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

Ziheng Yang

Department of Genetics, Evolution, and Environment

University College London

Outline

- Bayesian implementation of introgression (MSC-I) and migration (MSC-M) models
- Bayesian test of gene flow
- Heuristic methods for inferring gene flow
- Impact of gene flow

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 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

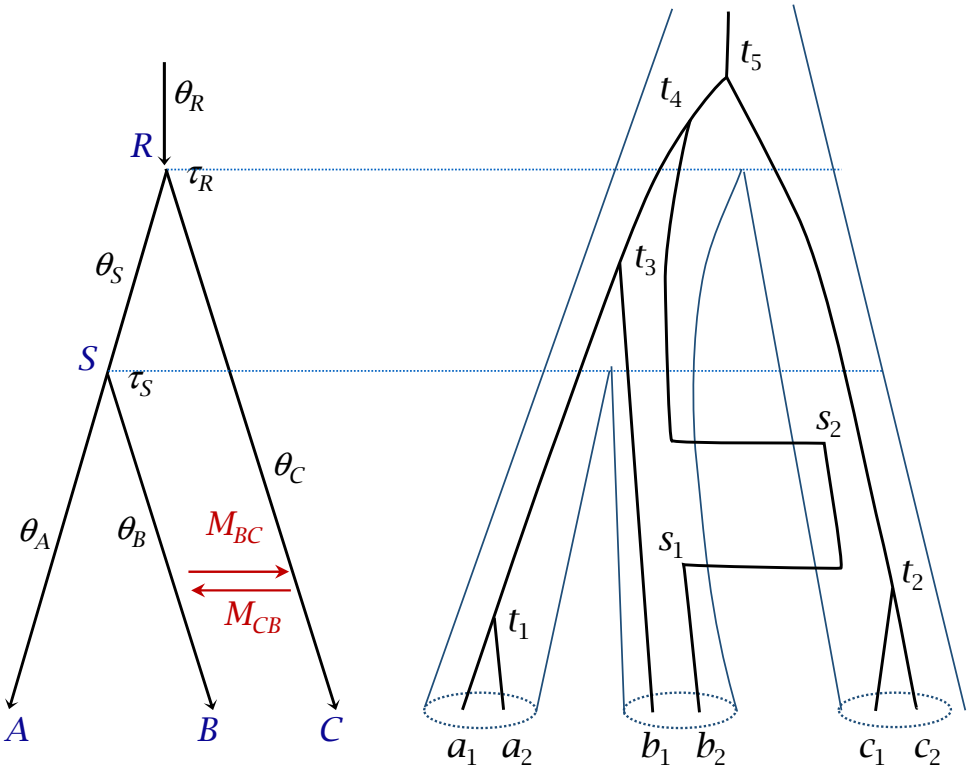
etc.

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

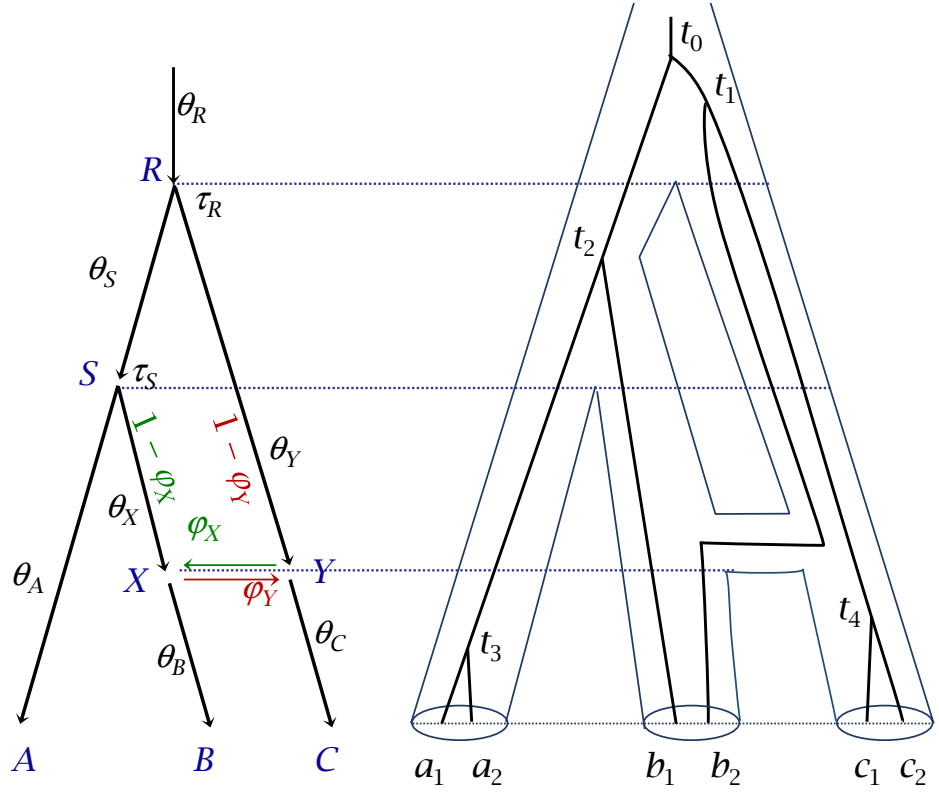
Some terminologies are confusing:

"*to distinguish hybridization from lineage sorting*"
"*investigate whether the conditions of applicability of coalescence-based methods are met …*"

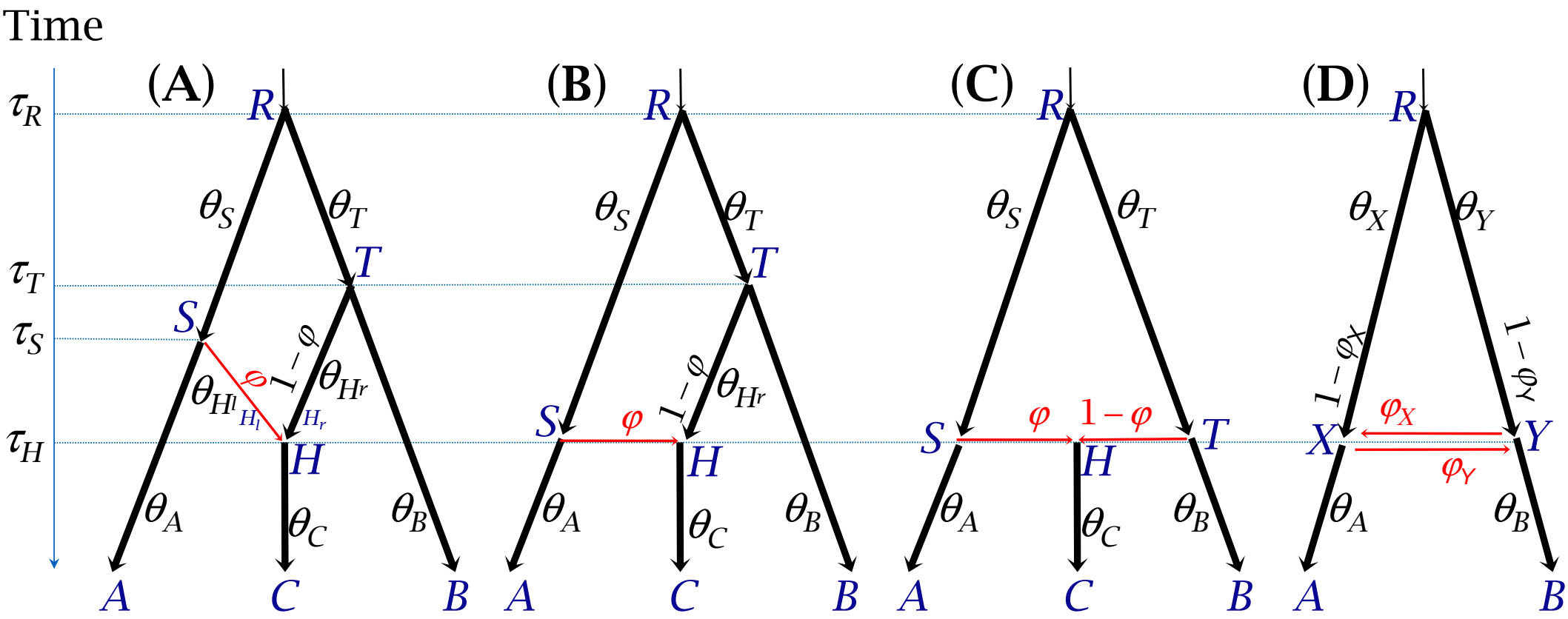
MSC-M (migration)



MSC-I (introgression)

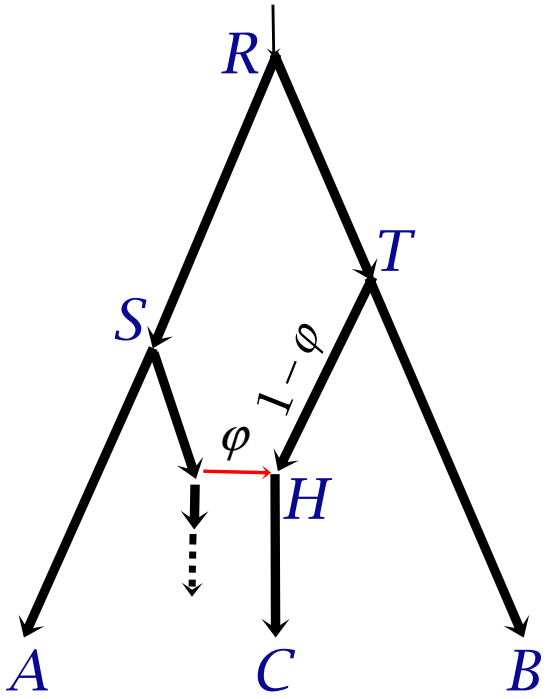


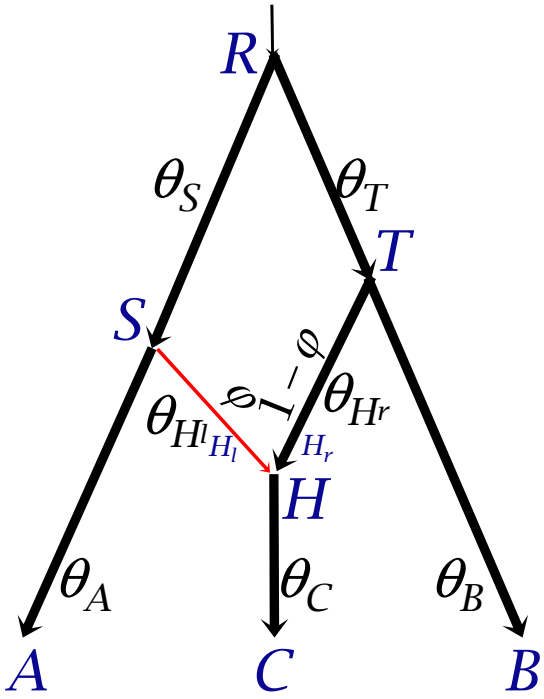
Different types of MSC-I models in BPP



$\gamma \equiv \varphi$

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





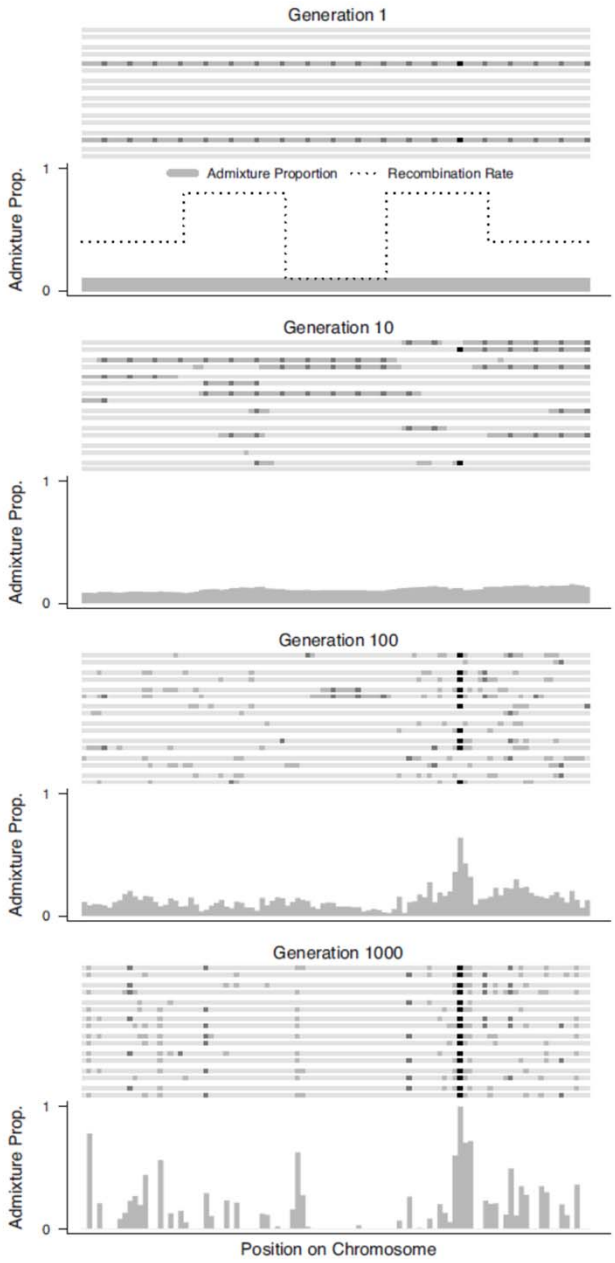
$$M_{AB} = N_B m_{AB}$$

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 (φ 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.

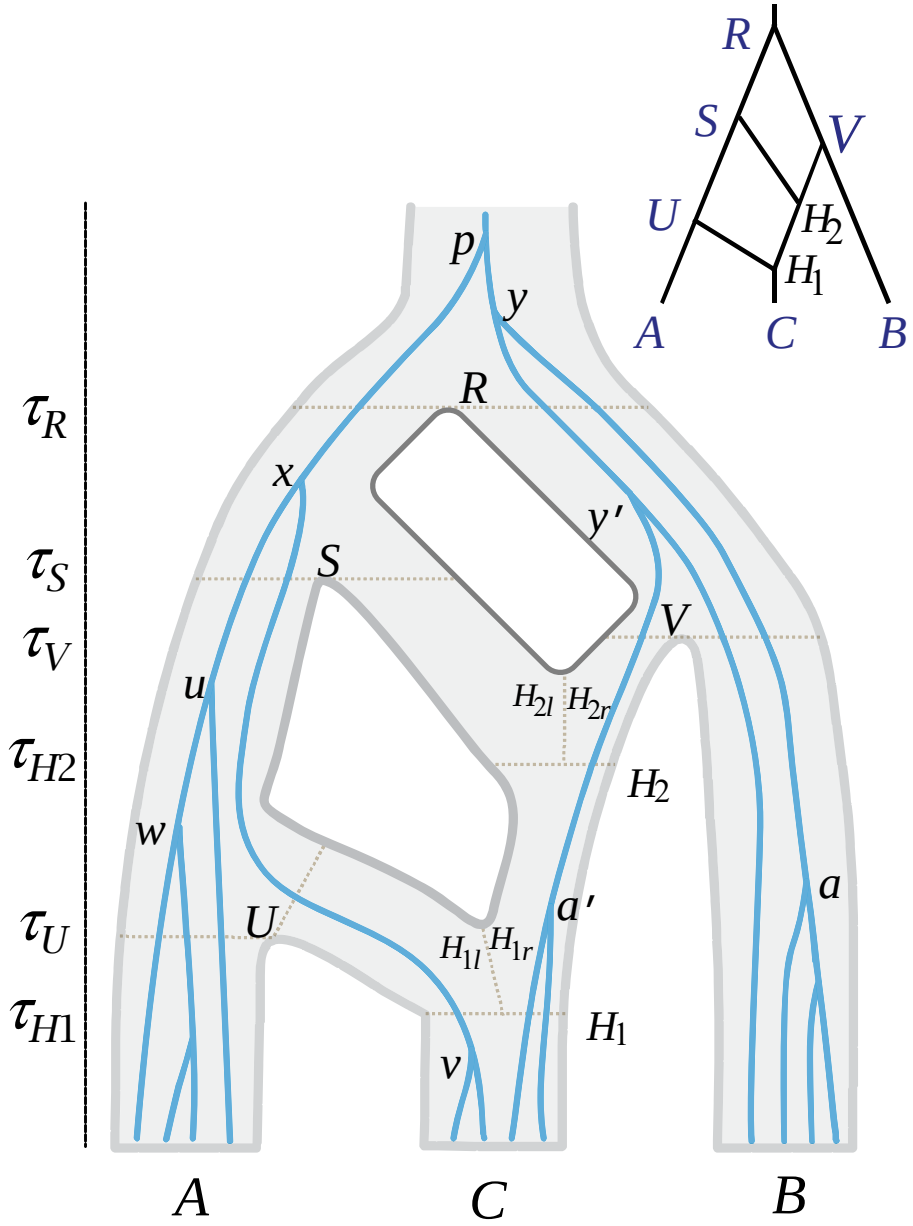


Martin SH, Jiggins CD. 2017. Interpreting the genomic landscape of introgression. *Curr Opin Genet Dev* **47**: 69-74.

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 $\{\theta, \tau, \varphi\}$, $\{G_i, \boldsymbol{t}_i\}$.
2. Iterate
   - change gene-tree node age.
   - change gene-tree topology (SPR).
   - change parameters (θs, τs, φs).
   - 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.

Bayesian test of gene flow using Bayes factor

Null model H_0 (no gene flow): $\varphi = 0$

Alternative model H_1 (gene flow): $\varphi > 0$

The Bayes factor is the ratio of marginal likelihood values for the two models.

$$BF_{10} = \frac{M_1}{M_0}$$

$BF_{10} > 100$ means strong rejection of H_0.

$BF_{10} < 0.01$ means strong rejection of H_1.

The marginal likelihood can be calculated using algorithms such as thermodynamic integration and stepping-stones.

Lartillot N, Philippe H. 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55:195-207.
Rannala B, Yang Z. 2017. Efficient Bayesian species tree inference under the multispecies coalescent. Syst. Biol. 66:823-842.

Bayesian test of introgression using Savage-Dickey density ratio

$$BF_{10} = \frac{\pi(\varphi=0)}{\pi(\varphi=0\,|X)} \approx \frac{\Pr(\varphi < \varepsilon)}{\Pr(\varphi < \varepsilon|X)}$$

Example. Suppose we use $\varepsilon = 0.01$, so that $\varphi < 0.01$ means "no gene flow".

Suppose the prior probability is $\Pr(\varphi < 0.01) = 1\%$.
The posterior probability $\Pr(\varphi < 0.01|X) = 0.008\%$.
Then

$$BF_{10} \approx \frac{1\%}{0.008\%} = 125 > 100$$

We use a cutoff of 100, similar to 1% significance in hypothesis testing.
Thus we have strong evidence against the null hypothesis of no gene flow.

Process the mcmc sample file to calculate $\Pr(\varphi < \varepsilon|X)$.
The prior probability $\Pr(\varphi < \varepsilon)$ is typically available analytically.

`mcmc.txt` file

Gen	phi_x
10	0.019626
20	0.011088
30	0.011088
40	0.017072
50	0.021010
60	0.021010
70	0.018433
...	

Ji J, Jackson DJ, Leache AD, Yang Z. 2023. Power of Bayesian and heuristic tests to detect cross-species introgression with reference to gene flow in the *Tamias quadrivittatus* group of North American chipmunks. *Syst. Biol.* 72:446-465.

Bayesian test of introgression using Savage-Dickey density ratio

$$BF_{10} = \frac{\pi(\varphi=0)}{\pi(\varphi=0\,|X)} \approx \frac{\Pr(\varphi < \varepsilon)}{\Pr(\varphi < \varepsilon|X)}$$

The approach based on the Savage-Dickey density ratio works only if the null and alternative hypotheses are nested.
Under MSC-I: H_0: $\varphi = 0$ against H_1: $\varphi > 0$.
Under MSC-M: H_0: $M = Nm = 0$ against H_1: $M > 0$.

Note that Bayes factor may lead to strong rejection of the alternative model H_1.

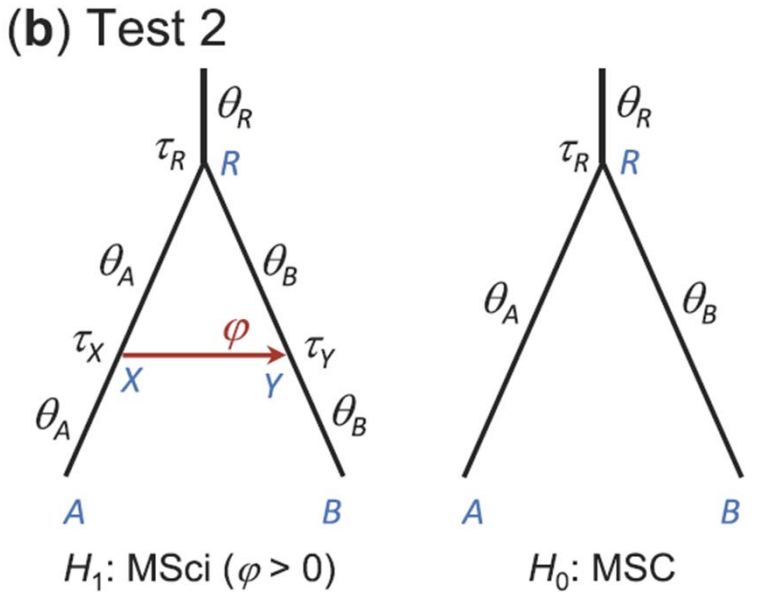


FIGURE 3. Parameters in the alternative and null hypotheses in two Bayesian tests of introgression (i.e., test of $H_0 : \varphi = 0$ against $H_1 : \varphi > 0$).

Ji *et al.* 2023. *Syst. Biol.* 72:446-465.

Mosquitoes

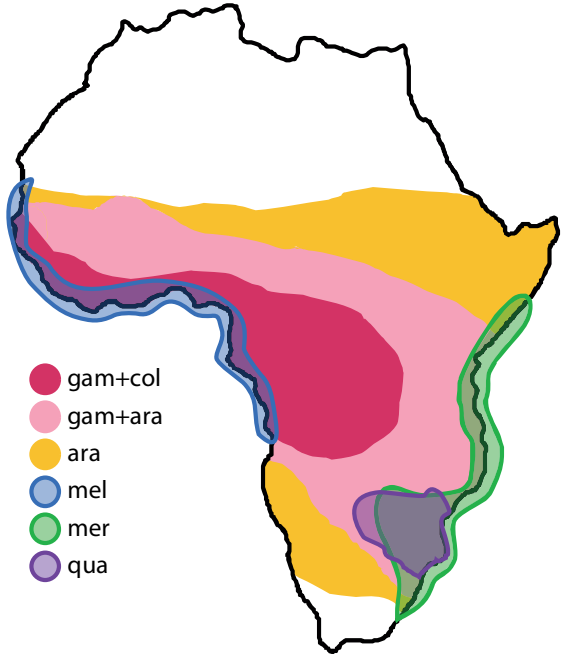
Mosquitoes infested by *Plasmodium* parasites bite humans, and humans get malaria.

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.

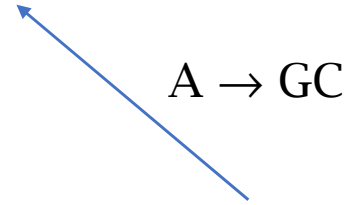
Dataset	Chromosome region										Total
	2L1	2La	2L2	2R	3L1	3La	3L2	3R	Xag	X2	
Noncoding	4134	6732	2330	17027	2496	6280	1823	14323	1825	622	57592
Coding	2223	2776	1362	6849	983	1998	764	4977	1179	394	23505

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.

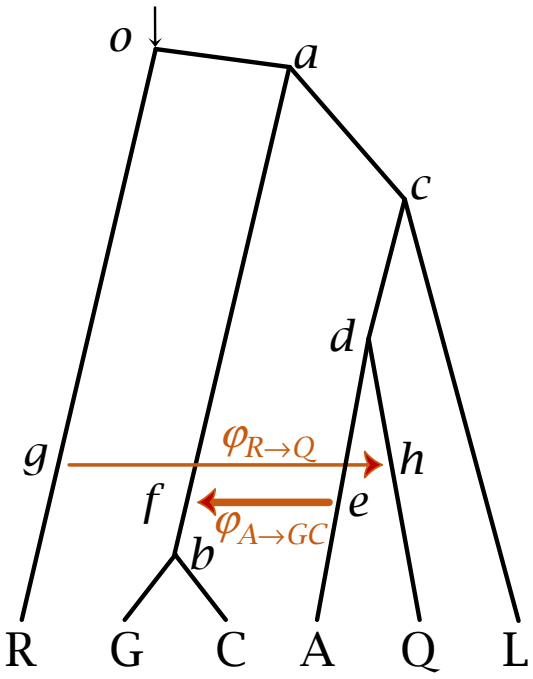
A → GC introgression in tree xi leads to tree ii.



The X chromosome tree (tree xi) is the true species tree.
The autosomes tree (tree ii) is the result of tree xi and A → GC gene flow.

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

Anopheles:
The rate of gene flow (φ and M) varies across the genome



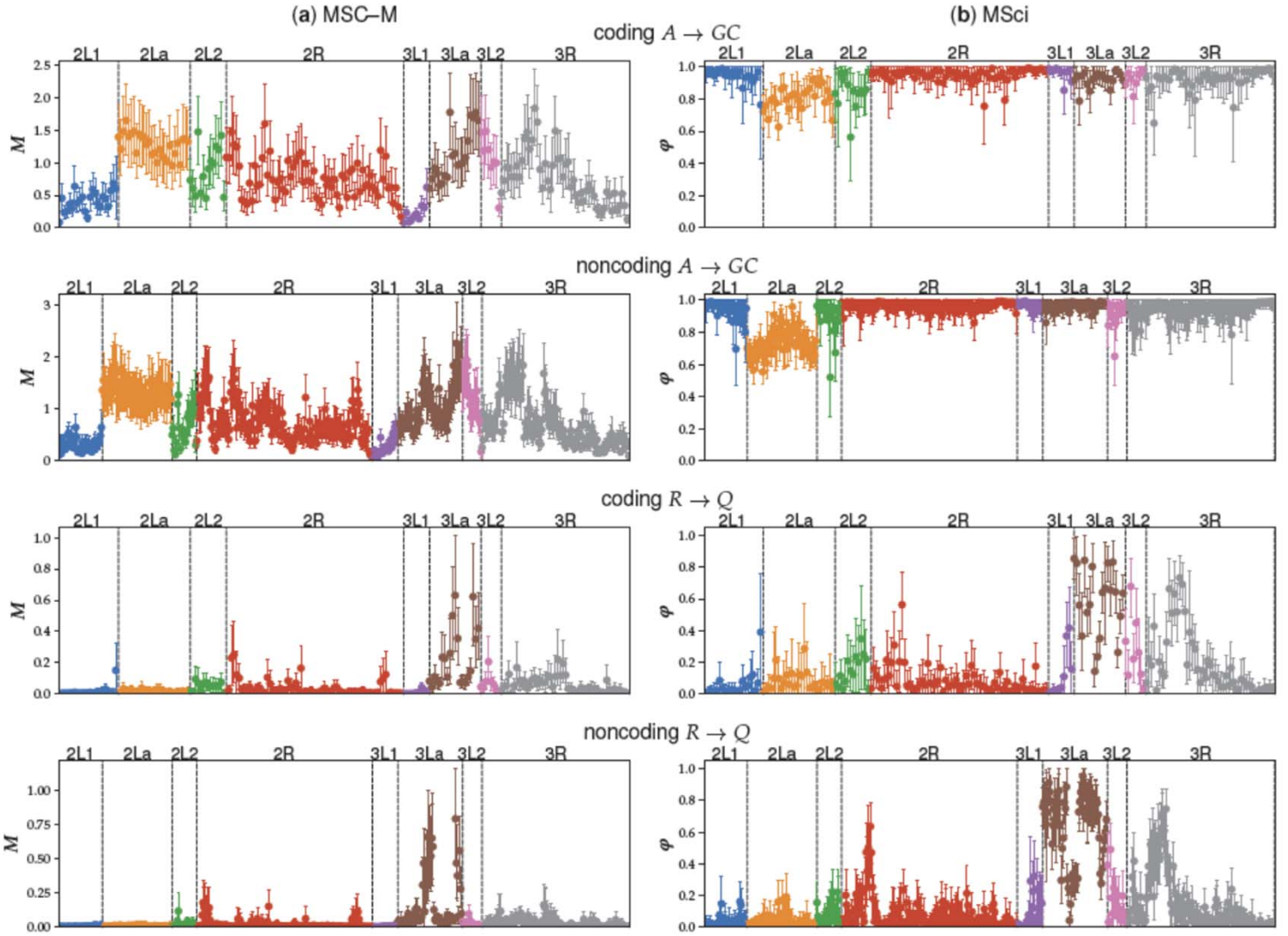
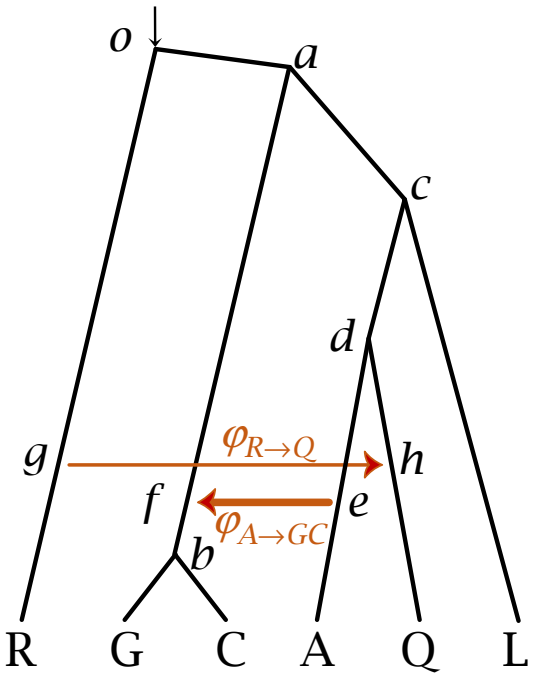


Fig. S8. (a) **MSC-M.** Posterior means and 95% HPD CIs of migration rates, $M_{A\to GC}$ and $M_{R\to Q}$ (fig. 6a), obtained from BPP analysis of the 100-loci blocks. (b) **MSC-I.** Introgression probabilities ($\varphi_{A\to GC}$, $\varphi_{R\to Q}$ in the MSC-I model, fig. 6b) under the MSC-I model. The MSC-I results are very similar to those of ref. (10, fig. 6), where inverse gamma priors were used for τ and θ. 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).

Flouri et al. 2023 PNAS 120:e2310708120

Anopheles: Test of gene flow.
A→GC throughout the genome, R→Q gene flow mostly on 3L and 3R



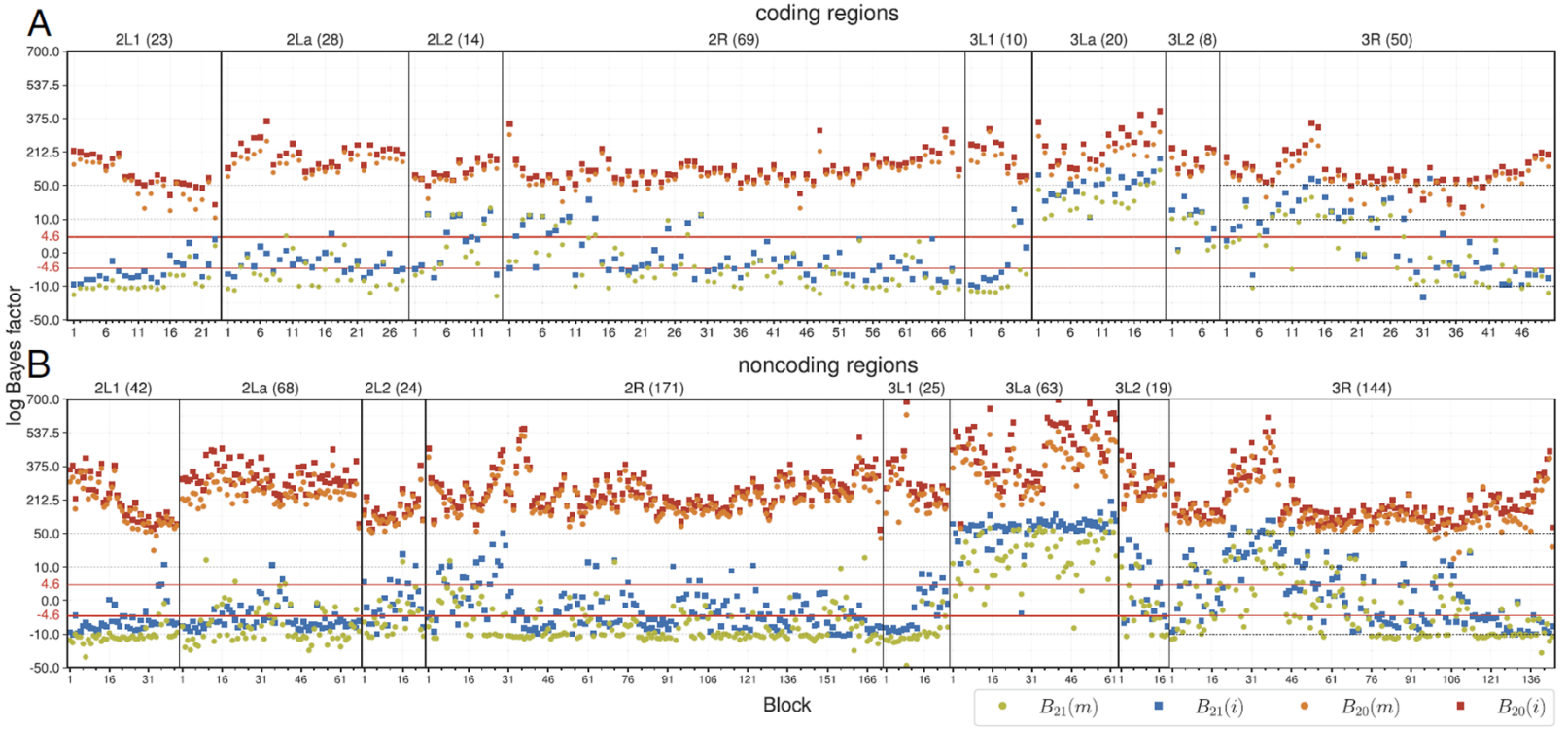


Fig. 7. The logarithm of the Bayes factor for testing gene flow obtained from BPP analysis of the 100-loci blocks of the (*A*) coding and (*B*) noncoding data from

Flouri et al. 2023 PNAS 120:e2310708120

If $B_{10} > 100$ (if $\log B_{10} > 4.6$), reject H_0 of no gene flow.
If $B_{10} < 0.01$ (if $\log B_{10} < -4.6$), reject H_1 of gene flow.

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 chromsomal regions between *Anopheles gambiae* and *An. arabiensis*. *Am. J. Trop. Med. Hyg.* 73(2): 326-335.

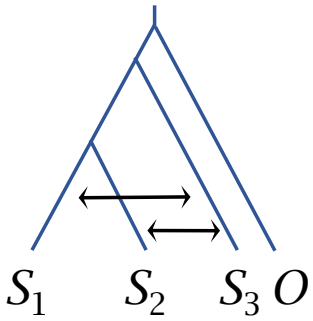
Heuristic/summary methods for inferring introgresison

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

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.

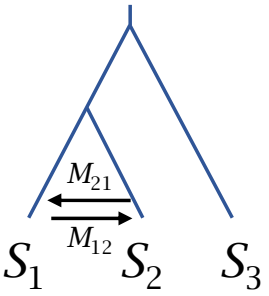
There are a number of heuristic methods.
ABBA-BABA test (*D* statistic) vs. bpp/3s



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

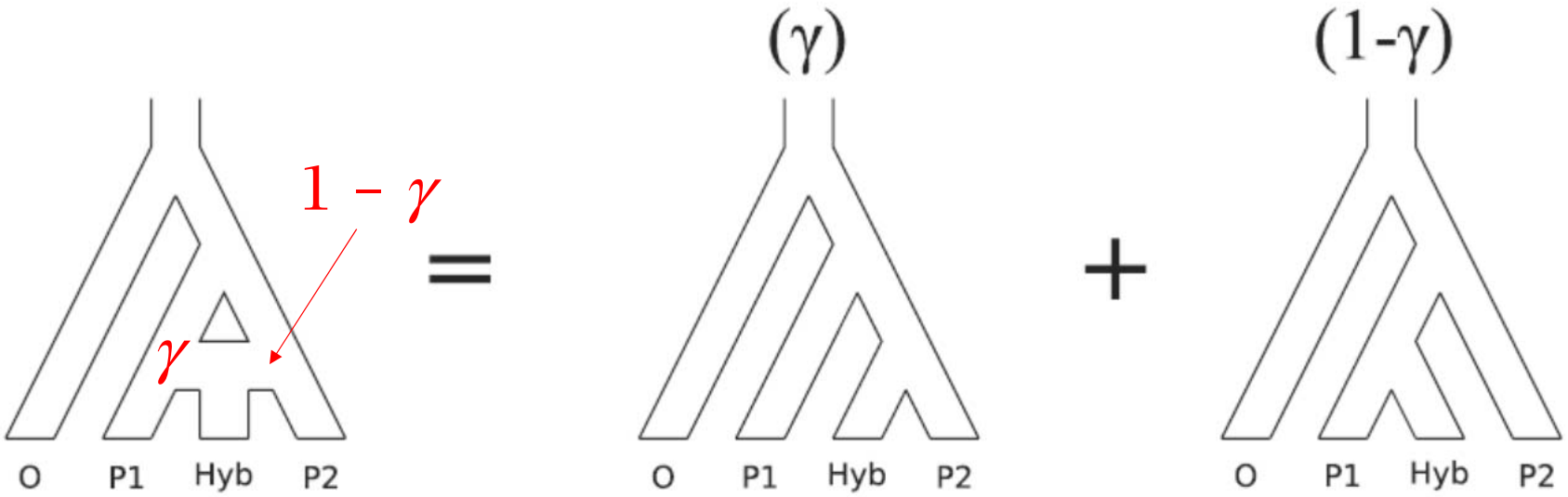
1. Uses genome-wide counts (averages).
2. Tests for gene flow between S_1 and S_3 or between S_2 and S_3, on a fixed species tree ((S_1, S_2), S_3), O).
3. Can't identify the direction of gene flow.
4. Information in the different gene genealogies across the genome is ignored.
5. Summary statistic lacks power.

3s (likelihood ratio test)



1. Uses short widely dispersed segments (loci) from the genome.
2. Tests for gene flow between S_1 and S_2, on a fixed species tree ((S_1, S_2), S_3).
3. Can estimate $M_{12} = N_2 m_{12}$ and $M_{21} = N_1 m_{21}$.
4. Information in the different gene genealogies across the genome is ignored.
5. LRT in theory uses all information in the data.

HyDe



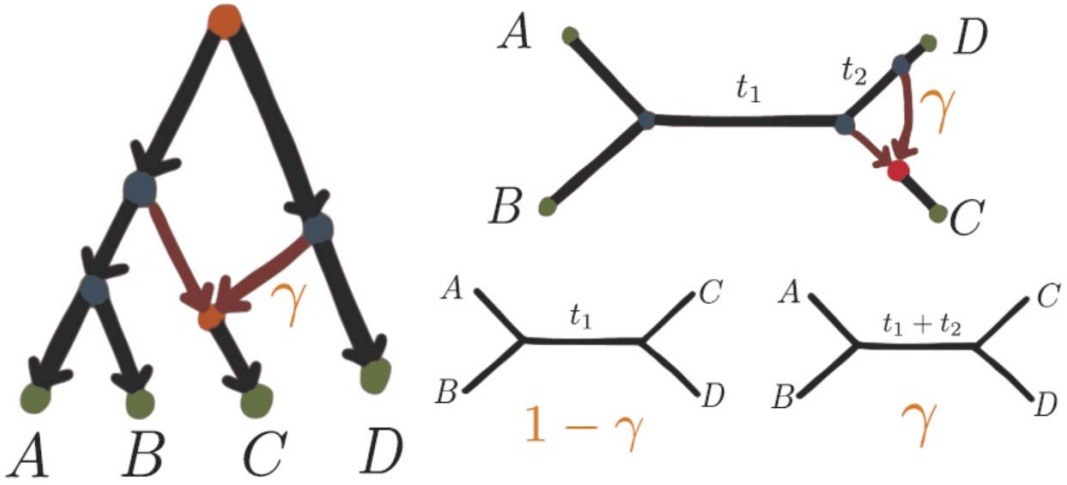
$$\frac{p_{xxyy} - p_{xyxy}}{p_{xyyx} - p_{xyxy}} = \frac{\gamma}{1 - \gamma}$$

$\gamma = \varphi$

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 (γ).



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

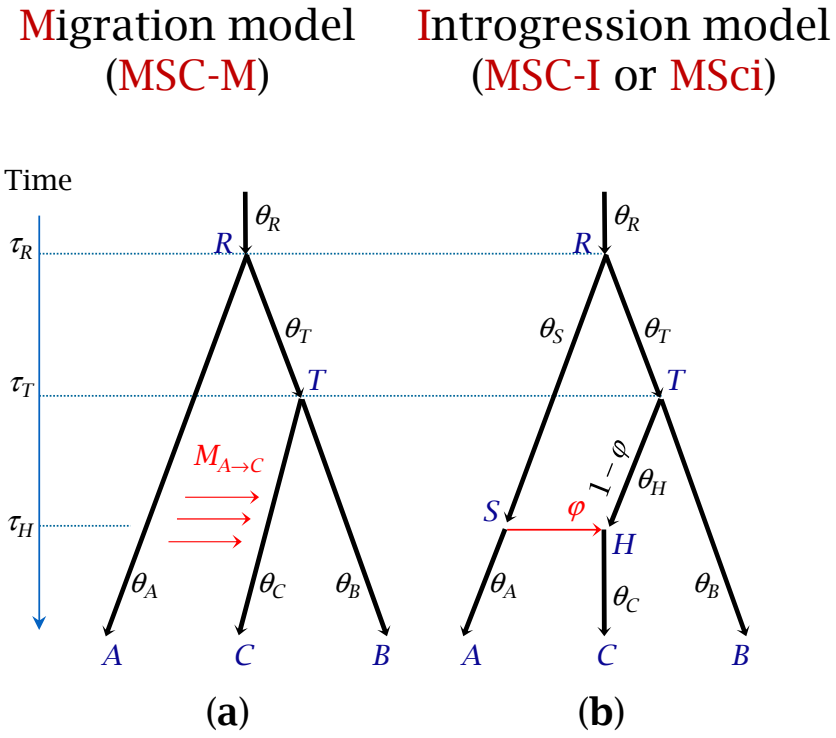
Features of summary methods

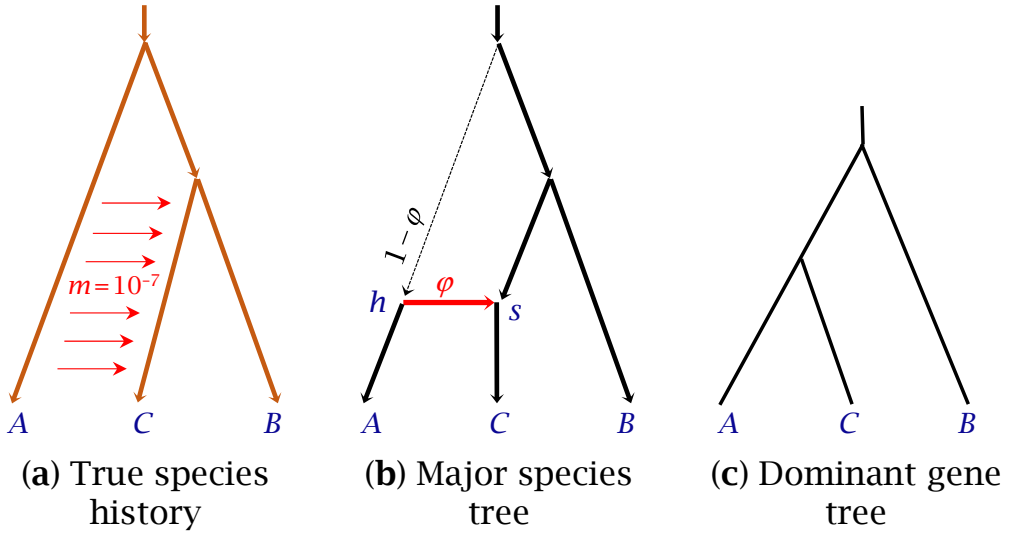
- They usually work for three species (under the clock) or four species (3+O).
- They can identify/estimate the introgression proportion and the internal branch lengths on the species tree (in coalescent units), but not other parameters in the model.
- They can't identify gene flow between sister lineages.

Impact of gene flow

Impact of gene flow on species tree estimation

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'.

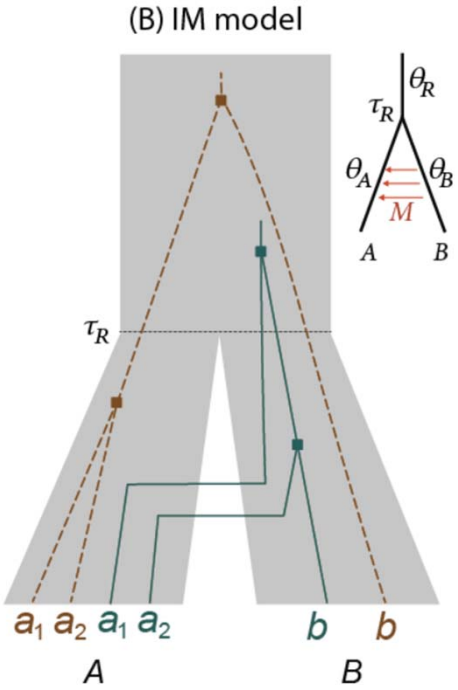




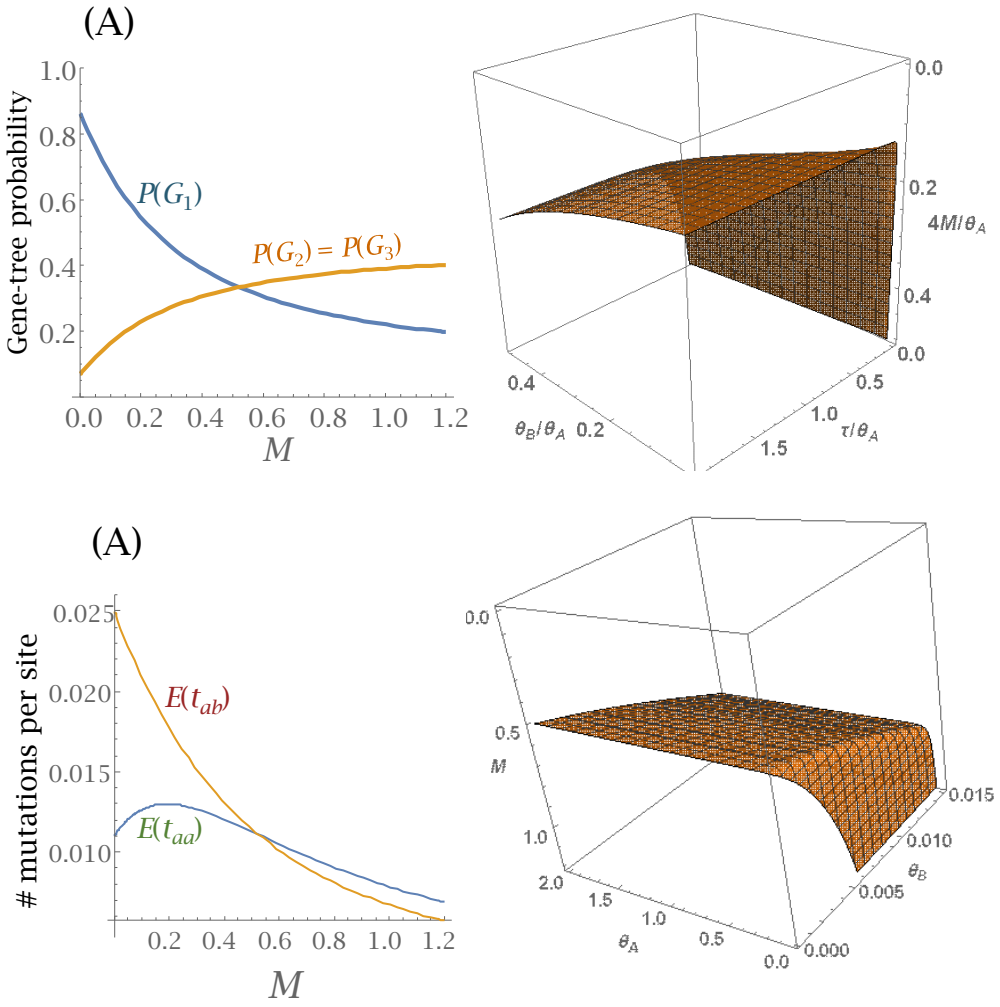
Jiao X, Flouri T, Rannala B, Yang Z. 2020. The impact of cross-species gene flow on species tree estimation. Syst Biol. 69: 830-847, 10.1093/sysbio/syaa001

Impact of gene flow on species definition

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



$\tau = 0.02$
$\theta_A = 0.025$
$\theta_B = 0.001$



(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

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

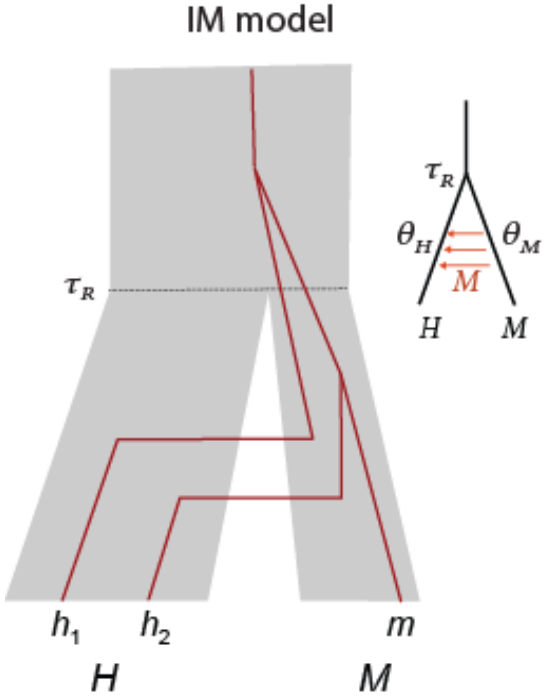
$N_H m_{M \to H} \approx 0.8$ immigrants per generation

(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.

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.

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