

# Inference of gene flow under MSC-I and MSC-M

**Ziheng Yang**

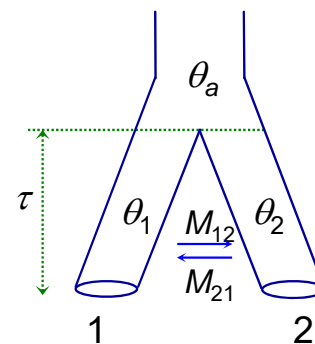
Department of Genetics, Evolution, and Environment

University College London

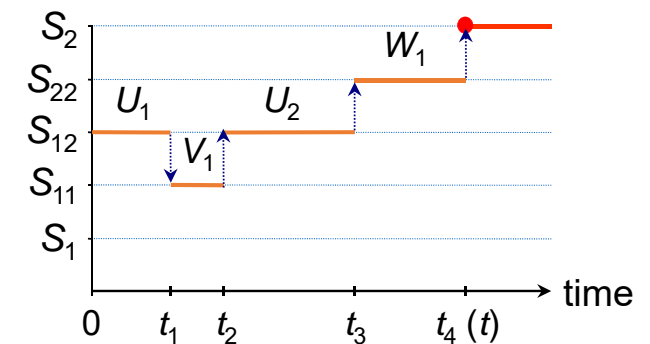
## MSC has many applications & extensions

- Inference of species divergences and population sizes
- Estimation of migration patterns and rates (IMa, etc.)
- Introgression & hybridization
- Species tree estimation (STEM, BEST, \*BEAST, BPP etc.)
- Species delimitation (BPP)
- ...

(a) Species tree



(b) Migration trajectory



MSC or coalescent is the biological process of reproduction viewed backwards in time

$H_0$ : MSC (null model)

$H_1$ : MSC + population structure

$H_2$ : MSC + hybridization

$H_3$ : MSC + recombination

$H_4$ : MSC + population structure + hybridization

etc.

(Degnan JH. 2018.  
Syst. Biol. 67:786-799)

Some terminologies are confusing:

*“to distinguish hybridization from lineage sorting”*

*“investigate whether the conditions of applicability of coalescence-based methods are met ...”*

# Gene flow

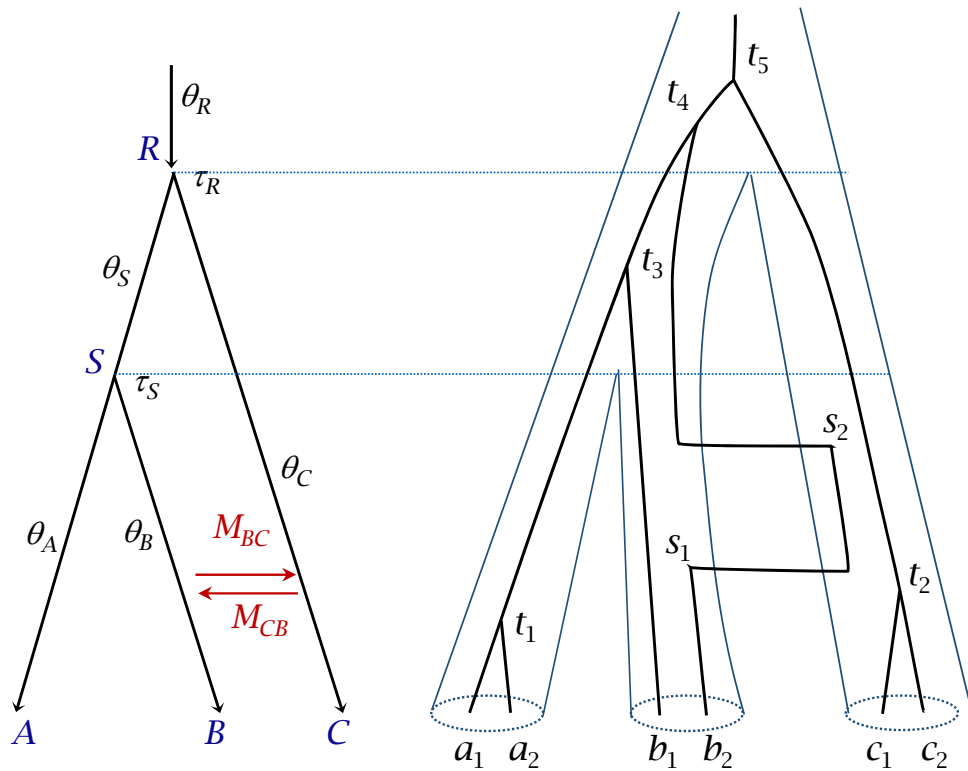
MSC + migration

MSC + introgression

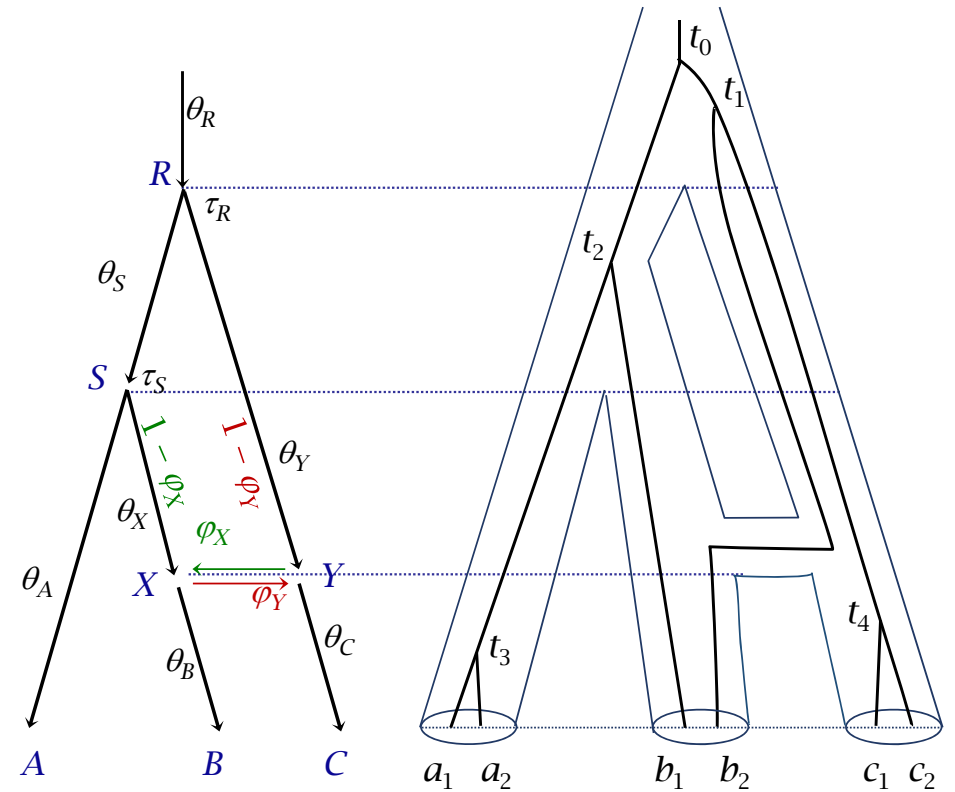
Hibbins MS, Hahn MW. 2022. Phylogenomic approaches to detecting and characterizing introgression. *Genetics* 220:10.1093/genetics/iyab1173.

Jiao X, Flouri T, Yang Z. 2021. Multispecies coalescent and its applications to infer species phylogenies and cross-species gene flow. *Nat Sci Rev* 8:DOI: 10.1093/nsr/nwab1127.

## MSC-M (migration)

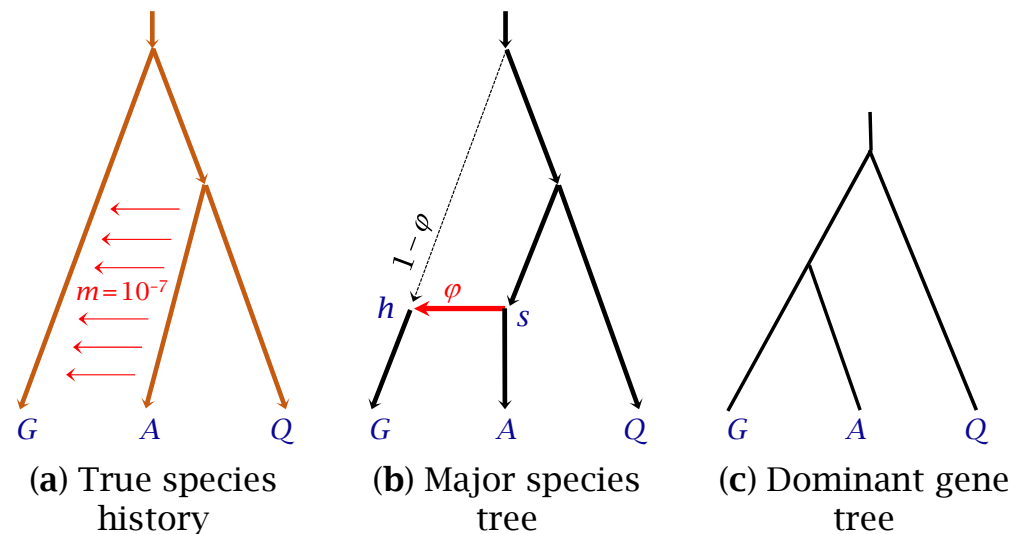
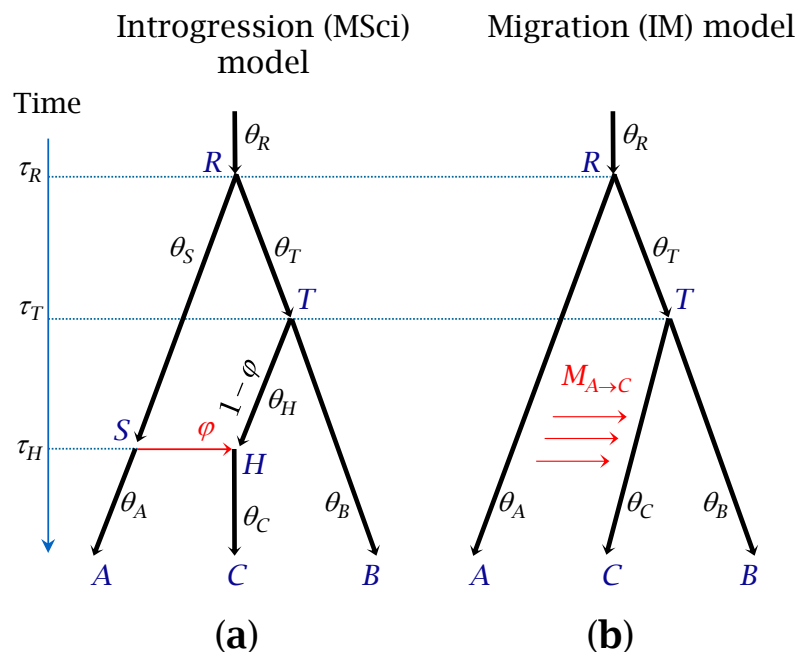


## MSC-I (introgression)



# Impact of gene flow on species trees

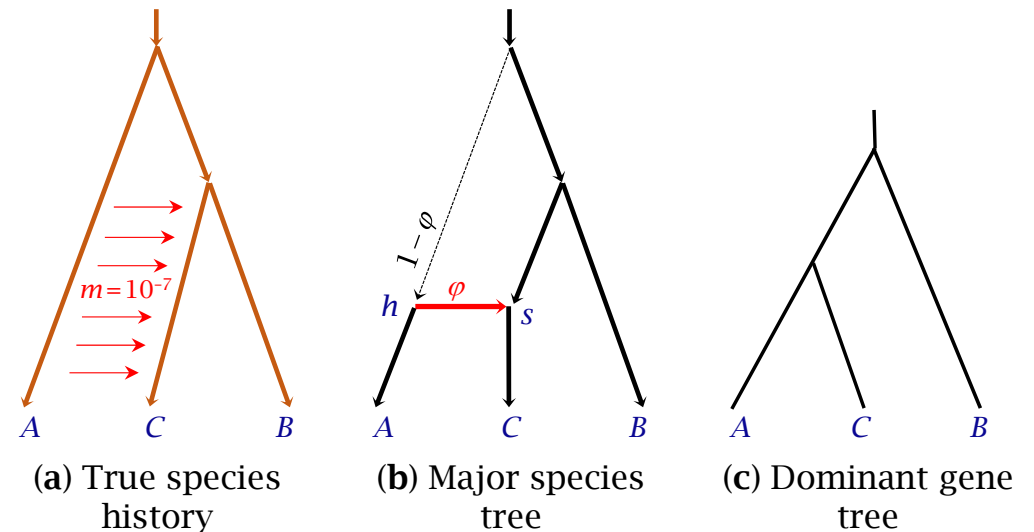
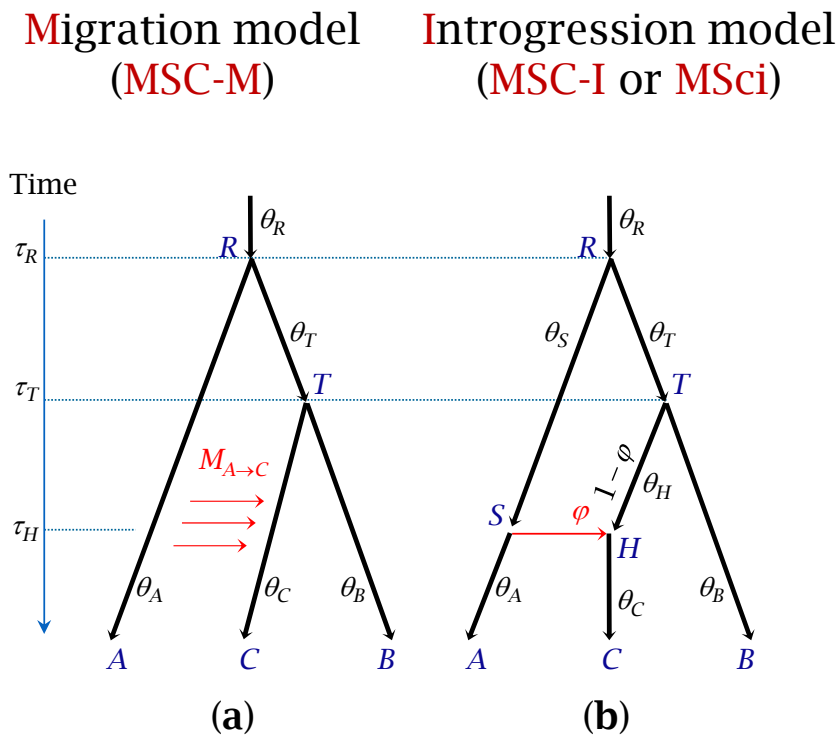
When the species tree is already a hard one (with short internal branches), even a small amount of gene flow ( $Nm < 1$  migrants per generation) can change the ‘genetic history’.



Jiao X, Flouri T, Rannala B, Yang Z. 2020. The impact of cross-species gene flow on species tree estimation. Syst Biol. 69: 830-847, 10.1093/sysbio/syaa001

# Impact of gene flow on species trees

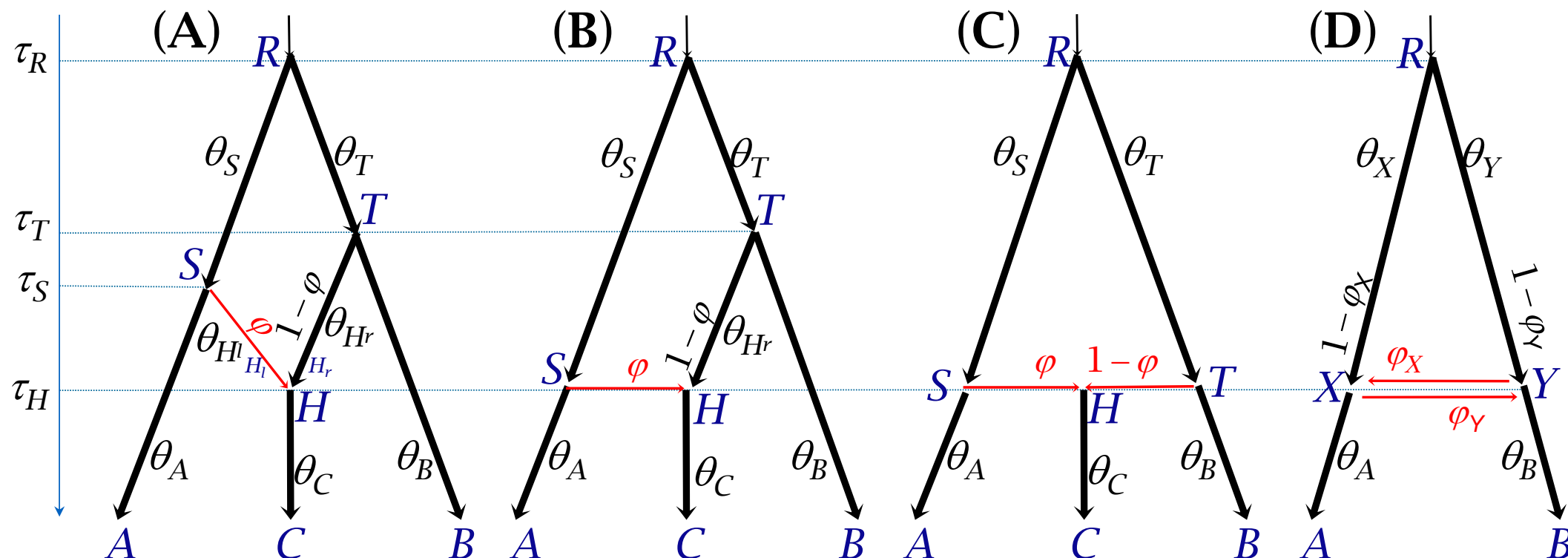
When the species tree is already a hard one (with short internal branches), even a small amount of gene flow ( $Nm < 1$  migrants per generation) can change the ‘genetic history’.



Jiao X, Flouri T, Rannala B, Yang Z. 2020. The impact of cross-species gene flow on species tree estimation. Syst Biol. 69: 830-847, 10.1093/sysbio/syaa001

# Different types of MSC-I models in BPP

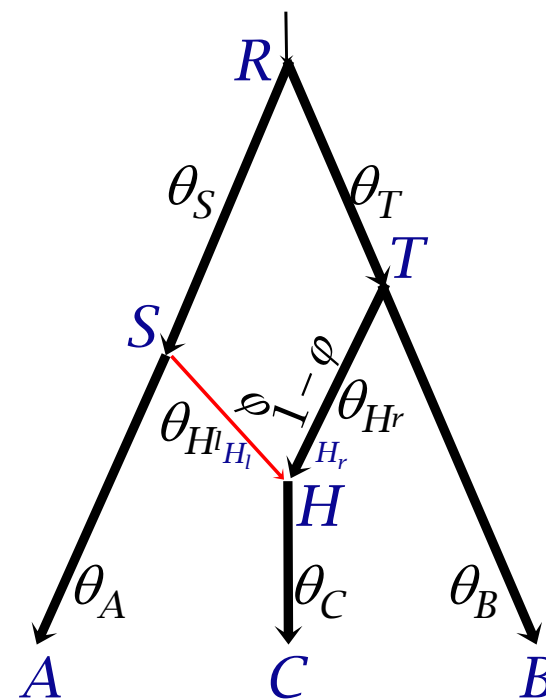
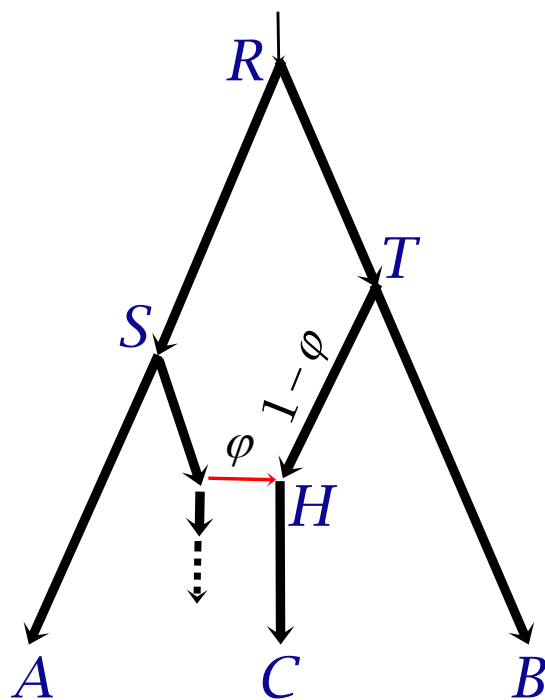
Time



$$\gamma \equiv \varphi$$



Ghost lineages (extinct or unsampled species) lead to model A



$$M_{AB} = N_B m_{AB}$$

$m_{AB}$  is the proportion of immigrants from  $A$  in the recipient population  $B$ , not the proportion of emigrants in donor population  $A$ .

(It does not matter what percentage of individuals leave population  $A$ , but it matters hugely what percentage of individuals in population  $B$  are aliens.)

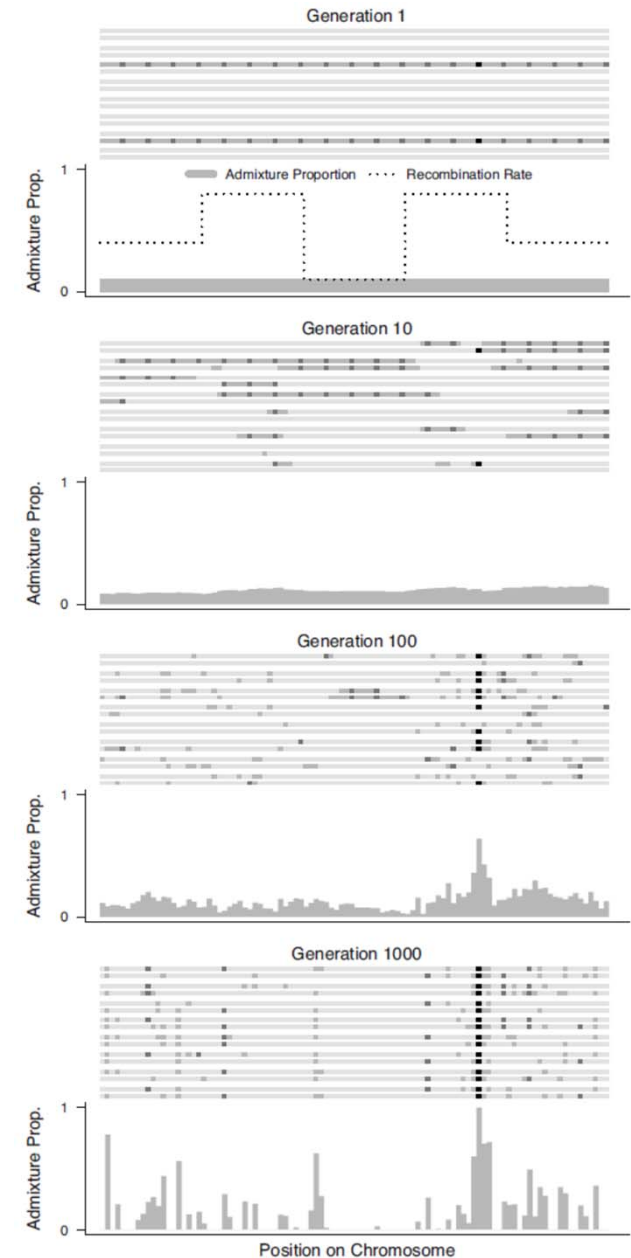
$M_{AB} = N_B m_{AB}$  is the expected number of migrants from  $A$  to  $B$  per generation.

Rates of gene flow estimated from genomic data ( $\phi$  in MSC-I or  $M$  in MSC-M) are effective rates.

They reflect the long-term effects of introgression, selection, and genetic drift, influenced by the local recombination rate.

Martin SH, Jiggins CD. 2017. Interpreting the genomic landscape of introgression. *Curr Opin Genet Dev* 47: 69-74.

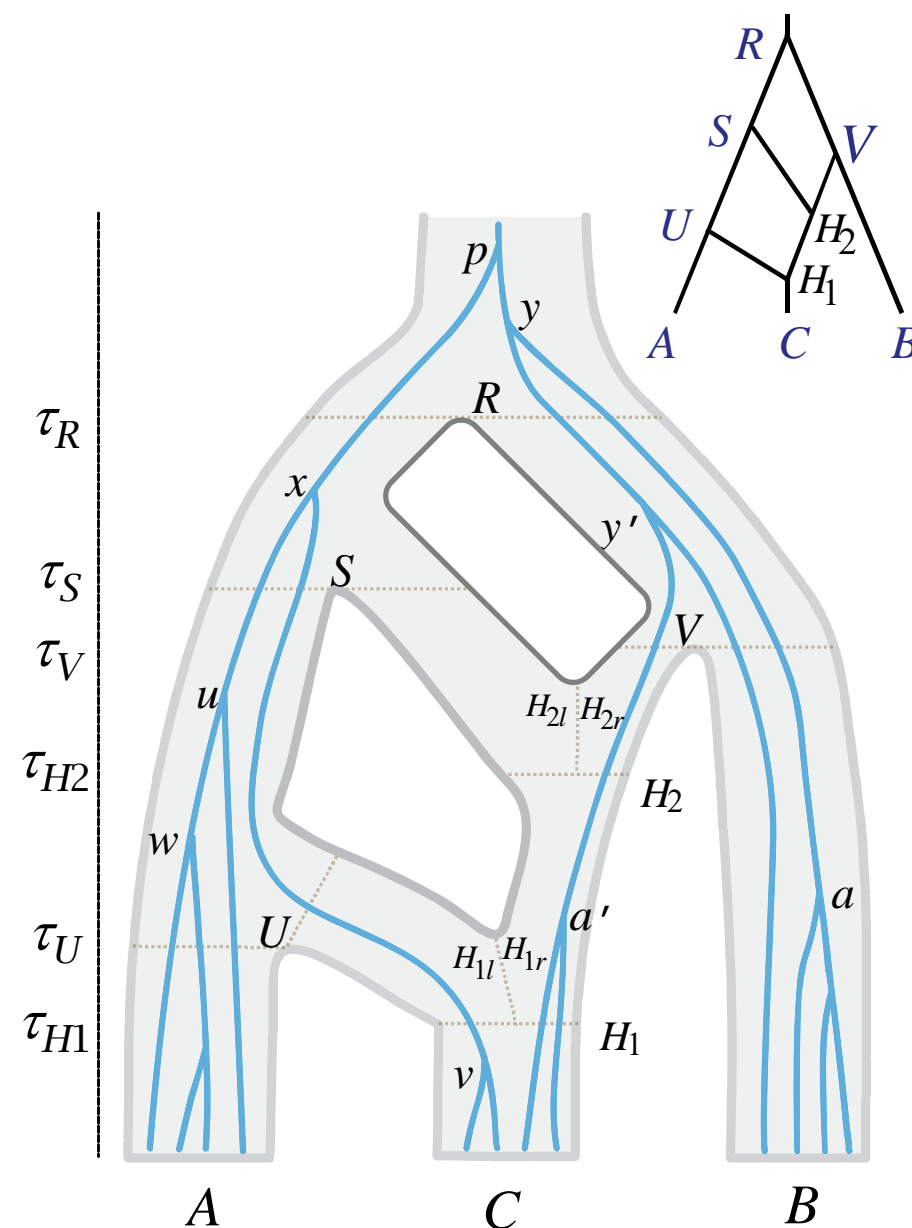
Westram AM, Stankowski S, Surendranadh P, Barton N. 2022 What is reproductive isolation? *J Evol Biol* 2022, 35: 1143-1164.



# MSC-I in BPP

1. Initialize  $\{\theta, \tau, \varphi\}, \{G_i, \mathbf{t}_i\}$ .
2. Iterate
  - change gene-tree node age.
  - change gene-tree topology (SPR).
  - change parameters ( $\theta_S, \tau_S, \varphi_S$ ).
  - Save every  $k$  iterations.

Flouri T, Jiao X, Rannala B, Yang Z. 2020. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Mol. Biol. Evol.* 37:1211-1223.



# Introgression inference programs

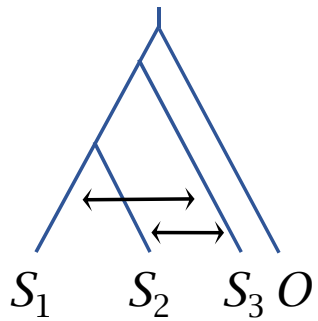
Gene tree topologies	Gene tree topologies and branch lengths	Sequence data
<p>Meng &amp; Kubatko (2009)</p> <p>PhyloNet (Yu et al. 2011, 2012)</p> <p>SnaQ/PhyloNetworks (Solís-Lemus &amp; Ane 2016).</p> <p>Unrooted gene tree topologies are used as data and pseudo-likelihood is used to fit to data of concordance factors.</p>	<p>Kubatko (2009)</p> <p>PhyloNet (Yu et al. 2014; Wen et al. 2016)</p> <p>These can be very sensitive to near zero branch-length estimates.</p>	<p><math>D</math> statistic or ABAB-BABA test (Green et al. 2010, Durand et al. 2011).</p> <p><math>D_{\text{FOIL}}</math> for 5 species (Pease &amp; Hahn 2015).</p> <p>Pools variable sites across the genome.</p>

SNaQ (Solís-Lemus and Ané 2016) is able to infer some rooted information (direction of some hybridization edges) in networks from unrooted trees.

In some cases, two networks might be indistinguishable using only gene tree topologies yet distinguishable using gene trees with branch lengths.

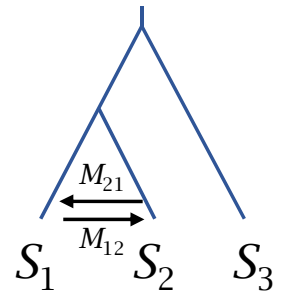
# There are a number of heuristic methods.

## ABBA-BABA test ( $D$ statistic) vs. bpp/3s



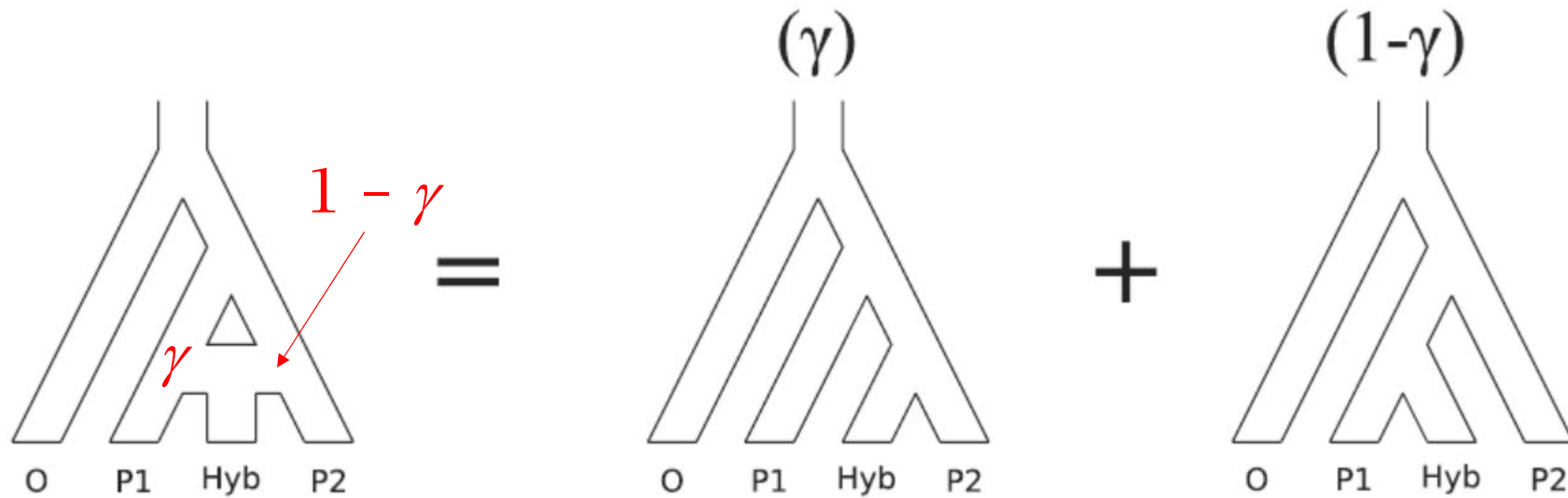
$$D = \frac{n_{ABBA} - n_{BABA}}{n_{ABBA} + n_{BABA}}$$

### 3s (likelihood ratio test)



- |   |  |
|---|--|
| <ol style="list-style-type: none"><li>1. Uses genome-wide counts (averages).</li><li>2. Tests for gene flow between <math>S_1</math> and <math>S_3</math> or between <math>S_2</math> and <math>S_3</math>, on a fixed species tree <math>((S_1, S_2), S_3), O</math>.</li><li>3. Can't identify the direction of gene flow.</li><li>4. Information in the different gene genealogies across the genome is ignored.</li><li>5. Summary statistic lacks power.</li></ol> | <ol style="list-style-type: none"><li>1. Uses short widely dispersed segments (loci) from the genome.</li><li>2. Tests for gene flow between <math>S_1</math> and <math>S_2</math>, on a fixed species tree <math>((S_1, S_2), S_3)</math>.</li><li>3. Can estimate <math>M_{12} = N_2 m_{12}</math> and <math>M_{21} = N_1 m_{21}</math>.</li><li>4. Information in the different gene genealogies across the genome is ignored.</li><li>5. LRT in theory uses all information in the data.</li></ol> |
|---|--|

# HyDe



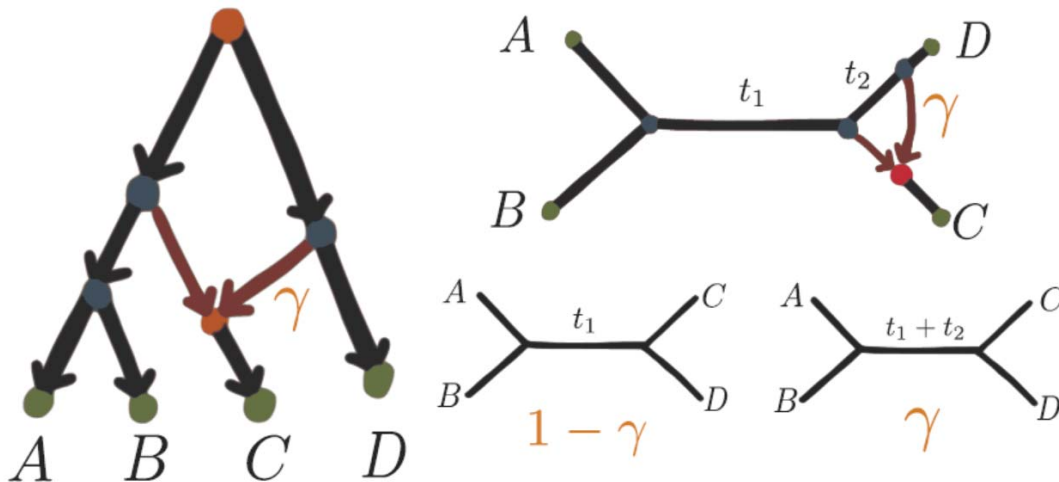
$$\frac{p_{xxyy} - p_{xyxy}}{p_{xyyx} - p_{xyxy}} = \frac{\gamma}{1 - \gamma}$$

$$\gamma = \varphi$$

Kubatko LS, Chifman J. 2019. BMC Evol Biol 19:112.  
Blischak PD, et al. 2018. Syst. Biol. 67:821-829

## SNaQ

Probabilities of the three (unrooted) quartet gene trees,  $P(G_1)$ ,  $P(G_2)$ ,  $P(G_3)$  can be used to estimate the introgression proportion ( $\gamma$ ).



Solis-Lemus C, Ane C. 2016. PLoS Genet 12:e1005896.  
Solis-Lemus C, et al. 2017. Mol Biol Evol 34:3292-3298.



## Features of summary methods

- They usually work for three species (under the clock) or four species (3+0).
- They can identify/estimate the introgression proportion and the internal branch lengths on the species tree (in coalescent units), but not other parameters in the model.
- They can't identify gene flow between sister lineages.

# Mosquitoes

Mosquitoes infested by *Plasmodium* parasites bite humans, and humans get malaria.

In 2015, there were 214 million malaria cases, 88% in Africa, 10% in SE Asia. 438K malaria deaths worldwide. (<https://www.who.int/gho/malaria/epidemic/cases/en/>).

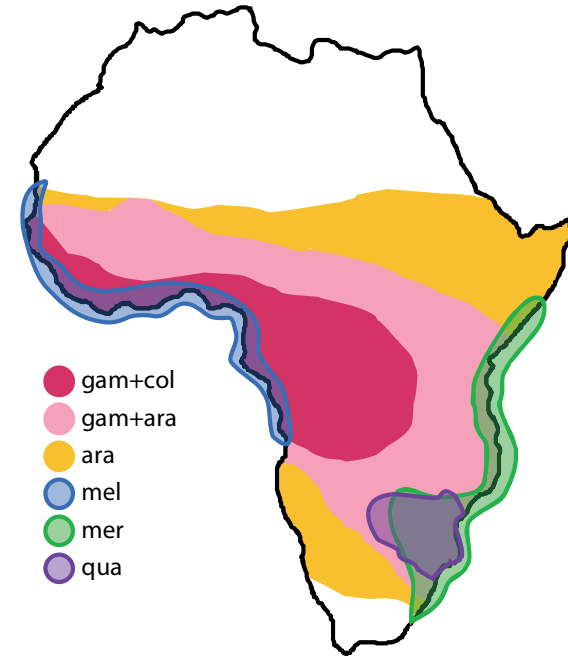
About 60 *anopheline* mosquito species can serve as vectors for five species of *Plasmodium* that produce illness in many animal species.



# The *Anopheles gambiae* species complex

Before the 1940s, there was one species *A. gambiae* recognised. Now there are 8:

*A. gambiae* (G) & *A. coluzzii* (C), *A. arabiensis* (A), *A. merus* (R), *A. melas* (L) & *A. bwambae*, and *A. quadriannulatus* (Q) & *A. amharicus*.



- *A. gambiae* and *A. coluzzii* are major malaria carriers, while *An. arabiensis* is a lesser vector.
- *A. melas* and *A. merus* are salt-tolerant, and have similar ecological and morphological characteristics, and are minor vectors.
- *A. quadriannulatus* bites animals and not humans.

# Data

Table 1: Number of loci in each chromosome region in noncoding and coding datasets.

Dataset	Chromosome region										Total
	2L1	2La	2L2	2R	3L1	3La	3L2	3R	Xag	X2	
Noncoding	4134	6732	2330	17027	2496	6280	1823	14323	1825	622	57592
Coding	2223	2776	1362	6849	983	1998	764	4977	1179	394	23505

Whole genome alignment from Fontaine *et al.* (2015).

We used twelve whole genomes for the six ingroup species: *A. gambiae* (G), *A. coluzzii* (C), *A. arabiensis* (A), *A. melas* (L), *A. merus* (R), and *A. quadriannulatus* (Q), and *A. christyi* (O) as outgroup.

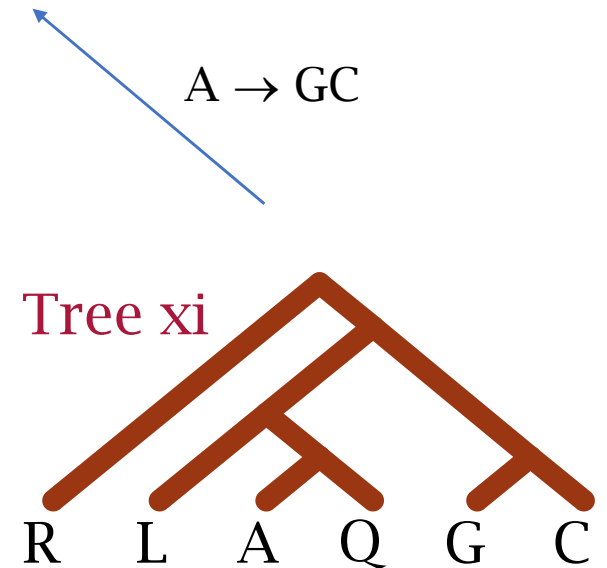
12 sequences per locus or 13 including outgroup.  
We compiled segments (loci) of 100-1000 bp, with a gap of >2 kb.

A  $\rightarrow$  GC introgression in tree xi leads to tree ii.



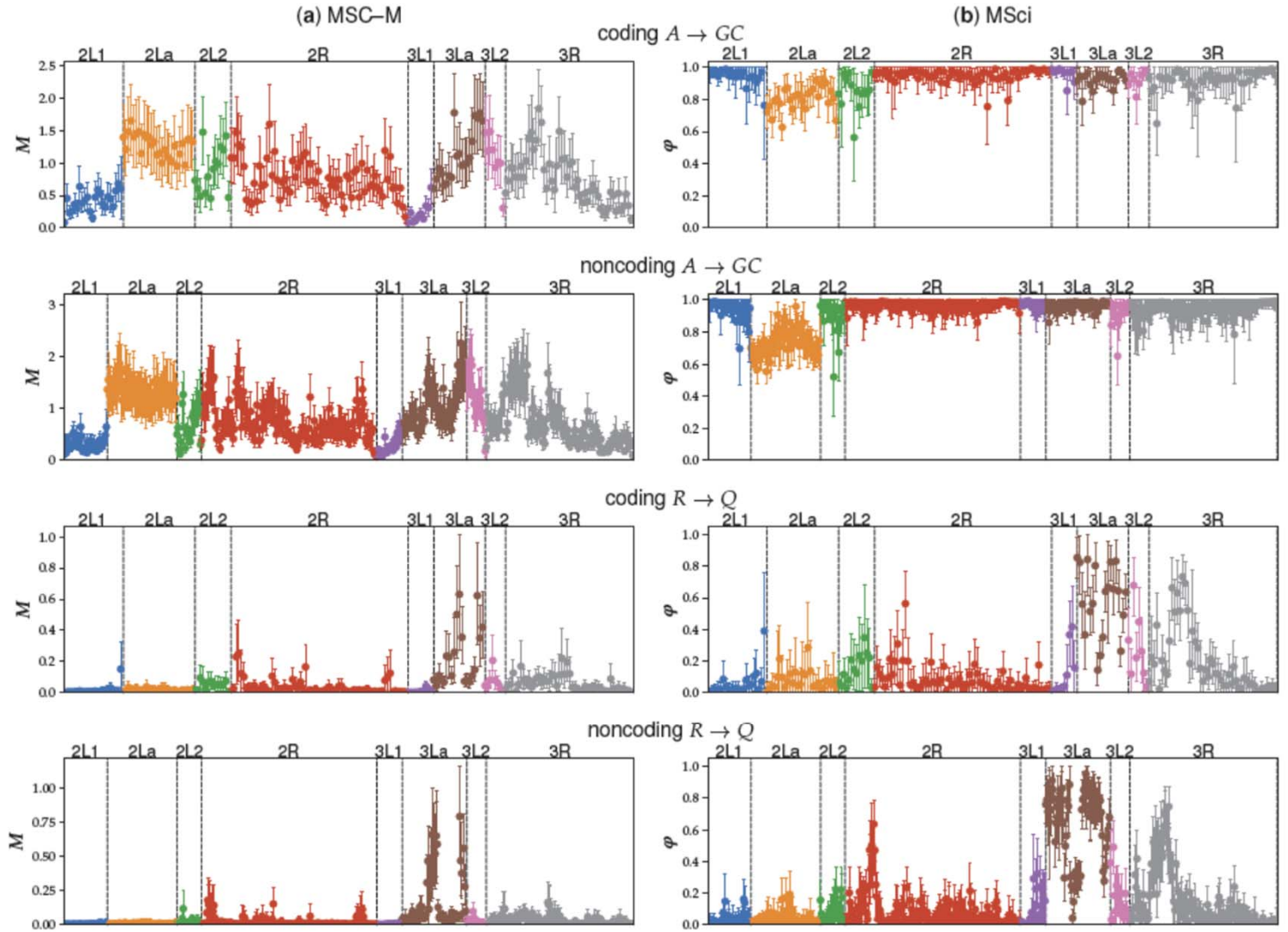
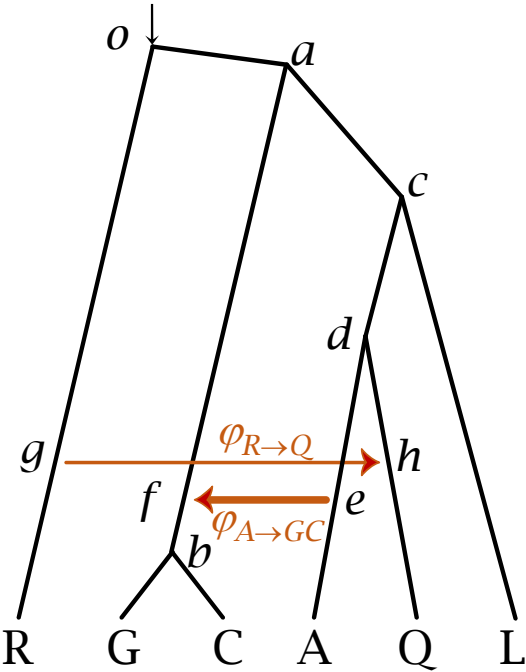
The X chromosome tree (**tree xi**) is the true species tree.

The autosomes tree (**tree ii**) is the result of **tree xi** and A  $\rightarrow$  GC gene flow.



Thawornwattana Y, et al. 2018. Mol Biol Evol 35:2512-2527.

*A. anopheles*:  
The rate of gene flow ( $\varphi$  and  $M$ ) varies across the genome



**Fig. S8.** (a) **MSC-M.** Posterior means and 95% HPD CIs of migration rates,  $M_{A \rightarrow GC}$  and  $M_{R \rightarrow Q}$  (fig. 6a), obtained from BPP analysis of the 100-loci blocks. (b) **MSC-I.** Introgression probabilities ( $\varphi_{A \rightarrow GC}$ ,  $\varphi_{R \rightarrow Q}$  in the MSC-I model, fig. 6b) under the MSC-I model. The MSC-I results are very similar to those of ref. (10, fig. 6), where inverse gamma priors were used for  $\tau$  and  $\theta$ . Here we used gamma priors, and assumed the same population size before and after each introgression event ( $\theta_R = \theta_g$ ,  $\theta_b = \theta_f$ , etc.; fig. 6b).

## Introgression between *A. Gambiae* and *A. arabiensis*

Slotman et al. (2005) used microsatellite markers to trace introgressed chromosomes from A → G.

- Introgressed X chromosomes were removed within two generations.
- After introgression from A into G, most introgressed alleles at third chromosome markers, particularly those on 3R, decreased steadily, indicating selection against them.
- Frequency of introgressed alleles on 2L were close to the original frequency even after 19 generations, whereas only two 2R markers showed a modest decrease.
- Attempts for G → A introgression were not successful.

Slotman, M. A., Della Torre, A., Calzetta, M., and Powell, J. R. 2005. Differential introgression of chromosomal regions between *Anopheles gambiae* and *An. arabiensis*. *Am. J. Trop. Med. Hyg.* 73(2): 326-335.

# A Martian theory of human evolution to illustrate the impact of gene flow on the concept of species...

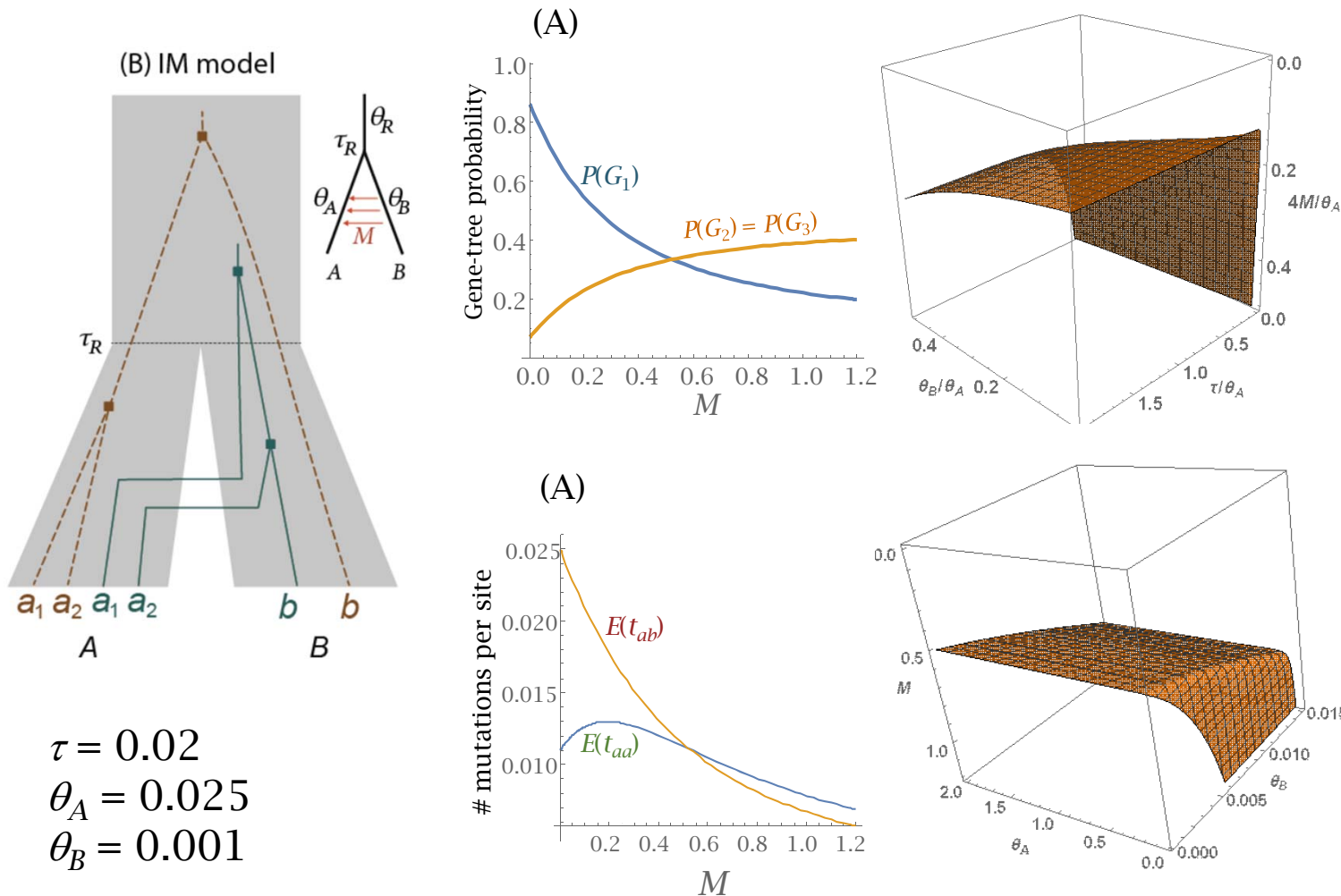


Jiao X, Yang Z. 2021. Defining species when there is gene flow. *Systematic Biology* **70**:108–119.



# Impact of gene flow on species definition

Small amount of gene flow ( $Nm < 1$  per generation) can have a drastic impact.



(1) Inside red tent,  
 $\Pr\{G_1\} < \Pr\{G_2\}$

$G_1 = ((a_1 a_2) b);$   
 $G_2 = ((a_1 b) a_2);$   
 $G_3 = ((a_2 b) a_1);$

(2) Inside red tent,  
 $E(t_{aa}) > E(t_{ab})$

Jiao X, Yang Z. 2021. Defining species when there is gene flow. *Systematic Biology* 70:108–119.

# A model of human/martian evolution

Suppose humans separated from martians  
1 myrs ago, and suppose

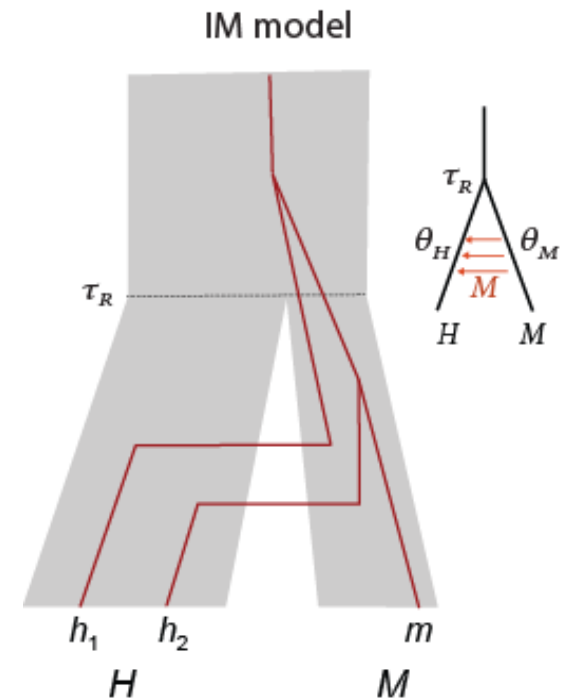
$$N_H = 10^6, N_M = 10^5,$$

$N_H m_{M \rightarrow H} \approx 0.8$  immigrants per generation

(These are not real estimates!)

Then we are all ‘genetically martian’ in that  
each of us is closer to a random martian  
than to another human:

- (i) Gene tree  $G_1 = ((h_1 h_2) m)$  is less probable  
than  $G_2 = ((h_1 m) h_2)$ ,
- (ii)  $E(t_{hh}) > E(t_{hm})$ .



# Implications?

- (i) DNA bar-coding. For example, the '10×' rule says that  $A$  and  $B$  are distinct species if the genetic distance (for *cytb* or *CO1*) is interspecific distance is 10× higher than the intraspecific distance.
- (ii) The genealogical divergence index (*gdi*) (Jackson et al. 2017) says that  $A$  and  $B$  are one species if  $gdi < 0.2$  or  $P_1 = \Pr\{G_1\} < 0.47$ . Here  $A$  and  $B$  can be two distinct species even if  $P_1 < 1/3$ .

Jackson ND, Carstens BC, Morales AE, O'Meara BC. 2017. Species delimitation with gene flow. *Systematic Biology* **66**:799-812.

# References

- Flouri T, Jiao X, Rannala B, Yang Z. 2018. Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Mol Biol Evol* 35:2585-2593.
- Flouri T, Rannala B, Yang Z. 2020. A tutorial on the use of BPP for species tree estimation and species delimitation. Pp. 5.6.1-16 in *Scornavacca C, Delsuc F, and Galtier N, eds. Phylogenetics in the Genomic Era*.
- Flouri T, Jiao X, Rannala B, Yang Z. 2020. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Mol Biol Evol* 37:1211-1223.
- Flouri T, Jiao X, Huang J, Rannala B, Yang Z. 2023. Efficient Bayesian inference under the multispecies coalescent with migration. *PNAS* 120:e2310708120.
- Jiao X, Yang Z. 2021. Defining species when there is gene flow. *Syst Biol* 70:108-119.
- Jiao X, Flouri T, Rannala B, Yang Z. 2020. The impact of cross-species gene flow on species tree estimation. *Syst Biol* 69:830-847.
- Thawornwattana Y, Dalquen DA, Yang Z. 2018. Coalescent analysis of phylogenomic data confidently resolves the species relationships in the *Anopheles gambiae* species complex. *Mol Biol Evol* 35:2512-2527.
- Zhu T, Yang Z. 2021. Complexity of the simplest species tree problem. *Mol Biol Evol* 10.1093/molbev/msab009

# Coalescence, introgression, and inversions cause a lot of challenges

- Deep coalescence (incomplete lineage sorting) due to radiative speciations or short branches in the species tree
- Introgresssion beween species & differential selection against introgressed alleles and chromosomes
- Chromosomal inversions
- Different chromosomes or genomic regions have different histories.
- Different methods produce different trees.
- Inversions & sequences produce different trees.