R-evolutionary insights

Deep dive into SNP-based population genomics

SESSION 9

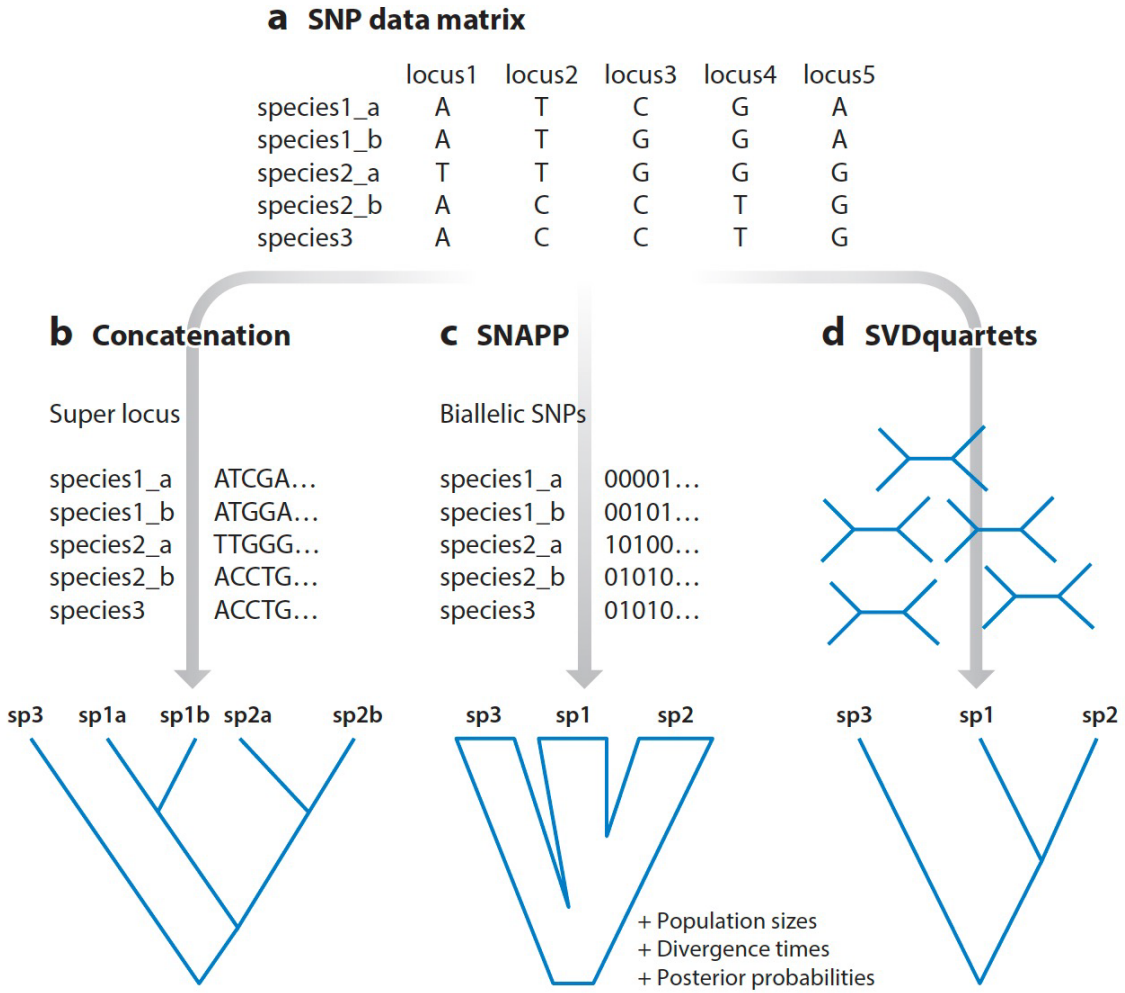
Lineage Divergence

- Craig Moritz – Introduction [15 mins]
- Sally Potter – Types of Analyses [30 mins]
- Arthur Georges – Identification of fundamental entities to work with, distances [15 mins]
- Sally Potter – ML method, SVDquarets, TreeMix [30 mins]
- Craig Moritz-- Discussion [15 mins]

Annual Review of Ecology, Evolution, and Systematics

The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics

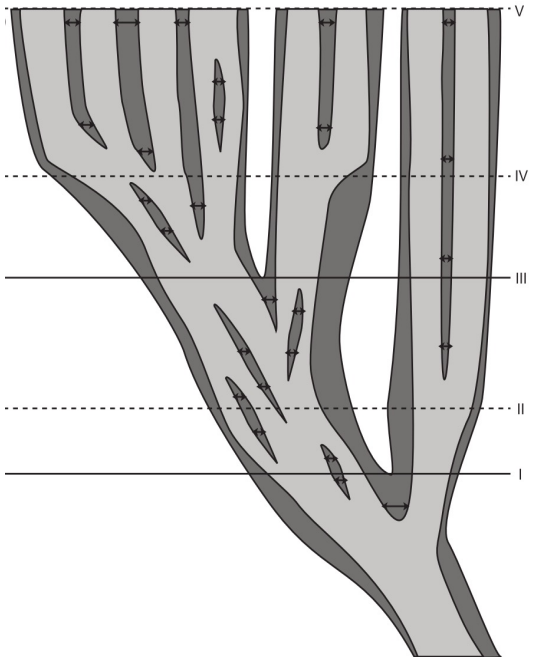
Adam D. Leaché[1] and Jamie R. Oaks[2]



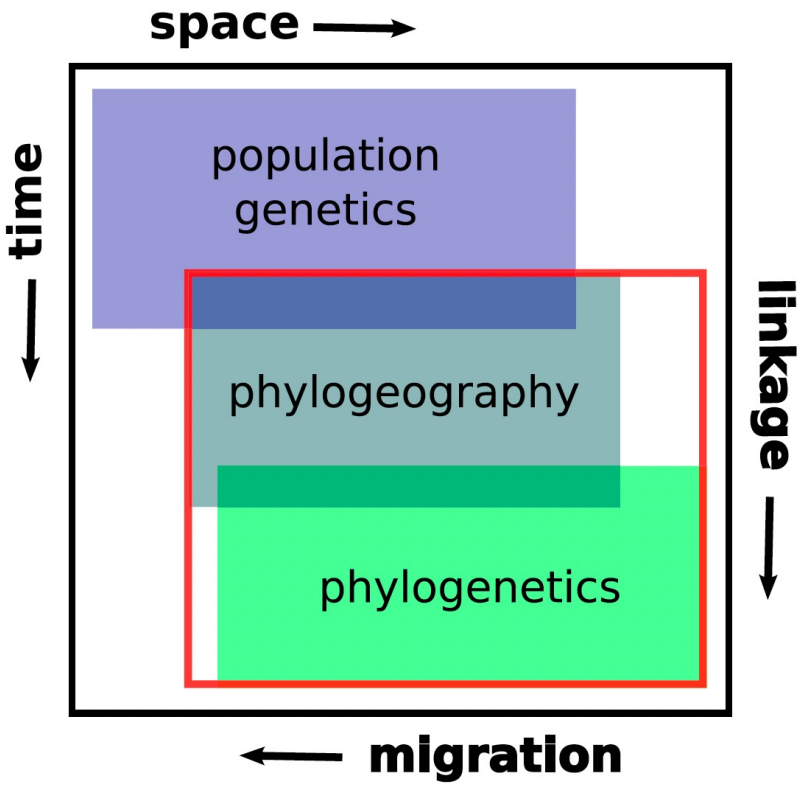
Reticulation, divergence, and the phylogeography–phylogenetics continuum

PNAS 2016

Scott V. Edwards[a,1], Sally Potter[b,c], C. Jonathan Schmitt[a], Jason G. Bragg[b,c], and Craig Moritz[b,c]

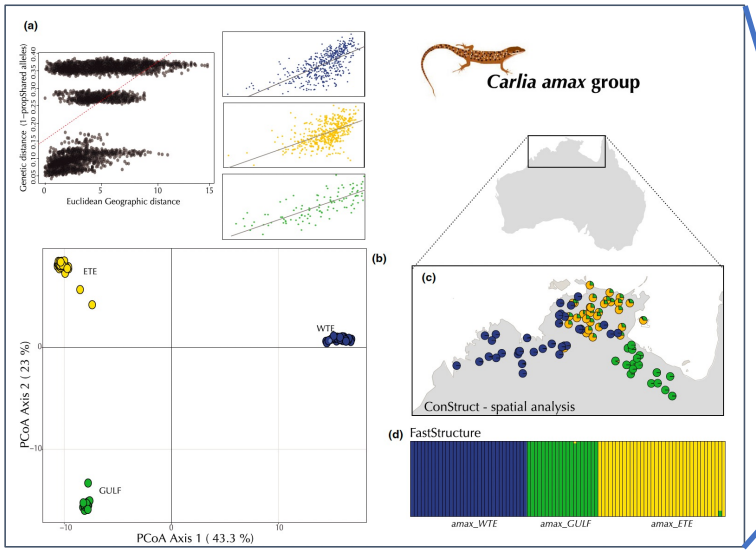


(Cutter 2013)

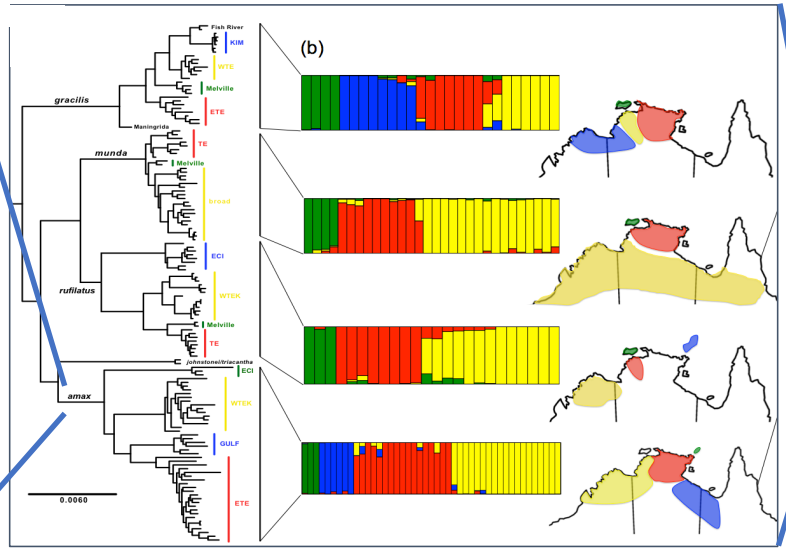


Scales of interest – populations to clades

Landscape and population genetics
SNPs Fenker et al. 2021



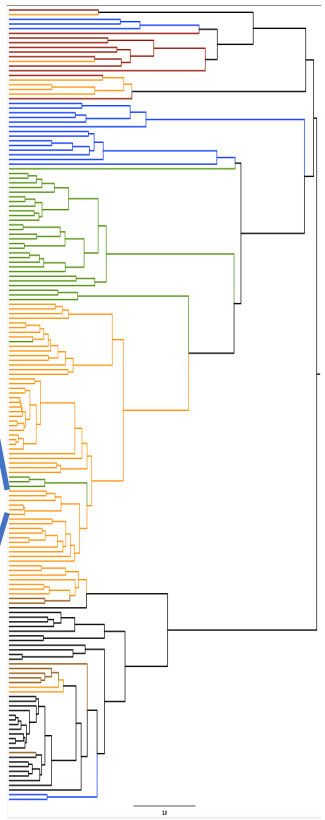
Comparative phylogeography & historical demography. Exon sequencing. Potter et al. 2018



Sampling unit = individuals

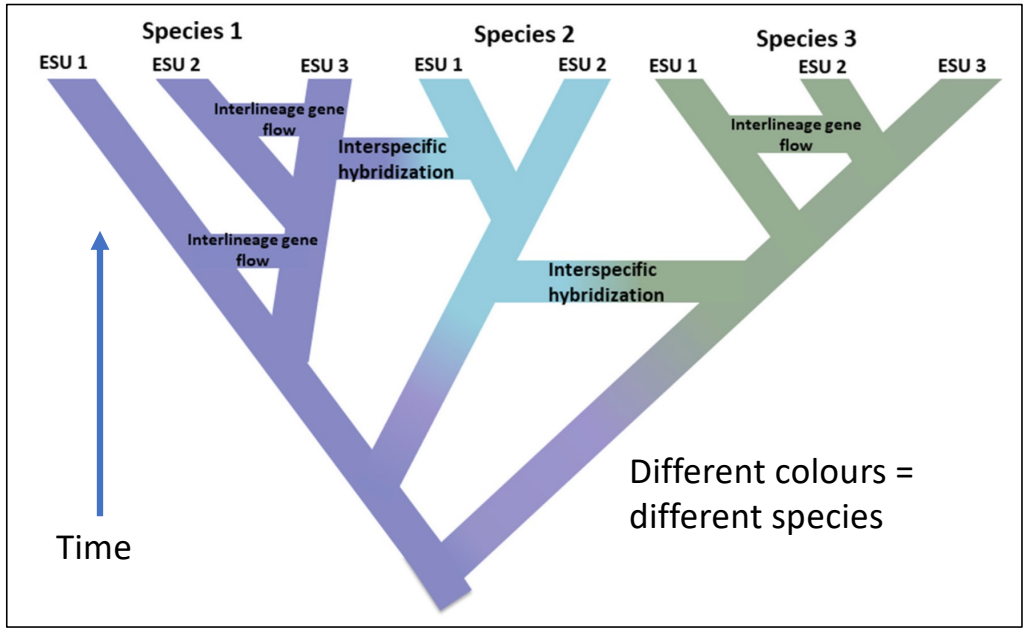
Species trees: Eugongylus skinks. Exon sequences

Bragg et al. in prep.



Sampling unit = separate genetic lineages (incl species + ESUs)

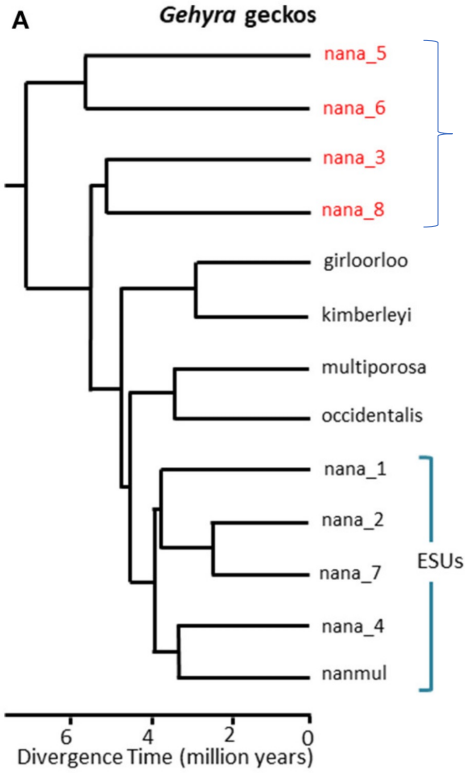
The speciation continuum



The reality is often more complicated:

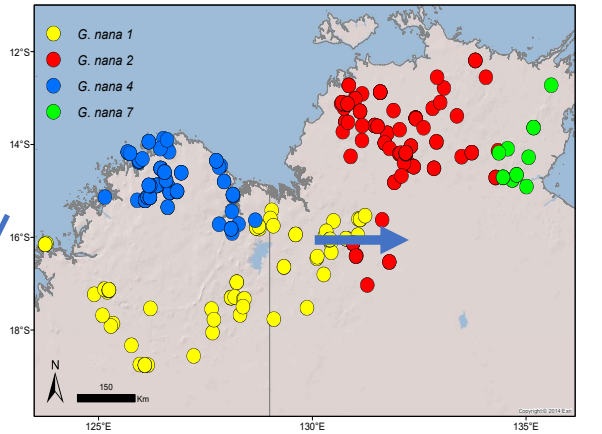
- Low dispersal species often have multiple, long-isolated sets of populations = Evolutionarily Significant Units (ESUs).
- Both species and ESUs might have exchanged genes during their history of divergent evolution

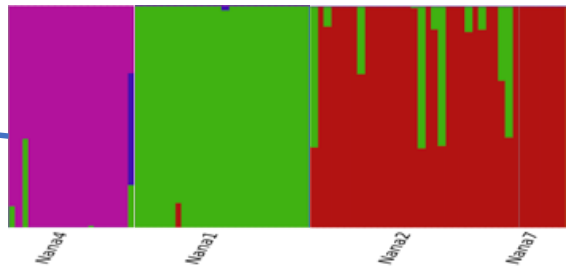
Unravelling a species complex of tropical geckos (Gehyra nana group - Moritz et al. 2018; Doughty et al. 2018)



4 new species described – no gene flow, morph differences

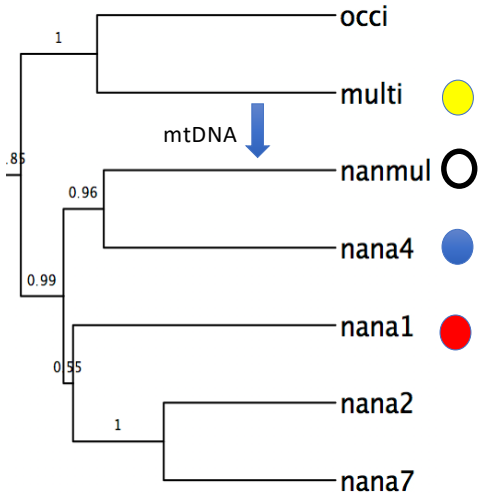
Left as one species with multiple ESUs

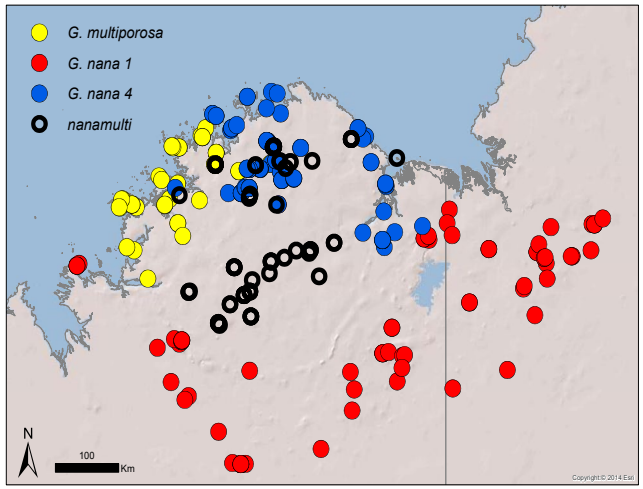




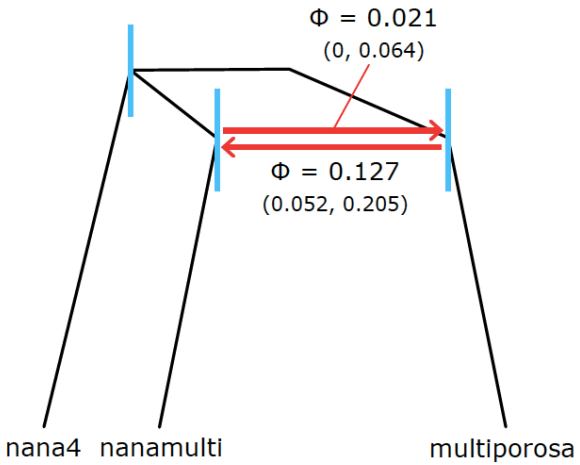
Some gene flow between adjacent ESUs

Now add deep introgression…

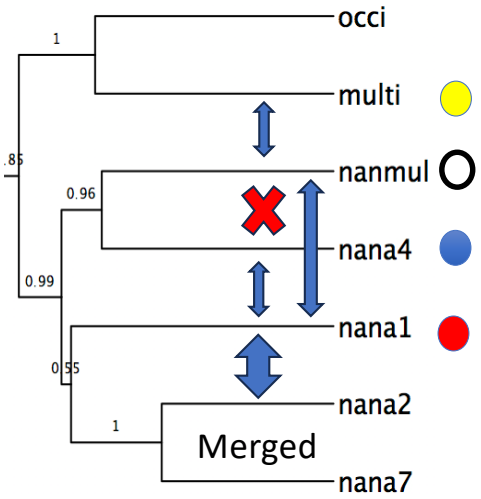


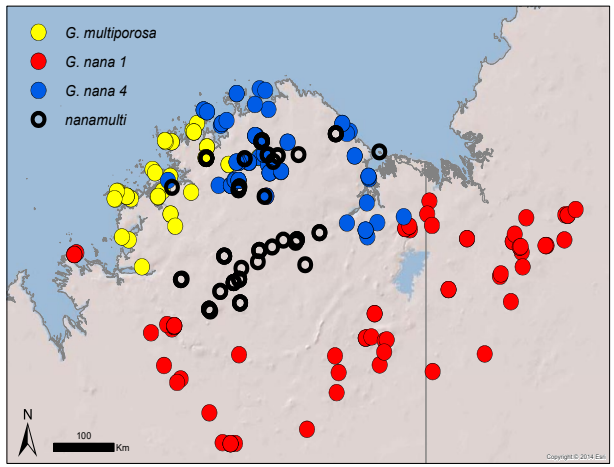


BPP test infers substantial nDNA introgression in the history of "nanamulti" (Read et al. in prep.)



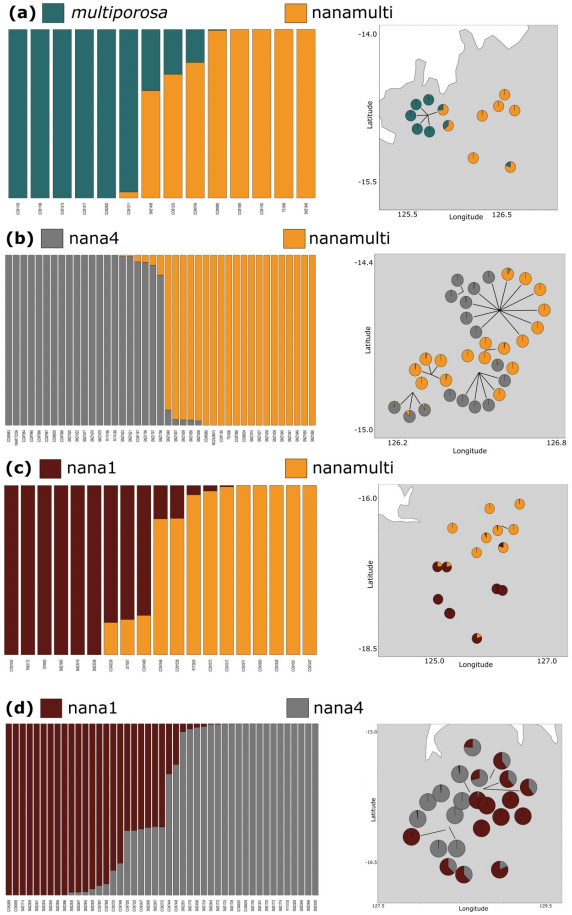
Localised admixture at some but not all lineage boundaries





Read et al. In prep

BPP test infers substantial nDNA introgression in the history of "nanamulti" (Read et al. in prep.)



Admixture among non-sister lineages can affect estimates of phylogenies

The Effect of Gene Flow on Coalescent-based Species-Tree Inference

COLBY LONG[1,*] AND LAURA KUBATKO[2]

Syst Biol 2016

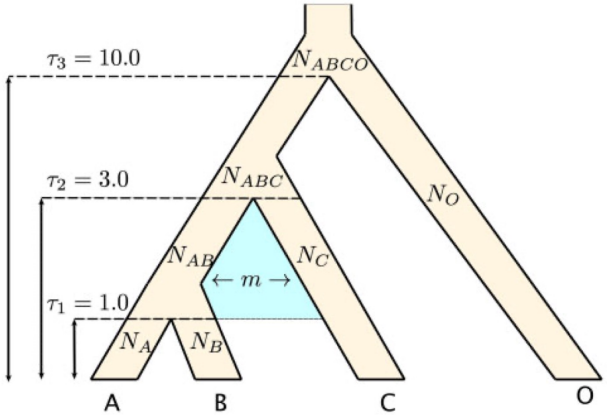
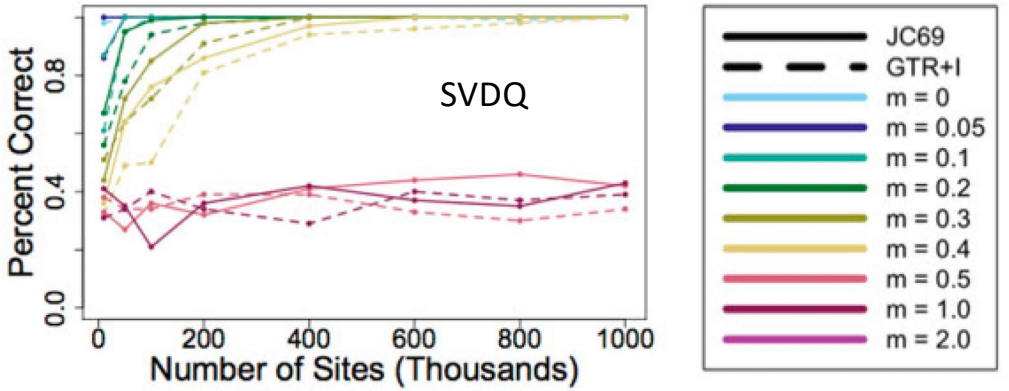
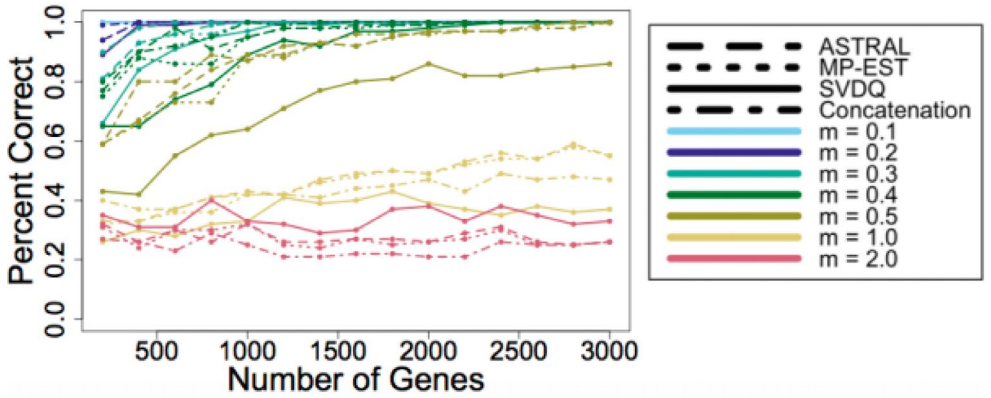


FIGURE 5. Model tree for the simulation studies.

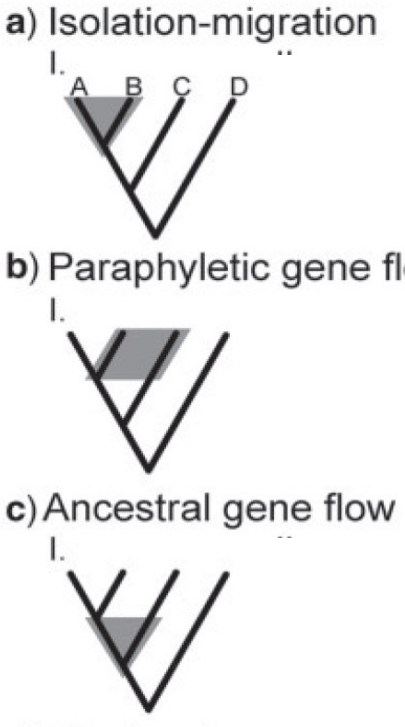


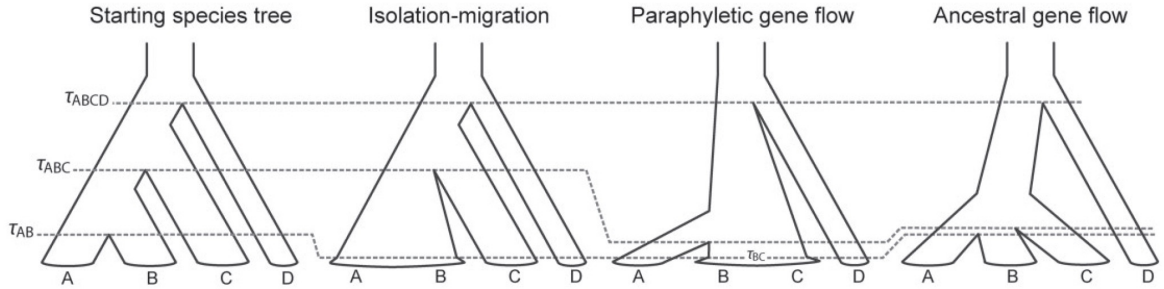


The Influence of Gene Flow on Species Tree Estimation: A Simulation Study

ADAM D. LEACHÉ[1,*], REBECCA B. HARRIS[1], BRUCE RANNALA[2,3], AND ZIHENG YANG[3,4]

Syst Biol 2013





IGURE 8. Species tree distortions caused by gene flow that can result from coalescent methods that only model ILS. Dashed lines illustrate cies tree compression, and the widening of branches illustrates species tree dilation in relation to the starting species tree.

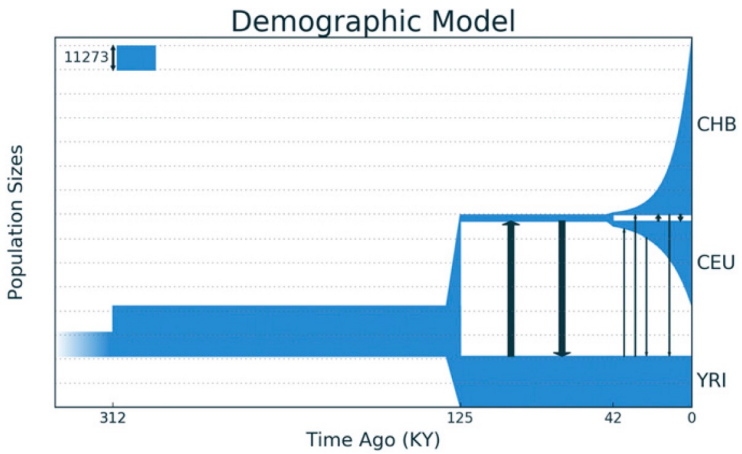
Including recently admixed, or incorrectly assigned, individuals can substantially alter inferred species trees

Testing models of population divergence using SNP data – eg. moments (python)

Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation

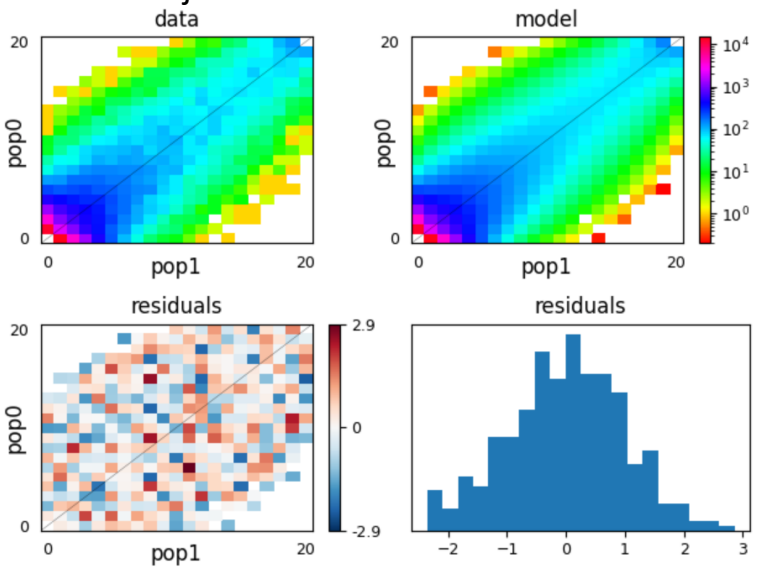
Genetics 2017

Julien Jouganous, Will Long,[1] Aaron P. Ragsdale, and Simon Gravel[2]



Also GADMA,

Observed vs simulation joint AFS with model fits



https://moments.readthedocs.io/en/main/

*Momi*2 – multi-population histories with divergence and pulsed migration

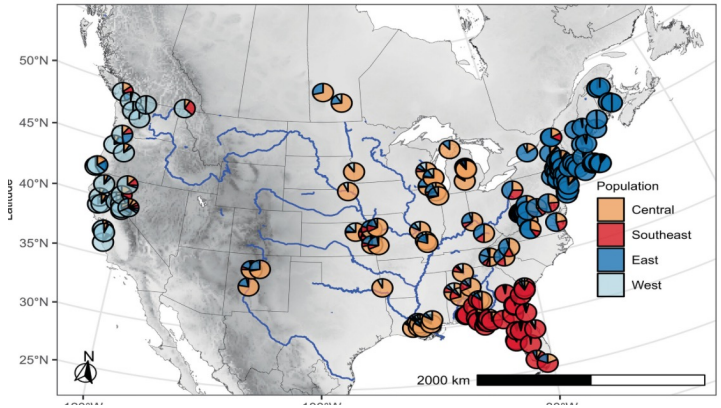
Efficiently Inferring the Demographic History of Many Populations With Allele Count Data

J. Am Stats Assoc 2020

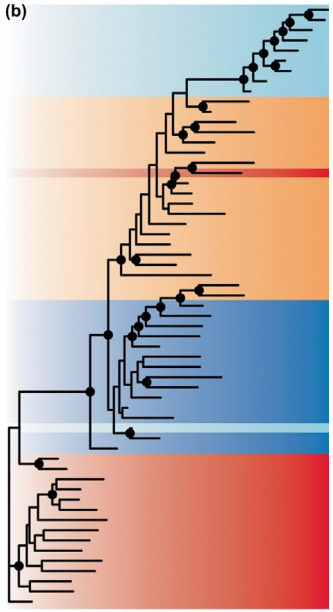
Jack Kamm[a,b,f], Jonathan Terhorst[c], Richard Durbin[a,b], and Yun S. Song[d,e,f]

Biogeographic barriers and historic climate shape the phylogeography and demography of the common gartersnake

Leonard N. Jones II, Adam D. Leaché, Frank T. Burbrink

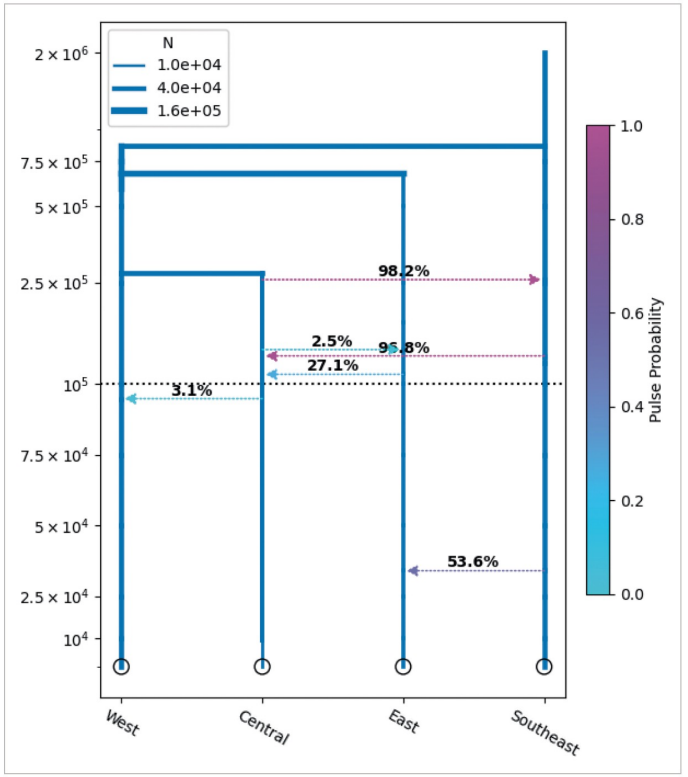


Example – ddRAD data across genetic lineages – TESS3r



IQTREE concat tree

Testing population models using *momi*2 – best fit of 13 divergence modes with splitting, admixture and pop. size change



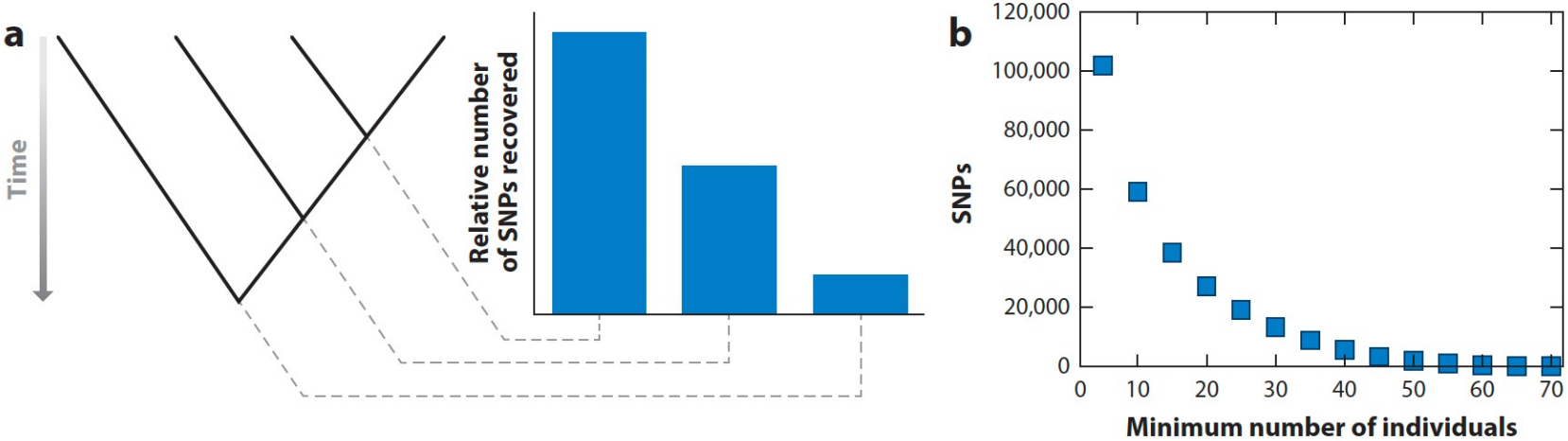
Annual Review of Ecology, Evolution, and Systematics

The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics

Adam D. Leaché[1] and Jamie R. Oaks[2]

An issue for ddRAD SNPs in phylogenetics:

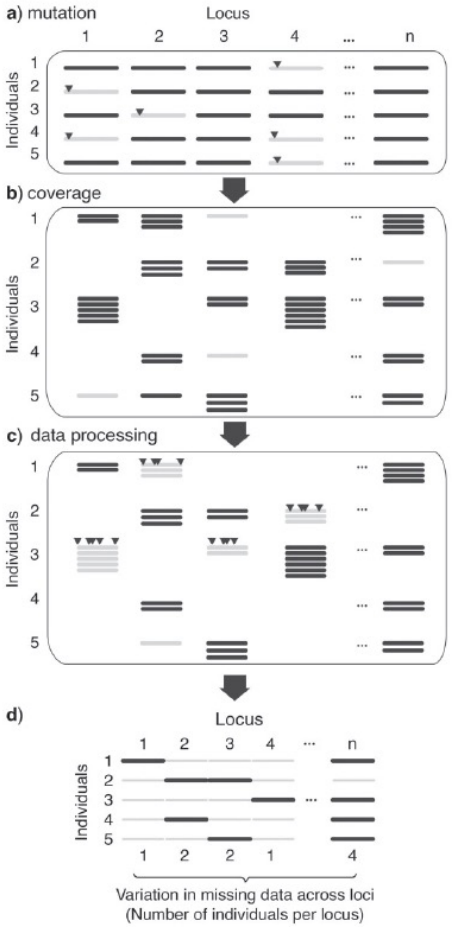
- Allele Drop Out (ADO) due to decay in shared restriction sites



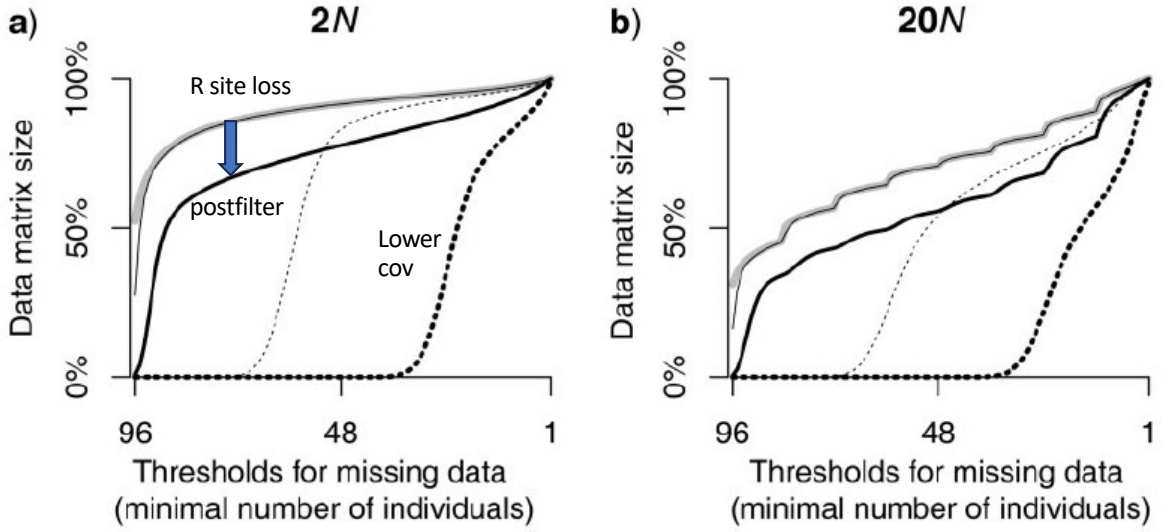
Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences

HUATENG HUANG* AND L. LACEY KNOWLES

Syst Biol 2016



Simulation study of ddRAD datasets – 8 taxa with total depth of 2N vs 20N generations
At moderate coverage (5x) more data lost for deeper divergences

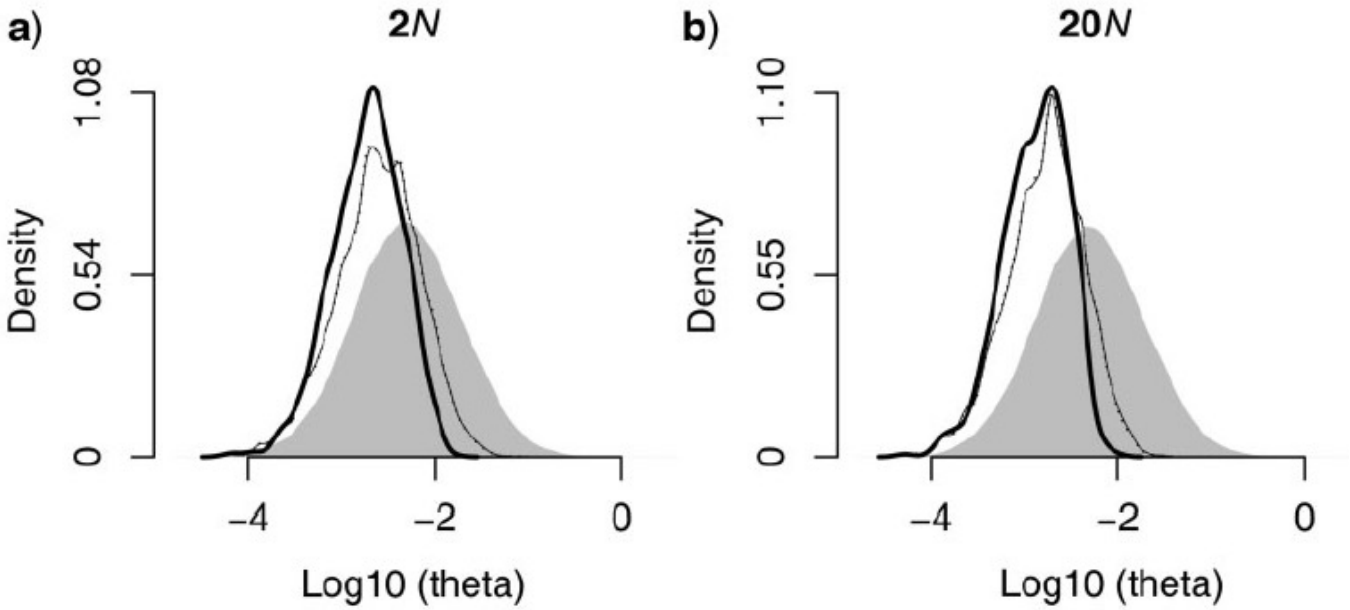


complete incomplete

Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences

HUATENG HUANG* AND L. LACEY KNOWLES

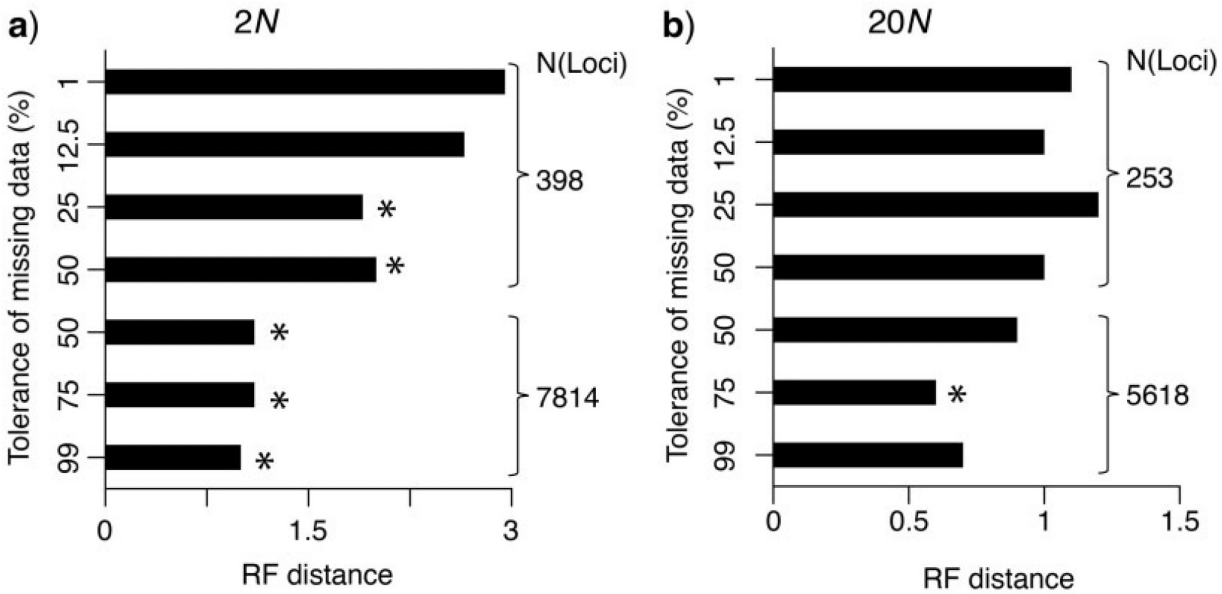
Filtering to remove loci with more missing data can bias against loci with higher mutation rates (theta)



Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences

HUATENG HUANG* AND L. LACEY KNOWLES

Including genes with more missing data yields trees closer to true tree (low RF distance)

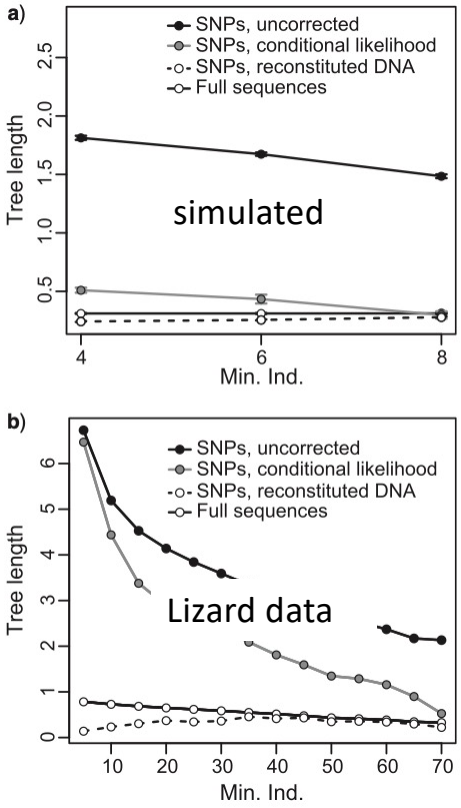


Poorly represented loci can still carry useful phylogenetic signal deep in the tree – even for quite old radiations (Eaton et al. 2017, Syst Biol. 66:399)

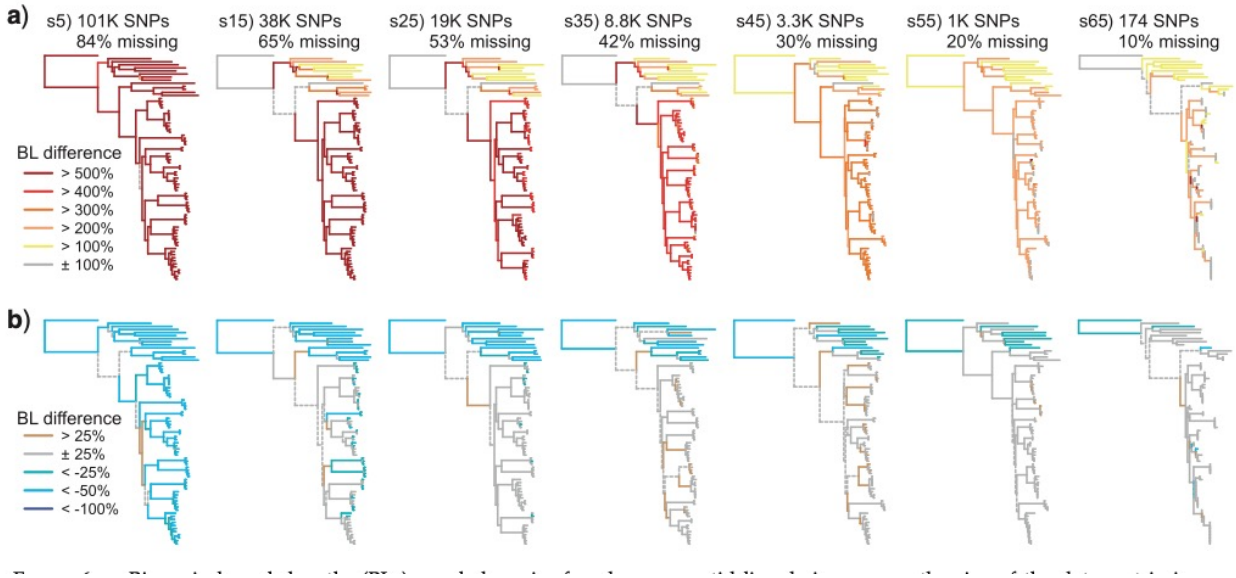
Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias Corrections for Inferring SNP Phylogenies

Syst Biol 2015

ADAM D. LEACHÉ[1,2,*], BARBARA L. BANBURY[1], JOSEPH FELSENSTEIN[1,3], ADRIÁN NIETO-MONTES DE OCA[4], AND ALEXANDROS STAMATAKIS[5,6]



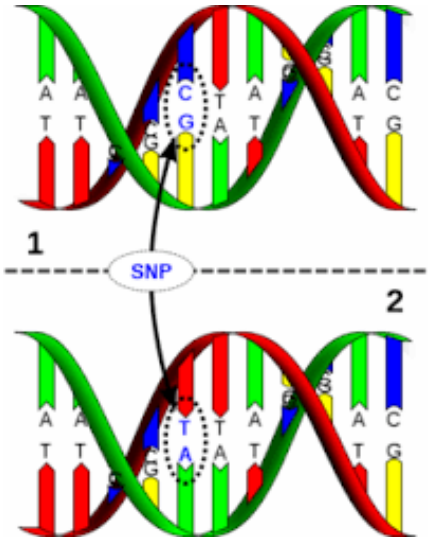
Failure to correct for the number of invariant sites results in massive inflation of branch lengths, especially when including loci with more missing data (left)

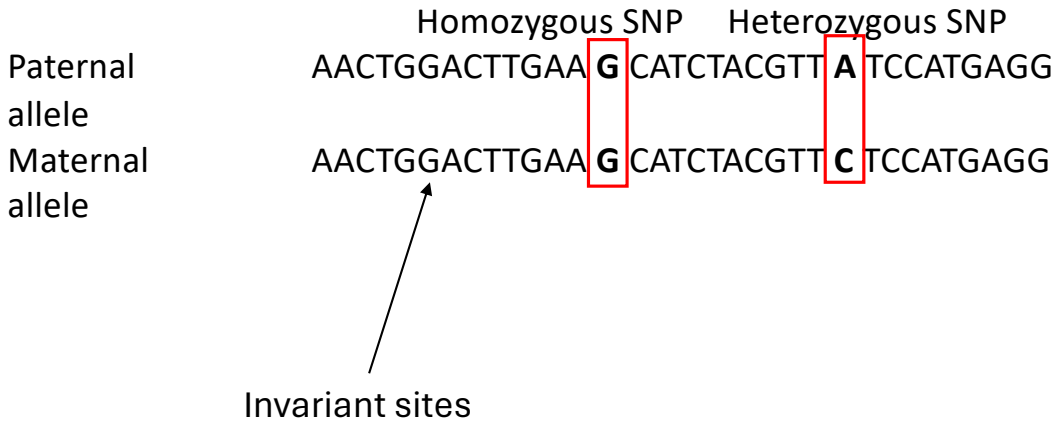


Key points from Intro

- Speciation continuum – popgen -> phylo. Where is your system?
- The value of sampling within as well as across species. Avouid recently admixed samples – eg in hybrid zones
- Topology can be quite robust to low levels of mixing, but branch lengths are strongly affected
- For limited numbers of taxa/populations, complex population histories can be explored using methods simulating multi-SFS
- Filtering SNPs – simulations suggest retaining SNPs with high drop-out; keep invariant sites.

Single Nucleotide Polymorphism vs Sequence data





Reduced representation SNP data – we don't know how frequent SNPs are…

Increasing independent genetic markers

- A single locus is only a single representation of history from highly stochastic ancestral (or genealogical) processes
- SNPs provide opportunity to increase independent markers
- No phasing of haplotypes and no intralocus recombination
- Accuracy of parameters – e.g., population size, divergence times, and migration rates, increases when more loci are sampled (Felsenstein 2006)
- More data for resolving difficult and/or recent phylogenetic relationships

Comparative studies of SNPs vs gene sequences

Empirical + simulation studies

- SNPs & gene sequences produce broadly concordant topological results.

*results can be sensitive to assumptions used during data assembly

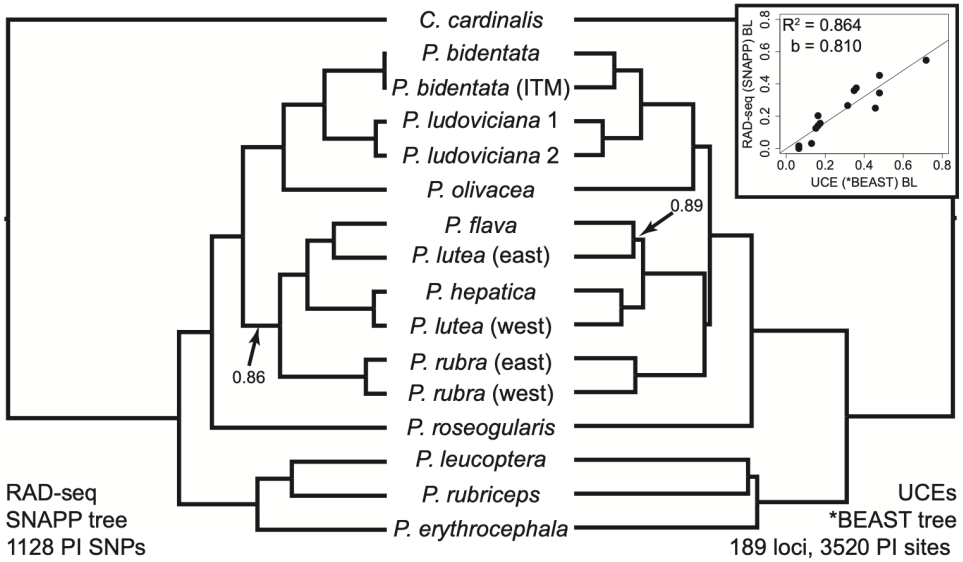
- Most data conflicts associated with a difficult clade comprising a rapid diversification

*important to accurately model heterogeneous data, such as gene tree heterogeneity

- Loci with high mutation rates were the most likely to be missing

*caution against removing data, which could result in biasing the data in favour of loci with low mutation rates

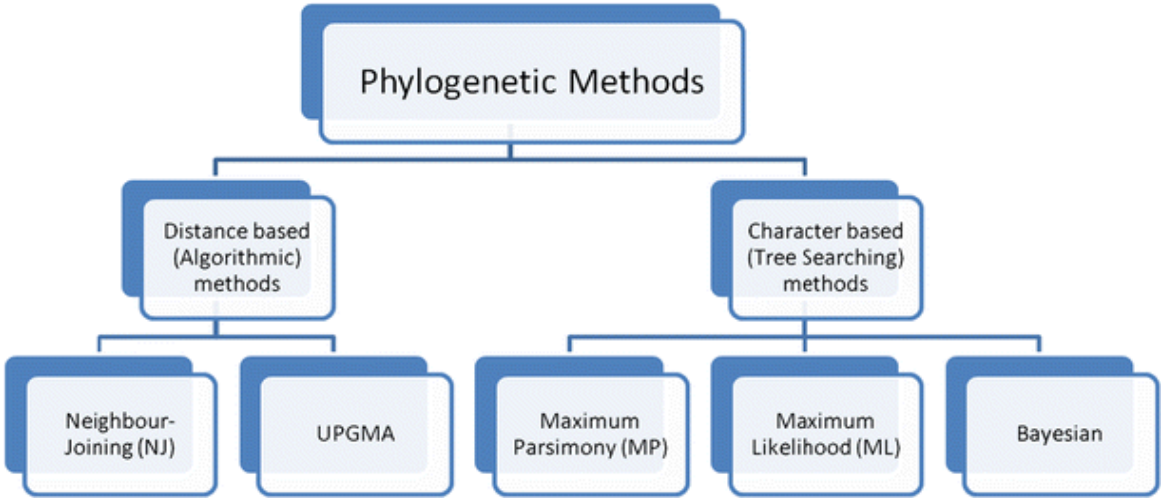
*loci with large amounts of missing data tend to produce discordant topologies with increased branch length errors

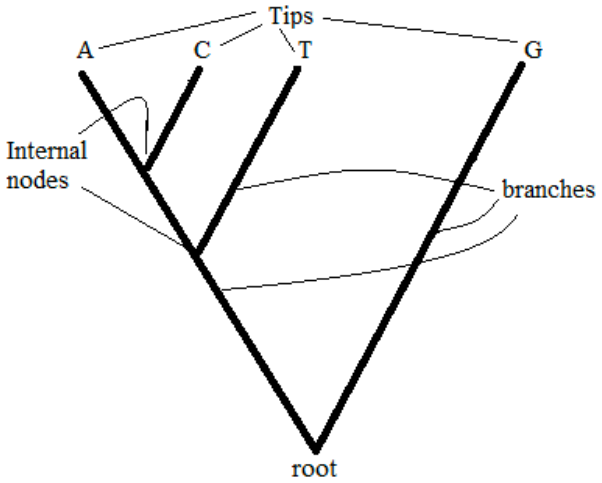


Leaché & Oaks Annu. Rev. Ecol. Evol. Syst. 2017. 48:69–84

Manthey et al. Syst. Biol. 2016. 65(4):640–650

Types of tree reconstruction

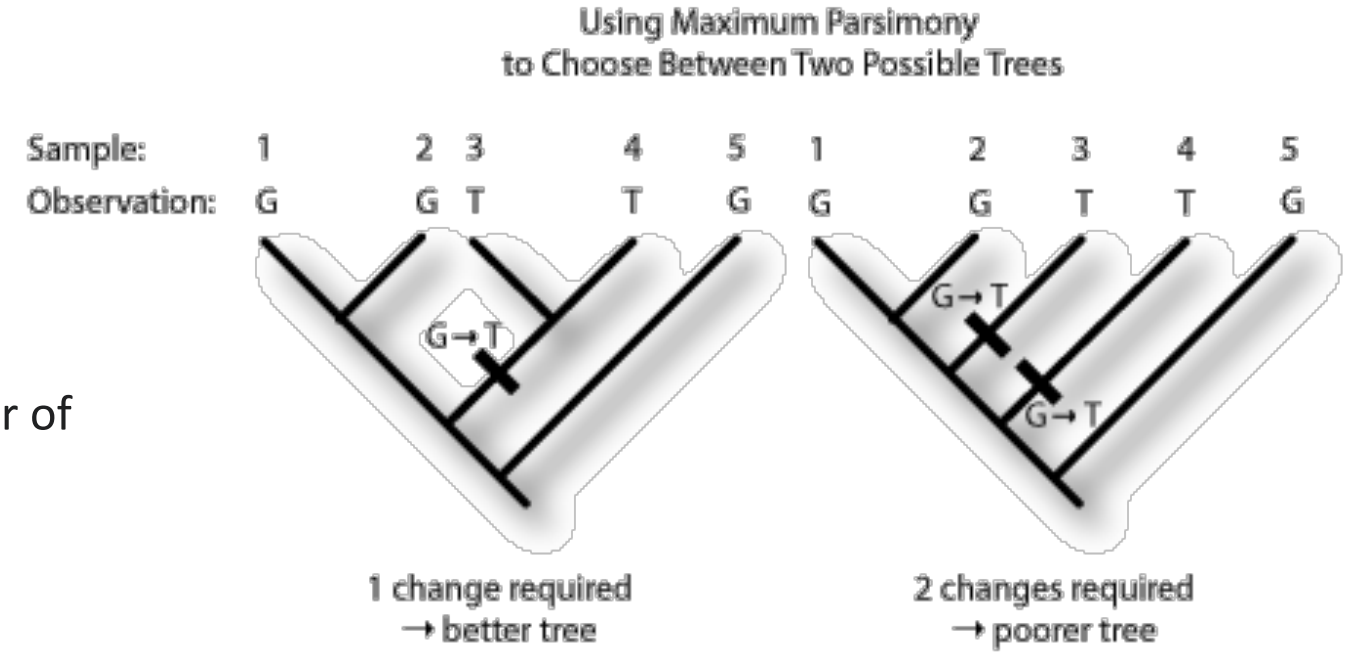




- Distance based methods : find a tree such that branch lengths of paths between sequences (species) fit a matrix of pairwise distances between sequences
- Maximum Parsimony : find a phylogenetic tree that explains the data, with as few evolutionary changes as possible.
- Maximum likelihood : find a tree that maximizes the probability of the genetic data given the tree.
- Bayesian : find a tree that maximises the probability of the tree given a prior probability distribution of parameters.

Maximum Parsimony

Maximum Parsimony minimises the total number of character-state changes to produce tree topology.



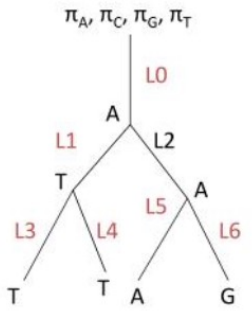
Maximum Likelihood

- Maximum Likelihood estimates the tree topology that optimises the likelihood of data given a defined model.
- The model considers that sequences are evolving according to a tree, based on:
  - a tree structure
  - branch lengths
  - parameters of the model of sequence evolution
- The likelihood of a set of data (i.e. sequences),

is the probability of the data, given a hypothesis

(i.e. a model of substitution with a set of

parameters θ and a tree τ)



- The desired probability is the product of the probabilities of the branches.
- L(tree) = L0 × L1 × L2 × L3 × L4 × L5 × L6

Bayesian inference

- Combines the information in the prior and in the data likelihood to create a posterior probability of trees
- Probability that the tree is correct given the data, the prior and the likelihood model.
- Posterior Probabilities are determined by integrating the product of the prior probabilities and the model over all possible parameter values.
- Uses Markov Chain Monte Carlo (MCMC) methods to searches parameter space and calculates the marginal estimation (volume under posterior probability curve).
- The MCMC algorithm follows a set of rules and takes ‘steps’ by making changes to the parameters including the topology and the branch lengths.

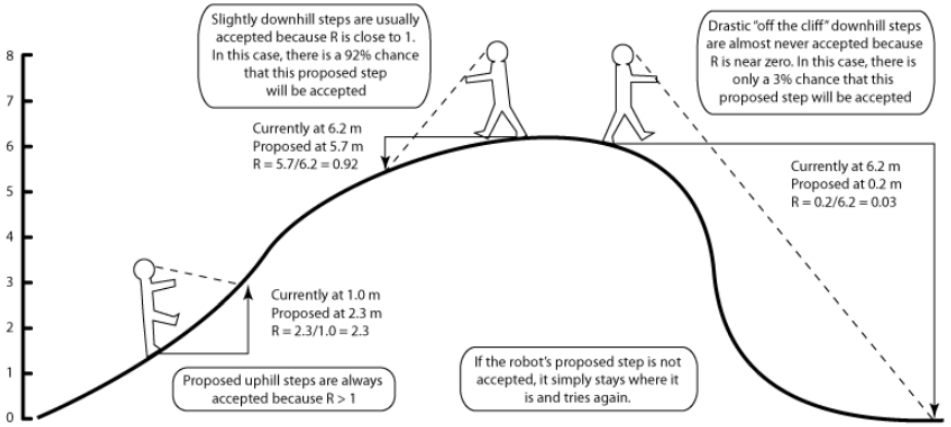
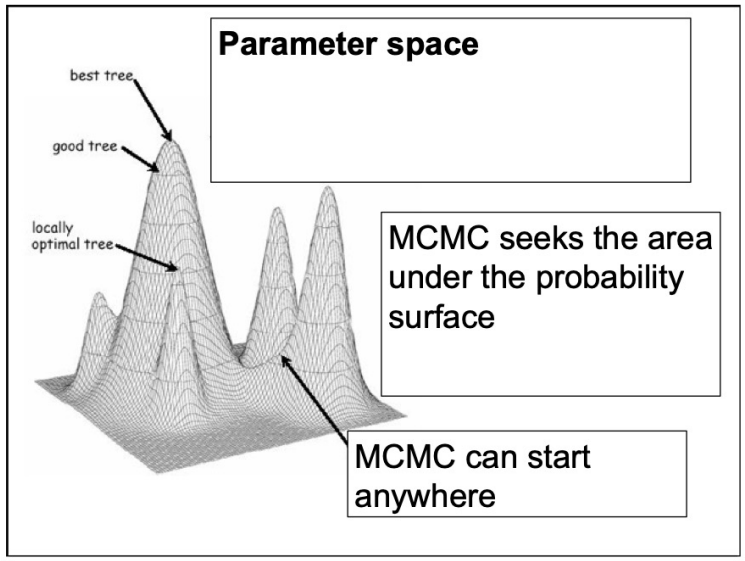
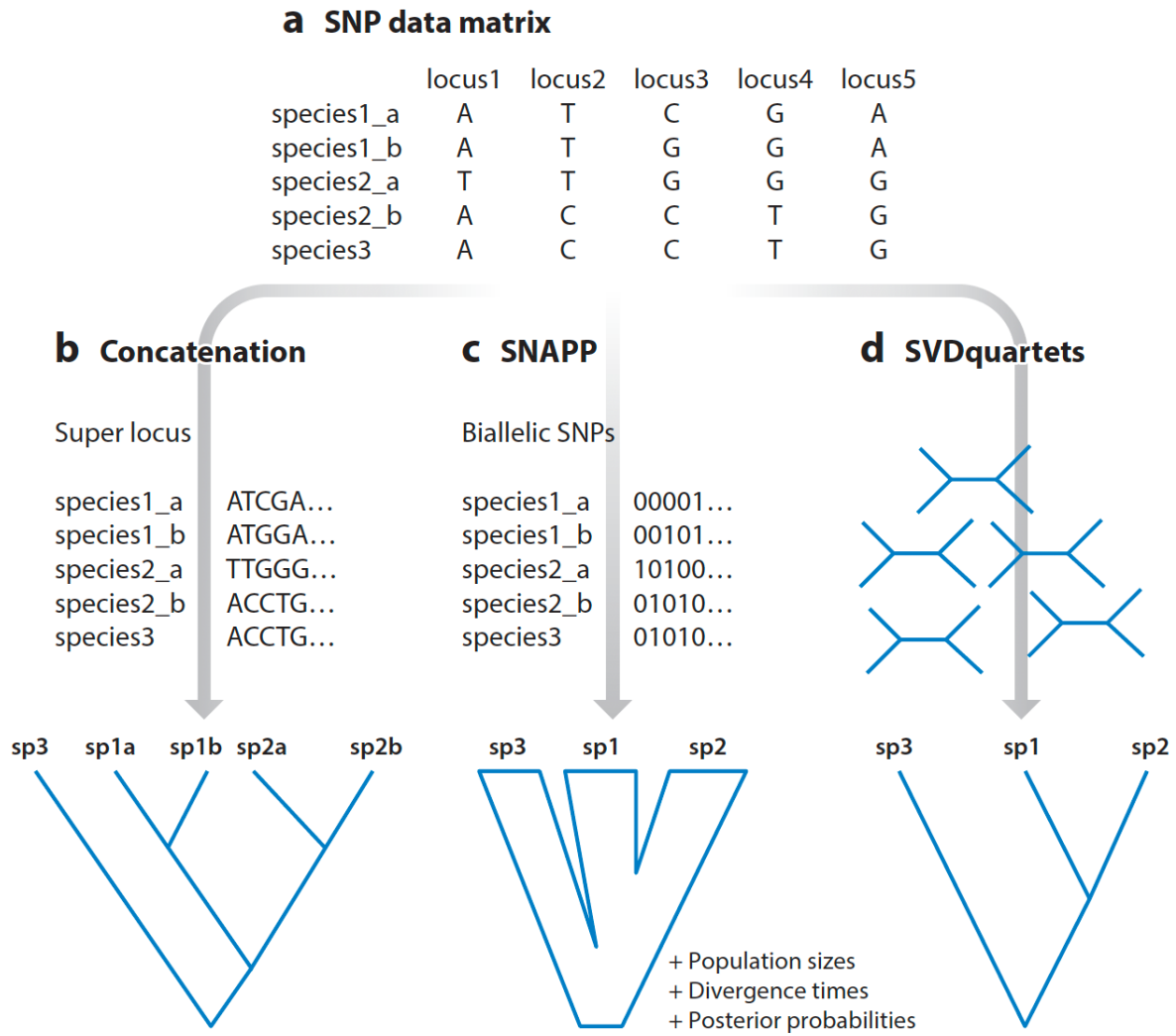


Illustration of MCMC method process (Lewis, 2011)



Concatenation vs Multi Species Coalescent



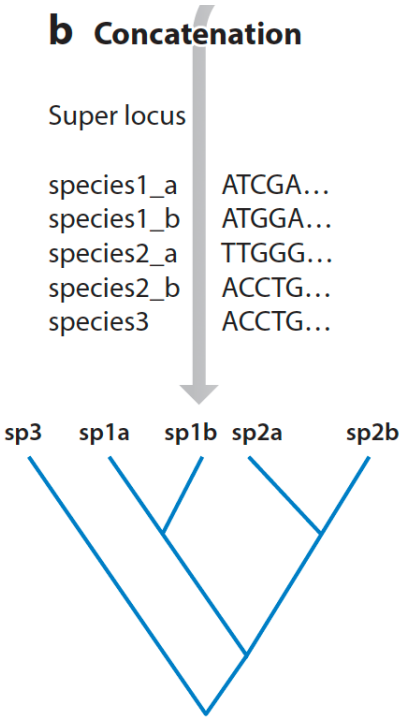
Leaché & Oaks Annu. Rev. Ecol. Evol. Syst. 2017. 48:69–84

Concatenation

- Stitching SNPs together to construct one “super locus”
- All data is analyzed simultaneously - **fast** and mostly accurate.
- Analysed using a variety of phylogenetic approaches
  - E.g., ML, MP, Bayesian inference
- Can be problematic - ignores incomplete lineage sorting

> *incorrect topology* when ILS is high

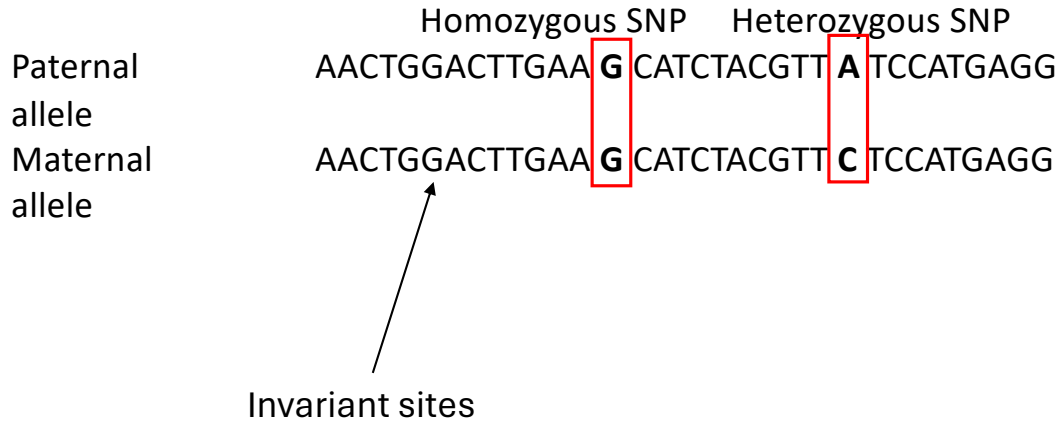
- Assumes all SNPs share the same coalescent history
- Can yield branch length errors (inflation) especially near the tips of the tree
- If you don’t take into account invariant sites > branch length errors



Concatenation

Studies suggest it is better to apply concatenation to the original sequence alignments from which the SNPs were extracted.

Removing constant sites results in acquisition bias that can inflate branch length estimates and, in extreme cases, produce an inaccurate phylogeny.

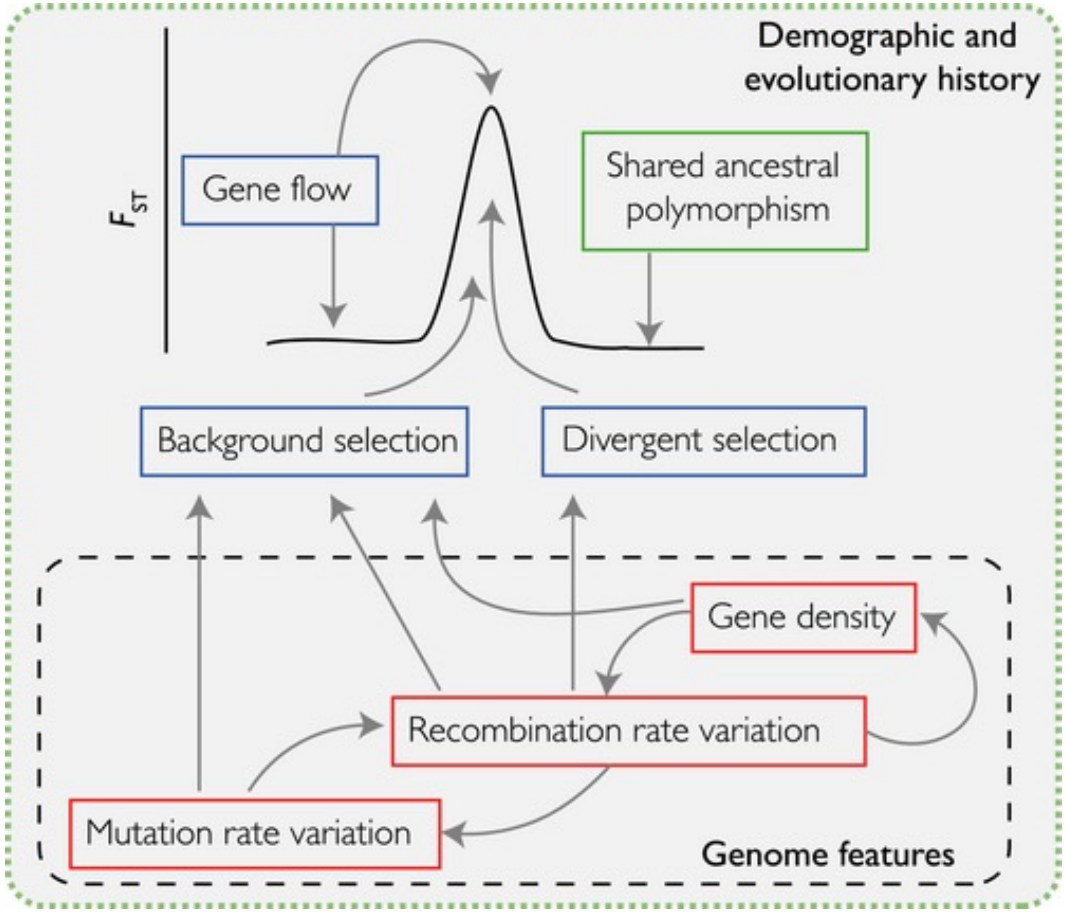


Leaché & Oaks Annu. Rev. Ecol. Evol. Syst. 2017. 48:69–84

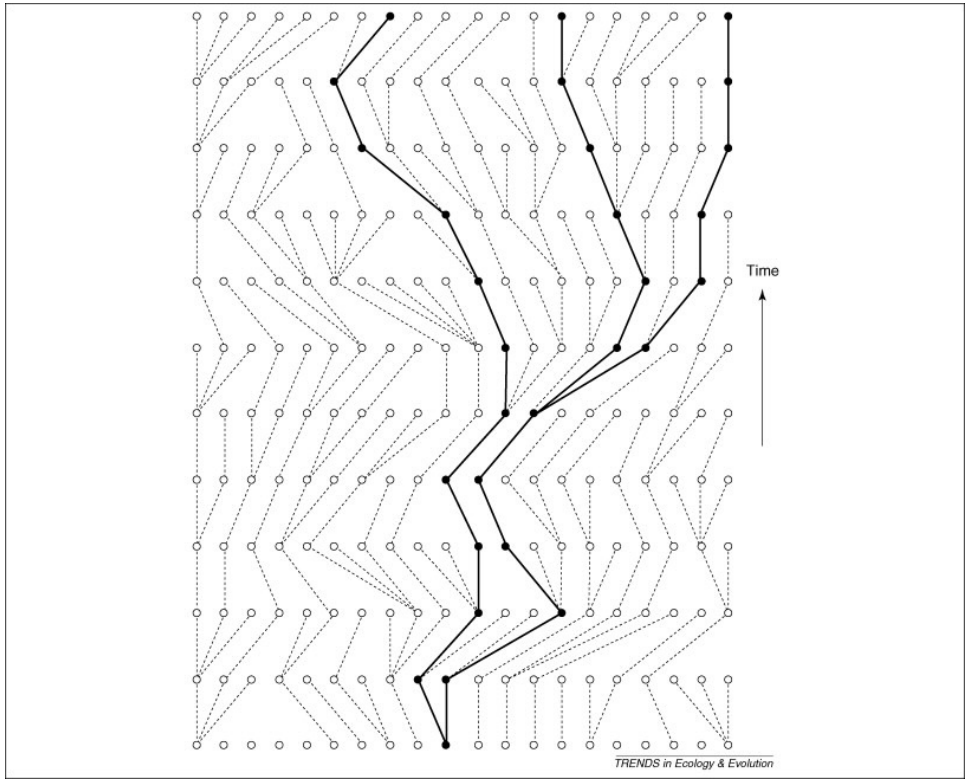
The genome is heterogeneous!

Gene tree ≠ species tree

- horizontal gene transfer
- gene duplication
- hybridization
- incomplete lineage sorting

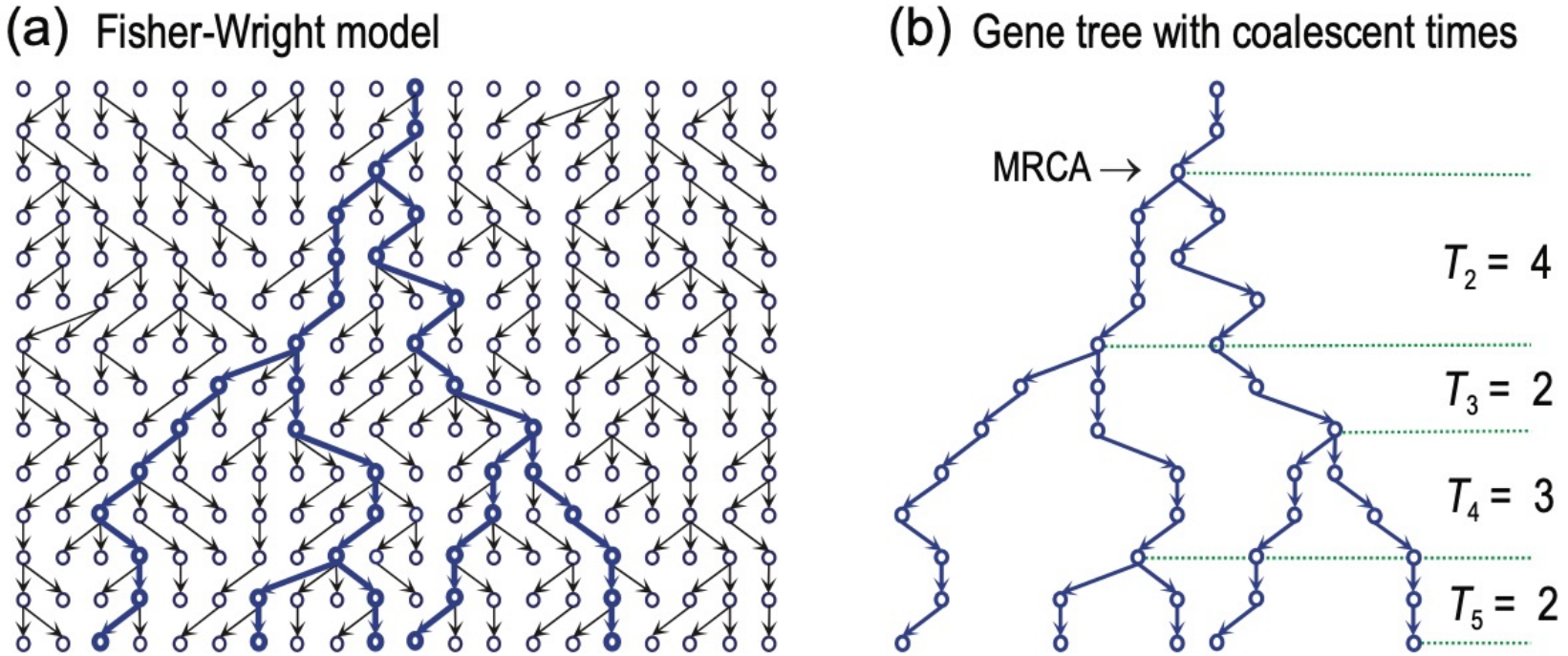


Coalescent history

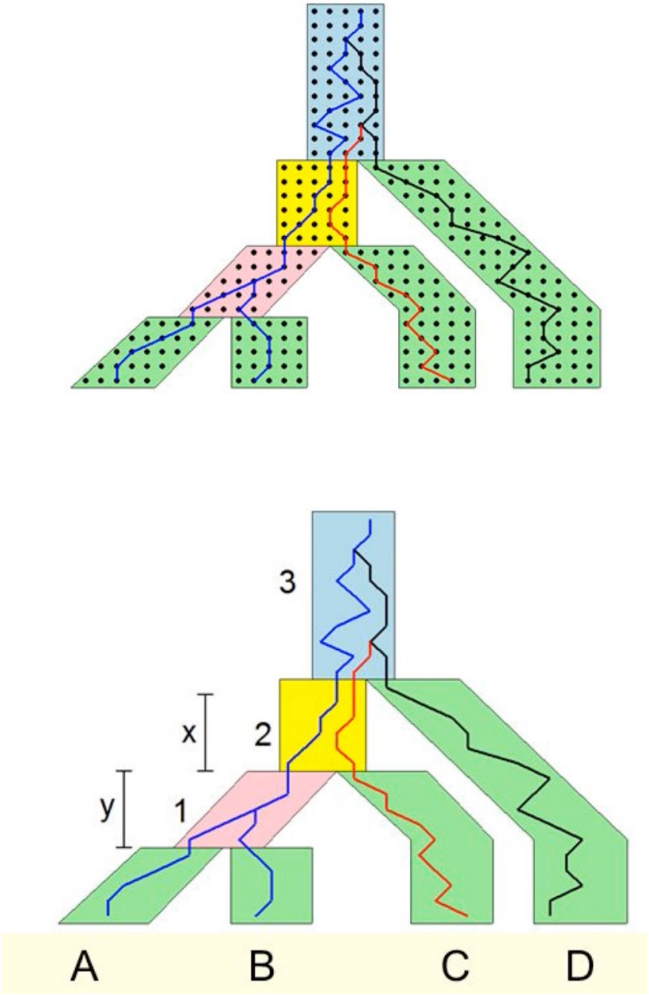


- The *coalescent* - mathematical concept representing the stochastic process of the genealogical history of a sample of DNA sequences
- It provides a way of looking back in time that allows ancestral inference

Coalescent theory - bridges pop gen <> phylo continuum



Whereas traditional population genetic models of drift and mutation describe changes in allele frequencies over generations in the *population*, the coalescent focuses on the *sample* and traces the genealogical history of lineage joining of the sampled sequences backwards in time. (Edwards et al. 2016)



MultiSpecies Coalescent

The multispecies coalescent (MSC) is an extension of the single-population coalescent to the case of multiple species.

- Integrates process of species divergences and the within-population process of drift and mutation

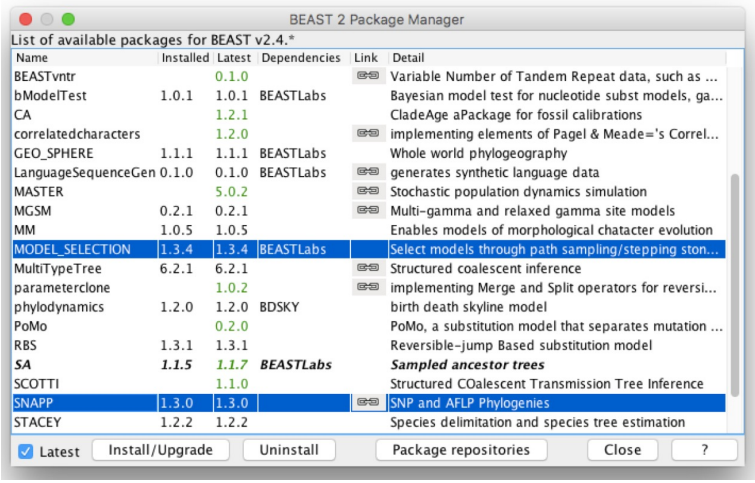
- Can estimate population parameters:
  - Species divergence times
  - Populations sizes (and changes – ancestors)
  - Rates of cross-species gene flow
  - Mutation rates
  - Delimit species boundaries

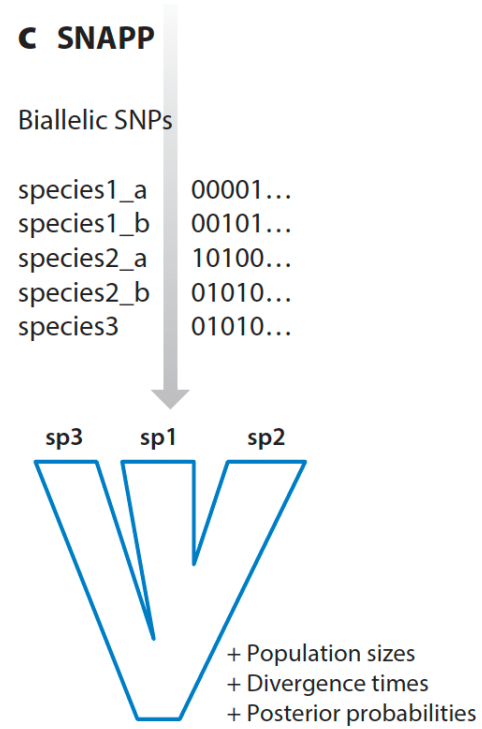
- Individual gene trees are estimated simultaneously or separately with a species tree as a means of estimating phylogenetic relationships

Full Bayesian Approach – SNAPP

- Jointly estimates the species-tree topology, speciation times and effective population sizes without first estimating gene trees for each locus
- SNAPP (Bryant et al., 2012) for biallelic SNP data
- Run through BEAST 2

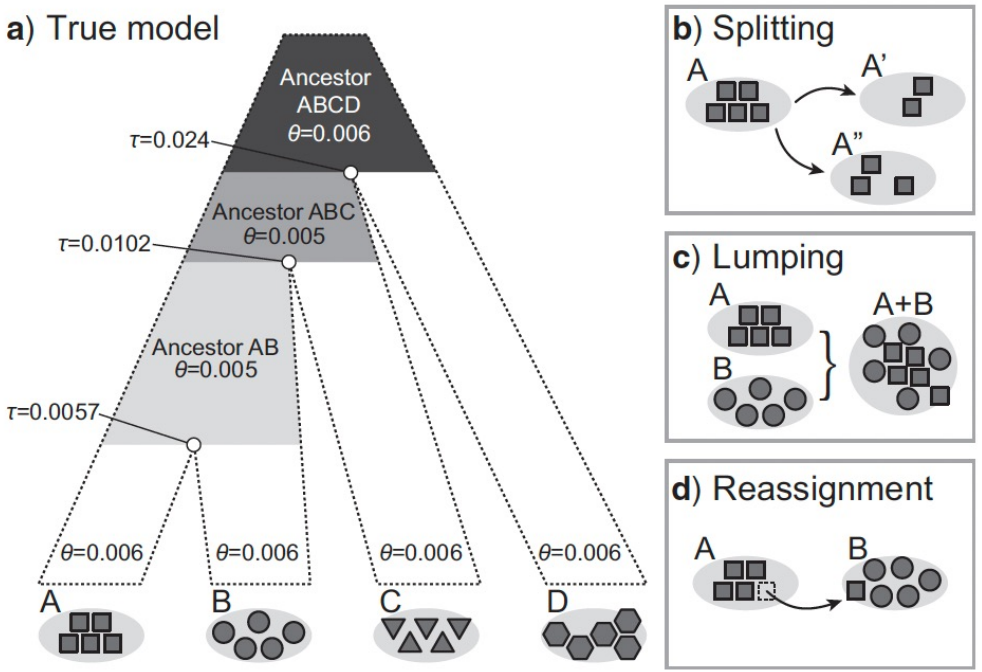
See tutorial http://evomics.org/wp-content/uploads/2018/01/BFD-tutorial.pdf





Species delimitation: SNAPP + BFD*

- Estimates species trees directly from biallelic markers - bypassing the need to explicitly integrate or sample the gene trees at each locus.
- Estimates the probability of allele frequency change across ancestor/descendent nodes
- Marginal likelihoods for alternative species delimitation models are not directly comparable > so add proportionality constants that bring marginal likelihoods to the same scale for comparing competing species assignments with Bayes factors.



Leaché et al. 2014 Syst. Biol.

Example

Species Delimitation: BFD*

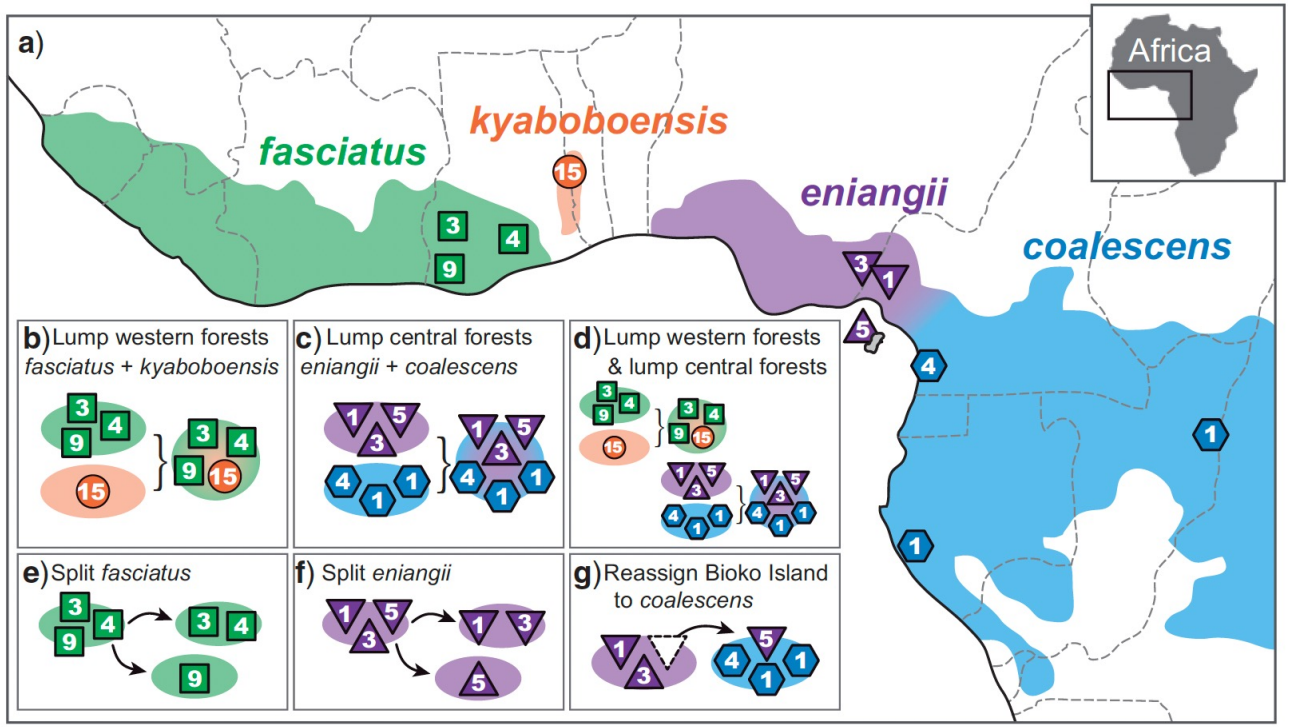


TABLE 2. Empirical results for BFD* species delimitation in the *Hemidactylus fasciatus* complex

		129 SNPs			1087 SNPs		
Model	Species	ML	Rank	BF	ML	Rank	BF
a. Current taxonomy	4	−1673.4	2	–	−12890.3	2	–
b. Lump western forests	3	−1724.2	5	+101.5	−15024.5	6	+4268.3
c. Lump central forests	3	−1788.0	6	+229.2	−14094.0	5	+2407.4
d. Lump western & central forests	2	−1842.9	7	+339.0	−16190.4	7	+6600.3
e. Split *fasciatus*	5	−1713.2	4	+79.7	−13088.0	3	+395.5
f. Split *eniangii*	5	−1625.9	1	−95.1	−12615.3	1	−550.0
g. Reassign Bioko Island	4	−1712.6	3	+78.4	−13434.4	4	+1088.2

Leaché et al. 2014 Syst. Biol.

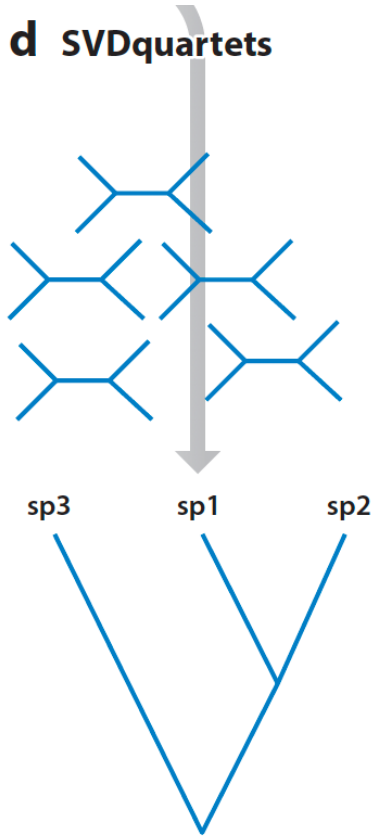
Table 1. Studies evaluating the robustness of species tree phylogenetic methods to various genetic forces and sampling schemes

Topic	Conclusions/comments[a]
General violation of multispecies coalescent model[104]	• Claims the majority of multilocus sequence data sets are a poor fit to the multispecies coalescent model, although much of the violation stems from fit of substitution model or unknown sources on a minority of genes.
Gene flow[73,105]	• The coalescent method is robust to low levels of gene flow • Concatenation performs poorly relative to the coalescent methods in the presence of gene flow. • Gene flow can lead to overestimation of population sizes and underestimation of species divergence times in species trees.
Sampling/mutation[33,74,94]	• Increased sampling of individuals per species can significantly improve the estimation of shallow species trees. • Sampling more individuals does not significantly improve accuracy in estimating deep species trees. Adding more loci can improve the estimation of deep relationships.
Recombination[39,106]	• Mutational variance is a major source of error in estimates of species trees. • Recombination has minor effect on species tree estimation except on extremely short species trees. • The negative effects of recombination can be easily overcome by increased sampling of alleles
Missing data[19,63,67]	• Missing data can decrease the support of species tree estimates • Missing data can significantly affect the accuracy of species tree estimation • Species tree methods are "remarkably resilient" to missing data[68]
Taxon sampling[15]	• Compared to concatenation, coalescent methods are more robust to poor taxon sampling
Long-branch attraction[29]	• Species tree methods are more resilient to the effects of long-branch attraction than concatenation methods
Random rooting of gene trees[36,37]	• Misrooting of gene trees can mimic the coalescent process
Other[80]	• Anomalous gene trees are unlikely to pose a significant danger to empirical phylogenetic study, in part because species trees in the anomaly zone are likely to be rare.

Influence of gene flow, sampling, missing data

Analyses that avoid gene trees

- Infer species trees without the requirement of first (or simultaneously) estimating gene trees for each locus
- Use full-likelihood approaches to inferring species trees directly from SNP data
- They analytically integrate gene trees out of the equation.



Quartet Methods - SVDquartets

- Algebraic geometry of site pattern probabilities on trees under continuous-time Markov chain models of sequence evolution.
- Coalescent-based method > multi-locus unlinked single-site data, infers the quartet trees for all subsets of four species, and then combines the set of quartet trees into a species tree using a quartet amalgamation heuristic
- SVDquartets (singular value decomposition for quartets; Chifman & Kubatko 2014)
- Several advantages over other coalescent methods for estimating species trees
  - statistically robust under very general models of molecular evolution and the coalescent
  - statistically consistent estimator of the species tree (variation in evolutionary rates and effective population sizes; Long & Kubatko 2017)
  - fast, produces a species tree with estimates of uncertainty from a bootstrapping procedure
  - subsampling quartets from a SNP data matrix - missing data largely circumvented since some data will typically be present at a locus for each of the four samples included in the quartet.
  - potential to provide more information for relationships deep in a phylogeny compared to those at the tips

Accounting for migration - TreeMix

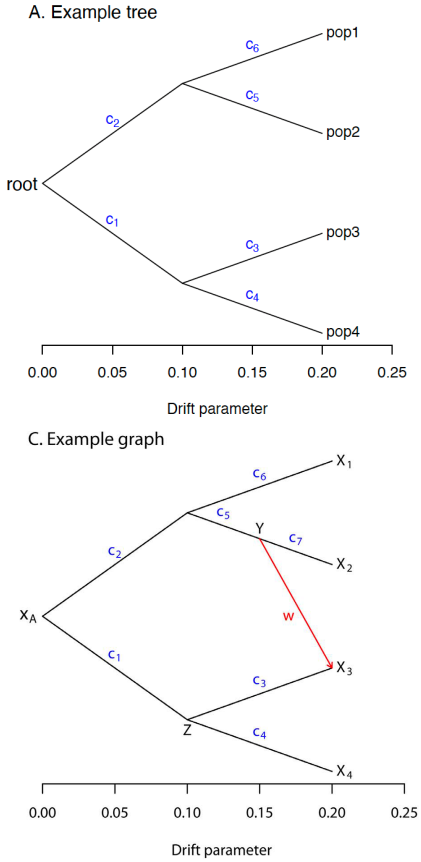
- Population/species history is often complex
  - population size change, gene flow
- TreeMix (Pickrell & Pritchard 2012)

- statistical framework for building population trees and testing for the presence of gene flow between diverged populations

- Relationship between populations is represented as a graph, allowing models of both population splits and gene flow
- Represent the topology of relationships between populations, rather than the precise times of demographic events

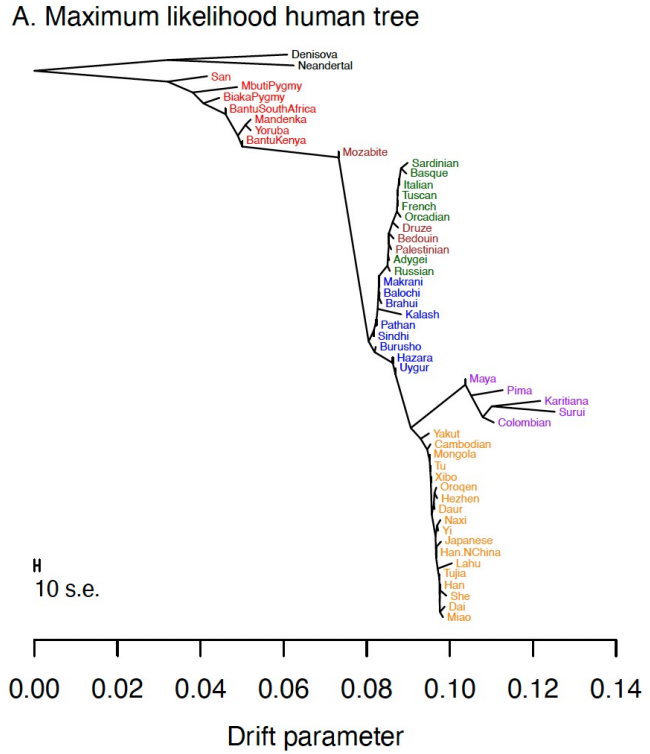
Step 1 : generate a maximum likelihood tree

Step 2: identify populations that are poor fits to the tree model, and model migration events involving these populations



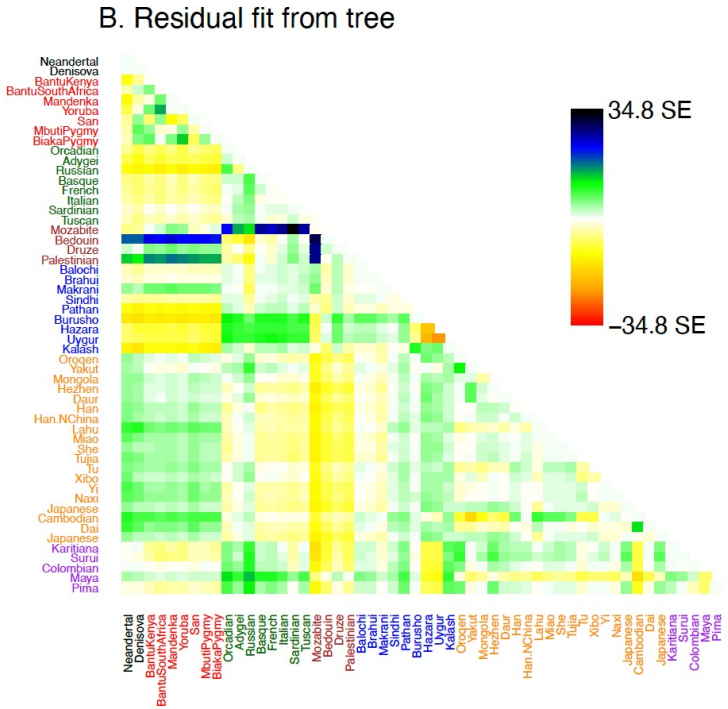
TreeMix Example

Humans



Pickrell & Pritchard (2012)

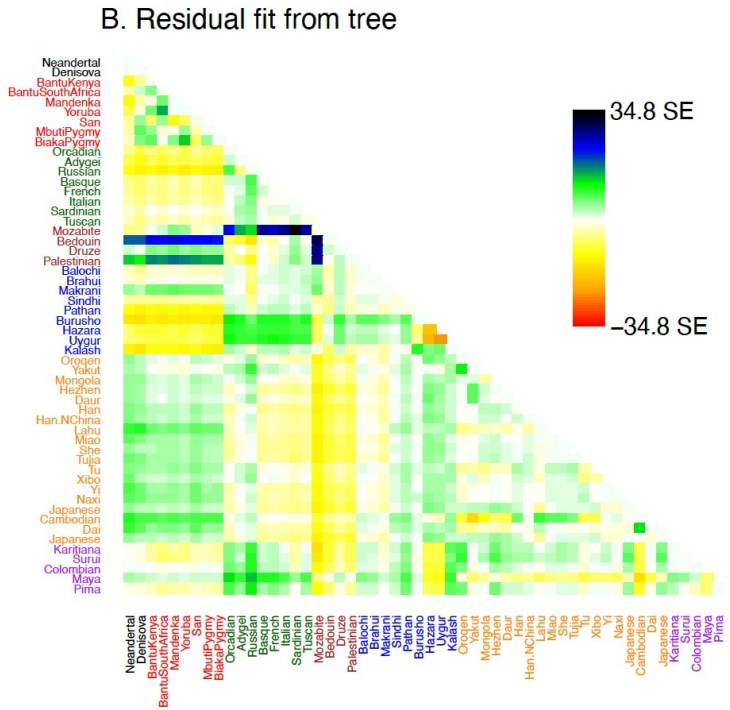
Residual distance between each pair of populations i and j divided by the average standard error across all pairs.

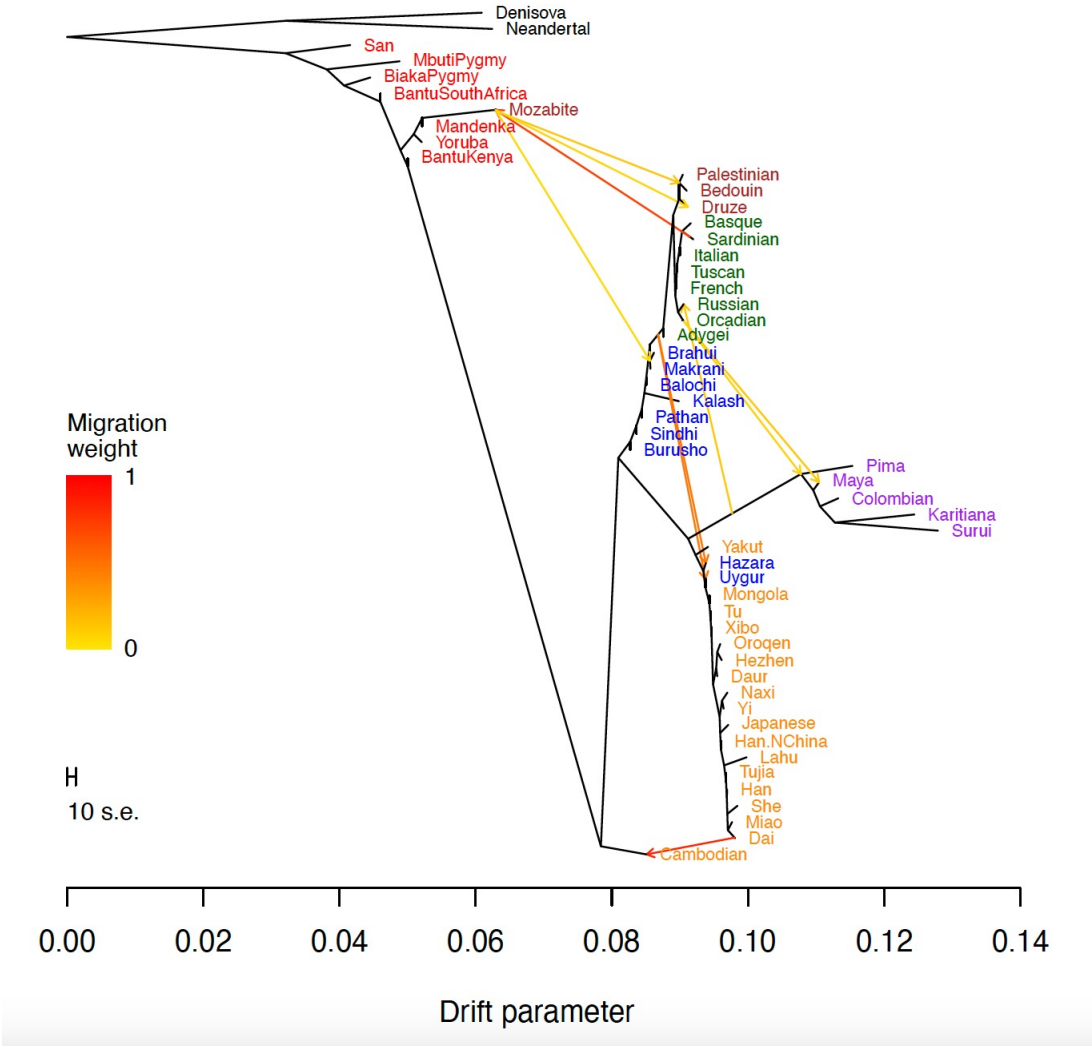


Residuals > 0 represent populations more closely related to each other in the data than in the best tree, and thus are candidates for admixture events.

TreeMix Example

Human migration





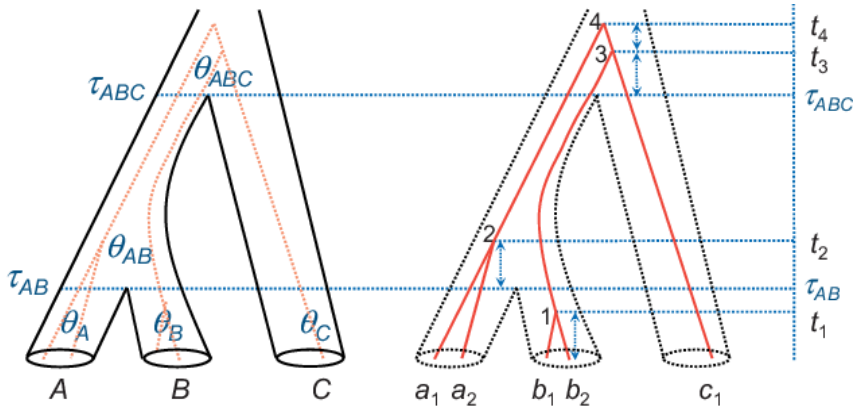
Pickrell & Pritchard (2012)

Putting divergence estimates on species trees

- Fossil or node calibrations OR estimate from the data
- ML estimates of branch lengths from a concatenated alignment

  > systematic errors - fails to account for variation in gene genealogies & associated gene divergence times

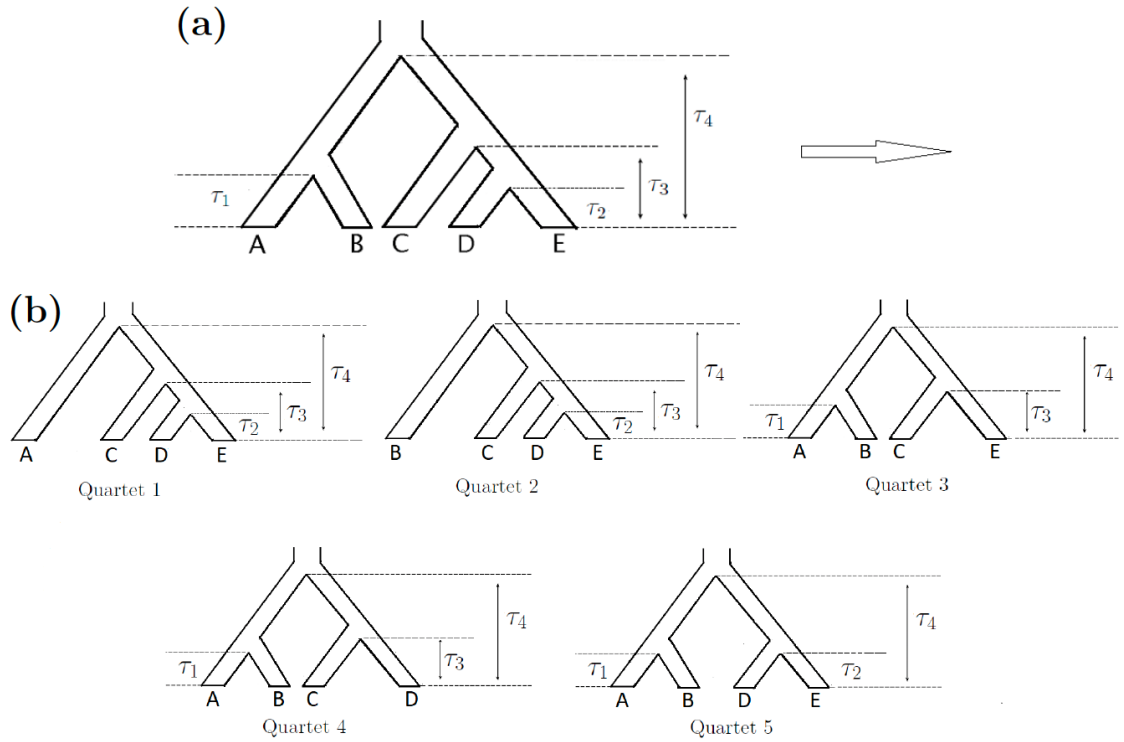
- MSC methods estimate either speciation times or internal branch lengths in addition to estimating the species-tree topology
- Calculate divergence estimates using maximum a posteriori estimator based on composite likelihood (MAPCL)



Flouri et al. 2018 Mol. Biol. Evol. 35(10):2585–2593

qAge - SVDquartets

- qAge - quartet-based method that operates under the MSC assumptions
- It can be used to estimate speciation times on trees obtained using any method!
- Compute composite likelihoods based on the fit of observed site-pattern probabilities to their expectations under the MSC model.
- Estimate tau and theta via Bayesian maximum a posteriori (MAP) estimation
- Assumptions (i) JC69 substitution model, and (ii) effective population sizes are constant throughout the tree Chifman and Kubatko (2015)



Peng et al. 2022, *Bioinformatics*, 38(23): 5182–5190

Genome-wide SNPs

Phylogenetics

- Bifurcating tree to represent the pattern of ancestry and descent
- Network analysis of relationships
- Spatially comprehensive sampling, but low replication at each site
- Applies from shallow to deep divergences
- Historically restricted to mtDNA (this is rapidly changing)

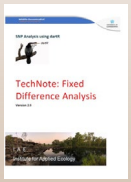
Lineage Diversity
(lineages on independent evolutionary trajectories)

Species Delimitation

Phylogeography

Popgen

- Contemporary and recent processes
- Population differentiation
- Geneflow and admixture
- Population structure (cryptic lineages?)
- Relationships (including kinship)
- Population assignment



Genome-wide SNPs

Phylogenetics

Popgen

Lineage Diversity
(lineages on independent evolutionary trajectories)

Species Delimitation

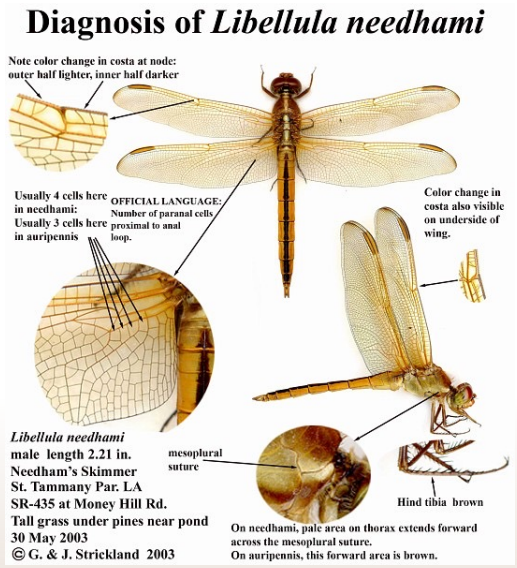
Phylogeography

Diagnosable Lineages

What is Diagnosability?

All individuals in a group of individuals possess a heritable character state that is absent in all individuals outside that group

Morphology



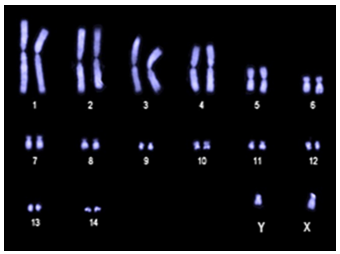
Why is Diagnosability Important?

Diagnosability cannot be sustained in the presence of admixture.

Provides evidence that two putative lineages are on independent evolutionary trajectories
(via either geographic or reproductive isolation)

Absence supports the Status Quo
(no evidence of independent evolutionary trajectory)

Chromosomal incompatability



Fixed Allelic Differences

"Putative Taxon"	Allele Frequency
Group A	0
Group B	100

Diagnosability → implies → Independent Trajectory

Analysis

Under the Hood

- Sample comprehensively
- Collect multiple individuals from each site (n ≥ 10)
- Remove sites with evidence of admixture (Bulloo River, Lake Dunn in Cooper Creek)
- Generate a phylogeny
- Identify diagnosable groupings
- Overlay diagnosable groups on the phylogeny
- Overlay diagnosable lineages on the physical map

Chambers and Hillis (2020, Systematic Biology, 69:184–193)

Adequate estimate of allele frequency profiles (2n = 20)

gl.pcoa(gl); gl.pcoa.plot(); gl.nhybrids()

Later in this Session – Sally Potter

gl.fixed.diff(); gl.collapse()

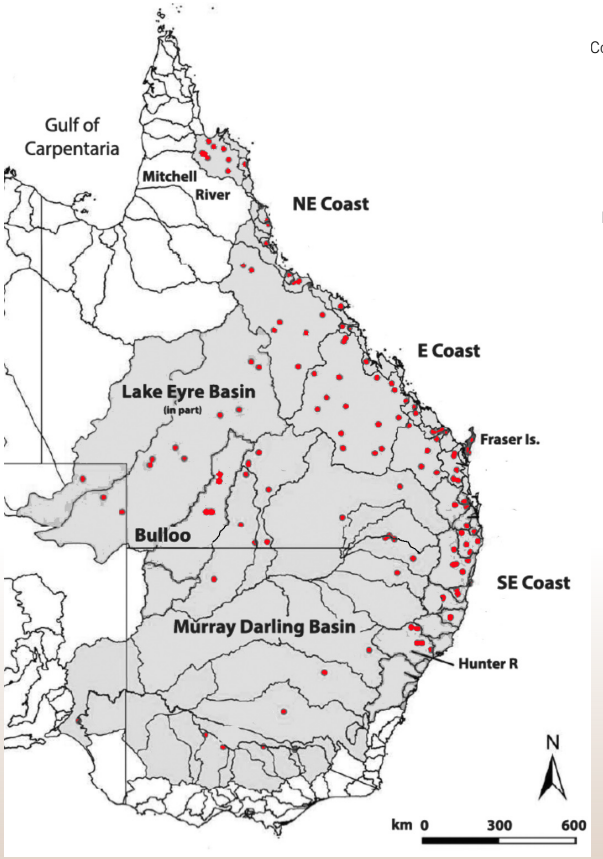
Interpretation

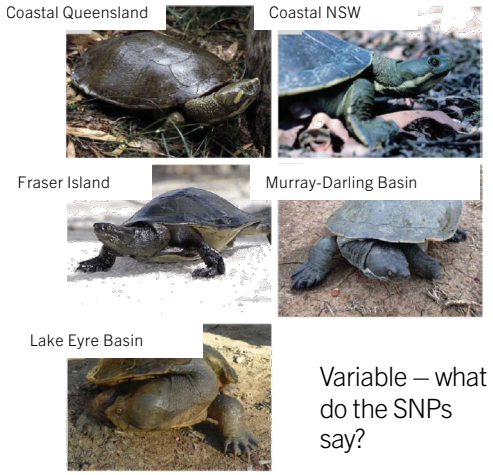
Refer Unmack et al. 2022. Biol. J. Linn. Soc. 135:117–137
Georges et al. 2018. Molecular Ecology 27:5195-5213

Analysis

Under the Hood

- Sample comprehensively
- Collect multiple individuals from each site (n ≥ 10)
- Remove sites with evidence of admixture (Bulloo River, Lake Dunn in Cooper Creek)
- Generate a phylogeny
- Identify diagnosable groupings
- Overlay diagnosable groups on the phylogeny
- Select diagnosable lineages as the representation of lineage diversity





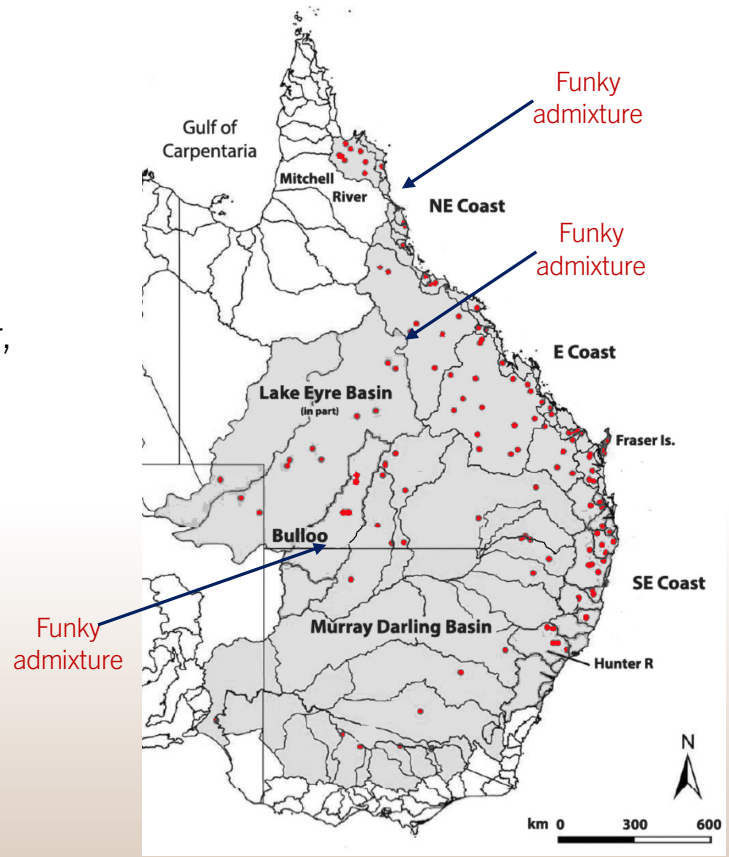
Variable – what do the SNPs say?

Target: n = 10 at each locality

Analysis

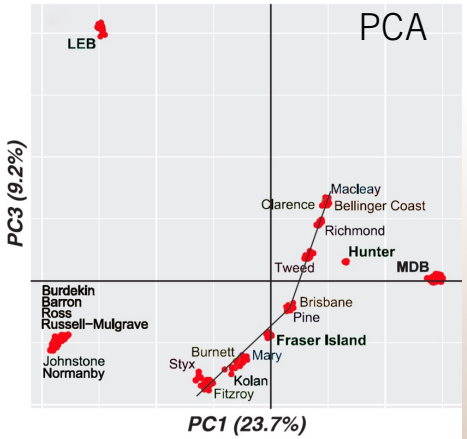
Under the Hood

- Sample comprehensively
- Collect multiple individuals from each site (n ≥ 10)
- Remove sites with evidence of contemporary admixture (Bulloo River, Lake Dunn in Cooper Creek)
- Generate a phylogeny
- Identify diagnosable groupings
- Overlay diagnosable groups on the phylogeny
- Select diagnosable lineages as the representation of lineage diversity



Info only – not to follow

```
# Identify admixture
pc <- gl.pcoa(gl)
gl.pcoa.plot(pc, gl)
gl.nhybrids()
# Drop recalcitrant pops
gl <- gl.drop.pop()
# Re-examine PCA
pc <- gl.pcoa(gl)
gl.pcoa.plot(pc, gl)
```

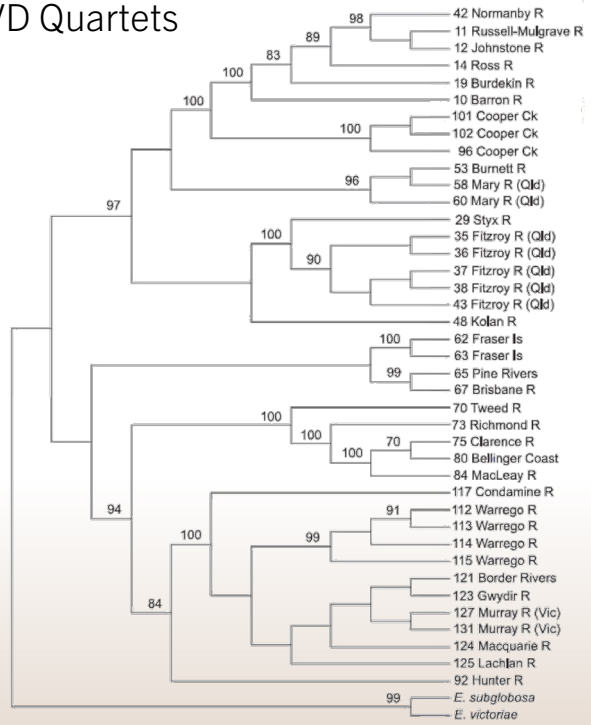


Analysis

Under the Hood

- Sample comprehensively
- Collect multiple individuals from each site (n ≥ 10)
- Remove sites with evidence of admixture (Bulloo River, Lake Dunn in Cooper Creek)
- Generate a phylogeny (SVDquartets)
- Identify diagnosable groupings
- Overlay diagnosable groups on the phylogeny
- Overlay diagnosable lineages on the physical map

SVD Quartets



Info only – not to follow

```
# Identify admixture
pc <- gl.pcoa(gl)
gl.pcoa.plot(pc, gl)
gl.nhybrids()
# Drop recalcitrant pops
gl <- gl.drop.pop()
# Re-examine PCA
pc <- gl.pcoa(gl)
gl.pcoa.plot(pc, gl)
# Generate nexus file
gl2svdquartets()
Run paup*
```

Fixed Difference Analysis

Fixed Difference Analysis

Under the Hood

- Sample comprehensively
- Collect multiple individuals from each site (n ≥ 10)
- Remove sites with evidence of admixture (Bulloo River, Lake Dunn in Cooper Creek)
- Generate a phylogeny (SVDquartets)
- Identify diagnosable groupings
- Overlay diagnosable groups on the phylogeny
- Overlay diagnosable lineages on the physical map

```
D <- gl.fixed.diff(gl)

D <- gl.collapse(D)   Run it recursively until no further improvement

Group:10_NECBarrCair+
[1] "10_NECBarrCair" "11_NECRussEube" "12_NECJohnWari" "14_NECRossRoss" "19_NECBurdMist"

Group:101_LEBCoopEulb+
[1] "101_LEBCoopEulb" "102_LEBCoopCull" "96_LEBCoopAvin"

Group:113_MDBWarrDart+
 [1] "113_MDBWarrDart" "114_MDBWarrSanf" "115_MDBWarrBiny" "117_MDBCondArch" "121_MDBBordGoon"
 [6] "123_MDBGwydBing" "124_MDBMacqCudg" "125_MDBLachForb" "127_MDBMurrAlbu" "131_MDBMurrMBri"
[11] "MDB_Warrego"

Group:35_ECFitzFair+
[1] "35_ECFitzFair" "36_ECFitzCarn" "37_ECFitzKorc" "38_ECFitzTaro" "43_ECFitzAlli" "58_ECMaryBoru"
[7] "53_ECBurnBara" "48_ECKolaKola" "60_ECMaryTiar"

Group:67_SECBrisWive+
[1] "67_SECBrisWive" "70_SECTweeUki"  "73_SECRichCasi" "SECClarence"    "84_SECMaclGeor"
[6] "SECBellinger"

Group:29_ECStyxStyx+
[1] "29_ECStyxStyx"  "35_ECFitzFair+"

No further amalgamation of populations at fd <= 0
```

Fixed Difference Analysis

Under the Hood

- Sample comprehensively
- Collect multiple individuals from each site (n ≥ 10)
- Remove sites with evidence of admixture (Bulloo River, Lake Dunn in Cooper Creek)
- Generate a phylogeny (SVDquartets)
- Identify diagnosable groupings
- Overlay diagnosable groups on the phylogeny
- Overlay diagnosable lineages on the physical map

D <- gl.fixed.diff(gl)

D <- gl.collapse(D) Run it recursively until no further improvement

Example only -- subset

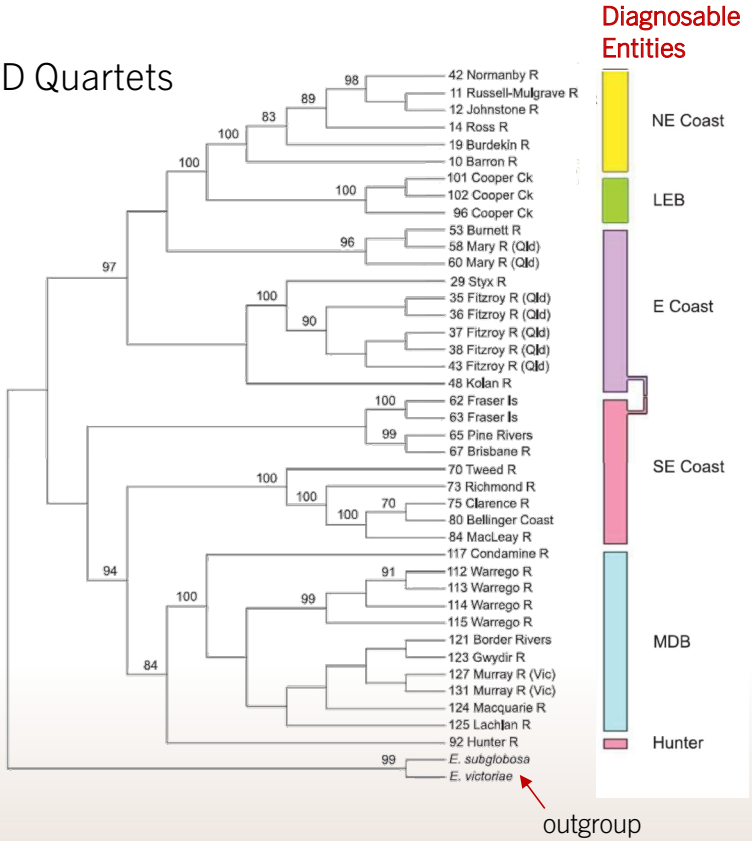
	NECNormanby	10_NECBarrCair+	101_LEBCoopEulb+	113_MDBWarrDart+	29_ECStyxStyx+	62_FRAFrasBowa	63_FRAFrasMack	67_SECBrisWive+	SECHunter	EmsubRope
10_NECBarrCair+	1									
101_LEBCoopEulb+	469	181								
113_MDBWarrDart+	385	66	333							
29_ECStyxStyx+	97	7	160	10						
62_FRAFrasBowa	535	146	508	103	11					
63_FRAFrasMack	650	205	610	138	16	3				
67_SECBrisWive+	311	34	258	7	1	16	22			
SECHunter	1261	609	1161	413	366	750	856	358		
EmsubRope	2380	581	2346	1784	1607	2133	2234	1714	2724	
EmvicVict	2489	593	2435	1871	1697	2227	2336	1786	2817	760

D <- gl.fixed.diff(D, test=TRUE)
gl <- gl.merge.pops(D$gl, poplist=c(###))

Info only – not to follow

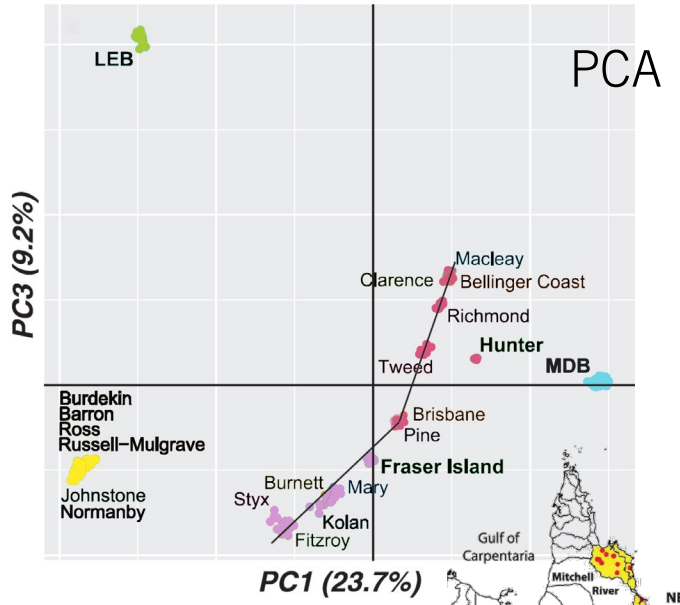
```
# Identify admixture
pc <- gl.pcoa(gl)
gl.pcoa.plot(pc, gl)
gl.nhybrids()
# Drop recalcitrant pops
gl <- gl.drop.pop()
# Re-examine PCA
pc <- gl.pcoa(gl)
gl.pcoa.plot(pc, gl)
# Generate nexus file
gl2svdquartets()
Run paup*

# Generate FD matrix
D <- gl.fixed.diff(gl)
# Collapse on no FD
D <- gl.collapse(D)
D <- gl.collapse(D)
D <- gl.collapse(D)
D <- gl.fixed.diff(D,
      test=TRUE)
```

SVD Quartets

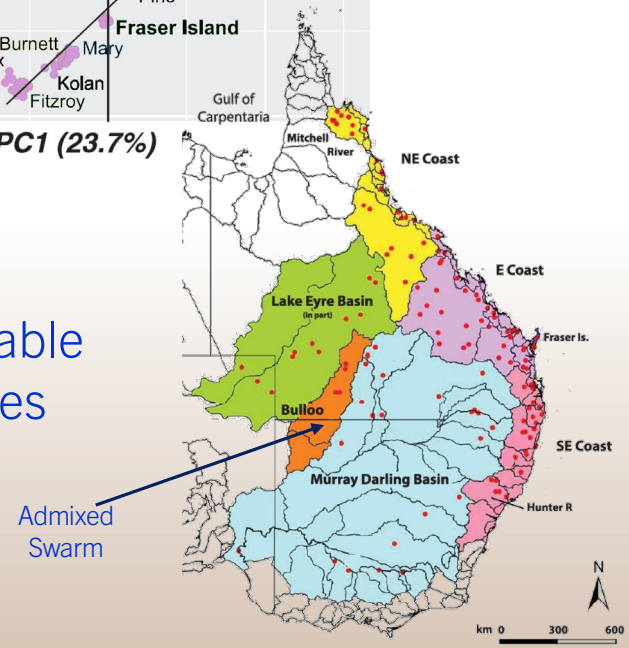


Which of these lineages (and their descendant clades) are on independent evolutionary trajectories?

PCA



Diagnosable Lineages

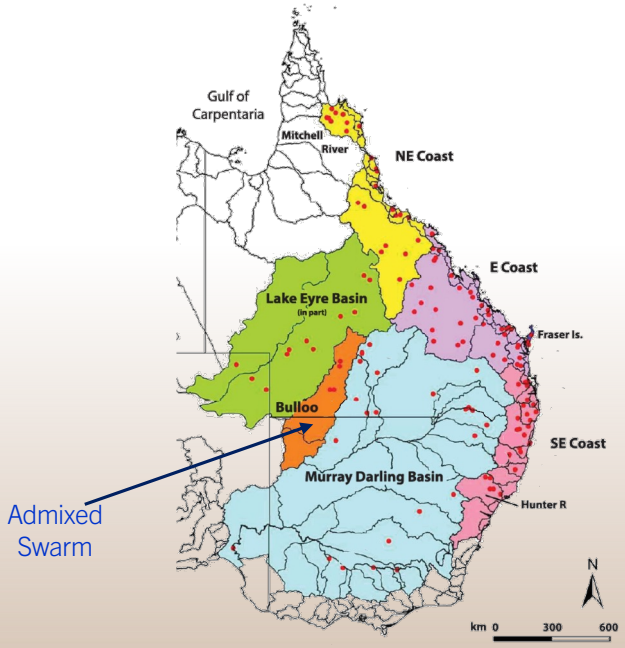


Admixed Swarm

Genome-wide SNPs

Phylogenetics

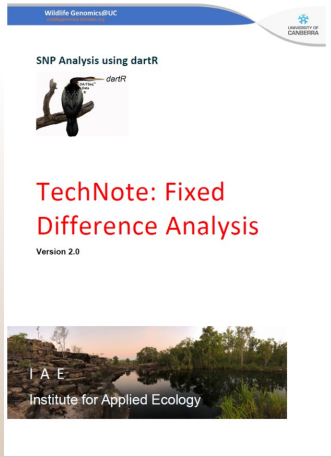
Popgen



Admixed Swarm

Lineage Diversity
(lineages on independent evolutionary trajectories)

Phylogeny
+
Diversity
=
Substantive Diagnosable Lineages



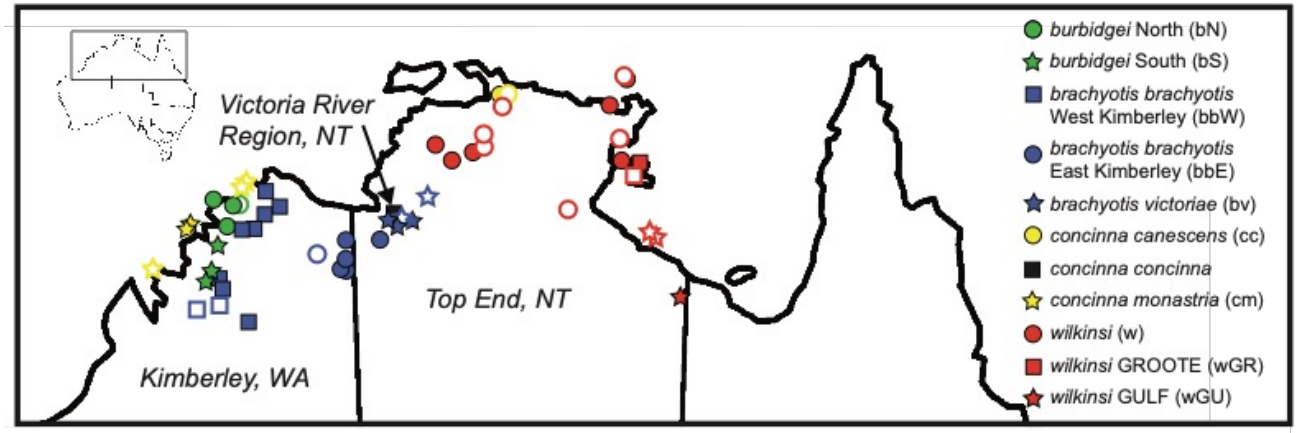
Refer

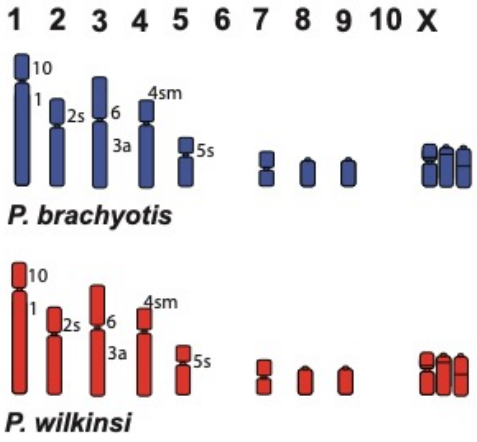
Unmack et al. 2022. Biol. J. Linn. Soc. 135:117–137

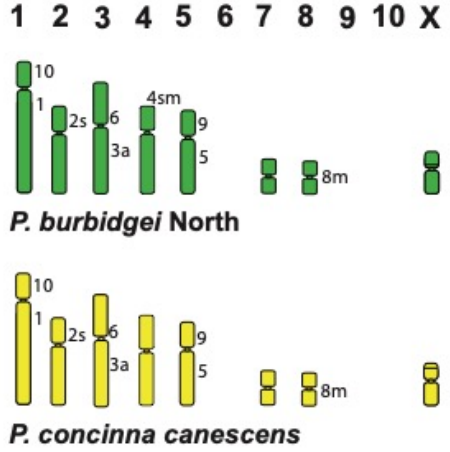
Georges et al. 2018. Molecular Ecology 27:5195-5213

IQTree2

The *brachyotis* group







Input file – phylip sequence format

```
 22 749814
ABTC103495_bbrachEK
TCCCGAGTCTCCTCCGACAATGGGCAGCGGTTATTGGACATGATTGAGACCCTGAAAATGGAGCACCACCTGGAGGCCCA
CAAGAATGTGCATGCCAACGGGCACTTGAGCATCAGTGCTGATGGCCCAGTGCCCGACAGCAGCTTCATCAGCAAGGCCA
AACACAACGGCGTGCCCAATGGCTACCAGGGGGGGTTGGAAATGATCAAGATCCCCATACCTGAACCCGAGCGCTCCCAG
TACCCCCGGGAGTGGGGCAAGACTCTGCTGGCTTTTCTGTATGCCCTCTGCTGCTTTGTGTTCACCACAGTGACCATCTC
GTTCGTCCATGAGCGAGTGCCTCCCAAGGAGATGCAGGCACCCCTACCGGACACCTTTTTTGACCATTTTAACCGGTATG
TCATTGCAGGGAGCCTGCCCCATCTGGGCCAGTGTCTGACTCGGTACTGGGAACAGAAGAAGTGCATGGCTCCTGGCTGT
GAACCCCAGGCTGTGCAGCGCATGATGGCTGCCCTCAAGCCCTATGTGTATGGCCAGAGCCTGGCAGGGGCTGGCGGTGG
CGGCTTCCTCTACCTGTTGACTAAACAGCCCCGGCAGAAAGCTACAAGTTTTCCTGTAGAGAGACTGAAGGAAGAAGCAA
AATGTTCAATCTGTCTGGACTTCTTAAAGGATGCAGTCAGTGTGGACTGTGGCCACAGTTTCTGCCTCCAGTGCATCACT
GAATTCTGTGAGAAATCAGACAGTTCTCAAGGGAATGTCTACTCCTGCCCRCAGTGCCGGAGTCAATTTCAACAGSATAG
CTTTCGTCCCAATAGRCAACTGGCCAGCATGGTGGAGAGCATCAAGCAGTTGAGTTTAAGCTCCGGTACAAAGAAAACCT
GGTTATGTGAGATTCATAAAGAGGAGCTGATCCATTTTTGTGAAGATGACCAAATCCCATTGTGCTGGATCTGTGACACT
TCCCCAGAGCACAGAACCCACAAGGCCCACCCCAAACTCCTCCTATCGGAAGATTACCAGCGGGTCCAGTTCTCCTACAA
ATGGCAGACCTCACCAGACAGTCCACAACGTTTTGATAGGGCCATCTGTGTCCTAGCCCAGGATGGCTTTGTGGGAGGGA
GGCACACGTGGGTGATAACTGTGGACCTGGCCCATGGGGGCAGCTGCACTGTGGGTGTGGTGAACCAGGATGTCCGACGG
AAAGGAGAACTCAGGCTTAGGCCAGAGGAAGGGATATGGGCAGTGCGATTGGCCTGGGGTTTTGTTTCAGCCTTGGGCTC
CTTCCCTACCAGGCTGTCCCTGGAAGAGAATCCCAGGCAAATCCGTGTCTCCCTTGATTATGAAGTAGGCTGGGTGACAT
TTAGCAATGCTGYCACTCAAGAGCCCATTTATACCTTCACTGCACCCTCAGTCCTGGGGTCCTGGAAGGTGATCAGCCCT
GAAAATGTTCAGGGTGTCAAAGGCTCCTGTCTGGTCATCCCCTGCACCTTTAGCTTCCCTAGTGATGTGAAAACAGACRA
```

Choose the substitution model

If you do not know which model is appropriate for your data, you can use ModelFinder to determine the best-fit model

-m MFP + ASC

ModelFinder compares log likelihoods of trees and chooses the model that minimises the BIC score (Bayesian Information Criterion). Can also use AIC & AICc – corrected Akaiki Information Criterion.

You do not know the constant (invariant sites) so this can bias the result. You can run an ascertainment bias correction **+ASC** to correct for this.

Bootstrapping

- To overcome the computational burden required by the nonparametric bootstrap, IQTREE introduces an ultrafast bootstrap approximation (UFBoot)

(Minh et al., 2013; Hoang et al., 2018)

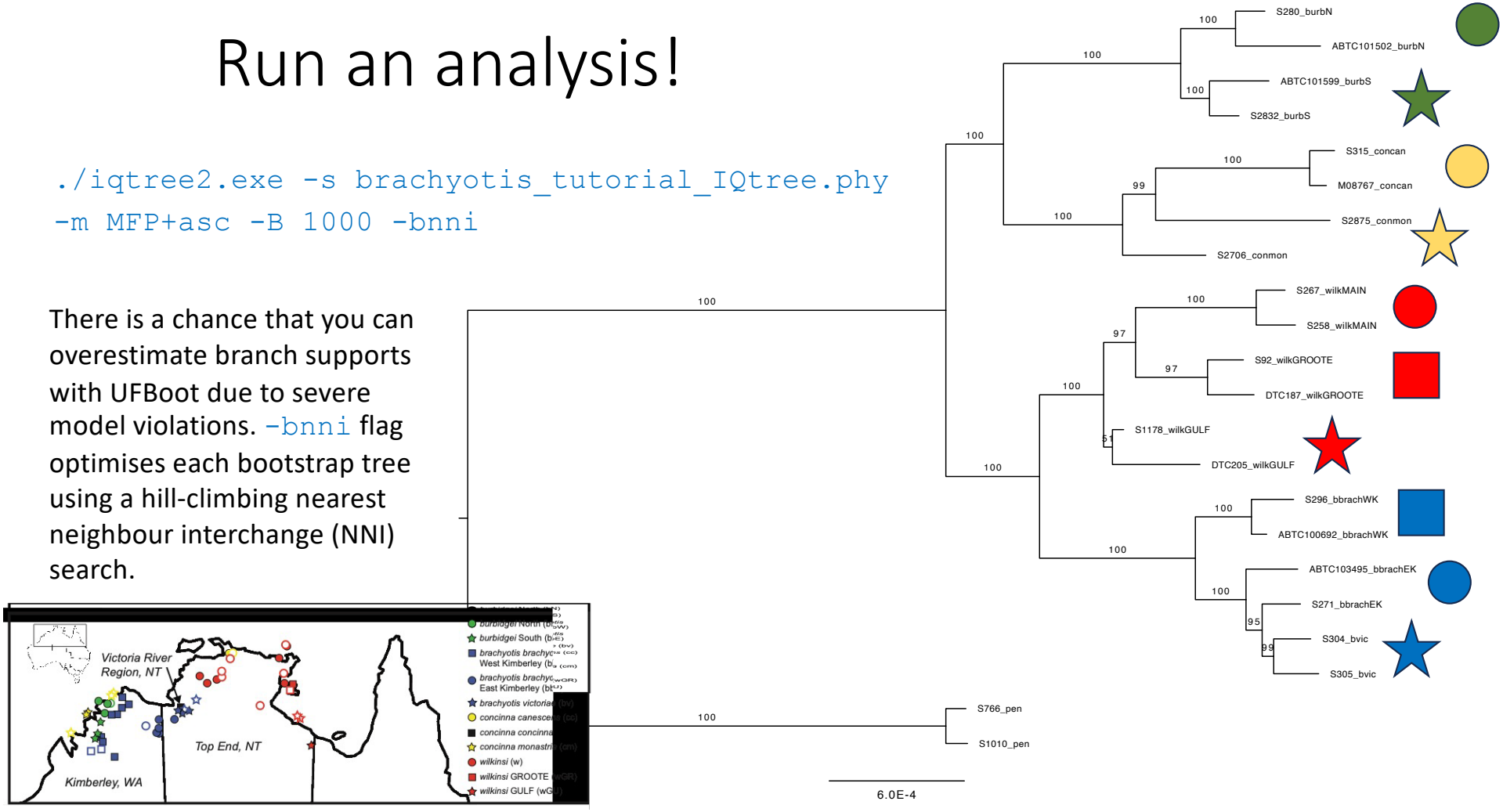
https://doi.org/10.1093/molbev/mst024

https://doi.org/10.1093/molbev/msx281

Run an analysis!

```
./iqtree2.exe -s brachyotis_tutorial_IQtree.phy
-m MFP+asc -B 1000 -bnni
```

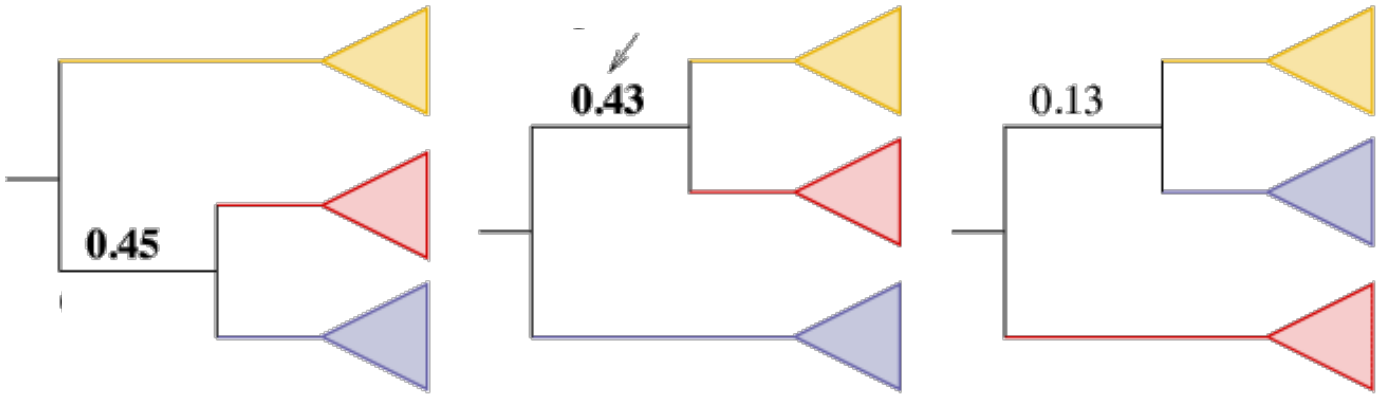
There is a chance that you can overestimate branch supports with UFBoot due to severe model violations. `-bnni` flag optimises each bootstrap tree using a hill-climbing nearest neighbour interchange (NNI) search.



Site Concordance Factors

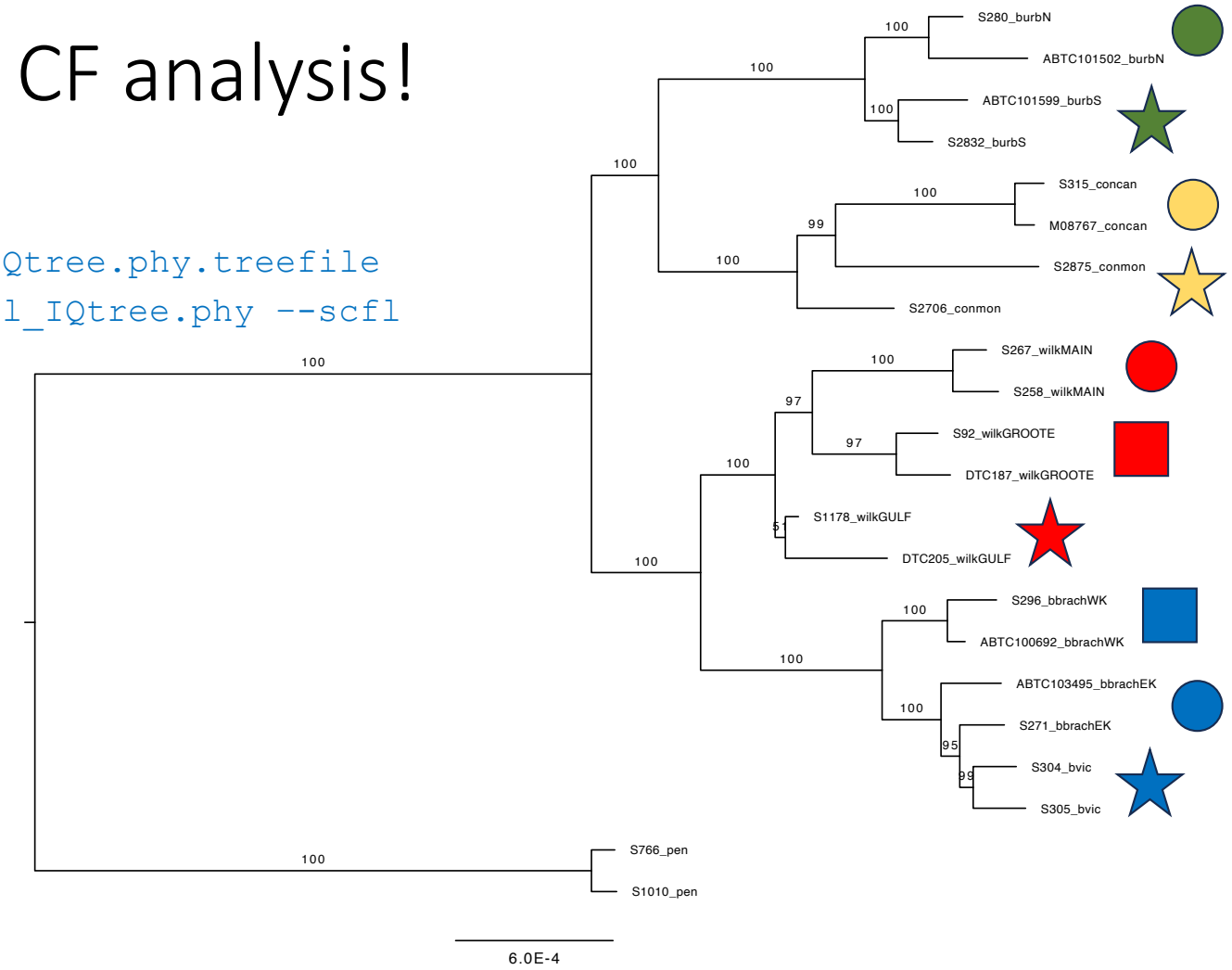
- For SNP datasets you can use the sCF – the percentage of decisive alignment sites supporting a branch in the ML tree.
- sCF gives an overview of the underlying disagreement between sites and can be informative when thinking about incomplete lineage sorting.

Confordance Factor:
43% of sites support this clade



Run an CF analysis!

```
./iqtree2.exe -t
brachyotis_tutorial_IQtree.phy.treefile
-s brachyotis_tutorial_IQtree.phy --scfl
100 -T AUTO -pre scf
```



Summary

- Familiarity with Maximum Likelihood analyses.
- Example of how to estimate boostrap support for a phylogeny.
- Example of how to estimate site concordance factors for a phylogeny.

In the tutorial:

- Show you how to estimate or assign different nucleotide models to analyses.
- Estimate phylogenetic trees from sequence data.
- Run bootstrap analyses & site concordance factor analyses.

SVDquartets

Estimating species/lineage trees

Adapted from Kubatko & Swofford 2017

Why SVDquartets

- Enables use of the full dataset (for large-scale data) in a computationally feasible analysis
- Uses a different model which requires less computation
- Infer relationships among quartets of taxa
- Uses a quartet assembly method to build a tree
- Can be used on SNP datasets and DNA sequence datasets

SVDquartets

- Each site assumed to evolve independently
- Assess site pattern probabilities

Taxon Sequence	
(A) Red Kangaroo	*TAGG T ACCCGATGCCGTA*
(B) Swamp Wallaby	*TAGG A ACCCGTTGCCGTA*
(C) Eastern Grey Kangaroo	*TACG T TCCGGAAGCGGAT*

- Probability of TAT

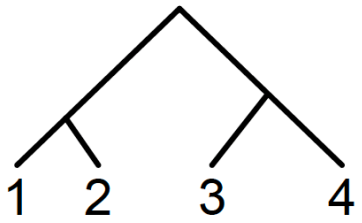
Species Tree Inference using SVDquartets

Laura Kubatko and Dave Swofford

Joint work with
Julia Chifman, American University
Colby Long, MBI at OSU

January 31, 2017

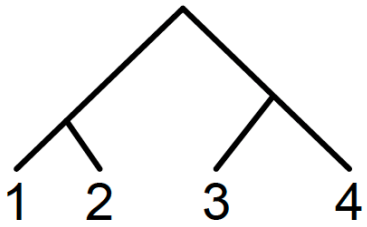
SVDquartets



Taxon	Sequence
1	ACCAATGCCGATGCCAAA
2	ACCATTGCCGATGCCATA
3	ACGAAAGCGGAAGCGAAA
4	ATGAAAGCGGAAGCCAAA

$$Flat_{12|34}(P) = \begin{pmatrix} & [AA] & [AC] & [AG] & [AT] & [CA] & \cdots \\ [AA] & p_{AAAA} & p_{AAAC} & p_{AAAG} & p_{AAAT} & p_{AACA} & \cdots \\ [AC] & p_{ACAA} & p_{ACAC} & p_{ACAG} & p_{ACAT} & p_{ACCA} & \cdots \\ [AG] & p_{AGAA} & p_{AGAC} & p_{AGAG} & p_{AGAT} & p_{AGCA} & \cdots \\ [AT] & p_{ATAA} & p_{ATAC} & p_{ATAG} & p_{ATAT} & p_{ATCA} & \cdots \\ [CA] & p_{CAAA} & p_{CAAC} & p_{CAAG} & p_{CAAT} & p_{CACA} & \cdots \\ [\cdots] & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

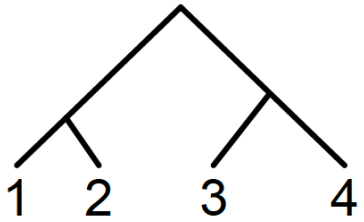
SVDquartets



Taxon	Sequence
1	ACCAATGCCGATGCCAAA
2	ACCATTGCCGATGCCATA
3	ACGAAAGCGGAAGCGAAA
4	ATGAAAGCGGAAGCCAAA

$$Flat_{12|34}(P) = \begin{pmatrix} & [AA] & [AC] & [AG] & [AT] & [CA] & \cdots \\ [AA] & \mathbf{5} & p_{AAAC} & p_{AAAG} & p_{AAAT} & p_{AACA} & \cdots \\ [AC] & p_{ACAA} & p_{ACAC} & p_{ACAG} & p_{ACAT} & p_{ACCA} & \cdots \\ [AG] & p_{AGAA} & p_{AGAC} & p_{AGAG} & p_{AGAT} & p_{AGCA} & \cdots \\ [AT] & p_{ATAA} & p_{ATAC} & p_{ATAG} & p_{ATAT} & p_{ATCA} & \cdots \\ [CA] & p_{CAAA} & p_{CAAC} & p_{CAAG} & p_{CAAT} & p_{CACA} & \cdots \\ [\cdots] & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

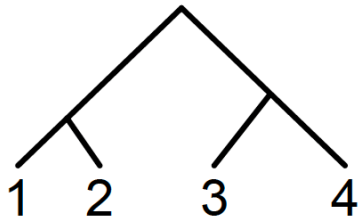
SVDquartets



Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

$$Flat_{12|34}(P) = \begin{pmatrix} & [AA] & [AC] & [AG] & [AT] & [CA] & \cdots \\ [AA] & \mathbf{5} & p_{AAAC} & p_{AAAG} & p_{AAAT} & p_{AACA} & \cdots \\ [AC] & p_{ACAA} & p_{ACAC} & p_{ACAG} & p_{ACAT} & p_{ACCA} & \cdots \\ [AG] & p_{AGAA} & p_{AGAC} & p_{AGAG} & p_{AGAT} & p_{AGCA} & \cdots \\ [AT] & p_{ATAA} & p_{ATAC} & p_{ATAG} & p_{ATAT} & p_{ATCA} & \cdots \\ [CA] & p_{CAAA} & p_{CAAC} & p_{CAAG} & \mathbf{2} & p_{CACA} & \cdots \\ [\cdots] & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

SVDquartets



Taxon	Sequence
1	ACCAATGCCGGAGCCCAAA
2	ACCATTGACGGAGCCAATA
3	ACGAAAGACGGAAGCAAAA
4	ATGAAAGTCGGAAGCTAAA

These two columns are identical - matrix rank is reduced by one

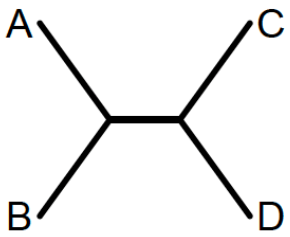
$$Flat_{12|34}(P) = \begin{pmatrix} & [AA] & [\mathbf{AC}] & [AG] & [AT] & [\mathbf{CA}] & \cdots \\ [AA] & \mathbf{5} & p_{\mathbf{AAAC}} & p_{AAAG} & p_{AAAT} & p_{\mathbf{AACA}} & \cdots \\ [AC] & p_{ACAA} & p_{\mathbf{ACAC}} & p_{ACAG} & p_{ACAT} & p_{\mathbf{ACCA}} & \cdots \\ [AG] & p_{AGAA} & p_{\mathbf{AGAC}} & p_{AGAG} & p_{AGAT} & p_{\mathbf{AGCA}} & \cdots \\ [AT] & p_{ATAA} & p_{\mathbf{ATAC}} & p_{ATAG} & p_{ATAT} & p_{\mathbf{ATCA}} & \cdots \\ [CA] & p_{CAAA} & p_{\mathbf{CAAC}} & p_{CAAG} & \mathbf{2} & p_{\mathbf{CACA}} & \cdots \\ [\cdots] & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

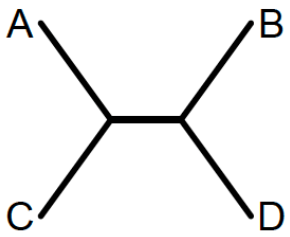
SVDquartets

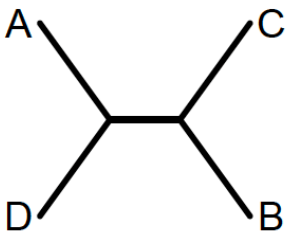
- *Data: aligned SNPs or DNA sequences for multiple loci*
- *Construct the flattening matrix*
- *Compute singular value decomposition (SVD) of the flattening matrix - SVD score for a split A|B*
- *Pick tree relationships that give the best value of SVD score*
- *not impacted by molecular clock or population sizes change along the tree (Long & Kubatko 2017)

SVDquartets

Which of these three options represents the true split of these four taxa?

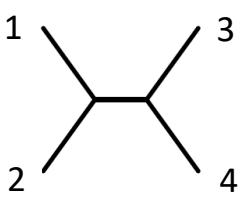


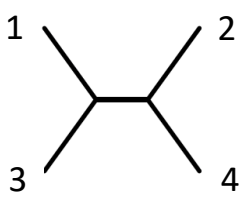


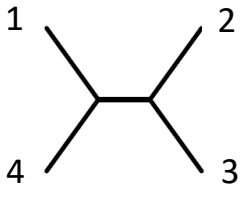


Use the observed site pattern distribution to estimate the true split.

*Compute a score for each split in a given quartet of taxa and choose the split with the best (lowest) score.

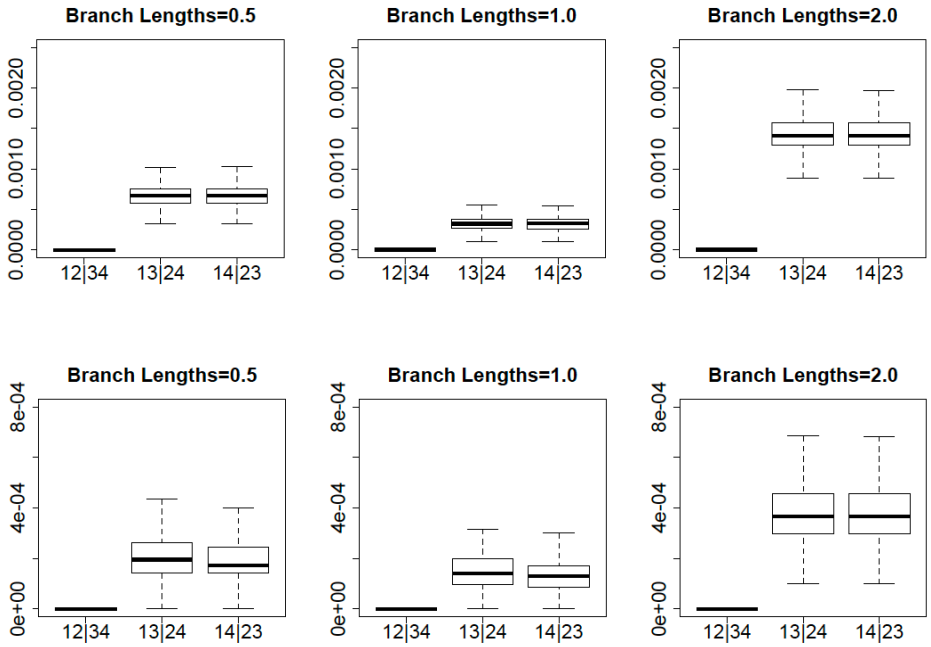


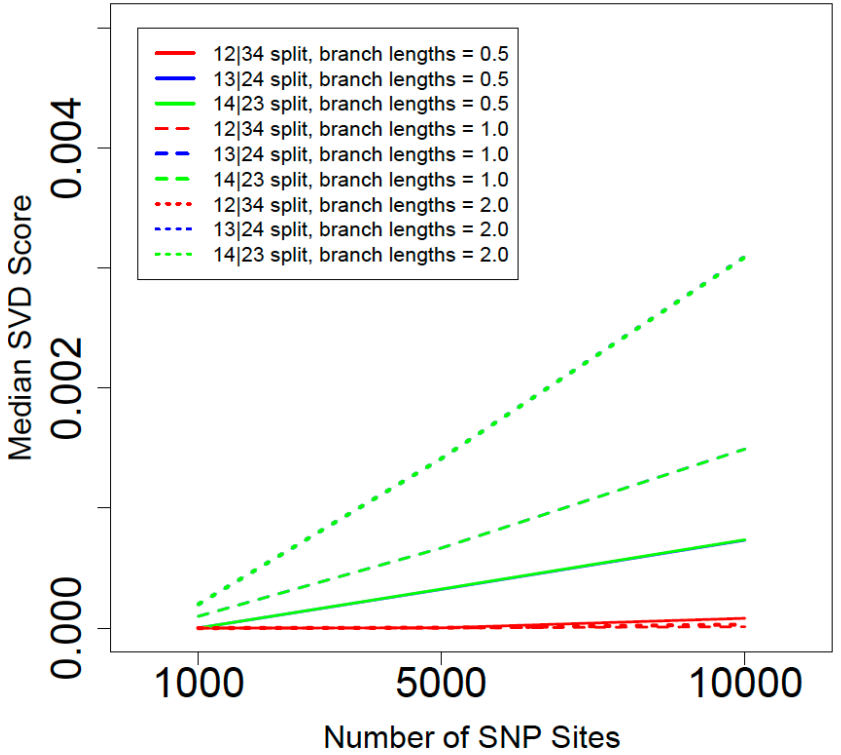




SVDquartets

Simulated data from Juke-Cantor & GTR+I+Γ nucleotide model for 4-taxon tree from 5000 SNPs





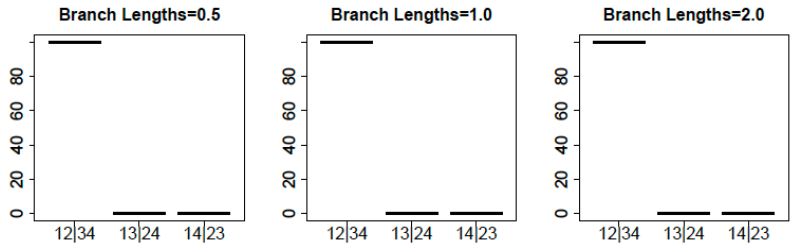
The lowest SVD score = best quartet model

SVDquartets

- Nonparametric bootstrap
  - procedure to measure confidence in split
- Generate bootstrap data sets from the original data matrix
- Compute split scores on all three splits for each bootstrap data matrix
- Record the number of bootstrap data sets for which each split is inferred, and use the proportion

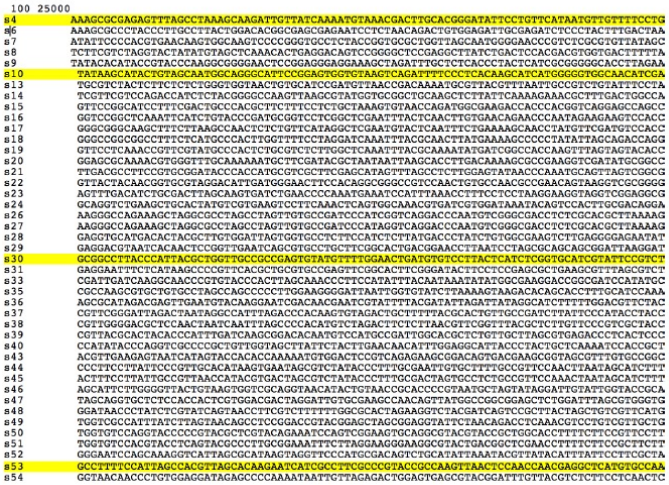
  of these as a bootstrap support

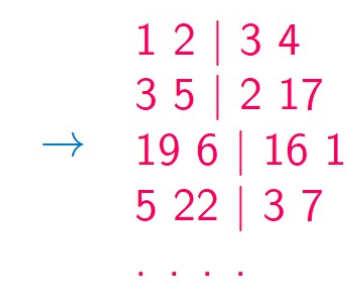
  measure

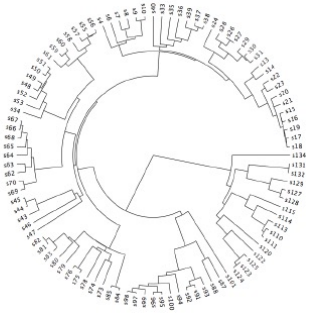


Algorithm

1. Generate all quartets (small problems) or sample quartets (large problems)
2. Estimate the correct quartet relationship for each sampled quartet
3. Use a quartet assembly method to build the tree - PAUP* uses the method of Reaz-Bayzid-Rahman (2014), called QFM, to build the tree.



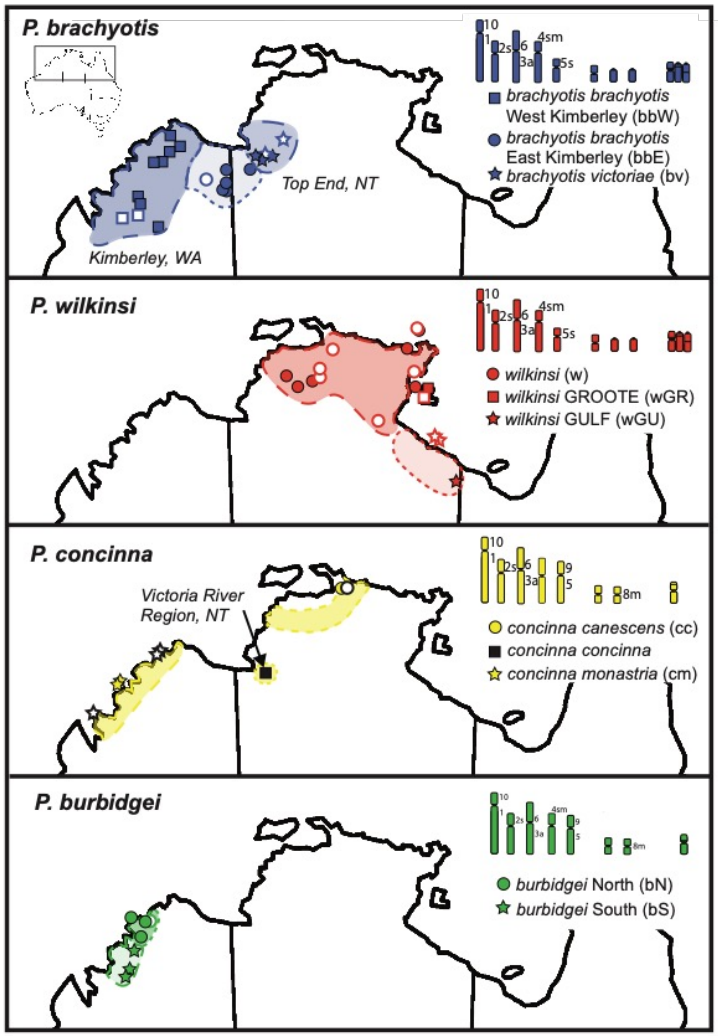




brachyotis group of rock-wallabies (phylogeography/phylogenetics)

Aim: resolve the phylogenetic relationships of lineages/OTUs

In this example – we use 2 individuals per OTU
Dataset is exon sequence data ~800K sites
(monomorphic & polymorphic), 10 OTUs + outgroup



Input file

Nexus input file – ntax, nchar
> Concatenated SNP data
Taxa partition block
> Assign individuals to taxon for species/lineage tree analysis

```
#NEXUS
begin taxa;
    dimensions ntax=22;
    taxlabels
    ABTC103495_bbrachEK
    S271_bbrachEK
    S296_bbrachWK
    ABTC100692_bbrachWK
    S304_bvic
    S305_bvic
    S280_burbN
    ABTC101502_burbN
    ABTC101599_burbS
    S2832_burbS
    S315_concan
    M08767_concan
    S2875_conmon
    S2706_conmon
    S267_wilkMAIN
    S258_wilkMAIN
    S92_wilkGROOTE
    DTC187_wilkGROOTE
    S1178_wilkGULF
    DTC205_wilkGULF
    S766_pen
    S1010_pen
;
end;

begin characters;
    dimensions nchar=749814;
    format datatype=dna missing=? gap=-;
    matrix
    ABTC103495_bbrachEK
TCCCGAGTCTCCTCCGACAATGGGCAGCGGTTATTGGACATGATTGAGACCCTGAAAATGGAGCACCACCTGGAGGCCCA
CAAGAATGTGCATGCCAACGGGCACTTGAGCATCAGTGCTGATGGCCCAGTGCCCGACAGCAGCTTCATCAGCAAGGCCA
AACACAACGGCGTGCCCAATGGCTACCAGGGGGGGTTGGAAATGATCAAGATCCCCATACCTGAACCCGAGCGCTCCCAG
TACCCCCGGGAGTGGGGCAAGACTCTGCTGGCTTTTCTGTATGCCCTCTGCTGCTTTGTGTTCACCACAGTGACCATCTC
GTTCGTCCATGAGCGAGTGCCTCCCAAGGAGATGCAGGCACCCCTACCGGACACCTTTTTTGACCATTTTAACCGGTATG
TCATTGCAGGGAGCCTGCCCCATCTGGGCCAGTGTCTGACTCGGTACTGGGAACAGAAGAAGTGCATGGCTCCTGGCTGT
GAACCCCAGGCTGTGCAGCGCATGATGGCTGCCCTCAAGCCCTATGTGTATGGCCAGAGCCTGGCAGGGGCTGGCGGTGG
CGGCTTCCTCTACCTGTTGACTAAACAGCCCCGGCAGAAAGCTACAAGTTTTCCTGTAGAGAGACTGAAGGAAGAAGCAA
AATGTTCAATCTGTCTGGACTTCTTAAAGGATGCAGTCAGTGTGGACTGTGGCCACAGTTTCTGCCTCCAGTGCATCACT
GAATTCTGTGAGAAATCAGACAGTTCTCAAGGGAATGTCTACTCCTGCCCRCAGTGCCGGAGTCAATTTCAACAGSATAG
```

```
BEGIN SETS;
   TAXPARTITION OTUs =
        bbrachEK: 1-2,
        bbrachWK: 3-4,
        bvic: 5-6,
        burbN: 7-8,
        burbS: 9-10,
        concan: 11-12,
        conmon: 13-14,
        wilkMAIN: 15-16,
        wilkGROOTE: 17-18,
        wilkGULF: 19-20,
        pen: 21-22;
 END;
```

PAUP*

```
P A U P *
Version 4.0a (build 168) for macOS (built on Aug  2 2020 at 06:41:30)
Wed Mar  6 09:39:35 2024

        ------------------------------NOTICE------------------------------
          This is a test version that is still changing rapidly.
          Please report bugs to dave@phylosolutions.com
        ------------------------------------------------------------------

Running on Apple M1
    8 CPU cores on 1 socket
    Executable built for Intel(R) 64 architecture
    Compiled using Intel compiler (icc) 11.1.0 (build 20091012)
    SSE vectorization enabled
    SSSE3 instructions supported
    Multithreading enabled using Pthreads

Note: Attempt to get current version info from server was not successful (timeout limit exceeded).
      You can do "checkVersion timeout=<number-of-seconds>;" at the PAUP* command prompt to get
      current version info from our server with a longer timeout period.

paup>
```

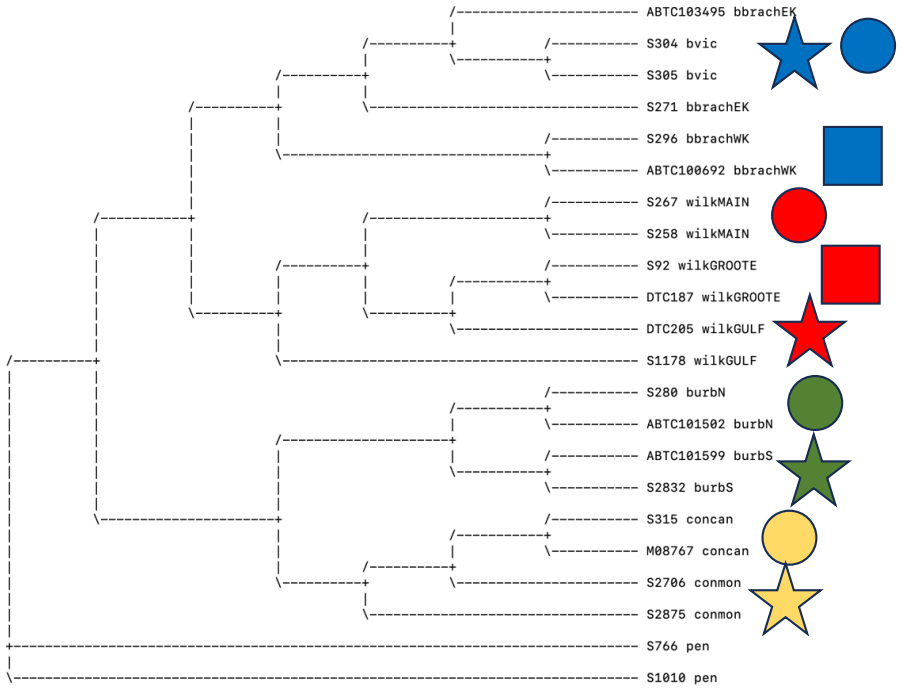
Load your nexus input file

exe brachyotis_tutorial_SVD.nex

Then run your analysis!

Individual/lineage analysis

```
Quartet assembly completed:
  Total weight of incompatible quartets = 693 (9.474%)
  Total weight of compatible quartets   = 6622 (90.526%)
  Time used for QFM = 0.01 sec (CPU time = 0.00 sec)

Tree from SVDQuartets analysis (also stored to tree buffer):
```



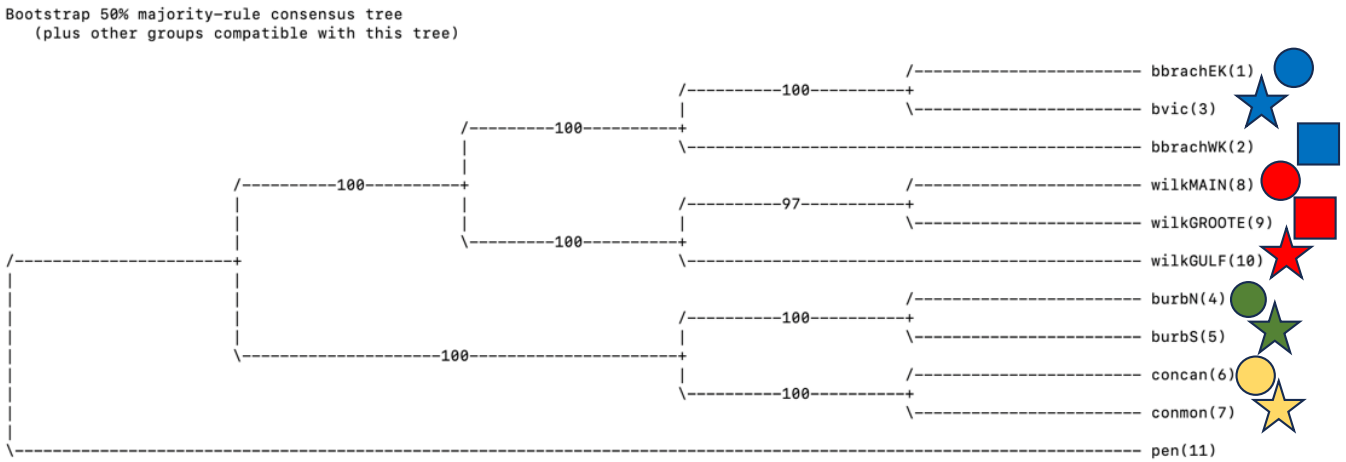

```
svdq taxpartition = none showScores=yes
seed=1234 treeFile=brach_SVD.tre;
```

SVD scores for each quartet

quartet	----split----				score
1	1	2	3	4	1.8281e-06
	1	3	2	4	2.3952e-05
	1	4	2	3	2.3869e-05
2	1	2	3	5	9.3960e-06
	1	3	2	5	4.0127e-06
	1	5	2	3	8.6807e-06
3	1	2	3	6	9.7673e-06
	1	3	2	6	4.4987e-06
	1	6	2	3	1.1045e-05
4	1	2	3	7	1.1316e-05
	1	3	2	7	1.7177e-05
	1	7	2	3	1.5422e-05
5	1	2	3	8	9.6282e-06
	1	3	2	8	1.7488e-05
	1	8	2	3	1.5547e-05
6	1	2	3	9	1.1211e-05
	1	3	2	9	1.9668e-05
	1	9	2	3	1.7083e-05
7	1	2	3	10	9.2679e-06
	1	3	2	10	1.5580e-05
	1	10	2	3	1.3362e-05
8	1	2	3	11	1.0145e-05
	1	3	2	11	1.9123e-05
	1	11	2	3	1.6987e-05
9	1	2	3	12	9.7229e-06
	1	3	2	12	1.8964e-05
	1	12	2	3	1.7108e-05
10	1	2	3	13	1.8840e-05
	1	3	2	13	1.8076e-05
	1	13	2	3	1.1219e-05

Species tree analysis

```
outgroup 21-22

svdq taxpartition=OTUs showScores=yes seed=1234568 bootstrap
nreps=100 treeFile=brach_SVD_OTUs_bs.tre;

rootTrees rootMethod=outgroup

savetrees file=brach_SVD_OTUs_rooted.tre
```



Summary

- Familiarity with site pattern probabilities and coalescent based quartet assembly analyses.
- Example of how to estimate species/lineage trees from SNP data using the quartet method.
- Example of how to simulate bootstrap support for a phylogeny.

In the tutorial:

- Show you how to assign different nucleotide models to analyses.
- Estimate divergence times given a species tree and sequence data.
- Convert divergence values to time using theta and mutation rate estimates.
- Additional information of how to get help for the analysis using the program & other resources.

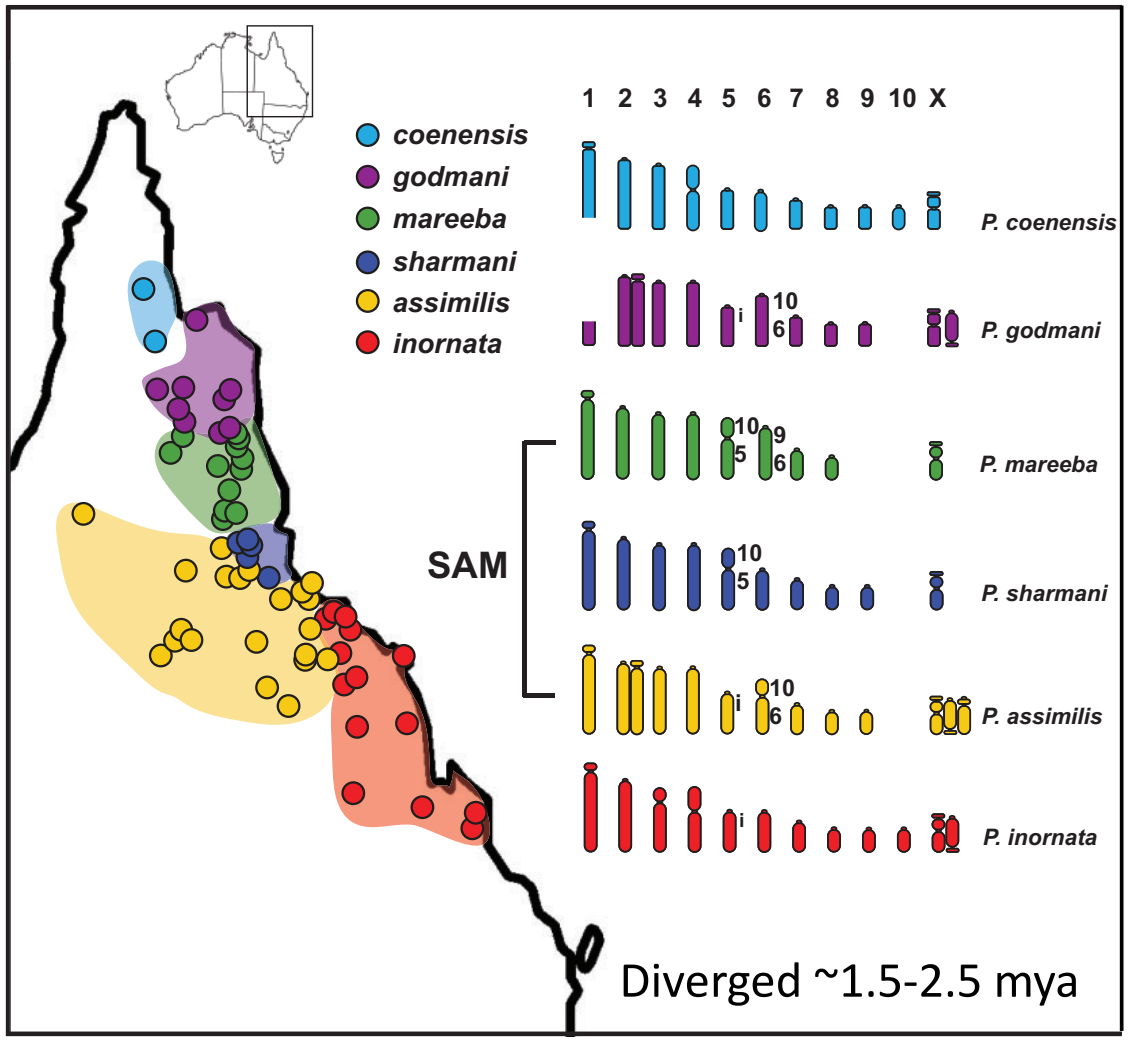
TreeMix

TreeMix in practice

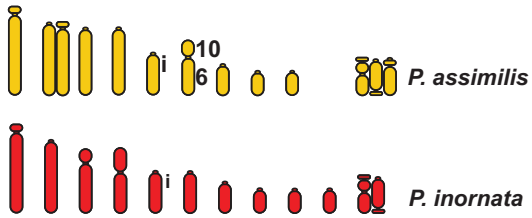
- Want to resolve phylogenetic relationships and estimate migration between species recently diverged (~1-2 million years divergence)

- Species complex
- Chromosomally divergent
- Are unresolved phylogenetic issues due to ILS or introgression?

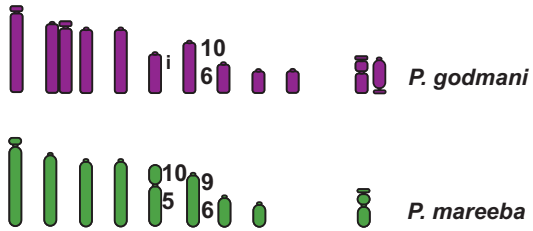
penicillata complex



Simple rearrangements

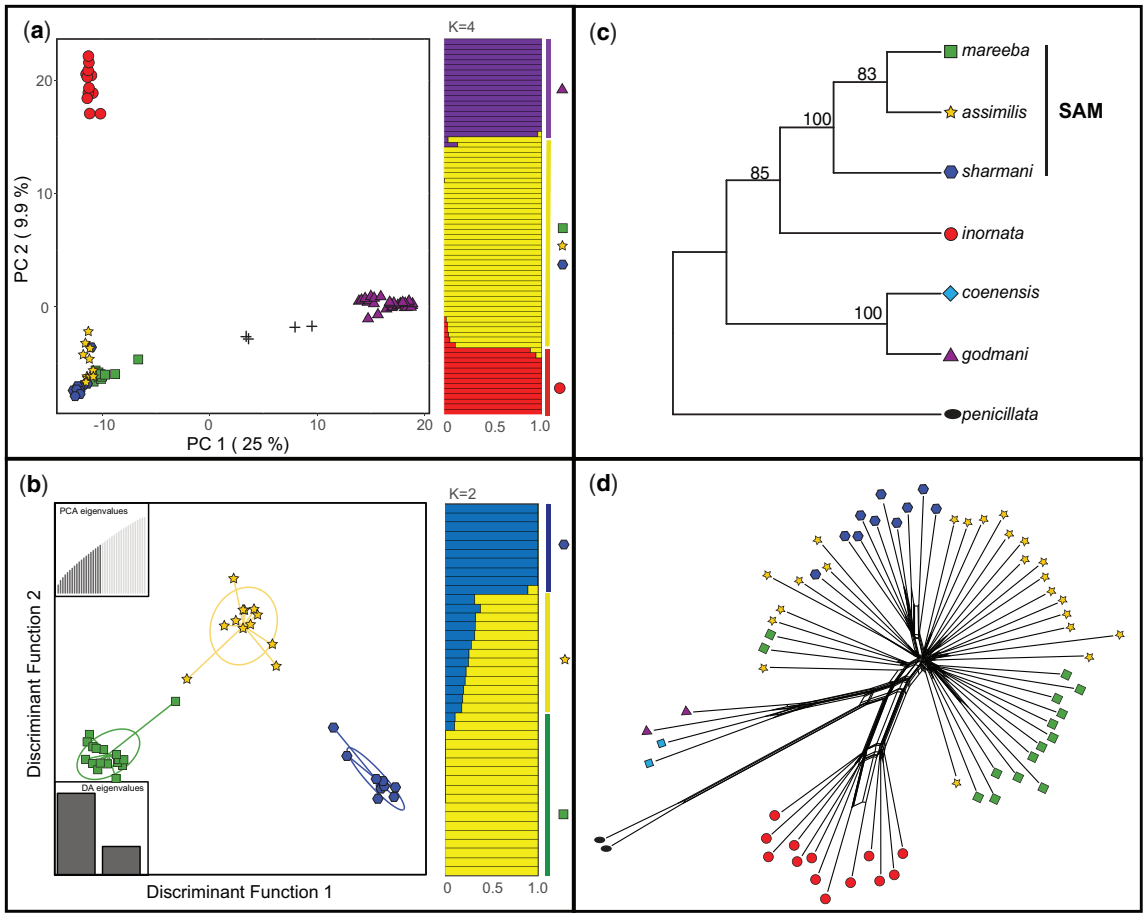


Complex rearrangements



Potter et al. 2022 *Mol Biol Evol*

Divergence and evolutionary history



Potter et al. 2022 MBE

Estimate a ML tree and test migration

First estimate a ML tree:

treemix -i SAMIG_tutorial_treemix.input.gz -root godmani -bootstrap 1000 -k 1 –o godmani_root_k1_b1000

Then run a bash script to estimate likelihoods of trees with varying levels of migration.

treemix -i SAMIG_tutorial_treemix.input.gz -root godmani -m 1 -k 1 –o godmani_root_m1_k1_b1000

```
for m in {1..5}
    do
    for i in {1..5}
        do
        treemix \
          -i SAMIG_tutorial_treemix.input.gz \
          -o SAMIG.${i}.${m} \
          -global \
          -m ${m} \
          -k 1 \
          -root godmani
        done
done
```

Run for different values of k

- groups SNPs to account for LD
- Accounts for nearby SNPs not being independent
- *recommend using a value that far exceeds the known extent of LD in the organism

Run OptM to estimate optimal number of migration events

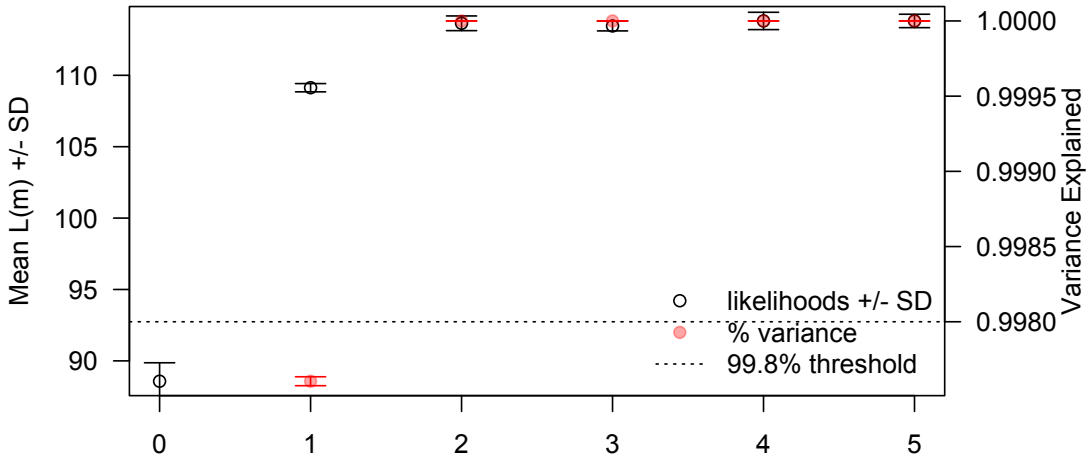
- OptM is a package in R

https://cran.r-project.org/web/packages/OptM/OptM.pdf

- Optimal value of m can be inferred from the second-order rate of change in likelihood (Δm) across incremental values of m.
- Uses Evanno, linear or SiZer approach to estimate optimal m

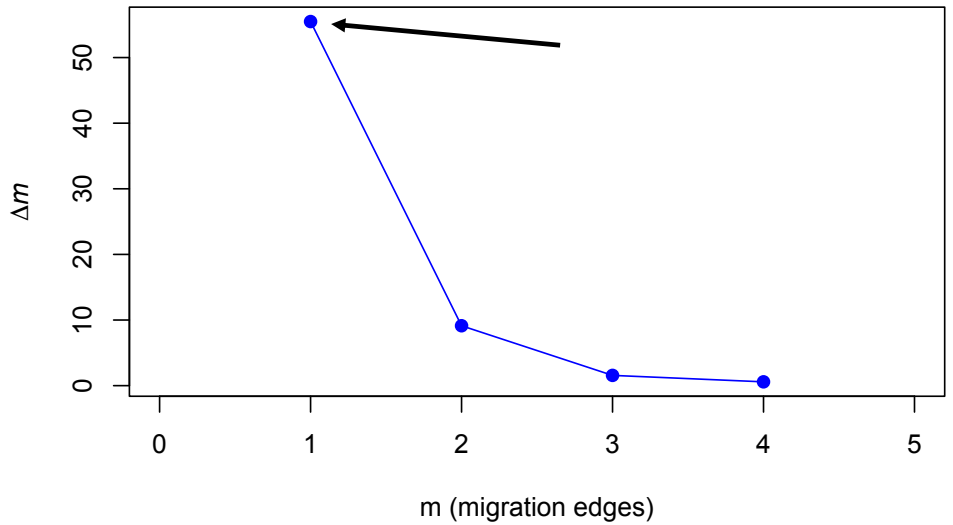
OptM Results

- The 99.8% threshold (horizontal dotted line) is recommended by Pickrell and Pritchard 2012 as good support for a model.



Evanno Method

- Select highest log likelihood (m=2) OR
- Select second-order rate of **Δ**m change across values of m (m=1)



Tree/Graph visualisation

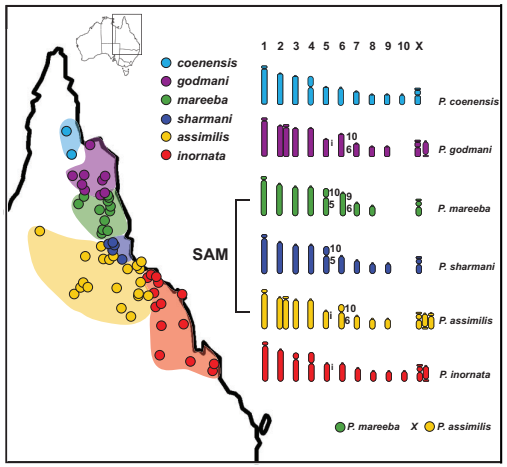
To visualise the graph

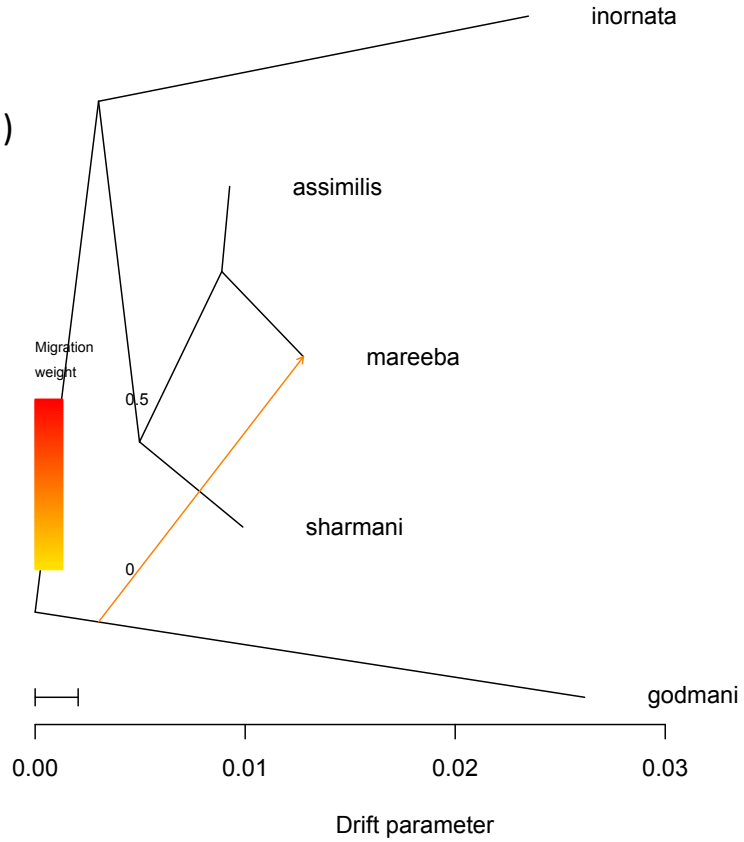
- R script plotting_funcs.R (in src folder in TreeMix download)

>source("src/plotting_funcs.R")

>plot_tree("path_to_file_with_correct_m_model")

e.g., plot_tree("SAMIG_k10.1.1")



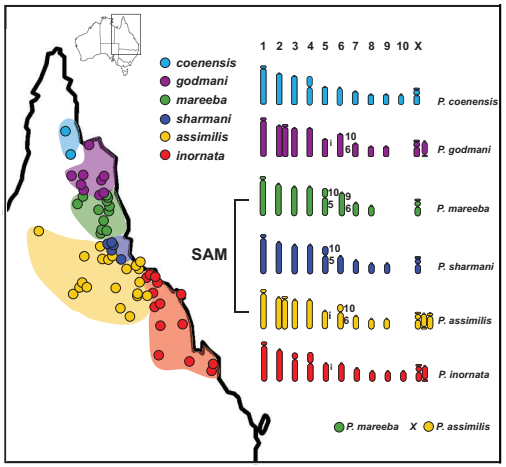


Residual visualisation

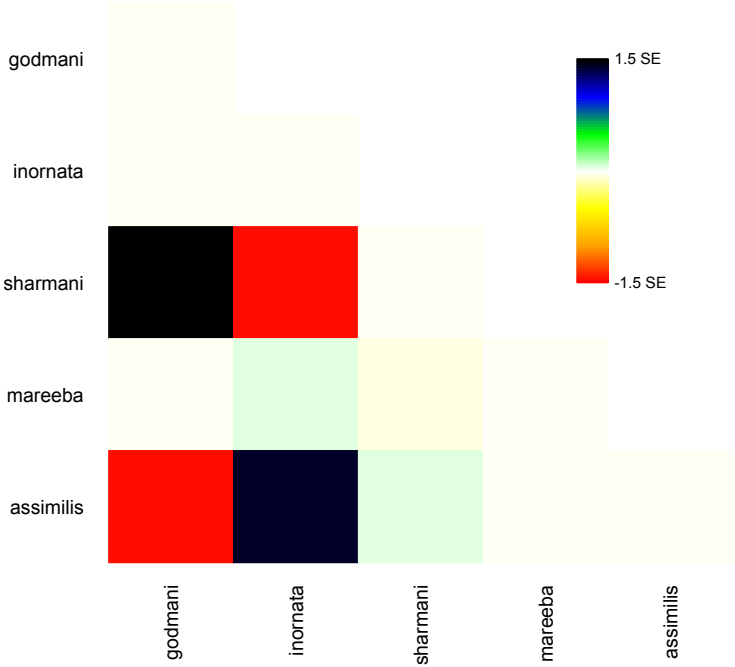
You can visualise the residuals from the fit of the model to the data.

Evaluating the fit of the model can help identify populations that are not well-modeled (e.g., due to additional migration)

>plot_residuals("SAMIG_k10.1.1", "poporder")



Residuals above zero represent populations that are more closely related to each other in the data than in the best-fit tree, and thus are candidates for admixture events.



Summary

- Familiarity with TreeMix analyses and interpretation.
- Example of how to estimate ML tree from SNP data and varying numbers of migration events.
- Example of how to simulate bootstrap support for a ML tree.

In the tutorial:

- Step you through how to build ML trees, with bootstrap support.
- Estimate number of migration events.
- Additional information of how to get help for the analysis using the program & other resources.