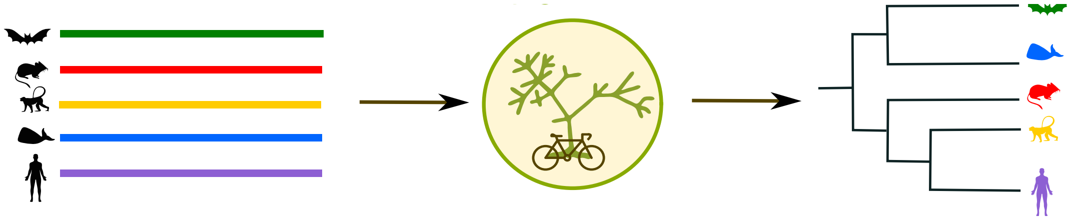
UC San Diego
JACOBS SCHOOL OF ENGINEERING

Accurate, scalable, and fully automated inference of species trees from raw genome assemblies using ROADIES

Anshu Gupta
Department of Computer Science and Engineering

July 15, 2024



Era of Phylogenomics: Influx of genomic data

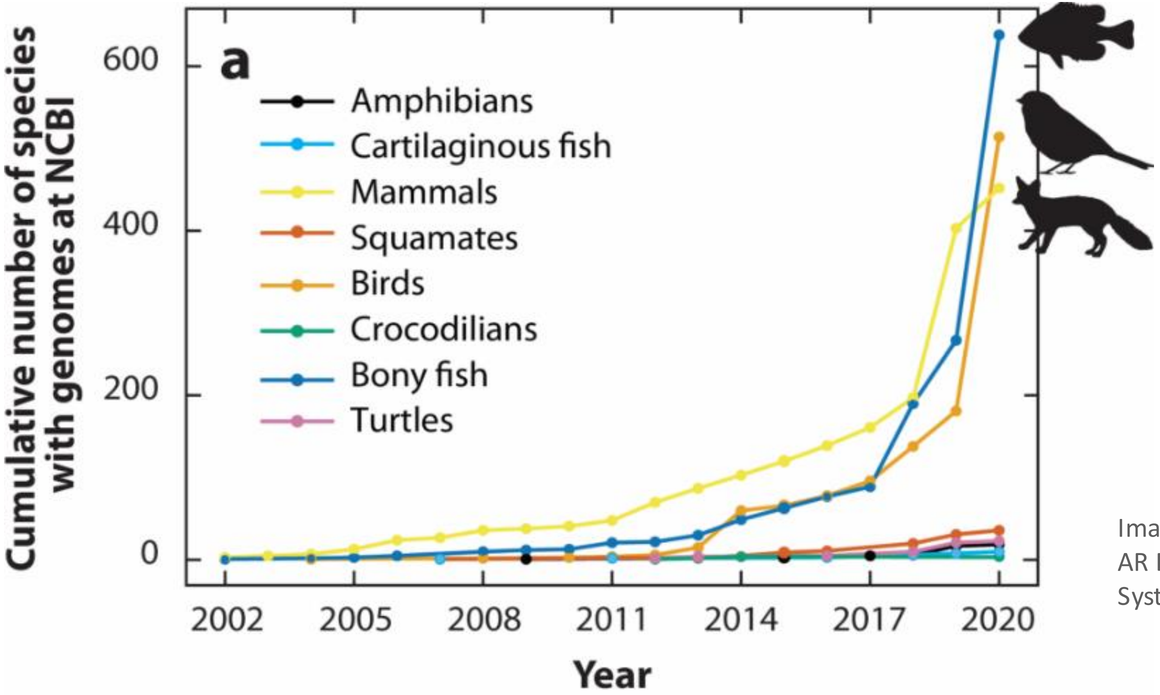
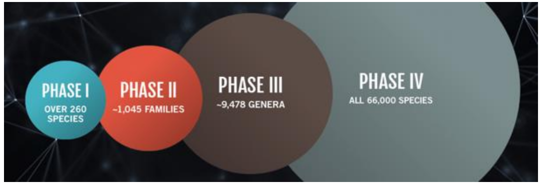


Image source: G A Bravo et al., AR Ecology, Evolution, and Systematics, 2021

Large-scale Genomic Sequencing

- Multiple consortiums aiming to sequence thousands to millions of species



VGP VERTEBRATE GENOMES PROJECT

A PROJECT OF THE G10K CONSORTIUM

GENOME 10K

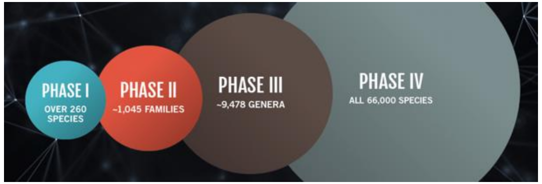
B10K

BAT 1K

EARTH BIOGENOME PROJECT

Large-scale Genomic Sequencing

- Multiple consortiums aiming to sequence thousands to millions of species



- Phylogenomic analyses of huge datasets solves various questions related to Tree of Life

VGP VERTEBRATE GENOMES PROJECT

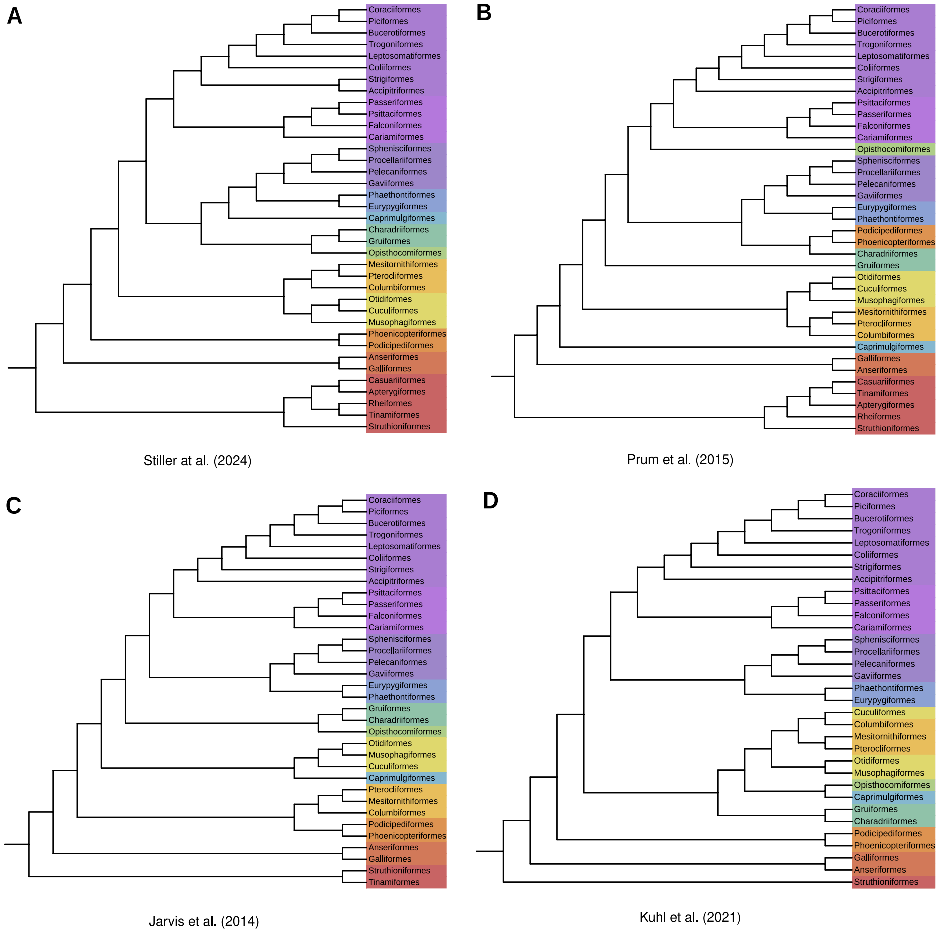
A PROJECT OF THE G10K CONSORTIUM

GENOME 10K

B10K

BAT 1K

EARTH BIOGENOME PROJECT



Stiller at al. (2024)

Prum et al. (2015)

Jarvis et al. (2014)

Kuhl et al. (2021)

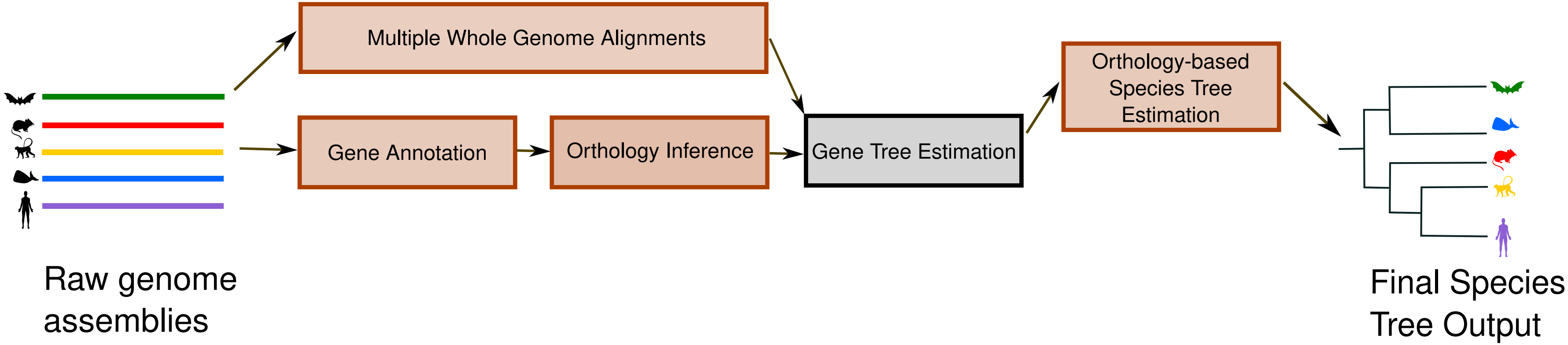
Evolutionary Trees are still debated!

Figure shows different avian phylogenies proposed by various groups

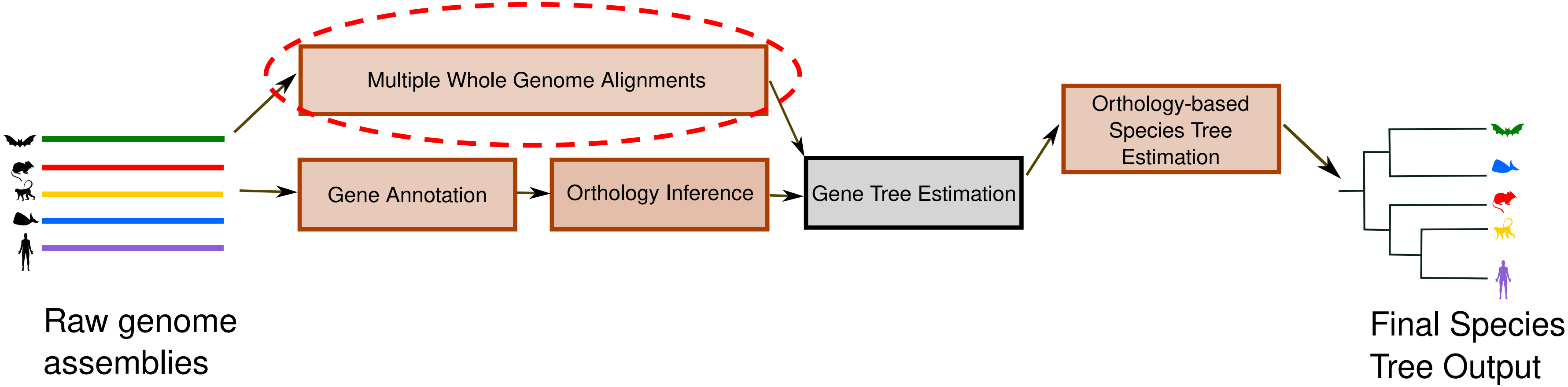
Image source:

A. Stiller, J. et al. Complexity of avian evolution revealed by family-level genomes. Nature (2024).
B. Prum, R., Berv, J., Dornburg, A. et al. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature (2015).
C. Jarvis, E. D. et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science (2014).
D. Kuhl et al., An Unbiased Molecular Approach Using 3’-UTRs Resolves the Avian Family-Level Tree of Life, MBE (2021).

Existing approaches to estimate species tree

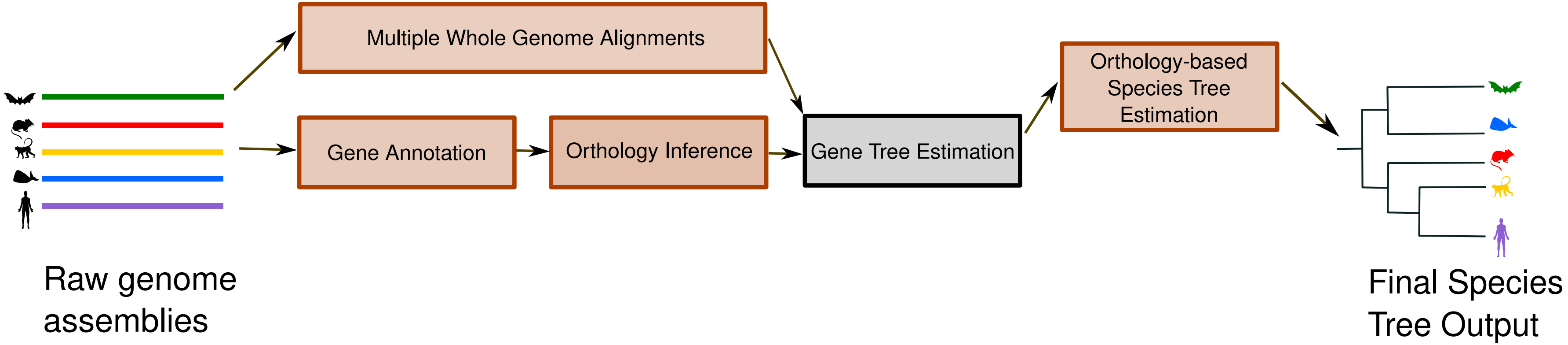

Multiple Whole Genome Alignments
Gene Annotation
Orthology Inference
Gene Tree Estimation
Orthology-based Species Tree Estimation
Raw genome assemblies
Final Species Tree Output

Existing approaches to estimate species tree

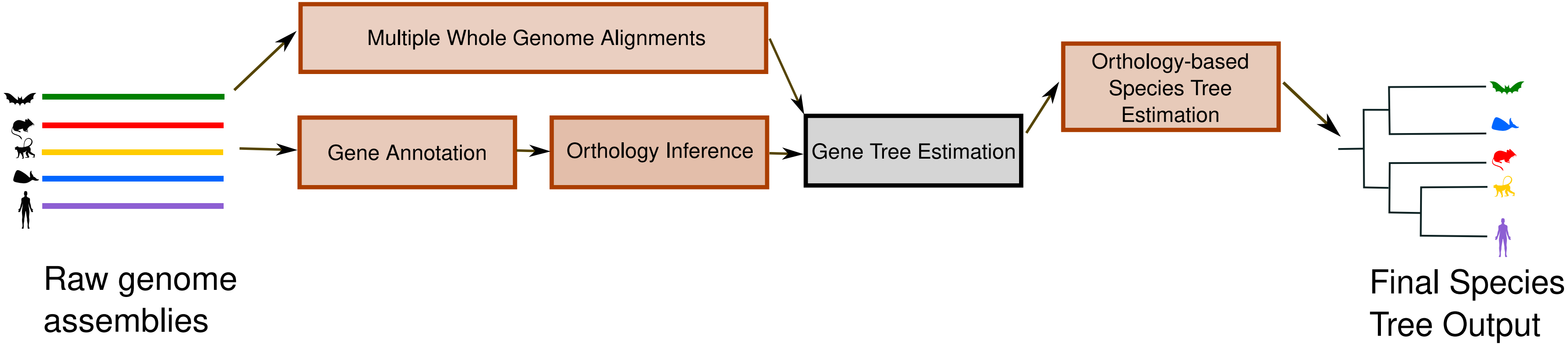

Multiple Whole Genome Alignments
Gene Annotation
Orthology Inference
Gene Tree Estimation
Orthology-based Species Tree Estimation
Raw genome assemblies
Final Species Tree Output

Existing approaches to estimate species tree

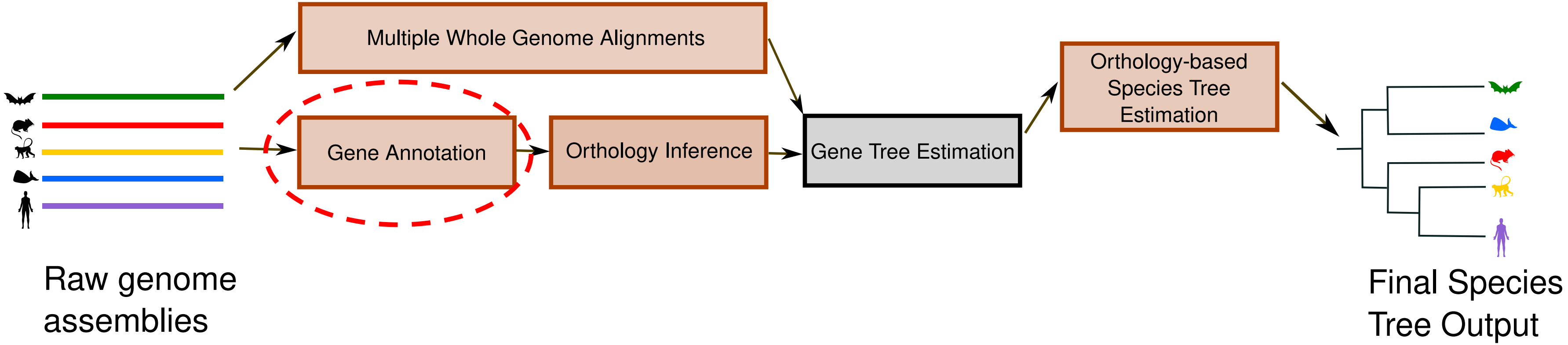
- Computationally expensive for larger datasets
- Requires guide tree as input


Multiple Whole Genome Alignments
Gene Annotation
Orthology Inference
Gene Tree Estimation
Orthology-based Species Tree Estimation
Raw genome assemblies
Final Species Tree Output

Existing approaches to estimate species tree


Multiple Whole Genome Alignments
Gene Annotation
Orthology Inference
Gene Tree Estimation
Orthology-based Species Tree Estimation
Raw genome assemblies
Final Species Tree Output

Existing approaches to estimate species tree



Multiple Whole Genome Alignments

Gene Annotation

Orthology Inference

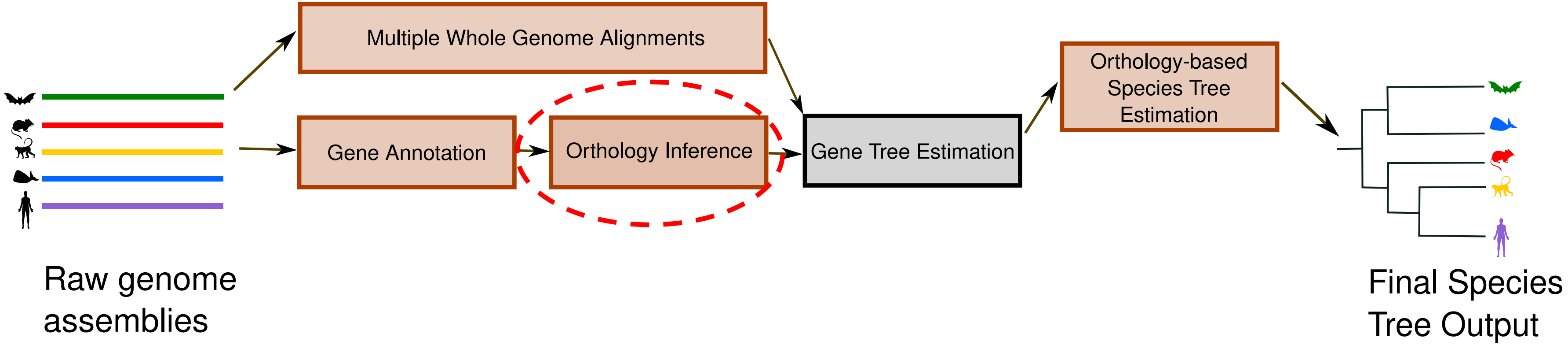
Gene Tree Estimation

Orthology-based Species Tree Estimation

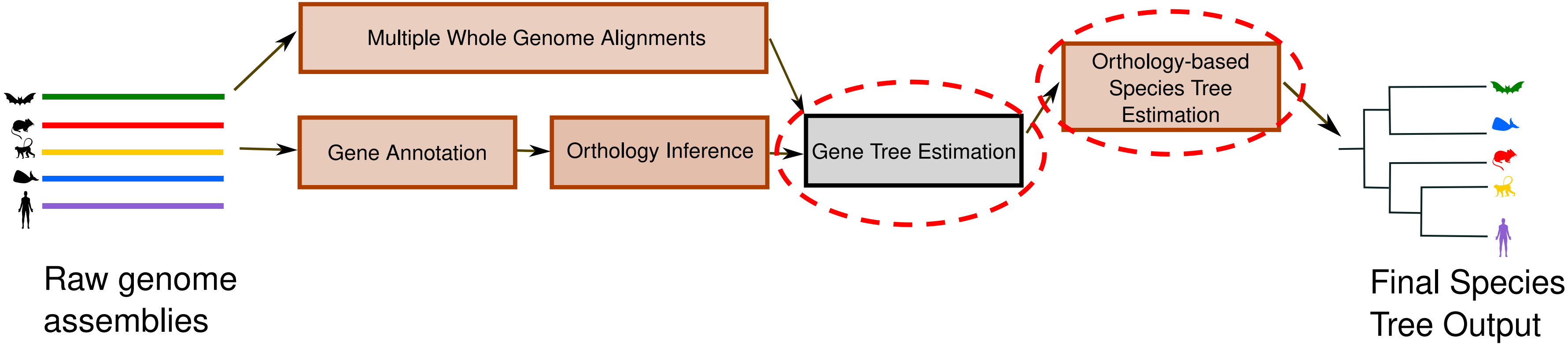
Raw genome assemblies

Final Species Tree Output

Existing approaches to estimate species tree

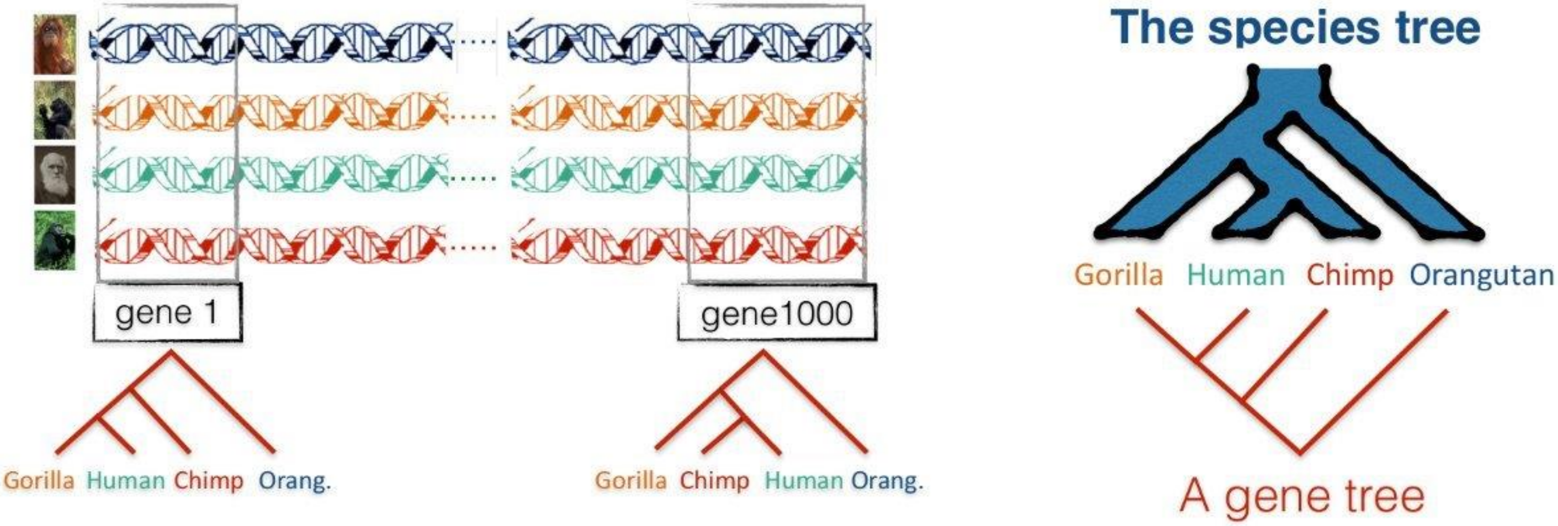

Multiple Whole Genome Alignments
Gene Annotation
Orthology Inference
Gene Tree Estimation
Orthology-based Species Tree Estimation
Raw genome assemblies
Final Species Tree Output

Existing approaches to estimate species tree


Multiple Whole Genome Alignments
Gene Annotation
Orthology Inference
Gene Tree Estimation
Orthology-based Species Tree Estimation
Raw genome assemblies
Final Species Tree Output

Gene Tree Discordances

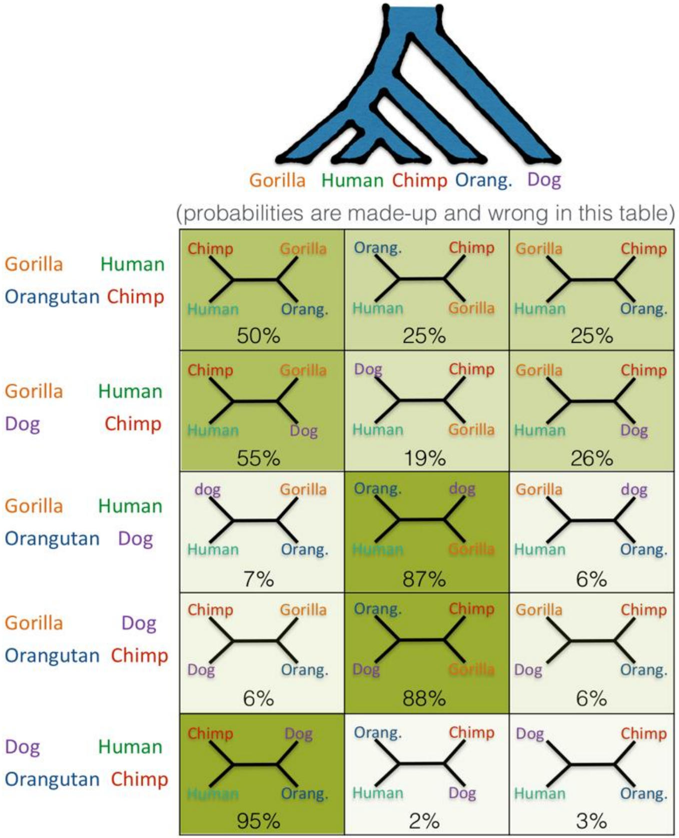
- Different parts of the genome can infer different phylogenies


gene 1
gene1000
Gorilla Human Chimp Orang.
Gorilla Chimp Human Orang.
The species tree
Gorilla Human Chimp Orangutan
A gene tree

Gene tree discordance aware tool - ASTRAL[1]

- Estimates an unrooted species tree given a set of unrooted gene trees.
- It is statistically consistent under the multi-species coalescent model

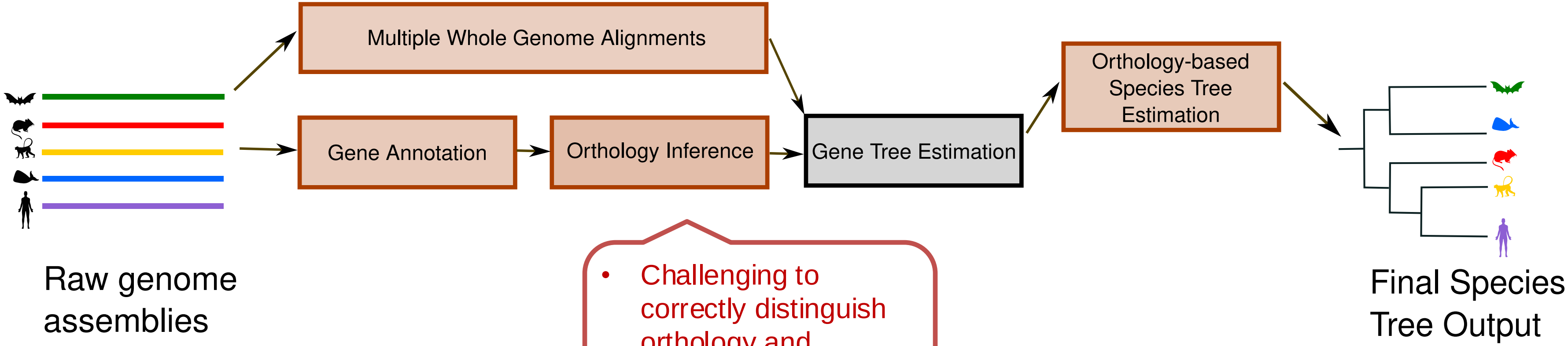
[1]Mirarab et al., ASTRAL: genome-scale coalescent-based species tree estimation, Bioinformatics (2014).


Gorilla Human Chimp Orang. Dog
Gorilla Human Orangutan Chimp
Chimp Gorilla Human Orang. 50%
Orang. Chimp Human Gorilla 25%
Gorilla Chimp Human Orang. 25%
Gorilla Human Dog Chimp
Chimp Gorilla Human Dog 55%
Dog Chimp Human Gorilla 19%
Gorilla Chimp Human Dog 26%
Gorilla Human Orangutan Dog
dog Gorilla Human Orang. 7%
Orang. dog Human Gorilla 87%
Gorilla dog Human Orang. 6%
Gorilla Dog Orangutan Chimp
Chimp Gorilla Dog Orang. 6%
Orang. Chimp Dog Gorilla 88%
Gorilla Chimp Dog Orang. 6%
Dog Human Orangutan Chimp
Chimp Dog Human Orang. 95%
Orang. Chimp Human Dog 2%
Dog Chimp Human Orang. 3%

(probabilities are made-up and wrong in this table)

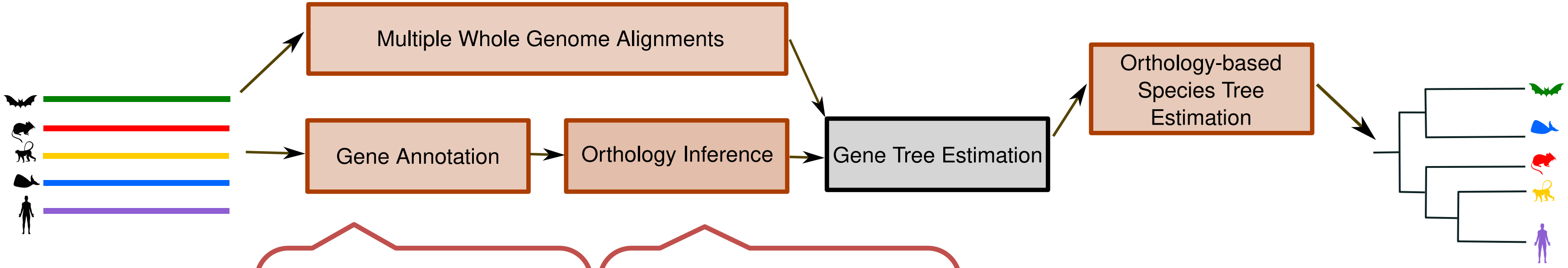
Image source: UCSD ECE 208 Lecture Slides

Existing approaches to estimate species tree


Multiple Whole Genome Alignments
Gene Annotation
Orthology Inference
Gene Tree Estimation
Orthology-based Species Tree Estimation
Raw genome assemblies
Final Species Tree Output

- Challenging to correctly distinguish orthology and paralogy

Existing approaches to estimate species tree


Multiple Whole Genome Alignments
Gene Annotation
Orthology Inference
Gene Tree Estimation
Orthology-based Species Tree Estimation

Raw genome assemblies

- Manual and error prone process
- Requires domain expertise

- Challenging to correctly distinguish orthology and paralogy

Final Species Tree Output

No automated yet accurate tool exists to infer phylogeny directly from raw genomes

- Separate tool exists for individual steps
- Accuracy depends on input tree/alignment -> error-prone
- Relies on domain expertise
- Takes months to complete

Objectives

To develop a

- Reference-free
- Orthology-free
- Alignment-free
- Discordance-aware

Approach for Estimation of Species tree

Objectives

To develop a

- **R**eference-free
- **O**rthology-free
- **A**lignment-free
- **DI**scordance-aware

Approach for **E**stimation of **S**pecies tree

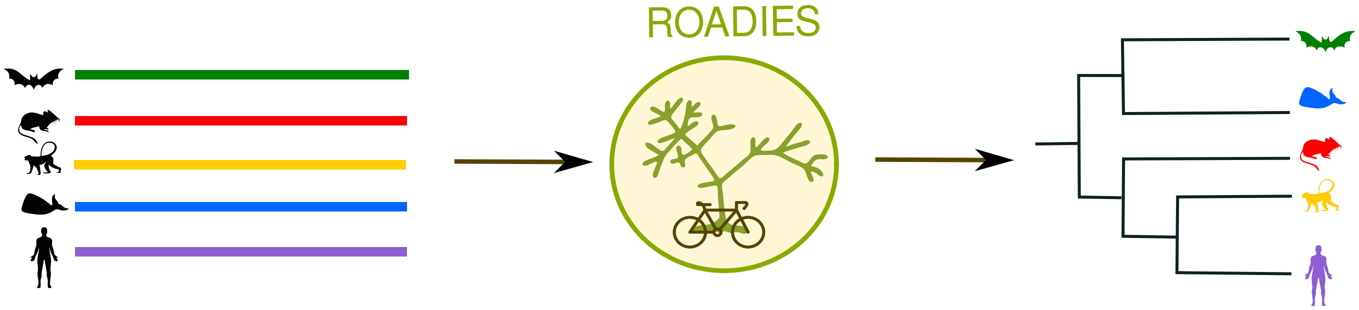
ROADIES – **R**eference-free **O**rthology-free **A**lignment-free **DI**scordance-aware **E**stimation of **S**pecies Tree

What is ROADIES?

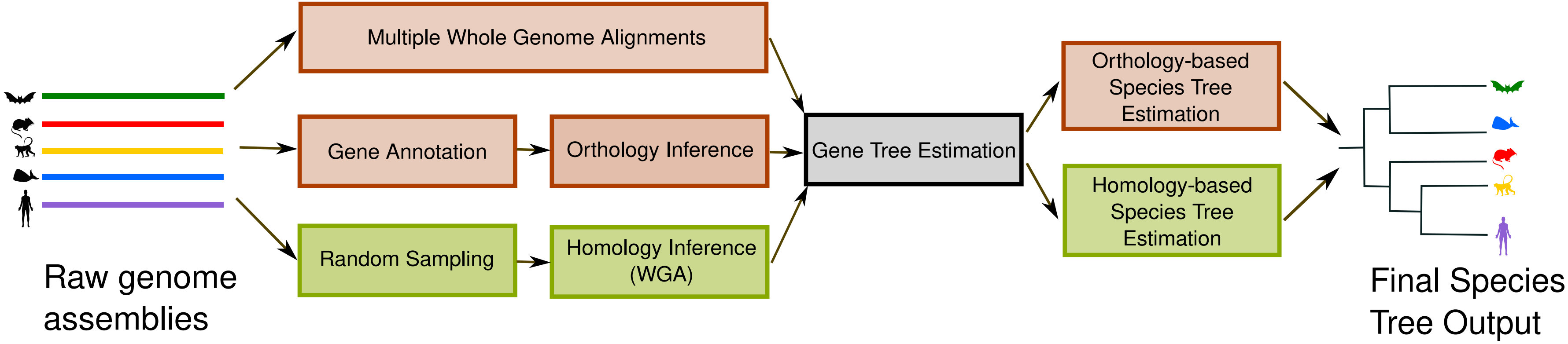
Automated tool which takes raw genomic assemblies as input and outputs species tree

Input Genomic Sequences

Final Species Tree Output



What ROADIES does?



Raw genome assemblies

Multiple Whole Genome Alignments

Gene Annotation

Orthology Inference

Random Sampling

Homology Inference (WGA)

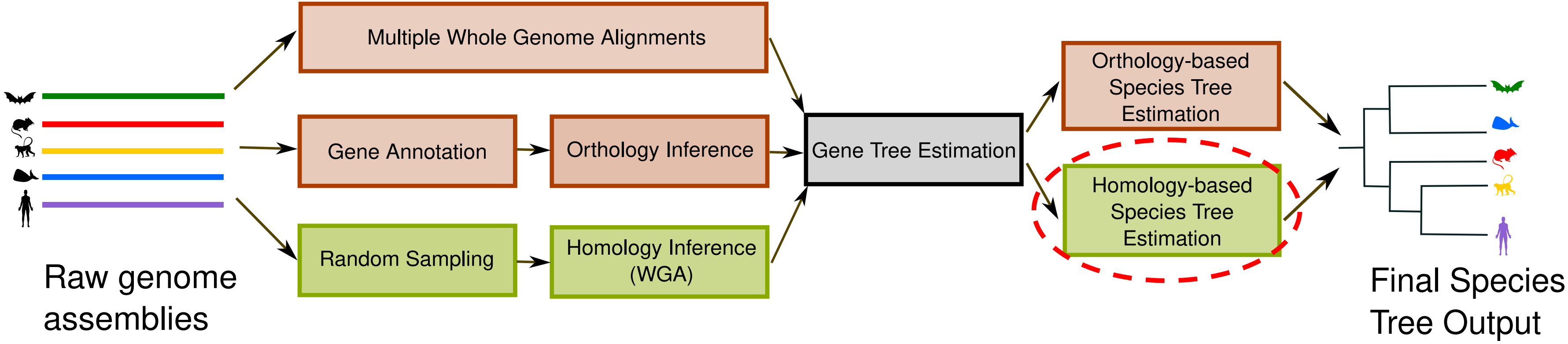
Gene Tree Estimation

Orthology-based Species Tree Estimation

Homology-based Species Tree Estimation

Final Species Tree Output

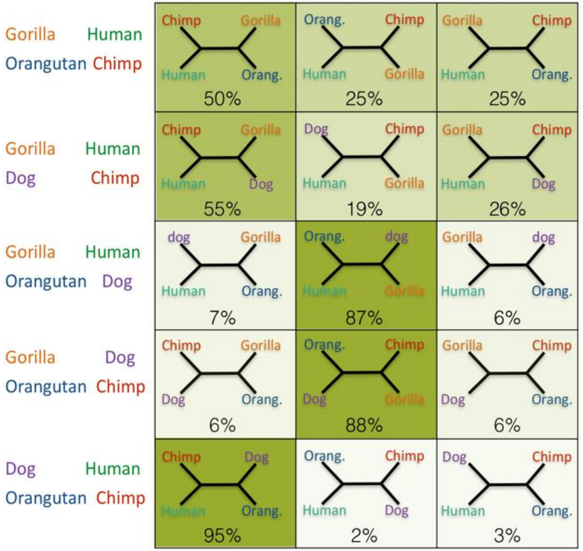
What ROADIES does?


Multiple Whole Genome Alignments
Gene Annotation
Orthology Inference
Random Sampling
Homology Inference (WGA)
Gene Tree Estimation
Orthology-based Species Tree Estimation
Homology-based Species Tree Estimation
Raw genome assemblies
Final Species Tree Output

Species Tree Estimation by ASTRAL-Pro[1]

- ASTRAL Pro → ASTRAL for PaRalogs and Orthologs
- Statistically consistent discordance aware tool
- Finds the best tree with maximum dominant quartets
- Does not require separation of orthologs and paralogs

(probabilities are made-up and wrong in this table)

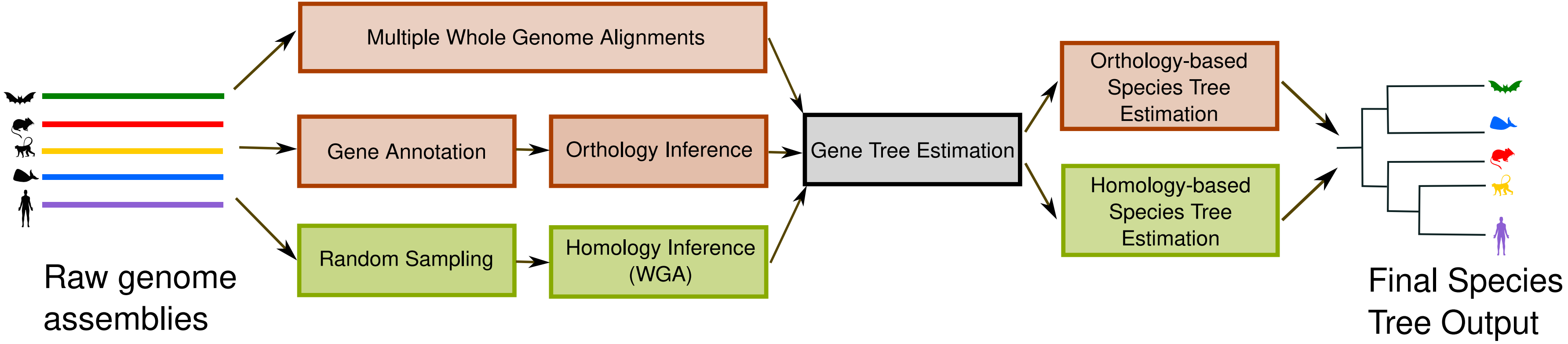


Gorilla Human Orangutan Chimp	Chimp Gorilla Human Orang. 50%	Orang. Chimp Human Gorilla 25%	Gorilla Chimp Human Orang. 25%
Gorilla Human Dog Chimp	Chimp Gorilla Human Dog 55%	Dog Chimp Human Gorilla 19%	Gorilla Chimp Human Dog 26%
Gorilla Human Orangutan Dog	dog Gorilla Human Orang. 7%	Orang. dog Human Gorilla 87%	Gorilla dog Human Orang. 6%
Gorilla Dog Orangutan Chimp	Chimp Gorilla Dog Orang. 6%	Orang. Chimp Dog Gorilla 88%	Gorilla Chimp Dog Orang. 6%
Dog Human Orangutan Chimp	Chimp Dog Human Orang. 95%	Orang. Chimp Human Dog 2%	Dog Chimp Human Orang. 3%

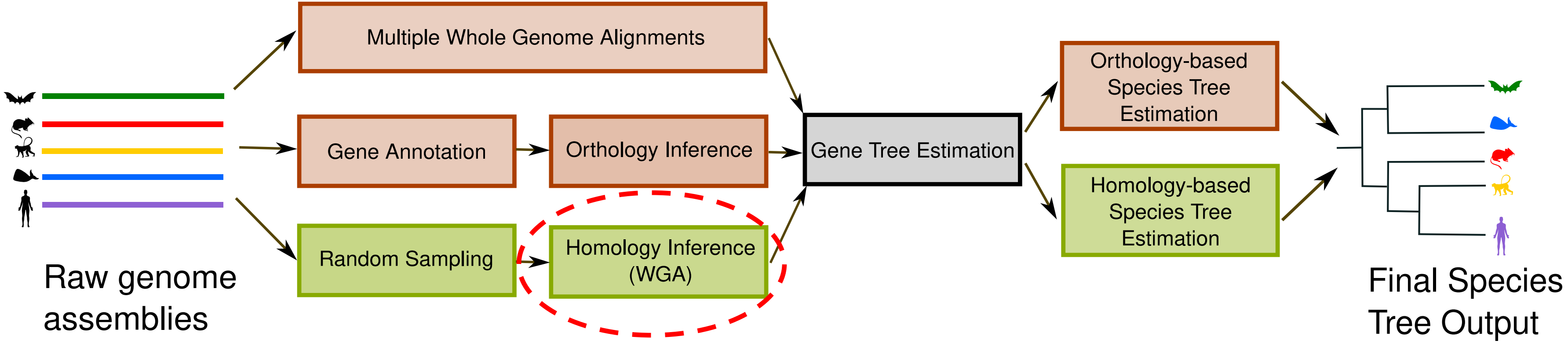
[1] Zhang et al. "ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy", MBE 2020

Image source: UCSD ECE 208 Lecture Slides

What ROADIES does?


Multiple Whole Genome Alignments
Gene Annotation
Orthology Inference
Random Sampling
Homology Inference (WGA)
Gene Tree Estimation
Orthology-based Species Tree Estimation
Homology-based Species Tree Estimation
Raw genome assemblies
Final Species Tree Output

What ROADIES does?



Multiple Whole Genome Alignments

Gene Annotation

Orthology Inference

Random Sampling

Homology Inference (WGA)

Gene Tree Estimation

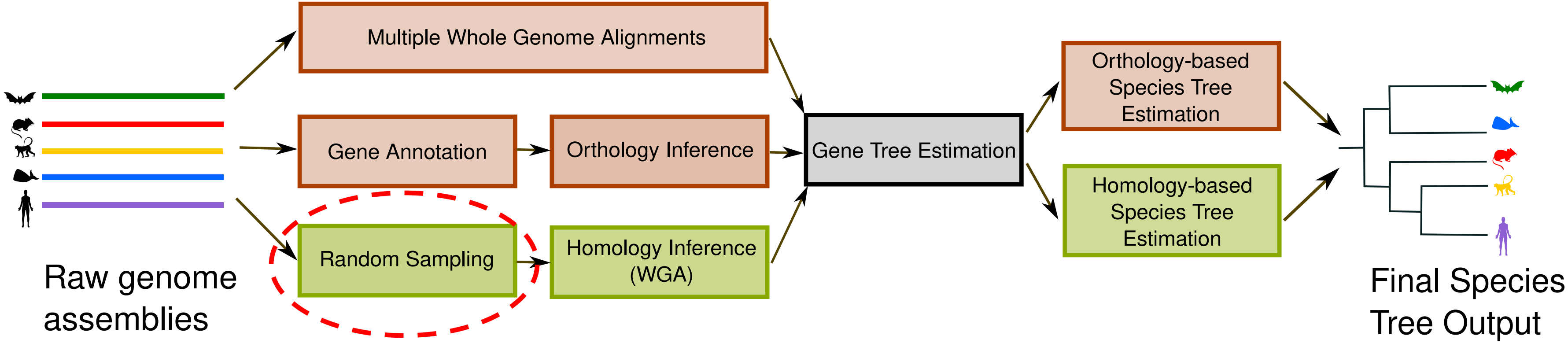
Orthology-based Species Tree Estimation

Homology-based Species Tree Estimation

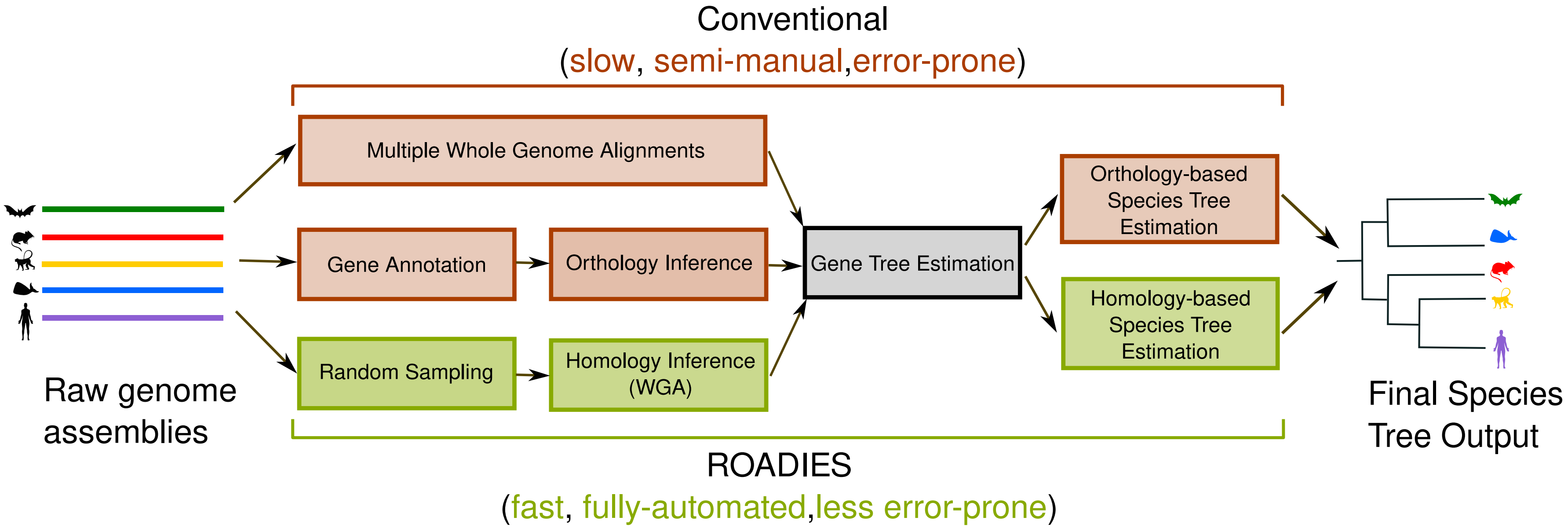
Raw genome assemblies

Final Species Tree Output

What ROADIES does?


Multiple Whole Genome Alignments
Gene Annotation
Orthology Inference
Random Sampling
Homology Inference (WGA)
Gene Tree Estimation
Orthology-based Species Tree Estimation
Homology-based Species Tree Estimation
Raw genome assemblies
Final Species Tree Output

What ROADIES does?


Conventional
(slow, semi-manual,error-prone)
Multiple Whole Genome Alignments
Gene Annotation
Orthology Inference
Random Sampling
Homology Inference (WGA)
Gene Tree Estimation
Orthology-based Species Tree Estimation
Homology-based Species Tree Estimation
Raw genome assemblies
Final Species Tree Output
ROADIES
(fast, fully-automated,less error-prone)

ROADIES pipeline

Random Sampling of genes from input genomic assemblies

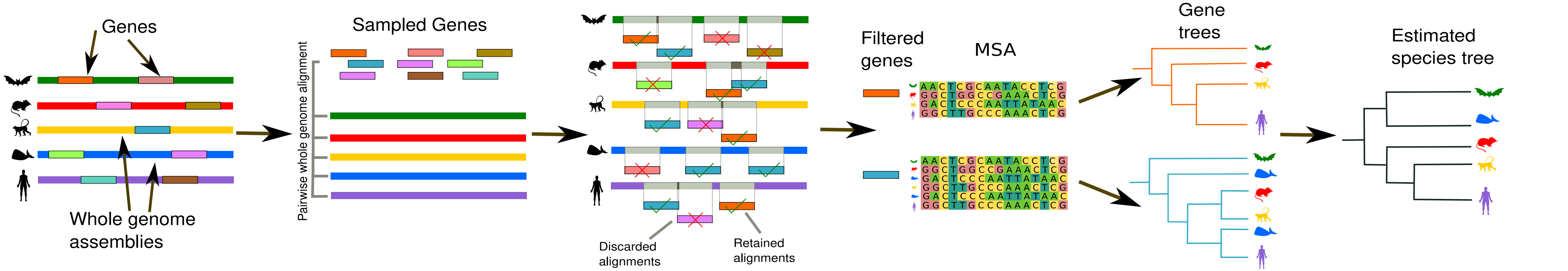
Pairwise alignment of sampled genes to other genomes

Filter low quality pairwise alignments and repeats

Perform Multiple sequence alignment of filtered genes across all species

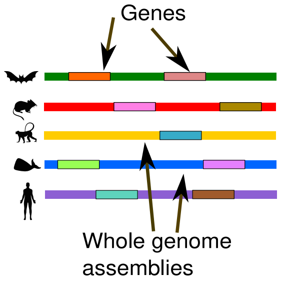
Estimate genes trees from MSAs

Estimate species tree from lists of gene trees



ROADIES pipeline

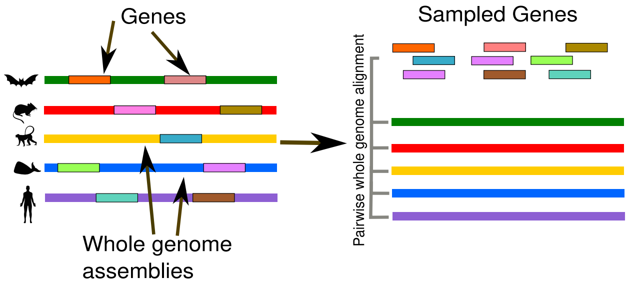
Random Sampling of genes from input genomic assemblies



ROADIES pipeline

Random Sampling of genes from input genomic assemblies

Pairwise alignment of sampled genes to other genomes

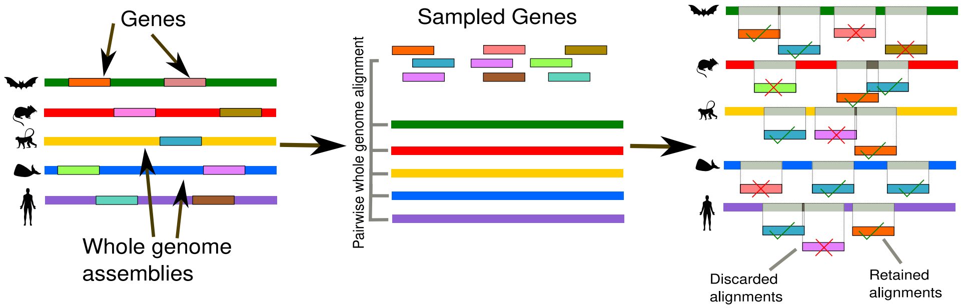


ROADIES pipeline

Random Sampling of genes from input genomic assemblies

Pairwise alignment of sampled genes to other genomes

Filter low quality pairwise alignments and repeats



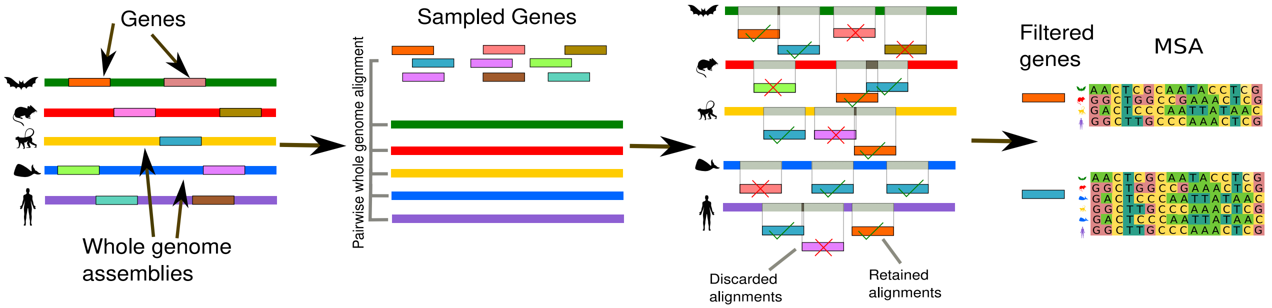
ROADIES pipeline

Random Sampling of genes from input genomic assemblies

Pairwise alignment of sampled genes to other genomes

Filter low quality pairwise alignments and repeats

Perform Multiple sequence alignment of filtered genes across all species



UC San Diego
JACOBS SCHOOL OF ENGINEERING

ROADIES pipeline

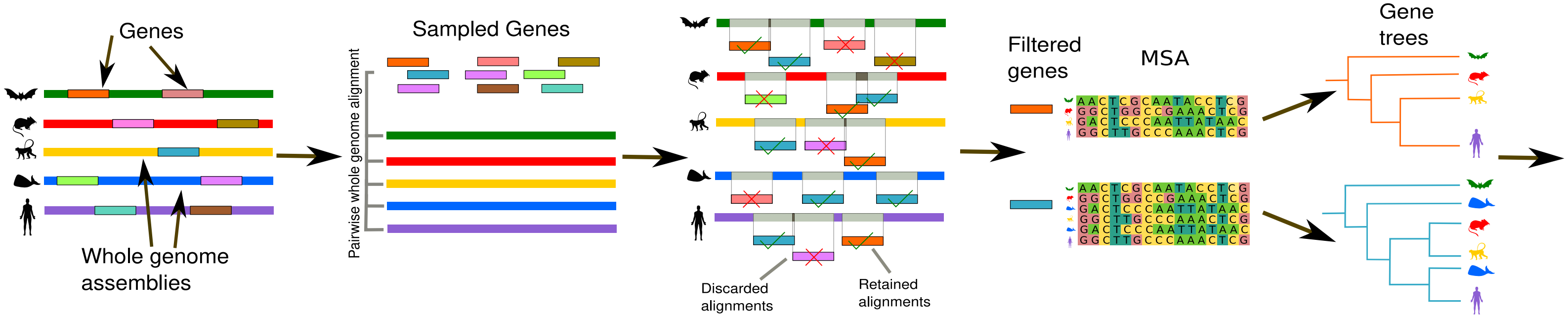
Random Sampling of genes from input genomic assemblies

Pairwise alignment of sampled genes to other genomes

Filter low quality pairwise alignments and repeats

Perform Multiple sequence alignment of filtered genes across all species

Estimate genes trees from MSAs



ROADIES pipeline

Random Sampling of genes from input genomic assemblies

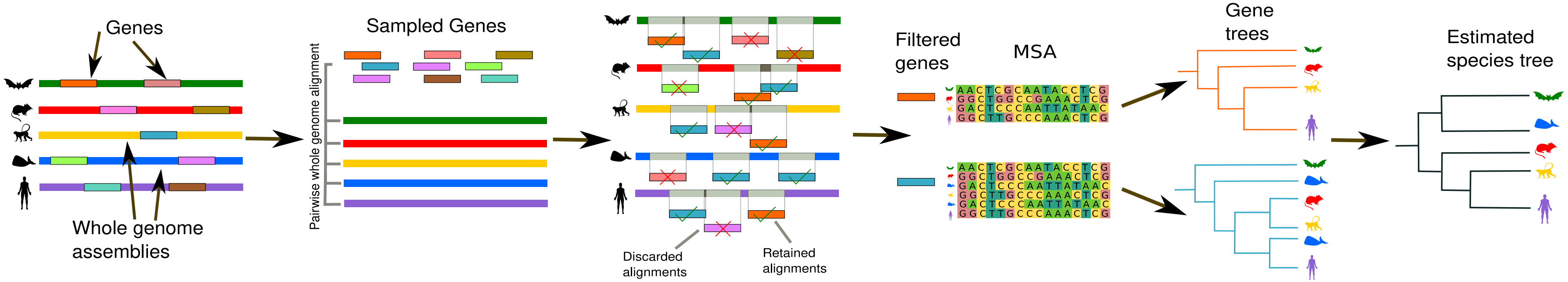
Pairwise alignment of sampled genes to other genomes

Filter low quality pairwise alignments and repeats

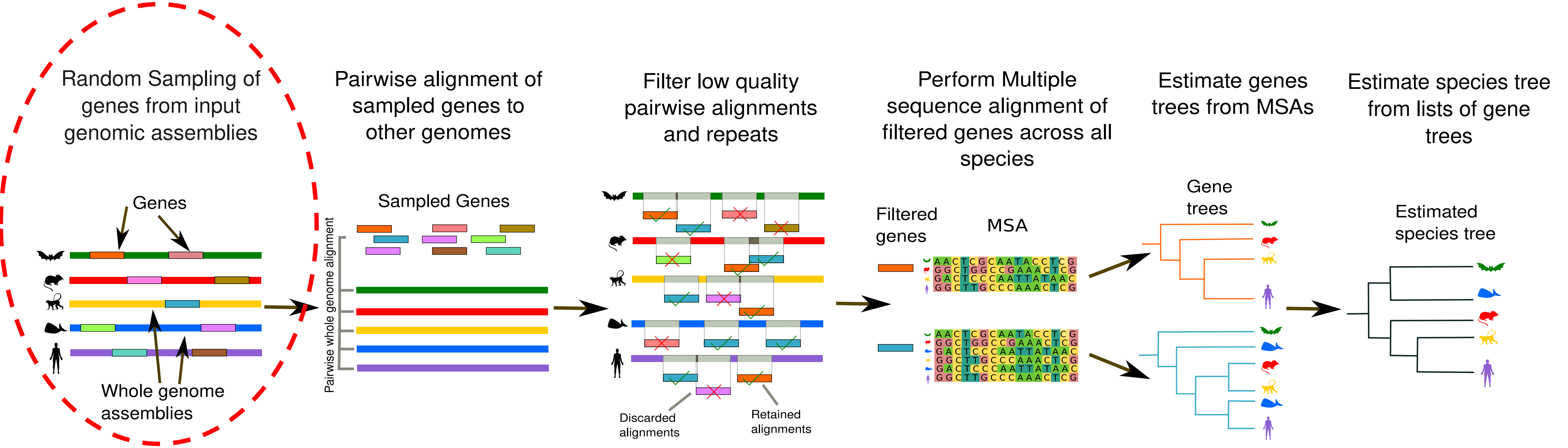
Perform Multiple sequence alignment of filtered genes across all species

Estimate genes trees from MSAs

Estimate species tree from lists of gene trees



Problem: How many genes to start with?



Random Sampling of genes from input genomic assemblies

Pairwise alignment of sampled genes to other genomes

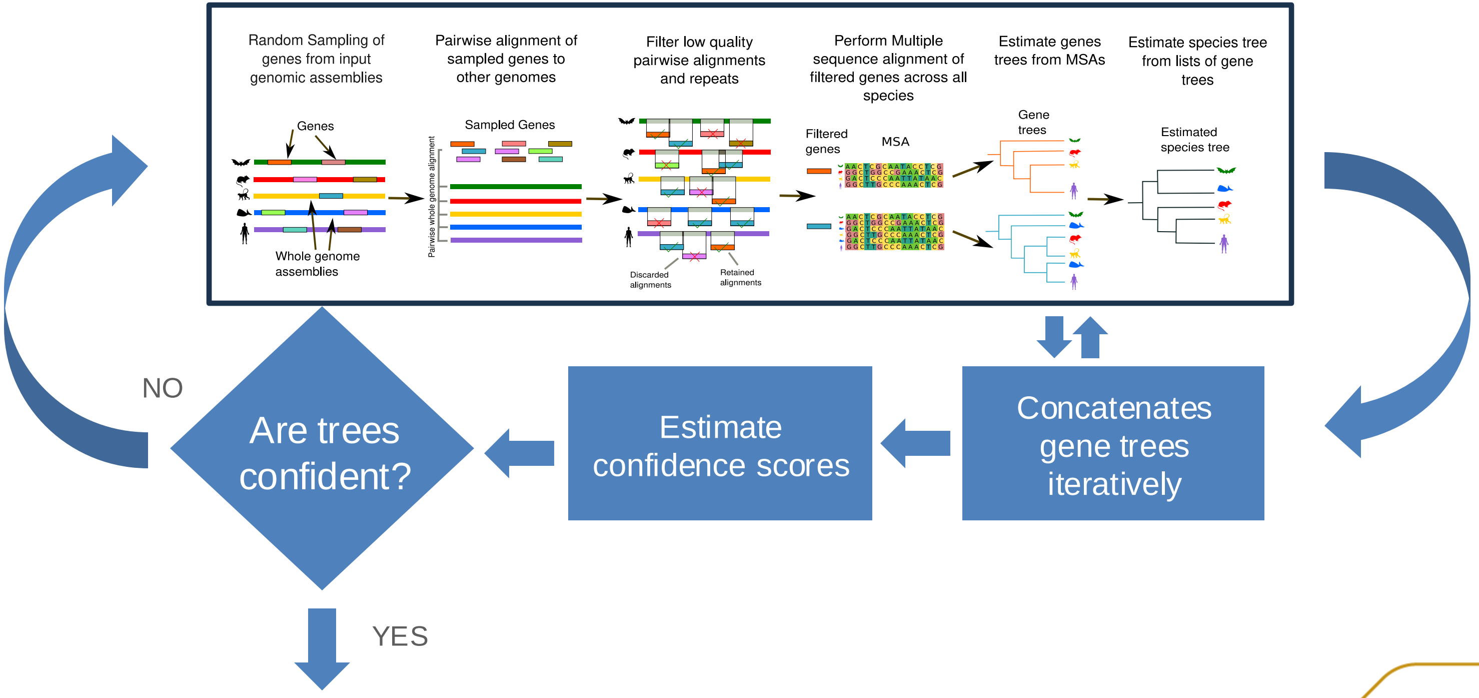
Filter low quality pairwise alignments and repeats

Perform Multiple sequence alignment of filtered genes across all species

Estimate genes trees from MSAs

Estimate species tree from lists of gene trees

ROADIES converges into accurate tree with more gene trees



Random Sampling of genes from input genomic assemblies

Pairwise alignment of sampled genes to other genomes

Filter low quality pairwise alignments and repeats

Perform Multiple sequence alignment of filtered genes across all species

Estimate genes trees from MSAs

Estimate species tree from lists of gene trees

Concatenates gene trees iteratively

Estimate confidence scores

Are trees confident?

NO

YES

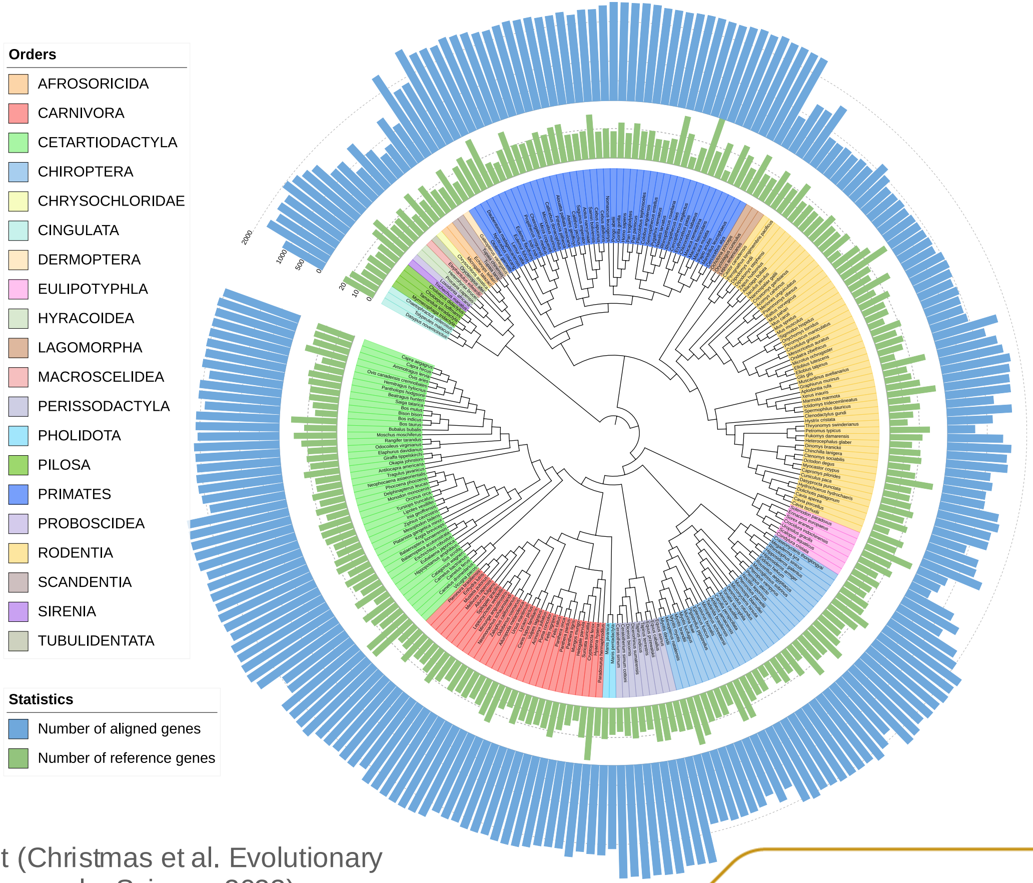
Final converged tree

Methodology

- **Datasets** – 240 Mammals[1], 100 Flies[2] and 363 Birds[3]
- **Normalized Robinson-Foulds distance**
  - Species-level
  - Orders/Group-level
- **Tree confidence metric**
  - Local posterior probability
  - Quartet scores

1 – Zoonomia 2020, 2- Kim et al. 2021, 3 – Birds 10k Genome Project (Feng et al. 2020), Image source: Wikipedia, Researchers Sequence Genomes of 363 Bird Species - SciNews

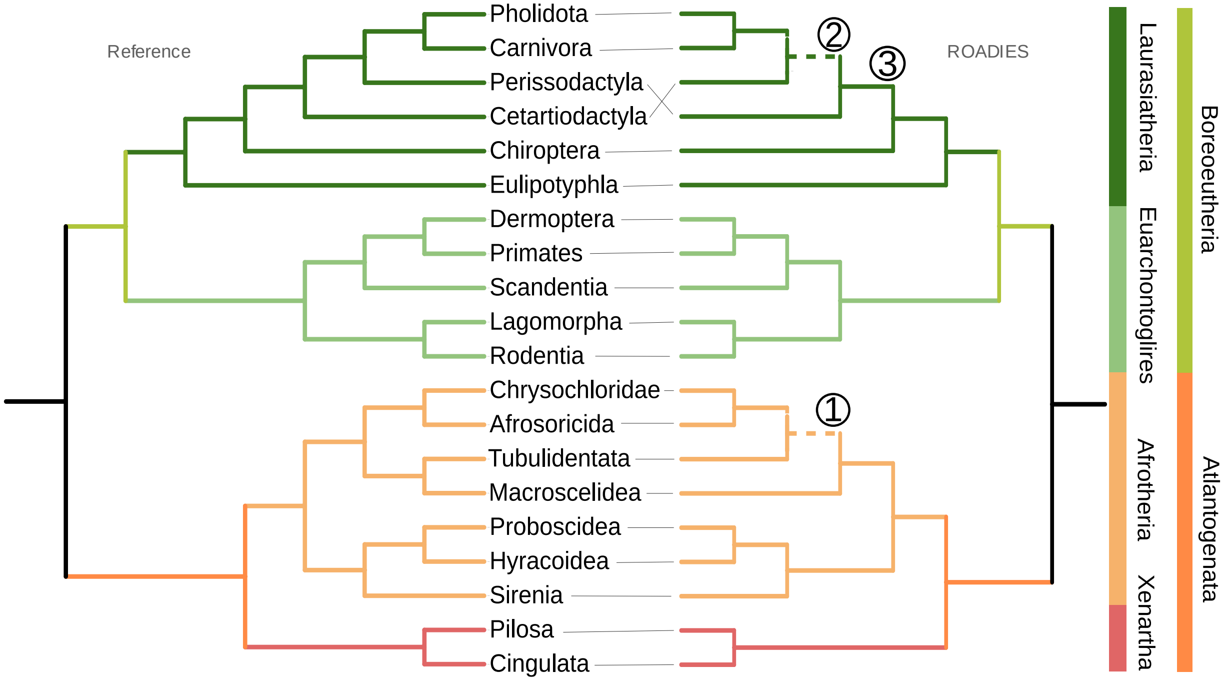
ROADIES estimates accurate phylogeny of 240 placental mammals



Species-level normalized RF distance with reference tree = 0.037

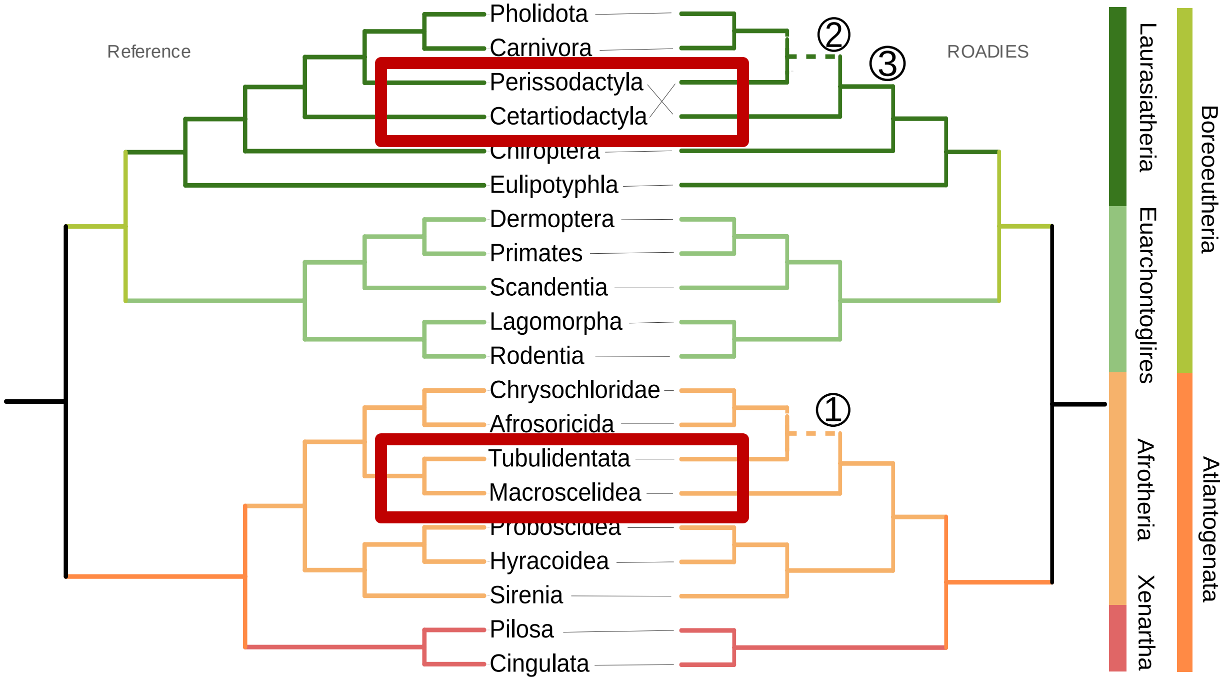
Reference tree and datasets taken from Zoonomia project (Christmas et al. Evolutionary constraint and innovation across hundreds of placental mammals, Science 2023).

ROADIES estimates accurate phylogeny of 240 placental mammals at order-level



Order-level normalized RF distance with reference tree = 0.11

ROADIES estimates accurate phylogeny of 240 placental mammals at order-level



Order-level normalized RF distance with reference tree = 0.11

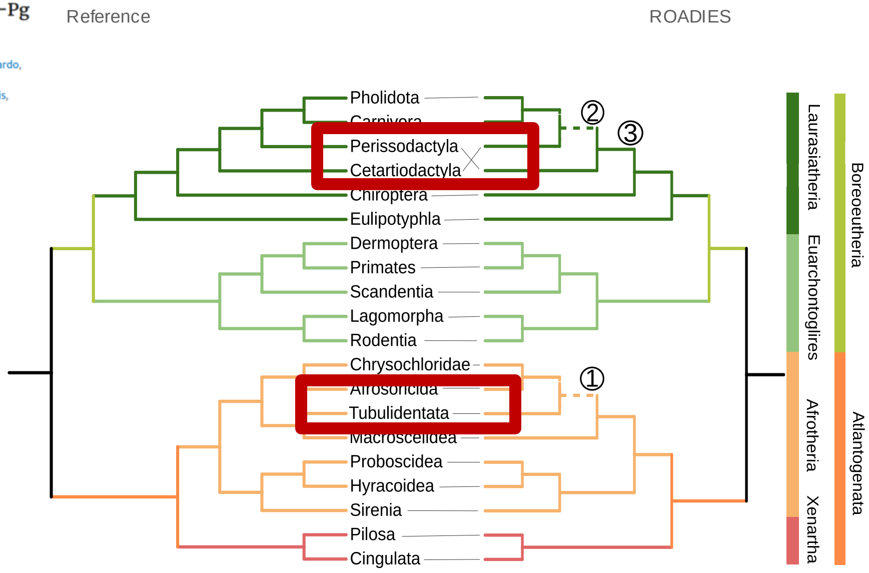
ROADIES estimates accurate phylogeny of 240 placental mammals at order-level

The placental mammal ancestor and the post-K-Pg radiation of placentals

Maureen A O'Leary [1], Jonathan I Bloch, John J Flynn, Timothy J Gaudin, Andres Giallombardo, Norberto P Giannini, Suzann L Goldberg, Brian P Kraatz, Zhe-Xi Luo, Jin Meng, Xijun Ni, Michael J Novacek, Fernando A Perini, Zachary S Randall, Guillermo W Rougier, Eric J Sargis, Mary T Silcox, Nancy B Simmons, Michelle Spaulding, Paúl M Velazco, Marcelo Weksler, John R Wible, Andrea L Cirranello

Affiliations + expand

PMID: 23393258 DOI: 10.1126/science.1229237



Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation

Nathan S. Upham, Jacob A. Esselstyn, Walter Jetz

Published: December 4, 2019 • https://doi.org/10.1371/journal.pbio.3000494

Research article | Open access | Published: 06 November 2019

The effect of alignment uncertainty, substitution models and priors in building and dating the mammal tree of life

Yan Du, Shaoyuan Wu, Scott V. Edwards & Liang Liu

BMC Evolutionary Biology 19, Article number: 203 (2019) | Cite this article

RESEARCH ARTICLE | BIOLOGICAL SCIENCES |

Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions

Hidenori Nishihara, Masami Hasegawa, and Norihiro Okada Authors Info & Affiliations

June 27, 2006 | 103 (26) 9929-9934 | https://doi.org/10.1073/pnas.0603797103

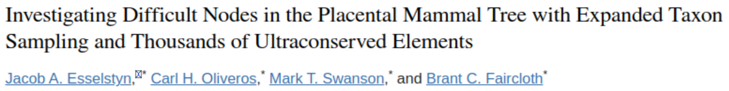
JOURNAL ARTICLE

Phylogenomic Resolution of the Phylogeny of Laurasiatherian Mammals: Exploring Phylogenetic Signals within Coding and Noncoding Sequences

Meng-Yun Chen, Dan Liang, Peng Zhang Author Notes

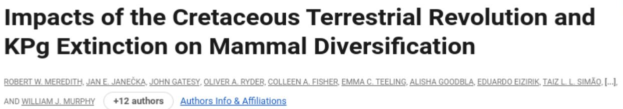
Genome Biology and Evolution, Volume 9, Issue 8, August 2017, Pages 1998–2012, https://doi.org/10.1093/gbe/evx147

Published: 02 August 2017 **Article history ▾**



Investigating Difficult Nodes in the Placental Mammal Tree with Expanded Taxon Sampling and Thousands of Ultraconserved Elements

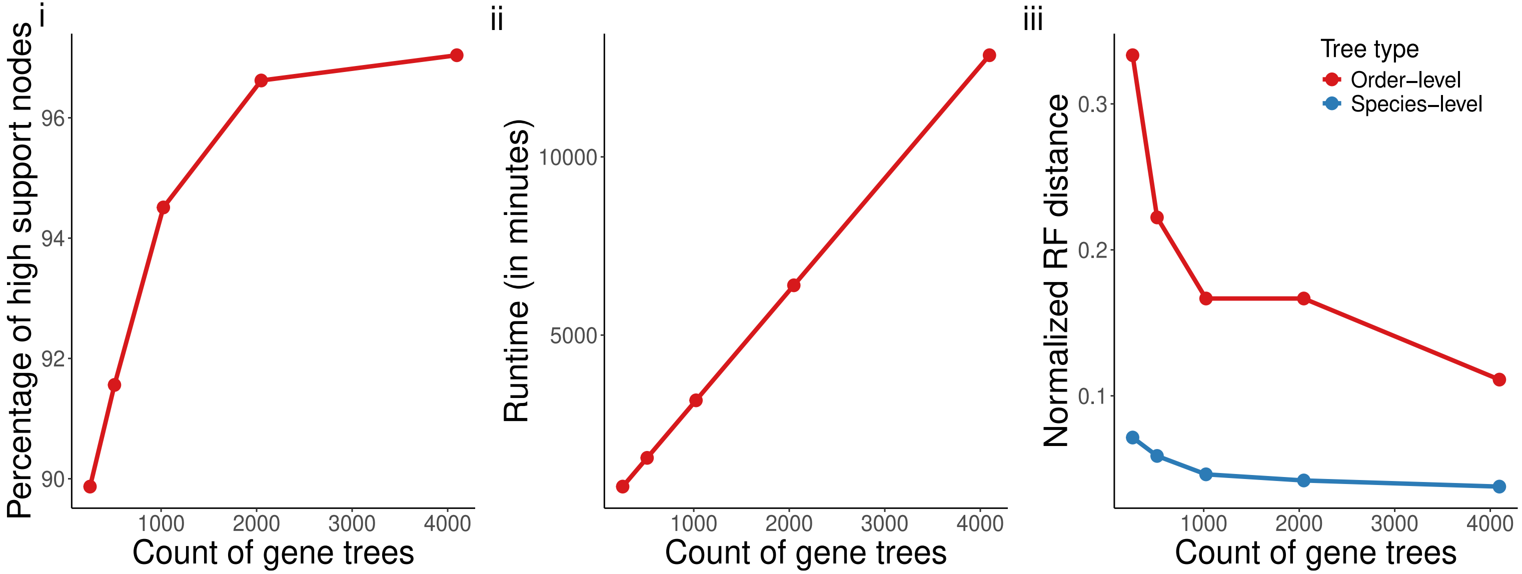
Jacob A. Esselstyn,* Carl H. Oliveros,* Mark T. Swanson,* and Brant C. Faircloth*



Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification

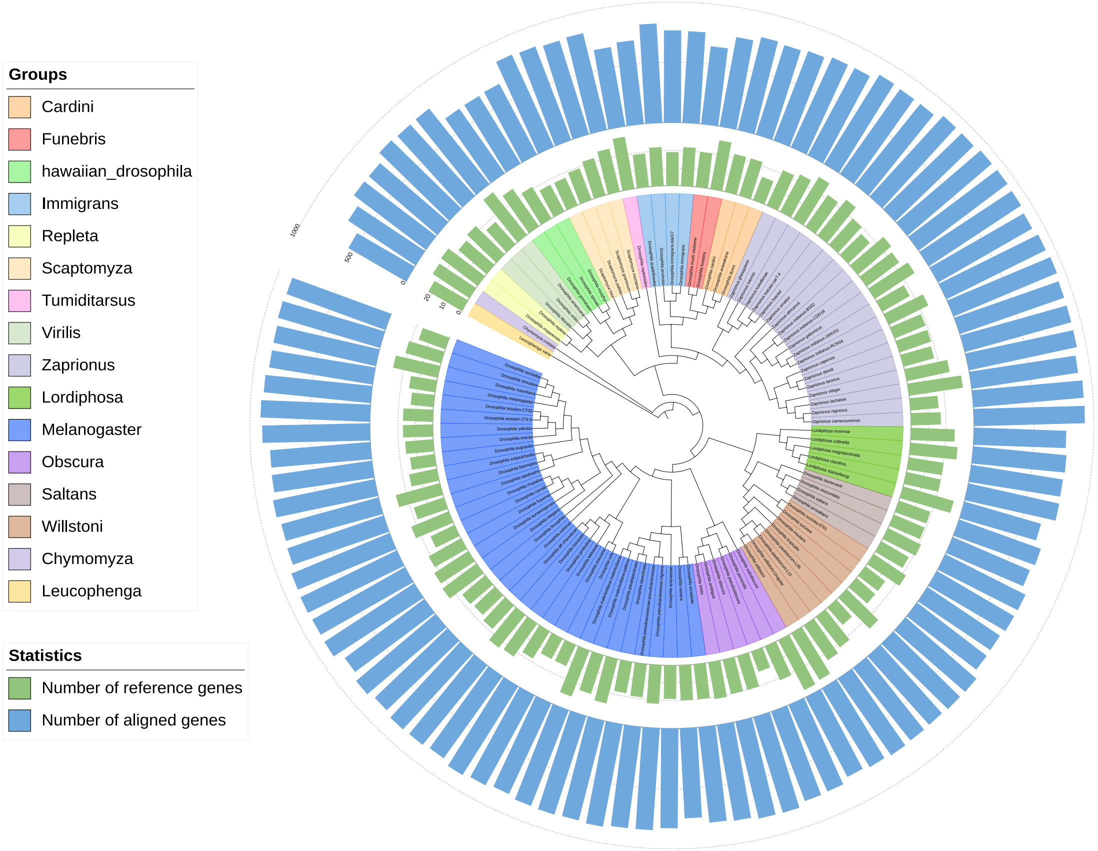
ROBERT W. MEREDITH, JAN E. JANEČKA, JOHN GATESY, OLIVER A. RYDER, COLLEEN A. FISHER, EMMA C. TEELING, ALISHA GOODBLA, EDUARDO EIZIRIK, TAIZ L. L. SIMÃO, [...], AND WILLIAM J. MURPHY +12 authors Authors Info & Affiliations

ROADIES converges with more gene trees



Experiments run on AWS R6a 16-core instances
Runtime is calculated as wall clock time
High support nodes: Nodes with localPP >= 0.95

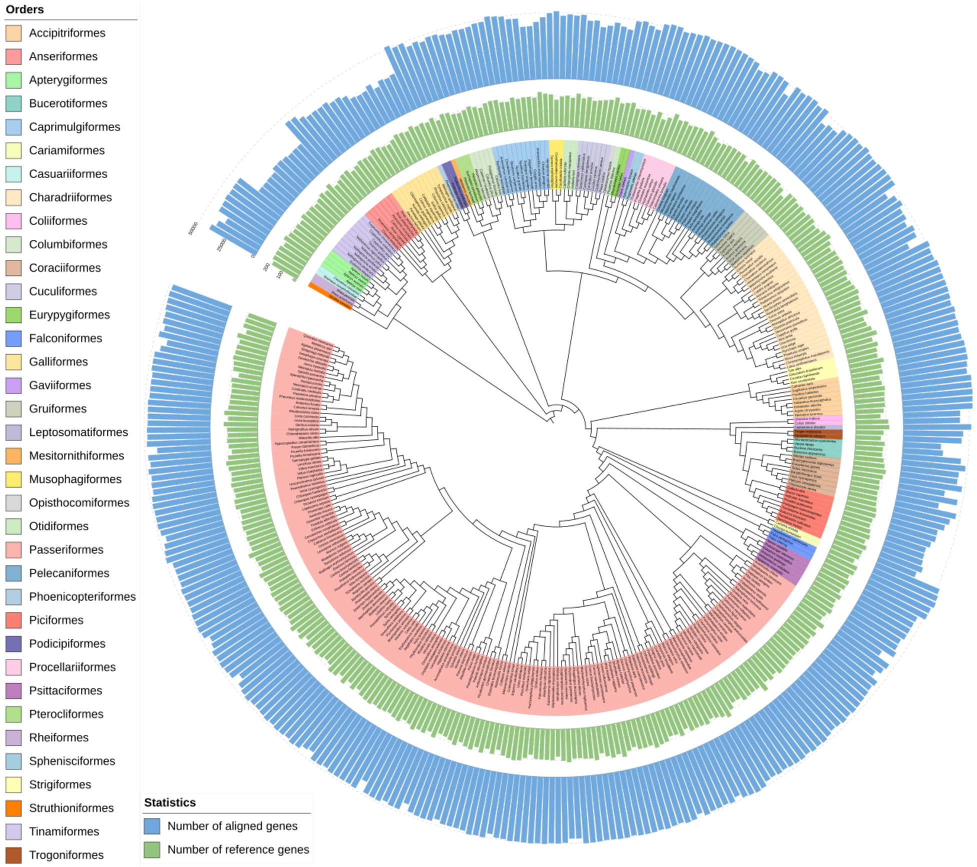
ROADIES estimates accurate phylogeny of 100 Drosophilid genomes (fruit flies)



Species-level normalized RF distance with reference tree = 0.062

Reference tree and datasets taken from the paper: Kim, B. Y. et al. Highly contiguous assemblies of 101 drosophilid genomes. ELife, 2021.

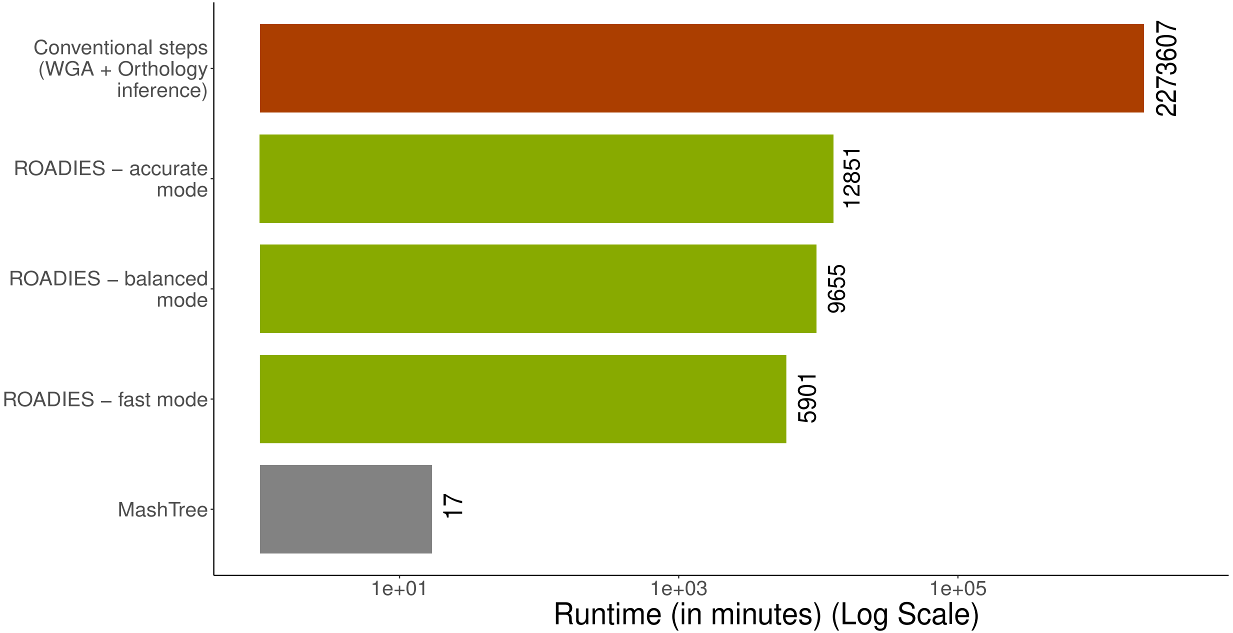
ROADIES estimates accurate phylogeny of 363 avian species



Species-level normalized RF distance with reference tree = 0.037

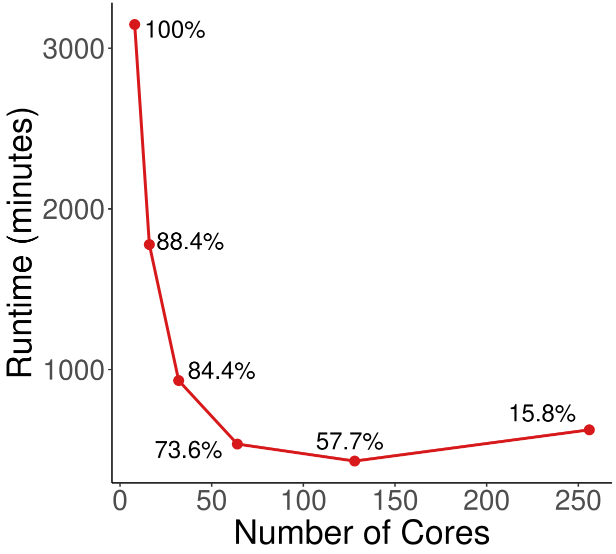
Reference tree and datasets taken from Stiller, J. et al. Complexity of avian evolution revealed by family-level genomes. Nature (2024) doi:10.1038/s41586-024-07323-1

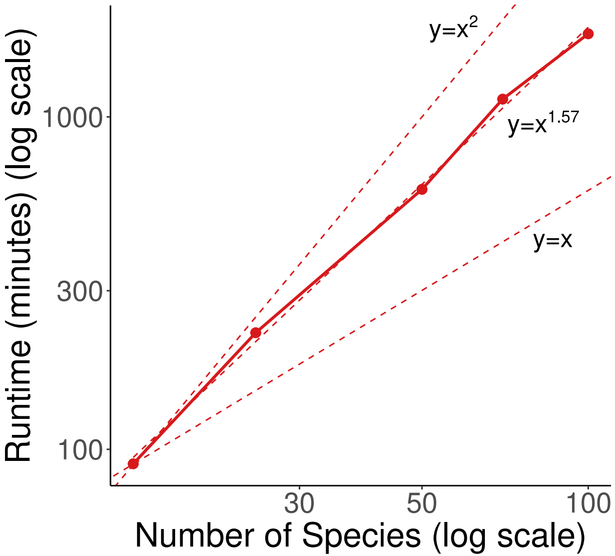
ROADIES is 176x faster than conventional approaches



Experiments run on AWS R6a 16-core instances
Runtime is calculated as wall clock time
Experiments are performed with mammals datasets

ROADIES scales well with increasing system cores and species count





ROADIES scales between linear and quadratic with more species count

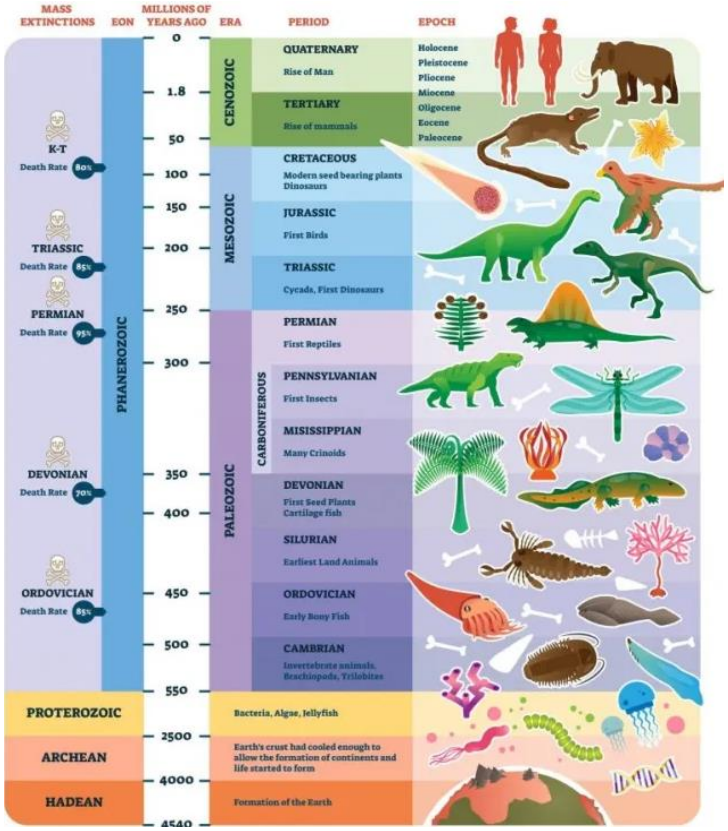
Experiments run on AWS R6a instances
Runtime is calculated as wall clock time

ROADIES aims to support diverse evolutionary timescales

- Mammals ~ 100 million years
- Flies ~ 40 million years
- Birds ~ 150 million years

- Fish ~ 500 million years
- Bacteria ~ 3 billion years
- SARS Cov2 ~ 3-4 years

Image Source: geologyscience.com/geology-branches/paleontology/geologic-time-scale/



Summary

- ROADIES is first-of-its-kind tool which automates species tree inference directly from raw genome assemblies
- Highly configurable and scalable
- 176x faster than conventional methods (for mammals dataset)
- Accurate results for mammals, flies and, birds dataset

Acknowledgments

Thank you for your attention.

Questions?

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

Advisors:

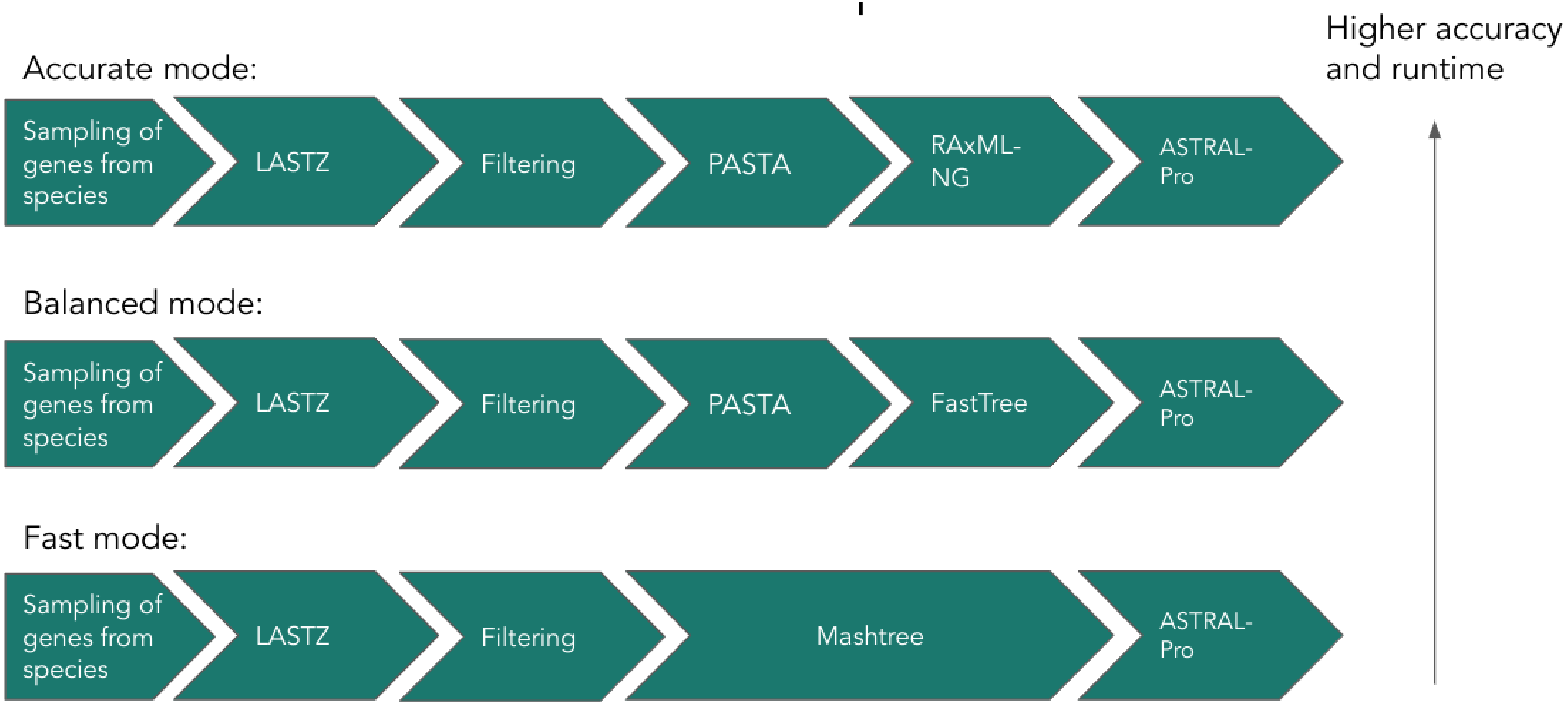
- Yatish Turakhia (UCSD)
- Siavash Mirarab (UCSD)

Collaborators:

- Tian (Kevin) Liu (UCSD)
- Hiram Clawson (UCSC)
- Guojie Zhang (Zhejiang University)
- Yulong Xie (Zhejiang University)
- Benedict Paten (UCSC)

Additional Slides

Modes of operation



Accurate mode:

Sampling of genes from species → LASTZ → Filtering → PASTA → RAxML-NG → ASTRAL-Pro

Balanced mode:

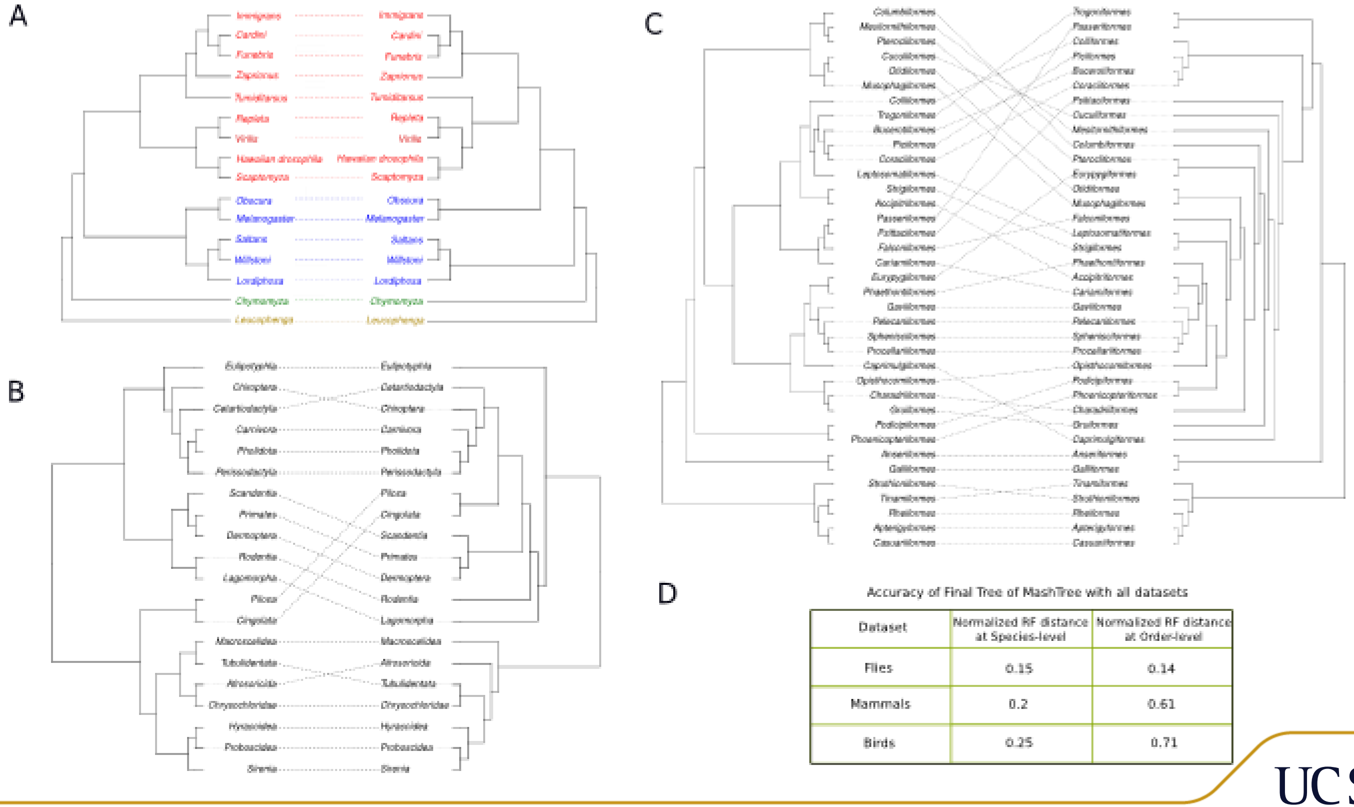
Sampling of genes from species → LASTZ → Filtering → PASTA → FastTree → ASTRAL-Pro

Fast mode:

Sampling of genes from species → LASTZ → Filtering → Mashtree → ASTRAL-Pro

Higher accuracy and runtime

ROADIES is more accurate than MashTree



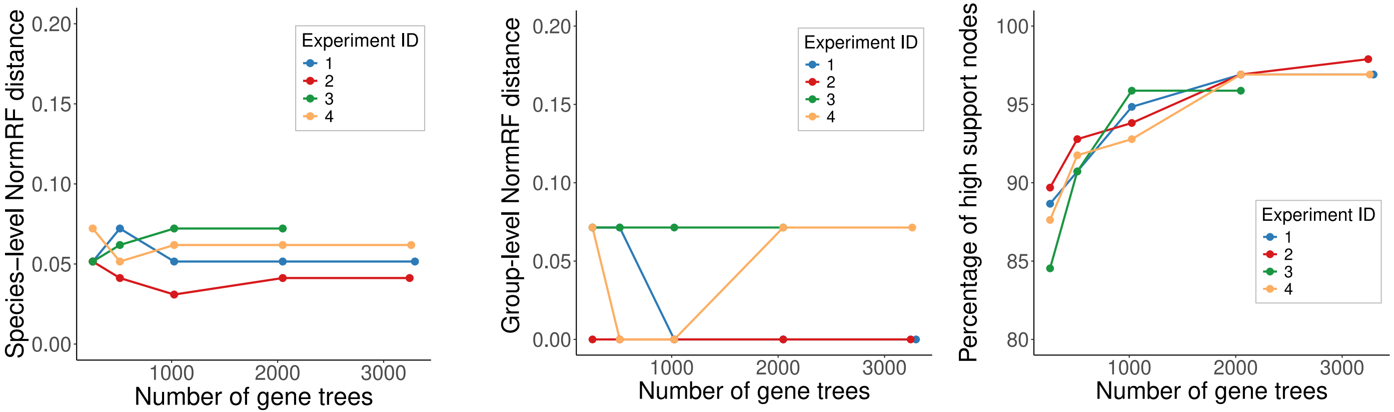
Accuracy of Final Tree of MashTree with all datasets

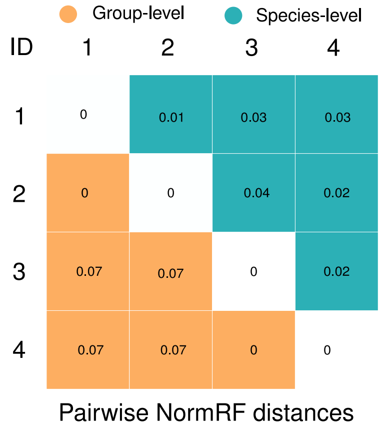
Dataset	Normalized RF distance at Species-level	Normalized RF distance at Order-level
Flies	0.15	0.14
Mammals	0.2	0.61
Birds	0.25	0.71

UC San Diego
JACOBS SCHOOL OF ENGINEERING

ROADIES gives stable results

Even if ROADIES randomly samples genes, results are consistent





● Group-level ● Species-level

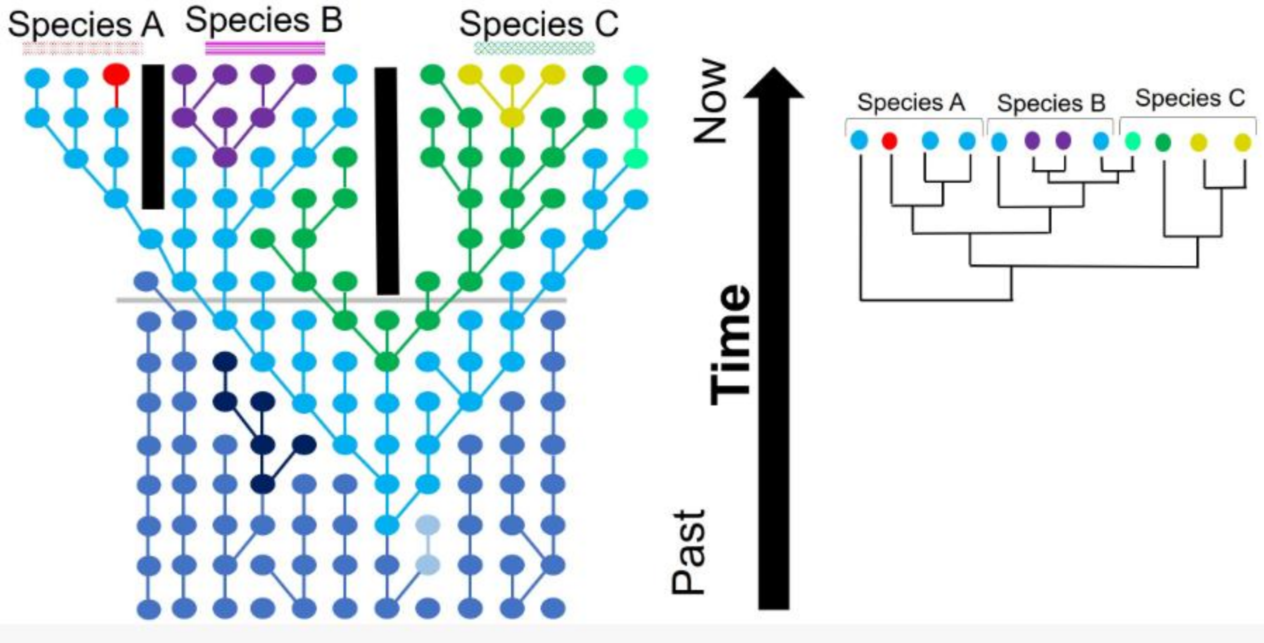
ID	1	2	3	4
1	0	0.01	0.03	0.03
2	0	0	0.04	0.02
3	0.07	0.07	0	0.02
4	0.07	0.07	0	0

Pairwise NormRF distances

Experiments run on AWS R6a instances
Runtime is calculated as wall clock time
Variance experiments are tested with Drosophila datasets

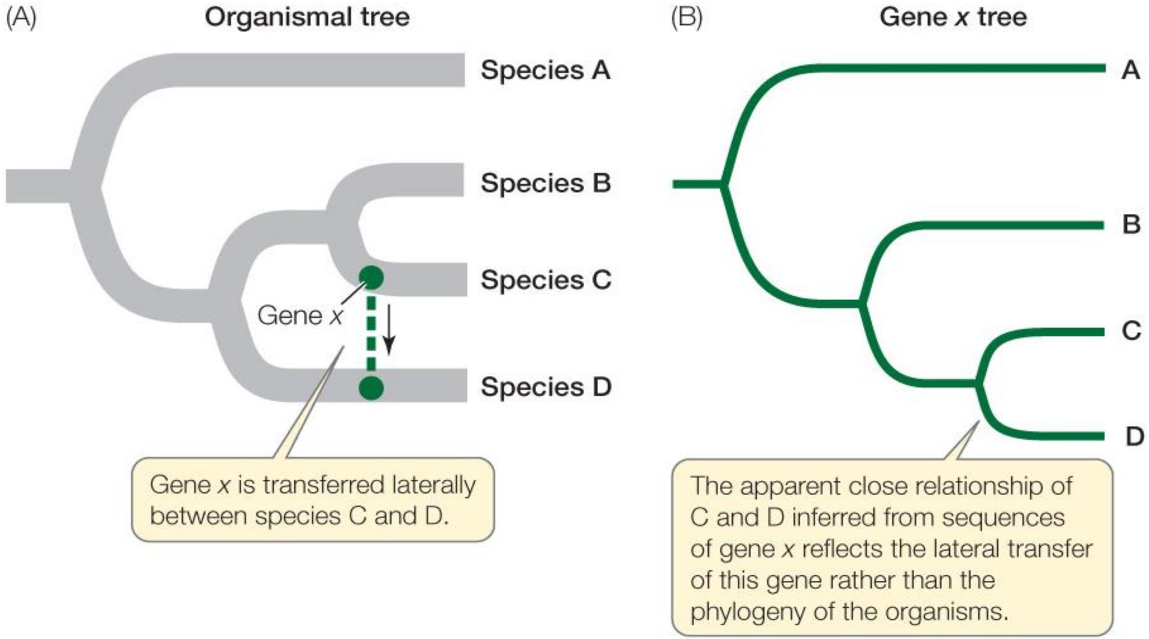
Causes of Gene Tree Discordance

Incomplete Lineage Sorting



Causes of Gene Tree Discordance

Horizontal Gene Transfer

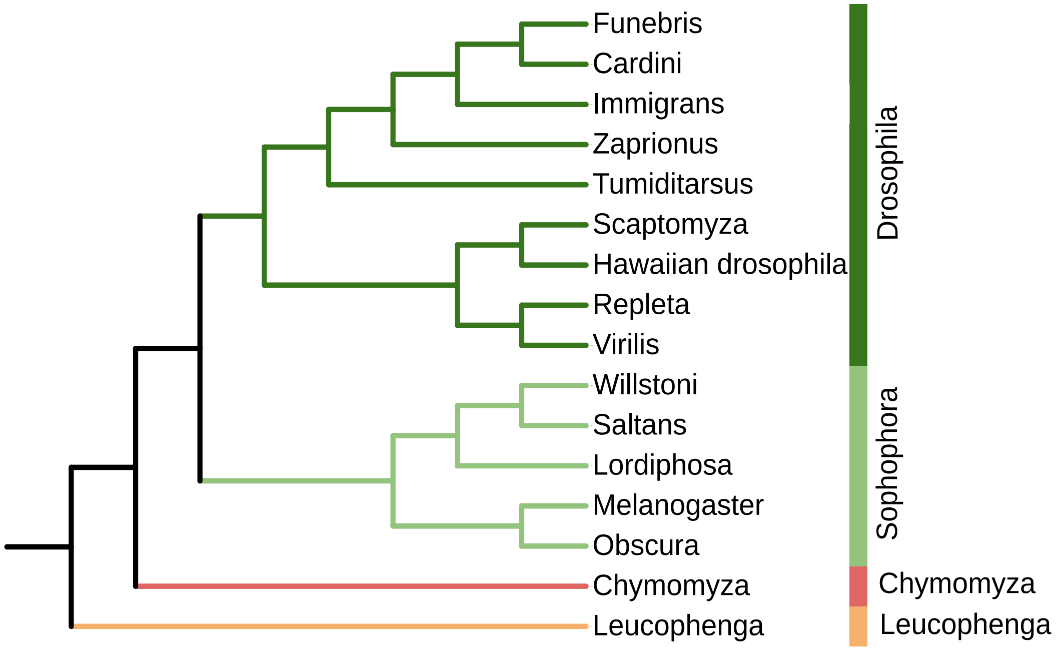


Gene *x* is transferred laterally between species C and D.

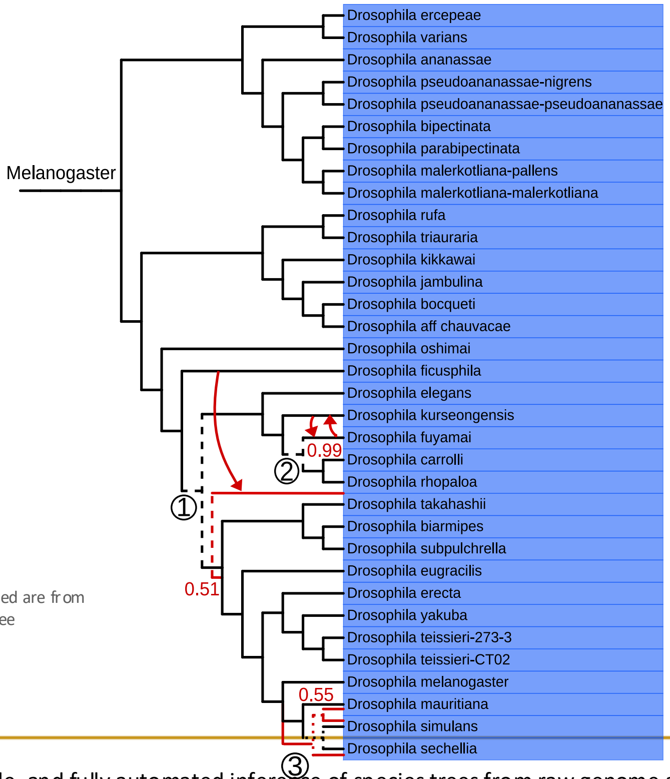
The apparent close relationship of C and D inferred from sequences of gene *x* reflects the lateral transfer of this gene rather than the phylogeny of the organisms.

ROADIES estimates exact same phylogeny of 100 Drosophilid genomes at group-level

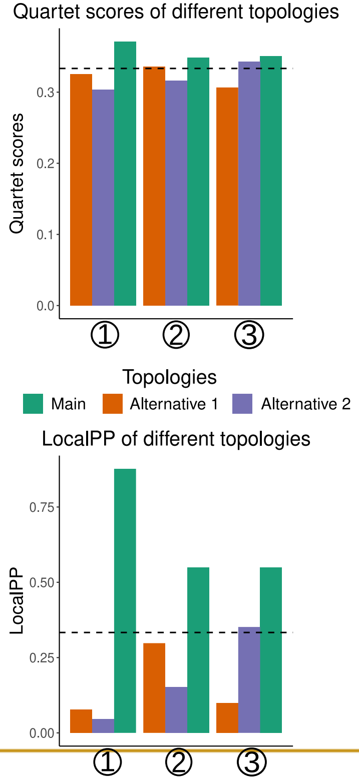
Group-level normalized RF distance with reference tree = 0



ROADIES' differences with reference is limited to low-confident branches



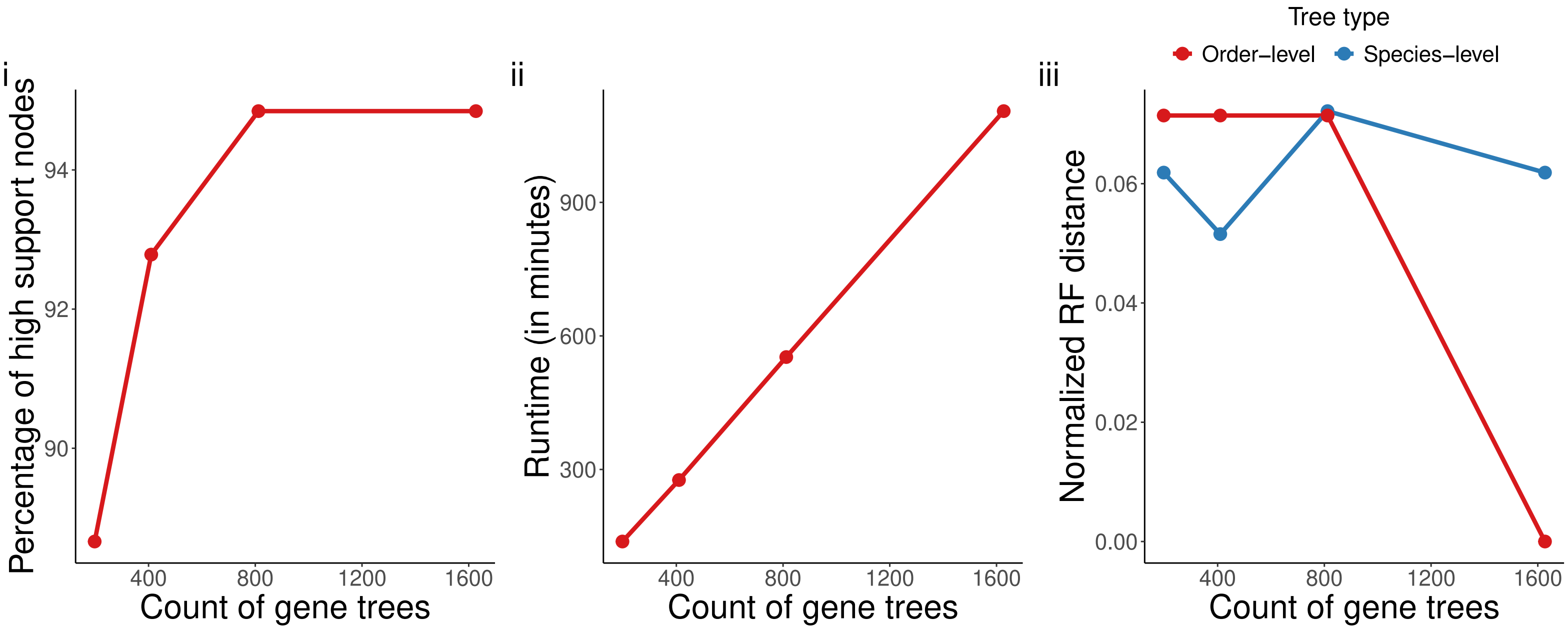
Nodes highlighted in red are from reference tree


Quartet scores of different topologies

Topologies

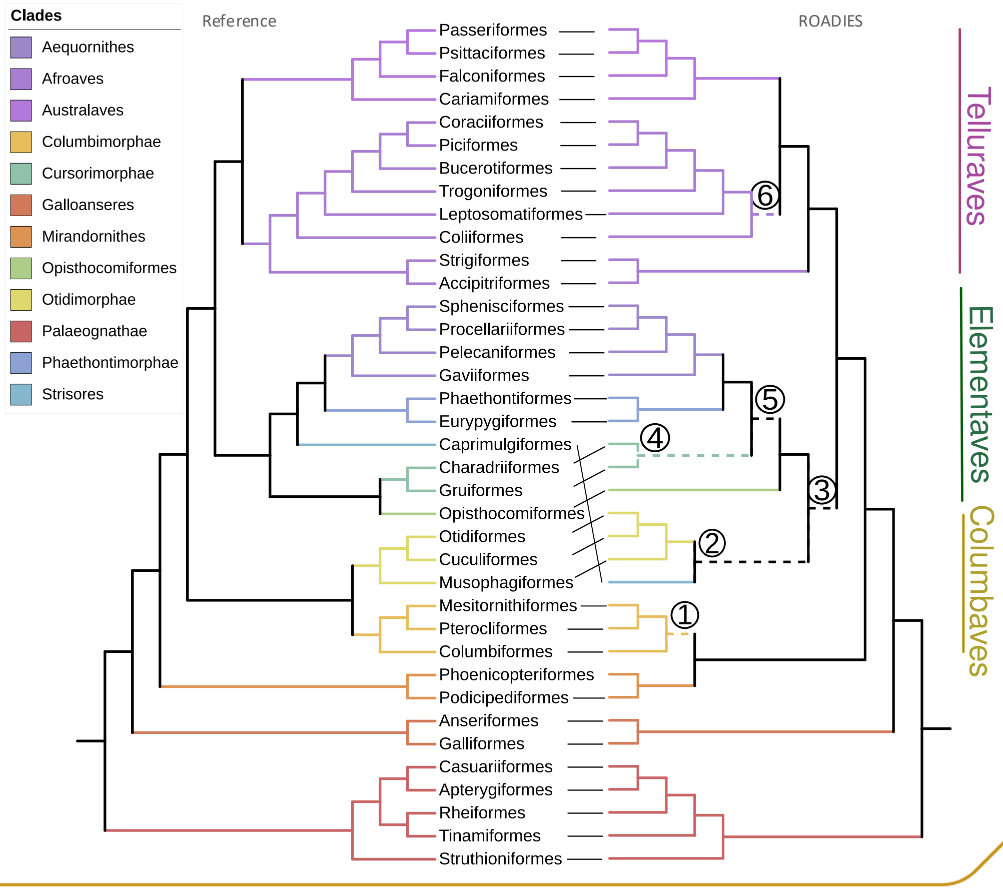
LocalPP of different topologies

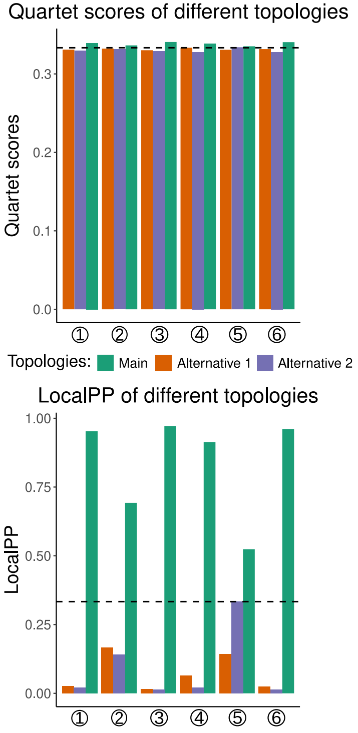
ROADIES convergence results



Experiments run on AWS R6a 16-core instances
Runtime is calculated as wall clock time

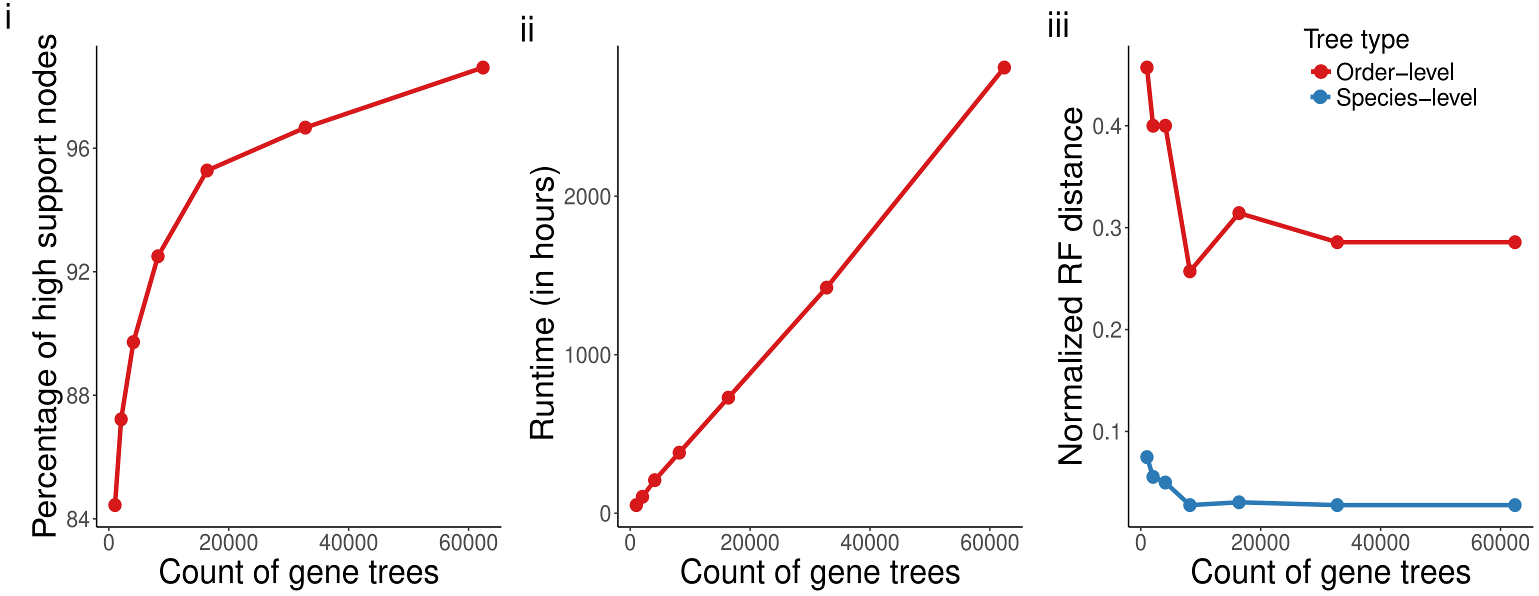
Order-level phylogeny





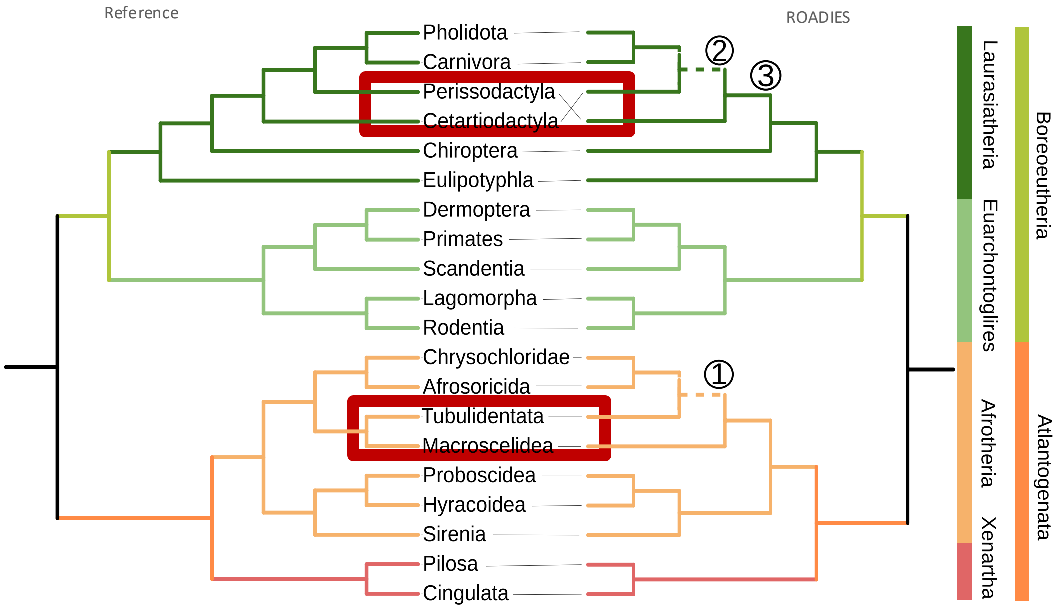
Order-level normalized RF distance with reference tree = 0.28

ROADIES convergence results

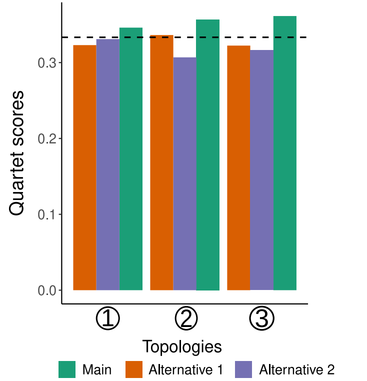


Experiments run on AWS R6a 16-core instances
Runtime is calculated as wall clock time

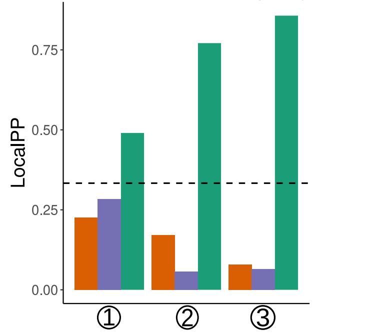
Low-confident branches are debatable



Quartet scores of different topologies



LocalPP of different topologies


LocalPP
①
②
③