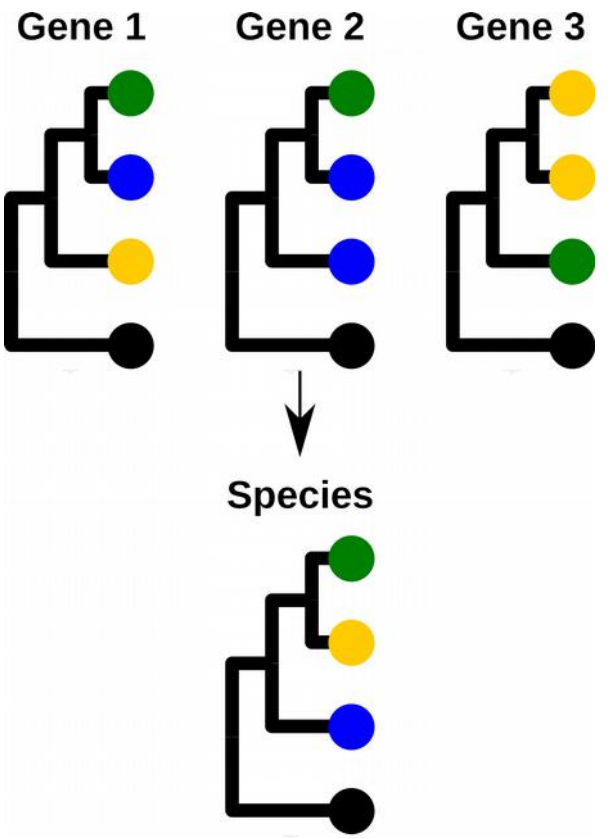
Conflict in multi-gene datasets: why it happens and what to do about it

Deep coalescence, paralogy, reticulation

Alexander N. Schmidt-Lebuhn, alexander.s-l@csiro.au



Genomics for Australian Plants:
Australian Angiosperm Tree of Life
Phylogenomics Workshop 2021

Australian BioCommons

Genomics for Australian Plants

BIOPLATFORMS AUSTRALIA

CENTRE FOR AUSTRALIAN NATIONAL BIODIVERSITY RESEARCH

CSIRO

Content

Premise: what is this about?

Sanger data versus capture/enrichment data

Deep Coalescence

Paralogy

Reticulation

Genomics for Australian Plants

www.genomicsforaustralianplants.com/

@PlantsAus

BIOPLATFORMS AUSTRALIA

https://bioplatforms.com/

@BioplatformsAus

Aims

- Develop genomics resources
- Understanding evolution & conservation of Australian flora
- Upskilling

Areas

- Reference genomes
- Phylogenomics
- Conservation genomics

Genomics for Australian Plants workshop and webinar series for analysing target capture datasets

Genomics for Australian Plants

BIOPLATFORMS AUSTRALIA

Australian BioCommons

https://asbs2021.bablglobal.com/workshop/

Australasian Systematic Botany Society
Annual Conference 2021

ASBS 2021 *virtually*

Biodiverse Futures – Systematics in a Changing World

Virtual ASBS conference from 12-16 July 2021

https://asbs2021.bablglobal.com/

Phylogenomics in GAP

Multiple low-copy nuclear genes

from sequence capture / target enrichment data

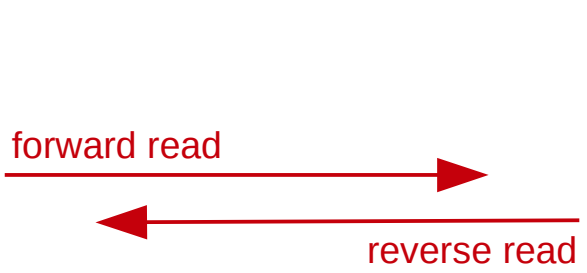
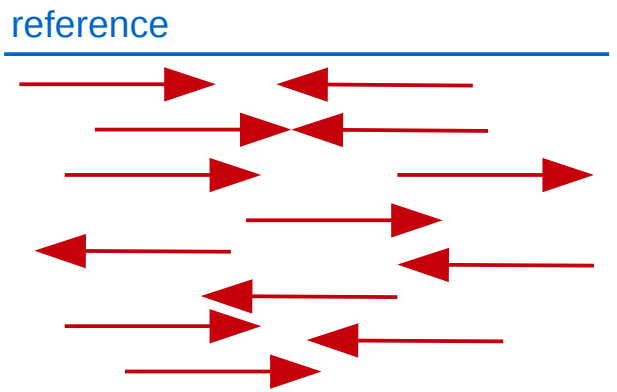
above species level, for phylogenetics

Syst. Biol. 68(4):594–606, 2019

DOI:10.1093/sysbio/syy086
Advance Access publication December 10, 2018

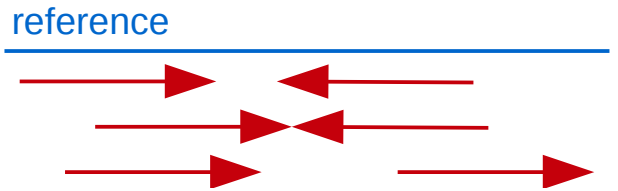
A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering

MATTHEW G. JOHNSON[1,2,*], LISA POKORNY[3], STEVEN DODSWORTH[3,4], LAURA R. BOTIGUÉ[3,5], ROBYN S. COWAN[3], ALISON DEVAULT[6], WOLF L. EISERHARDT[3,7], NIROSHINI EPITAWALAGE[3], FÉLIX FOREST[3], JAN T. KIM[3], JAMES H. LEEBENS-MACK[8], ILIA J. LEITCH[3], OLIVIER MAURIN[3], DOUGLAS E. SOLTIS[9,10], PAMELA S. SOLTIS[9,10], GANE KA-SHU WONG[11,12,13], WILLIAM J. BAKER[3], AND NORMAN J. WICKETT[2,14]

	Ye olde PCR & Sanger data	Enrich/capture & NGS
Assembly & contig building		
Amount of raw data	Few Mb trace files per sample	≥100s of Mb raw NGS reads per sample
Number of regions	Usually 1-5 per study	100s (angiosperm kit: 353)
Behaviour of loci	Usually 2 phylogenies: ribosomal & plastid	Each nuclear gene inherited +/- independently
Type of seq data	Often non-coding spacers (ITS, trnL-trnF, psbA-trnH)	Often protein-coding genes → consider codon positions

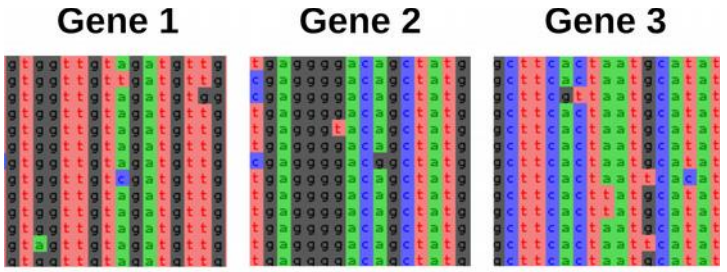
Assembly of reads

(HybPiper)

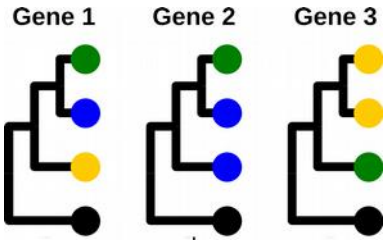


Contig(s)

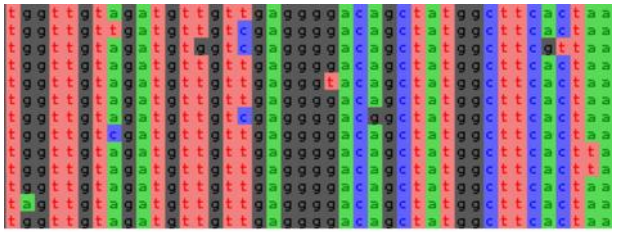
Alignments



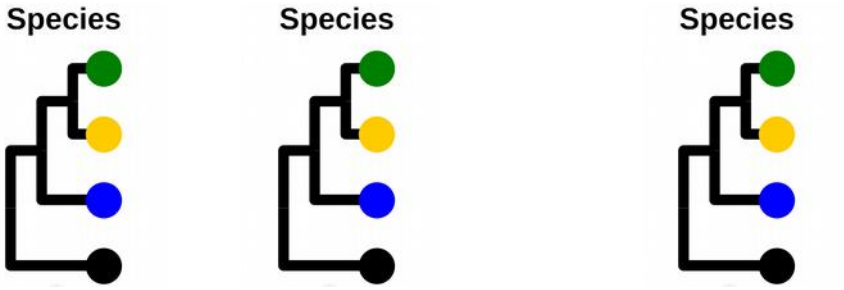
Gene trees



Concatenation



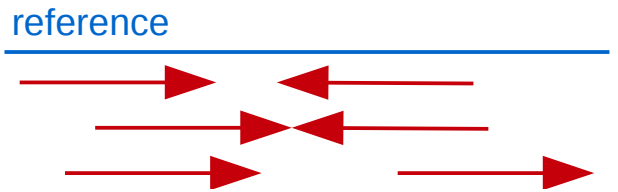
Species phylogeny



’short-cut’ StarBEAST concatenated

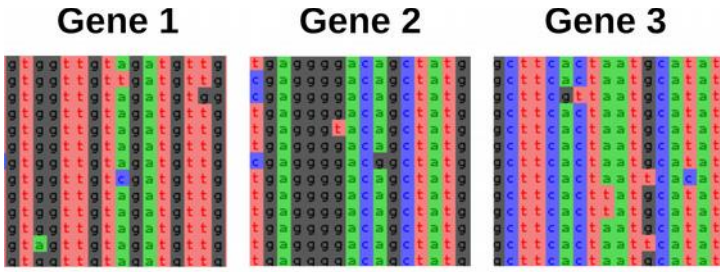
Assembly of reads

(HybPiper)

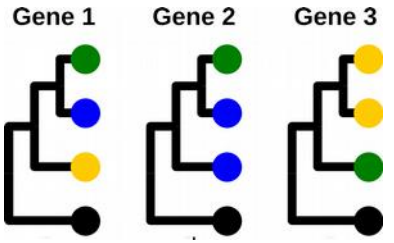


Contig(s)

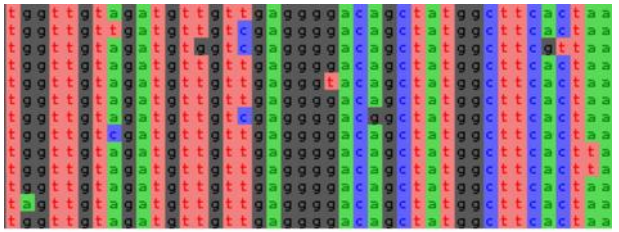
Alignments



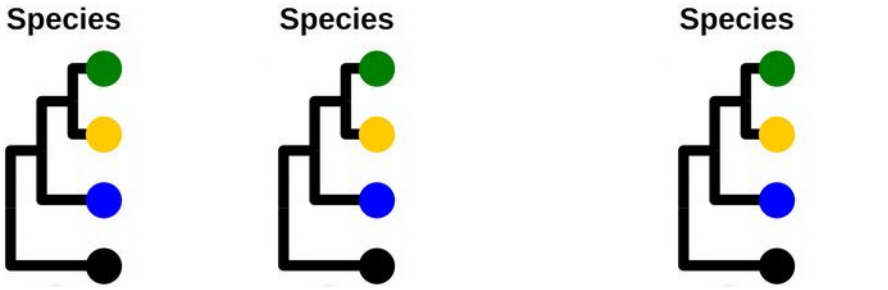
Gene trees



Concatenation



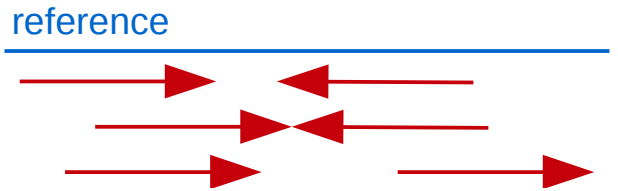
Species phylogeny



’short-cut’ StarBEAST concatenated

Assembly of reads

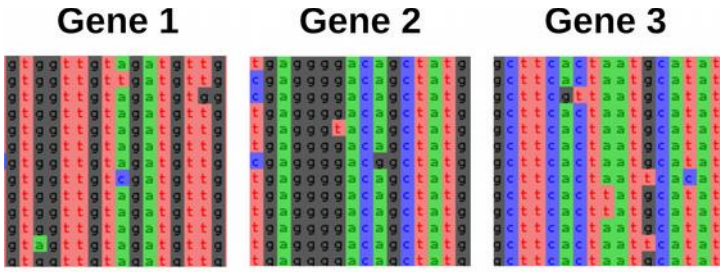
(HybPiper)



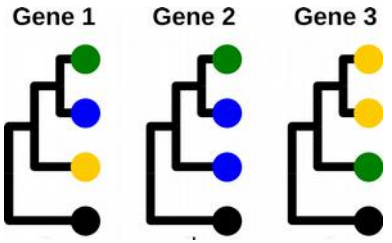
No chimeric contigs!
Paralog Finder!

Contig(s)

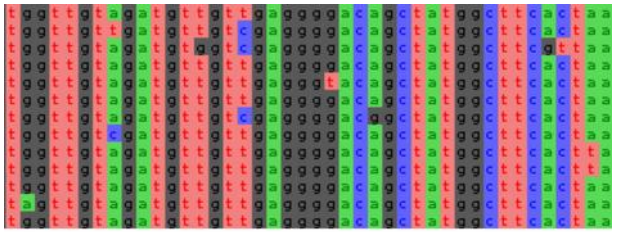
Alignments



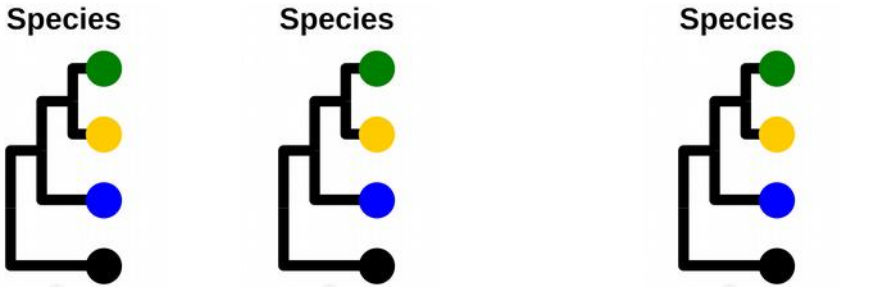
Gene trees



Concatenation



Species phylogeny



’short-cut’ StarBEAST concatenated

Why don’t all gene trees agree with each other?

(And with the species tree?)

Conflicts in the data are common

Three main processes:

- Lineage sorting / coalescence
- Gene duplication and loss
- Reticulation

Good early summary:
Maddison 1997, Syst. Biol.

To consider in each case

What happened?

= hypothesised biological process

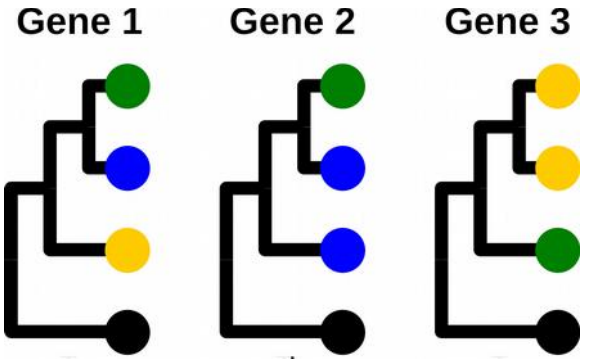
To consider in each case

What happened?

= hypothesised biological process

How does the process present in the data?

‘ideal case’



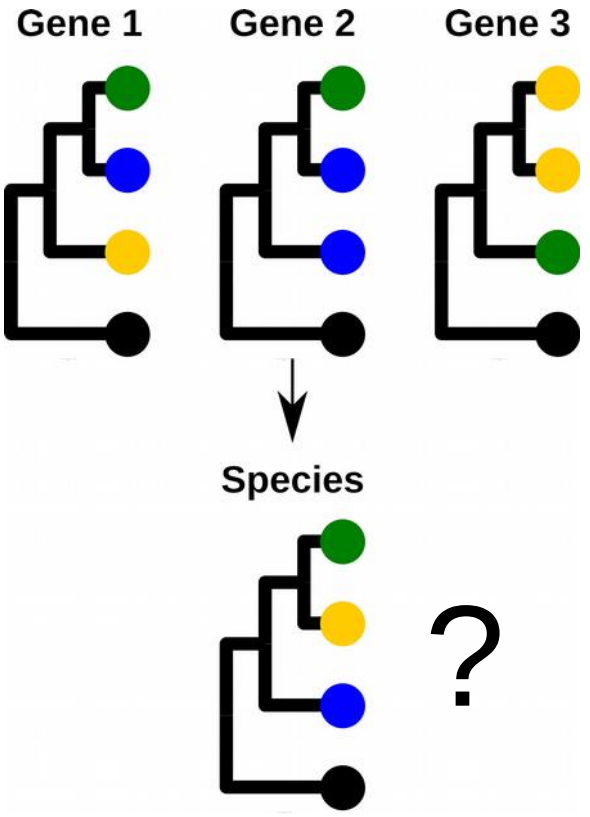
To consider in each case

What happened?

= hypothesised biological process

How does the process present in the data?

'ideal case'



How to infer the species tree?

= approaches and software

Deep coalescence

Random sampling of alleles into descendent species lineages

Lineage sorting / coalescence

What is happening?

Ancestral allele diversity is sampled by descendants

Alleles paraphyletic to those in sister species

= ‘Incomplete Lineage Sorting’

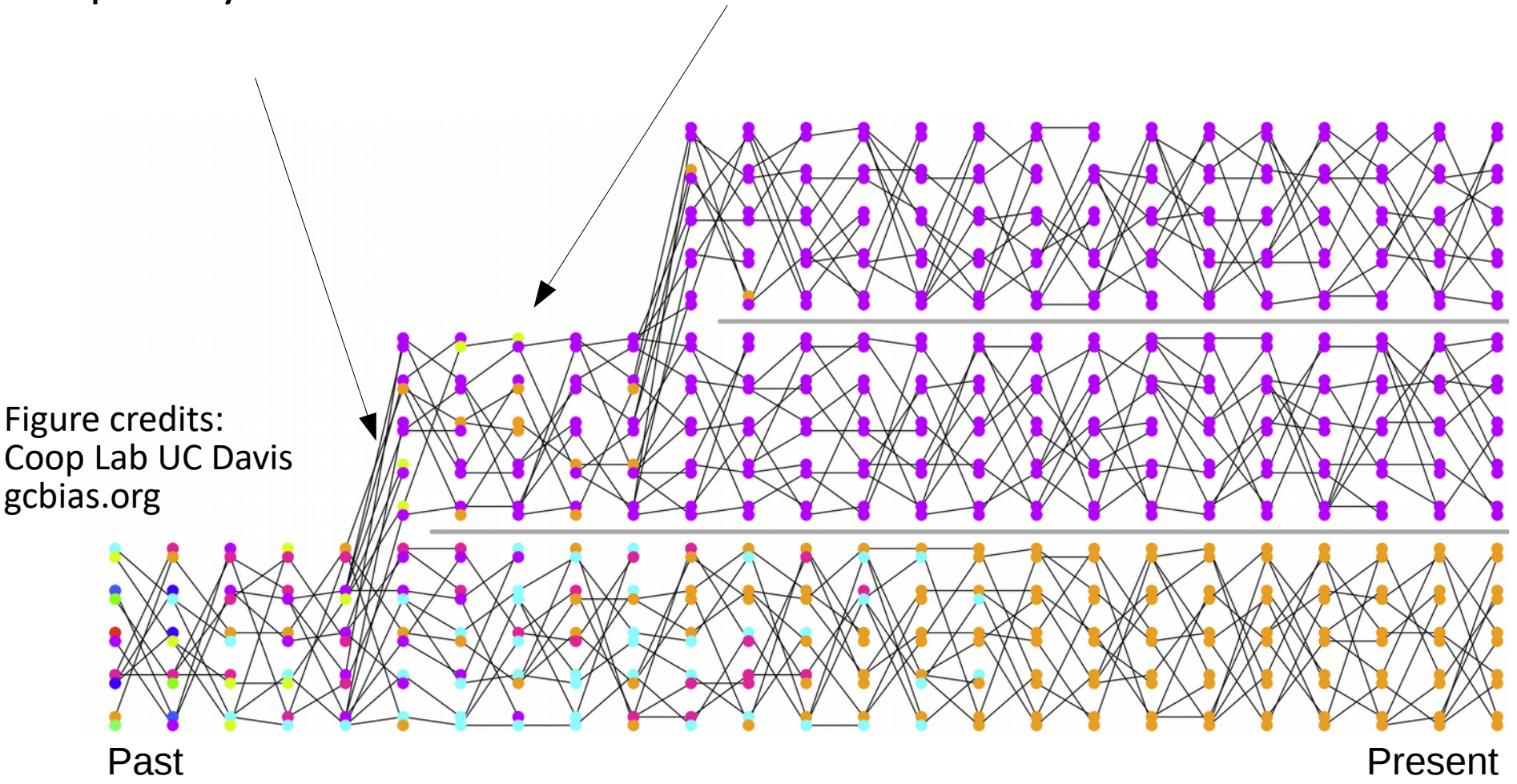


Figure credits:
Coop Lab UC Davis
gcbias.org

Side note on classification

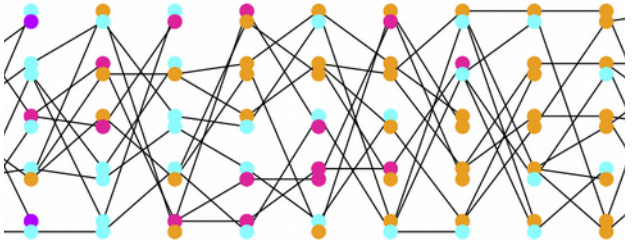
Does not (necessarily) mean that the species is badly circumscribed

"This species is non-monophyletic" is <u>not</u> a meaningful sentence if sexually reproducing

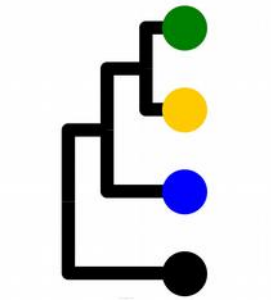
(See: Hennig 1970, *Phylogenetic Systematics*)

We don't classify gene copies, but specimens

Genealogy inside species



≠



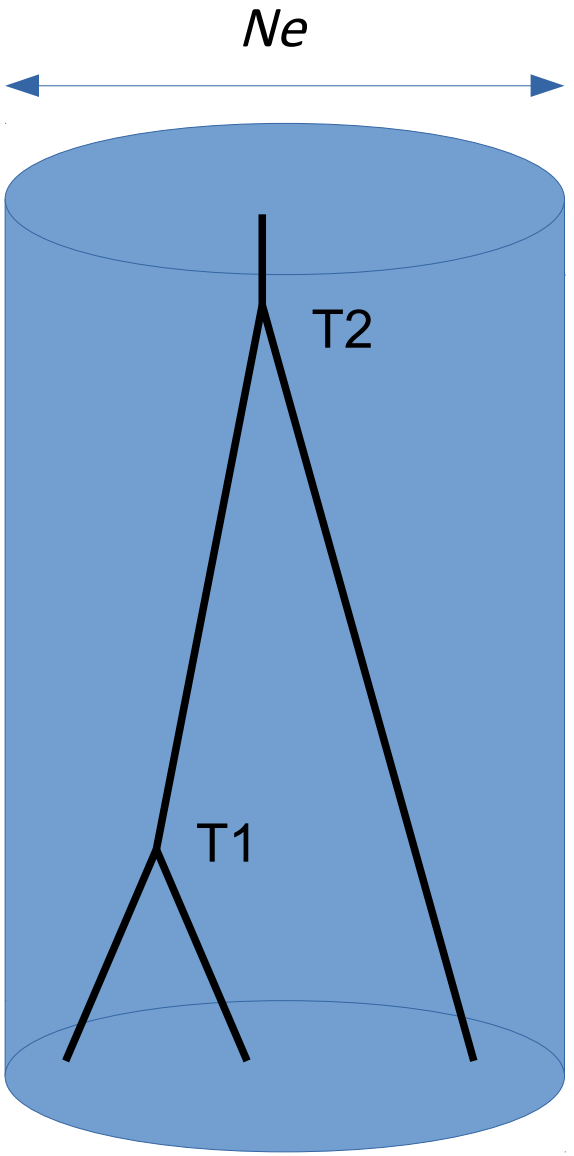
Species relationships

Lineage sorting / coalescence

Lineage sorting = alleles in species becoming monophyletic through genetic drift

Coalescence = extant alleles merging into ancestors back in time

Coalescent Model = estimating coalescent times based on effective population size (*Ne*), which is genetic diversity divided by 4 x mutation rate (μ)



Lineage sorting / coalescence

Consequences in phylogenetics:

Incomplete Lineage Sorting = non-monophyly of alleles in species is resolved by Genetic Drift, but...

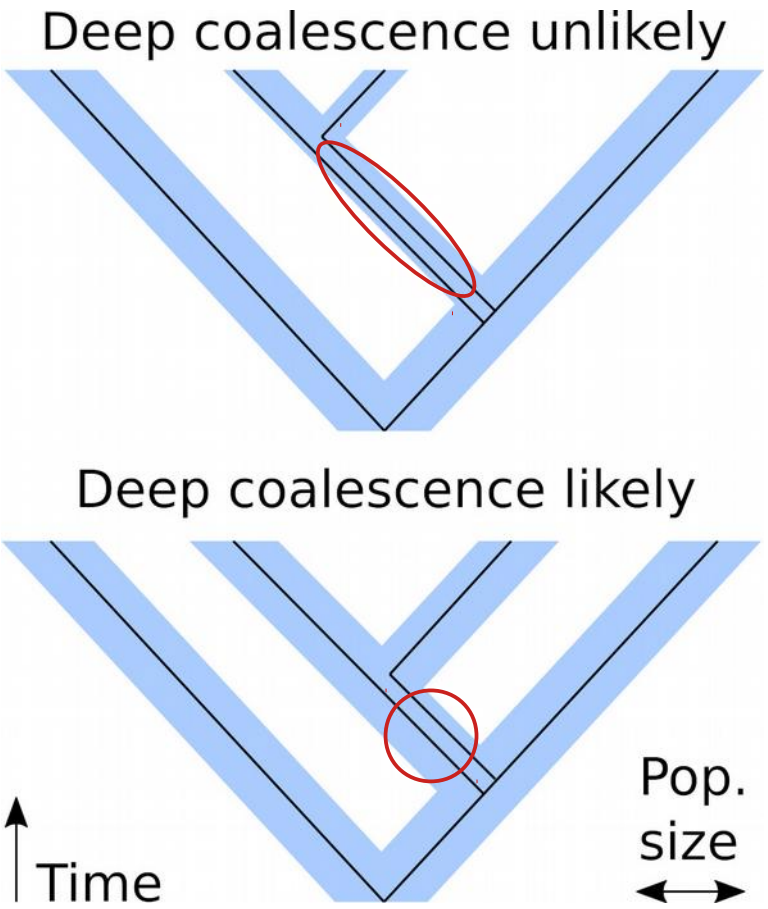
...short time between speciation events

+

large effective population size

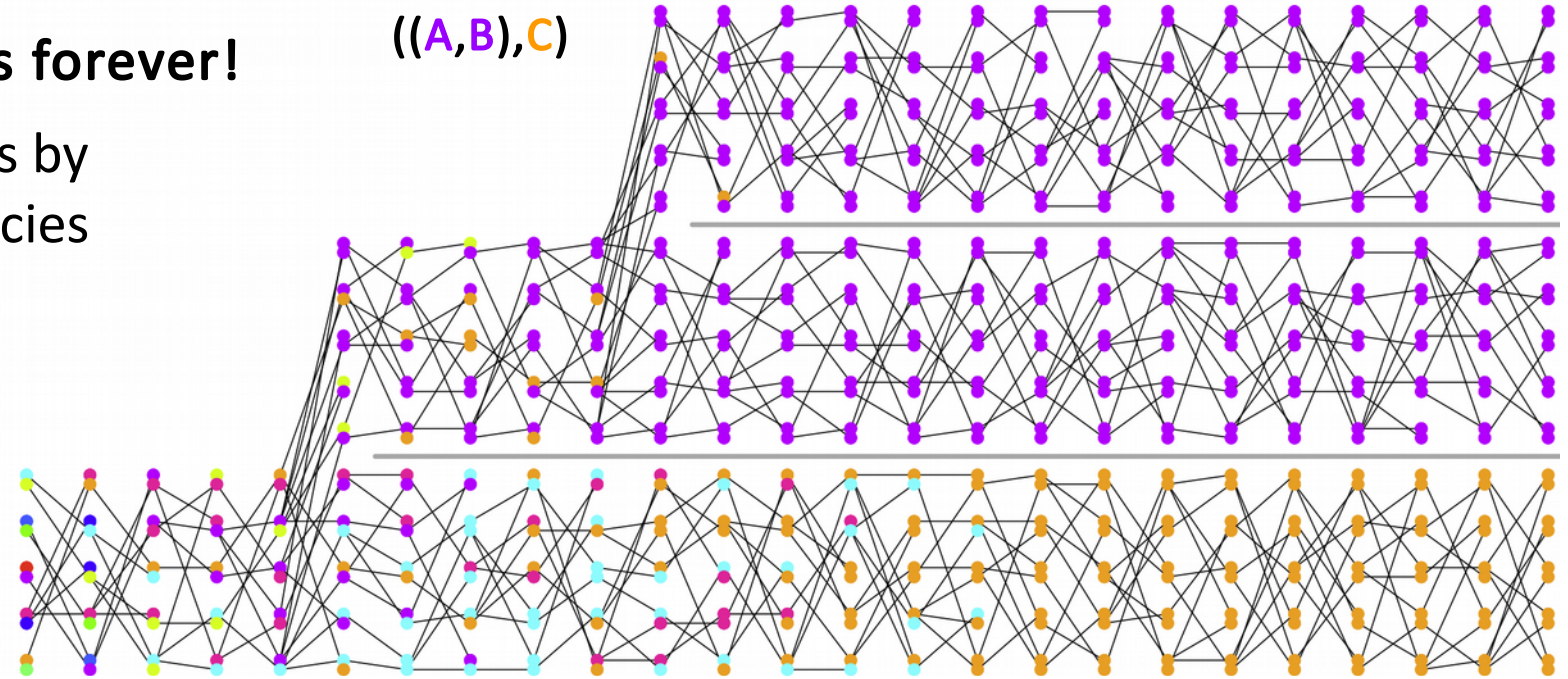
=

Deep Coalescence



Incongruence is forever!

Only resolution is by extinction of species



(Less of a problem in deep phylogenies – most lineages go extinct)

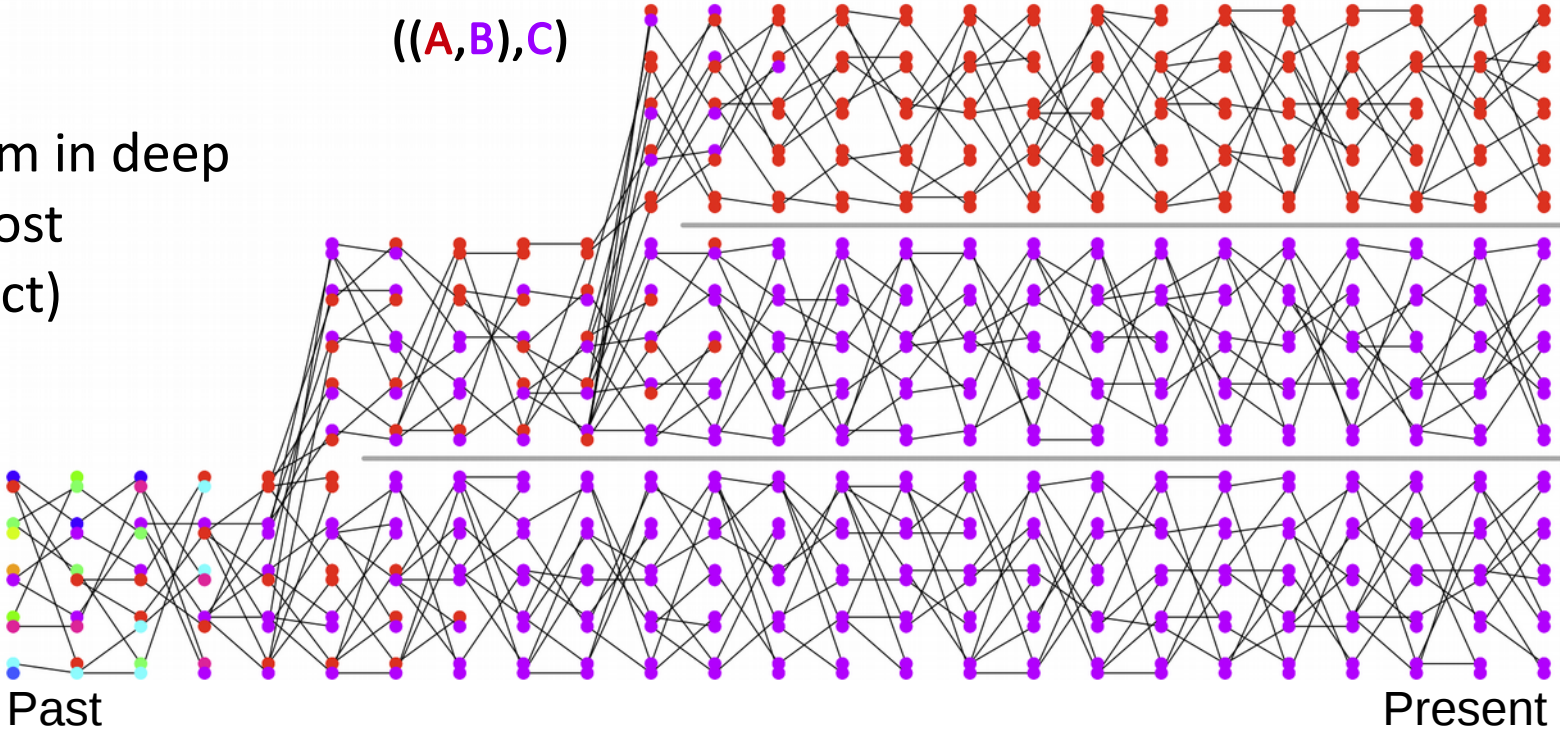


Figure credits:
Coop Lab UC Davis
gcbias.org

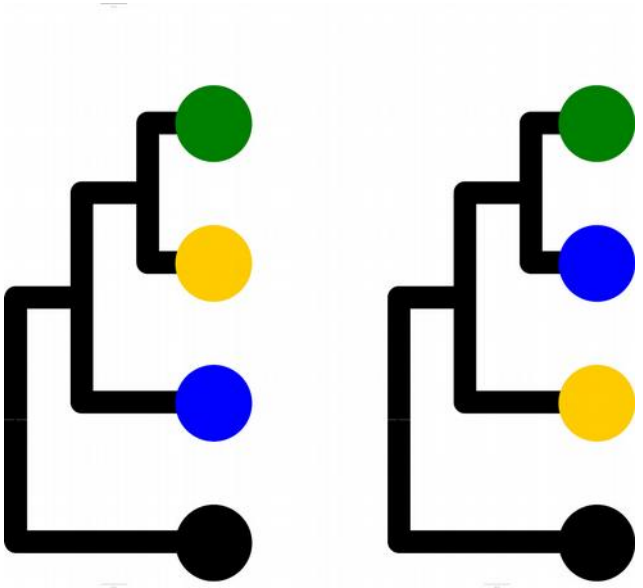
Deep coalescence

How does it present in the data?

Gene tree incongruence

But if this is the only issue:

Alleles from each species are relatively closely related



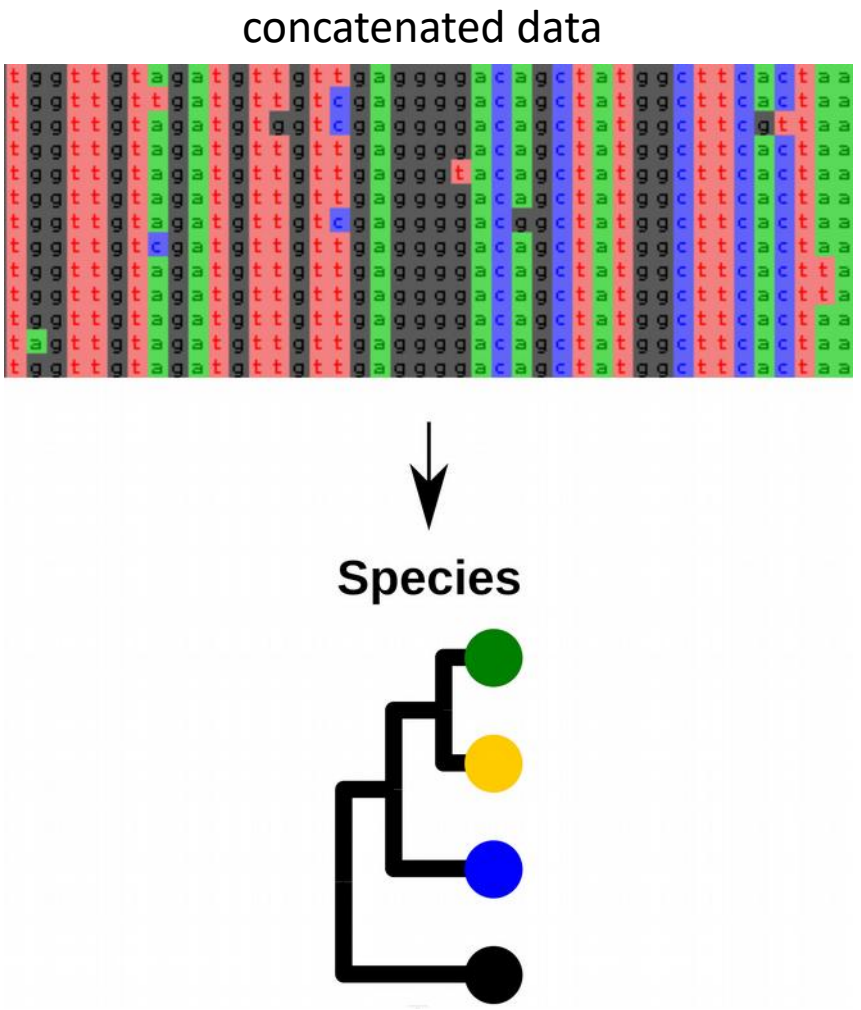
Deep coalescence

How to infer the species tree?

Easiest option:

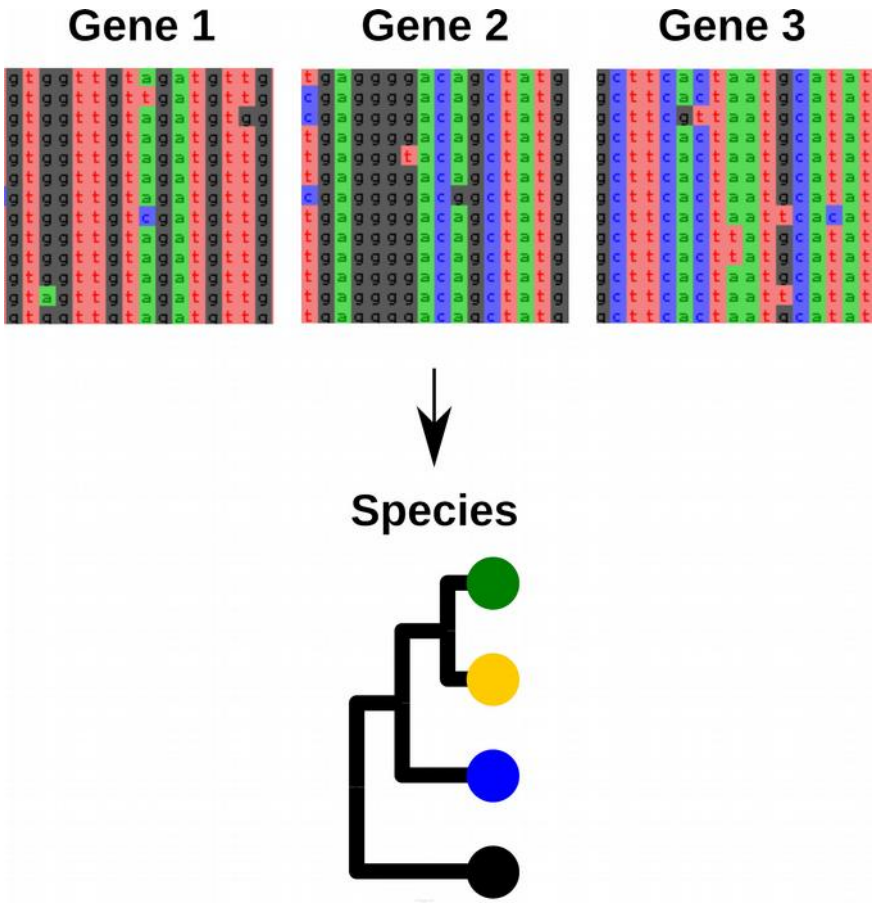
Ignore the problem and use concatenation

Often works well enough (Smith & Hahn, 2020)



Bayesian multi-gene coalescent

- StarBEAST, www.beast2.org
- Estimates species tree and all gene trees at the same time
- Ultrametric tree, thus rooted



Downsides:

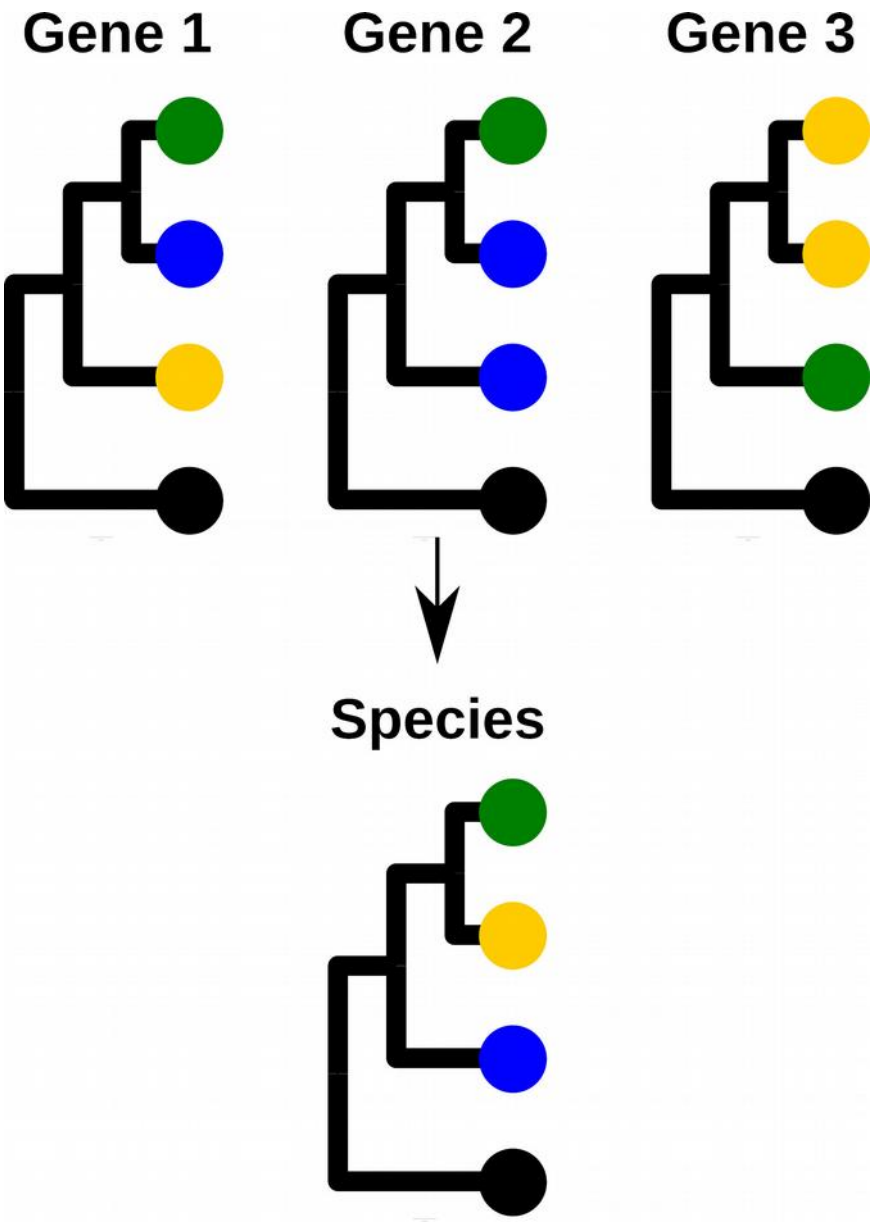
- Computationally intensive, slow
- Problems with missing data / patchy matrix

Short-cut methods

- Many options, e.g. ASTRAL https://github.com/smirarab/ASTRAL
- Infer species tree from gene trees
- Can deal with missing data
- Extremely fast

Downsides:

- Gene tree topologies fixed
- Less reliable for deeper phylogenetics (Smith & Hahn, 2020)
- Branch lengths often meaningless
- Needs outgroup rooting



Gene duplication and loss

Paralogs and orthologs

Gene duplication and loss

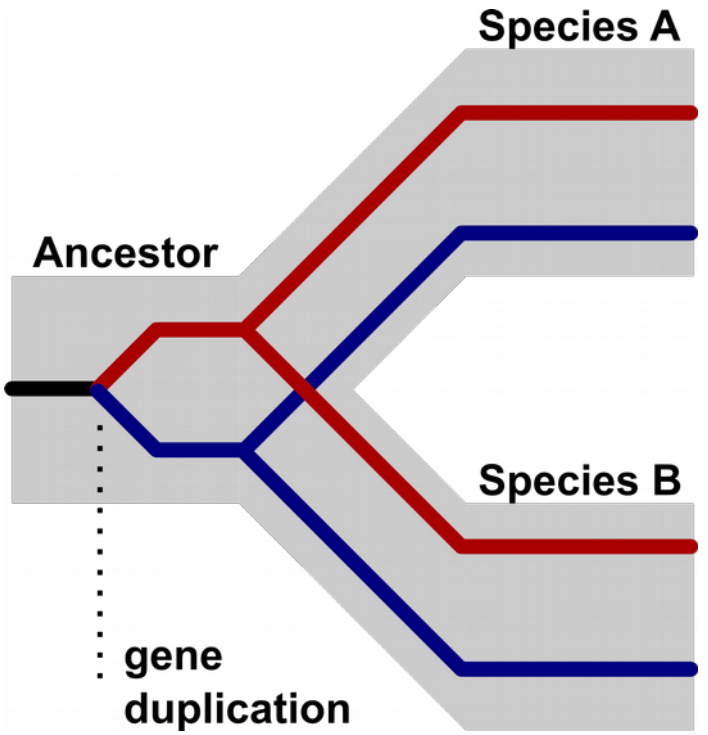
What happened?

- Gene duplication or
- Genome duplication

Both copies of original gene retained in the descendants

Potentially specialisation of gene functions

Genes can also, of course, be lost

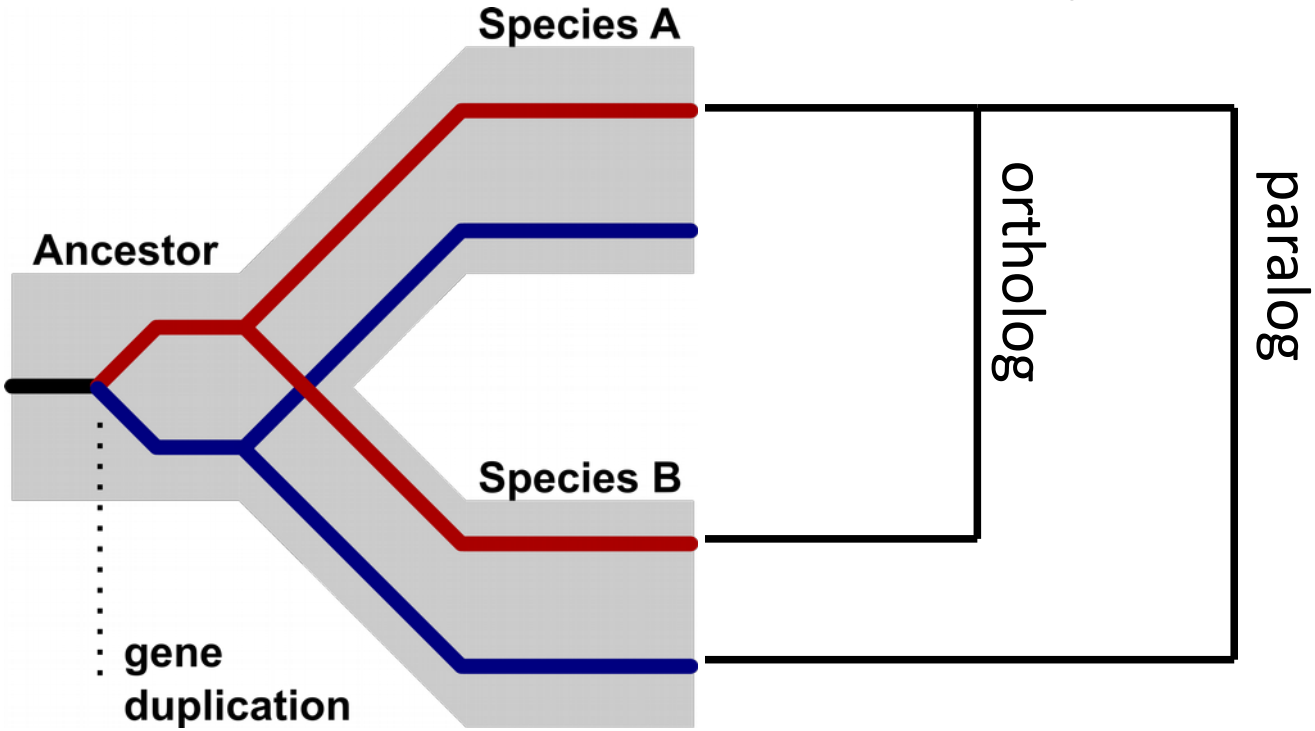


Gene duplication and loss

Orthologs = descended from same ancestral sequence after gene duplication

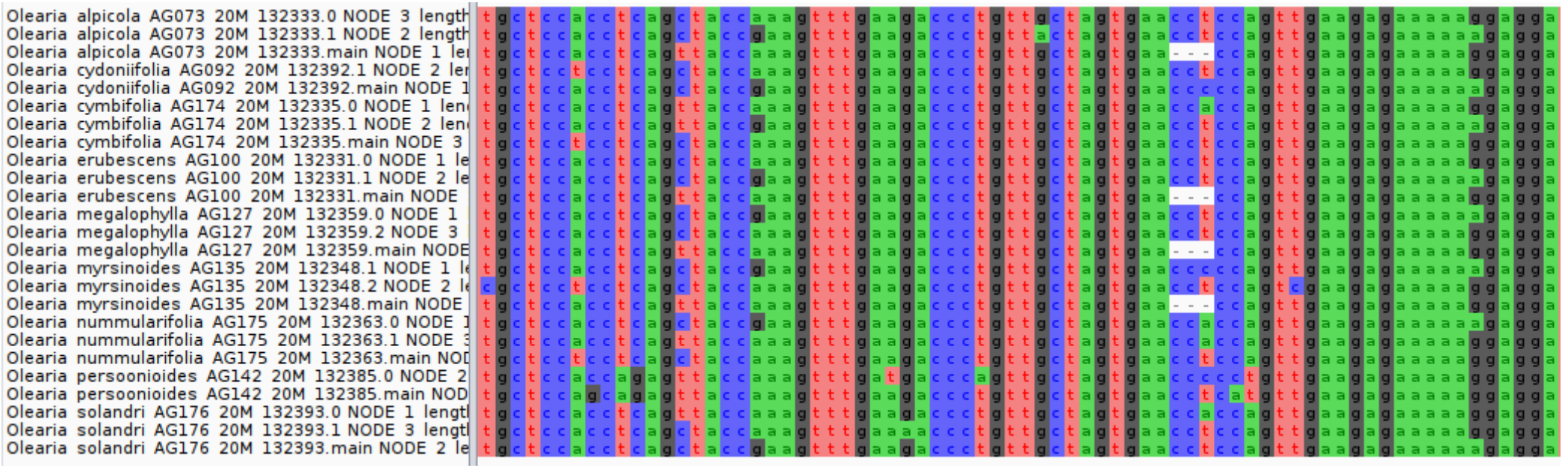
Paralogs = descended from different ancestral sequences after gene duplication

→ gene tree interferes with species tree



Example alignment

Note how the same patterns reoccur in several samples.

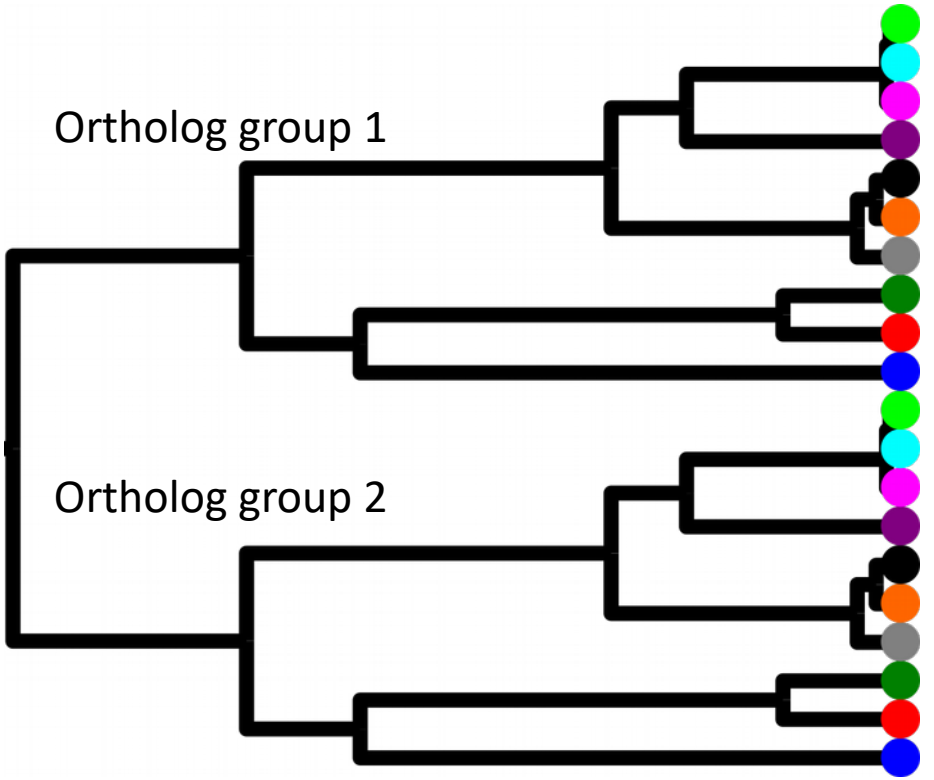


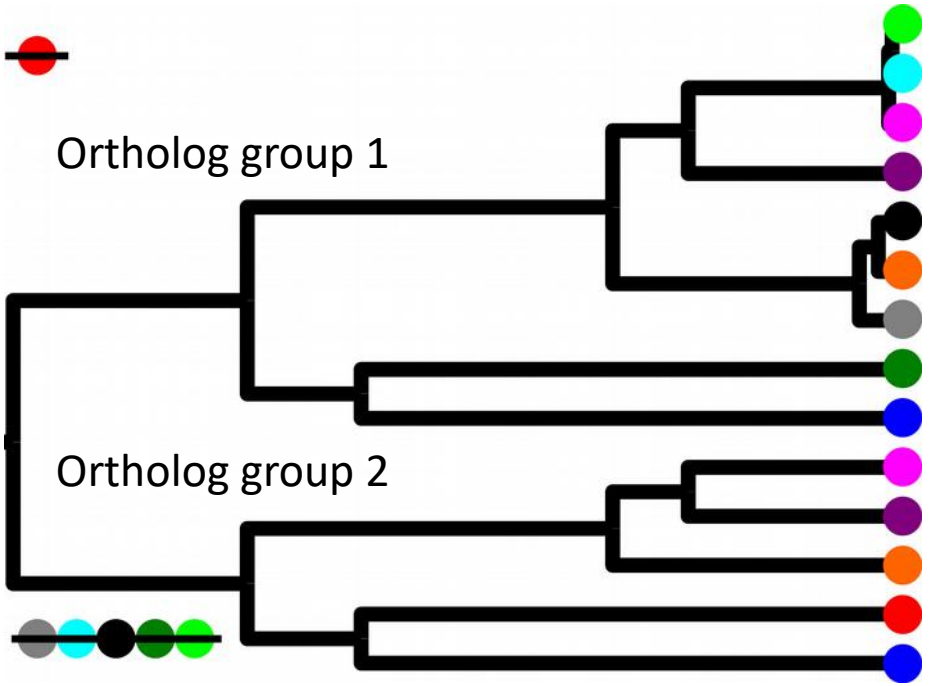
Gene duplication and loss

How does it present in the data?

Ideally, species replicated in N parts of gene tree for N gene duplication events

Realistically, gene losses and failure to amplify make gene trees less complete →



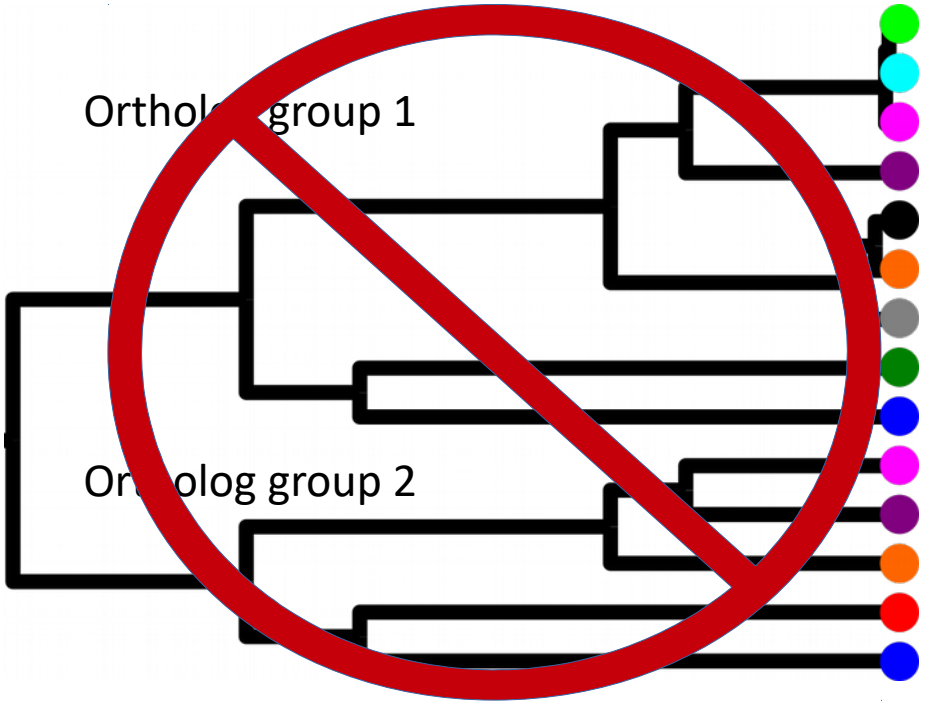


Gene duplication and loss

How to infer the species tree

Easiest option:

Chuck out genes with paralogs

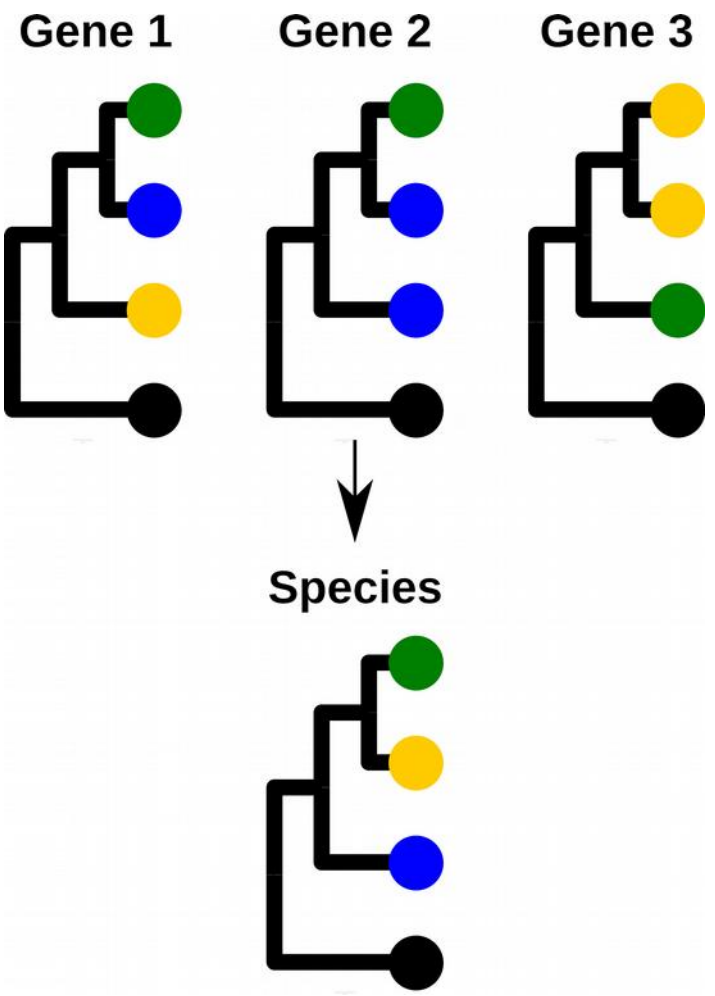


Gene duplication and loss

How to infer the species tree

Shortcut methods

- Infer species tree directly from gene trees incl. paralogs (same problem as before).
- Minimise Gene Duplications and Losses, e.g. iGTP.
- Likelihood, e.g. GeneRax.

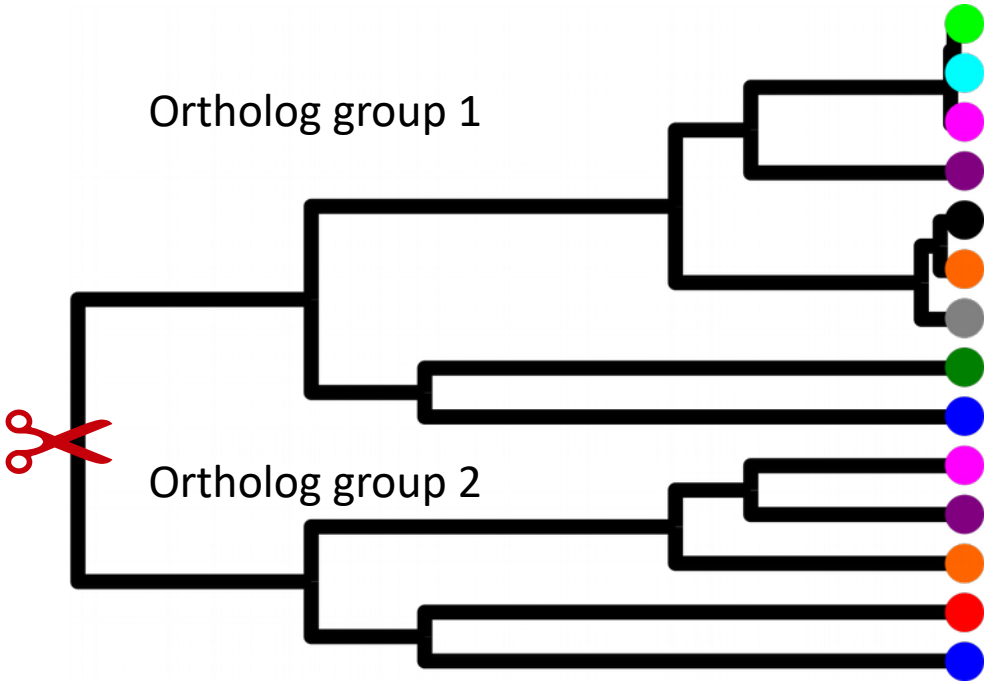


Gene duplication and loss

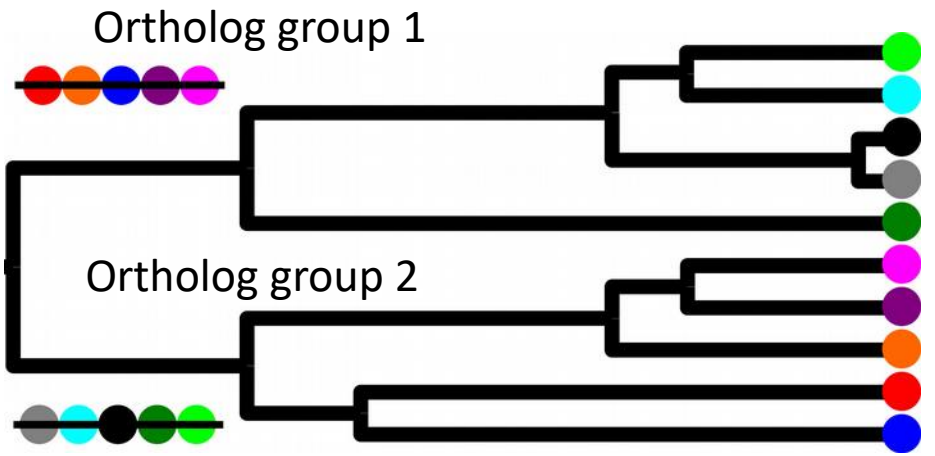
How to infer the species tree

Bioinformatically separate ortholog groups using gene tree topologies

(Yang & Smith, 2014)



Problem if gene tree too incomplete →

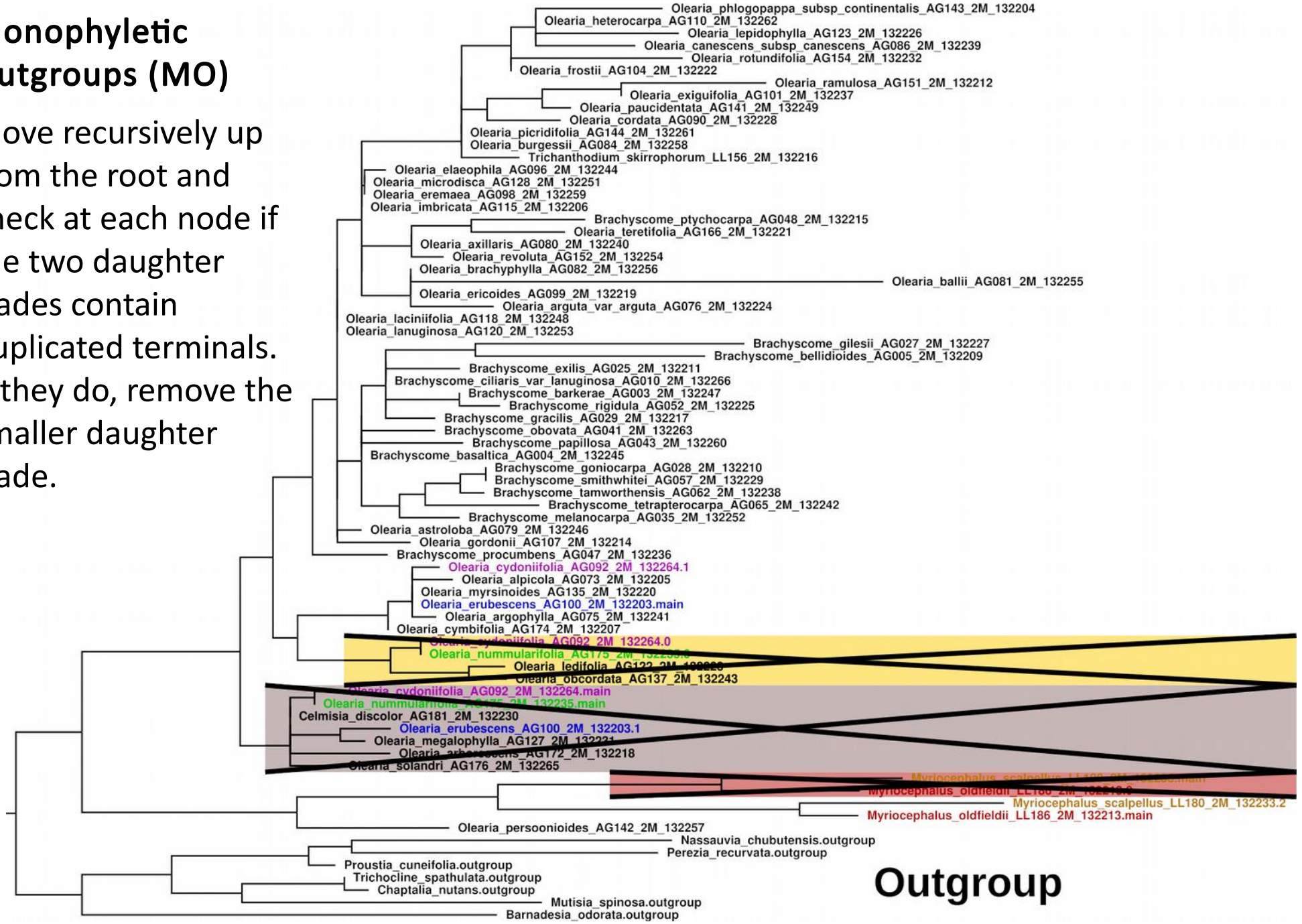


...but can't tell, because no sample 2x in gene tree

Approach #1:

Monophyletic Outgroups (MO)

Move recursively up from the root and check at each node if the two daughter clades contain duplicated terminals. If they do, remove the smaller daughter clade.



Approach #3:

Maximum Inclusion (MI)

Dissects the unrooted gene tree into pieces without duplicated terminals, starting with largest clade

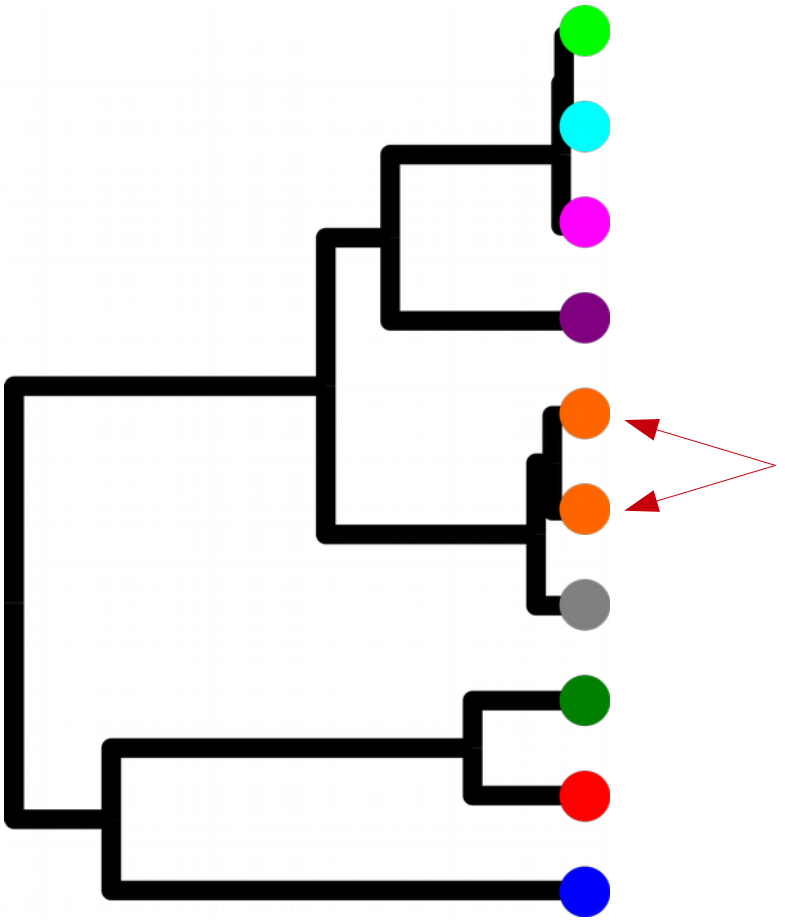
Not recommended by Yang & Smith

When paralogy is not an issue

Duplication in terminal branch of phylogeny is irrelevant:

Will present as indistinguishable from different alleles

Cannot mislead phylogenetic analysis



Reticulation

Hybridisation, introgression, allopolyploidy, chloroplast capture

Reticulation

What happened?

- Allopolyploidy / hybrid speciation
- Introgression / back-crossing / admixture
- Chloroplast (organelle) capture

Allopolyploid speciation

What happened?

Genome duplication restores fertility of hybrid, produces hybridogenic species

Example:

Spearmint (*Menta spicata*)

= *M. longifolia* x *suaveolens*

x

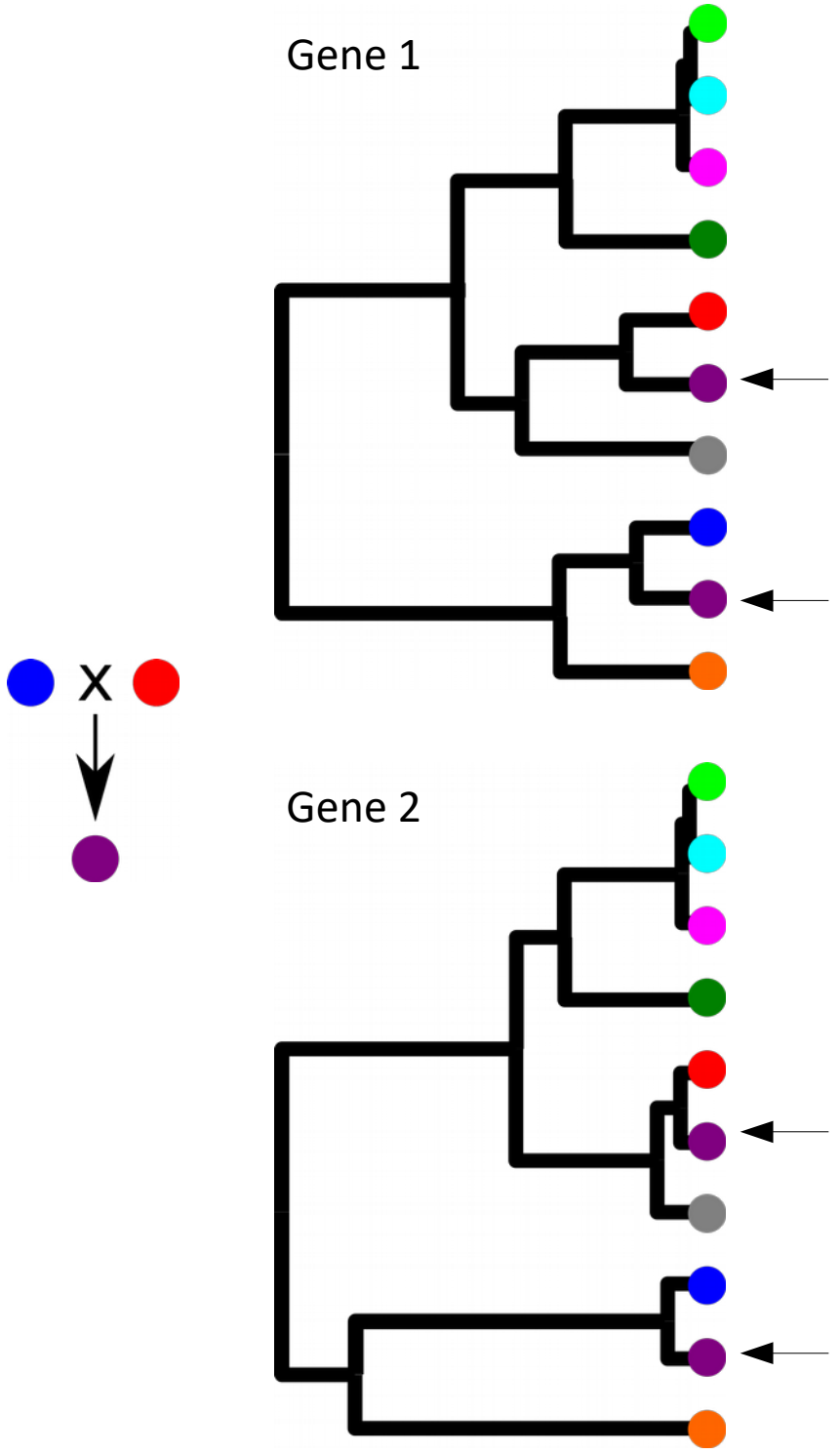
genome duplication

Allopolyploid speciation

How does it present in the data?

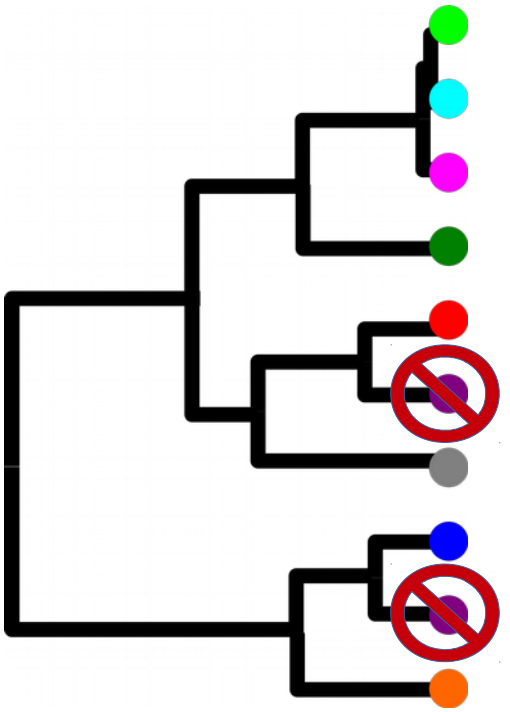
Ideally, species or clade placed with two parental species across all gene trees

But: gene losses, gene tree incongruence



How to infer the species tree

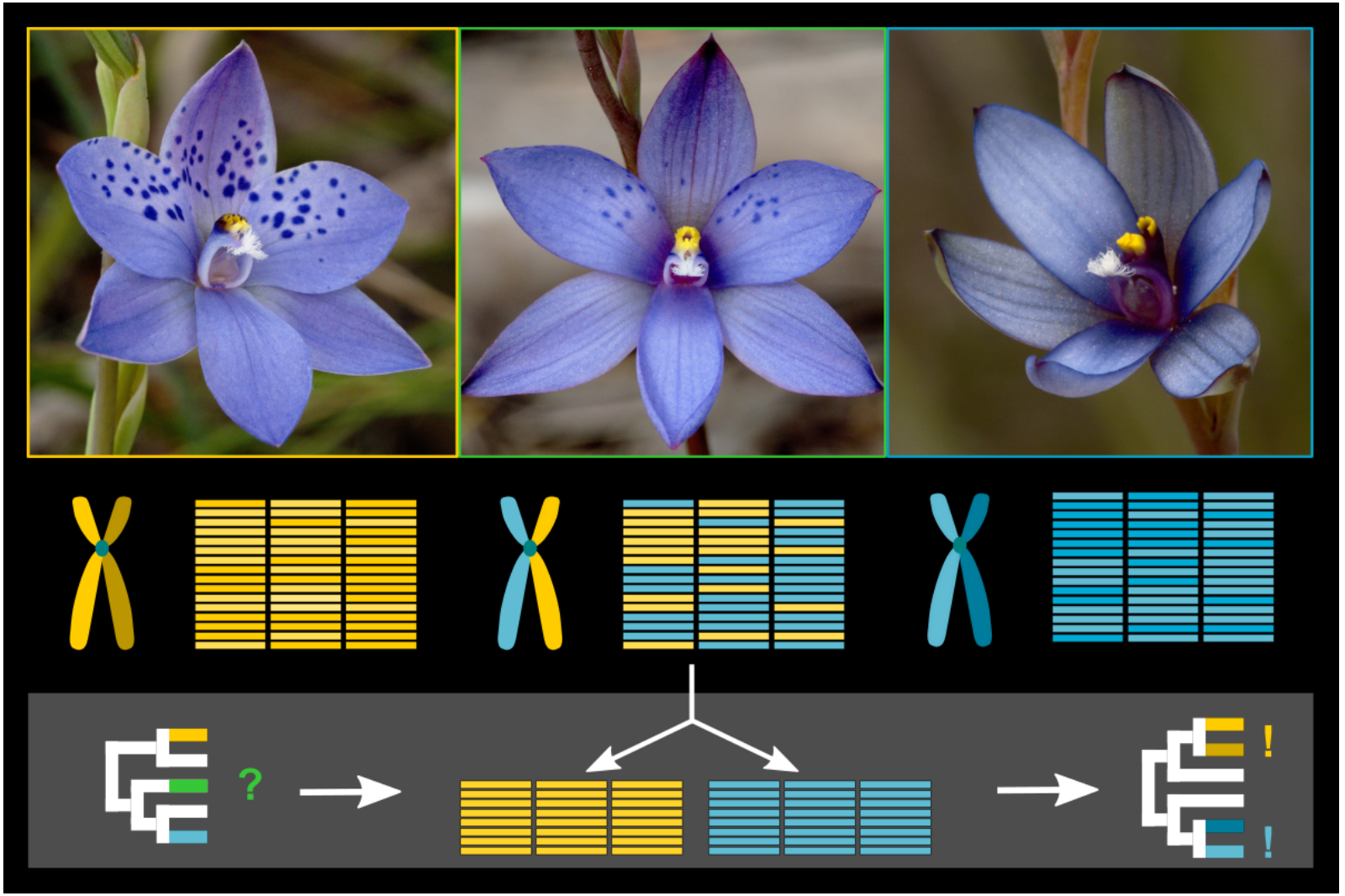
Easiest option: Remove hybrids & hybridogenic lineages from analyses that assume tree-like structure of data



Allele phasing

HybPhaser pipeline, separate talk by Lars Nauheimer

https://doi.org/10.1101/2020.10.27.354589



Australian
BioCommons

WEBINAR: Detection of and phasing of hybrid accessions in a target capture dataset

Thursday, 10 June 2021

12:00 pm – 1:00 pm

Showcasing HybPhaser, a novel bioinformatics workflow for detecting and phase hybrids in target capture datasets.

Presenter: Dr Lars Nauheimer, Australian Tropical Herbarium

More information here

Register here

https://www.biocommons.org.au/events/hybphaser

Introgression

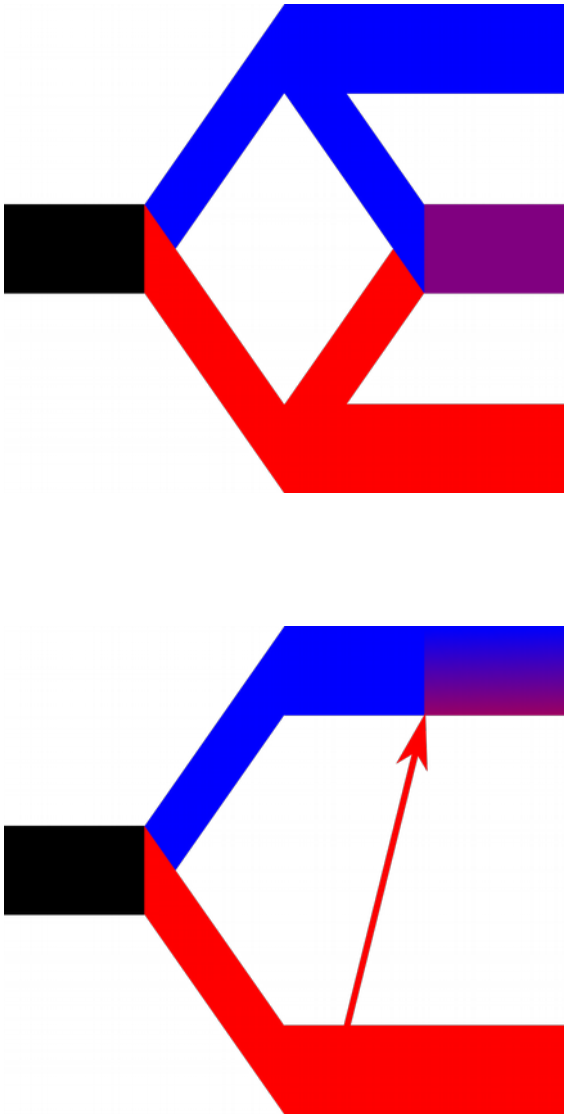
Limited gene flow between species (back-crossing hybrids)

How does it present in the data?

Only few genes affected →

Problem:

How to distinguish from deep coalescence?



Distinguishing deep coalescence and reticulation

"ABBA BABA" test

Phylogeny of three species and outgroup should have many (((A,A),B),B)

Do we have more (((A,B),B),A) and (((B,A),B),A) than expected?

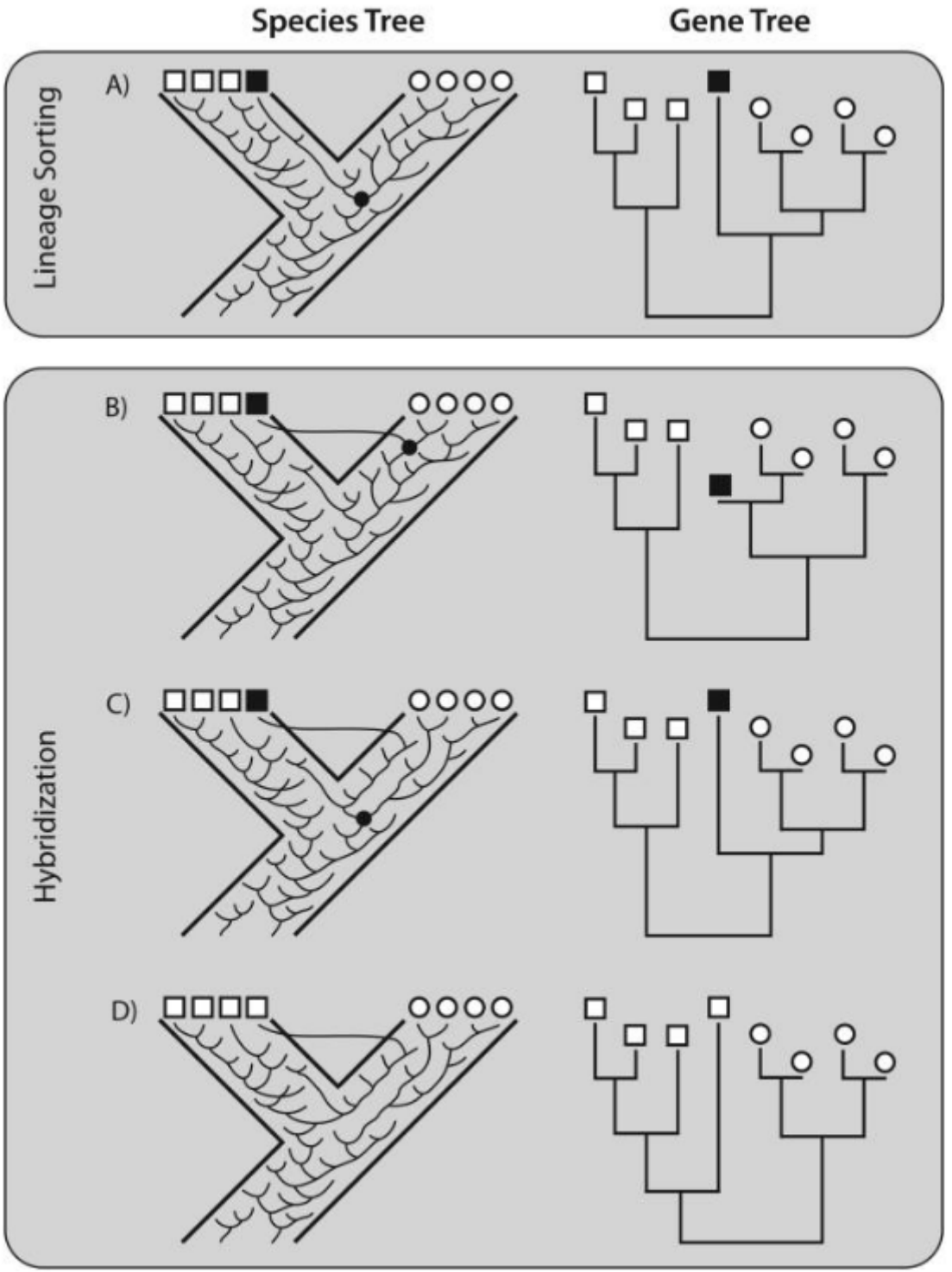
Uses allele frequencies: best for SNP datasets and multiple individuals per species.

http://evomics.org/learning/population-and-speciation-genomics/2018-population-and-speciation-genomics/abba-baba-statistics/

Distinguishing deep coalescence and reticulation

Simulation test using age of gene tree coalescence versus age of species tree coalescence (Joly et al. 2009, https://doi.org/10.1086/600082)

But doesn't always work, see cases C, D in Joly's figure



Phylogenetic network analyses

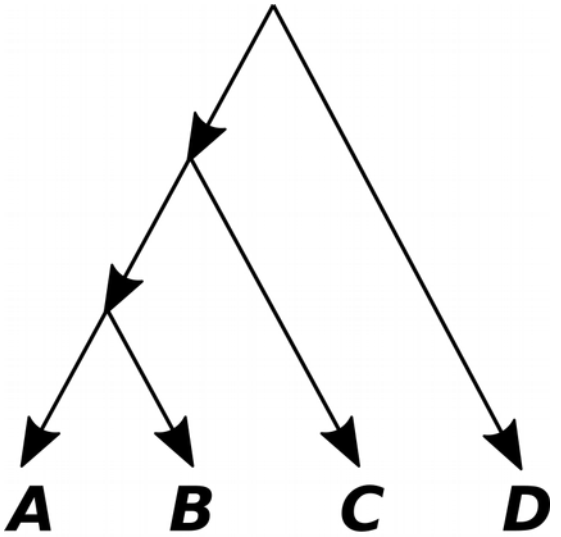
Analyses modelling species tree
with and without gene flow, e.g.

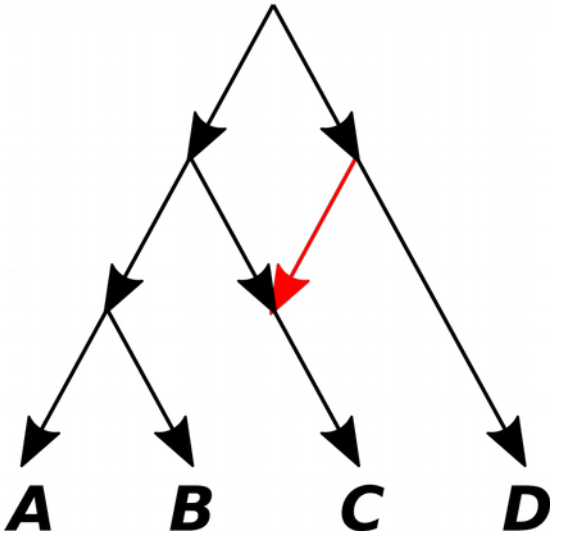
BPP, https://github.com/bpp

Species Networks applying Quartets
(SNaQ; Solís-Lemus & Ané, 2016)
https://github.com/crsl4/PhyloNetworks.jl

Computationally intensive,
slow and limited to few species

Various methods in PhyloNet:
https://bioinfocs.rice.edu/phylonet





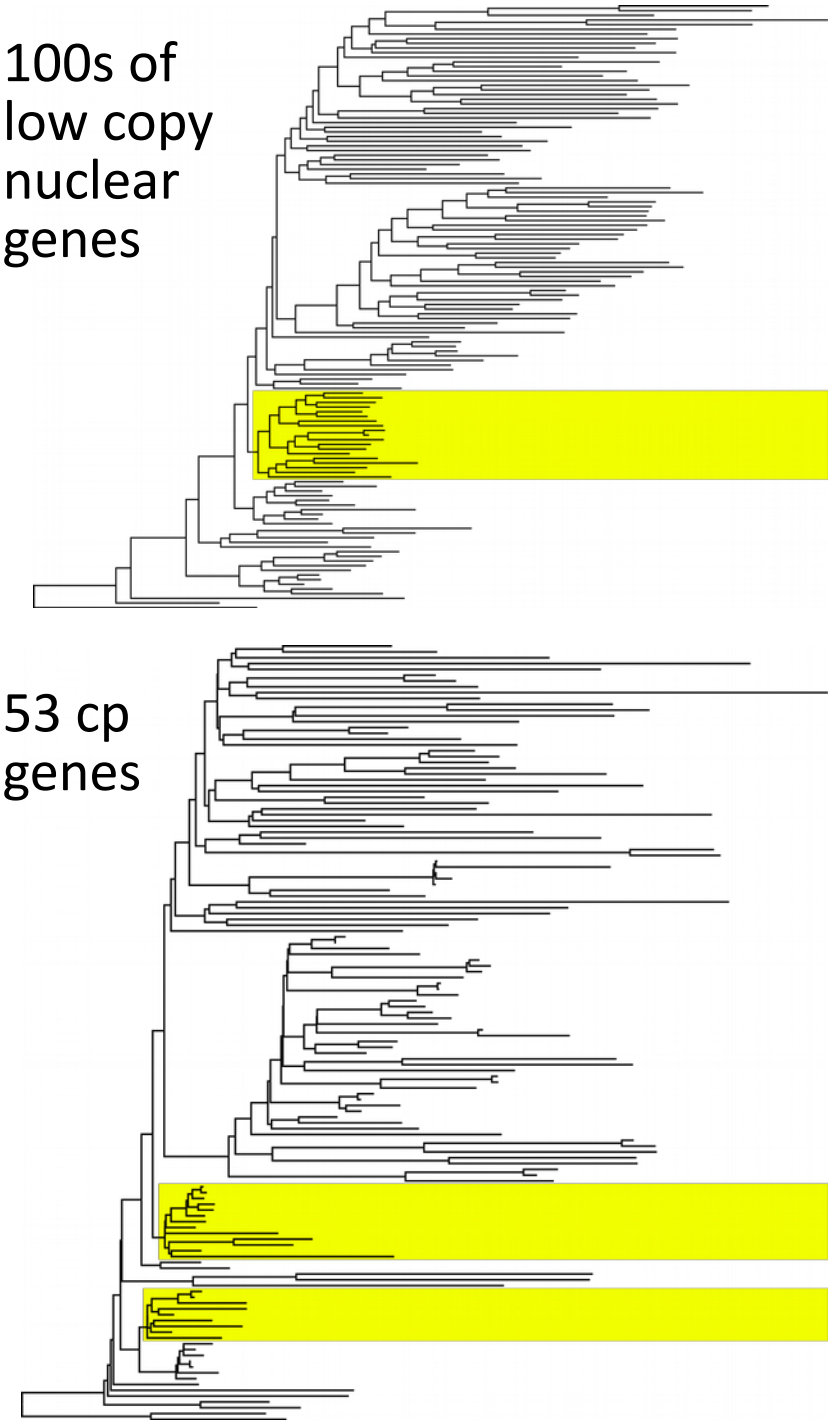
Chloroplast capture

Organelles jump species more easily than nuclear genes (Stegemann et al., 2012)

How does it present in the data?

Nuclear data ± consistently support one topology, chloroplast another

E.g., *Cassinia* group in Australian Asteraceae →



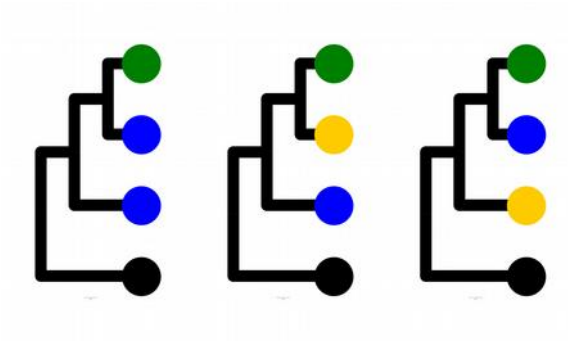
What if case isn’t clear?

Consider biological plausibility

Choice of approach depends on assumptions

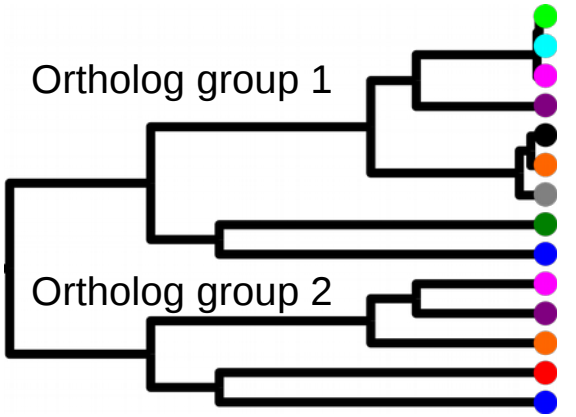
The more distant, the less likely are hybridisation and deep coalescence

Summary



Deep coalescence:

- Large pop sizes & rapid speciations
- Phylogenetic methods available

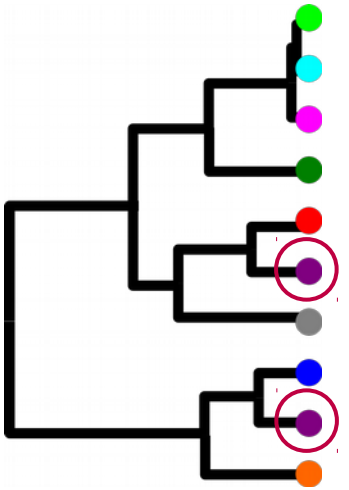


Paralogy

- Ancestral gene/genome duplication
- Bioinformatic & shortcut methods

Reticulation

- Hybrid speciation or introgression
- Phylogenetic methods for few-species cases



Ideal cases easy to recognise, but IRL...

And all of them can happen in the same phylogeny!

Background, overview and theory

Altenhoff AM, Dessimoz C, 2012. Inferring Orthology and Paralogy, pp. 259-279 in: Maria Anisimova (ed.), Evolutionary Genomics: Statistical and Computational Methods, Volume 1,Methods in Molecular Biology, vol. 855, Springer. https://people.inf.ethz.ch/adriaal/orthology-bookchapter.pdf

Barker MS, Arrigo N, Baniaga AE, Li Z, Levin DA, 2016. On the relative abundance of autopolyploids and allopolyploids. New Phytologist 210: 391–398. https://doi.org/10.1111/nph.13698

Folk RA, Soltis PS, Soltis DE, Guralnick R, 2018. New prospects in the detection and comparative analysis of hybridization in the tree of life. Amer. J. Bot. 105: 364-375. https://doi.org/10.1002/ajb2.1018

Maddison WP, 1997. Gene trees in species trees. Syst. Biol. 46: 523-536. https://doi.org/10.1093/sysbio/46.3.523

Nakhleh L, 2013. Computational approaches to species phylogeny inference and gene tree reconciliation. Trends in Ecology & Evolution 28: 719-728. https://doi.org/10.1016/j.tree.2013.09.004

Payseur, B. A., and L. H. Rieseberg. 2016. A genomic perspective on hybridization and speciation. Molecular Ecology 25: 2337–2360. https://doi.org/10.1111/mec.13557

Peer YV de, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. Nature Reviews Genetics 18: 411–424. http://dx.doi.org/10.1038/nrg.2017.26

Smith ML, Hahn MW, 2020. New approaches for inferring phylogenies in the presence of paralogs. Trends in Genetics 1721. https://doi.org/10.1016/j.tig.2020.08.012

Stegemann S, Keuthe M, Greiner S, Bock R, 2012. Horizontal transfer of chloroplast genomes between plant species. Proc. Natl. Acad. Sci. 109: 2434–2438. https://doi.org/10.1073/pnas.1114076109

Individual methods and software

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. https://doi.org/10.1093/molbev/msz296

Heled J, Drummond AJ, 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27: 570-580. https://doi.org/10.1093/molbev/msp274

Johnson MG, et al., 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. Syst. Biol. 68: 594-606. https://doi.org/10.1093/sysbio/syy086

Joly S, McLenachan PA, Lockhart PJ, 2009. A statistical approach for distinguishing hybridization and incomplete lineage sorting. Amer. Nat. 174: E54-E70. https://doi.org/10.1086/600082

Morel B, Schade P, Lutteropp S, Williams TA, Szöllősi GJ, Stamatakis A, 2021. SpeciesRax: A tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss. https://doi.org/10.1101/2021.03.29.437460

Nauheimer L, Weigner N, Joyce E, Crayn D, Clarke C, Nargar K, 2020. HybPhaser: a workflow for the detection and phasing of hybrids in target capture datasets. BioRxiv. https://doi.org/10.1101/2020.10.27.354589

Solís-Lemus C, Ané C, 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genetics 12: e1005896. https://doi.org/10.1371/journal.pgen.1005896

Yang Y, Smith SA, 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. Mol. Biol. Evol. 31: 3081–3092. https://doi.org/10.1093/molbev/msu245

Zhang C, Rabiee M, Sayyari E, Mirarab S, 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19: 153. https://doi.org/10.1186/s12859-018-2129-y