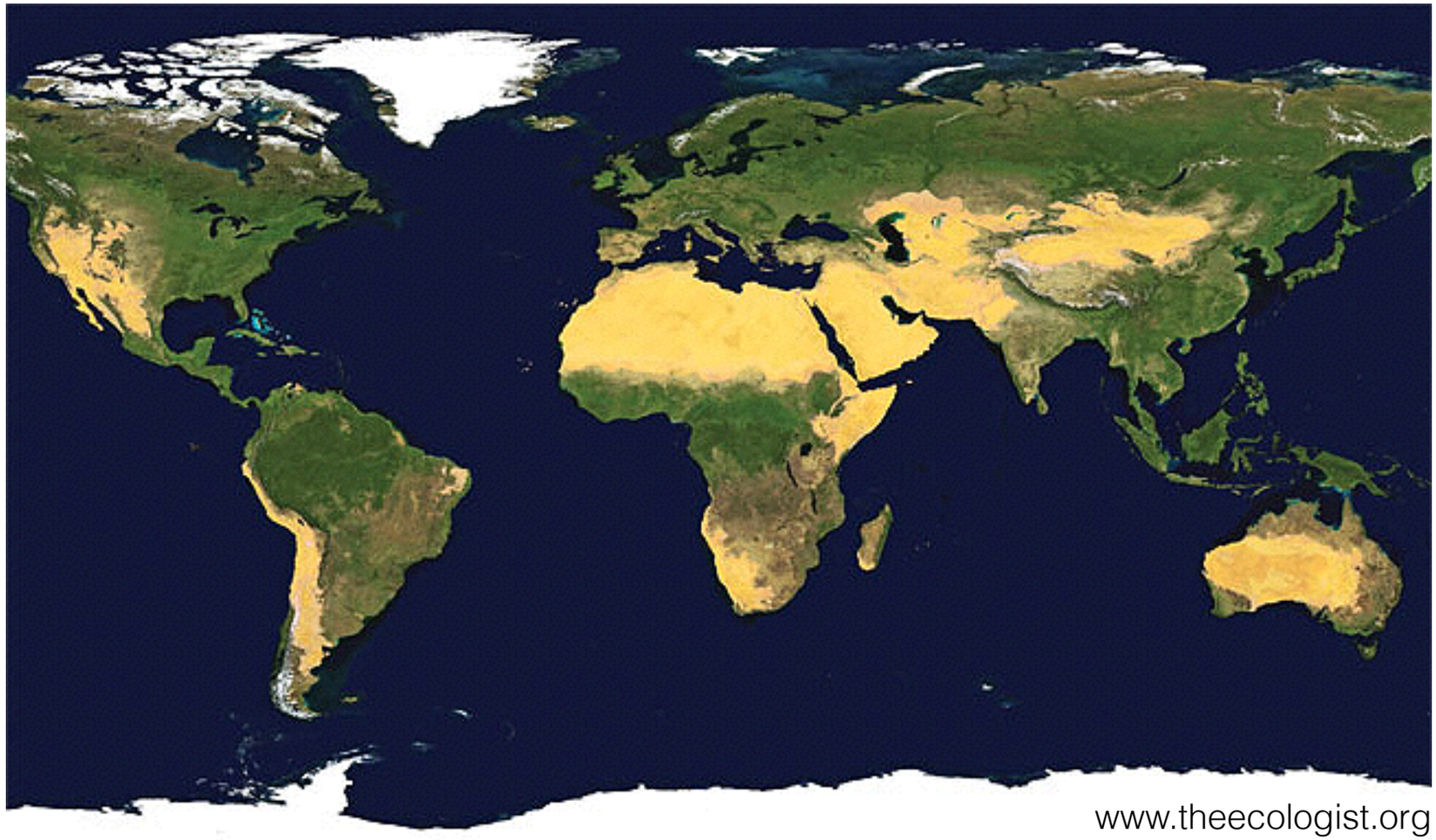
Re-examining the Hemiptera phylogeny using supermatrices and phylogenomic data sets

Christopher L. Owen
Postdoctoral Research Associate
GWU Computational Biology Institute

OUTLINE

1. Introduction
2. Methods
3. Preliminary Results
4. Future directions

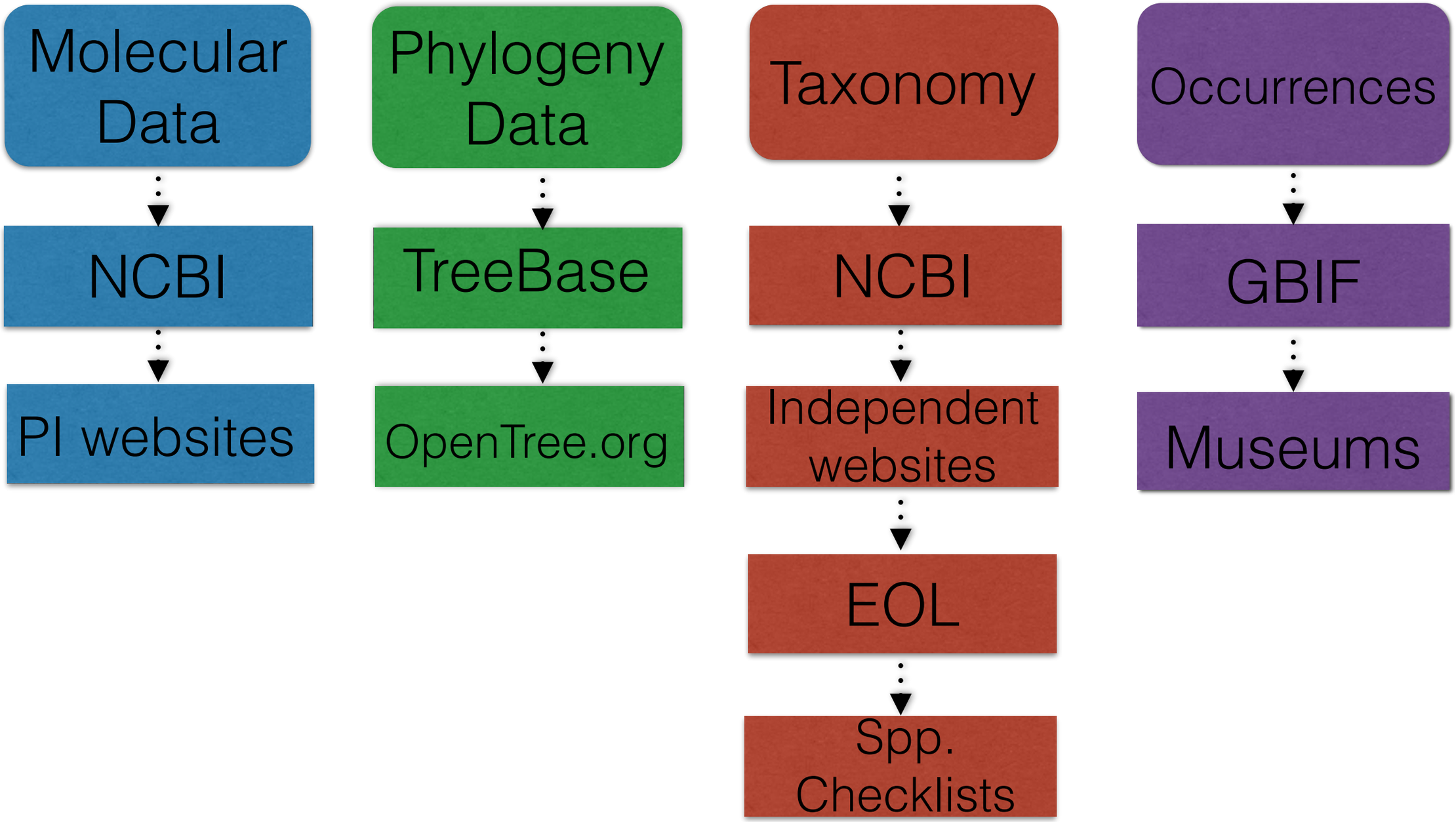
Introduction: Southern Hemisphere biogeography and diversification



www.theecologist.org

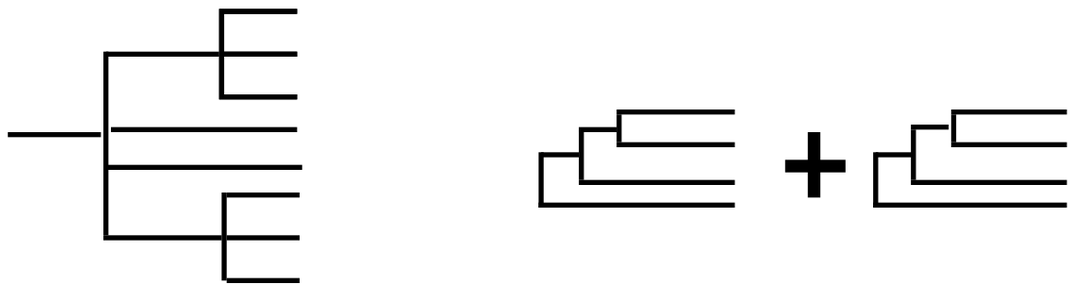
- Three major episodes of cooling since the Eocene
- Floral turnover
- Signatures of both changes in diversification and floral turnover in Australian cicadas

Goal: synthesize all available Hemiptera data to test hypotheses

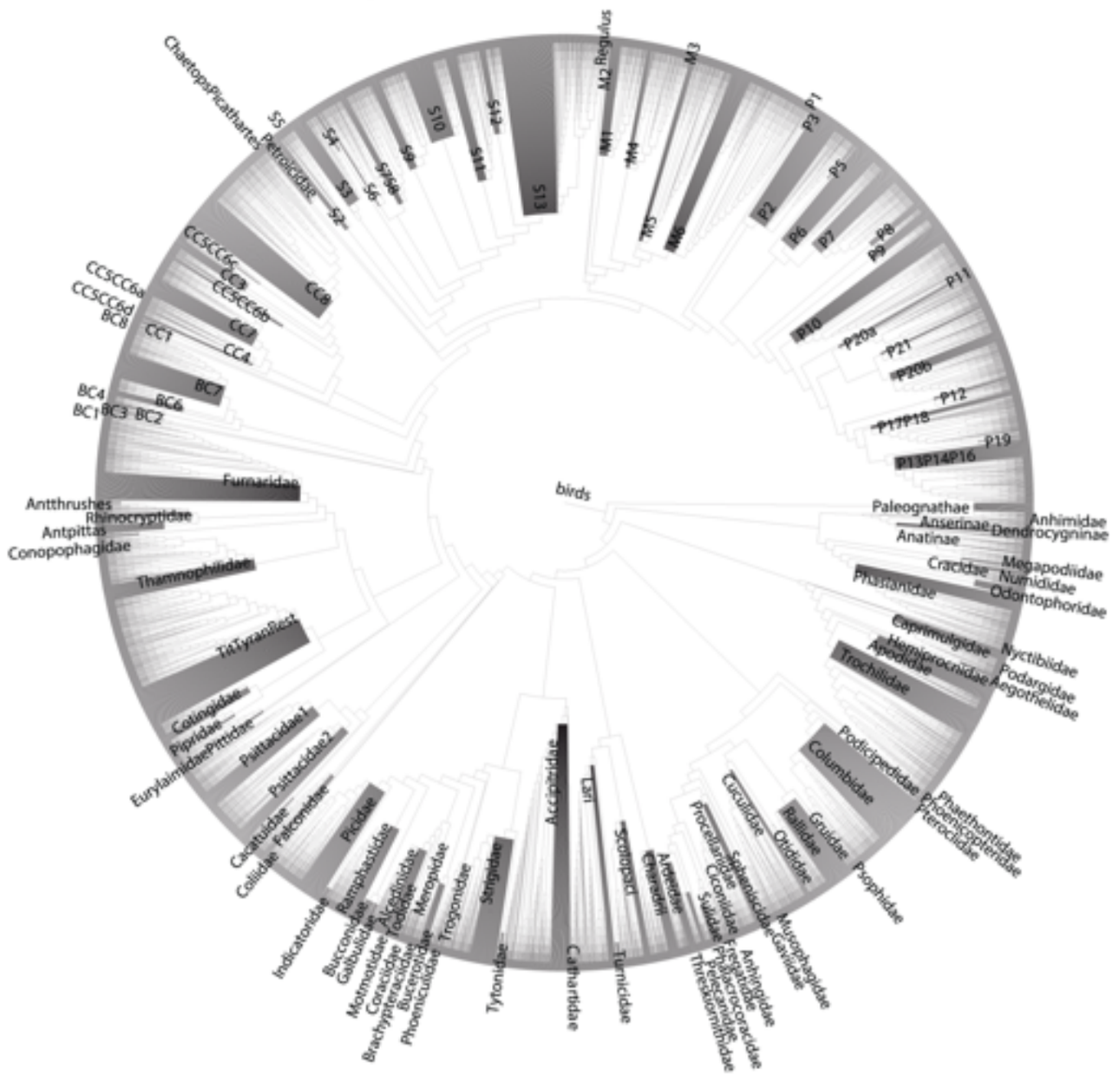


AVAToL: Open Tree of Life

Published Taxonomies + Published Phylogenies

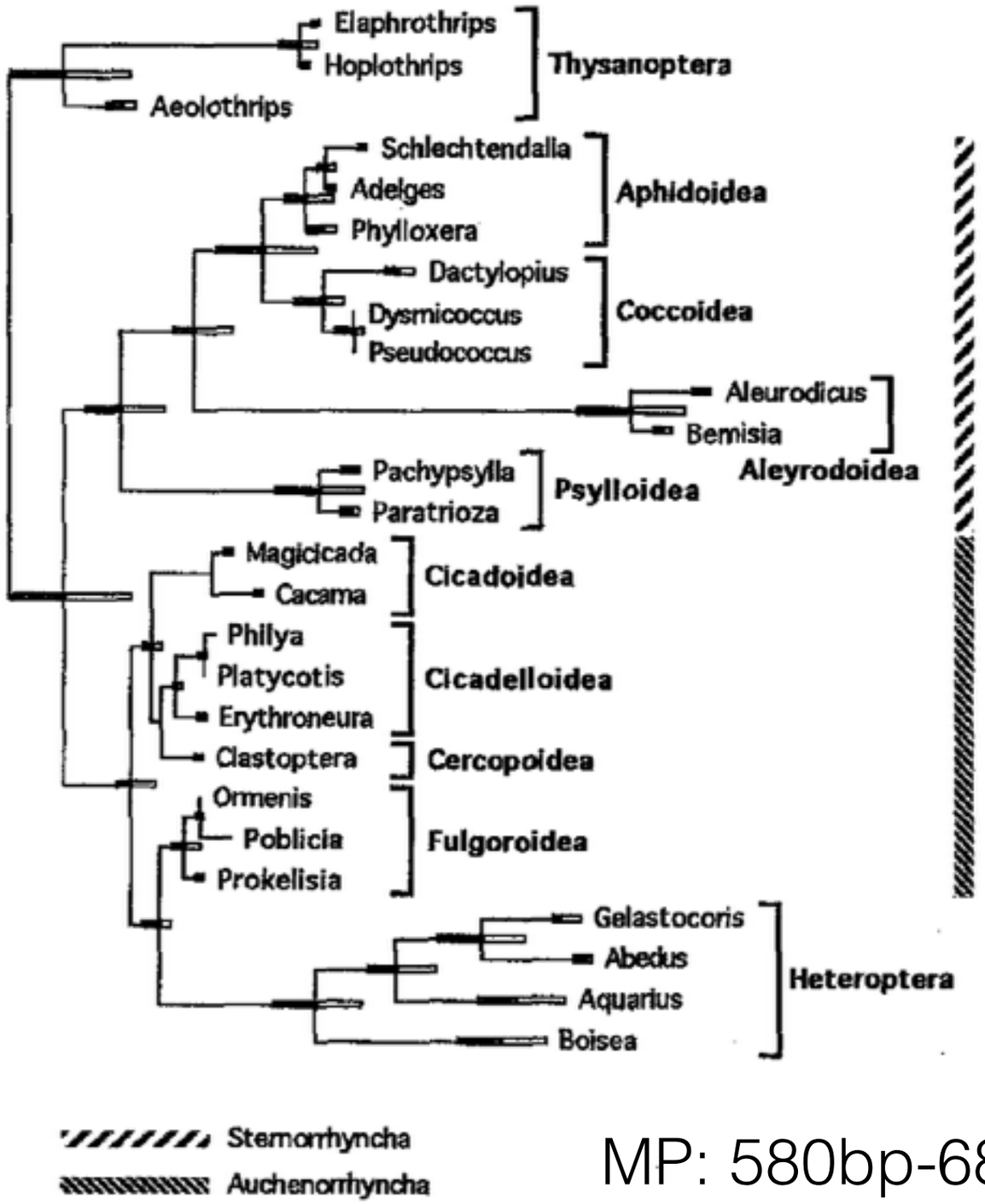


- Phylogeny of all life
- Accessible to the public
- Public curation
- Add/Extract phylogenies
- First release: early 2014
- Two additional AVAToL groups



Smith et al. 2013 PLOS Comp. Biol.

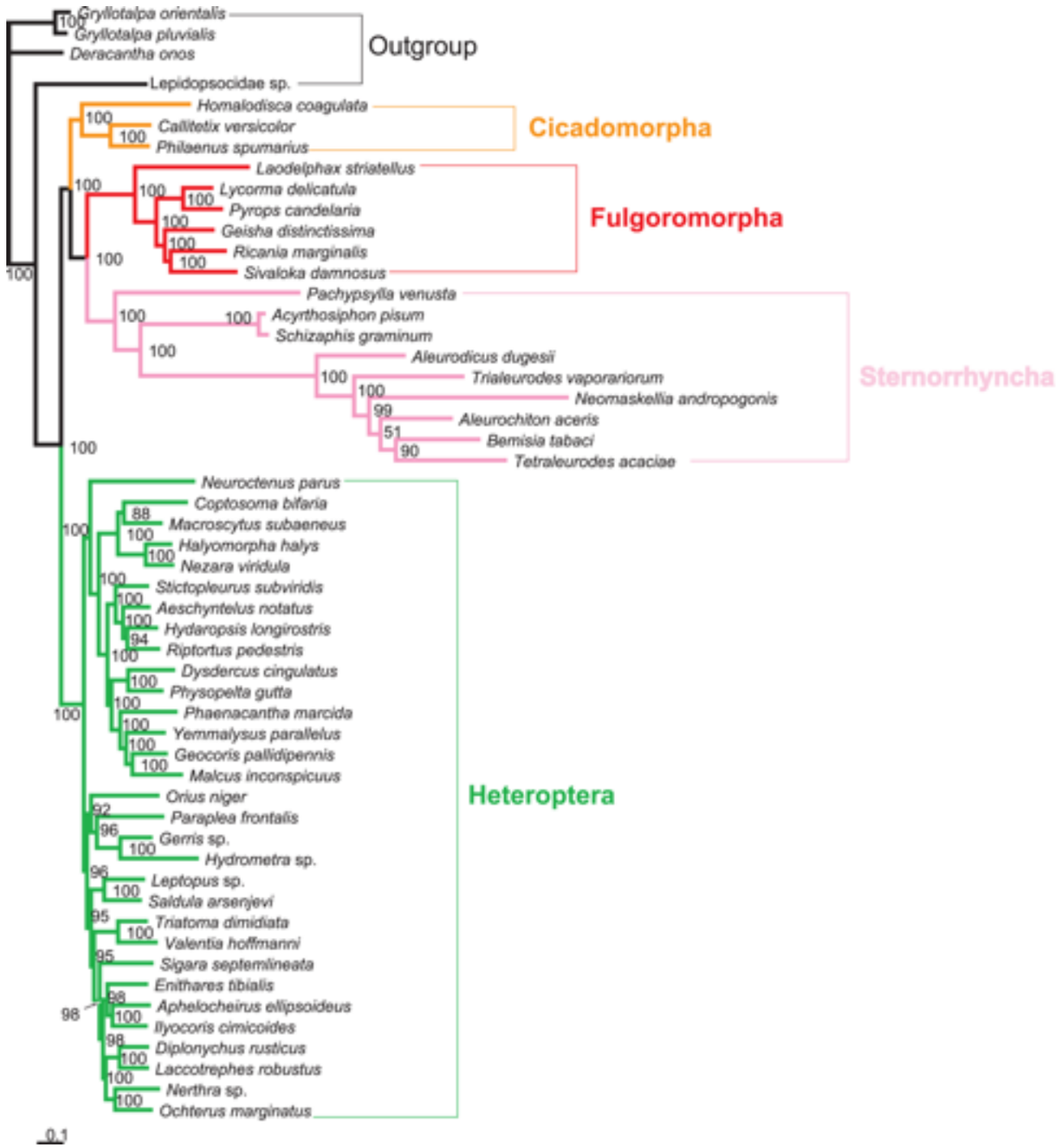
Hemiptera Phylogenies:
von Dohlen & Moran 1995 J. Mol. Evol.



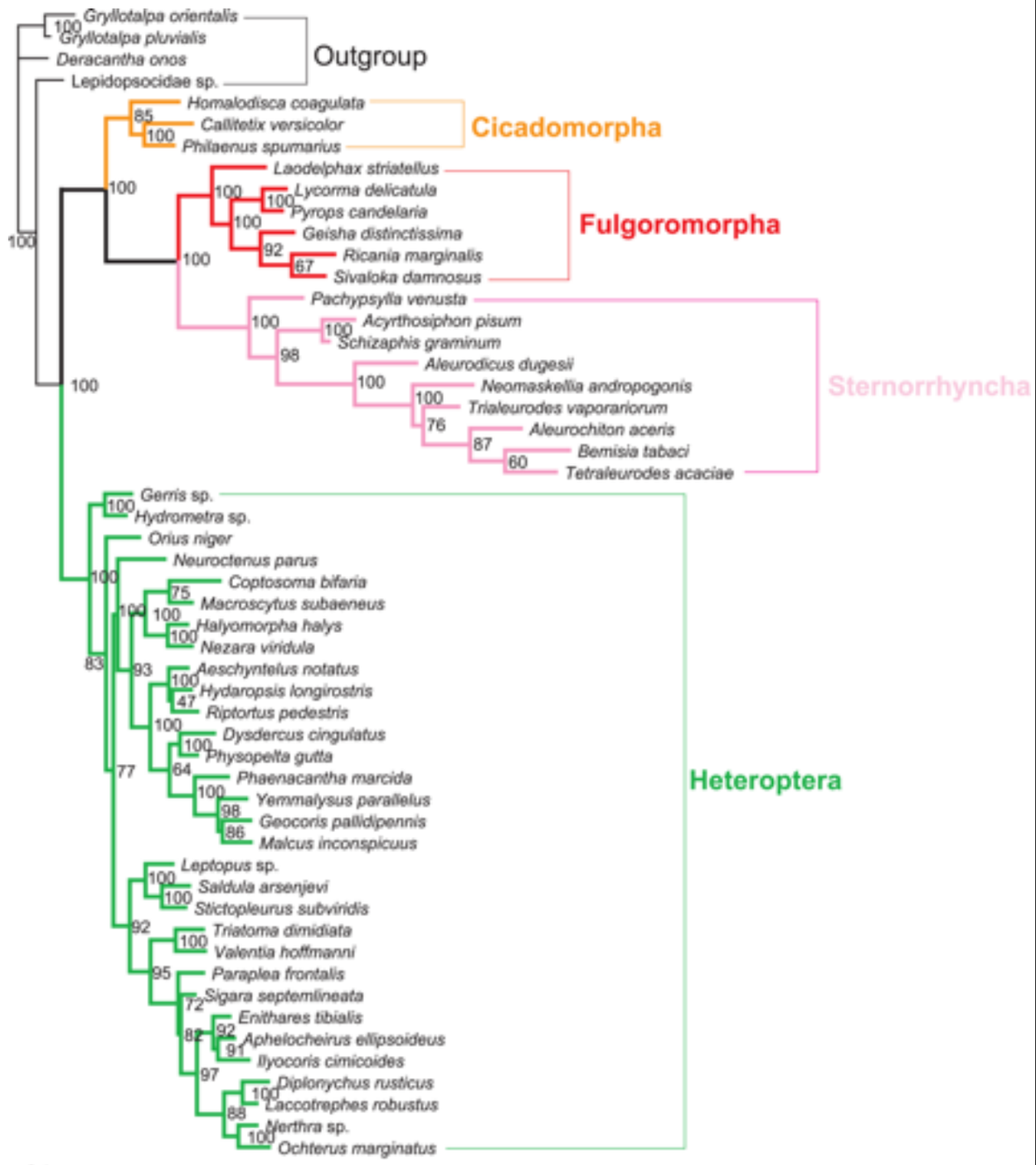
MP: 580bp-680bp 18S rDNA

Recent Hemiptera Phylogenies: Song et al. 2012 PLOS One

BI Tree:1st & 2nd postions, 2 rRNA genes

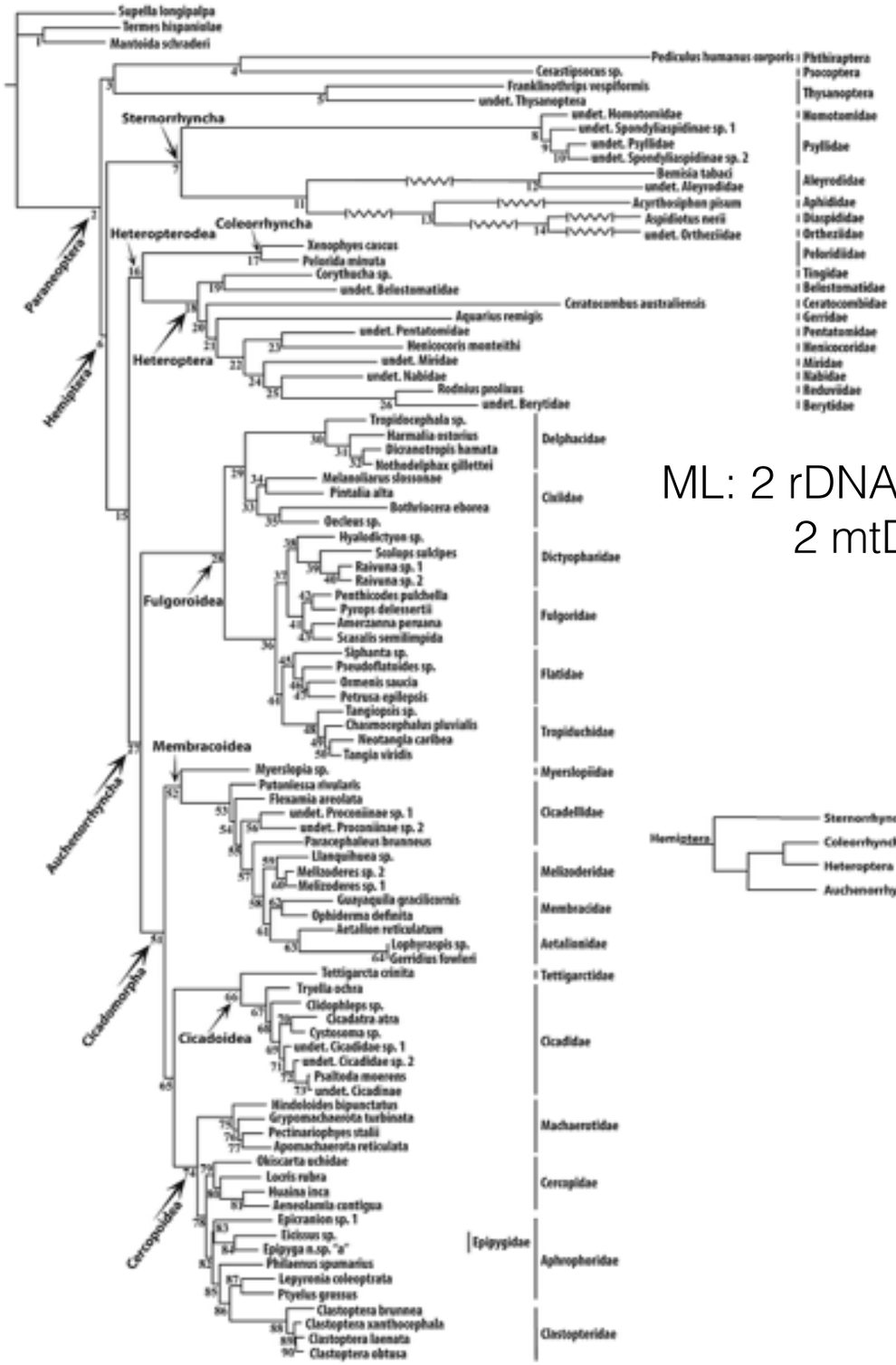


ML Tree:1st & 2nd postions, 2 rRNA genes



Recent Hemiptera Phylogenies: Cryan and Urban 2012 J. Syst. Ent.

- Well-supported backbone for most higher taxa
- Lacking support for Hemiptera
- Additional molecular evidence for long branches in Sternorrhyncha



ML: 2 rDNA, 3 nDNA, 2 mtDNA

Methods: assemble transcriptome data gene alignments

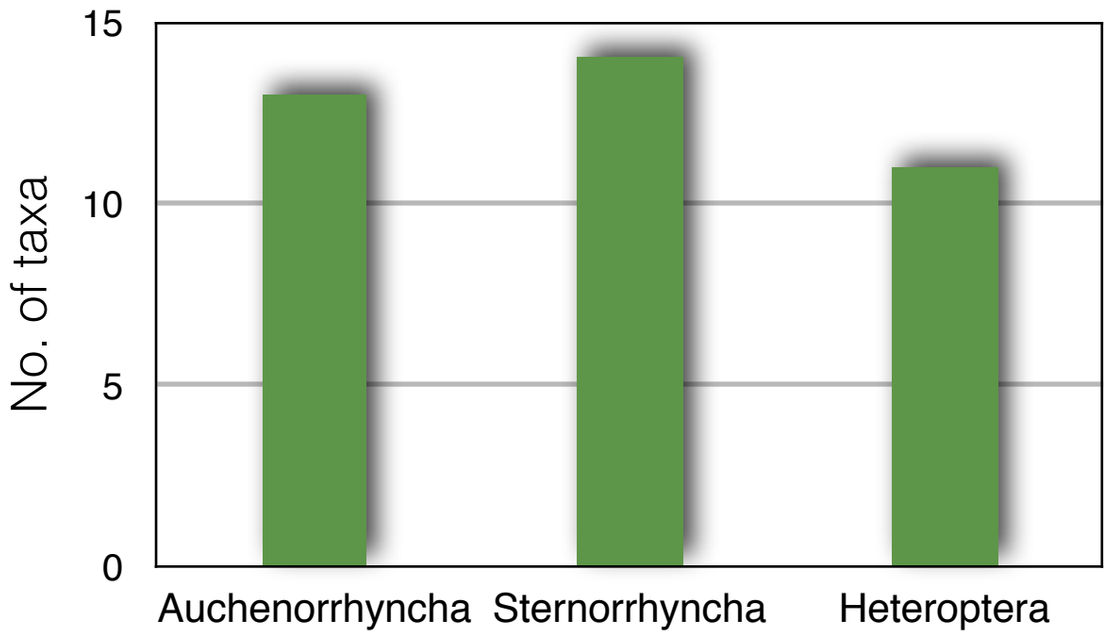
- Developed a linux and python pipeline to automate the retrieval and assembly of SRA data (gsAssembler and Trinity)
- Individual sequencing runs inspected individually using FastQC and PRINSEQ for QC (adaptors, low complexity, poly-A tails, etc.)
- Removal of bacterial, viral, and Ribosomal RNA sequences using DeconSeq and Pathoscope
- Assemblies were quickly check by mapping reads to back to assemblies using BWA
- HaMStR pipeline used to assign orthologs using 4103 Paraneoptera single-copy orthologs from OrthoDB (*Pediculus humanus, Rhodnius prolixus*,* and *Acyrthosiphon pisum*)

Methods: assemble NCBI dbEST data gene alignments

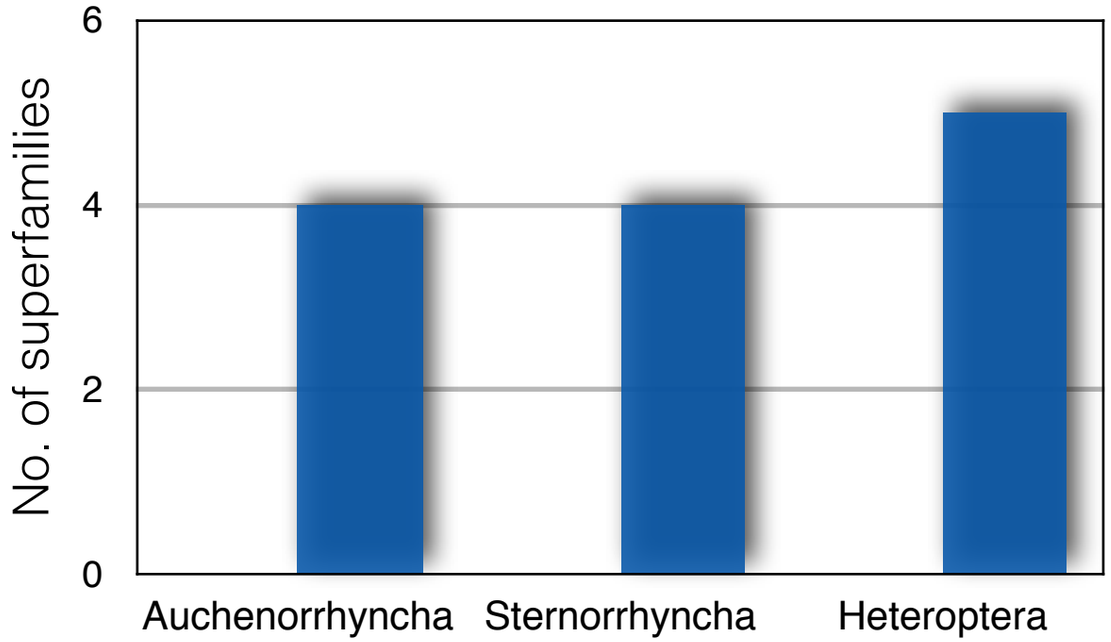
- Taxa with > 6,000 reads
- removed vector, low-complexity reads, and Ribosomal RNA
- assembled and identified orthologs the same as NGS transcriptomes

Results: taxonomic diversity of dbEST and transcriptome data

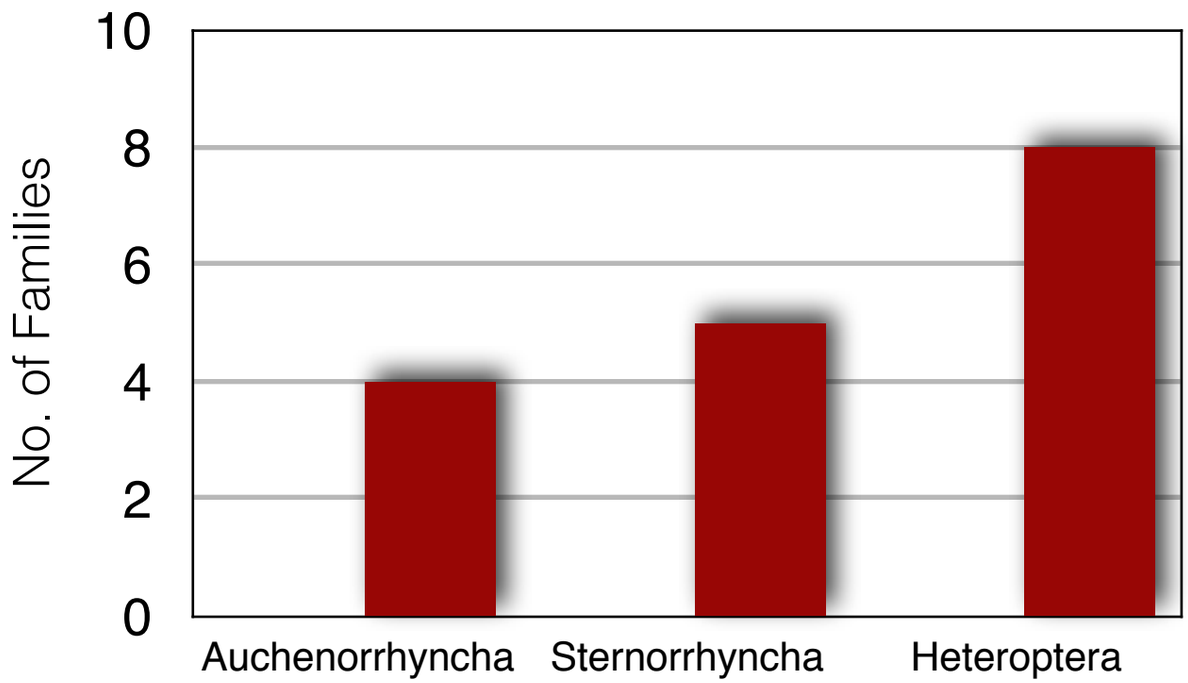
Taxa distributed among suborders



Superfamly diversity within suborders

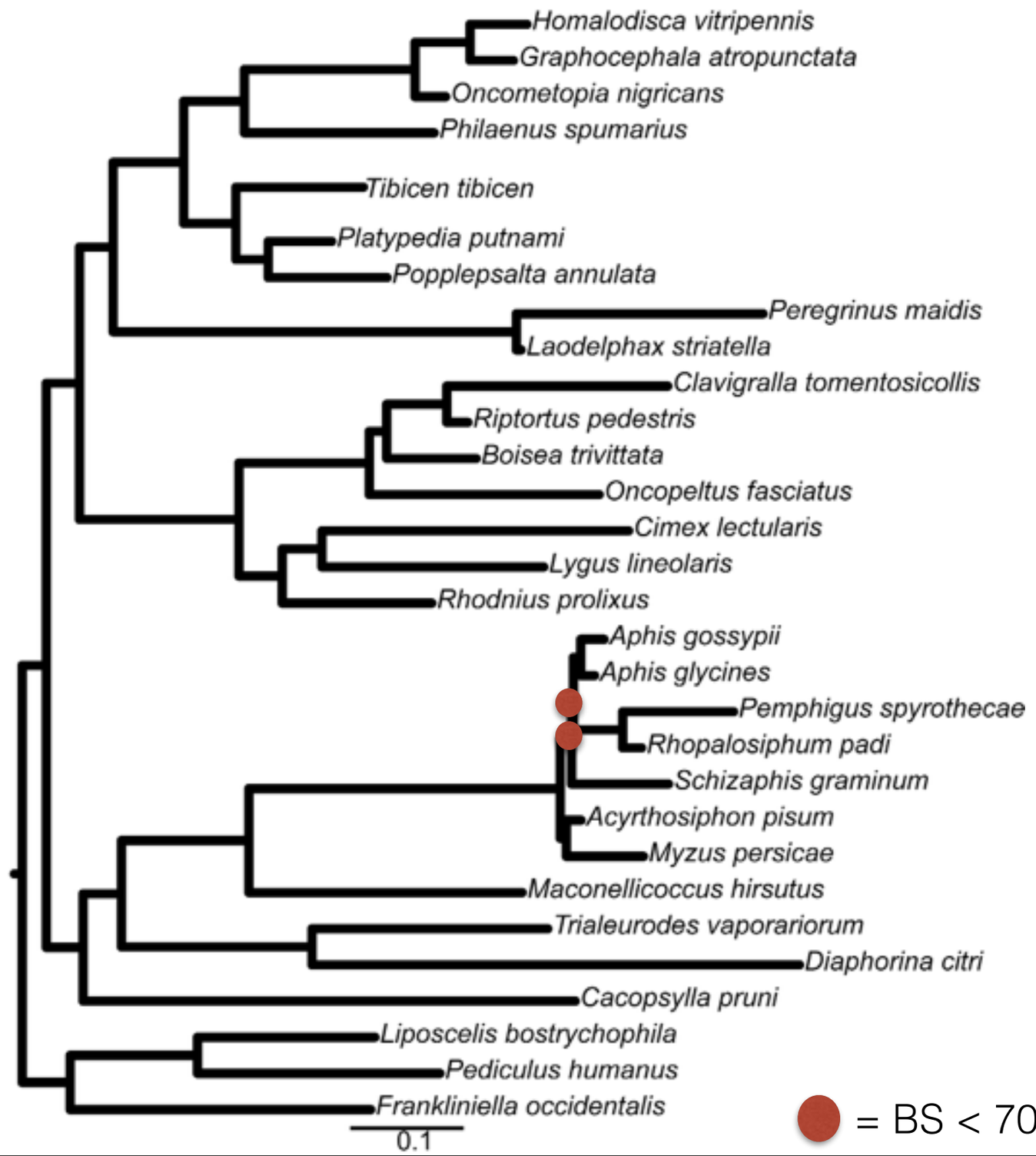


Family diversity within suborders



Hemiptera phylogeny: transcriptome & dbEST data

ML: 35 loci; 15,508AA; partitioned; 100BS reps



Auchenorrhyncha

Graminella nigrifrons, Nilaparvata lugens, Magicicada septendecim, Sogatella furcifera

Heteroptera

Arma chinensis, Lygaeus kalmii, Lygus hesperus, Triatoma rubida

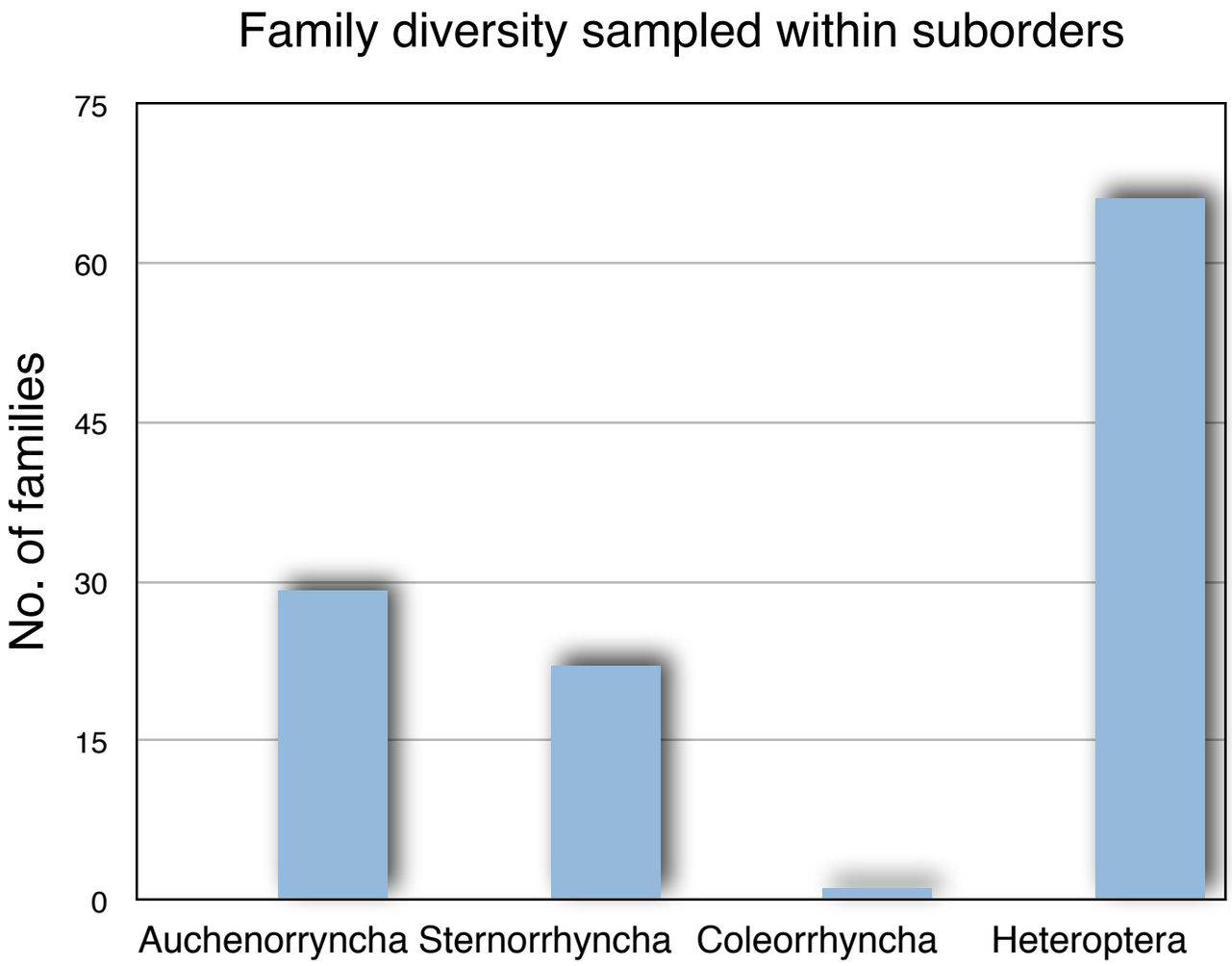
Sternorrhyncha

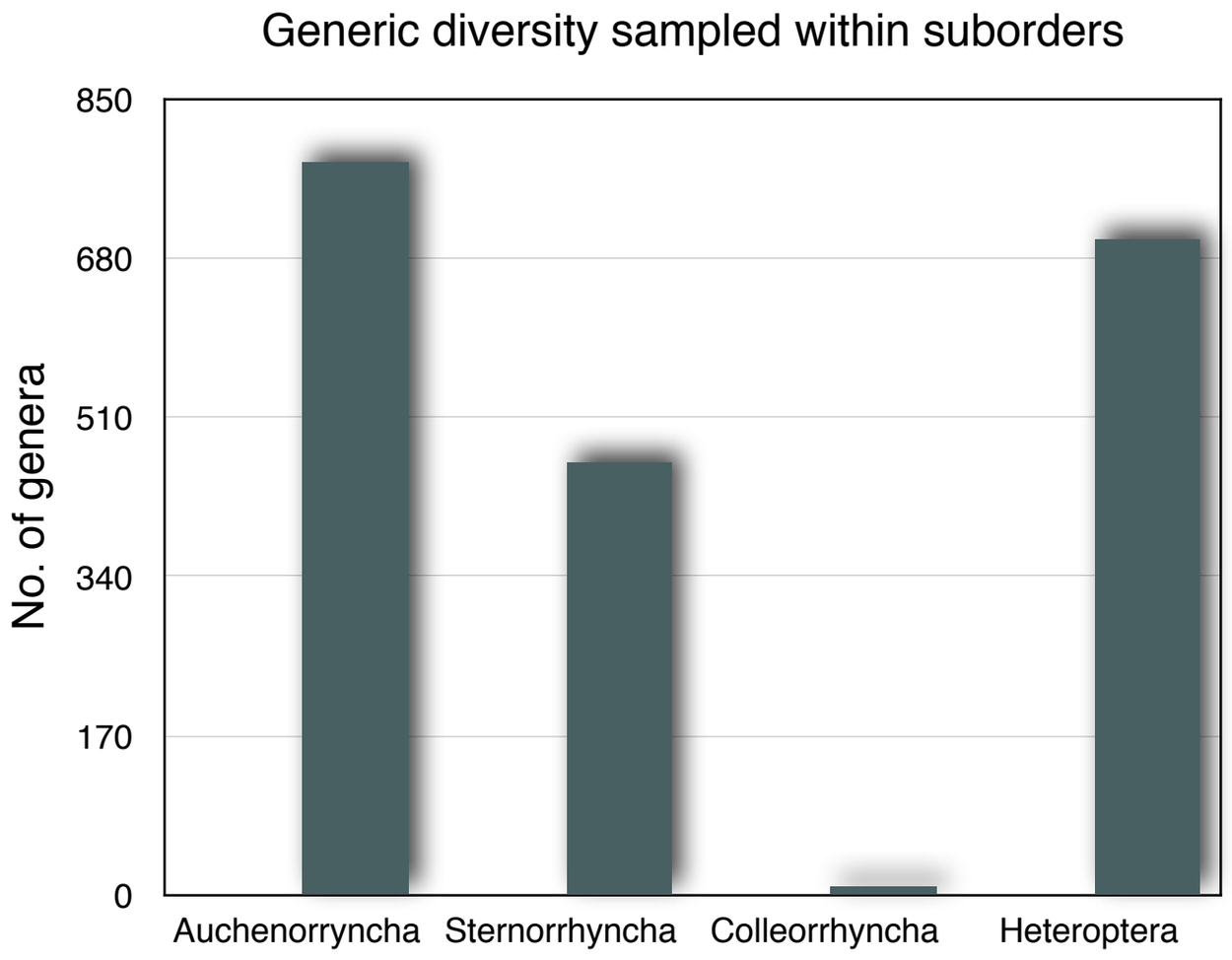
Aphis nerii, Pachypsylla venusta, Sitobion avenae

Methods: assemble Sanger data gene alignments

- Mega-phylogeny method using Phlawd
- GenBank release 194 (February 2013)
- 8 Loci: COI, 12S, 16S, elongation factor 1 alpha (EF1a), wingless (Wg), Histone 3 (H3), 18S, 28S
- BLAST all sequences against a non-redundant database (Genbank Release 198)
- Align protein-coding nucleotides with AA alignment and ribosomal loci by secondary structure in Mafft
- Partitions and evolutionary models estimated in PartitionFinder
- ML nucleotide and AA phylogenies estimated in RAxML and FastTree
- Outgroups: Psocoptera, Thysanoptera, and Phthiraptera

Results: taxonomic diversity of Sanger data



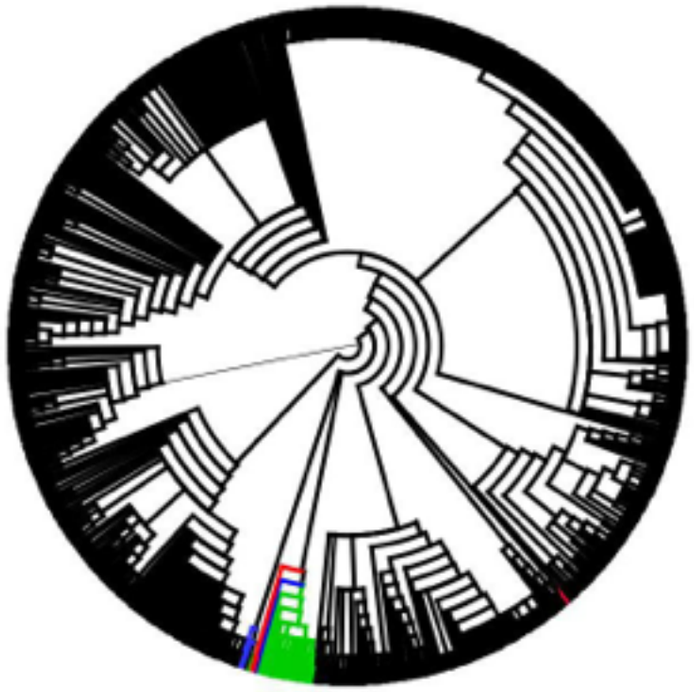


Results: Sanger data quality

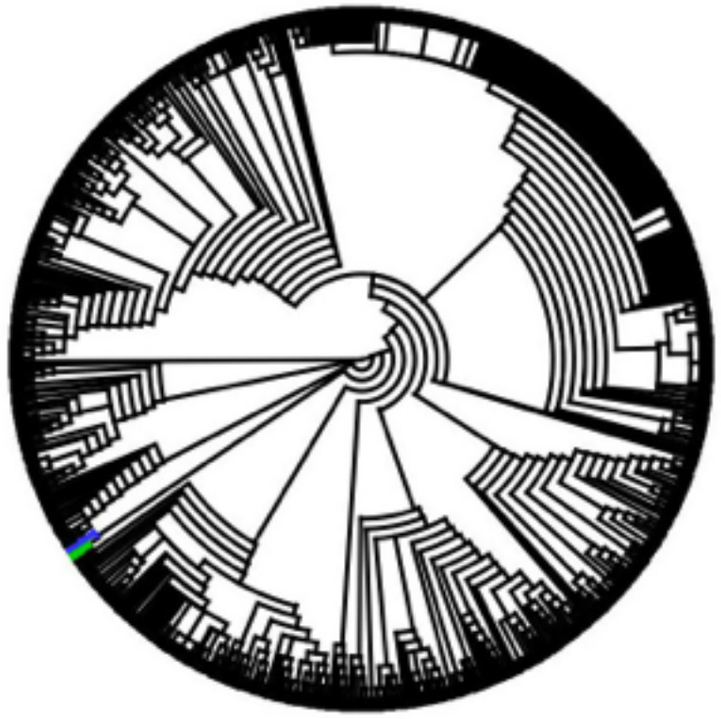
- Removed ~ 5% of sequences
    - stop codons in the middle of coding regions
    - best BLAST hit to bacteria
    - incorrect locus labels

Results: gene trees and non-monophyletic suborders

Wg 1st & 2nd pos;
ML partitioned; 747 taxa



Wg RY-code 3rd pos;
ML partitioned; 747 taxa



Psylids

Outgroup

Heteroptera

Wg 1st & 2nd pos;
rogue taxa removed;
ML partitioned; 747 taxa

Results: concatenated phylogeny

Nucleotides: 5248 taxa, 5956 bp; 75% missing data

- Outgroup non-monphyletic
- Suborders non-monophyletic

Amino acids: 3759 taxa, 1083 AA; 70% missing data

- Outgroup non-monphyletic
- Suborders non-monophyletic

Future directions

- Continue working at the gene tree level, including the transcriptome and EST data
- Try mixture models to accommodate rate heterogeneity.
- Estimate phylogenetic decisiveness and prune taxa accordingly
- Compare nucleotide bias among loci and taxa
- Estimate a synthetic tree using Bayesian phylogeny estimates for genera, families, and tribes to use with backbone of NGS/est data

Questions?