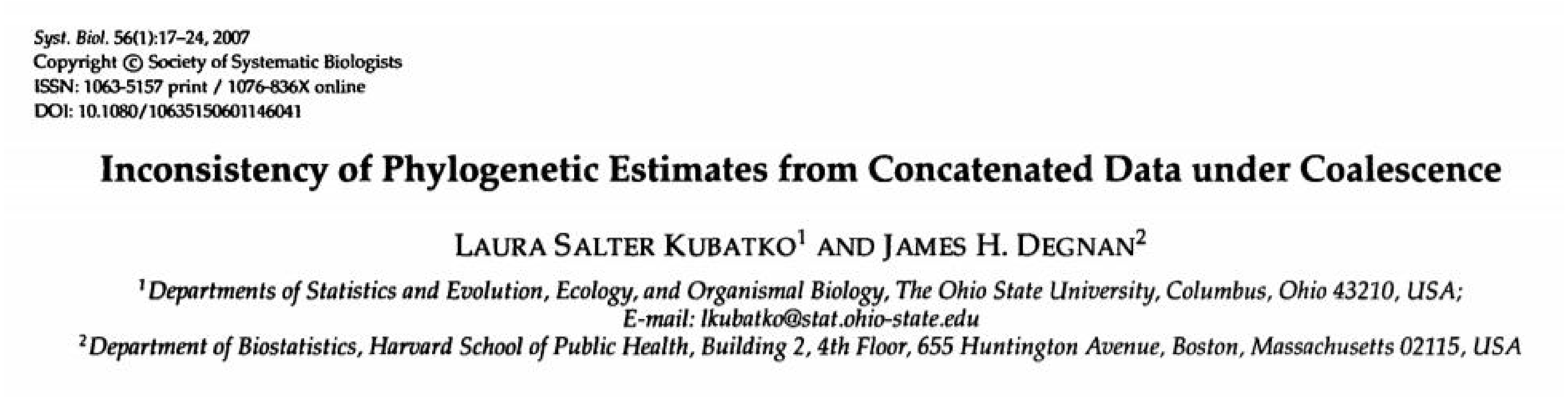
Species tree estimation & the multispecies coalescent

Back to biology - Are JC69's assumptions realistic?

Assumptions of JC69:

1. All substitutions equally likely
2. Base frequencies equal
3. Every site has equal probability of substitution
4. Process is constant through time
5. Sites are independent of each other
6. Substitution is Markovian (memoryless)
7. All sites have the same evolutionary history

Concatenated gene sequences - assumes every gene has same evolutionary history



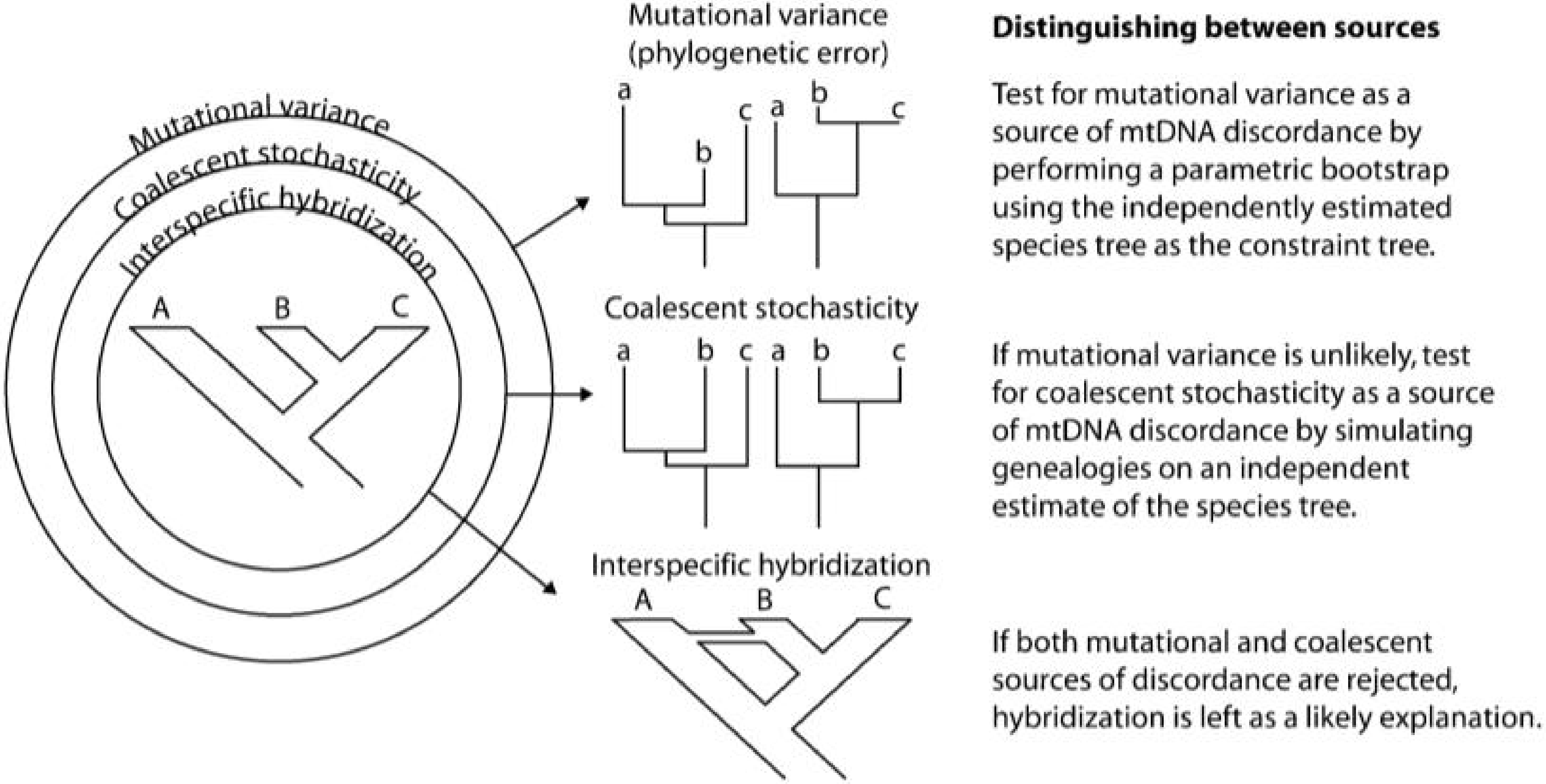
Syst. Biol. 56(1):17–24, 2007
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150601146041

Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence

LAURA SALTER KUBATKO[1] AND JAMES H. DEGNAN[2]

[1] *Departments of Statistics and Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, Ohio 43210, USA;*
E-mail: lkubatko@stat.ohio-state.edu
[2] *Department of Biostatistics, Harvard School of Public Health, Building 2, 4th Floor, 655 Huntington Avenue, Boston, Massachusetts 02115, USA*

Stochasticity from:



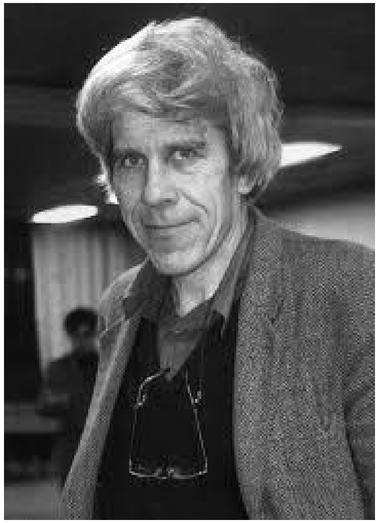
Distinguishing between sources

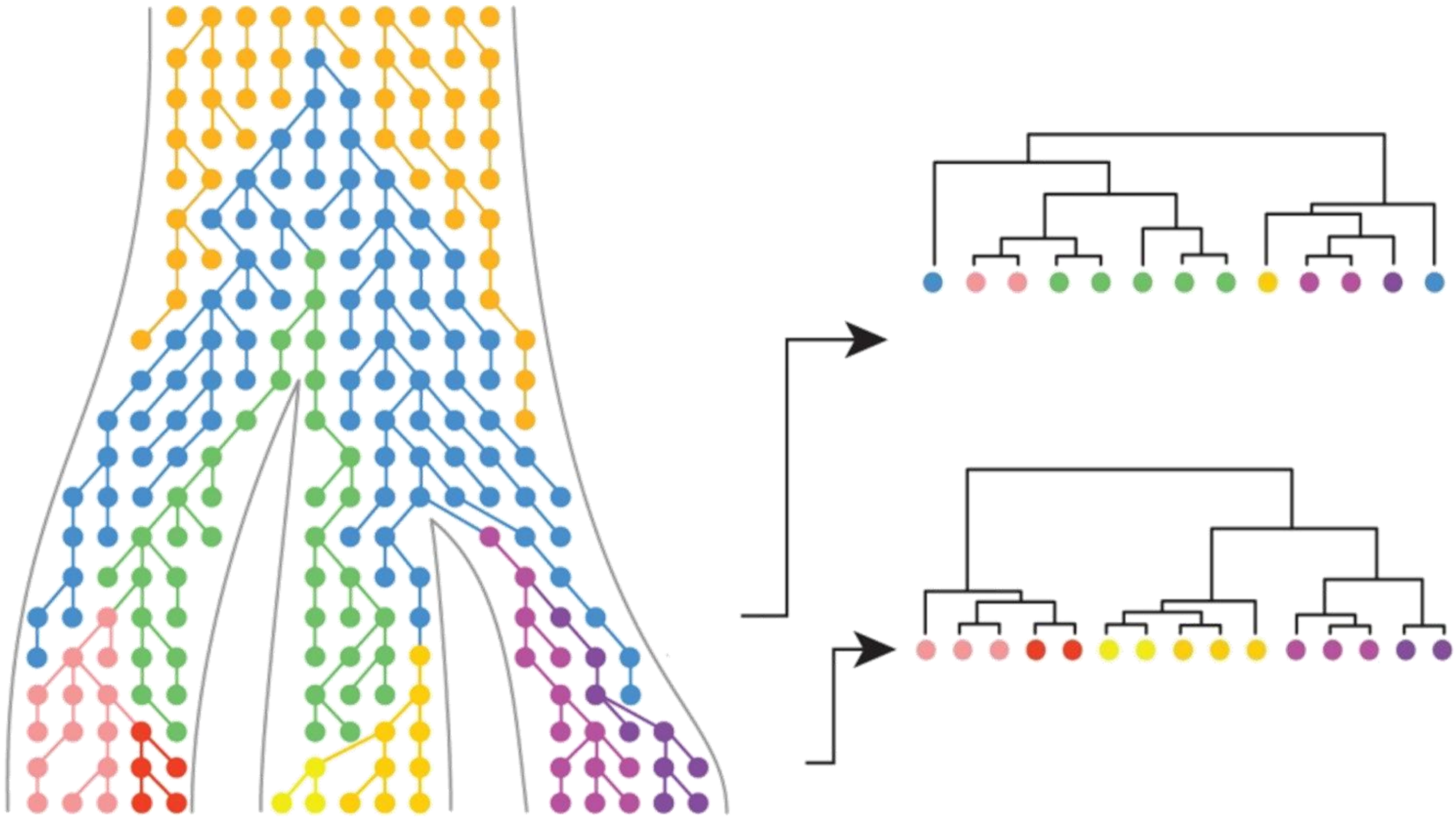
Test for mutational variance as a source of mtDNA discordance by performing a parametric bootstrap using the independently estimated species tree as the constraint tree.

If mutational variance is unlikely, test for coalescent stochasticity as a source of mtDNA discordance by simulating genealogies on an independent estimate of the species tree.

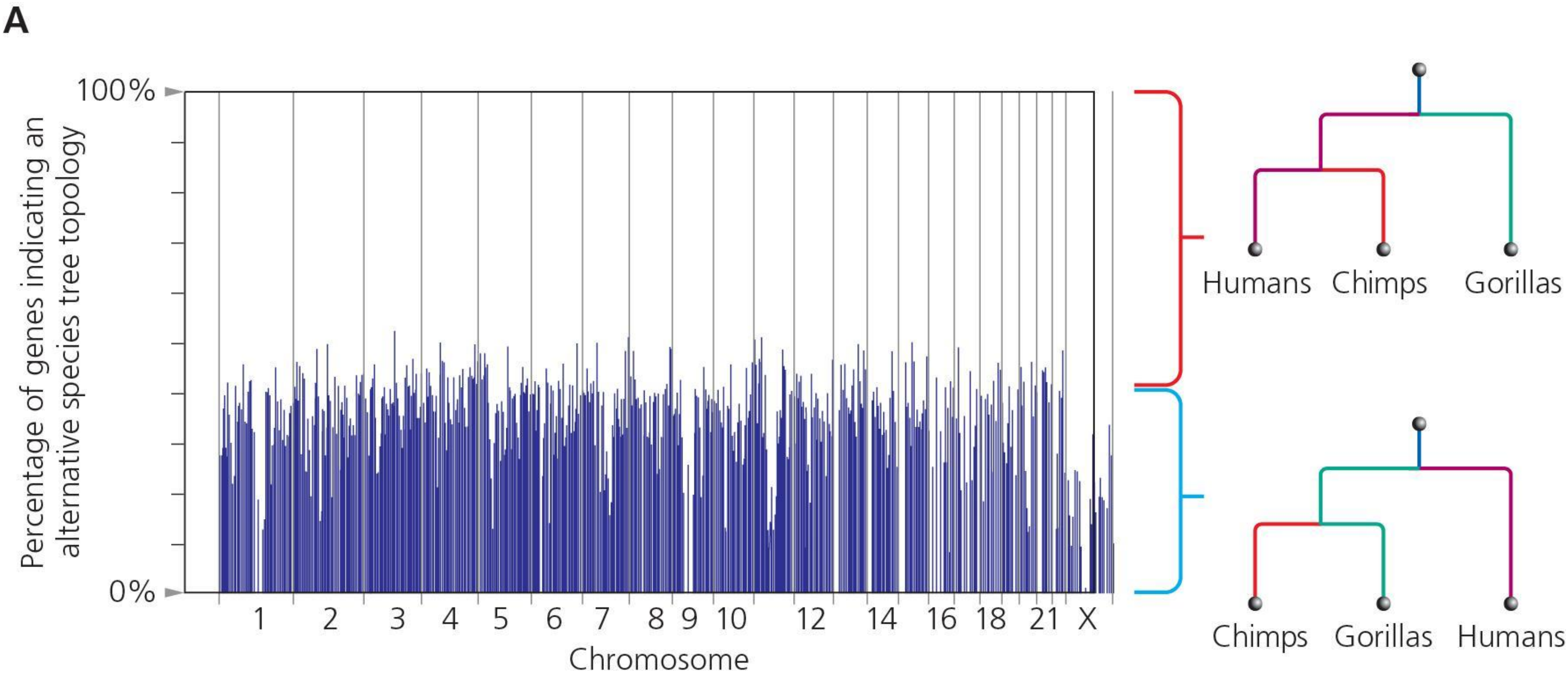
If both mutational and coalescent sources of discordance are rejected, hybridization is left as a likely explanation.

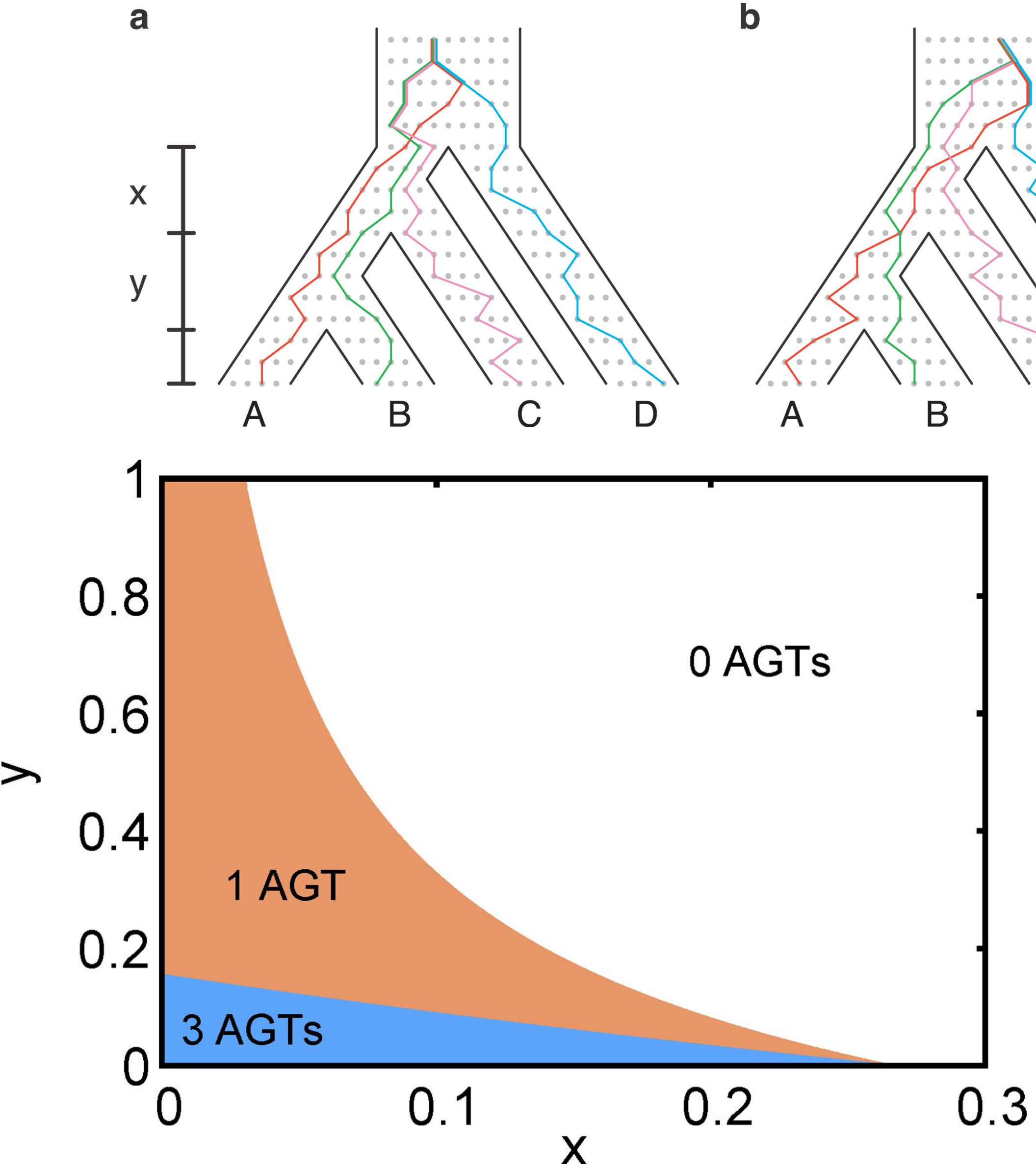
"As I write these words, even so as to be able to write them, I am pretending to a unity that, deep inside myself, I now know does not exist. I am fundamentally mixed, male with female, parent with offspring, warring segments of chromosome that interlocked in strife millions of years before Europe existed or saw any of the human violence that became later, for sure, embedded in my ancestry." - WD Hamilton





Incomplete lineage sorting


A
Percentage of genes indicating an alternative species tree topology
100%
0%
1
2
3
4
5
6
7
8
9
10
12
14
16
18
21
X
Chromosome
Humans
Chimps
Gorillas
Chimps
Gorillas
Humans



The “Anomaly zone”

AGT - Anomalous gene tree with higher likelihood than the tree species topology

ILS will result when branches are short & population sizes are large

(often mistakenly thought only to occur in recent radiations, but ancient short internal branches just as at risk!)

The Coalescent

Coalescence - MRCA of a pair of genes

Built on standard population genetics (e.g. Wright-Fisher model)

Key parameter - Effective population size
(May be very different from “census population size)

Expected # of generations to coalescence - 2Ne

The Coalescent

What is the probability two copies of a gene in a randomly breeding population have an ancestor 1 generation ago?

$$1/(2Ne)$$

What about j generations ago?

$$(1-1/(2Ne))^{j} * 1/(2Ne)$$

Genetic drift & the molecular clock

The chance of someone's copy of an allele drifting to fixation is:

$1/(2N_e)$

Let *v* be the rate of neutral mutations per allele, per generation

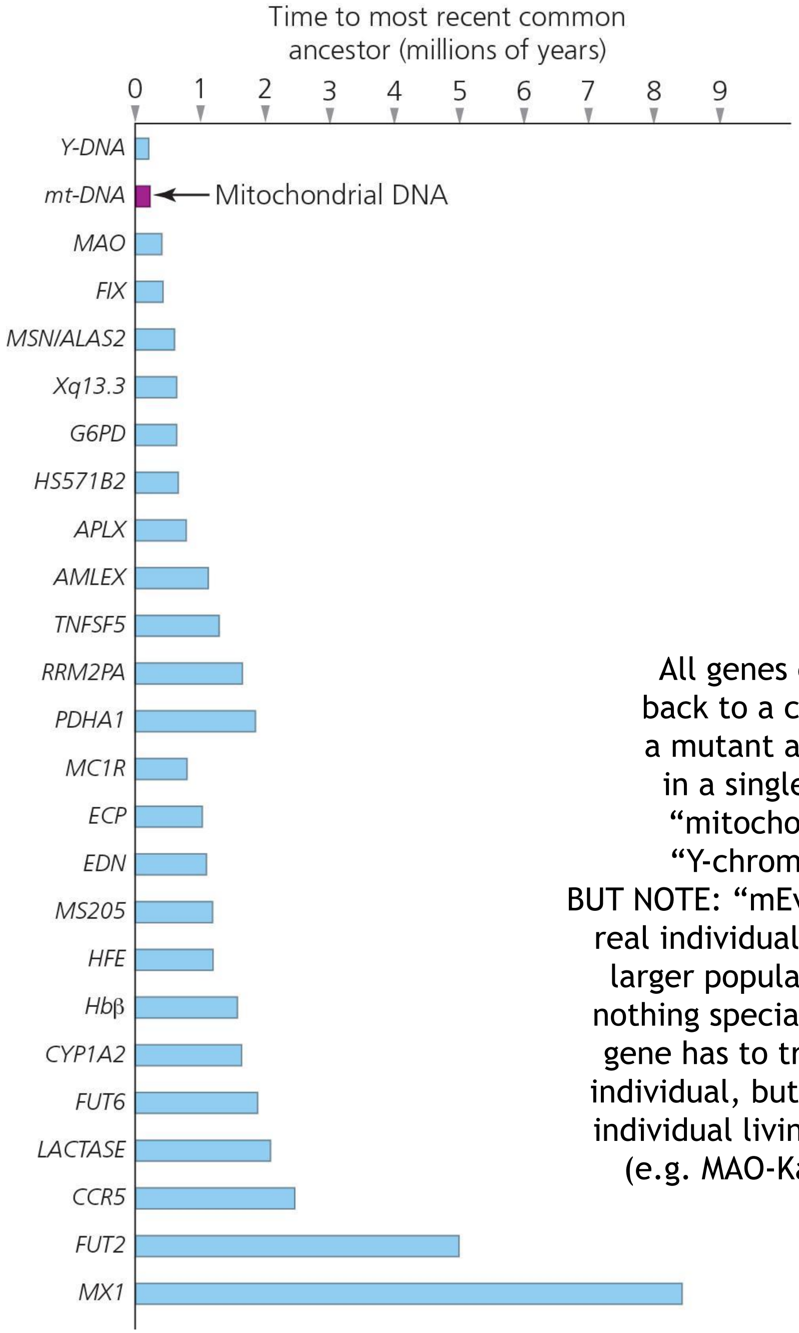
Every generation there will be:

$2N_e v$ mutations

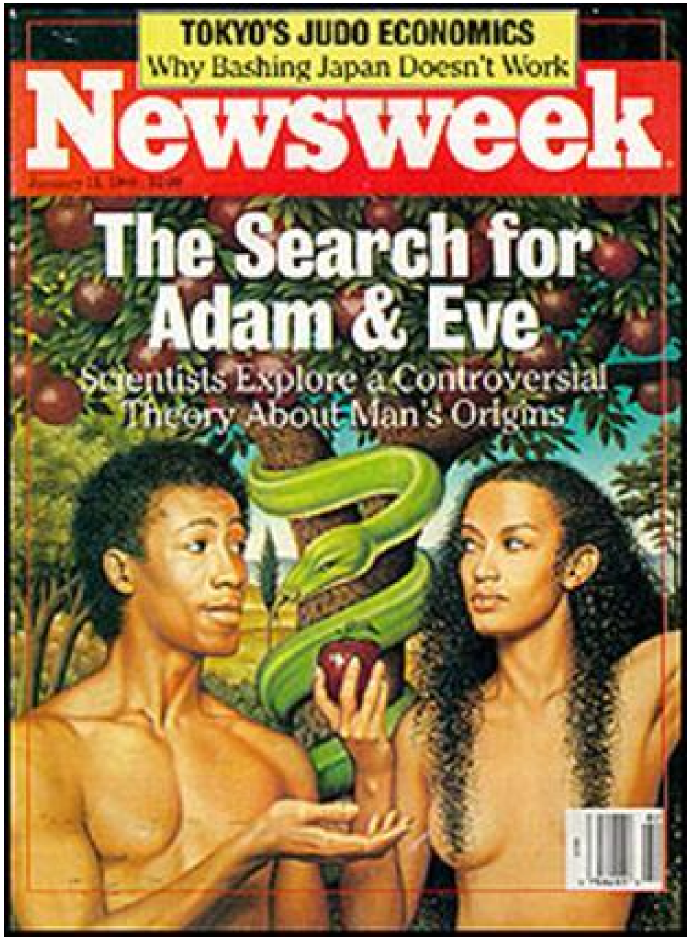
Therefore, alleles turn over at rate *v*

Coalescent times vary among genes

Why??



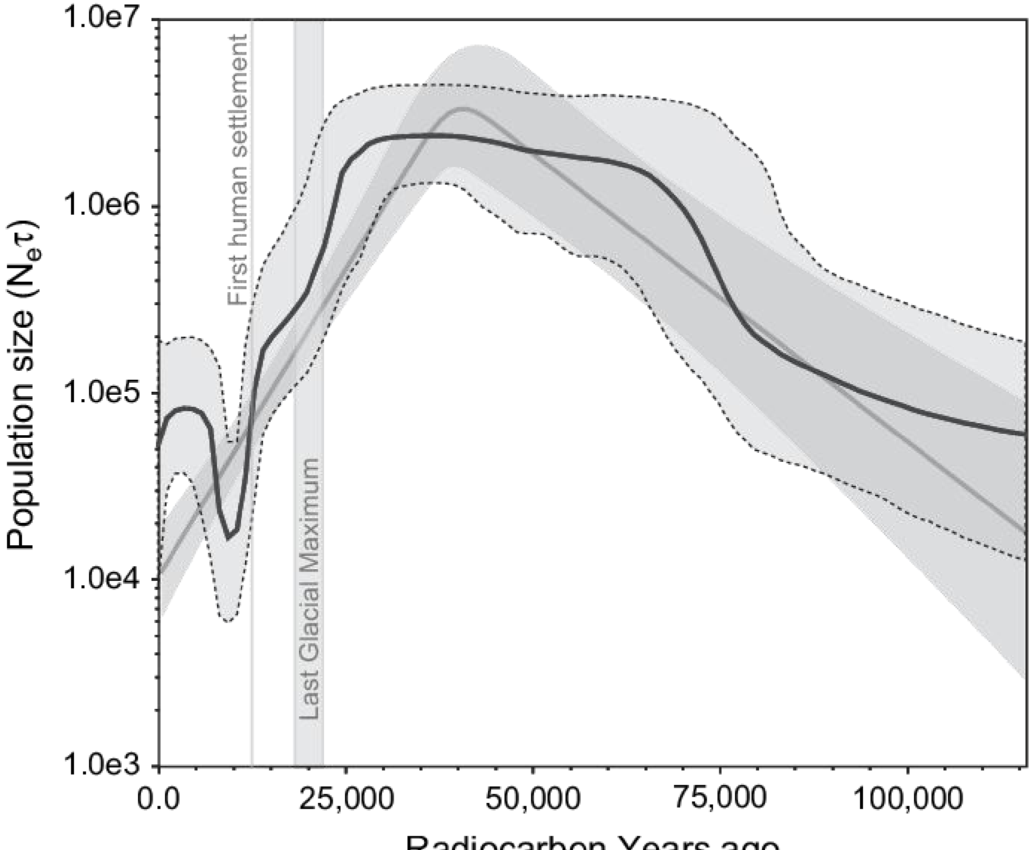
All genes eventually trace back to a common ancestor: a mutant allele that existed in a single individual (i.e. “mitochondrial Eve” and “Y-chromosome Adam”).
BUT NOTE: “mEve” and “yAdam” were real individuals that were part of a larger population. And there was nothing special about them...every gene has to trace back to a single individual, but it will be a different individual living at a different time (e.g. MAO-Karen, APLX-Gengis)

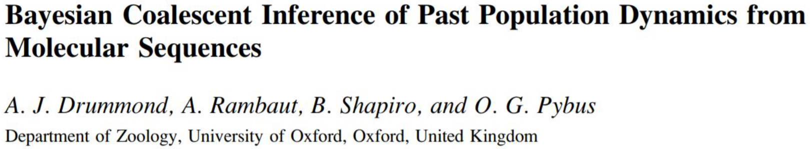


TOKYO'S JUDO ECONOMICS
Why Bashing Japan Doesn't Work
Newsweek
The Search for Adam & Eve
Scientists Explore a Controversial Theory About Man's Origins

The Coalescent

Coalescence-time measured in units of *Ne*





Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences

A. J. Drummond, A. Rambaut, B. Shapiro, and O. G. Pybus

Department of Zoology, University of Oxford, Oxford, United Kingdom

The Multispecies Coalescent

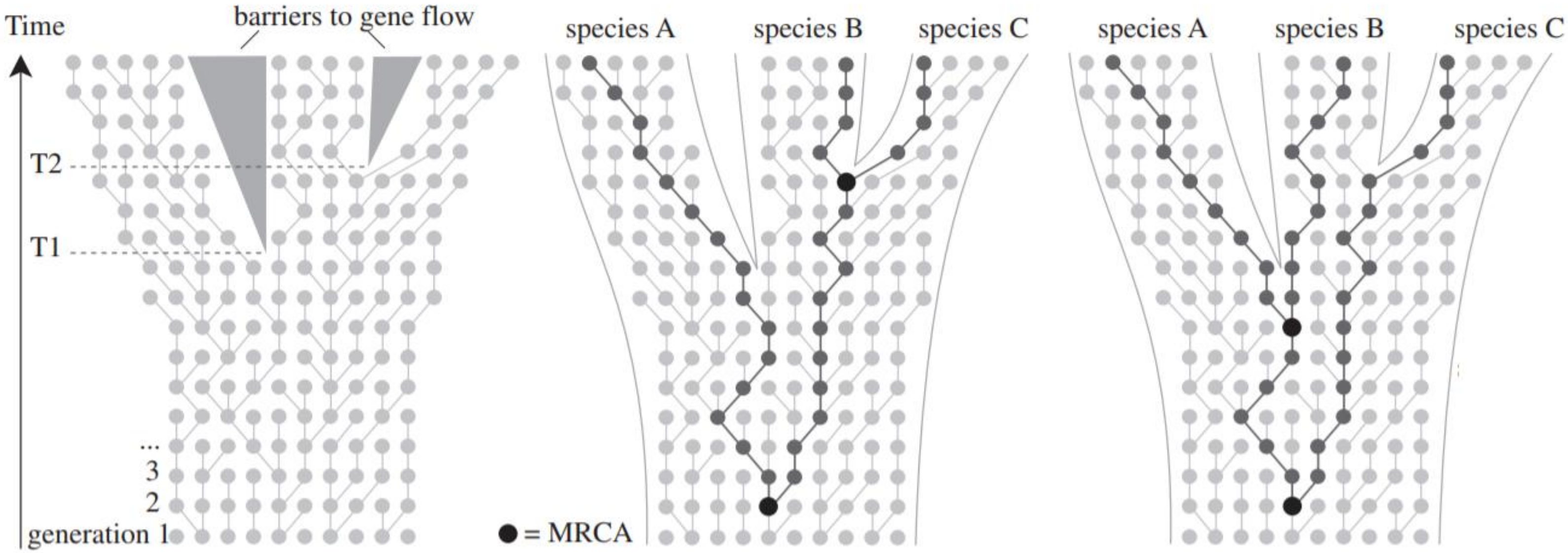
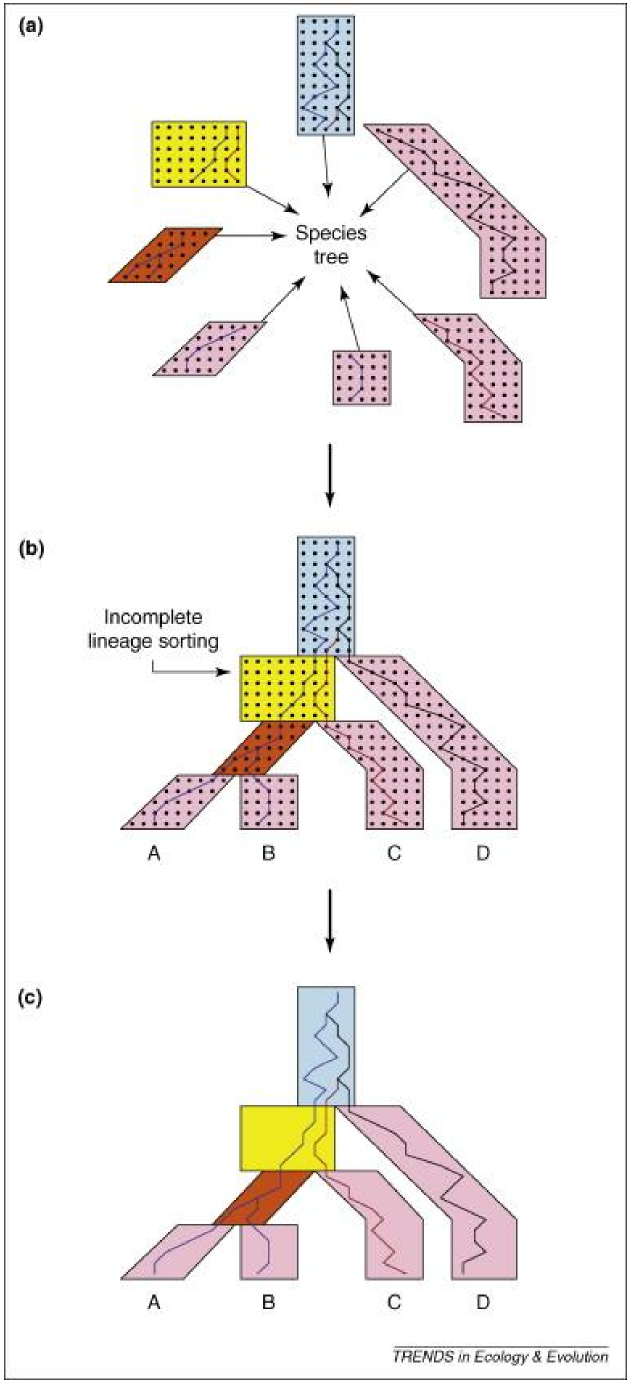


Image: Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., López-Bautista, J. M., Zuccarello, G. C., & De Clerck, O. (2014). DNA-based species delimitation in algae. European journal of phycology, 49(2), 179-196.



TRENDS in Ecology & Evolution

$P(D_i \mid Q_i, \pi_i, \boldsymbol{\psi_i})$ = standard likelihood of gene tree

$P(\boldsymbol{\psi_i} \mid S)$ = Likelihood of gene tree given the species tree

"AND" rule:

$$\mathrm{P}(\mathrm{D}_1, \mathrm{D}_2 \ldots \mathrm{D}_\mathrm{n} \mid \mathrm{S}) = P(D_1 \mid Q_1, \pi_1, \boldsymbol{\psi_1}) * P(\boldsymbol{\psi_1} \mid S) \times \ldots \times P(D_n \mid Q_n, \pi_n, \boldsymbol{\psi_n}) * P(\boldsymbol{\psi_n} \mid S)$$

Methods & software

Parsimony - “MDC” species tree that minimizes deep coalescences (can be inconsistent estimator)

ML - STEM (Kubatko & Degnan 2007). Requires gene trees to be well-estimated and clock-like

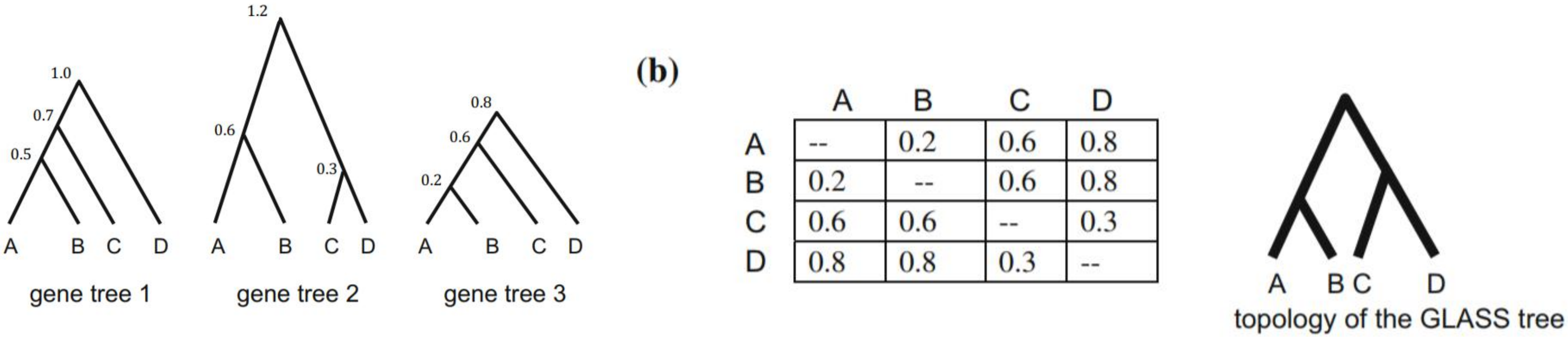
Bayesian - BEST, *BEAST, BPP. Bayesian approaches that integrate over uncertainty in gene trees. Great models...but complex and hard to converge!

$$P(S \mid D) \propto \int_G (\prod_{i=1} P(d_i \mid g_i) P(g_i \mid S)) P(S) dG,$$

Other methods

Concordance analysis- BCA/BUCKy. Semi-parametric clustering of gene trees into “concordance blocks” without regard to process

Summary methods- Uses properties of multispecies coalescent to summarize gene trees. STAR/STEAC/GLASS

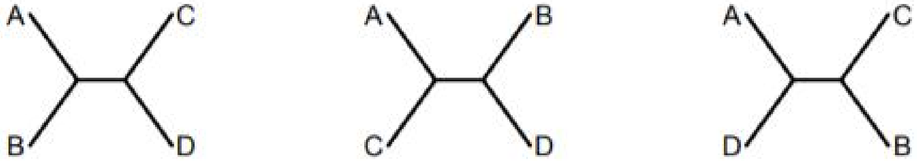


	A	B	C	D
A	--	0.2	0.6	0.8
B	0.2	--	0.6	0.8
C	0.6	0.6	--	0.3
D	0.8	0.8	0.3	--

Other methods

Quartets approaches: ASTRAL/SVDQuartets

Avoids calculating full likelihood, instead focuses on site patterns over 4 taxon combinations. Good for SNPs and genomic scale data



Species Tree Inference Summary – Comparison of Methods

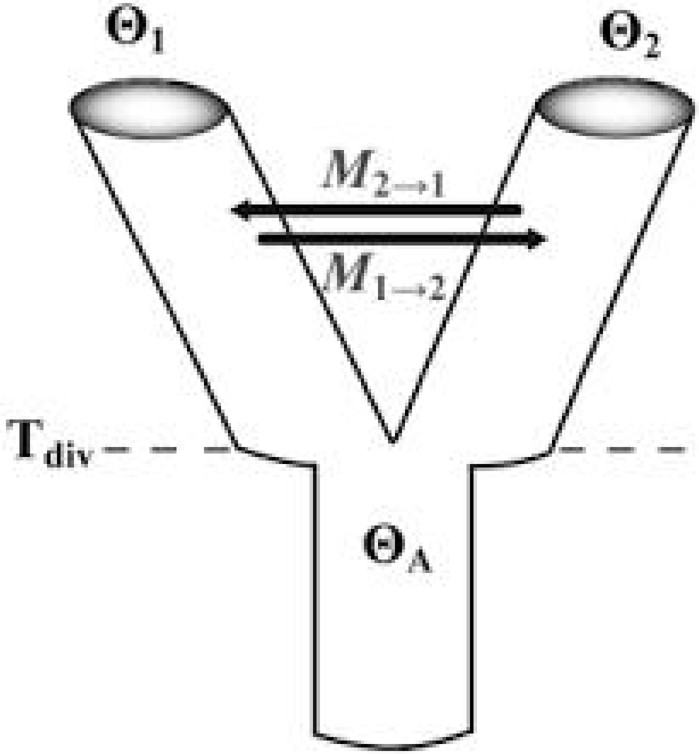
Software	Data Type	Measure of Uncertainty	Computation Time	Models Included
BEST	multilocus	posterior probability	long; can be run in parallel	coalescent; all reversible substitution models
*BEAST	multilocus	posterior probability	intermediate; can be run in parallel	coalesent; all reversible substitution models; relaxed clock; variable population sizes
BPP	multilocus	posterior probability	long	coalescent; JC69 model only; species delimitation
SVDQ	multilocus; SNP	bootstrap	short	coalescent; all reversible substitution models; parameter estimation ?
SNAPP	biallelic SNP; AFLP	posterior probability	long; can be run in parallel	coalescent; two-state substitution model; Bayes factor delimitation
ASTRAL	unrooted gene trees	bootstrap	short given gene trees	no specific model assumed
MP-EST	rooted gene trees	bootstrap	short given gene trees	coalescent model

Adding gene flow...

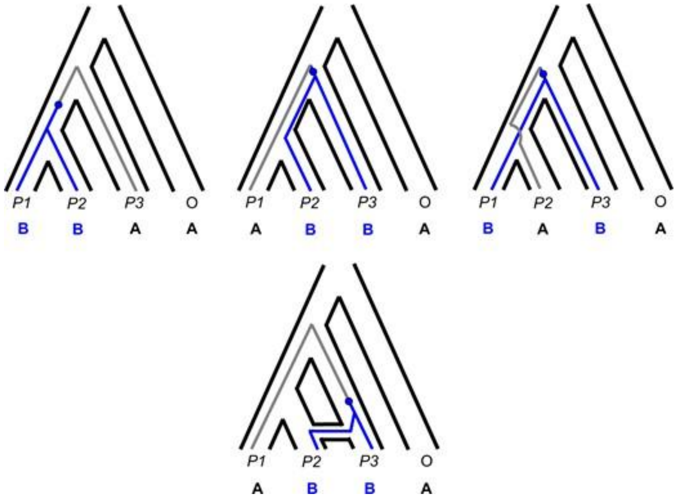
Often limited to a small number of species using multispecies coalescent models

e.g. IM & MSNC based models

(Wakeley & Hey 98, Nielsen & Wakeley 2001; Meng & Kubatko 2009)



ABBA - BABA tests



Quartet based method:

(((Sp1,Sp2),Sp3),Out)

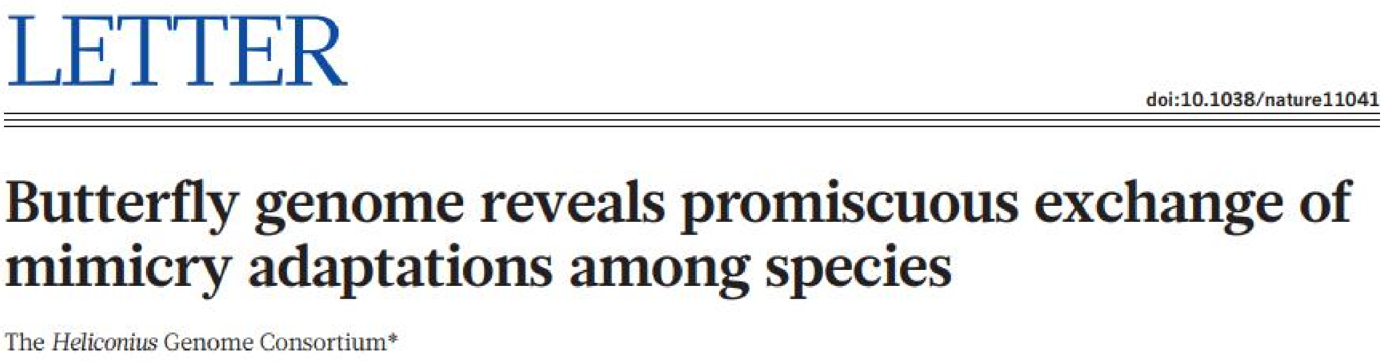
A B B A

B A B A

D = [sum(ABBA) – sum(BABA)] / [sum(ABBA) + sum(BABA)]

H0: f(ABBA) = f(BABA)

H1: Hybridization increases either ABBA or BABA (Sp2 x Sp3) or (Sp1 x Sp3) respectively.

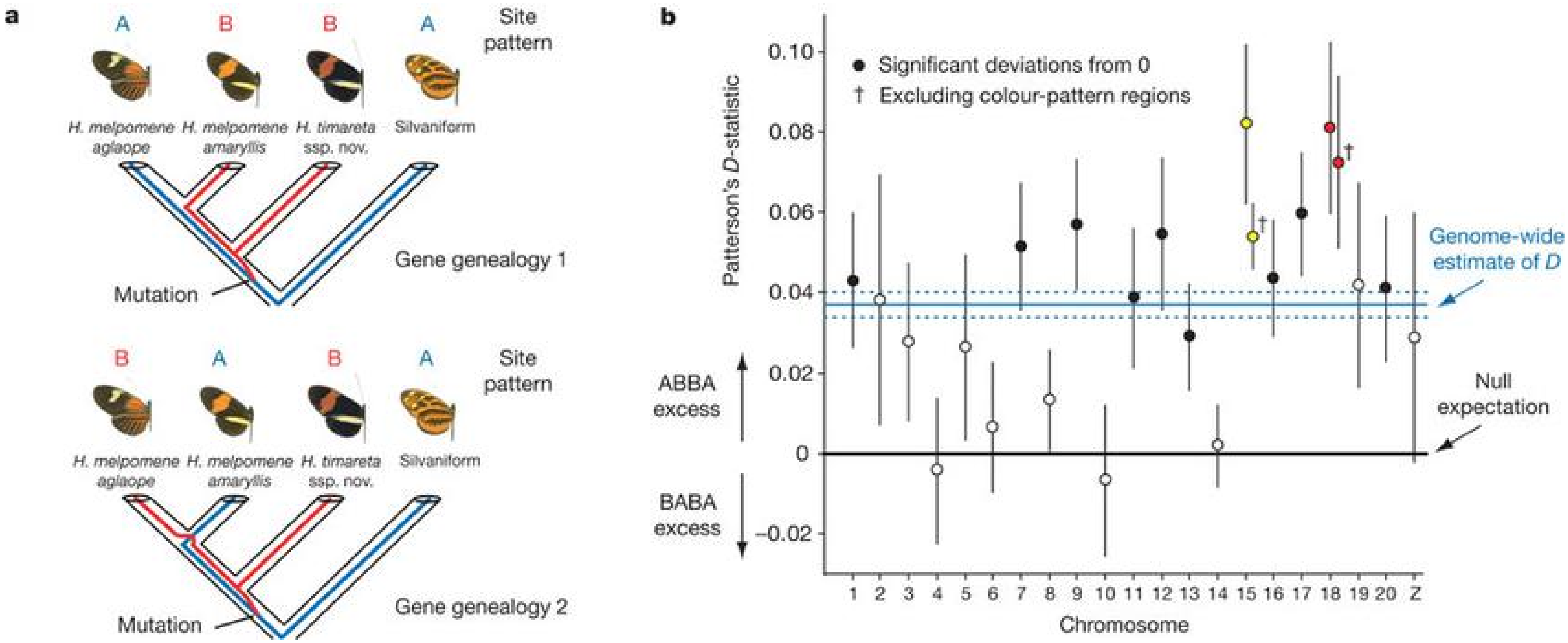


LETTER

doi:10.1038/nature11041

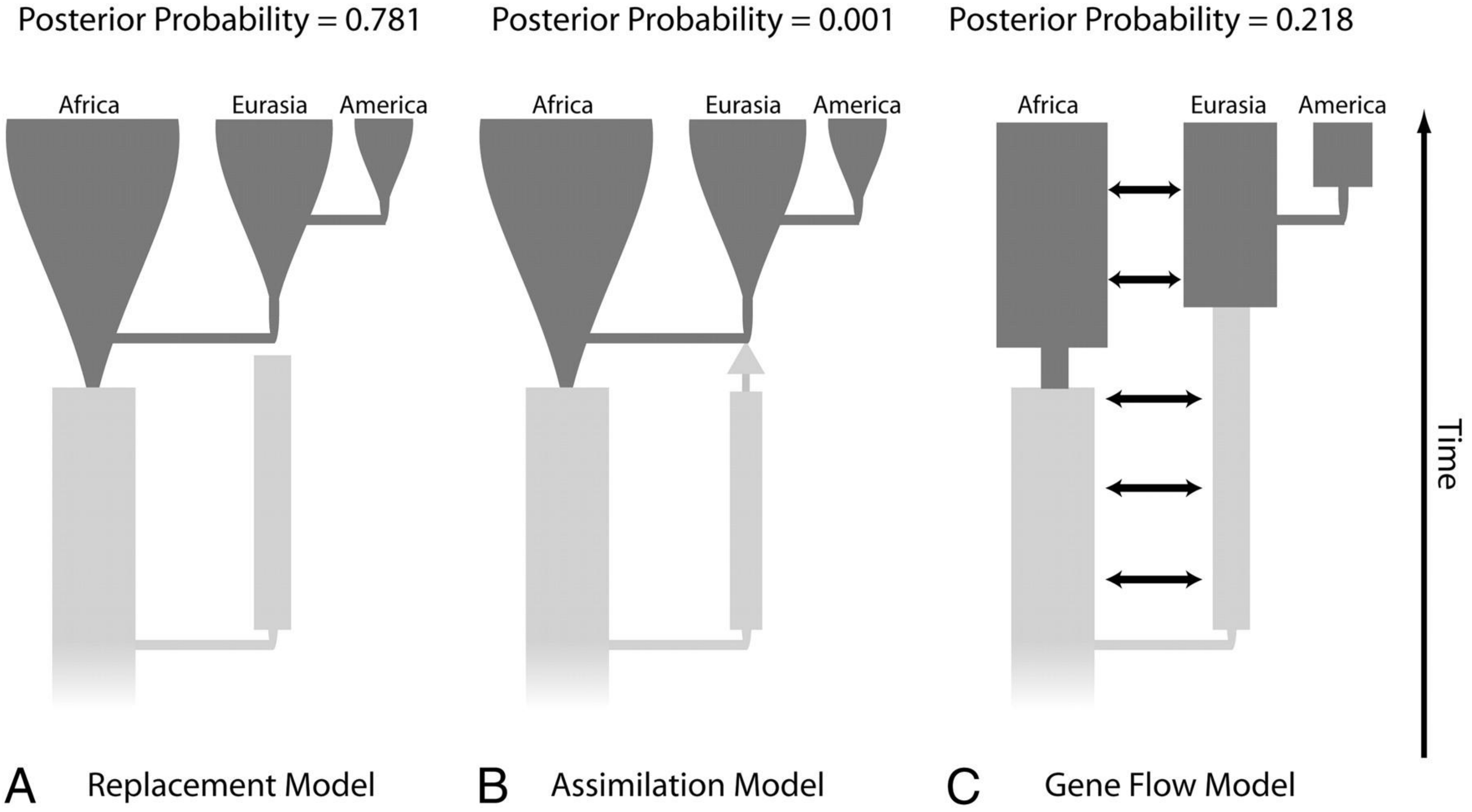
Butterfly genome reveals promiscuous exchange of mimicry adaptations among species

The *Heliconius* Genome Consortium*

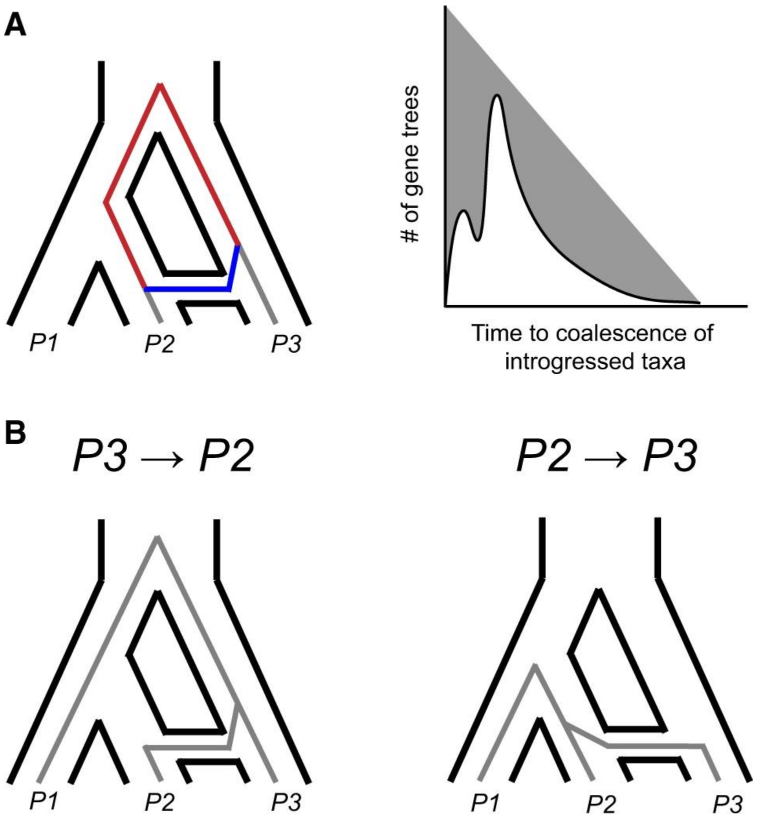


Robust & powerful for closely related species (w/o convergence), not great for quantifying amount of introgression or directionality.

Hypothesis testing

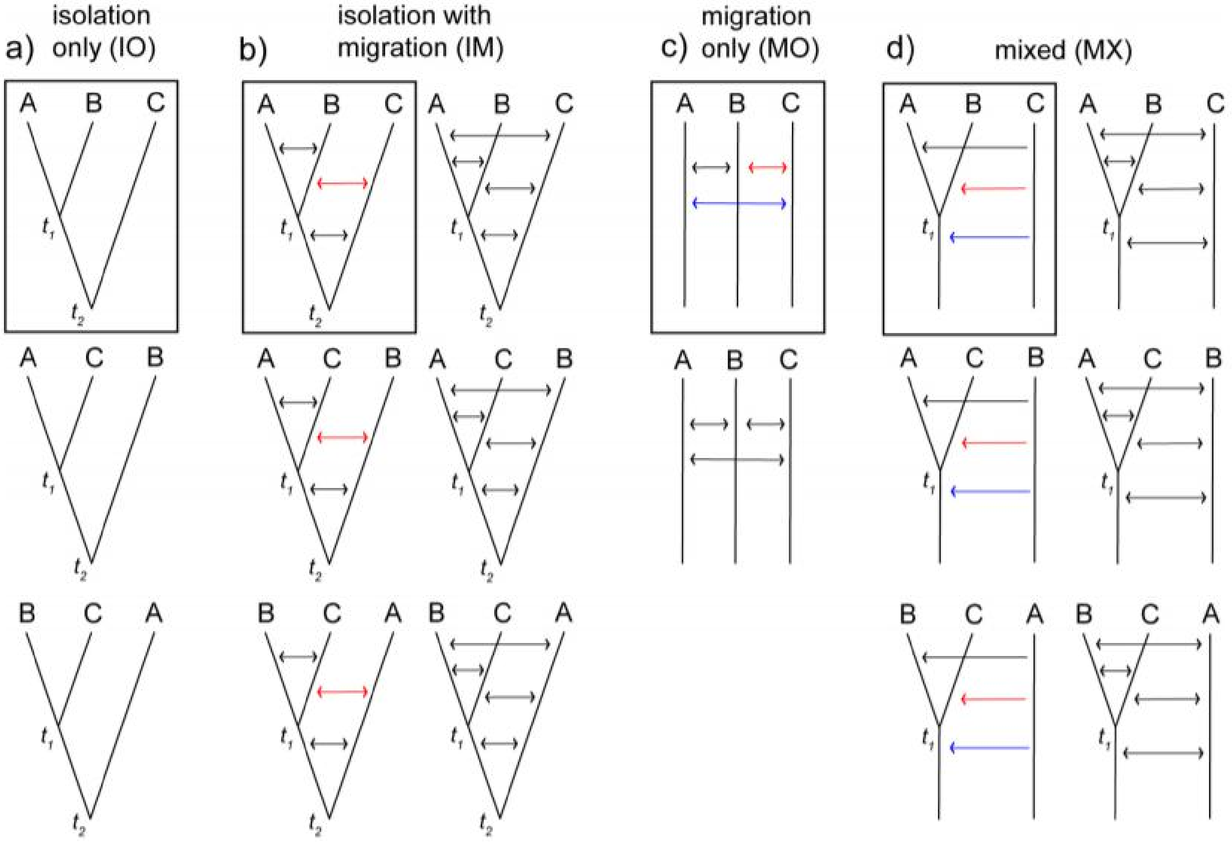


Coalescence times provide information on the timing, direction, and presence of introgression. (A) Postspeciation introgression between P2 and P3 allows them to coalesce more quickly at introgressed loci (blue). This reduces their whole-genome divergence relative to P1 and P3, an asymmetry that can be used to test for introgression. Since coalescence can now occur at one of two times, after introgression (blue) or after speciation (red), it also results in a bimodal distribution of coalescence times across loci (right figure). The more recent peak of this distribution can be used to estimate the timing of introgression. (B) The direction of introgression between P2 and P3 affects the time to coalesce of P1 and P3 at introgressed loci. P2 → P3 introgression allows P1 and P3 to coalesce more quickly (right), reducing their divergence at introgressed loci.



Genetics, Volume 220, Issue 2, February 2022, iyab173, https://doi.org/10.1093/genetics/iyab173

Search among all possible models…
PHRAPL (Jackson et al. 2017)



Some general thoughts...

Scaling multispecies coalescent to genomic scale is hard, adding more data doesn't necessarily improve estimation

Filter genes to those with strong phylogenetic signal

Interrogate your data

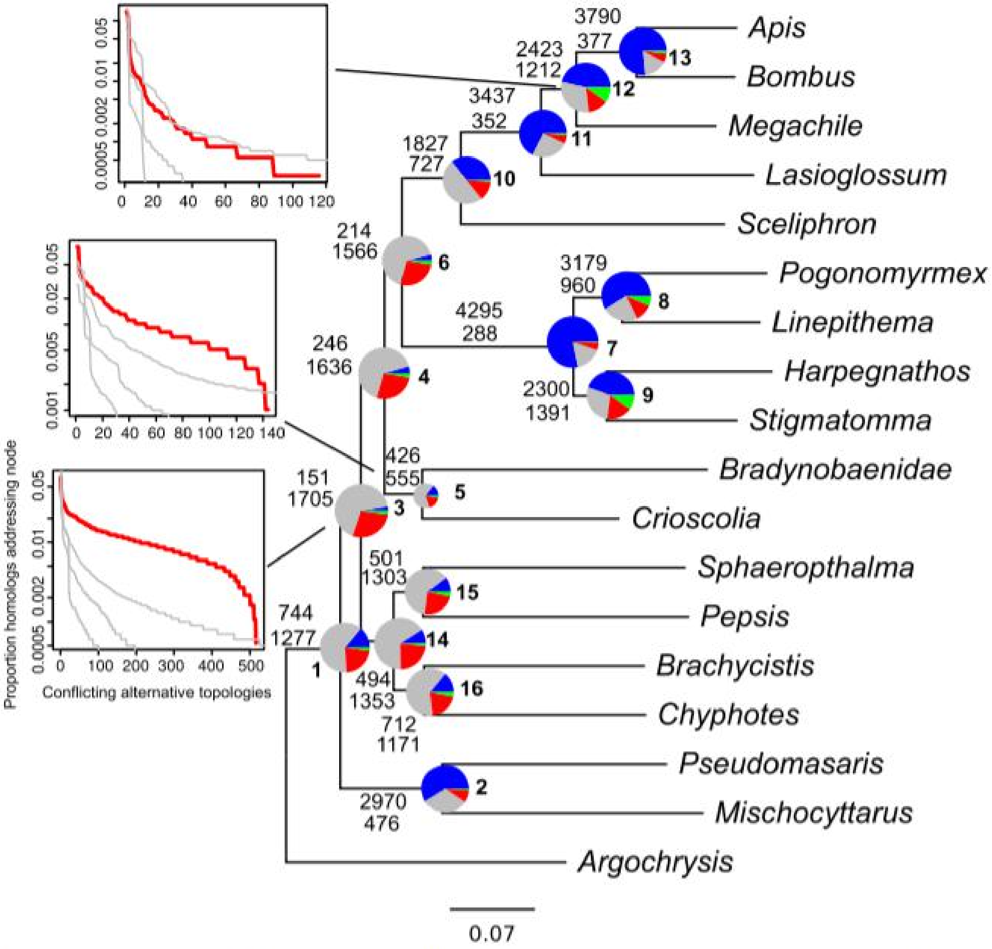


Fig. 2 Combined ML (species tree) topology for Hymenoptera, with summary of conflicting and concordant homologs. For each branch, the top number indicates the number of homologs concordant with the species tree at that node, and the bottom number indicates the number of homologs in conflict with that clade in the species tree. The pie charts at each node present the proportion of homologs that support that clade (blue), the proportion that support the main alternative for that clade (green), the proportion that support the remaining alternatives (red), and the proportion that inform (conflict or support) this clade that have less than 50 % bootstrap support (grey). The histograms show, for three nodes, the proportion of the total homologs that support each conflicting alternative resolution for the clade in question, sorted from largest to smallest. Grey lines represent distributions of conflicting alternative resolutions based on coalescent simulations generated with three tree heights. The histograms for other nodes are presented in Additional file 2: Figure S5