Phylogenetic Network Estimation for *Wolbachia*

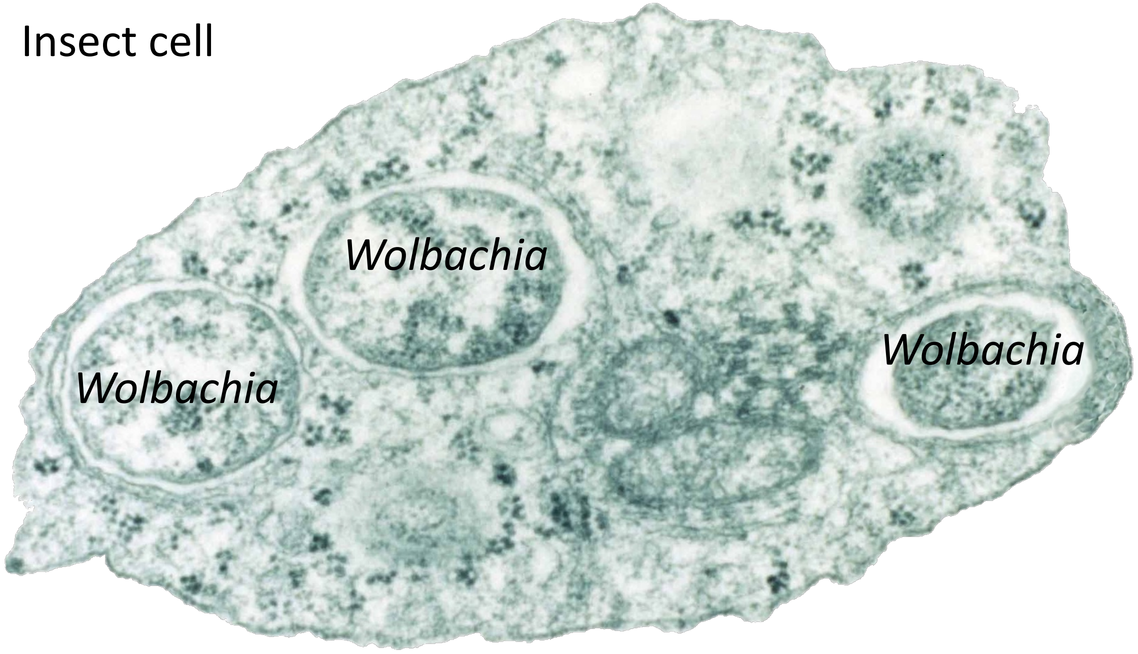
Paul Zaharias, Danny Rice, Tandy Warnow, and Irene Newton

The parasitic bacteria *Wolbachia*

- Infects the reproductive tracts of arthropods and nematods

- 25 to 70% of all insect species are potential hosts

- Impacts on human health (e.g., Dengue fever)



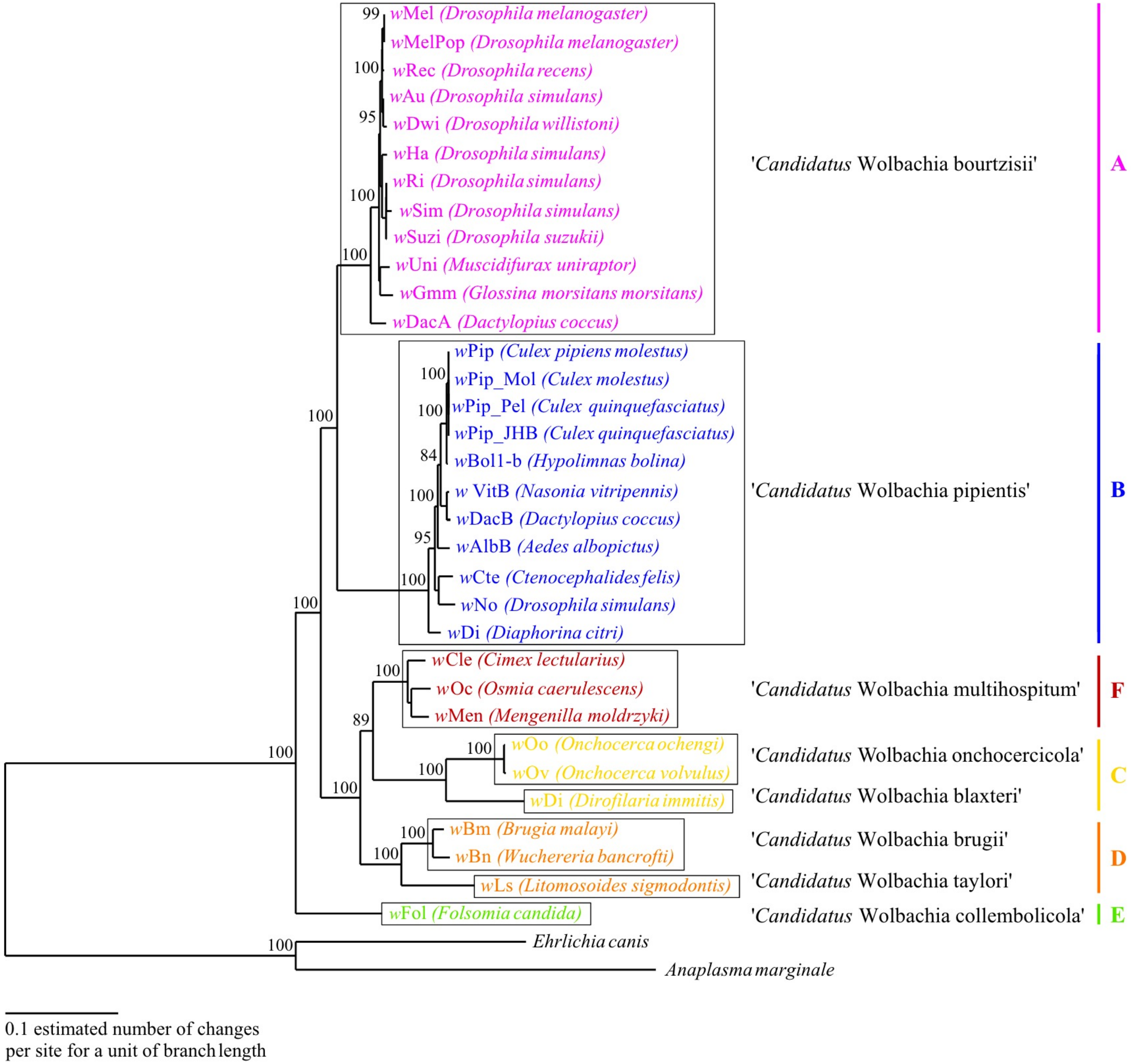
Scott O'Neill — *Genome Sequence of the Intracellular Bacterium Wolbachia.* PLoS Biol 2/3/2004: e76. [doi:10.1371/journal.pbio.0020076](https://doi.org/10.1371/journal.pbio.0020076)

Questions of interest

- What genes are under selection?
- Can we improve species delimitation?
- What is the percentage of horizontally-transferred genes in *Wolbachia?*

One of the trees from the literature

Ramírez-Puebla et al.,
Syst. and Appl. Microbiology, 2016



99
wMel (Drosophila melanogaster)
wMelPop (Drosophila melanogaster)
100
wRec (Drosophila recens)
wAu (Drosophila simulans)
95
wDwi (Drosophila willistoni)
wHa (Drosophila simulans)
wRi (Drosophila simulans)
100
wSim (Drosophila simulans)
wSuzi (Drosophila suzukii)
100
wUni (Muscidifurax uniraptor)
wGmm (Glossina morsitans morsitans)
wDacA (Dactylopius coccus)
'Candidatus Wolbachia bourtzisii'
A
wPip (Culex pipiens molestus)
100
wPip_Mol (Culex molestus)
100
wPip_Pel (Culex quinquefasciatus)
wPip_JHB (Culex quinquefasciatus)
84
wBol1-b (Hypolimnas bolina)
100
w VitB (Nasonia vitripennis)
wDacB (Dactylopius coccus)
95
wAlbB (Aedes albopictus)
100
wCte (Ctenocephalides felis)
wNo (Drosophila simulans)
wDi (Diaphorina citri)
100
100
'Candidatus Wolbachia pipientis'
B
wCle (Cimex lectularius)
100
wOc (Osmia caerulescens)
wMen (Mengenilla moldrzyki)
'Candidatus Wolbachia multihospitum'
F
89
100
wOo (Onchocerca ochengi)
100
wOv (Onchocerca volvulus)
wDi (Dirofilaria immitis)
'Candidatus Wolbachia onchocercicola'
'Candidatus Wolbachia blaxteri'
C
100
100
wBm (Brugia malayi)
100
wBn (Wuchereria bancrofti)
wLs (Litomosoides sigmodontis)
'Candidatus Wolbachia brugii'
'Candidatus Wolbachia taylori'
D
100
wFol (Folsomia candida)
'Candidatus Wolbachia collembolicola'
E
100
Ehrlichia canis
Anaplasma marginale
0.1 estimated number of changes
per site for a unit of branch length

Ramírez-Puebla, Shamayim T., et al. "A response to Lindsey et al." Wolbachia pipientis should not be split into multiple species: A response to Ramírez-Puebla et al."." *Systematic and applied microbiology* 39.3 (2016): 223-225.

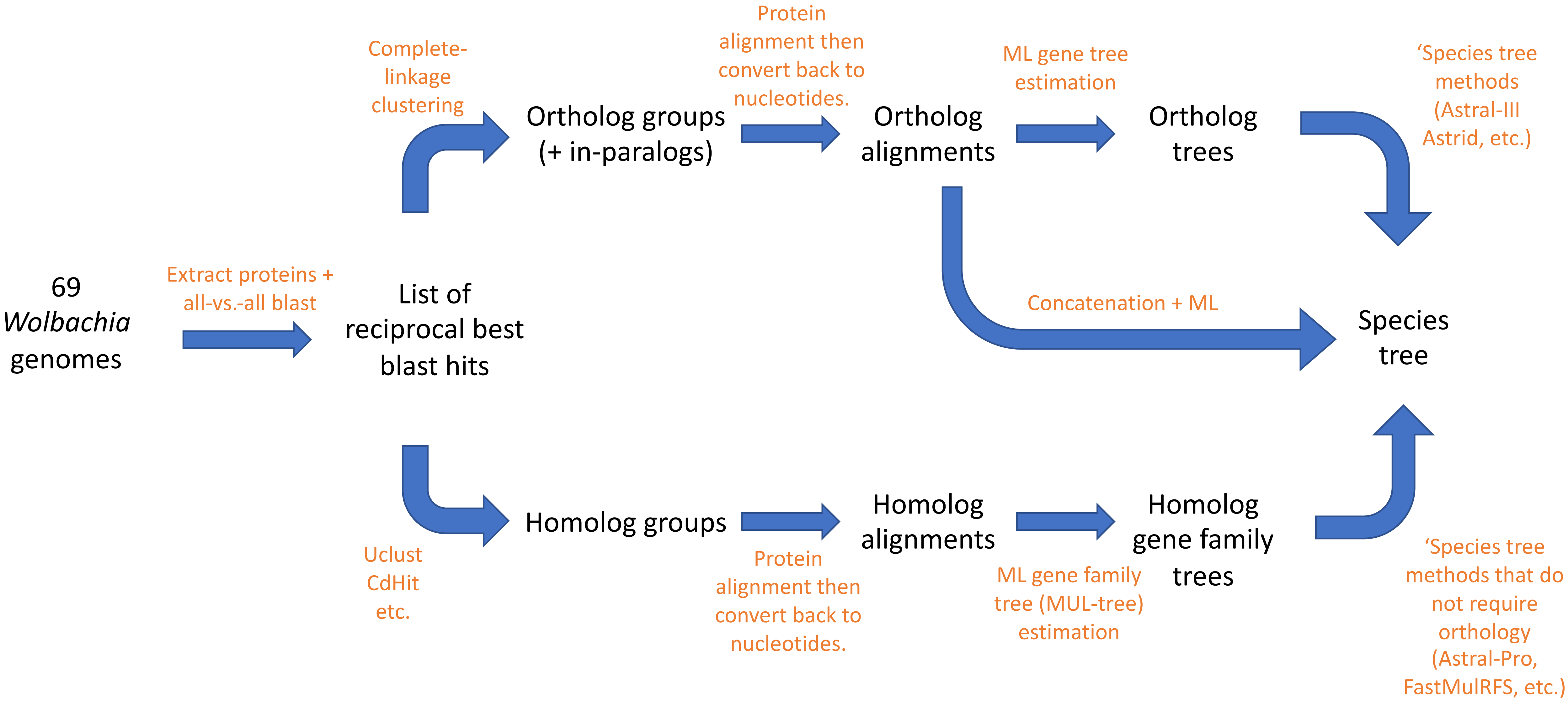
Species tree estimation options

- From orthologs:
    - Concatenation using maximum likelihood: statistically inconsistent, computationally intensive, can be accurate
    - Summary methods (that combine gene trees), such as ASTRAL, ASTRID: faster than concatenation, statistically consistent if given true gene trees, can be accurate
    - Site-based methods, such as LilyQ and SVDquartets/SVDquest: statistically consistent, limited to perhaps 100 leaves due to computational challenges
- The choice between methods depends on the data

Species tree estimation options

- From homologs:
  - Construct gene family trees and combine
    - ASTRAL-Pro (variant of ASTRAL to handle gene family trees)
    - DISCO+ASTRID (decomposes gene family trees into single copy trees, then runs ASTRID)
    - FastMulRFS (supertree method applied to MUL-trees)
  - Variant of concatenation
    - Construct gene family trees, decompose using DISCO, then use concatenation
- Statistical consistency without knowing orthology is achieved in some cases
- The choice between methods depends on the data

What we did: genomes to species trees



69 *Wolbachia* genomes

Extract proteins + all-vs.-all blast

List of reciprocal best blast hits

Complete-linkage clustering

Ortholog groups (+ in-paralogs)

Protein alignment then convert back to nucleotides.

Ortholog alignments

ML gene tree estimation

Ortholog trees

‘Species tree methods (Astral-III Astrid, etc.)

Concatenation + ML

Species tree

Uclust
CdHit
etc.

Homolog groups

Protein alignment then convert back to nucleotides.

Homolog alignments

ML gene family tree (MUL-tree) estimation

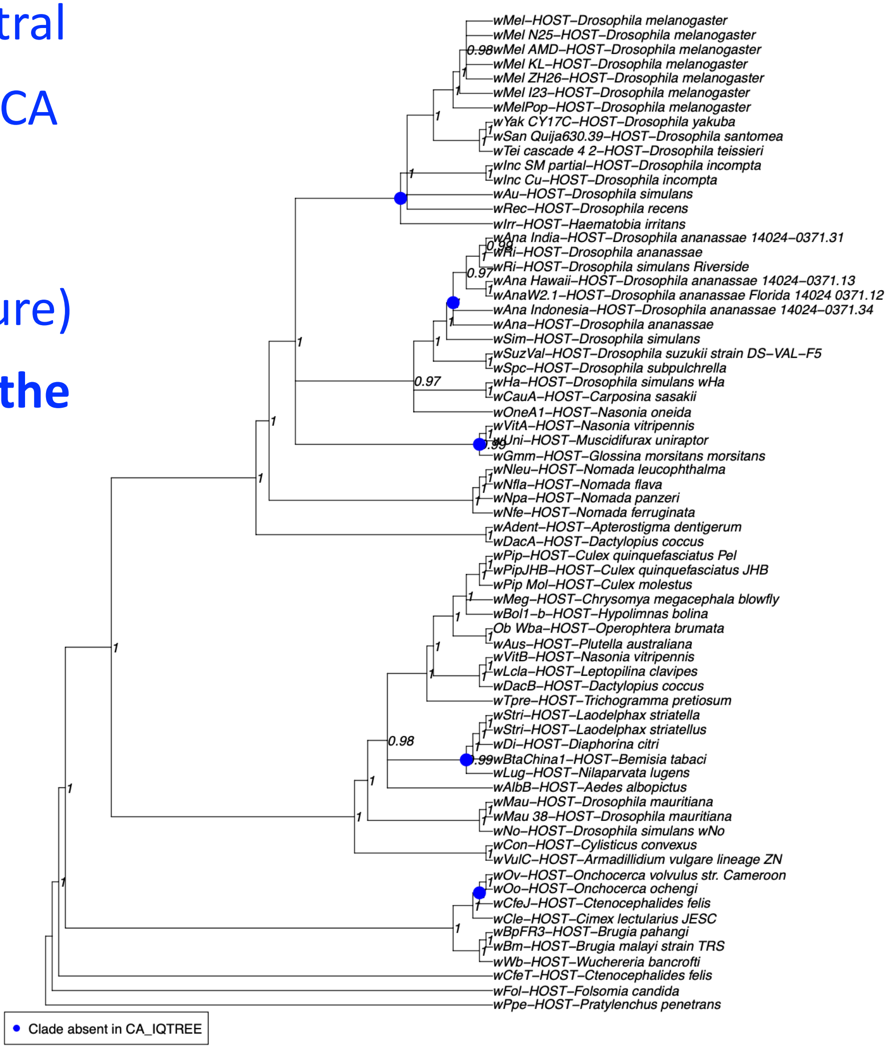
Homolog gene family trees

‘Species tree methods that do not require orthology (Astral-Pro, FastMulRFS, etc.)

ASTRAL and Concatenation-IQTree: very similar

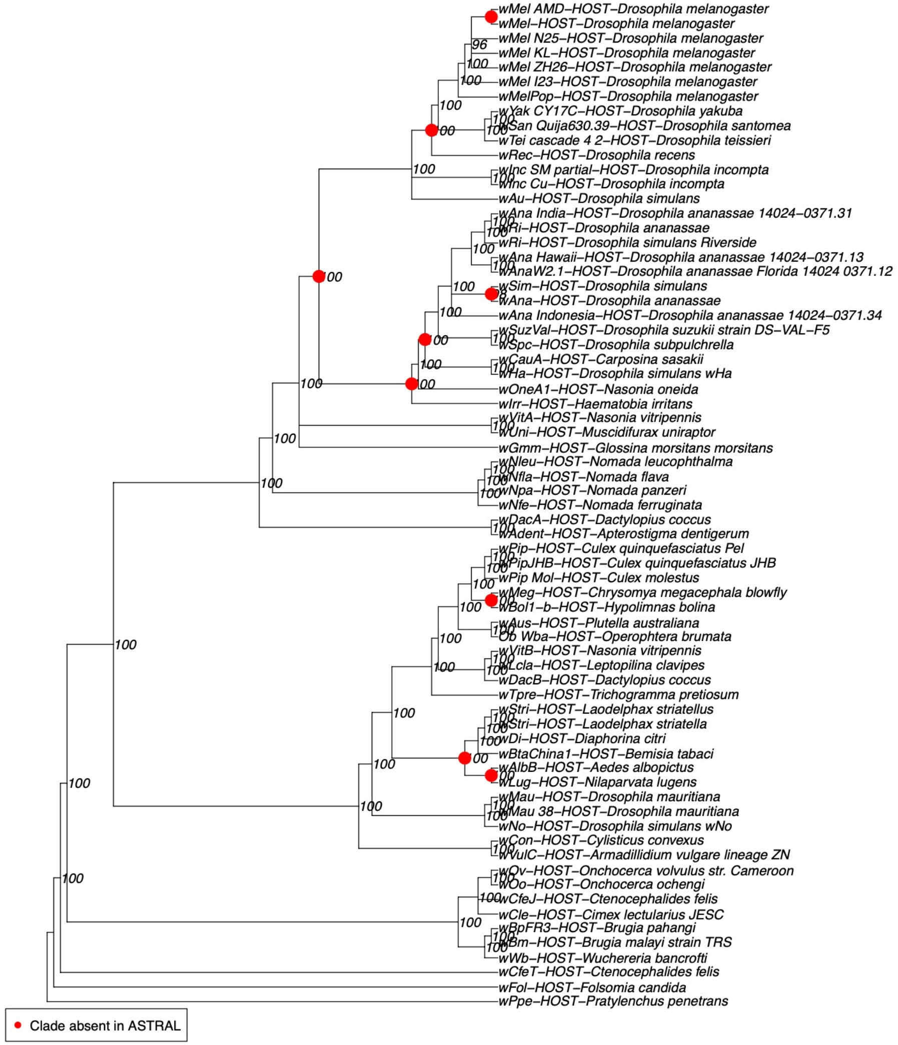
- Collapsed pp < 0.95 for Astral
- Collapsed UfBoot < 95 for CA
- Both trees are very well supported, but with some differences (marked in figure)
- **Are this incompatibilities the result of HGT events or something else?**

Astral



wMel–HOST–Drosophila melanogaster
wMel N25–HOST–Drosophila melanogaster
wMel AMD–HOST–Drosophila melanogaster
wMel KL–HOST–Drosophila melanogaster
wMel ZH26–HOST–Drosophila melanogaster
wMel I23–HOST–Drosophila melanogaster
wMelPop–HOST–Drosophila melanogaster
wYak CY17C–HOST–Drosophila yakuba
wSan Quija630.39–HOST–Drosophila santomea
wTei cascade 4 2–HOST–Drosophila teissieri
wInc SM partial–HOST–Drosophila incompta
wInc Cu–HOST–Drosophila incompta
wAu–HOST–Drosophila simulans
wRec–HOST–Drosophila recens
wIrr–HOST–Haematobia irritans
wAna India–HOST–Drosophila ananassae 14024–0371.31
wRi–HOST–Drosophila ananassae
wRi–HOST–Drosophila simulans Riverside
wAna Hawaii–HOST–Drosophila ananassae 14024–0371.13
wAnaW2.1–HOST–Drosophila ananassae Florida 14024 0371.12
wAna Indonesia–HOST–Drosophila ananassae 14024–0371.34
wAna–HOST–Drosophila ananassae
wSim–HOST–Drosophila simulans
wSuzVal–HOST–Drosophila suzukii strain DS–VAL–F5
wSpc–HOST–Drosophila subpulchrella
wHa–HOST–Drosophila simulans wHa
wCauA–HOST–Carposina sasakii
wOneA1–HOST–Nasonia oneida
wVitA–HOST–Nasonia vitripennis
wUni–HOST–Muscidifurax uniraptor
wGmm–HOST–Glossina morsitans morsitans
wNleu–HOST–Nomada leucophthalma
wNfla–HOST–Nomada flava
wNpa–HOST–Nomada panzeri
wNfe–HOST–Nomada ferruginata
wAdent–HOST–Apterostigma dentigerum
wDacA–HOST–Dactylopius coccus
wPip–HOST–Culex quinquefasciatus Pel
wPipJHB–HOST–Culex quinquefasciatus JHB
wPip Mol–HOST–Culex molestus
wMeg–HOST–Chrysomya megacephala blowfly
wBol1–b–HOST–Hypolimnas bolina
Ob Wba–HOST–Operophtera brumata
wAus–HOST–Plutella australiana
wVitB–HOST–Nasonia vitripennis
wLcla–HOST–Leptopilina clavipes
wDacB–HOST–Dactylopius coccus
wTpre–HOST–Trichogramma pretiosum
wStri–HOST–Laodelphax striatella
wStri–HOST–Laodelphax striatellus
wDi–HOST–Diaphorina citri
wBtaChina1–HOST–Bemisia tabaci
wLug–HOST–Nilaparvata lugens
wAlbB–HOST–Aedes albopictus
wMau–HOST–Drosophila mauritiana
wMau 38–HOST–Drosophila mauritiana
wNo–HOST–Drosophila simulans wNo
wCon–HOST–Cylisticus convexus
wVulC–HOST–Armadillidium vulgare lineage ZN
wOv–HOST–Onchocerca volvulus str. Cameroon
wOo–HOST–Onchocerca ochengi
wCfeJ–HOST–Ctenocephalides felis
wCle–HOST–Cimex lectularius JESC
wBpFR3–HOST–Brugia pahangi
wBm–HOST–Brugia malayi strain TRS
wWb–HOST–Wuchereria bancrofti
wCfeT–HOST–Ctenocephalides felis
wFol–HOST–Folsomia candida
wPpe–HOST–Pratylenchus penetrans
● Clade absent in CA_IQTREE

CA (IQ-TREE)



wMel AMD–HOST–Drosophila melanogaster
wMel–HOST–Drosophila melanogaster
wMel N25–HOST–Drosophila melanogaster
wMel KL–HOST–Drosophila melanogaster
wMel ZH26–HOST–Drosophila melanogaster
wMel I23–HOST–Drosophila melanogaster
wMelPop–HOST–Drosophila melanogaster
wYak CY17C–HOST–Drosophila yakuba
wSan Quija630.39–HOST–Drosophila santomea
wTei cascade 4 2–HOST–Drosophila teissieri
wRec–HOST–Drosophila recens
wInc SM partial–HOST–Drosophila incompta
wInc Cu–HOST–Drosophila incompta
wAu–HOST–Drosophila simulans
wAna India–HOST–Drosophila ananassae 14024–0371.31
wRi–HOST–Drosophila ananassae
wRi–HOST–Drosophila simulans Riverside
wAna Hawaii–HOST–Drosophila ananassae 14024–0371.13
wAnaW2.1–HOST–Drosophila ananassae Florida 14024 0371.12
wSim–HOST–Drosophila simulans
wAna–HOST–Drosophila ananassae
wAna Indonesia–HOST–Drosophila ananassae 14024–0371.34
wSuzVal–HOST–Drosophila suzukii strain DS–VAL–F5
wSpc–HOST–Drosophila subpulchrella
wCauA–HOST–Carposina sasakii
wHa–HOST–Drosophila simulans wHa
wOneA1–HOST–Nasonia oneida
wIrr–HOST–Haematobia irritans
wVitA–HOST–Nasonia vitripennis
wUni–HOST–Muscidifurax uniraptor
wGmm–HOST–Glossina morsitans morsitans
wNleu–HOST–Nomada leucophthalma
wNfla–HOST–Nomada flava
wNpa–HOST–Nomada panzeri
wNfe–HOST–Nomada ferruginata
wDacA–HOST–Dactylopius coccus
wAdent–HOST–Apterostigma dentigerum
wPip–HOST–Culex quinquefasciatus Pel
wPipJHB–HOST–Culex quinquefasciatus JHB
wPip Mol–HOST–Culex molestus
wMeg–HOST–Chrysomya megacephala blowfly
wBol1–b–HOST–Hypolimnas bolina
wAus–HOST–Plutella australiana
Ob Wba–HOST–Operophtera brumata
wVitB–HOST–Nasonia vitripennis
wLcla–HOST–Leptopilina clavipes
wDacB–HOST–Dactylopius coccus
wTpre–HOST–Trichogramma pretiosum
wStri–HOST–Laodelphax striatellus
wStri–HOST–Laodelphax striatella
wDi–HOST–Diaphorina citri
wBtaChina1–HOST–Bemisia tabaci
wAlbB–HOST–Aedes albopictus
wLug–HOST–Nilaparvata lugens
wMau–HOST–Drosophila mauritiana
wMau 38–HOST–Drosophila mauritiana
wNo–HOST–Drosophila simulans wNo
wCon–HOST–Cylisticus convexus
wVulC–HOST–Armadillidium vulgare lineage ZN
wOv–HOST–Onchocerca volvulus str. Cameroon
wOo–HOST–Onchocerca ochengi
wCfeJ–HOST–Ctenocephalides felis
wCle–HOST–Cimex lectularius JESC
wBpFR3–HOST–Brugia pahangi
wBm–HOST–Brugia malayi strain TRS
wWb–HOST–Wuchereria bancrofti
wCfeT–HOST–Ctenocephalides felis
wFol–HOST–Folsomia candida
wPpe–HOST–Pratylenchus penetrans
● Clade absent in ASTRAL

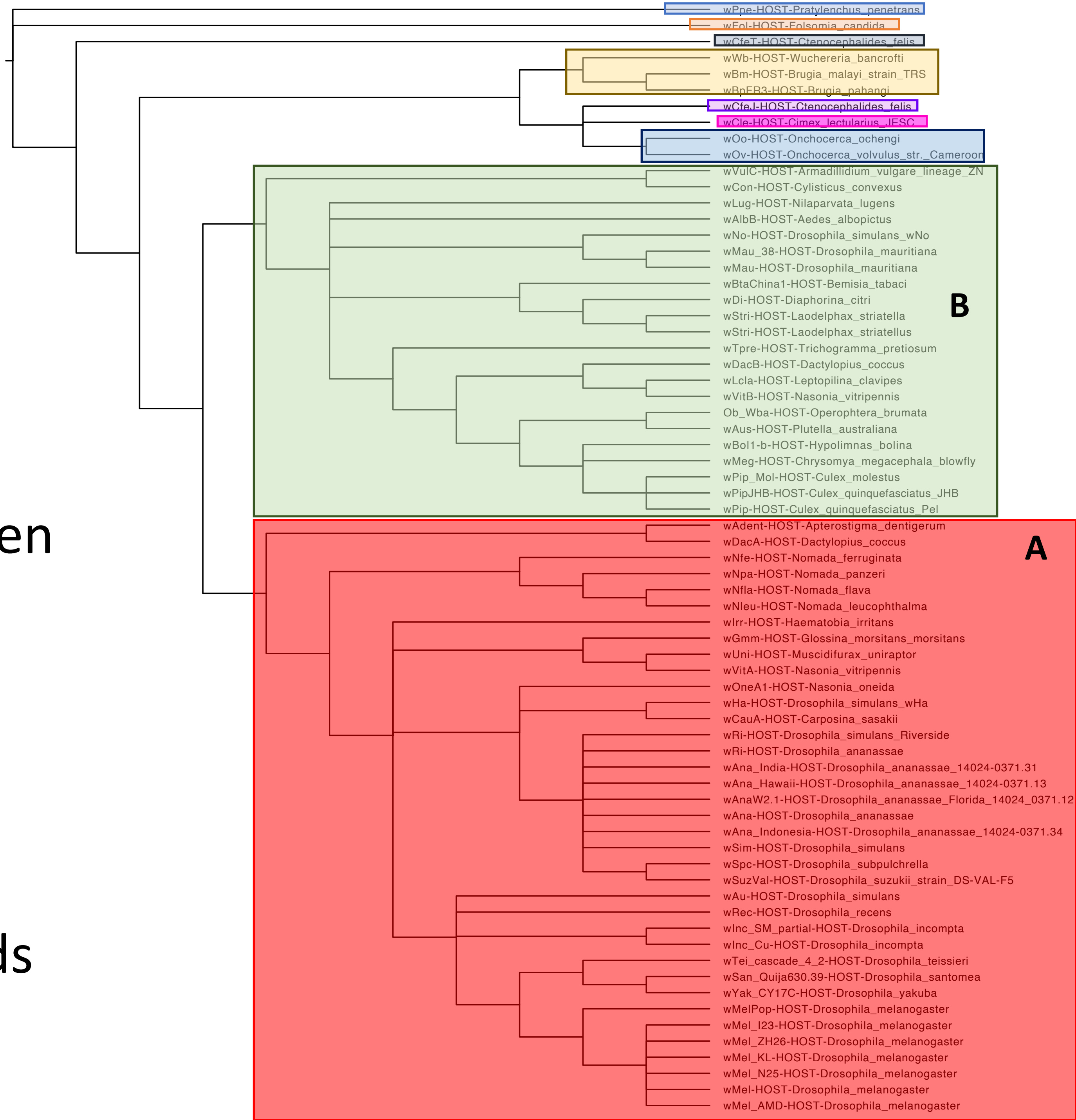
Reticulation events

- Consensus tree

- Between supergroups (= ‘species’) ? In particular, between supergroup A and B (most studied)

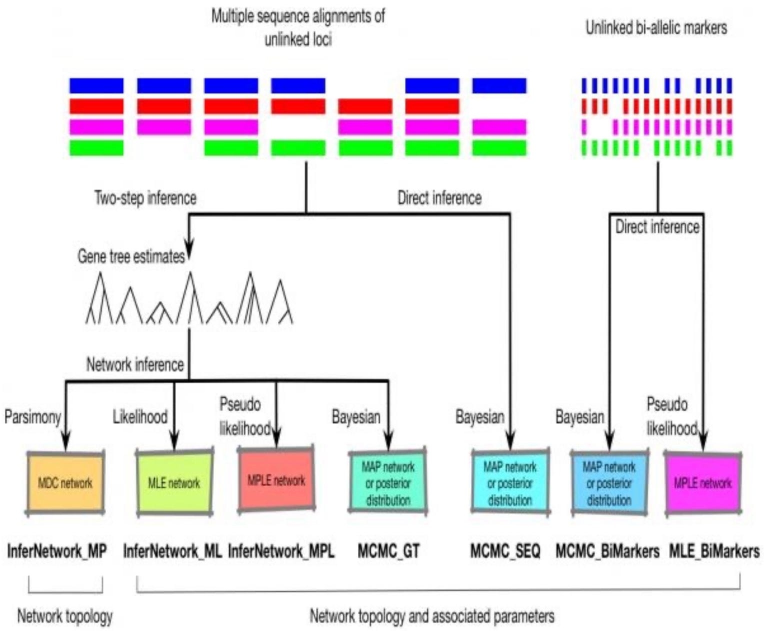
- Within ‘species’ ?

→ Phylogenetic network methods


wPpe-HOST-Pratylenchus_penetrans
wFol-HOST-Folsomia_candida
wCfeT-HOST-Ctenocephalides_felis
wWb-HOST-Wuchereria_bancrofti
wBm-HOST-Brugia_malayi_strain_TRS
wBpFR3-HOST-Brugia_pahangi
wCfeJ-HOST-Ctenocephalides_felis
wCle-HOST-Cimex_lectularius_JESC
wOo-HOST-Onchocerca_ochengi
wOv-HOST-Onchocerca_volvulus_str._Cameroon
wVulC-HOST-Armadillidium_vulgare_lineage_ZN
wCon-HOST-Cylisticus_convexus
wLug-HOST-Nilaparvata_lugens
wAlbB-HOST-Aedes_albopictus
wNo-HOST-Drosophila_simulans_wNo
wMau_38-HOST-Drosophila_mauritiana
wMau-HOST-Drosophila_mauritiana
wBtaChina1-HOST-Bemisia_tabaci
wDi-HOST-Diaphorina_citri
wStri-HOST-Laodelphax_striatella
wStri-HOST-Laodelphax_striatellus
wTpre-HOST-Trichogramma_pretiosum
wDacB-HOST-Dactylopius_coccus
wLcla-HOST-Leptopilina_clavipes
wVitB-HOST-Nasonia_vitripennis
Ob_Wba-HOST-Operophtera_brumata
wAus-HOST-Plutella_australiana
wBol1-b-HOST-Hypolimnas_bolina
wMeg-HOST-Chrysomya_megacephala_blowfly
wPip_Mol-HOST-Culex_molestus
wPipJHB-HOST-Culex_quinquefasciatus_JHB
wPip-HOST-Culex_quinquefasciatus_Pel
B
wAdent-HOST-Apterostigma_dentigerum
wDacA-HOST-Dactylopius_coccus
wNfe-HOST-Nomada_ferruginata
wNpa-HOST-Nomada_panzeri
wNfla-HOST-Nomada_flava
wNleu-HOST-Nomada_leucophthalma
wIrr-HOST-Haematobia_irritans
wGmm-HOST-Glossina_morsitans_morsitans
wUni-HOST-Muscidifurax_uniraptor
wVitA-HOST-Nasonia_vitripennis
wOneA1-HOST-Nasonia_oneida
wHa-HOST-Drosophila_simulans_wHa
wCauA-HOST-Carposina_sasakii
wRi-HOST-Drosophila_simulans_Riverside
wRi-HOST-Drosophila_ananassae
wAna_India-HOST-Drosophila_ananassae_14024-0371.31
wAna_Hawaii-HOST-Drosophila_ananassae_14024-0371.13
wAnaW2.1-HOST-Drosophila_ananassae_Florida_14024_0371.12
wAna-HOST-Drosophila_ananassae
wAna_Indonesia-HOST-Drosophila_ananassae_14024-0371.34
wSim-HOST-Drosophila_simulans
wSpc-HOST-Drosophila_subpulchrella
wSuzVal-HOST-Drosophila_suzukii_strain_DS-VAL-F5
wAu-HOST-Drosophila_simulans
wRec-HOST-Drosophila_recens
wInc_SM_partial-HOST-Drosophila_incompta
wInc_Cu-HOST-Drosophila_incompta
wTei_cascade_4_2-HOST-Drosophila_teissieri
wSan_Quija630.39-HOST-Drosophila_santomea
wYak_CY17C-HOST-Drosophila_yakuba
wMelPop-HOST-Drosophila_melanogaster
wMel_I23-HOST-Drosophila_melanogaster
wMel_ZH26-HOST-Drosophila_melanogaster
wMel_KL-HOST-Drosophila_melanogaster
wMel_N25-HOST-Drosophila_melanogaster
wMel-HOST-Drosophila_melanogaster
wMel_AMD-HOST-Drosophila_melanogaster
A

Phylonet (Nakhleh) and PhyloNetworks (Solis-Lemus)

Phylonet (Luay Nakhleh)



PhyloNetworks (Claudia Solís-Lemus)

PhyloNetworks.jl

snaQ

Phylogenetic Network Pipeline

- In all analyses, assumption is you have gene trees and you are seeking a phylogenetic network
- Three variants of the problem:
  - Given an underlying species tree, add reticulations (extra edges) to create network (easier)
  - Given an underlying species tree, add reticulations and allow tree to change (a bit harder)
  - Construct the phylogenetic network directly (much harder)

Phylogenetic Network Pipeline

- In all analyses, assumption is you have gene trees and you are seeking a phylogenetic network
- Three variants of the problem:
  - Given an underlying species tree, add reticulations (extra edges) to create network (easier)
  - Given an underlying species tree, add reticulations and allow tree to change (a bit harder)
  - Construct the phylogenetic network directly (much harder)

How do you proceed?

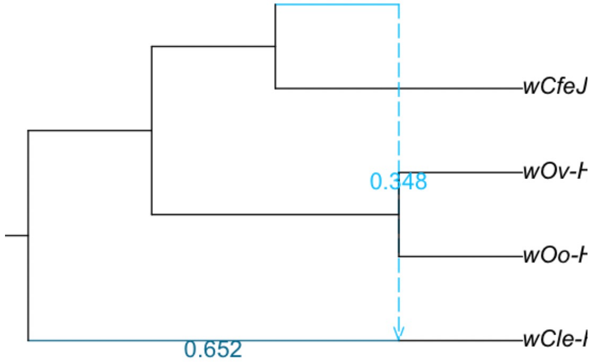
- Maximum Parsimony, Maximum Likelihood, Maximum Pseudo-Likelihood:
  - Fix a maximum number of ‘hybridization’ events and optimize a criterion
  - Choosing between different networks with different numbers of hybridization events is not well solved
  - Knowing the number of hybridization events is also not well solved

- Bayesian approaches: deal with the model complexity but limited to very small datasets (maybe 4 or 5 leaves)
  - **MCMC_GT** : input rooted gene trees
  - **MCMC_SEQ** : input alignments (direct inference)

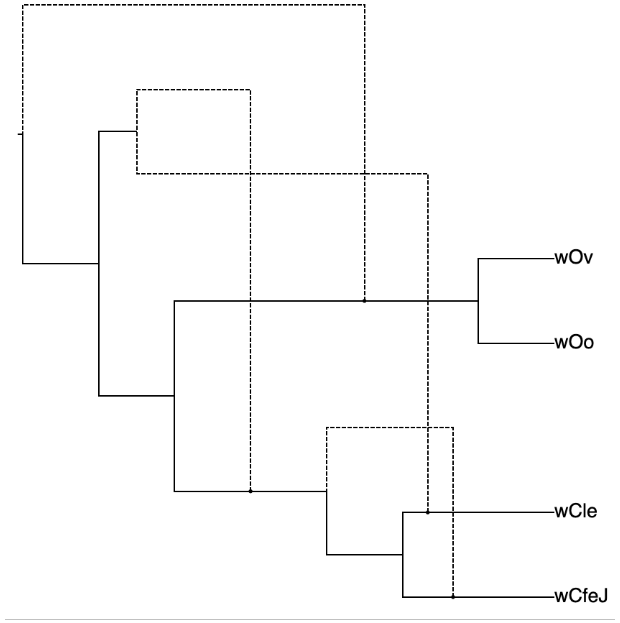
Problem n°1 : different methods return different number of reticulations

Only considering one 4-taxon 'species':

- SNaQ → 1 reticulation
  (too conservative?)

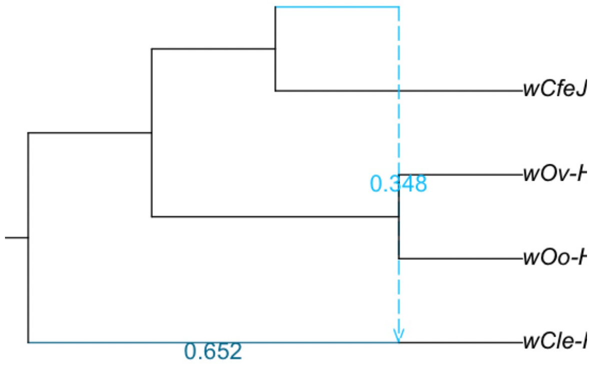


- MCMC_SEQ (direct Bayesian inference) →
  converges to 4 reticulations
  (too aggressive?)

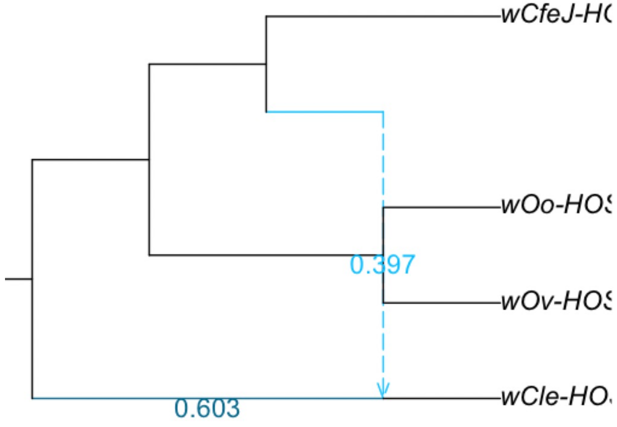


Problem n°2: For fixed number of reticulations, different methods return different networks

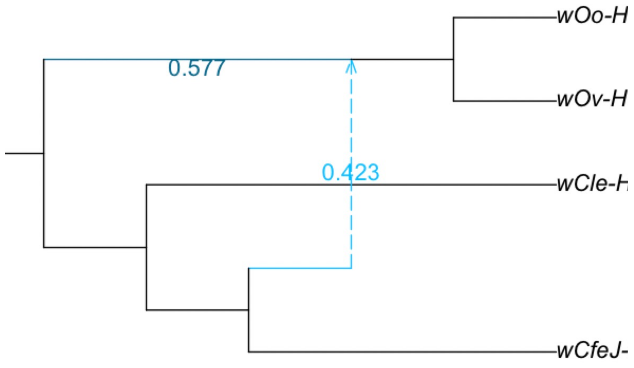
SNaQ



PhyloNet ML

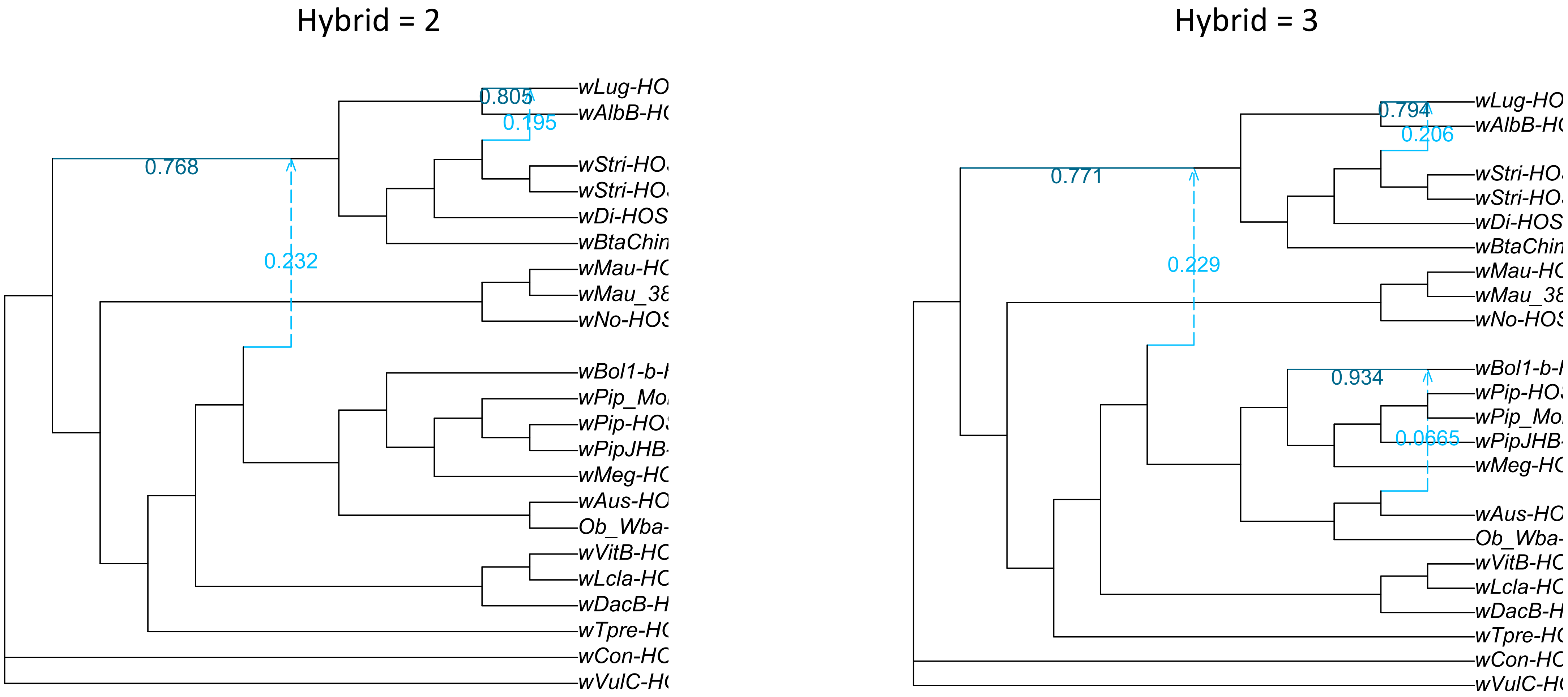


PhyloNet MPL



→ How do you score networks?

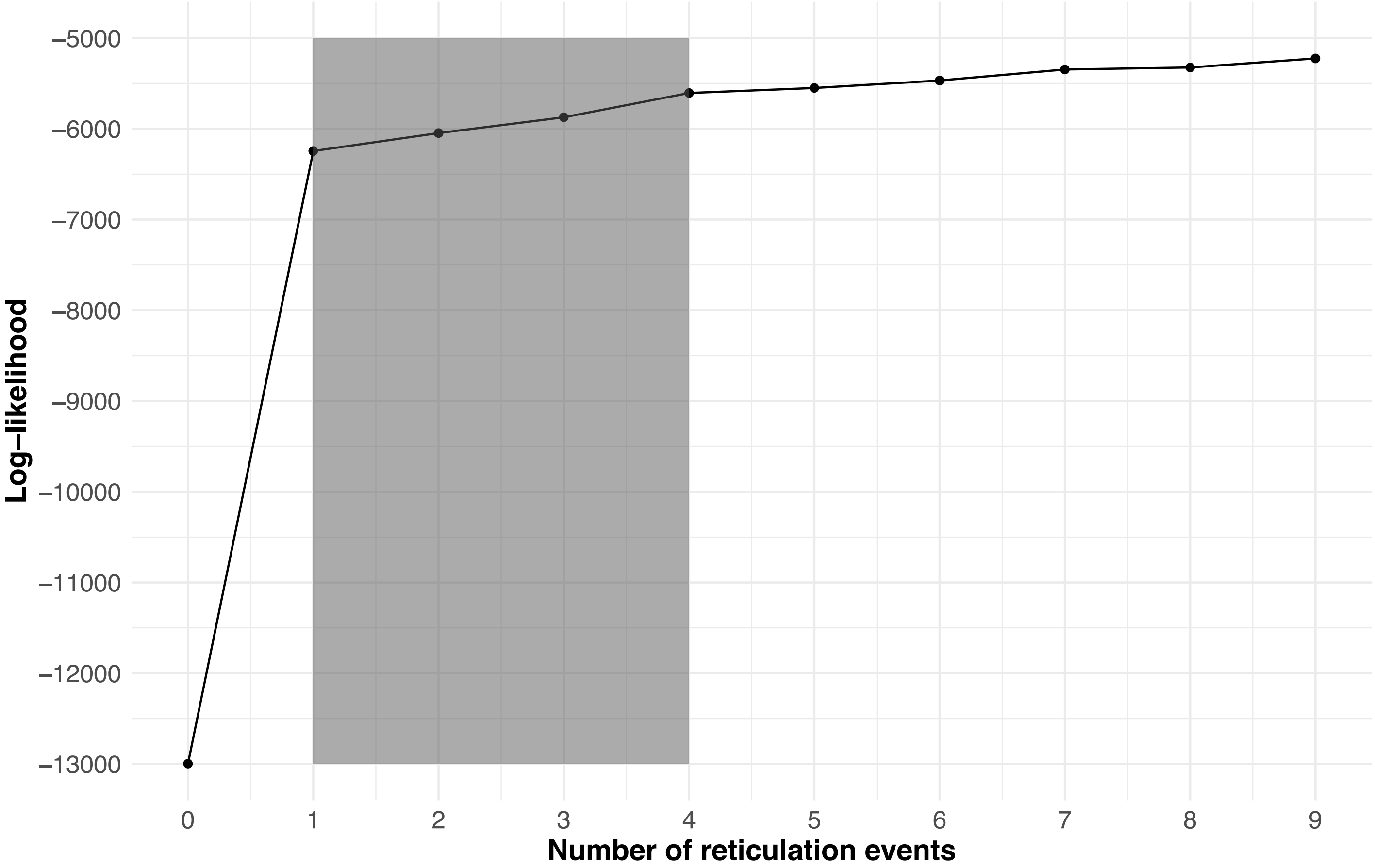
Problem n°3: how to choose number of reticulations?



SNaQ analysis with 2 vs. 3 hybridization events

(Pseudo) Log-likelihood improves as you increase the number of reticulation events

- Risk of falsely detecting extra reticulations
- Choice of "best" number of hybridization events usually done in a hand-wavy way (i.e. within the greyzone hereafter)
- Few statistical tests available (but see Cai & Ané, 2020)



Conclusion

- Species tree inference is hard, as it requires careful processing of the genomes (assembly pipeline), alignment procedure and tree estimation

- Phylogenetic network inference is much harder, as we are limited by the number of methods and most of them are highly time-consuming

- Wolbachia seems to have a limited amount of major hybridization events but we cannot quantify it with certainty