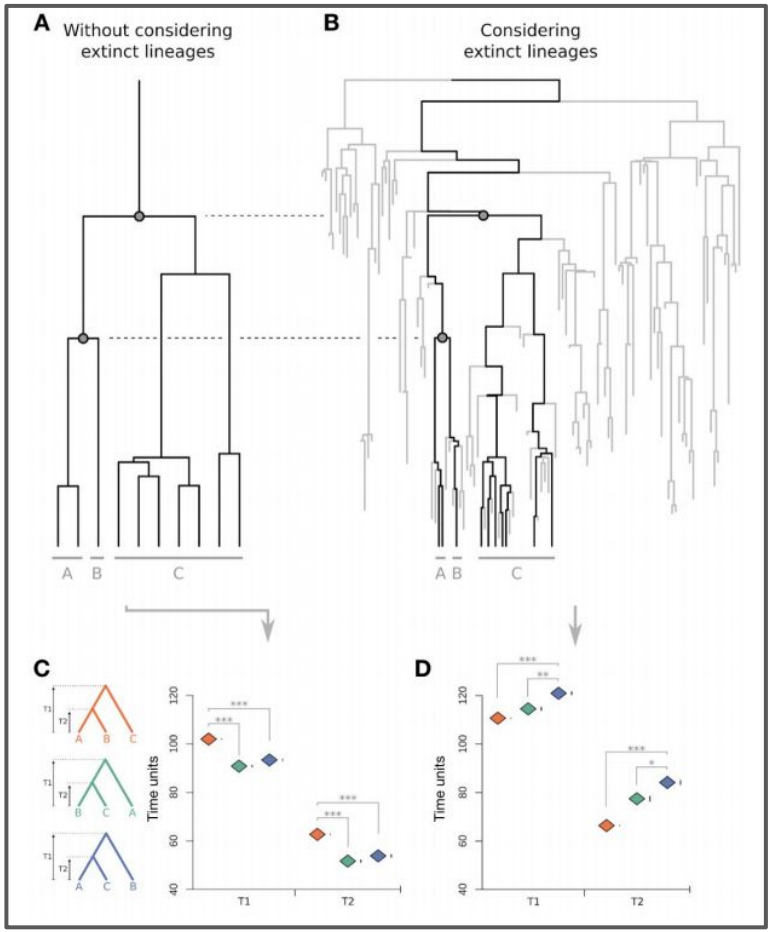
ILS or introgression P1,P2 (no phylogenetic conflict): D expected to be 0.
Introgression P2,P3: D expected to be >0.
Introgression P1,P3: D expected to be <0.
Does not indicate the introgression direction (D_FOIL does it).

The D-statistic for the *Gambia* complex



[Fontaine 2015]

The impact of missing species



The evolution of genomes (3000 genes) was simulated on two trees, one with only extant species (A), the other with all species (the Complete Species Tree, B). Mean T1 and T2 were computed for all genes supporting each one of the three possible topologies (orange, green and blue trees).The true topology (orange) is the one supported by the trees with the highest mean T1 and T2 when considering only extant species (C), but when considering also extinct lineages (D), the gene trees with the highest T1 and T2 support instead an incorrect topology.

[Davin 2018]

The impact of missing species

Title: A new species in the *Anopheles gambiae* complex reveals new evolutionary relationships between vector and non-vector species

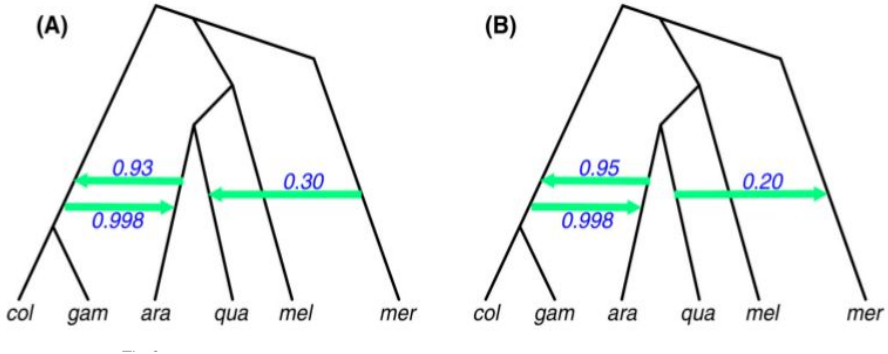
Running title: Vectorial evolution in the *An. gambiae* complex.

Authors

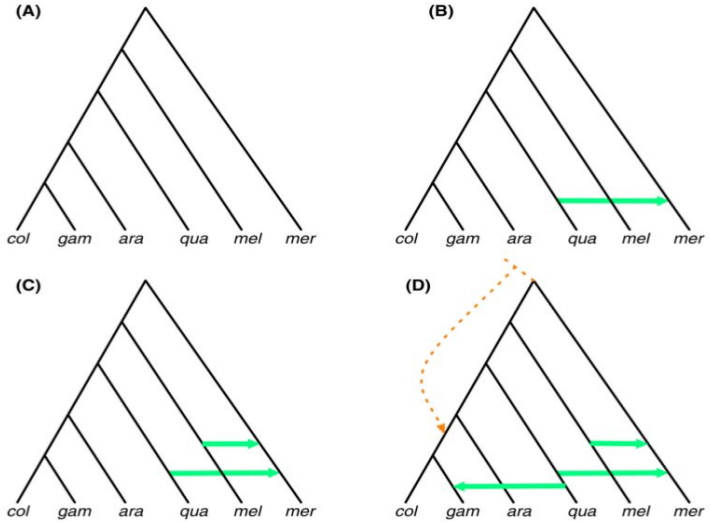
Maite G Barron[1], Christophe Paupy[2], Nil Rahola[2,3], Ousman Akone-Ella[3], Marc F. Ngangue[3,4], Theodel A. Wilson-Bahun[3], Marco Pombi[5], Pierre Kengne[2], Carlo Costantini[2], Frédéric Simard[2], Josefa Gonzalez[1,*] & Diego Ayala[2,3,*].

Phylogenetic networks to detect introgression

Introgression is a form of phylogenetic network. So one can use network inference methods to try to detect introgression events.
Wen, Yue, Hahn, Nakhleh (2016) used PhyloNet, a maximum likelihood method taking as input a a number of reticulations, a set of **single-copy gene trees** and optionally a species tree, and accounting for ILS.



Using the X chromosome tree to start



With no starting tree

The *Gambia* phylogeny, still controversial

Coalescent Analysis of Phylogenomic Data Confidently Resolves the Species Relationships in the *Anopheles gambiae* Species Complex

Yuttapong Thawornwattana,[1,2] Daniel Dalquen,[1] and Ziheng Yang*[,1,3]
[1]Department of Genetics, Evolution and Environment, University College London, London, United Kingdo
[2]Department of Microbiology, Faculty of Science, Mahidol University, Bangkok, Thailand
[3]Radcliffe Institute for Advanced Studies, Harvard University, Cambridge, MA
***Corresponding author:** E-mail: z.yang@ucl.ac.uk.
Associate editor: Koichiro Tamura

Abstract

Deep coalescence and introgression make it challenging to infer phylogenetic relationships amon that arose through radiative speciation events. Despite numerous phylogenetic analyses and t genomes, the phylogeny in the *Anopheles gambiae* species complex has not been confidently re over 80, 000 coding and noncoding short segments (called loci) from the genomes of six member and use a Bayesian method under the multispecies coalescent model to infer the species tree, w genealogical heterogeneity across the genome and uncertainty in the gene trees. We obtained a species tree from the distal region of the X chromosome: (*A. merus*, ((*A. melas*, (*A. arabiensis*, *A. gambiae*, *A. coluzzii*))), with *A. merus* to be the earliest branching species. This species tree agree inversion phylogeny and provides a parsimonious interpretation of inversion and introgress informed by the real data suggest that the coalescent approach is reliable while the sliding-win previous phylogenomic study generates artifactual species trees. Likelihood ratio test of gen evidence of autosomal introgression from *A. arabiensis* into *A. gambiae* (at the average rate generation), but not in the opposite direction, and introgression of the 3L chromosomal regi *A. quadriannulatus*. Our results highlight the importance of accommodating incomplete lineage s in phylogenomic analyses of species that arose through recent radiative speciation events.

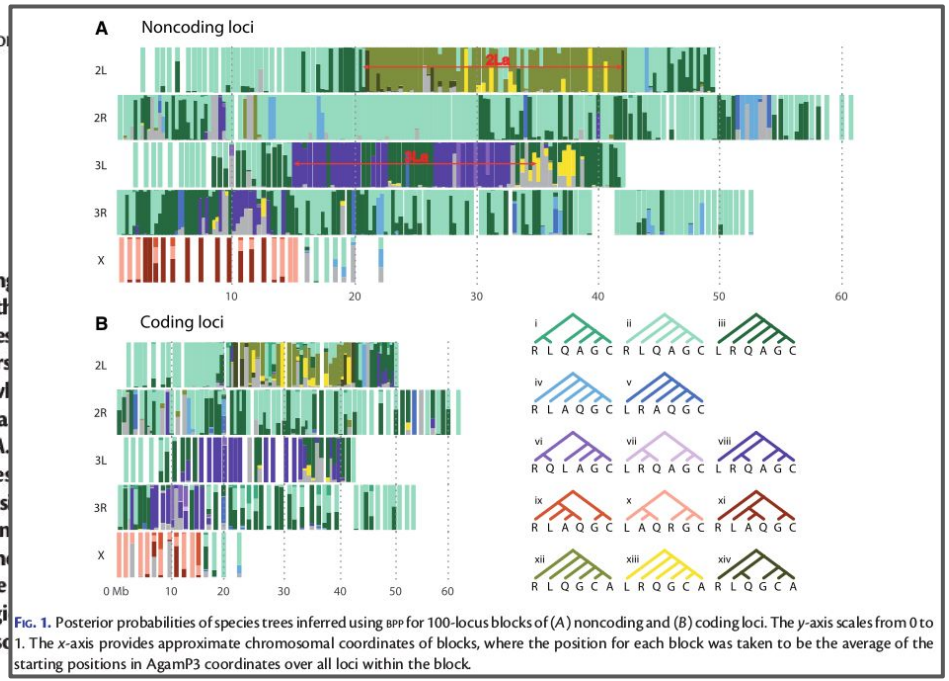


FIG. 1. Posterior probabilities of species trees inferred using BPP for 100-locus blocks of (*A*) noncoding and (*B*) coding loci. The *y*-axis scales from 0 to 1. The *x*-axis provides approximate chromosomal coordinates of blocks, where the position for each block was taken to be the average of the starting positions in AgamP3 coordinates over all loci within the block.

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

Cedric Chauve, Jingxue (Grace) Feng, Liangliang Wang

Departments of Mathematics and Statistics, Simon Fraser University

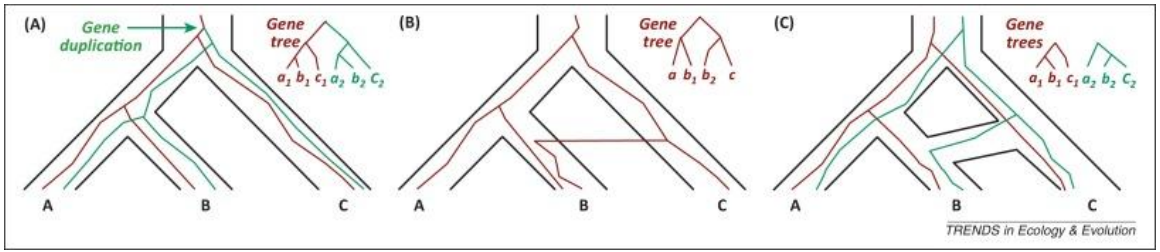
RECOMB-CG 2018

SFU
SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Introduction: gene tree/ species tree reconciliation

Observation: Introgression leaves a phylogenetic signal similar to HGT.

Reconciliation aims to explain the **discordance** between the **gene tree** of a gene family and a given **species phylogeny** by a parsimonious/ML evolutionary scenario involving non-speciation events, **gene duplication**, **gene loss** and **HGT** events (the **DTL model**).



[Nakleh 2013]

Hypothesis: Reconciliation provides an **efficient** approach to detect introgression events, when a **species phylogeny** is given, using the **full gene complement** of a dataset.

Important issue: reconciliation alone can not always distinguish **ILS** from **HGT**.

Data/Method: overview

Data: 14 Anopheles genomes; gene clustered into ~ 17,000 families; Multiple Sequence Alignment (MSA) for each family.

Step 1: Computing reconciled gene trees and "HGT" events

- MrBayes: sampling gene trees from the MSA.
- ALE: sampling amalgamated reconciled gene trees and recording HGT events.

Step 2: Assessing HGT events time consistency using MaxTiC

- Checking that the frequent HGTs are not time inconsistent; proxy for HGT accuracy.

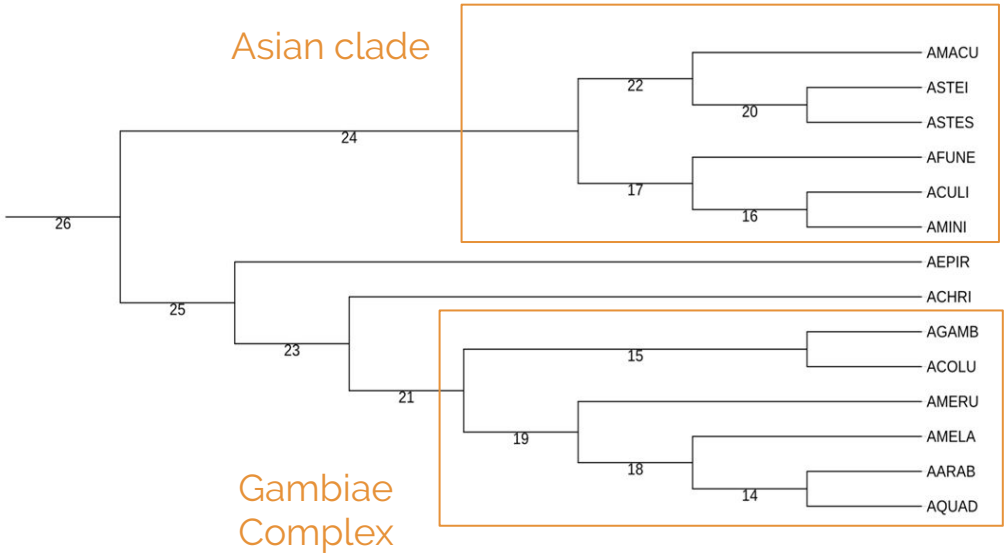
Step 3: Calling introgression events

- Detecting pairs of species with a high number of frequent HGTs
- Statistical test to detect genome regions with a large number of HGT-genes, that are unlikely to result from ILS.

Data: species tree

The species tree is the “X-phylogeny” used in [Anselmetti 2018]

Undated species phylogeny within the gambiae complex is taken from [Fontaine 2015]

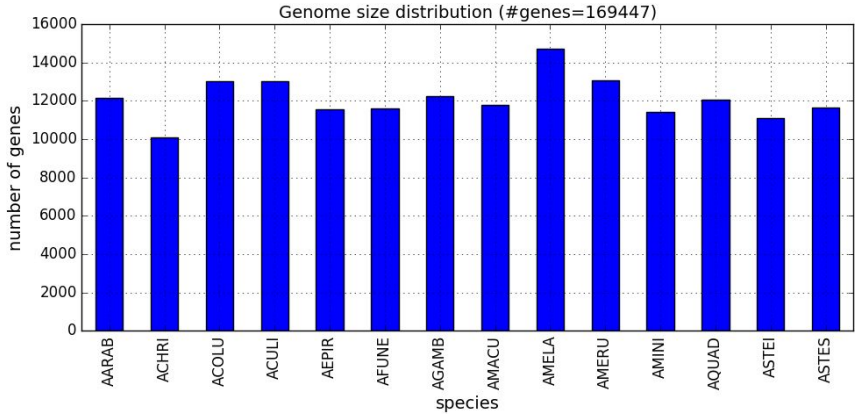


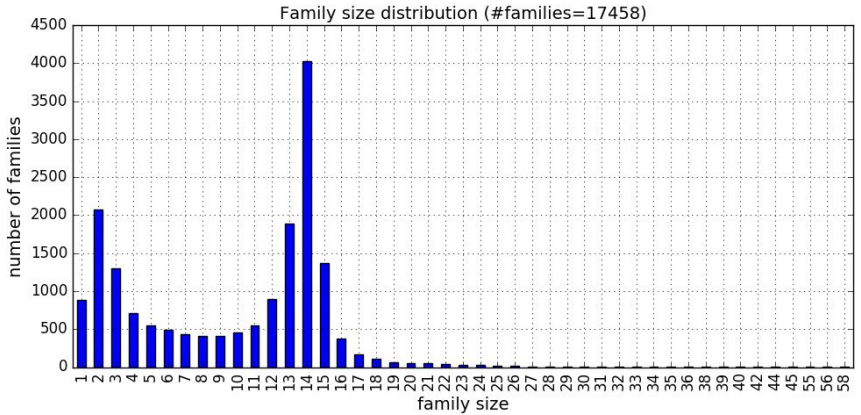
Data

The genomes of 14 African and Asian *Anopheles* mosquitoes, spanning roughly 40M years of evolution.

Genomes come from fully assembled (AGAMB) to highly fragmented (AMACU, ~10K scaffolds).

Genes are clustered into families using the OrthoDB, and MSA are taken from [Anselmetti 2018].





Method: Reconciliation and HGT

MrBayes: Bayesian reconstruction of gene trees from sequence data [Ronquist 2012]

- Input: one MSA per gene family
- Output: two samples of up to 20,000 gene trees (2 independent MCMC runs)
- Discarded: families with less than 4 genes or uncertain MCMC convergence in at least one chain or less than 5,000 sampled trees.

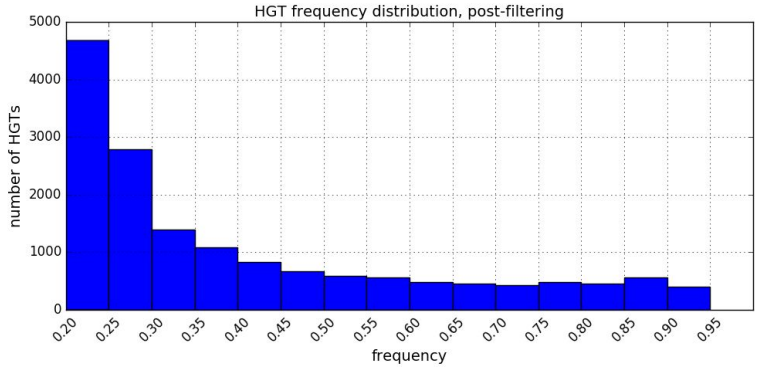
ALE: Amalgamated likelihood estimation, exploration of the space of reconciled gene trees [Szöllősi 2013]

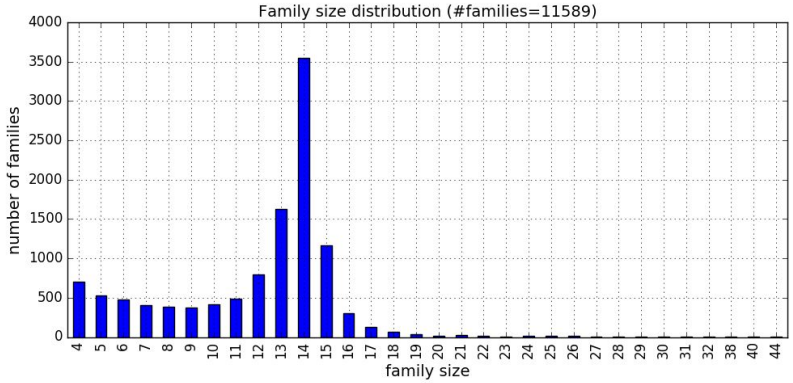
- Input: clades of a set of sampled gene trees, undated species tree
- Output: bayesian sampling of reconciled gene trees built from these clades together with one maximum likelihood amalgamated reconciled gene tree
- Discarded: families for which both ML reconciled gene trees are different.

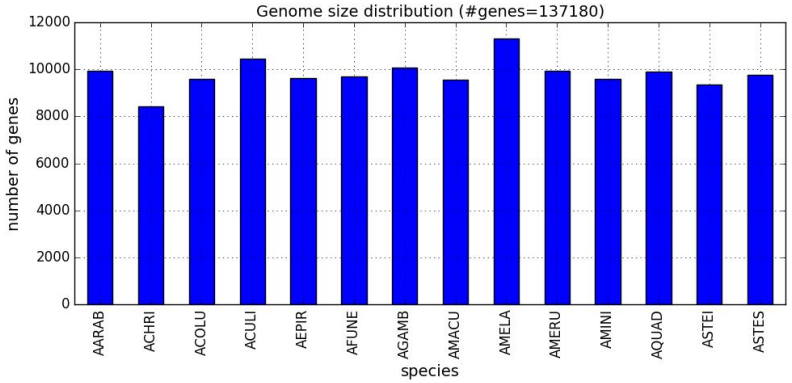
Results: “HGT” inference

Bottom (Left and Right):
Post MrBayes+ALE filtering of genes has a limited impact.

Top Right:
Most called “HGTs” are sampled with a low frequency, likely False Positives.

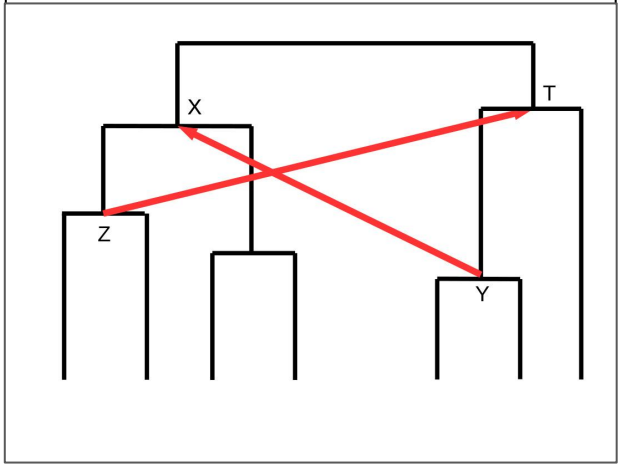






Method: Time inconsistency in called HGTs

Principle: time consistency used as a **proxy for HGT accuracy**; if the HGTs we call are wrong, they will likely result in time inconsistency.



MaxTiC: from a set of weighted time constraints inferred from HGTs, compute a maximum weight set of constraints that is time consistent. [Chauve 2017]

Consistency ratio =
Weight kept constraints / weight discarded constraints.
Weight of a constraint = sum of the HGTs frequencies defining this constraint.

Application: for each gene family, filter out all HGTs below a given **frequency threshold *t*** and run MaxTiC on the remaining HGTs. A high consistency ratio reinforces the confidence in the unfiltered HGTs accuracy.

Results: “HGTs” time consistency

	kept_constraints	weight	discarded_constraints	weight	weight_ratio
20	39	4464.978	35	447.840	0.908843
25	32	4172.676	31	377.050	0.917127
30	29	3877.516	23	315.326	0.924794
35	25	3626.852	19	285.850	0.926943
40	21	3359.890	17	245.530	0.931900
45	20	3156.614	16	220.198	0.934791
50	18	2983.730	14	196.538	0.938201
55	15	2823.456	13	164.976	0.944795
60	13	2658.032	10	138.578	0.950448
65	10	2473.694	10	117.980	0.954477
70	7	2288.016	6	90.642	0.961894
75	7	2083.748	3	67.658	0.968552
80	6	1806.846	3	48.958	0.973619
85	4	1561.264	2	22.772	0.985624
90	3	1186.730	1	1.800	0.998486
95	1	671.194	0	0.000	1.000000

The consistency ratio (last column) is high even at low values of *t* (first column).

The number of discarded time constraints (“HGTs” creating time inconsistency) is comparable to the kept ones but with a cumulated low weight.

Method: Calling potential introgression events

Principle: a potential introgression event **from species *d* to species *r*** is called if

- ➢ we observe at least 50 gene families having “HGTs” from *d* to *r* sampled by ALE with frequency at least 50%
- ➢ the sum of the frequencies of all such “HGTs” is at least 50.

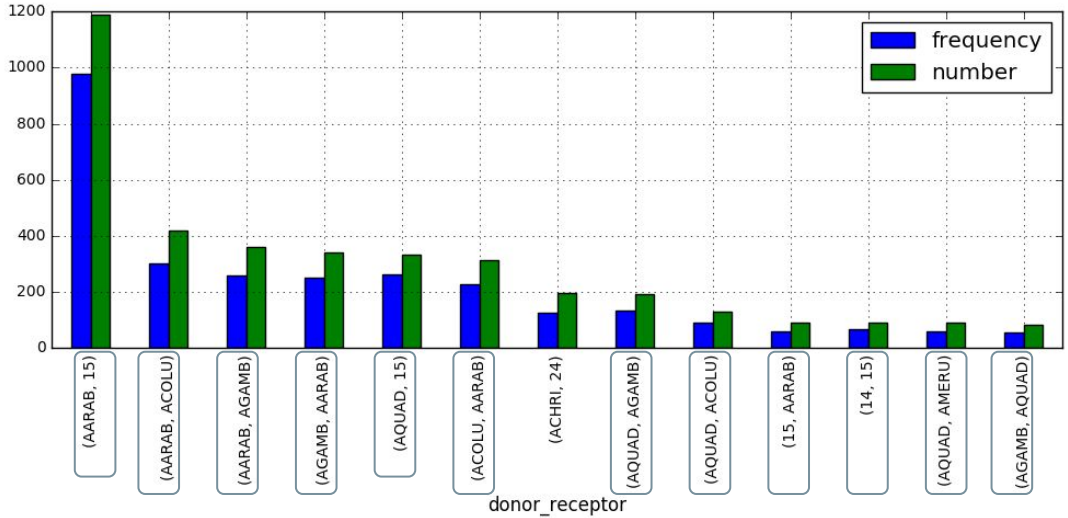
Pitfall: disentangling introgression from ILS

- ➢ ILS can also result in HGT-like patterns in reconciled gene trees;
- ➢ **Hypothesis**: introgression impacts multi-genes genome segments [Sousa 2017] while ILS is expected to impact genes randomly located along chromosomes.

Method: testing the spatial clustering of HGT genes along the chromosome of *A. gambia*

- ➢ *A. gambia* used as it is the only completely assembled genome
- ➢ Windows of 20 genes
- ➢ Benjamini-Yukieteli method to control the False Discovery Rate (FDR set at 1%)

Results: potential introgression events



15 = ancestor of AGAMB+ACOLU
14 = ancestor of AARAB+AQUAD
24 = ancestor of asian mosquitoes

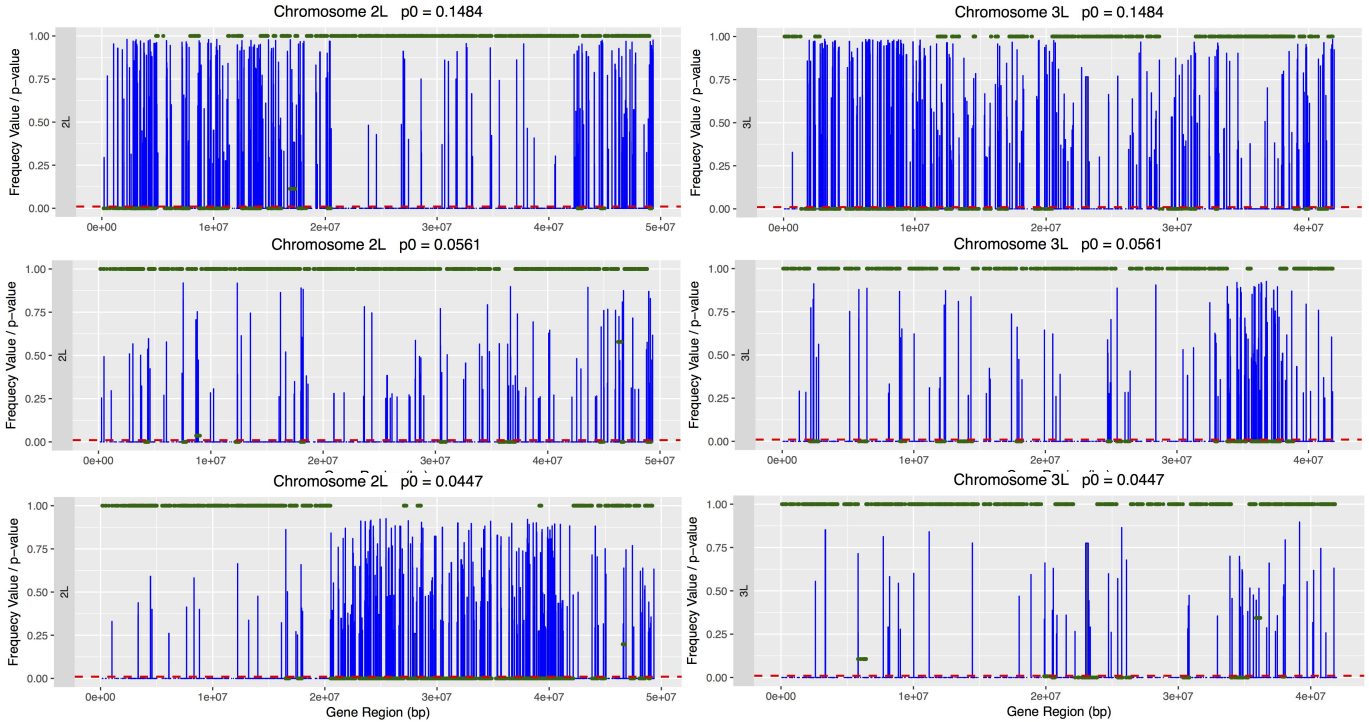
Several potential events supported by hundreds of gene families.

All but one are within the **African Gambia complex**, known for a high level of introgression [Fontaine 2015, Wen 2016] which seems confirmed here.

The question of a potential introgression event from the lineage of *A. christyi* (1 species) to ancestral species 24 is new.

Results: spatial location of HGT genes, *An. arabiensis*

All chromoplots found at https://github.com/cchauve/Anopheles_introgression_RECOMBCG_2018



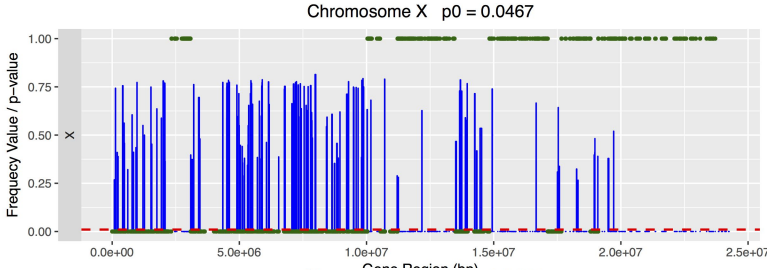
An. arabiensis
⇩
15

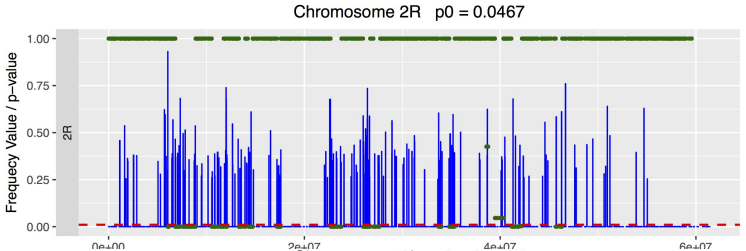
An. arabiensis
⇩
An gambia

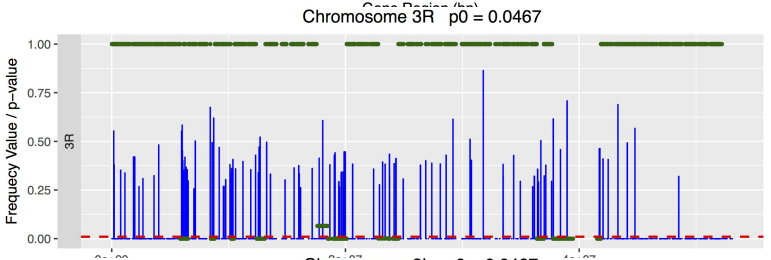
An. arabiensis
⇩
An coluzzi

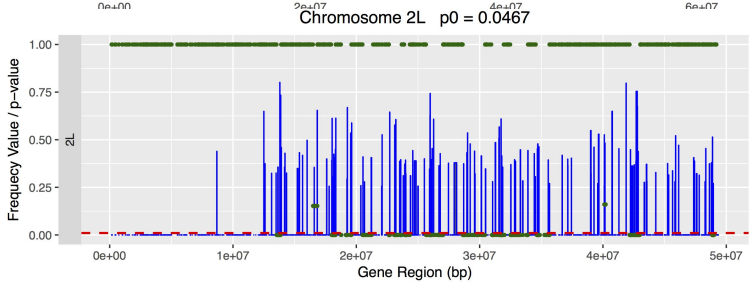
Results: A. christy -> 24

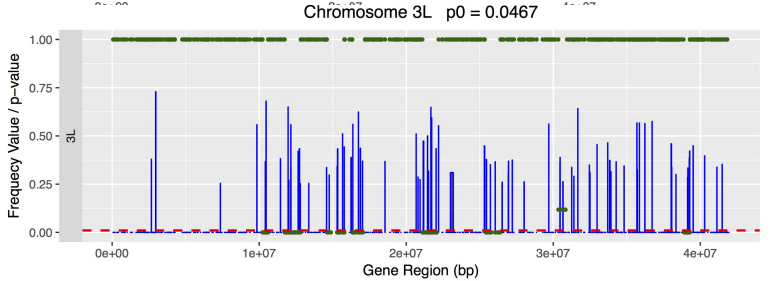
Similar support with other recent events and established events. Large segments of chromosome X seem to be introgressed.











Conclusion: summary

- We provide an **illustration/proof of concept** on a well studied dataset that reconciliation methods can be used to detect traces of introgression efficiently (resources) and without the need to rely on orthologous genes/markers, in a large dataset spanning a significant evolutionary time.

- Important methodological elements:
  - **Sampling** gene trees and HGTs, using sampling frequency to discard likely false positive HGTs
  - Statistical test of **genomic co-localization** of HGT genes (ILS vs. introgression).

- Results:
  - Extensive introgression within the Gambia complex.
  - One potential ancient introgression event.

Conclusion: future work

- Our work is an illustration on a specific dataset that an existing reconciliation algorithm (ALE) seems to be able to detect introgression. It does **not** claim to be a thoroughly tested method to detect traces of introgression.

- Next stage:
  - Simulations in a model including ILS and introgression, in order to assess the accuracy of high frequency HGT.
  - Assessing the accuracy of testing spatial co-localization is likely difficult.
  - Testing the impact of using fragmented assemblies of extant or ancestral species [Anselmetti 2018] instead of An. gambiae on the spatial co-localization test.
  - Robustness to the chosen reconciliation algorithm:
    - DTL: Ranger-DTL [Bansal 2018], ecceTERA [Jacox 2016], GTAC [Noutahi 2018]
    - DTL+ILS: NOTUNG [Stolzer 2012], [Chan 2017]
  - Simulations to assess the impact of missing species.

Appendix: testing spatial co-localization

Notations:

- **p**: the true probability of observing a gene whose evolution involves a (d, r) HGT within a gene family
- $\mathbf{p_0}$: the average of (d, r) HGT frequencies inferred from ALE for all the genes on the whole genome
- For each chromosome arm, let ***Xi*** be the number of observed HGTs from **d** to **r** in the **s** ALE-sampled reconciled gene trees for the **i**-th gene in the **j**-th window, where i = 1,...,n; j = 1,.., m

Appendix: testing spatial co-localization

Testing procedure:

- **Within j-th window, testing H_0: $p = p_0$ versus H_a: $p > p_0$**
  - $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{sn}$, $\mathrm{VAR}(\hat{p}) = \frac{p(1-p)}{sn}$
  - Test statistics $Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/(sn)}}$
  - p-value = $P(Z \geq Z_{obs})$
- **Control FDR using BY method to deal with dependencies**:

a. We want to control FDR at level $\alpha = 0.05$
b. Calculate p-value for the hypothesis test and order $p_{(1)} \leq \cdots\cdots \leq p_{(g)} \leq \cdots\cdot \leq p_{(m)}$
c. Multiply each $p_{(g)}$ by its adjustment factor $a_g = lm / g$, with $l = \sum_{k=1}^{m} 1/k$ for g = 1,...,m
d. If the multiplication in step c. does not follow the orininal ordering, apply a step-up method: $\tilde{p}_{(g)} = \min_{j=1,\ \ldots,\ m} a_j p_j$
e. Set $\tilde{p}_{(g)} = min(\tilde{p}_{(g)}, 1)$ for all g.