

Custom bioinformatic pipelines for community DNA barcoding

Aug 2014

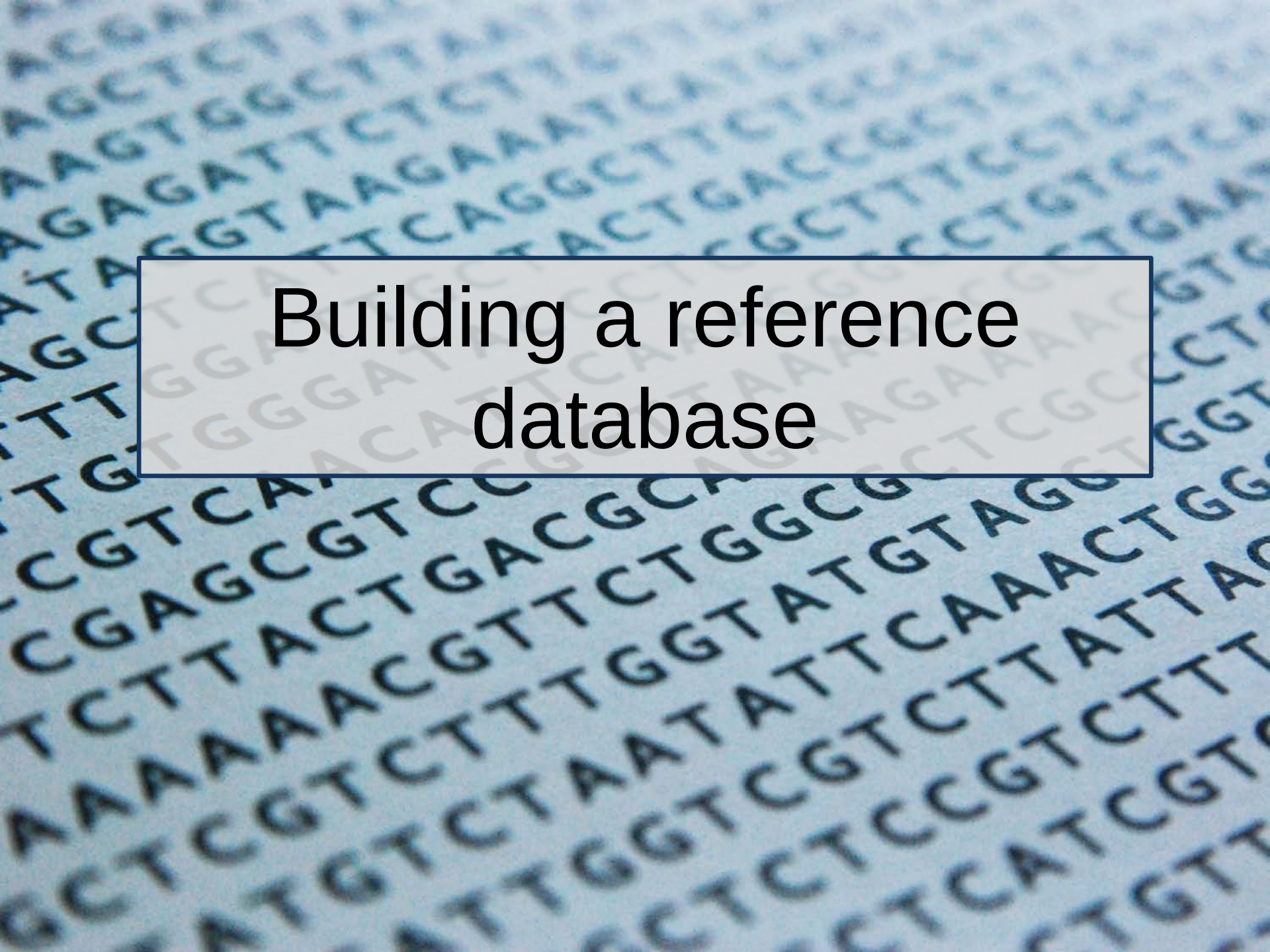
Douglas Chesters,
中国科学院动物研究所



中国科学院
CHINESE ACADEMY OF SCIENCES

Sections

- Building a reference library
- Taxonomic assignment to queries
 - Distance based
 - Phylogenetic based

The background of the slide is a dense, overlapping pattern of DNA sequence characters (A, T, C, G) in various shades of blue and grey, creating a textured, scientific appearance.

Building a reference database

Customized DNA barcoding

Step 1

Step 2

Step 3

Reference / Library



Apis mellifera
AAAAAATTTGGG



Apis cerana
AAAAAAAAAAGGG



Bombus terrestris
TTTTTTTTTTGGG



Bombus terrestris
TTTTTTTTTTGGA



Lasioglossum zephyrum
GGGGGGGGGGGG

Queries



Specimen 1
AAAAATTTTGGG

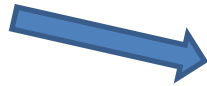


Specimen 2
AAACCAAAGGG



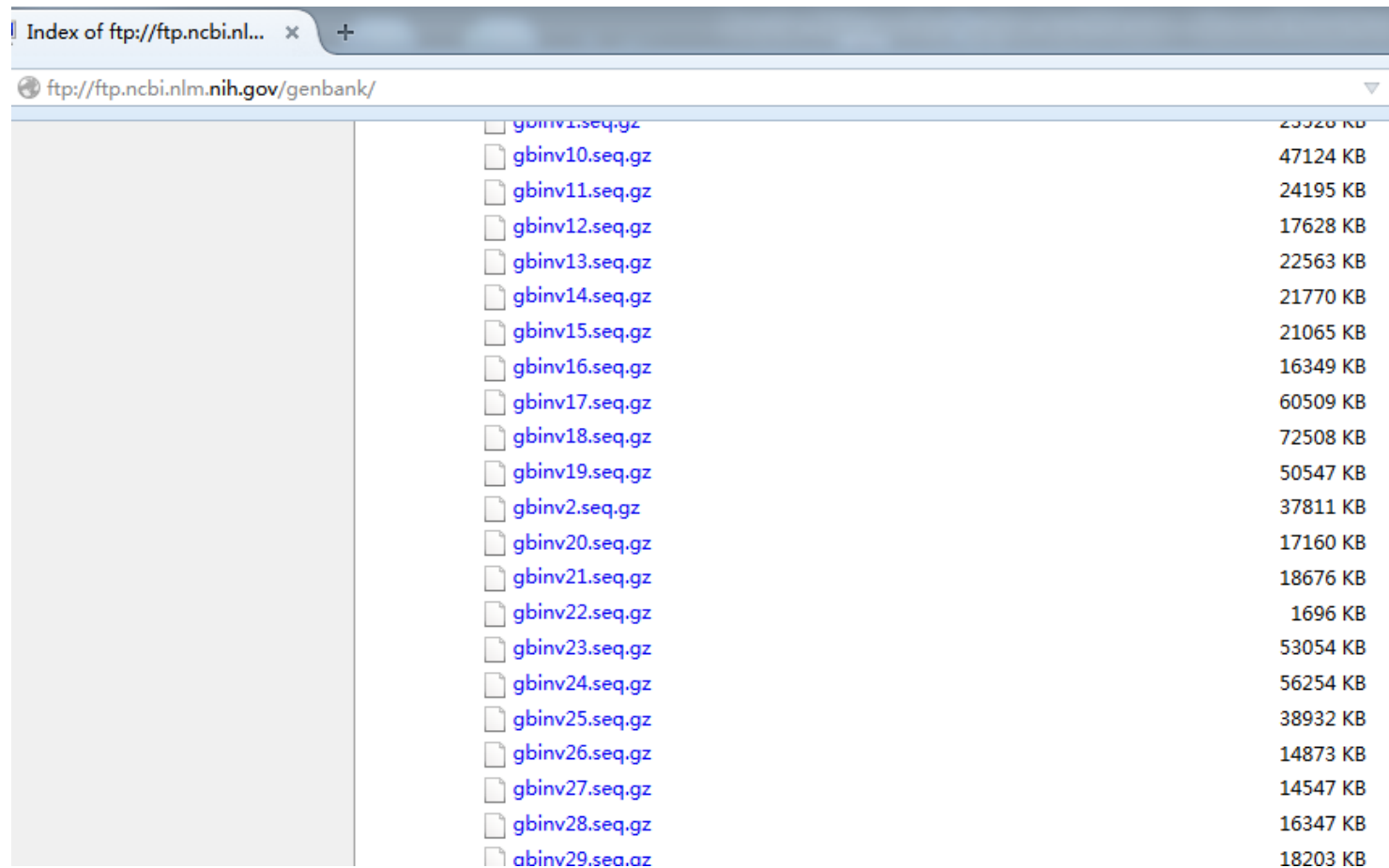
Specimen 3
TTGGGGGGGGGG

Where do we get data, Queries



ATTTTCGCGGGTCGC

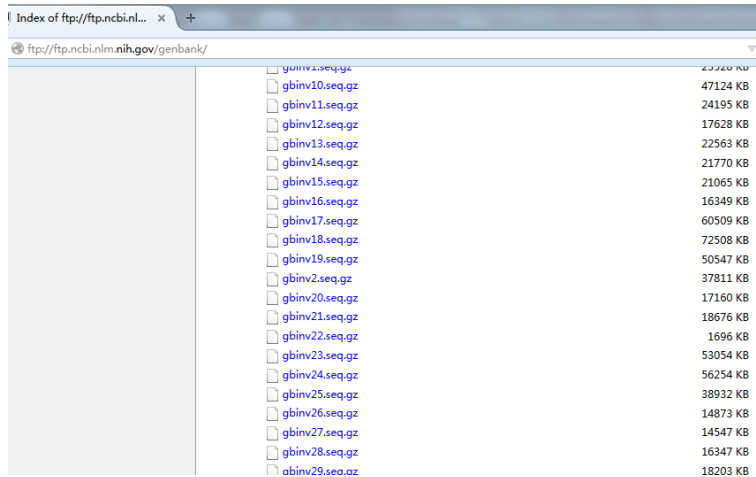
Where do we get data, References (the data mining approach)



A screenshot of a web browser displaying an FTP directory listing for the GenBank database. The browser's address bar shows the URL `ftp://ftp.ncbi.nlm.nih.gov/genbank/`. The directory listing shows a series of files named `gbinv1.seq.gz` through `gbinv29.seq.gz`, with the last file being `abinv29.sea.az`. Each file is preceded by a small folder icon and followed by its size in KB. The files are listed in a non-sequential order, starting with `gbinv1.seq.gz` at the top, followed by `gbinv10.seq.gz` through `gbinv19.seq.gz`, then `gbinv2.seq.gz` through `gbinv28.seq.gz`, and finally `abinv29.sea.az` at the bottom.

File Name	Size (KB)
gbinv1.seq.gz	23328 KB
gbinv10.seq.gz	47124 KB
gbinv11.seq.gz	24195 KB
gbinv12.seq.gz	17628 KB
gbinv13.seq.gz	22563 KB
gbinv14.seq.gz	21770 KB
gbinv15.seq.gz	21065 KB
gbinv16.seq.gz	16349 KB
gbinv17.seq.gz	60509 KB
gbinv18.seq.gz	72508 KB
gbinv19.seq.gz	50547 KB
gbinv2.seq.gz	37811 KB
gbinv20.seq.gz	17160 KB
gbinv21.seq.gz	18676 KB
gbinv22.seq.gz	1696 KB
gbinv23.seq.gz	53054 KB
gbinv24.seq.gz	56254 KB
gbinv25.seq.gz	38932 KB
gbinv26.seq.gz	14873 KB
gbinv27.seq.gz	14547 KB
gbinv28.seq.gz	16347 KB
abinv29.sea.az	18203 KB

Building a reference library from mined data



File Name	Size
gbinv1.seq.gz	23720 KB
gbinv10.seq.gz	47124 KB
gbinv11.seq.gz	24195 KB
gbinv12.seq.gz	17628 KB
gbinv13.seq.gz	22563 KB
gbinv14.seq.gz	21770 KB
gbinv15.seq.gz	21065 KB
gbinv16.seq.gz	16349 KB
gbinv17.seq.gz	60509 KB
gbinv18.seq.gz	72508 KB
gbinv19.seq.gz	50547 KB
gbinv2.seq.gz	37811 KB
gbinv20.seq.gz	17160 KB
gbinv21.seq.gz	18676 KB
gbinv22.seq.gz	1696 KB
gbinv23.seq.gz	53054 KB
gbinv24.seq.gz	56254 KB
gbinv25.seq.gz	38932 KB
gbinv26.seq.gz	14873 KB
gbinv27.seq.gz	14547 KB
gbinv28.seq.gz	16347 KB
abinv29.seq.gz	18203 KB

Reference / Library



Apis mellifera
AAAAAATTTGGG



Apis cerana
AAAAAAAAAAGGG



Bombus terrestris
TTTTTTTTTTGGG



Bombus terrestris
TTTTTTTTTTGGA



Lasioglossum zephyrum
GGGGGGGGGGGGG

Building a reference library from mined data

MOLECULAR ECOLOGY RESOURCES

Molecular Ecology Resources (2014)

doi: 10.1111/1755-0998.12256

Automated DNA-based plant identification for large-scale biodiversity assessment

ANNA PAPADOPOULOU,* DOUGLAS CHESTERS,† INDIANA CORONADO,‡ GISSELA DE LA CADENA,* ANABELA CARDOSO,* JAZMINA C. REYES,‡ JEAN-MICHEL MAES,§ RICARDO M. RUEDA‡ and JESÚS GÓMEZ-ZURITA*

**Animal Biodiversity and Evolution, Institut de Biologia Evolutiva (CSIC-Univ. Pompeu Fabra), 08003 Barcelona, Spain, †Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China, ‡Herbario y Jardín Botánico Ambiental, Universidad Nacional Autónoma de Nicaragua, León, Nicaragua, §Museo Entomológico de León, León, Nicaragua*

Download data

- Automate using wget (linux), curl (OSX)

Process database

- Format, .gb to fasta
- Sequence IDs, accession or GI number, Species name
- Remove sequences too short / long. Remove duplicates

Select taxon

- NCBI taxonomy IDs
 - 2759=Eukaryota, 33208=Metazoa, 50557=Insecta

Find homologs in database

- Which queries?
- Software options, Blast or Uclust
- To consider, algorithm, strand

Process homologs

- Trimming and Orientation
- De-replication: one per species or retain variation; exemplars, random, most complete, or most representative

Download data

- Automate using wget (linux), curl (OSX)

Process database

- Format, .gb to fasta
- Sequence IDs, accession or GI number, Species name
- Remove sequences too short / long. Remove duplicates

Select taxon

- NCBI taxonomy IDs
 - 2759=Eukaryota, 33208=Metazoa, 50557=Insecta

Find homologs in database

- Software options, Blast or Uclust
- To consider, algorithm, strand

Process homologs

- Trimming and Orientation
- De-replication: one per species or retain variation; exemplars, random, most complete, or most representative

Download data

- Automate using wget (linux), curl (OSX)

Process database

- Format, .gb to fasta
- Sequence IDs, accession or GI number, Species name
- Remove sequences too short / long. Remove duplicates

Select taxon

- NCBI taxonomy IDs
 - 2759=Eukaryota, 33208=Metazoa, 50557=Insecta

Find homologs in database

- Software options, Blast or Uclust
- To consider, algorithm, strand

Process homologs

- Trimming and Orientation
- De-replication: one per species or retain variation; exemplars, random, most complete, or most representative

Download data

- Automate using wget (linux), curl (OSX)

Process database

- Format, .gb to fasta
- Sequence IDs, accession or GI number, Species name
- Remove sequences too short / long. Remove duplicates

Select taxon

- NCBI taxonomy IDs
 - 2759=Eukaryota, 33208=Metazoa, 50557=Insecta

Find homologs in database

- Software options, Blast or Uclust
- To consider, algorithm, strand

Process homologs

- Trimming and Orientation
- De-replication: one per species or retain variation; exemplars, random, most complete, or most representative

Download data

- Automate using wget (linux), curl (OSX)

Process database

- Format, .gb to fasta
- Sequence IDs, accession or GI number, Species name
- Remove sequences too short / long. Remove duplicates

Select taxon

- NCBI taxonomy IDs
 - 2759=Eukaryota, 33208=Metazoa, 50557=Insecta

Find homologs in database

- Which queries?
- Software options, Blast or Uclust
- To consider, algorithm, strand

Process homologs

- Trimming and Orientation
- De-replication: one per species or retain variation; exemplars, random, most complete, or most representative

Download data

- Automate using wget (linux), curl (OSX)

Process database

- Format, .gb to fasta
- Sequence IDs, accession or GI number, Species name
- Remove sequences too short / long. Remove duplicates

Select taxon

- NCBI taxonomy IDs
 - 2759=Eukaryota, 33208=Metazoa, 50557=Insecta

Find homologs in database

- Software options, Blast or Uclust
- To consider, algorithm, strand

Process homologs

- Trimming and Orientation
- De-replication: one per species or retain variation; exemplars, random, most complete, or most representative

Reference / Library



Apis mellifera
AAAAAATTG



Apis cerana
AAAAAAAAG



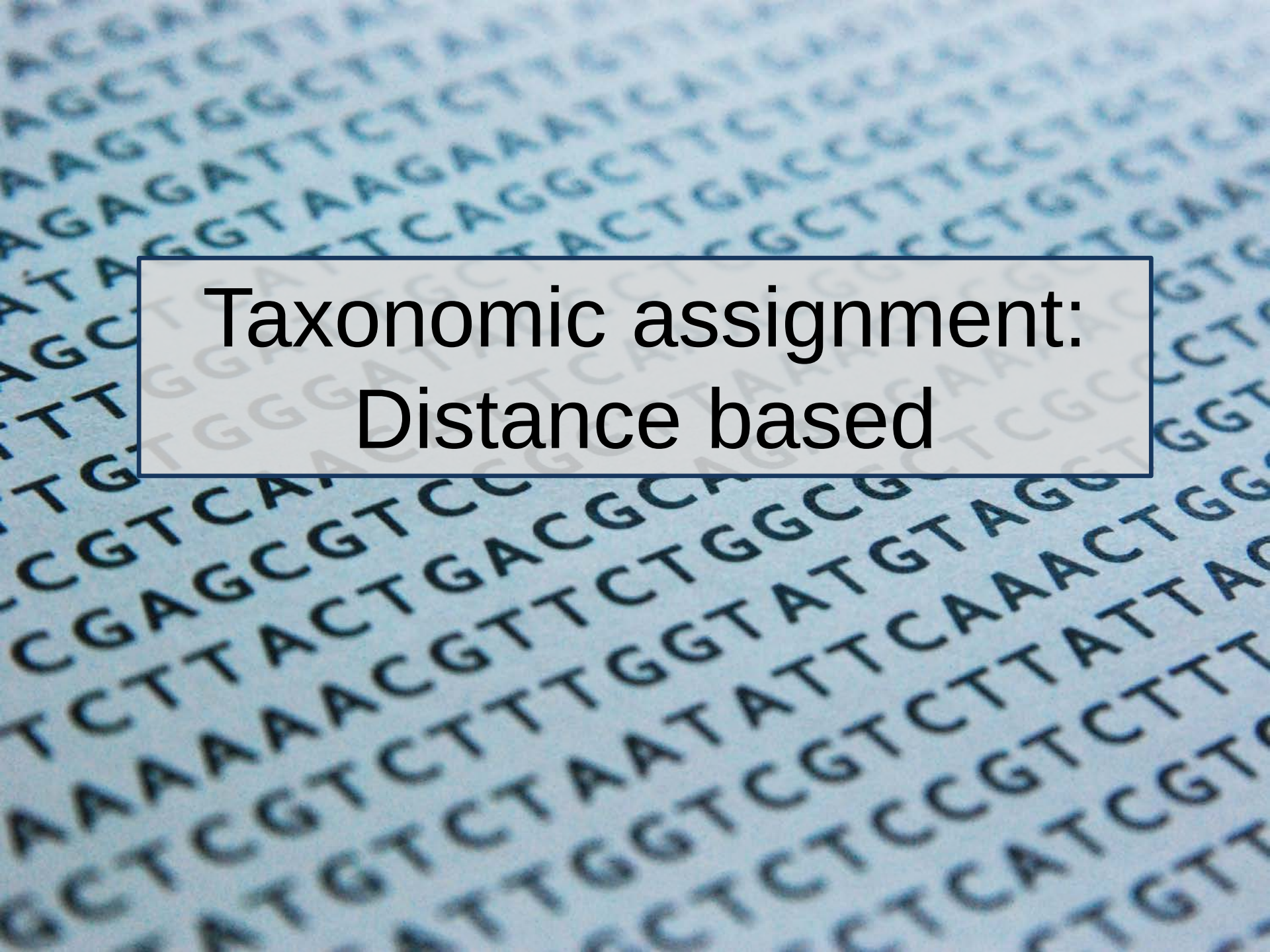
Bombus terrestris
TTTTTTTTT



Bombus terrestris
TTTTTTTTTG



Lasioglossum zephyrum
GGGGGGGGGG

The background of the slide is a light blue color with a pattern of DNA sequence characters (A, T, C, G) and their corresponding base pairs (A-T, C-G) scattered across it. The text is centered within a white rectangular box with a thin dark blue border.

Taxonomic assignment: Distance based

Taxonomic assignment: Distance based

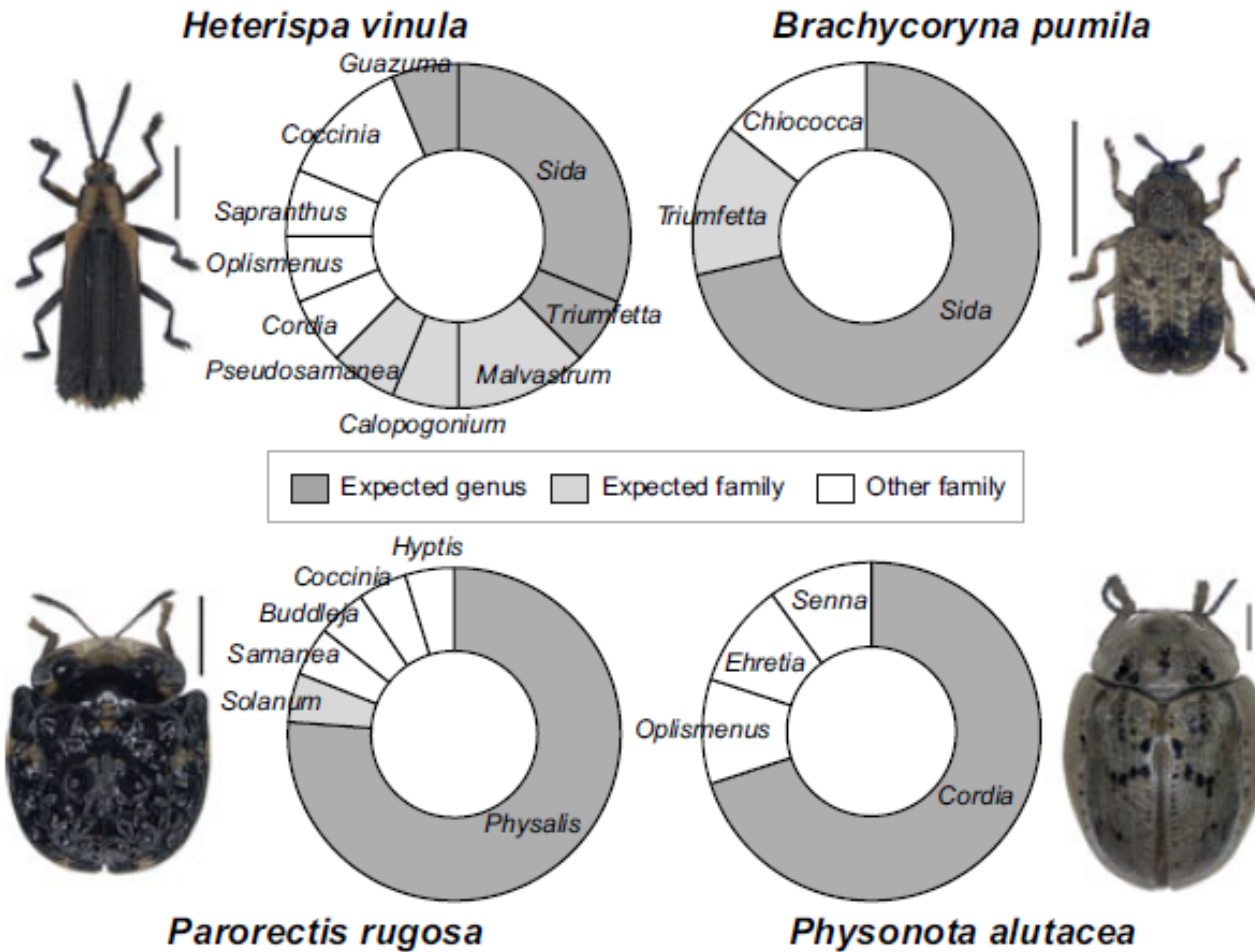
- Advantage
 - MSA not required
 - Computational
- Disadvantage
 - accuracy
- Blast and Uclust
 - local versus global pairwise alignment
 - Distance measure and gaps
 - Gaps as single event, ignore terminal gaps, ambiguities

Improve scoring of a query to reference pair

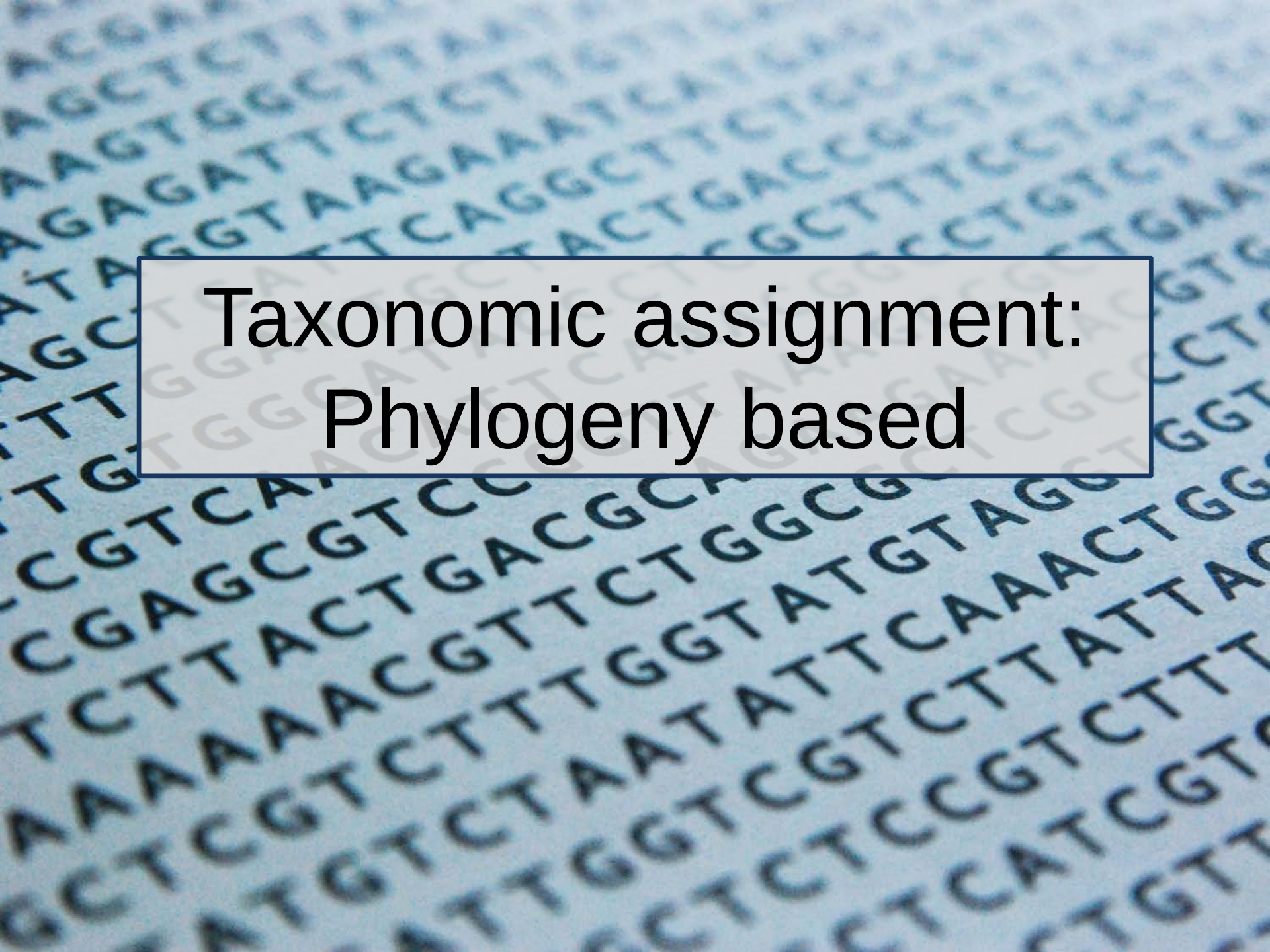
REFERENCE-Apis cerana	CCCCCAAAAATTTTGGGCTGGCAATT
QUERY-Specimen 1	-----AANAATTTT---CTGGCAA---



The image displays a sequence alignment between a reference sequence (Apis cerana) and a query sequence (Specimen 1). The reference sequence is shown as a continuous string of nucleotides: CCCCCAAAATTTTGGGCTGGCAATT. The query sequence is shown as a string of nucleotides with gaps: -----AANAATTTT---CTGGCAA---. The alignment is visualized using colored blocks: blue for the first five nucleotides (CCCCC), green for the next three (AAA), red for the next four (TTTT), black for the next three (GGG), blue for the next two (CT), red for the next two (GG), black for the next two (CA), green for the next two (AA), and red for the last two (TT). The query sequence has gaps (dashes) before the first green block, between the green and red blocks, between the red and black blocks, between the black and blue blocks, and after the last green block.



Papadopoulou et al. 2014.
Molecular Ecology Resources

The background of the slide is a light blue field filled with a dense, overlapping pattern of DNA sequence characters (A, T, C, G) in a darker blue font. The text is slightly blurred and oriented diagonally, creating a sense of depth and scientific context.

Taxonomic assignment: Phylogeny based

Retrieve 16S homologs

- Download mtgenomes for mammal refseq taxa
- Extract 16S rRNA (lrRNA)
- Use these as Blast queries, retrieve only full length entries (>1500)

Species filtering

- Random exemplar / longest / most representative sequence (Chesters and Zhu 2014)

Multiple Sequence Alignment

- References – Muscle (Edgar 2004)
- Queries to Reference profile – Pynast (Caporaso et al. 2010)
- MSA refinement

Infer Phylogenetic relationships

- RAxML (Stamatakis 2006)
- Branch swaps for references constrained according to high quality previously published tree (Bininda-Emonds et al. 2007)

Taxonomic assignment

- Requires complete taxonomies for reference members of the tree
- These are assigned to queries on the tree

Retrieve 16S homologs

- Download mtgenomes for mammal refseq taxa
- Extract 16S rRNA (lrRNA)
- Use these as Blast queries, retrieve only full length entries (>1500)

Species filtering

- Random exemplar / longest / most representative sequence (Chesters and Zhu 2014)

Multiple Sequence Alignment

- References – Muscle (Edgar 2004)
- Queries to Reference profile – Pynast (Caporaso et al. 2010)
- MSA refinement

Infer Phylogenetic relationships

- RAxML (Stamatakis 2006)
- Branch swaps for references constrained according to high quality previously published tree (Bininda-Emonds et al. 2007)

Taxonomic assignment

- Requires complete taxonomies for reference members of the tree
- These are assigned to queries on the tree

Retrieve 16S homologs

- Download mtgenomes for mammal refseq taxa
- Extract 16S rRNA (lrRNA)
- Use these as Blast queries, retrieve only full length entries (>1500)

Species filtering

- Random exemplar / longest / most representative sequence (Chesters and Zhu 2014)

Multiple Sequence Alignment

- References – Muscle (Edgar 2004)
- Queries to Reference profile – Pynast (Caporaso et al. 2010)
- MSA refinement

Infer Phylogenetic relationships

- RAxML (Stamatakis 2006)
- Branch swaps for references constrained according to high quality previously published tree (Bininda-Emonds et al. 2007)

Taxonomic assignment

- Requires complete taxonomies for reference members of the tree
- These are assigned to queries on the tree

Retrieve 16S homologs

- Download mtgenomes for mammal refseq taxa
- Extract 16S rRNA (lrRNA)
- Use these as Blast queries, retrieve only full length entries (>1500)

Species filtering

- Random exemplar / longest / most representative sequence (Chesters and Zhu 2014)

Multiple Sequence Alignment

- References – Muscle (Edgar 2004)
- Queries to Reference profile – Pynast (Caporaso et al. 2010)
- MSA refinement

Infer Phylogenetic relationships

- RAxML (Stamatakis 2006)
- Branch swaps for references constrained according to high quality previously published tree (Bininda-Emonds et al. 2007)

Taxonomic assignment

- Requires complete taxonomies for reference members of the tree
- These are assigned to queries on the tree

Retrieve 16S homologs

- Download mtgenomes for mammal refseq taxa
- Extract 16S rRNA (lrRNA)
- Use these as Blast queries, retrieve only full length entries (>1500)

Species filtering

- Random exemplar / longest / most representative sequence (Chesters and Zhu 2014)

Multiple Sequence Alignment

- References – Muscle (Edgar 2004)
- Queries to Reference profile – Pynast (Caporaso et al. 2010)
- MSA refinement

Infer Phylogenetic relationships

- RAxML (Stamatakis 2006)
- Branch swaps for references constrained according to high quality previously published tree (Bininda-Emonds et al. 2007)

Taxonomic assignment

- Requires complete taxonomies for reference members of the tree
- These are assigned to queries on the tree

Retrieve 16S homologs

- Download mtgenomes for mammal refseq taxa
- Extract 16S rRNA (lrRNA)
- Use these as Blast queries, retrieve only full length entries (>1500)

Species filtering

- Random exemplar / longest / most representative sequence (Chesters and Zhu 2014)

Multiple Sequence Alignment

- References – Muscle (Edgar 2004)
- Queries to Reference profile – Pynast (Caporaso et al. 2010)
- MSA refinement

Infer Phylogenetic relationships

- RAxML (Stamatakis 2006)
- Branch swaps for references constrained according to high quality previously published tree (Bininda-Emonds et al. 2007)

Taxonomic assignment

- Requires complete taxonomies for reference members of the tree
- These are assigned to queries on the tree

Overview:

Phylogeny for taxonomic assignment

- De novo tree of both references and queries
 - bagpipe_phylo.pl
- New queries added to existing phylogeny
 - EPA
 - pplacer

Taxonomic assignment on de novo tree of references and queries

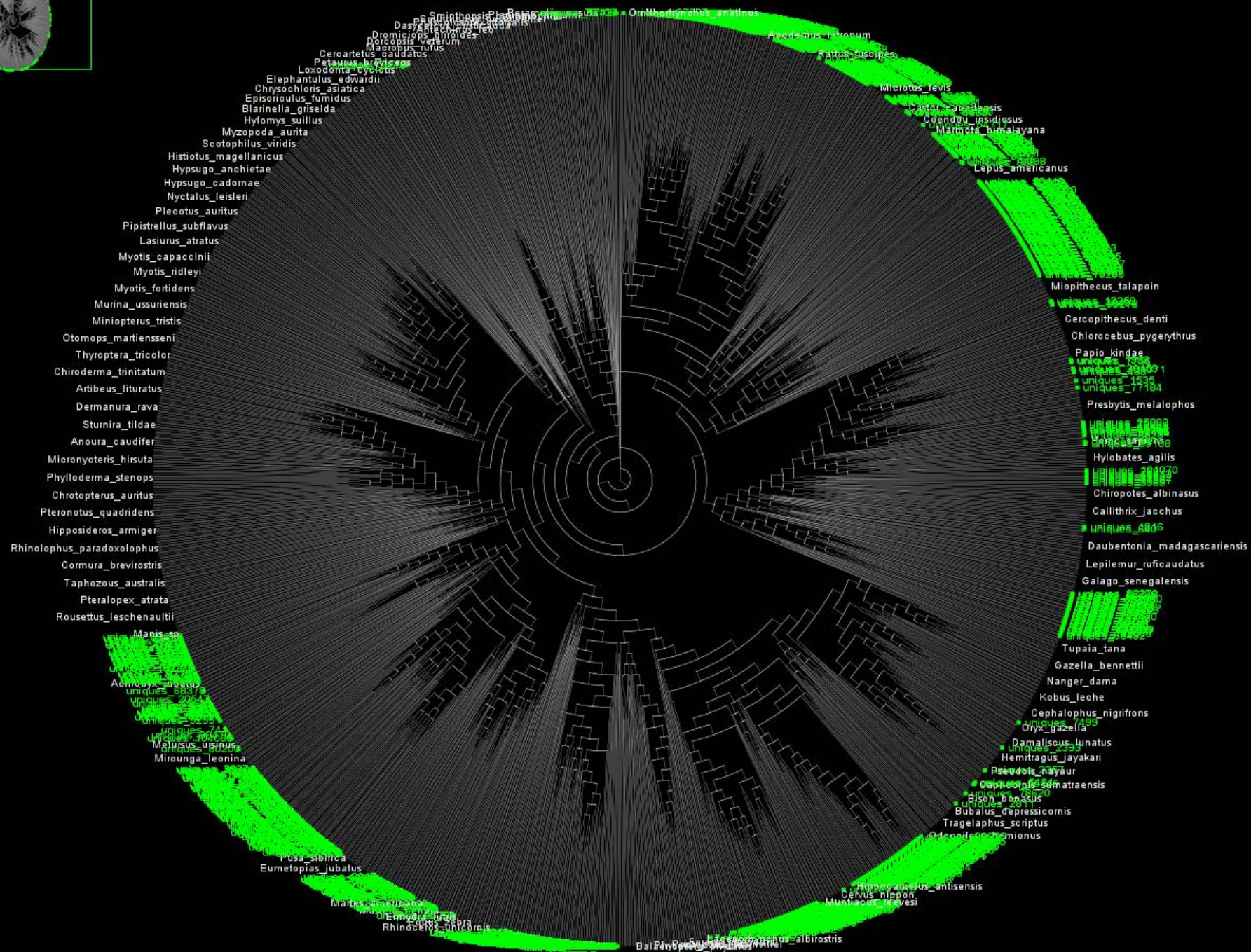
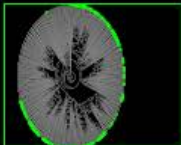
- bagpipe_phylo
 - New standalone implementation of a key component of BAGpipe (Papadopoulou et al. 2014)
 - Requires phylogeny of references and queries, and the NCBI taxonomy database
 - Will be available soon ...

Example: Next Generation Sequencing of Leech Gut Contents

- Mammal 16S rDNA
- Degraded DNA
- Only short fragments retrievable.



unpublished,
in collaboration with
Douglas Yu (KIZ)



uniques_4232
uniques_27219
uniques_77074
uniques_76849
uniques_589
uniques_3042
uniques_34552
uniques_29120
uniques_52485
uniques_44622
uniques_99582
uniques_46469
uniques_2209
uniques_91979
uniques_84830
uniques_59184
uniques_95770
uniques_102340
HYAENIDAE_Hyaena_hyaena
HYAENIDAE_Crocuta_crocuta
uniques_82159
uniques_37721
uniques_87
uniques_76209
uniques_58989
uniques_53380
FELIDAE_Acinonyx_jubatus
FELIDAE_Lynx_rufus
FELIDAE_Neofelis_nebulosa
FELIDAE_Uncia_uncia
FELIDAE_Panthera_tigris
uniques_68370
FELIDAE_Panthera_onca
FELIDAE_Panthera_pardus
FELIDAE_Felis_catus
FELIDAE_Panthera_leo
uniques_20547
FELIDAE_Prionailurus_bengalensis
FELIDAE_Puma_concolor
uniques_45630
uniques_57
uniques_8807
uniques_2397
uniques_2971
uniques_98210
uniques_42257
uniques_34634
uniques_82037
uniques_102131
HERPESTIDAE_Herpestes_javanicus
uniques_52551
CANIDAE_Vulpes_zerda
CANIDAE_Vulpes_vulpes
CANIDAE_Vulpes_corsac
CANIDAE_Nyctereutes_procyonoides
CANIDAE_Cuon_alpinus
uniques_741
CANIDAE_Canis_latrans
CANIDAE_Canis_lycaon
CANIDAE_Canis_lupus
uniques_80470
CANIDAE_Chrysocyon_brachyurus
uniques_104080
URSIDAE_Ailuropoda_melanoleuca
URSIDAE_Tremarctos_ornatus
URSIDAE_Arctodus_simus
URSIDAE_Melursus_ursinus
URSIDAE_Helarctos_malayanus
URSIDAE_Ursus_thibetanus
uniques_80200
URSIDAE_Ursus_maritimus
URSIDAE_Ursus_arctos
URSIDAE_Ursus_deningeri
URSIDAE_Ursus_spelaeus
URSIDAE_Ursus_americanus

bagpipe_phylo output:

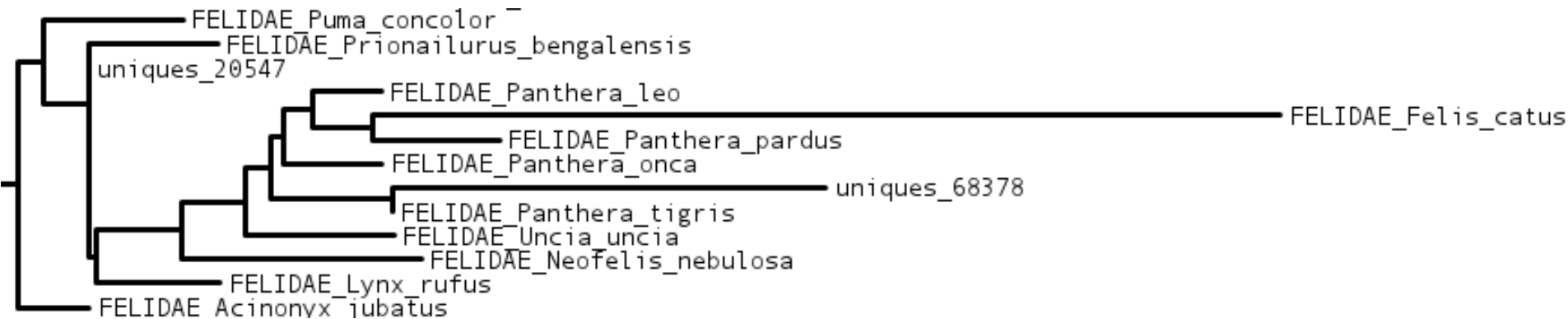
Top 14 ...

Query	Support	Shared Taxon	Shortest Distance to Reference Leaf
uniques_1307	44	Melogale_moschata	0.0000080
uniques_2811	65	Bos_indicus	0.0000080
uniques_5424	22	Capricornis_sumatraensis	0.0000080
uniques_617	94	Martes_flavigula	0.0000080
uniques_961	79	Homo_sapiens	0.0000080
uniques_80200	77	Ursus_thibetanus	0.01261555
uniques_741	48	Canis_latrans	0.013661082
uniques_7756	87	Papio_anubis	0.026794699
uniques_1993	94	Papio_anubis	0.037582712
uniques_4232	69	Manis_pentadactyla	0.043805173
uniques_20547	14	Felidae	0.04743098
uniques_2229	25	Rattus_rattus	0.04824355
uniques_2357	33	Caprinae	0.051507736
uniques_468	30	Macaca	0.07541466

bagpipe_phylo results:

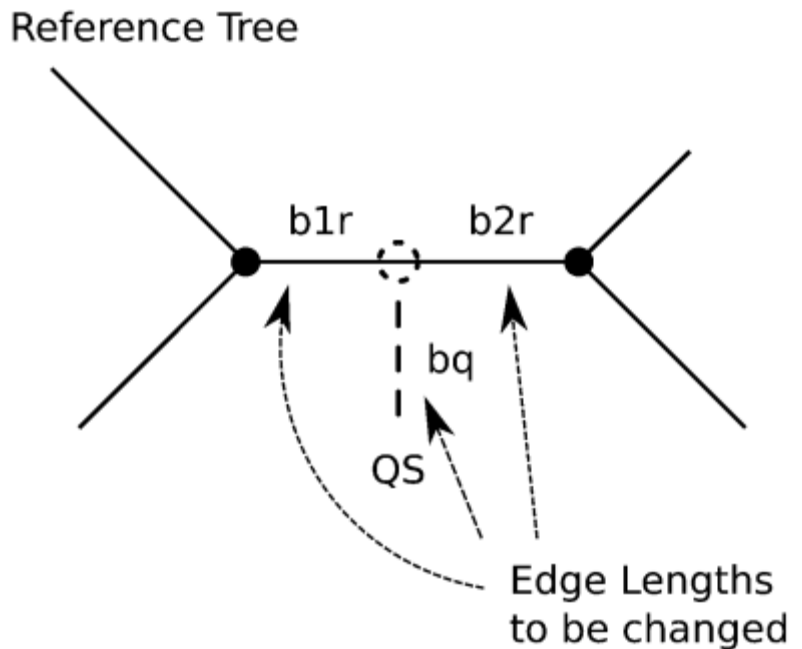
shared taxonomies assigned to queries

Query	Support	Shared Taxon	Shortest Distance to Reference Leaf
uniques_20547	14	Felidae	0.04743098
uniques_68378	31	Panthera_tigris	0.159440927

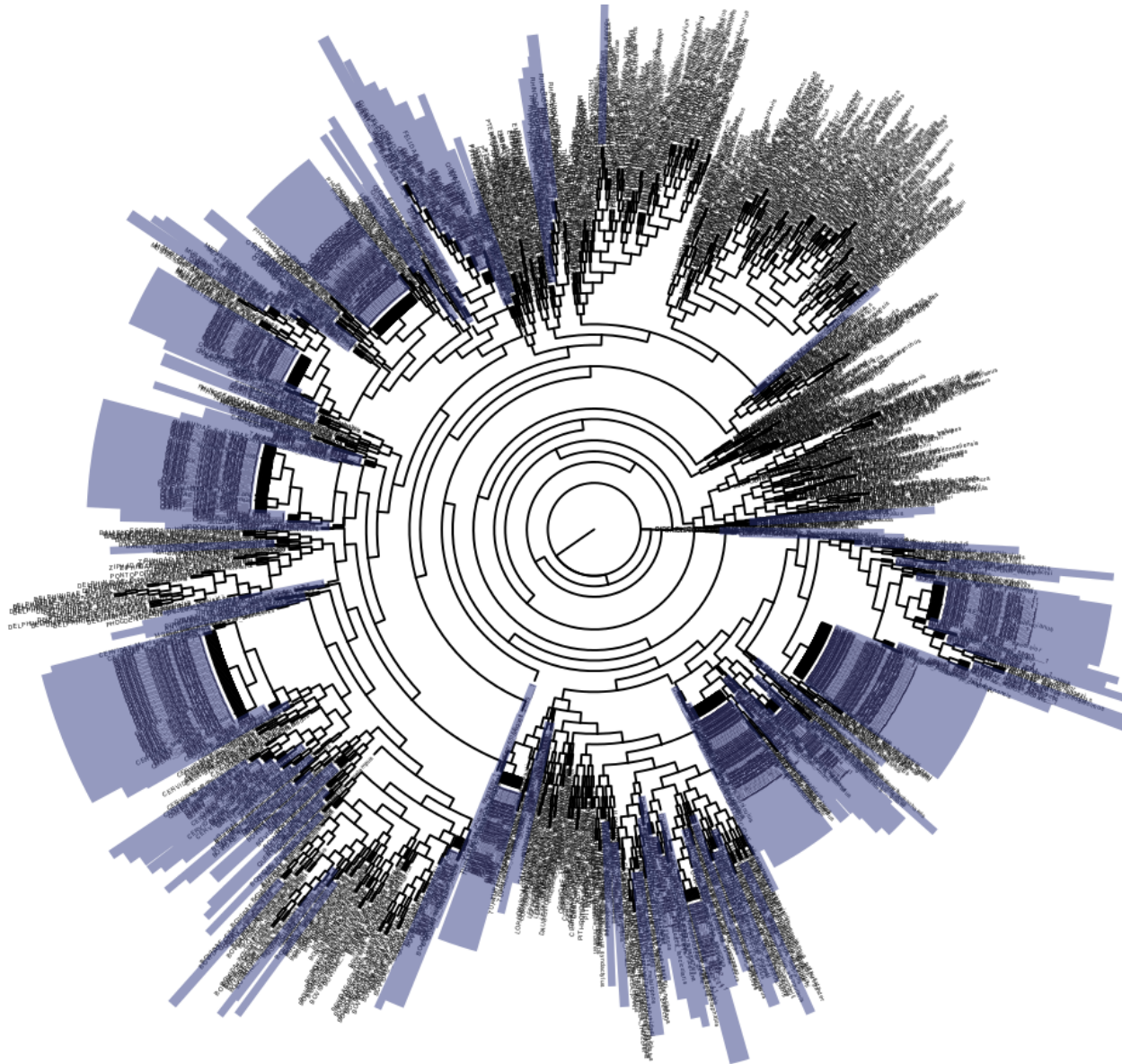


Evolutionary Placement Algorithms

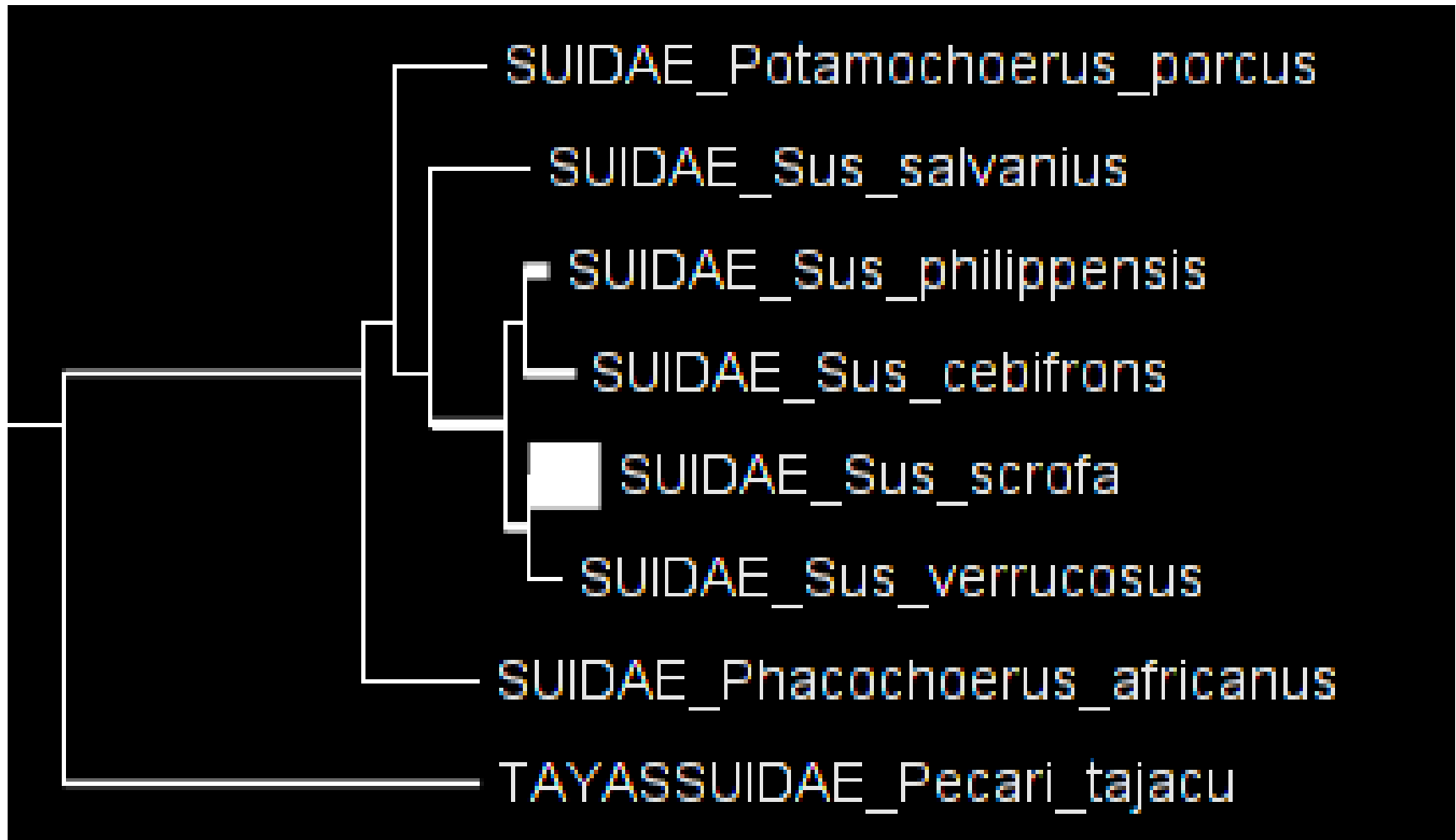
- New data assigned to existing tree
 - EPA in RAXML
 - pplacer



Evolutionary Placement in RAXML



Evolutionary Placement using pplacer



Conclusion

- Building an automated pipeline allows for easy replication / updating
- Customization:
 - can avoid the fitting an unsuited method
 - novelty!

Acknowledgments

- PI: 朱朝东
- Collaborators for the current work:
 - Anna Papadopoulou & Jesus Gomez-Zurita
 - Douglas Yu



中国科学院
CHINESE ACADEMY OF SCIENCES



Thanks for listening



中国科学院
CHINESE ACADEMY OF SCIENCES