Phylogenomic estimate of the Cicadidae (Hemiptera: Cicadoidea): Identifying contaminated/paralogous locus copies and exploring the utility of Hemiptera and cicada 1:1 ortholog sets in pest Hemiptera lineages

Christopher L. Owen

Computational Biology Institute

George Washington University

Co-authors: David C. Marshall, Katherine B.R. Hill, Elizabeth Wade, Geert Goemans, Alan Lemmon, Emily Lemmon, Chris Simon

Presentation outline

Part 1

1) Phylogenomic estimate of the Cicadidae phylogeny using anchored hybrid enrichment data

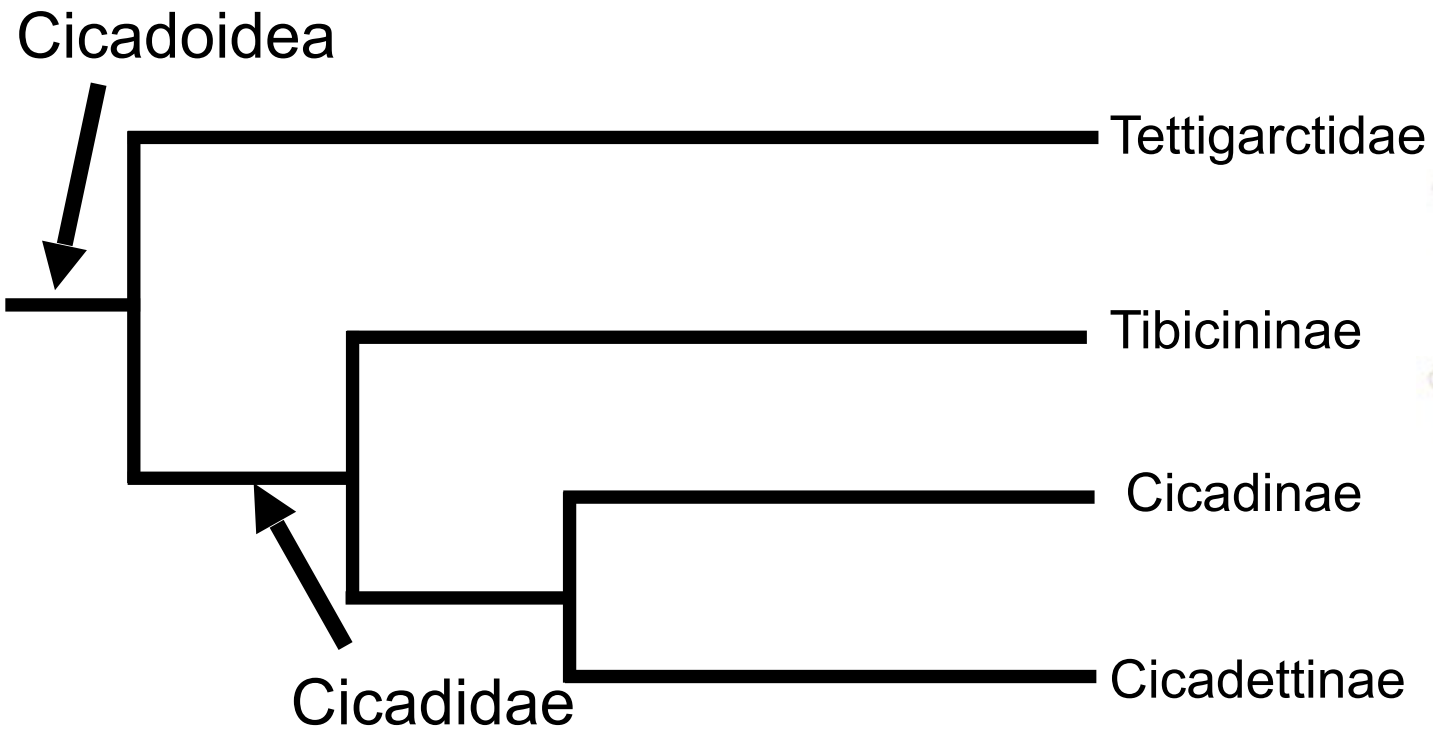
   a) Identifying contamination in phylogenomic datasets

Part 2

2) Utility of deep 1:1 orthologs to resolve shallow relationships in *Bemisia tabaci* Biotypes

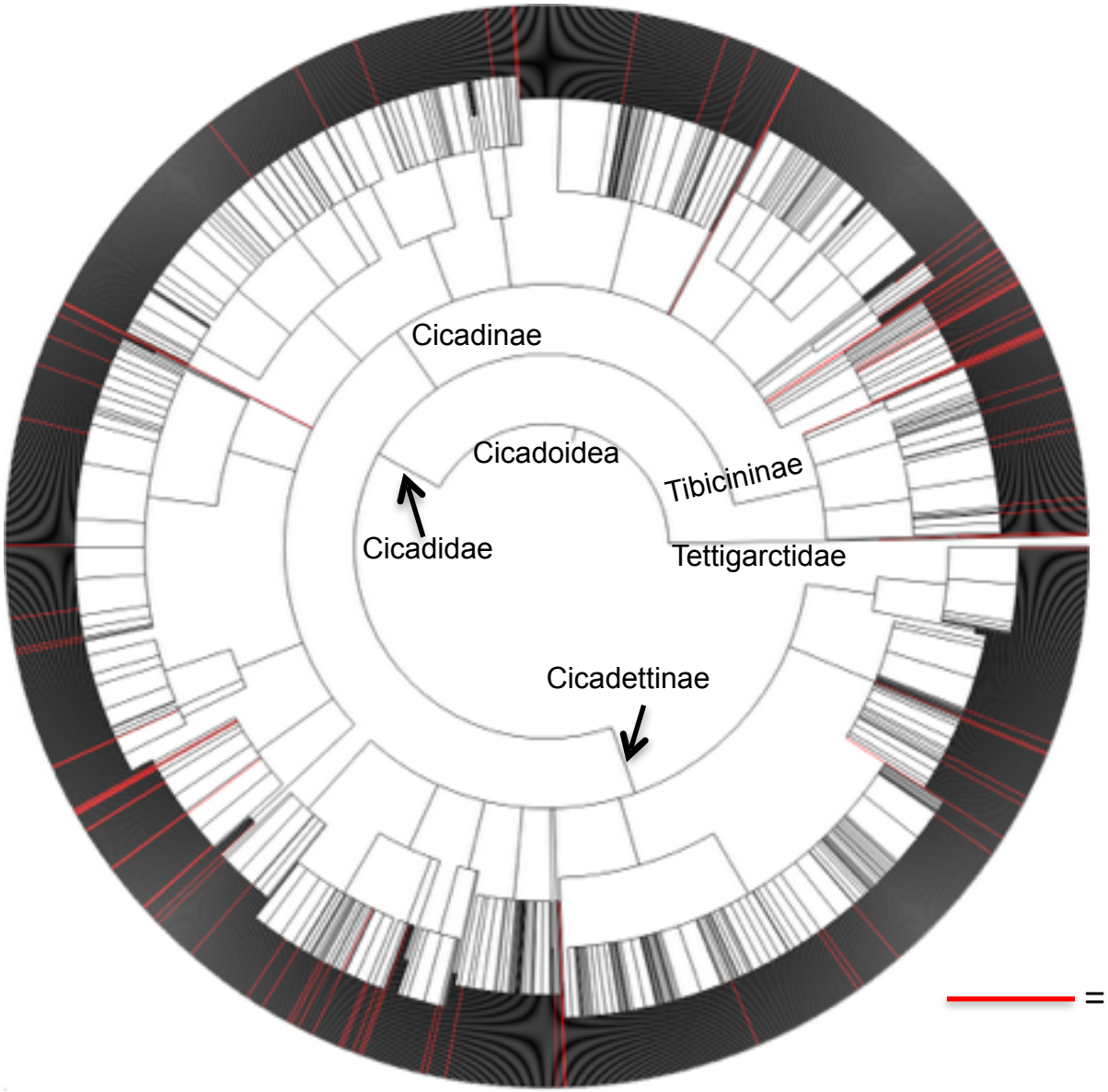
History of the Cicadidae phylogeny

- No molecular hypotheses of subfamily and tribal relationships
- Moulds (2005) proposed subfamily relationships and Australian tribal relationships (117 morphological characters)



Cicadidae taxonomy and extant diversity

Sanborn & Dmitriev (2016)
Subfamilies: 3
Tribes: 39
Subtribes: 38
Genera: 444
Species: 3044

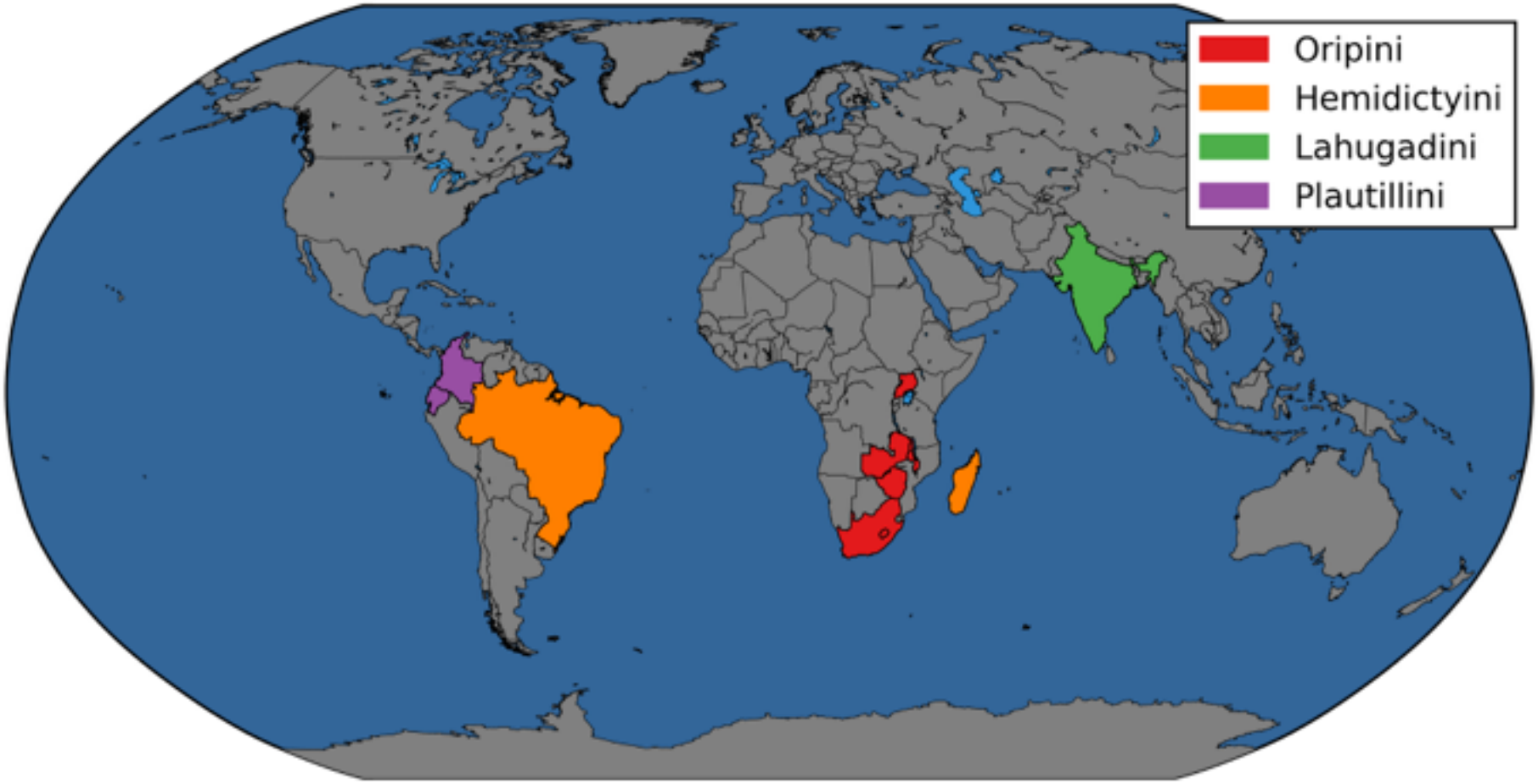


This Study
Subfamilies: 3
Tribes: 35
Subtribes: 21
Genera: 93
Species: 96

= sampled species

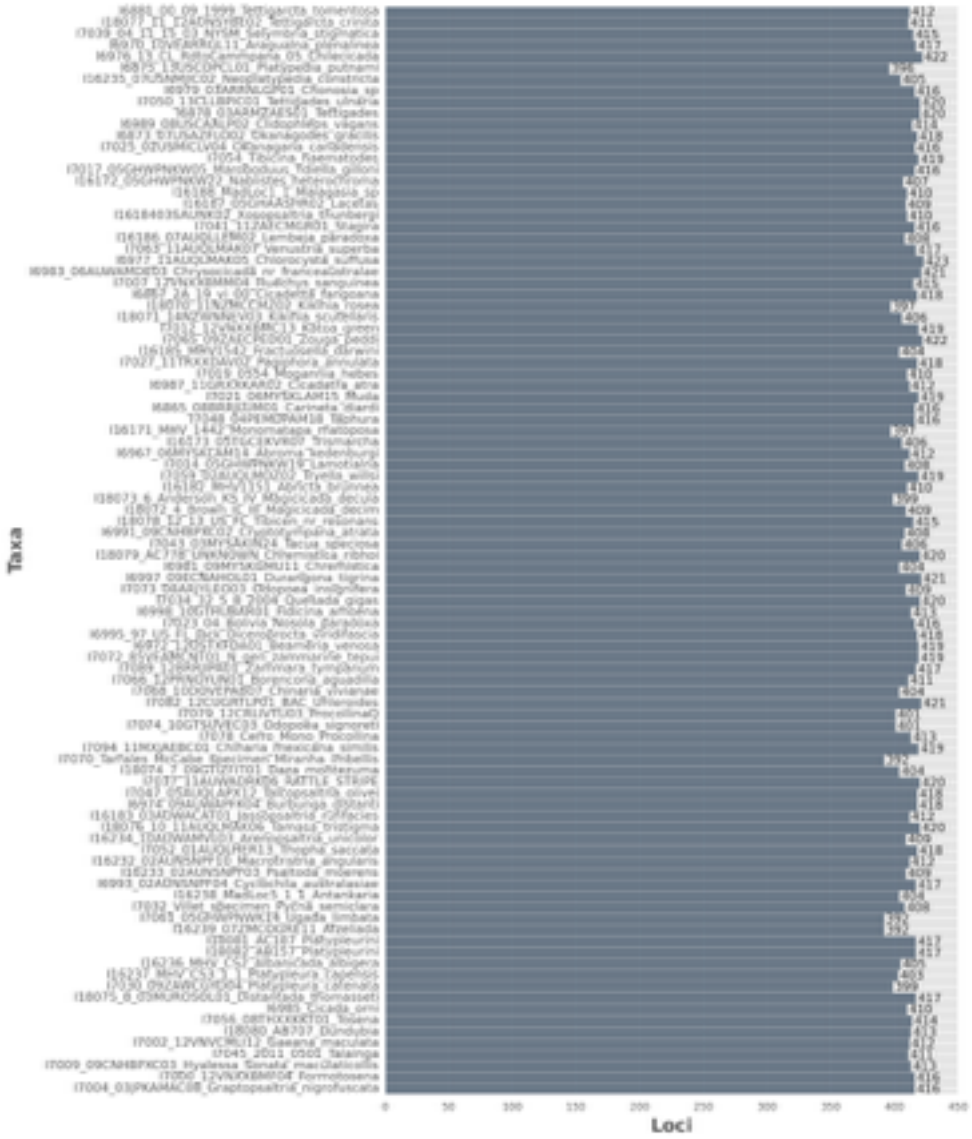
Taxonomy courtesy of D. Dmitriev

Tribes missing from then phylogenomic dataset

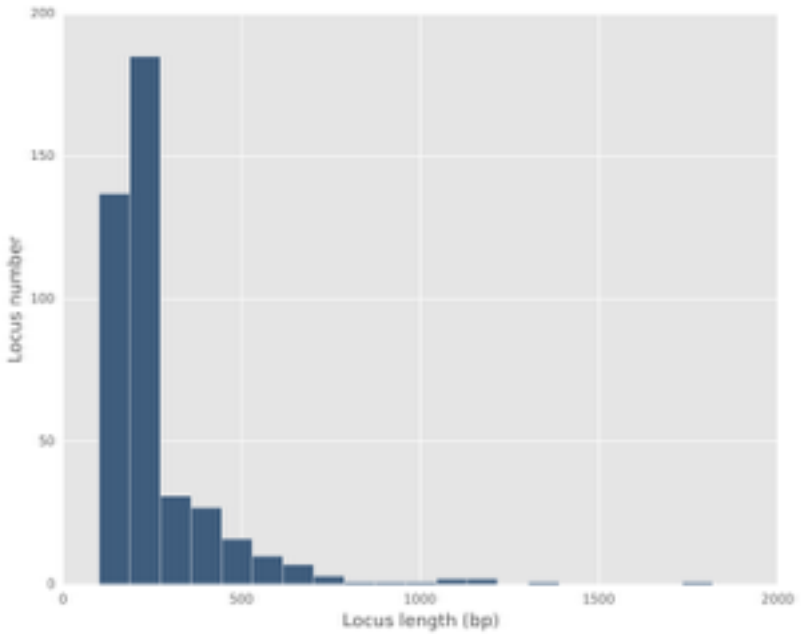


Cicadidae anchored hybrid enrichment phylogenomics dataset

A) Taxon locus occupancy (96 taxa)



B) Loci length distribution (425 loci)



Locus representation per taxon

Low: 392/425 loci (92%)

High: 423/425 loci (99%)

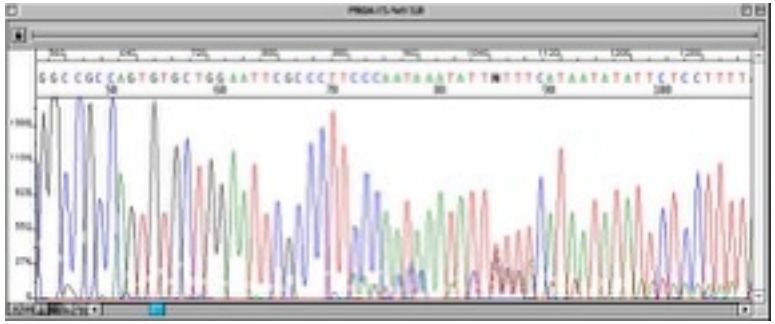
Loci length

- mode: 250bp
- concatenated length = 113,473bp

Methods: models and phylogenies

1. **Identify coding and non-coding regions for each locus**
   a. BLASTX cicada loci against curated db of cicada transcriptome ref proteins
   b. Convert BLASTX output to annotations for PartitionFinder
   c. Check loci annotations by hand
2. **Partitioning schemes and molecular models**
   a. PartitionFinder v1.1(Lanfear et al. 2012)
   b. Parameters: *search = all; model_selection = BIC*
   c. PartitionFinder subsets: non-coding, codon_pos1, codon_pos2, codon_pos3
3. **Gene trees**
   a. Garli v2.01(Zwickl 2006)
   b. Best ML tree from 5 independent search replicates
   c. 100 bootstrap replicates per locus
4. **Species tree**
   a. RAxML (not partitioned at the moment)
   b. Astral-II (Mirarab & Warnow 2015; Sayyari & Mirarab 2016)

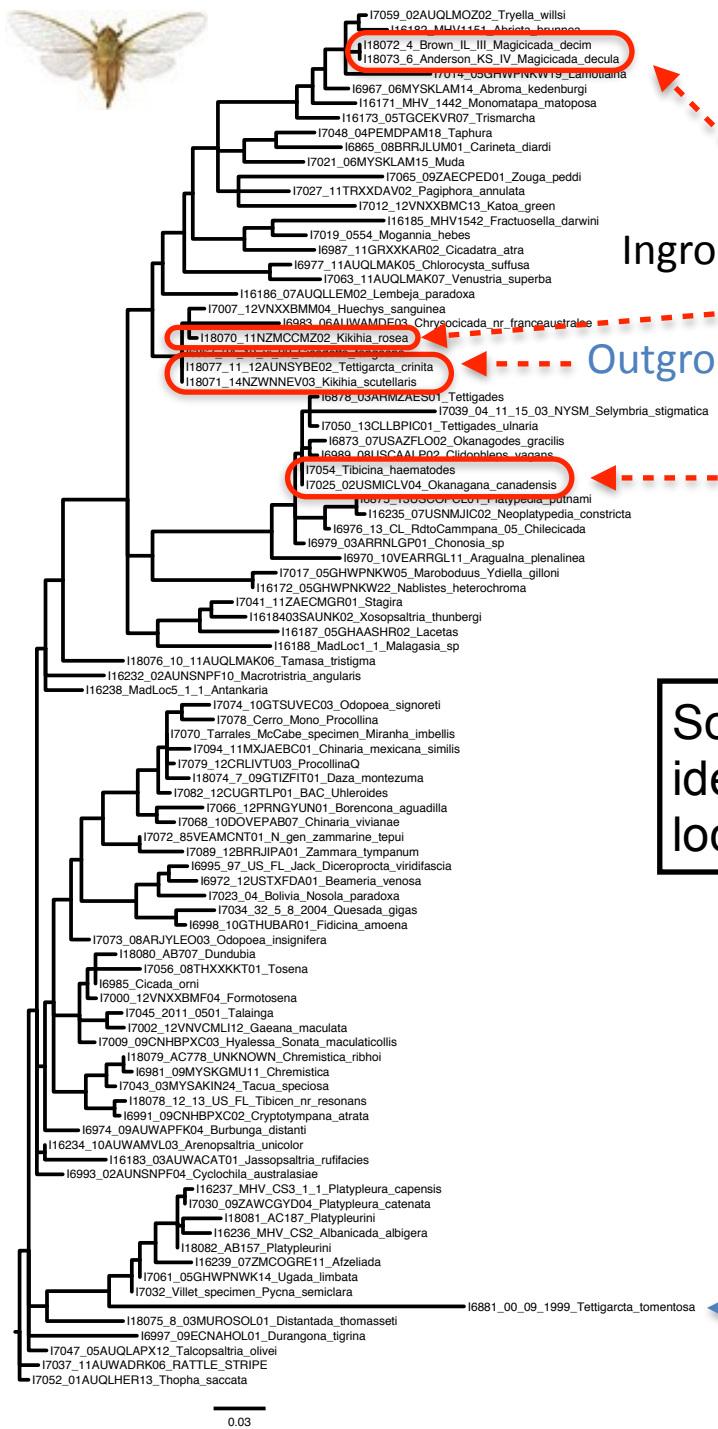
Contamination in phylogenomics datasets



http://ampliconexpress.com/wp-content/uploads/2014/04/Contaminated_Plasmid_Prep.jpg

- > 20% of non-primate genomes contain human DNA (Longo et al. 2011)
- Majority of our phylogenomics projects contain taxa that have not had more than a few genes sequenced
- How do we identify contamination in our phylogenomics projects (hybrid capture & transcriptome)?

Contamination example



Ingroup: *Magicicada* spp.

Ingroup: *Kikihia rosea*

Outgroup & ingroup: *Tettigarcta crinita* and *Kikihia scutellaris*

Ingroup: *Tibicina* and *Okanagana*

Some unrelated taxa, based on taxonomy, share identical sequence when nearly all other taxa in the locus differ to varying degrees

Outgroup: *Tettigarcta tomentosa*

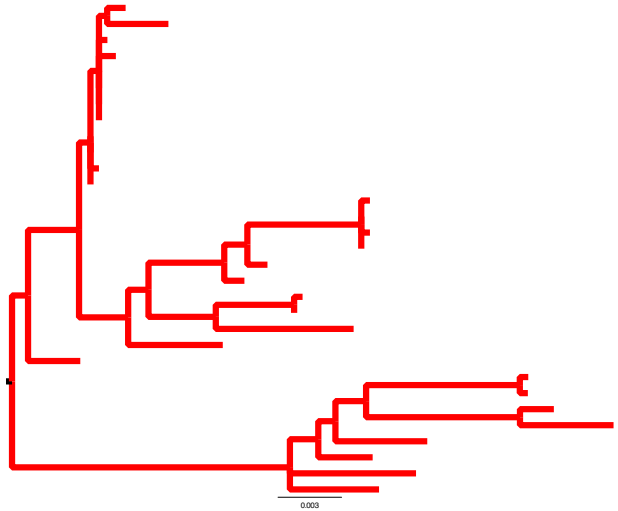
Methods: contamination detection

For each taxon pair for all gene trees:

Patristic Distance

÷

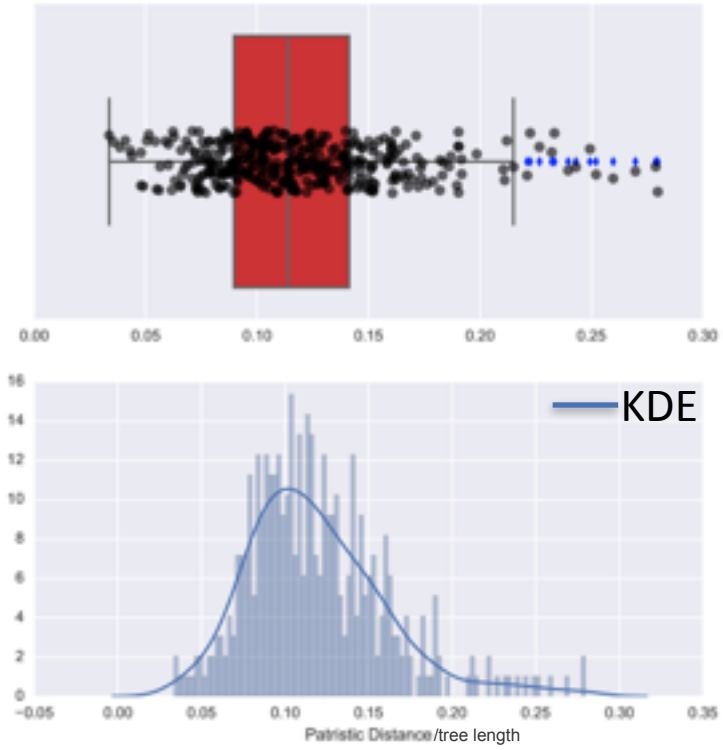
Gene tree length



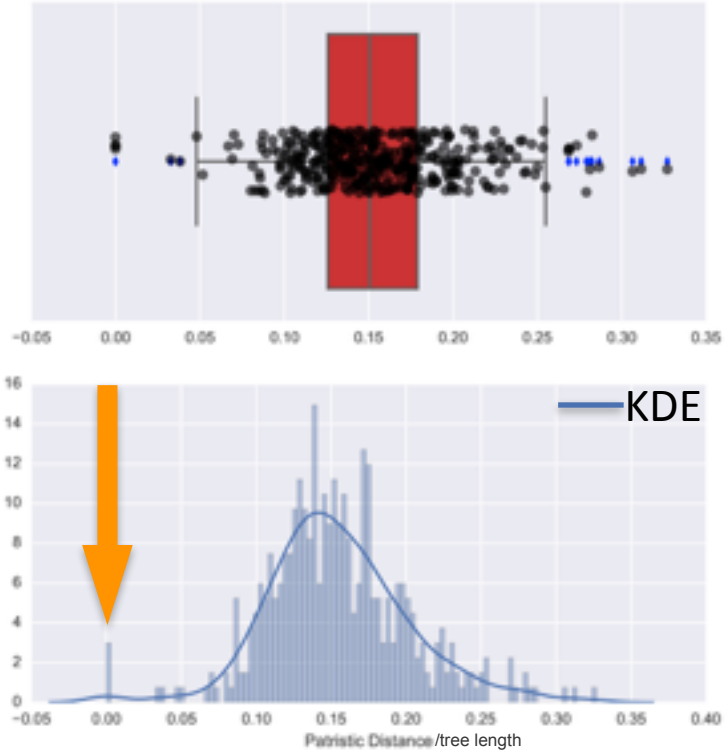
Calculations performed in Python using ETE 3 (Huerta-Cepas et al 2016)

Methods: contamination identification

Probably Not Contaminated
SpeciesA X SpeciesB

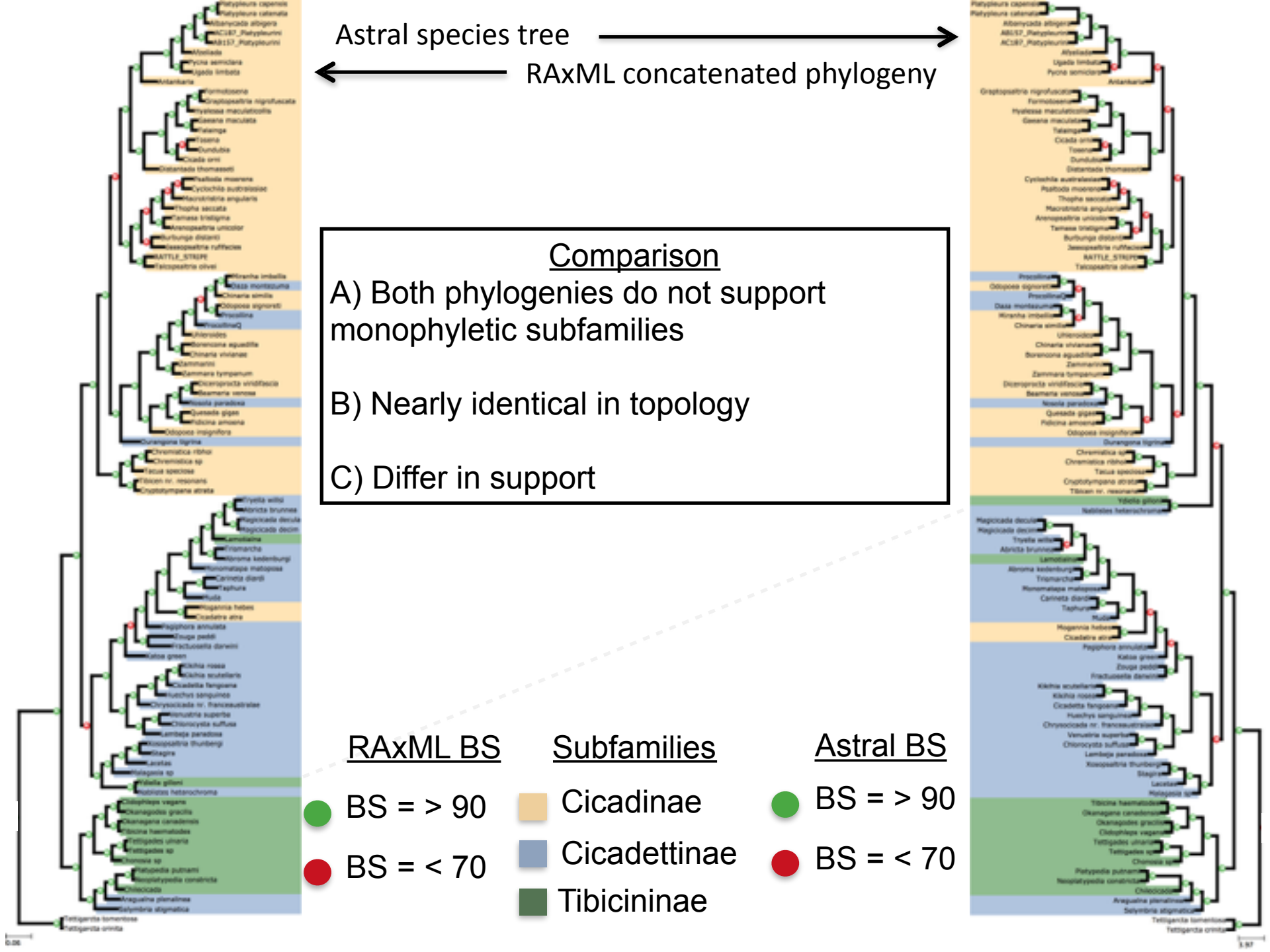


Possibly Contaminated
SpeciesA X SpeciesB

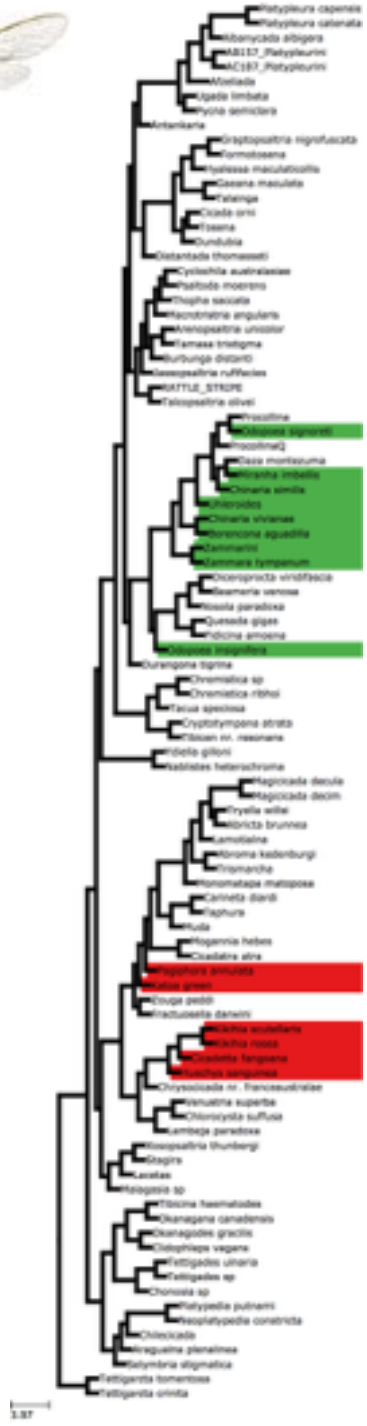


- Total number of comparisons $\frac{(N-1)N}{2}$, where N = 96: 4,560
- Related taxa not in ingroup or outgroup?
- False positives, but seems to capture the 'contaminated' sequences

Removed 530 sequences from 130 loci


Astral species tree
RAxML concatenated phylogeny
Comparison
A) Both phylogenies do not support monophyletic subfamilies
B) Nearly identical in topology
C) Differ in support
RAxML BS
BS = > 90
BS = < 70
Subfamilies
Cicadinae
Cicadettinae
Tibicininae
Astral BS
BS = > 90
BS = < 70

Is the tribal taxonomy congruent with the molecular phylogeny?

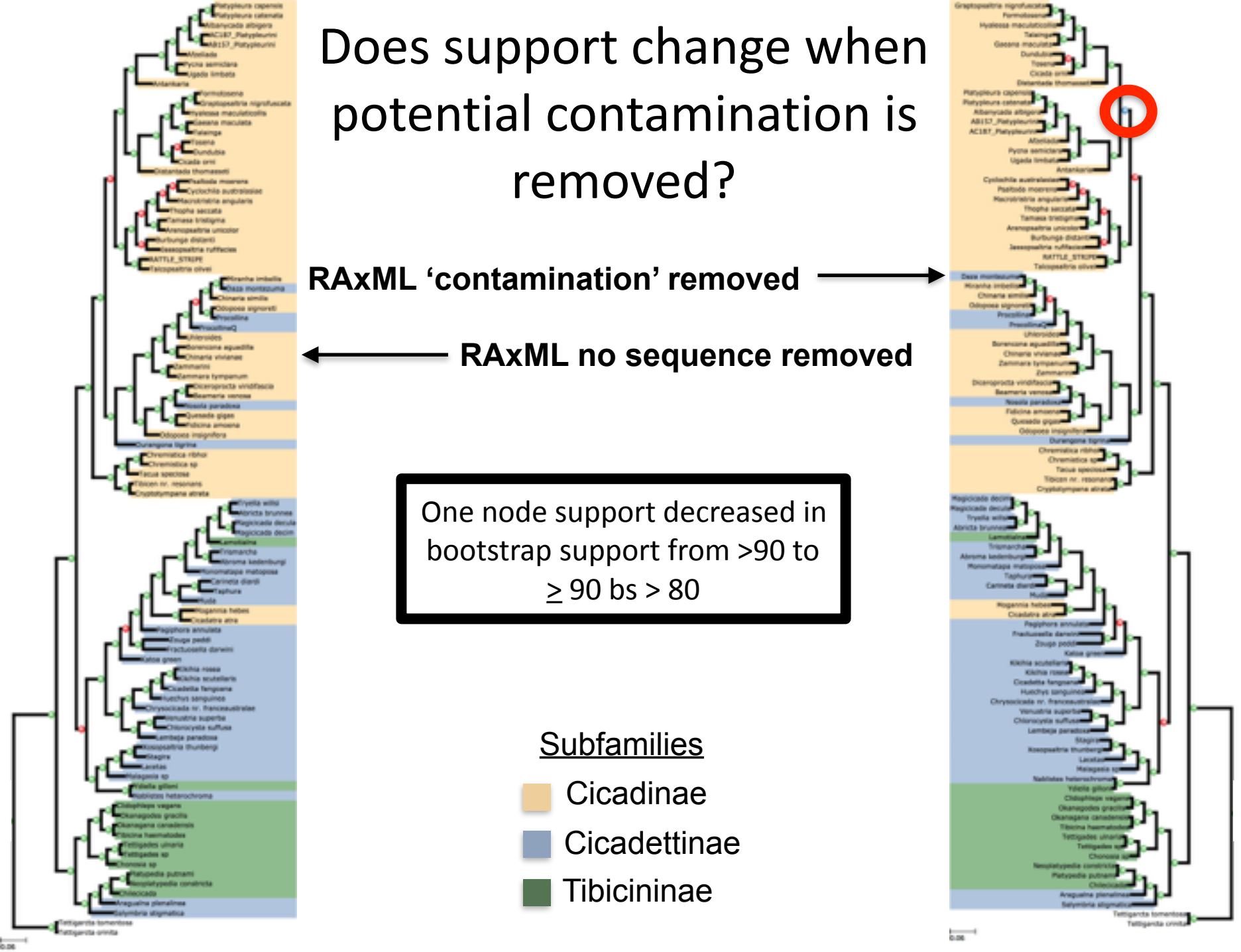


Tribe taxonomy is also not congruent with the molecular phylogeny:

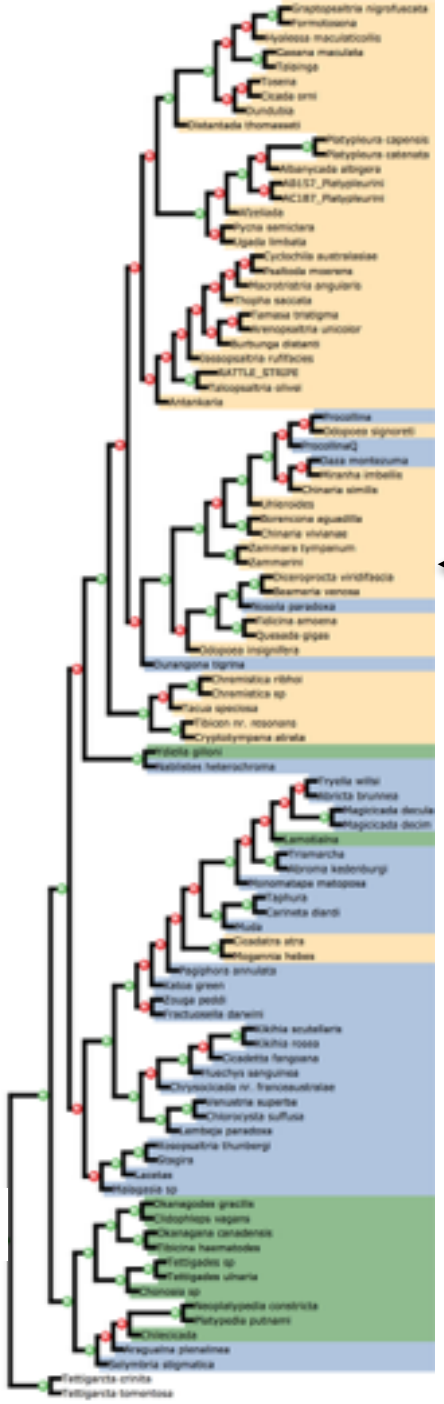
= Zammarini

= Cicadettini

Does support change when potential contamination is removed?

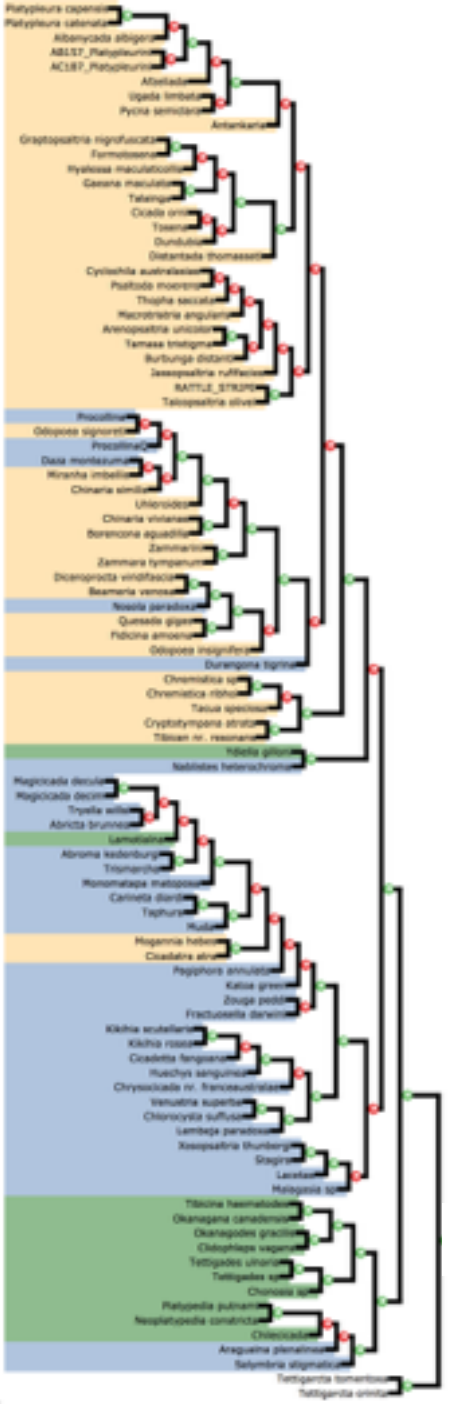


Does quartet support change after removing potentially contaminated sequences?



Astral species tree without 'contamination'

Astral species tree with 'contamination'



- Removing 'contaminated' sequences does not change quartet support
- Need to determine if the alternative signals are abundant vs. noise
- ML gene trees by themselves do not inform many relationships in the species tree

● ≤ 50%

● > 50%

Conclusions and future directions

1) Our molecular phylogeny does not support the subfamily taxonomy and some tribes

2) We are able to identify and eliminate contamination at the expense of some good data

3) Removing obvious contamination does not alter bootstrap support or quartet support

4) Further investigate alternative signal(s) from the quartet support: noise?, short branches?

Utility of deep 1:1 orthologs in pest Hemiptera lineages

Motivation

1. Can deep orthologs resolve closely related species?
2. At what age does the phylogenetic signal degrade for shallow divergences, where the support may be governed by the conflict?
3. Do we need to develop study specific 1:1 orthologs for each project?

Test example: *Bemisia tabaci*

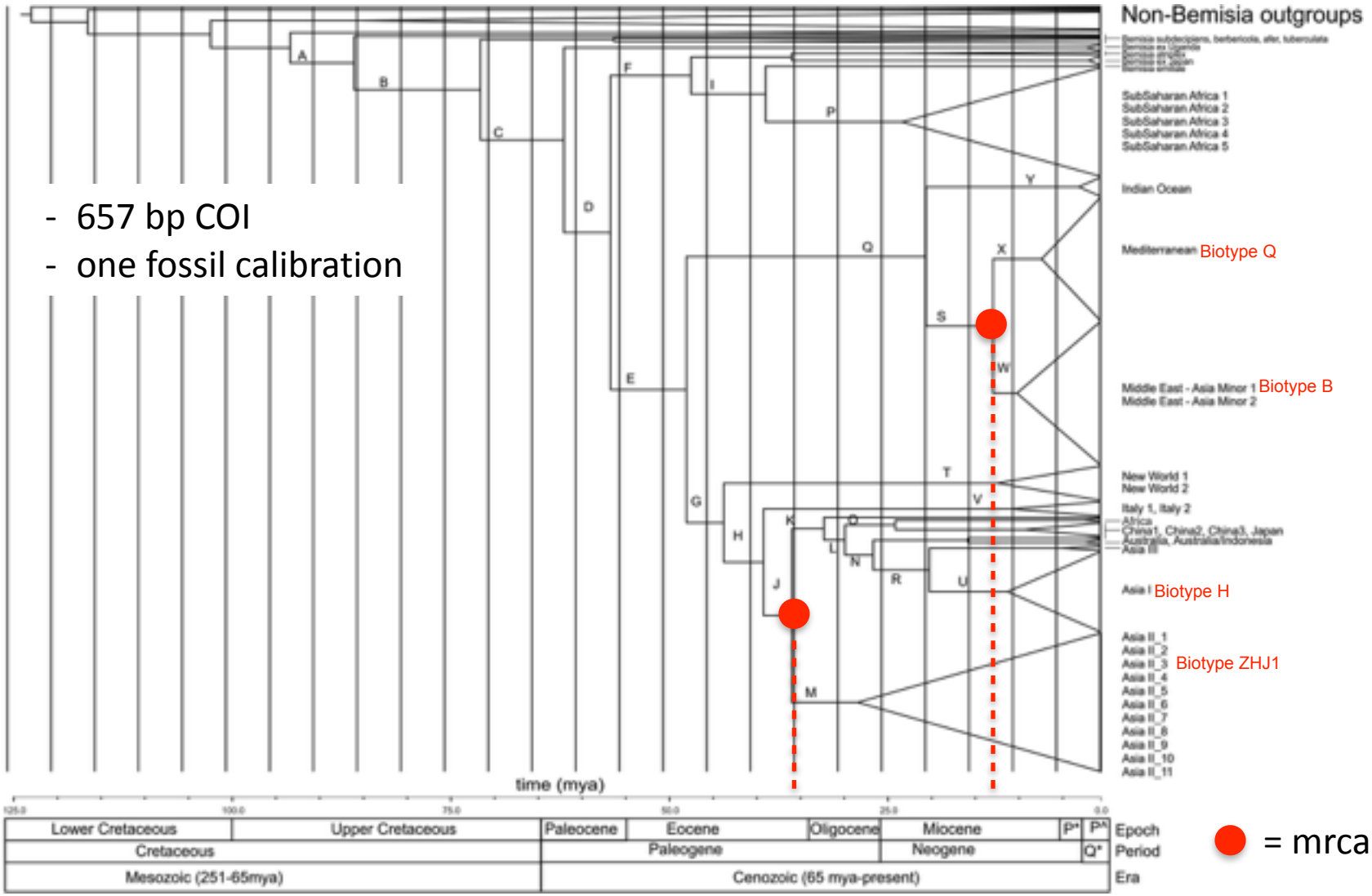
1) Cost the US > 1 billion dollars in crop & ornamental damages

2) > 500 different hosts

3) Adults morphologically cryptic

4) Few genetic markers to infer relationships: COI, 16S, ITS1, RNA polymerase II, pre-mRNA processing factor 8, AFLPs, and microsatellites

Bemisia tabaci Biotype phylogeny



Boykin et al. 2013

Bemisia tabaci data

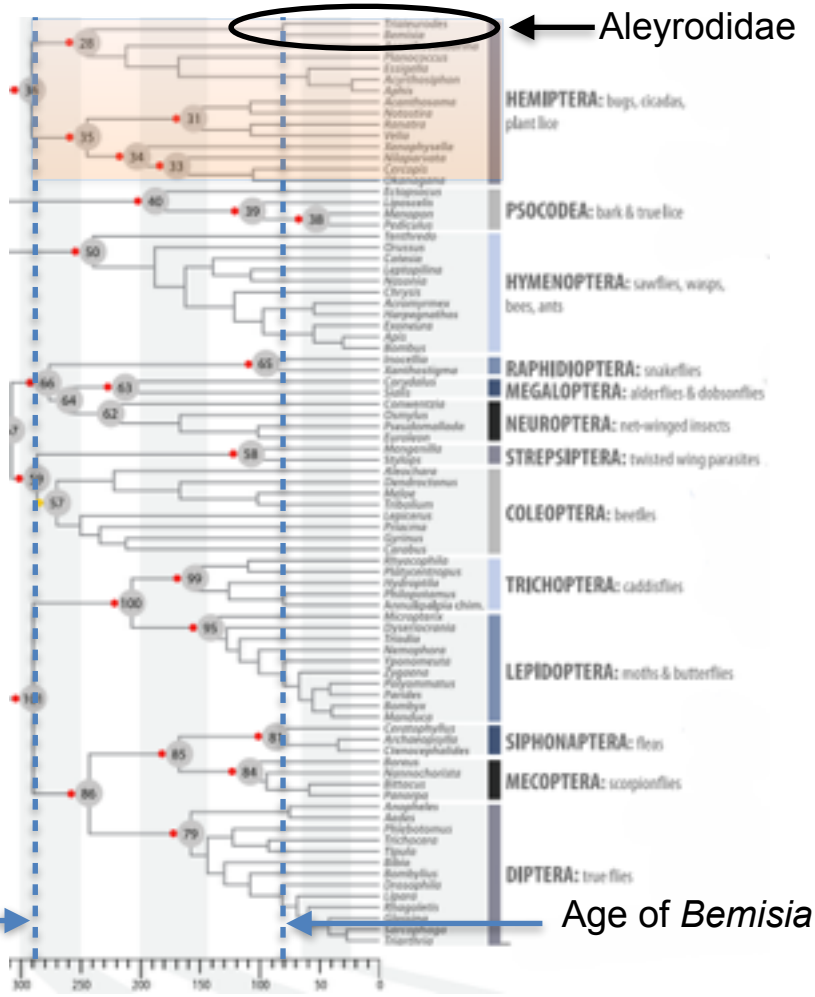
Ortholog database

-1754 1-to-1 orthologs predicted using tree-based orthology prediction (Yang and Smith 2014) and 8 Hemiptera genomes (I5K)

Data

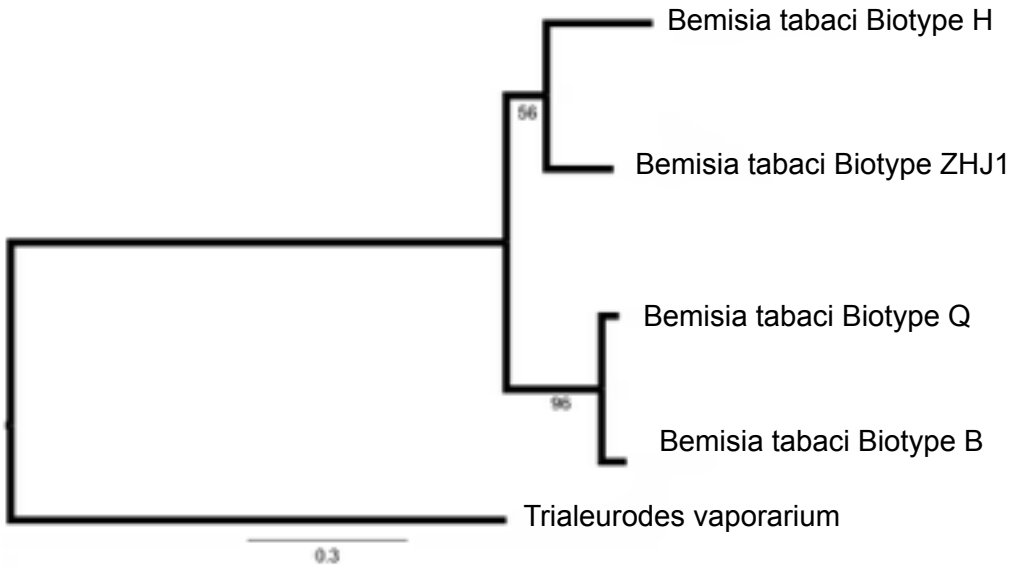
- Four *Bemisia tabaci* transcriptomes
- Two outgroups (Aleyrodidae spp.)
- 50 orthologs (HaMStR pipeline)
- COI data (760 bp)

Misof et al. 2014

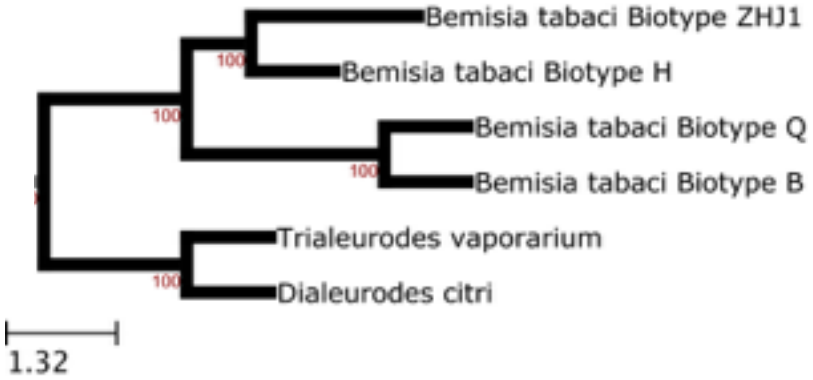


Bemisia tabaci phylogenies

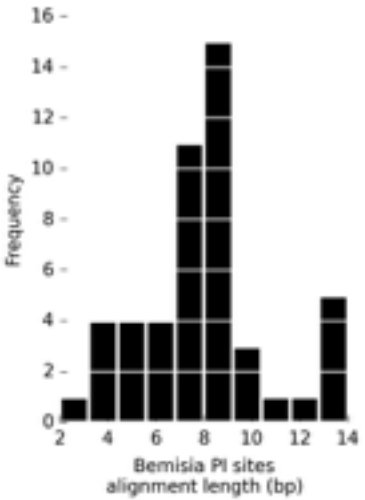
COI ML Garli phylogeny



Phylogenomic Astral species tree



nDNA Parsimony informative sites



mtDNA Parsimony informative sites

1st codon position: 18
2nd codon position: 3
3rd codon position: 57

78 Total PI sites

Conclusions and future directions

1) Deep Hemiptera orthologs do offer phylogenetically informative sites for shallow divergences (35 Ma - 15 Ma), but few.

2) Remove outgroups

3) Predict orthologs from transcriptomes and compare to deep orthologs

Flickr: Edithvale-Australia Insects and Spiders


Questions?