Inference of the tree of life with applications to plants

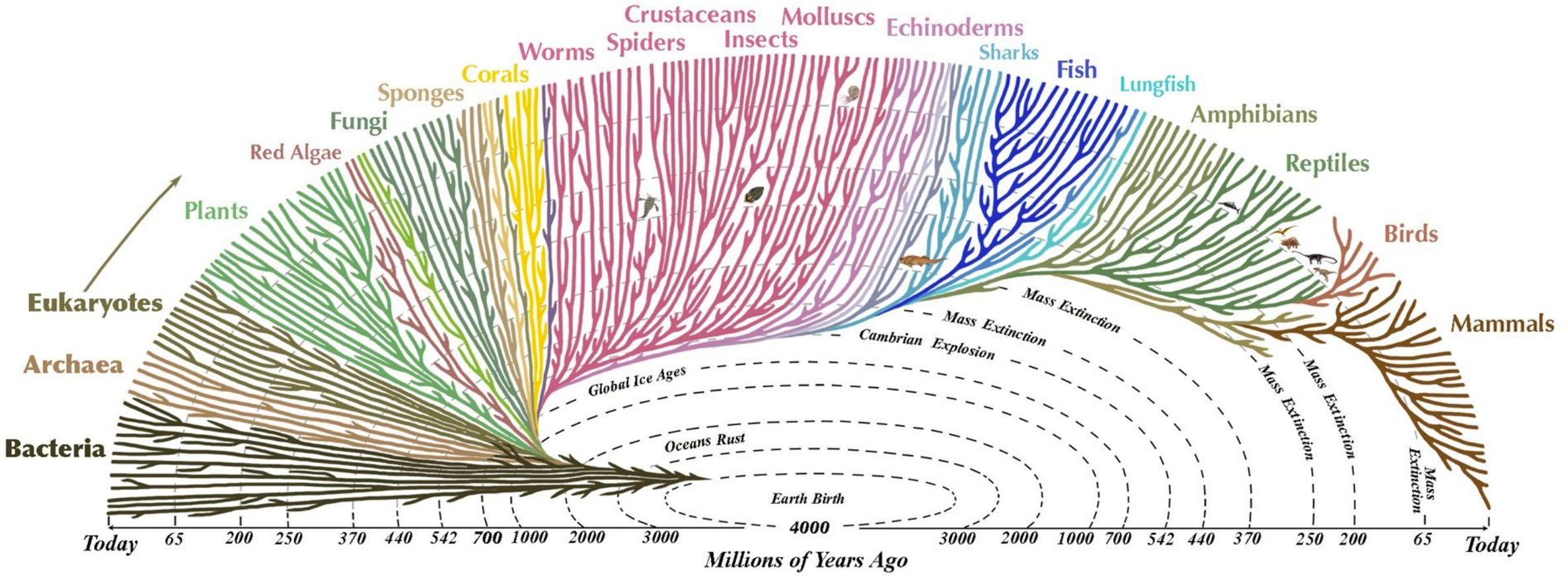
Siavash Mirarab

Electrical and Computer Engineering

University of California, San Diego

Tree of life

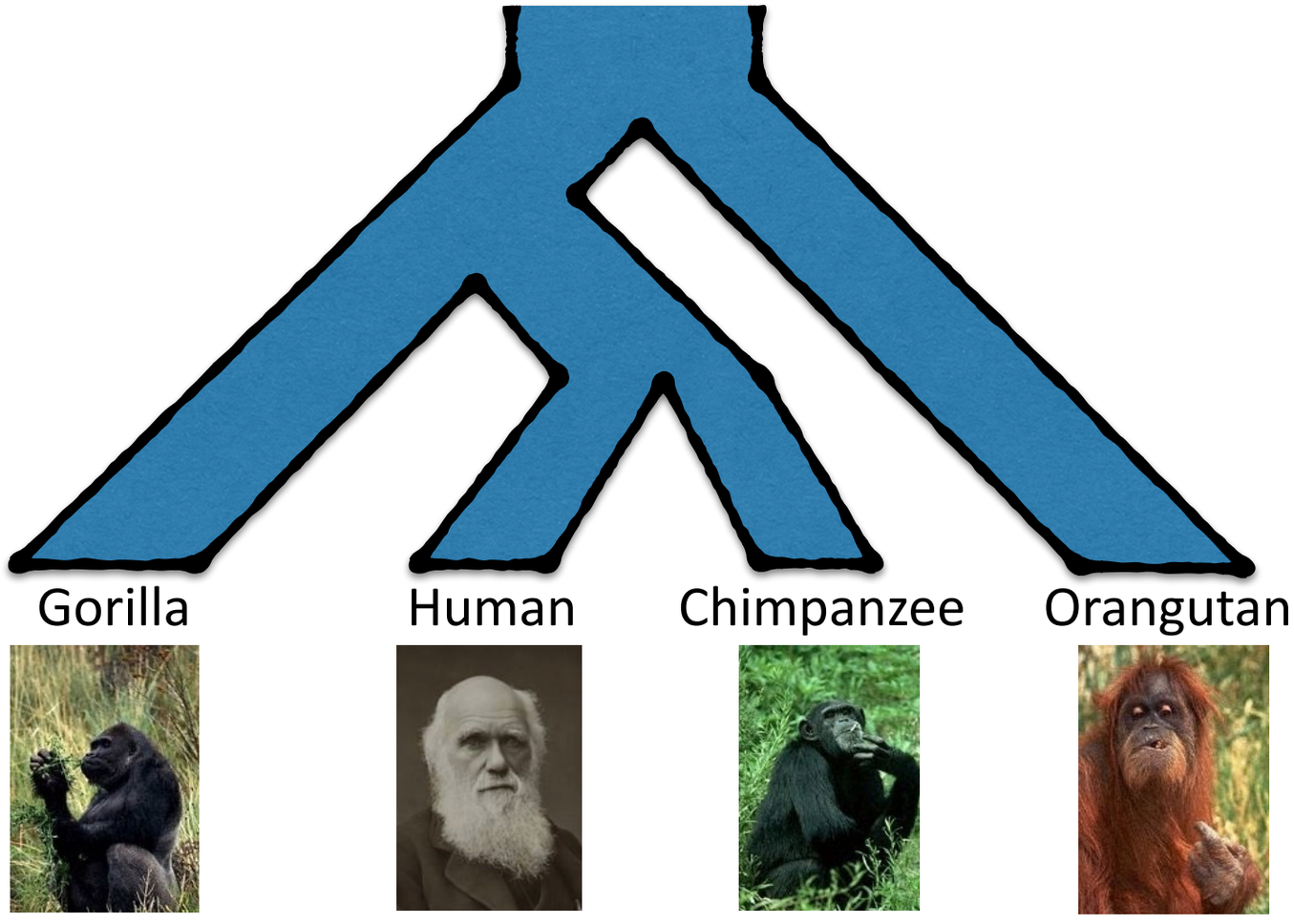
"***Nothing in biology makes sense except in the light of evolution***." Dobzhansky, 1973



source: http://www.evogeneao.com/

Phylogeny

Leaves are Species (a.k.a. **taxa**)



The branching structure shows evolutionary relationships

Why useful?

It ENABLEs downstream analyses

Order of biological innovation

Zym.: algae-like

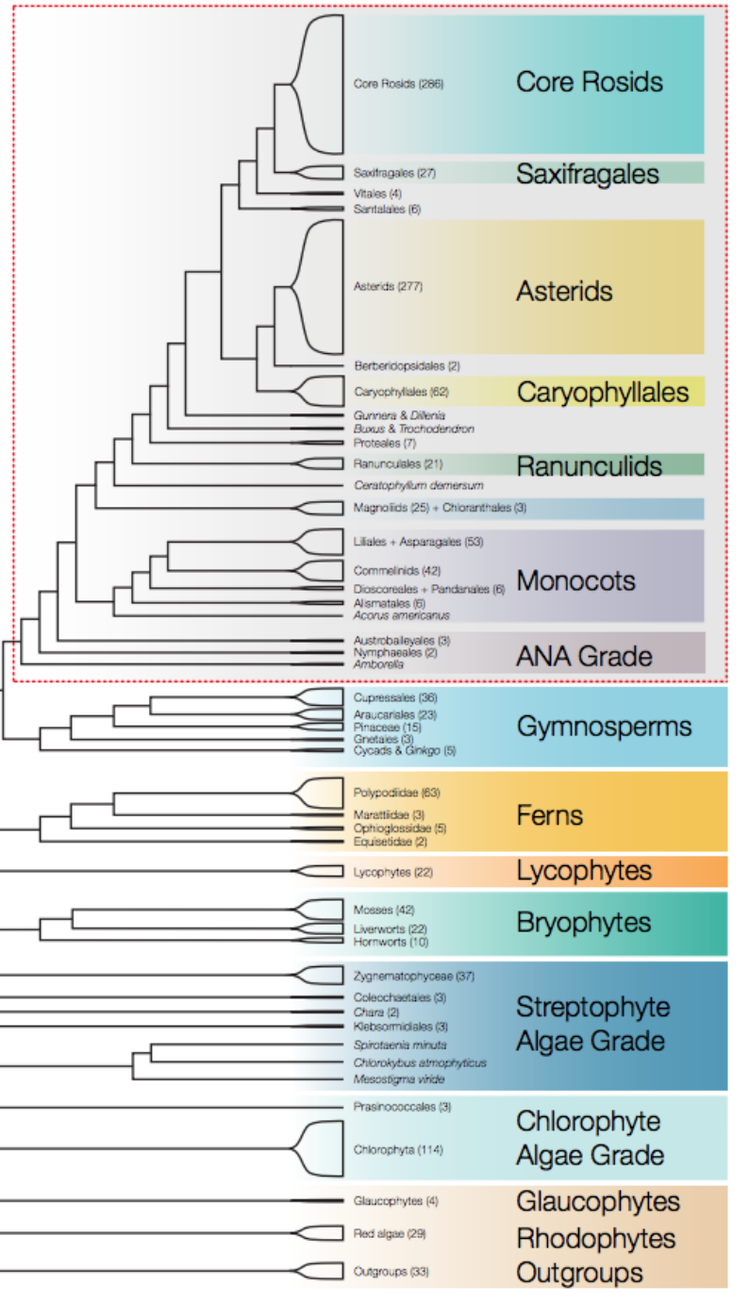
Chara: land plant-like

1KP project [unpublished results]

Order of biological innovation

Zym.: algae-like

Chara: land plant-like

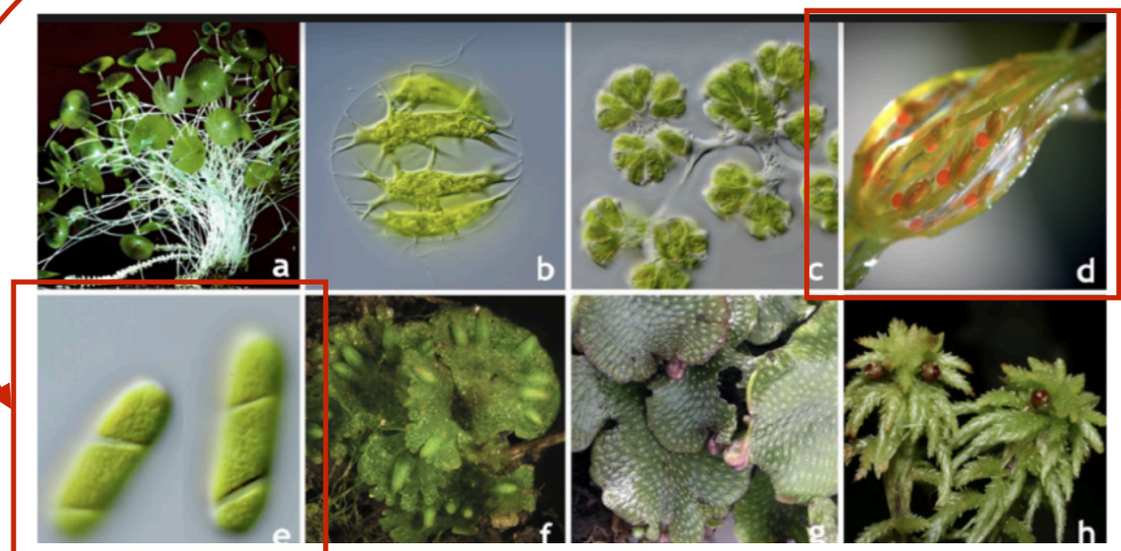


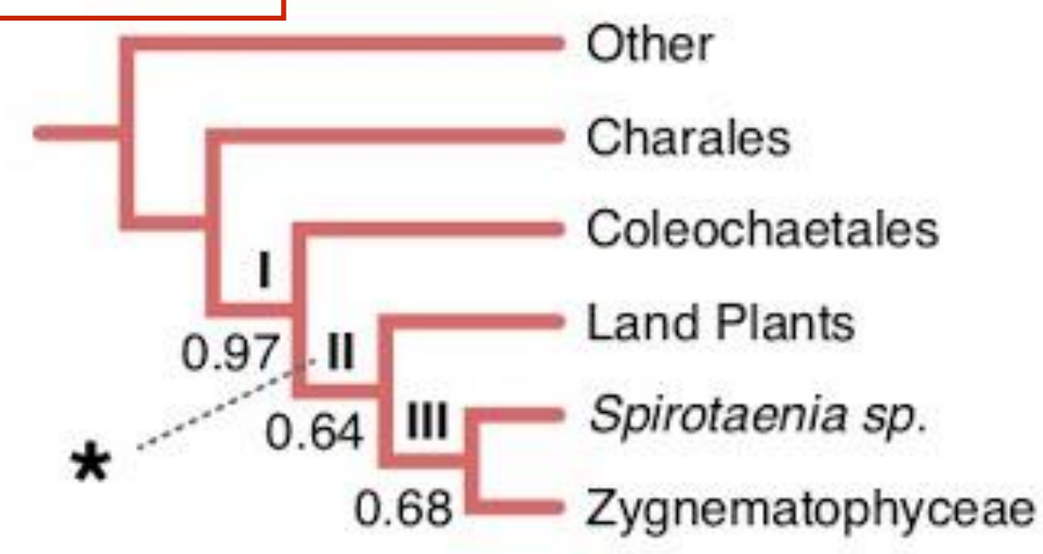
1KP project [unpublished results]

Order of biological innovation

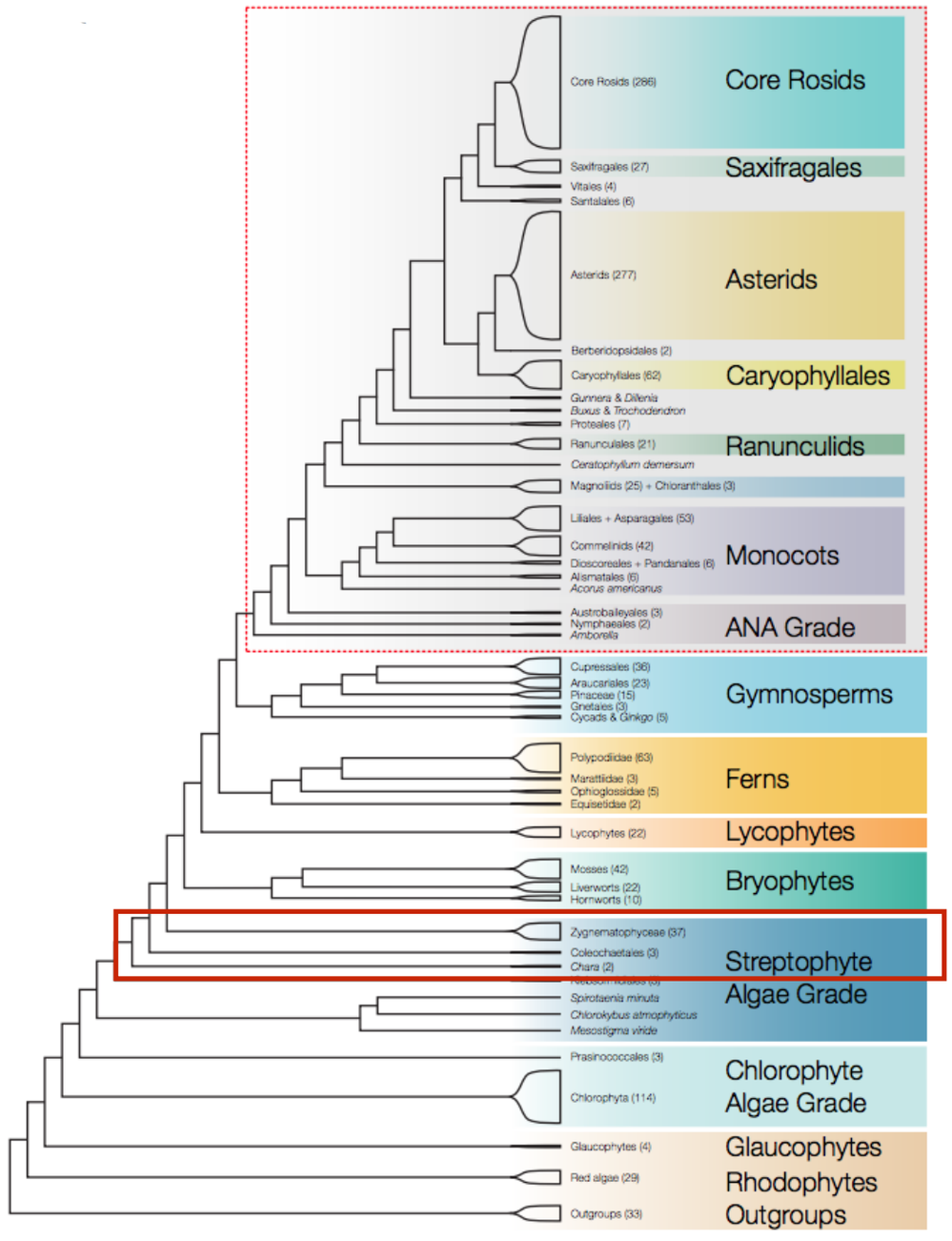
Zym.: algae-like

Chara: land plant-like

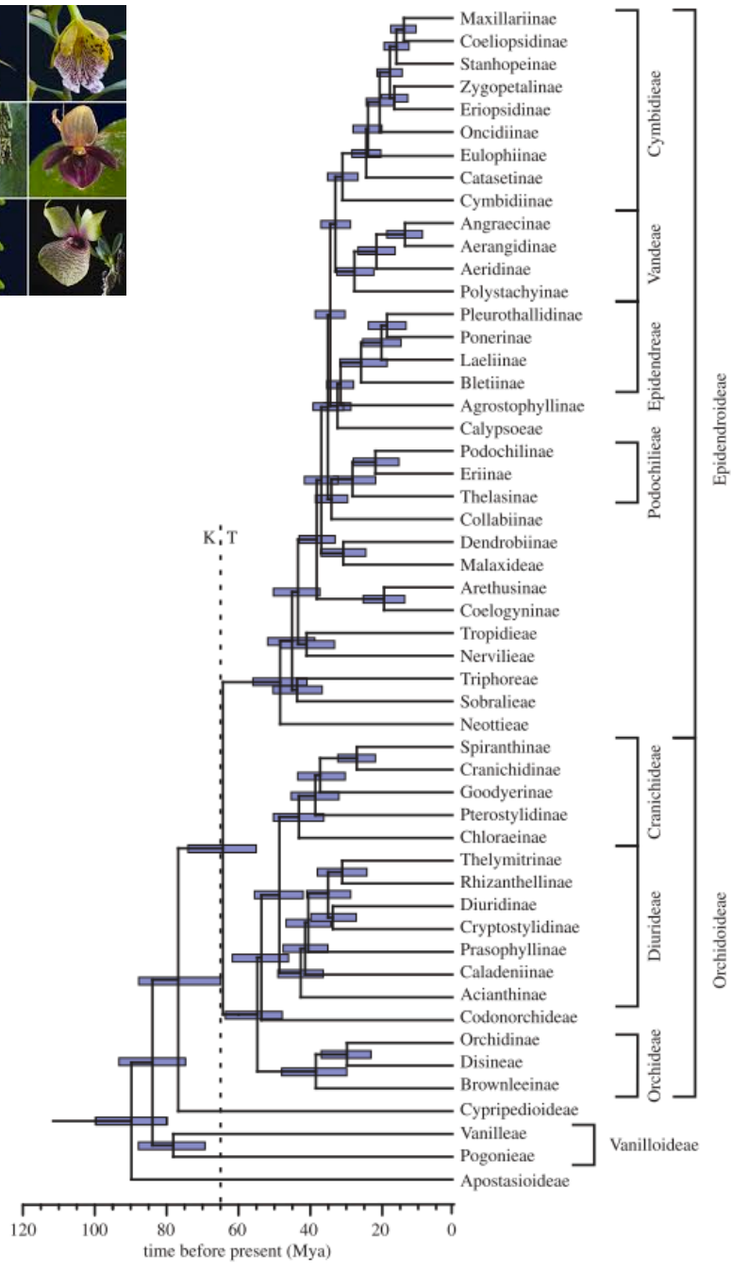




1KP project [unpublished results]

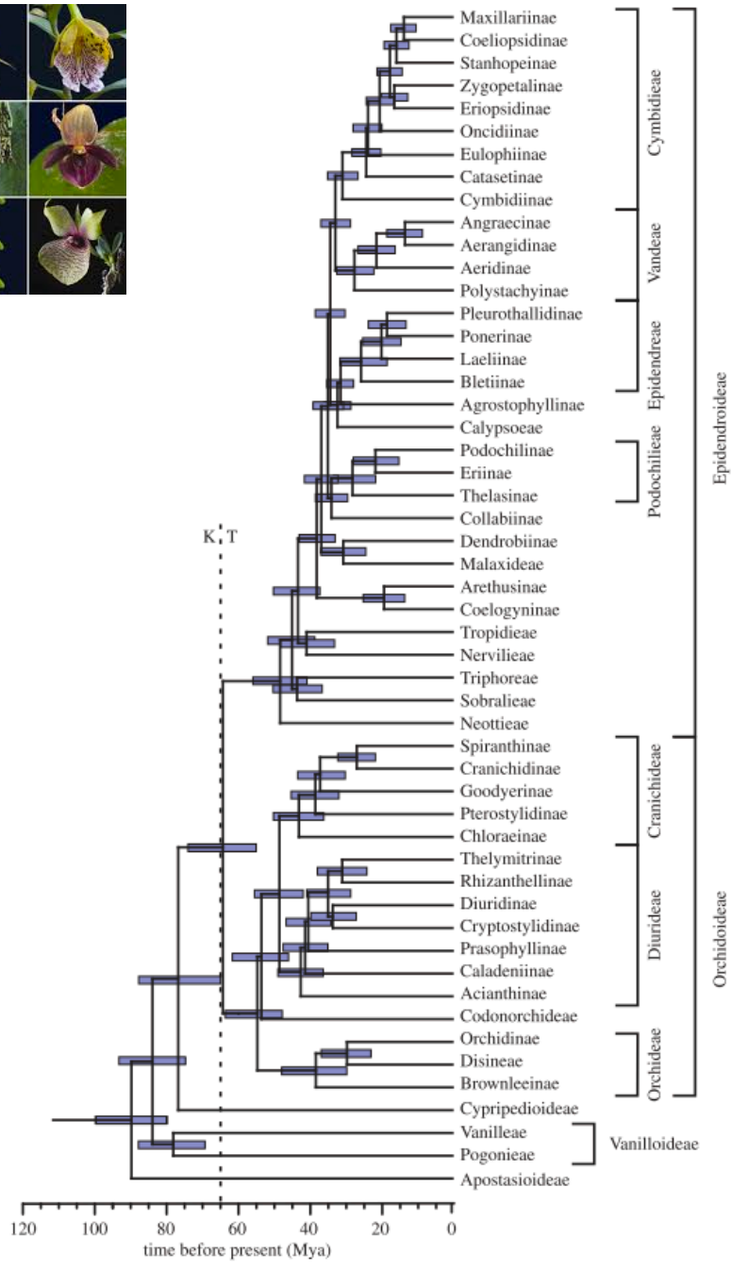


Understand timing of events

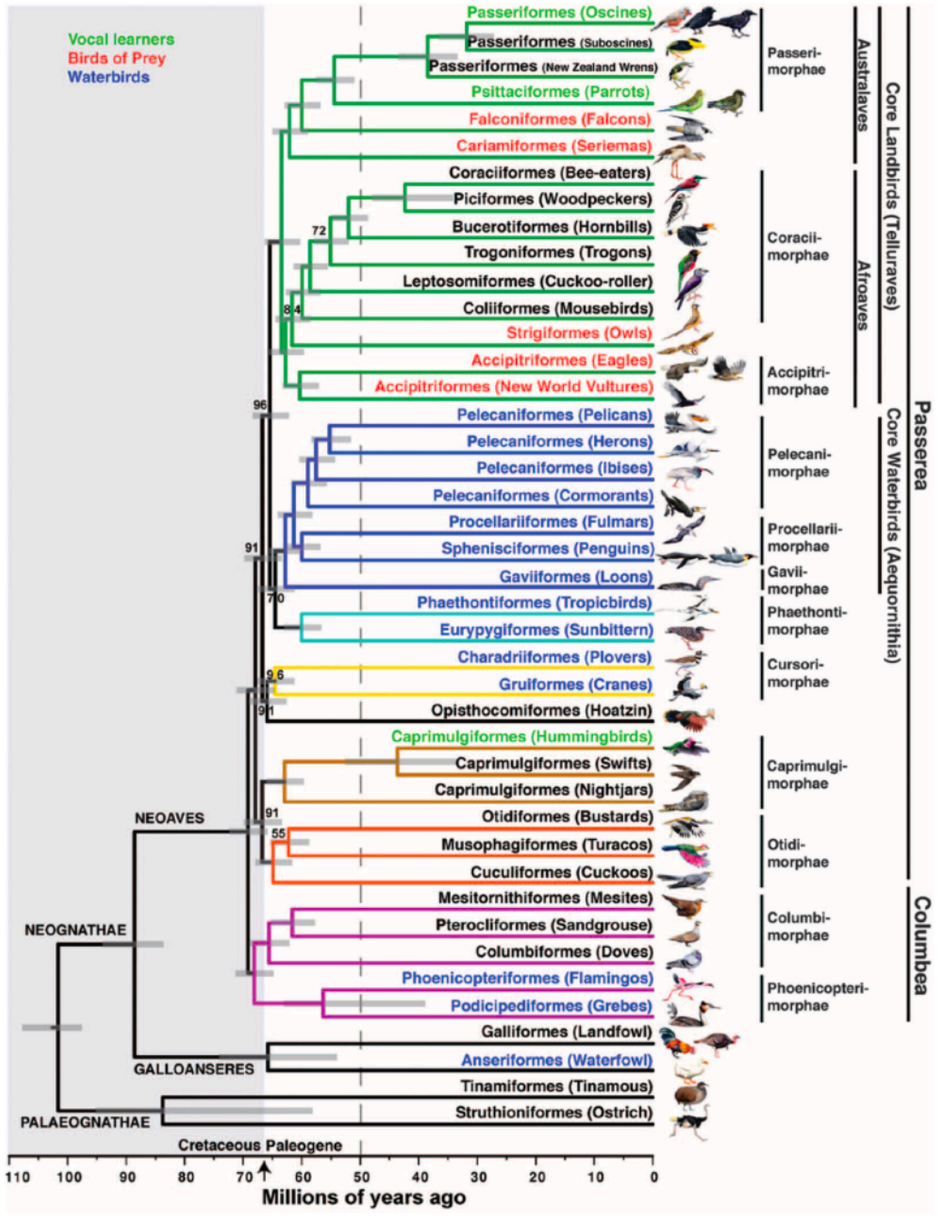


Givnish *et al*, Royal Society B, 2015

Understand timing of events



Givnish *et al*, Royal Society B, 2015



Jarvis, Mirarab *et al*, Science, 2014

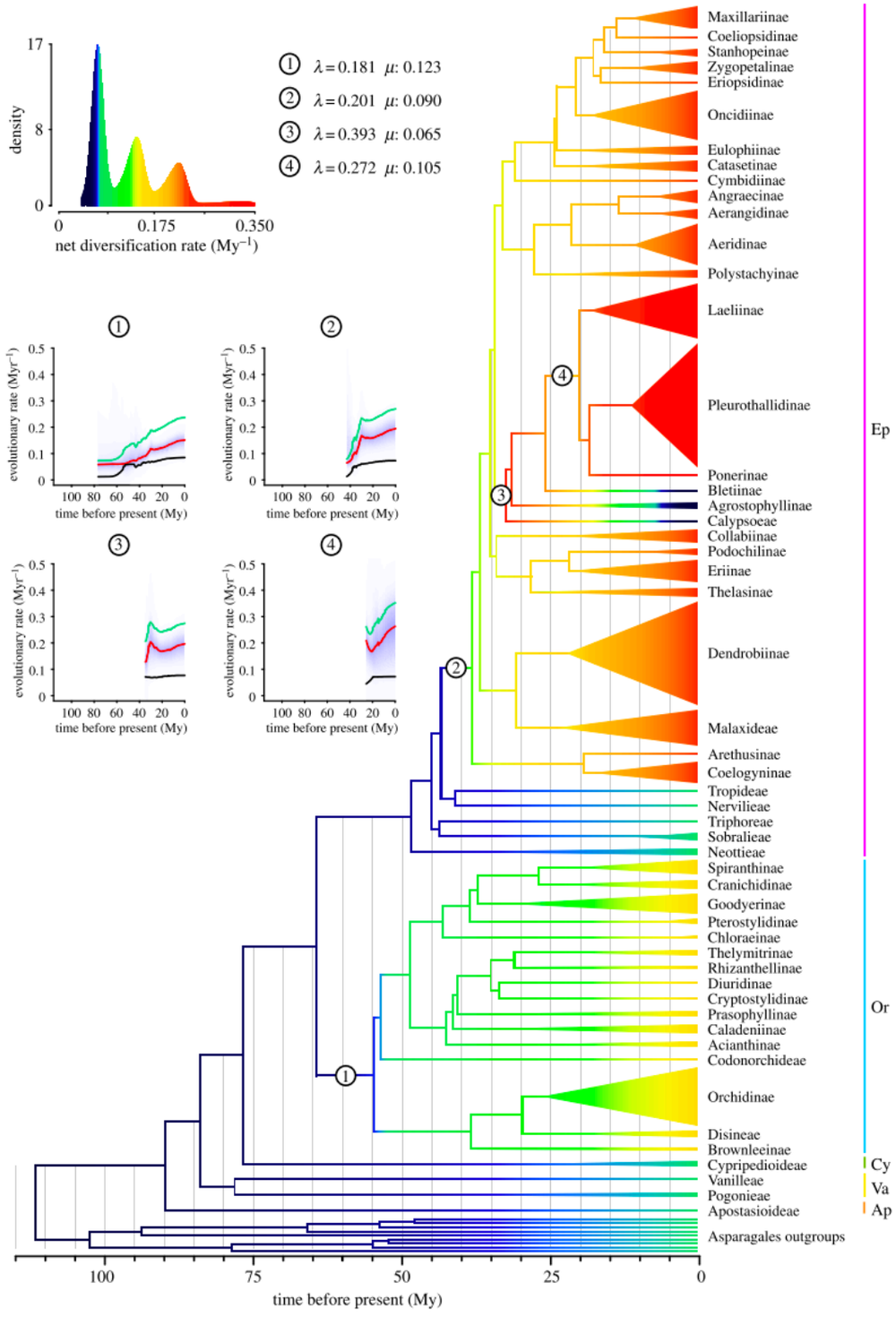
Understand drivers of diversity

- Orchids have more species than mammals, birds and reptiles combined

Understand drivers of diversity

- Orchids have more species than mammals, birds and reptiles combined

- Three significant shifts of diversification rate



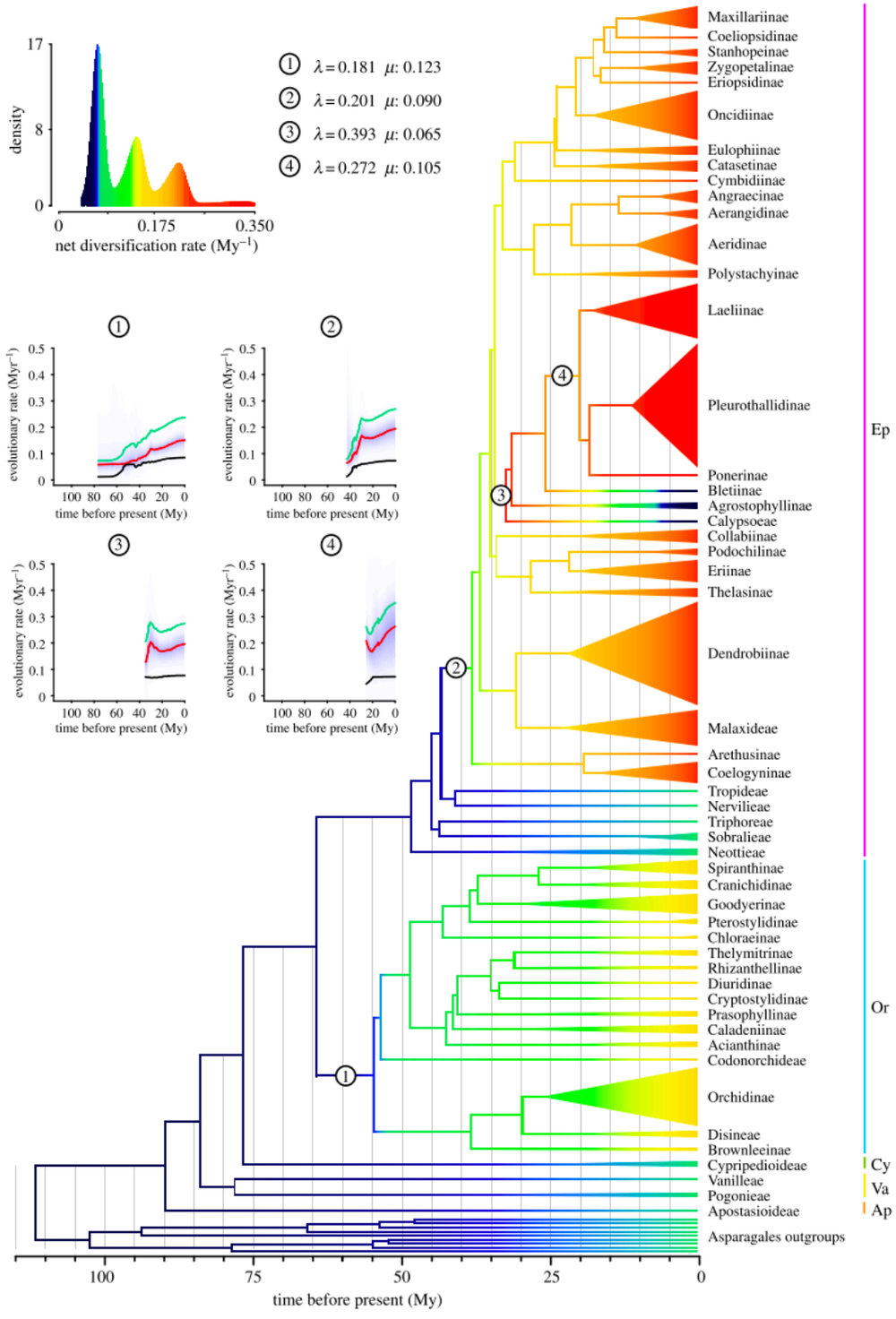
Givnish et al, Royal Society B, 2015

Understand drivers of diversity

- Orchids have more species than mammals, birds and reptiles combined

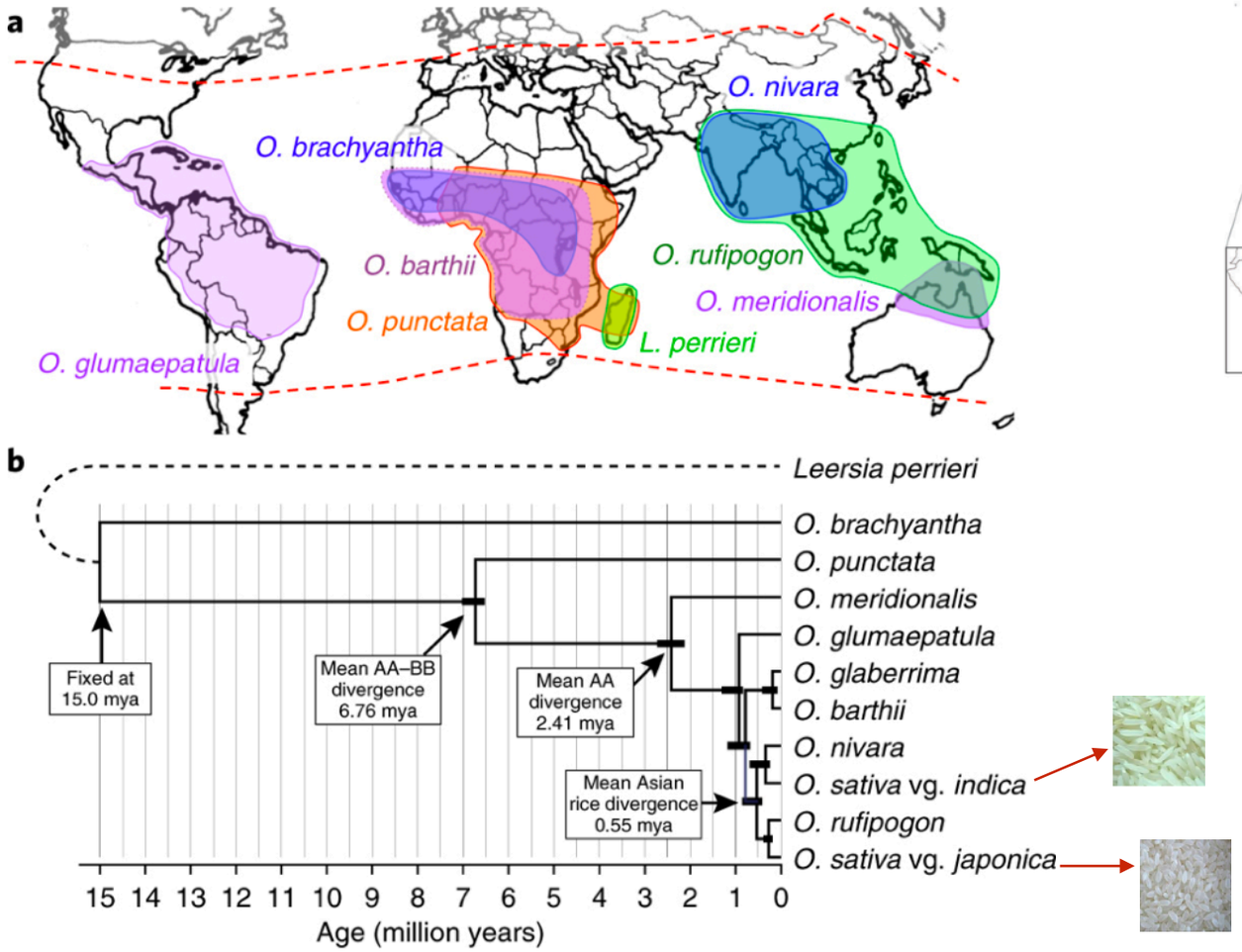
- Three significant shifts of diversification rate

- These coincide with evolution of pollinia, the epiphytic habit, CAM photosynthesis, tropical distribution, and pollination via butterflies or bees

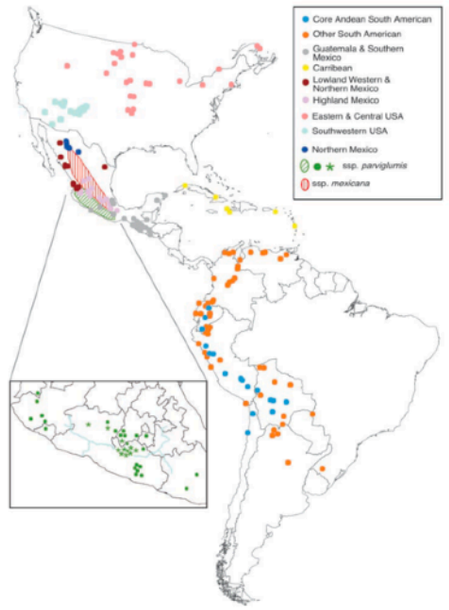


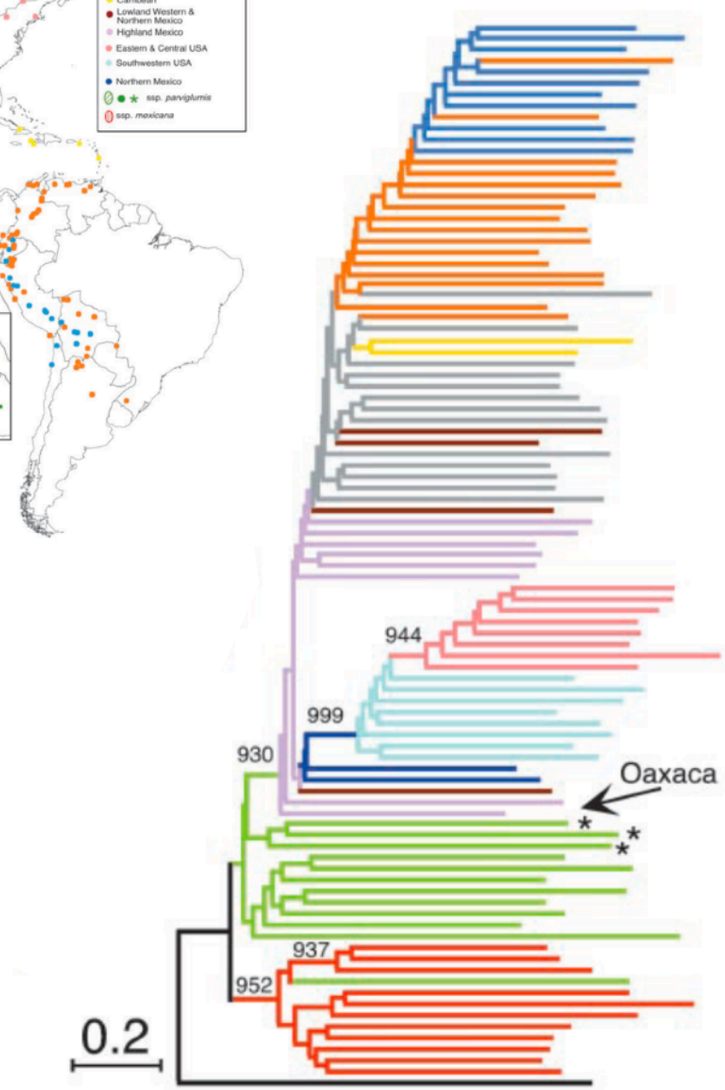
Givnish et al, Royal Society B, 2015

Domestication history



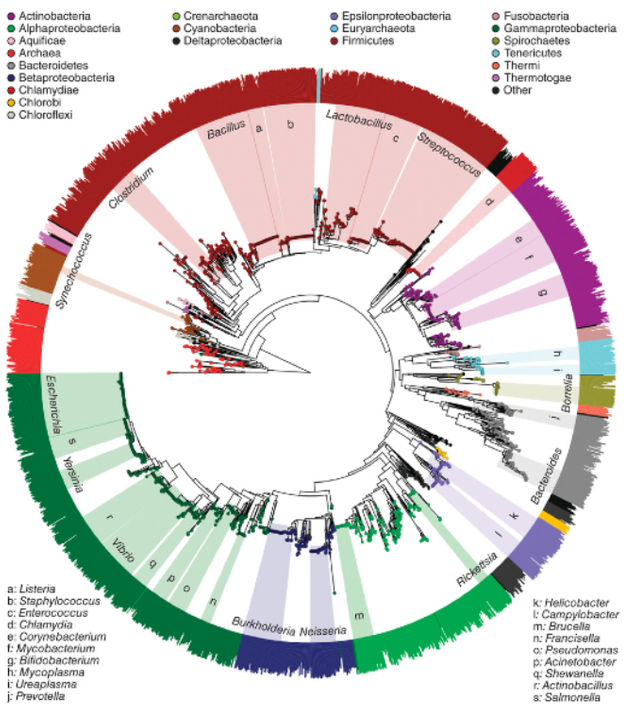
Stein *et al*, Nature genetics, 2018





Matsuoka et al, PNAS, 2002

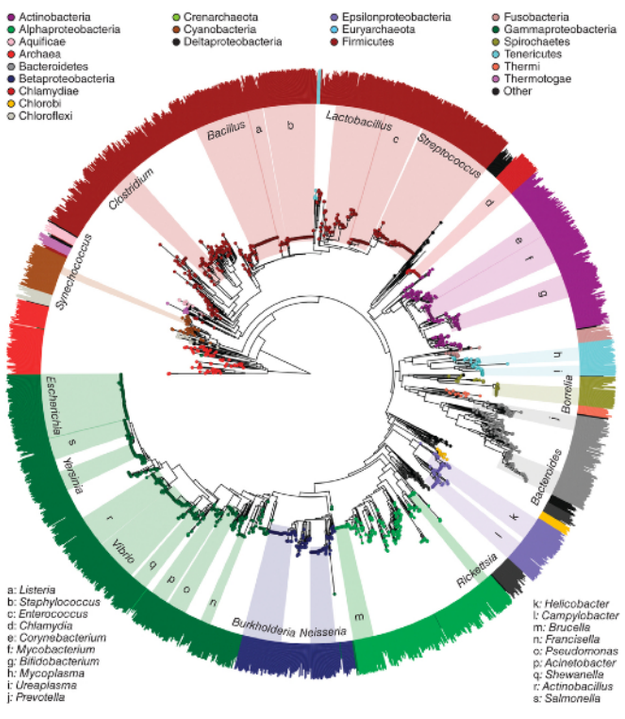
Biomedical Applications



source: Langille et al., Nature, 2013

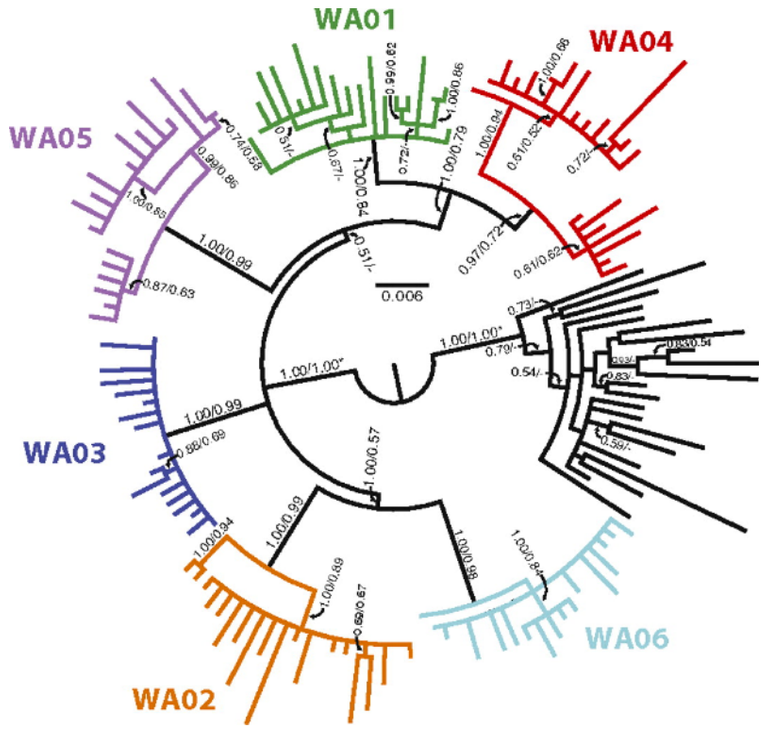
Microbiome

Biomedical Applications



source: Langille et al., Nature, 2013

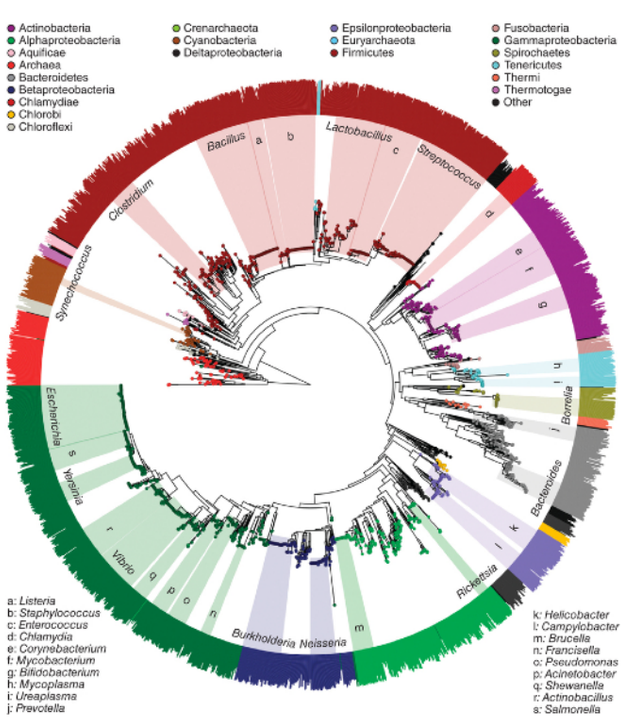
Microbiome



source: Scaduto et al., PNAS, 2010

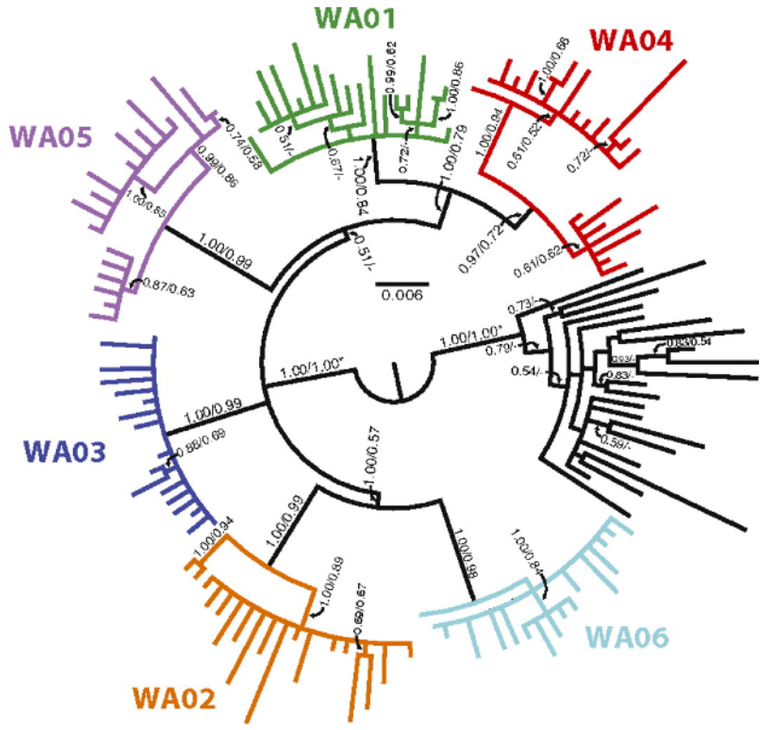
Forensic

Biomedical Applications



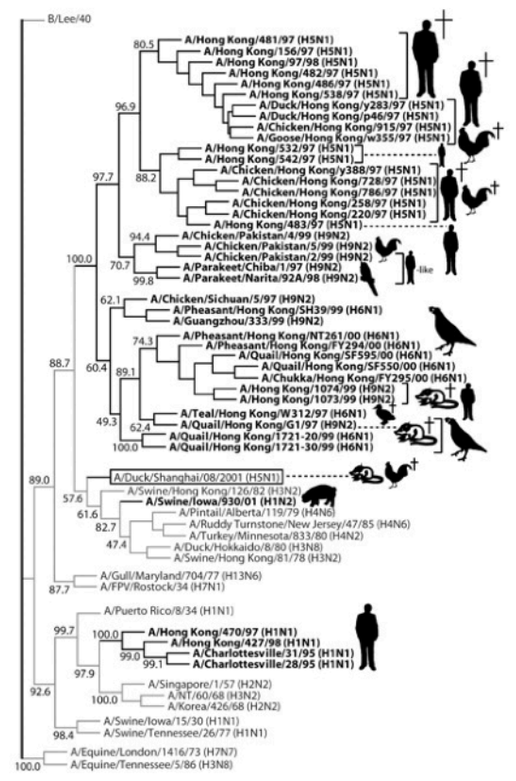
source: Langille et al., Nature, 2013

Microbiome



source: Scaduto et al., PNAS, 2010

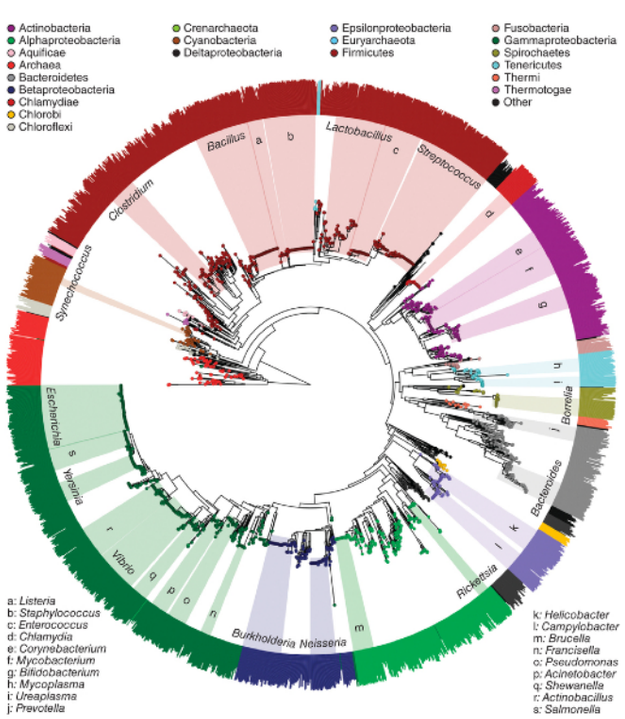
Forensic



Avian influenze

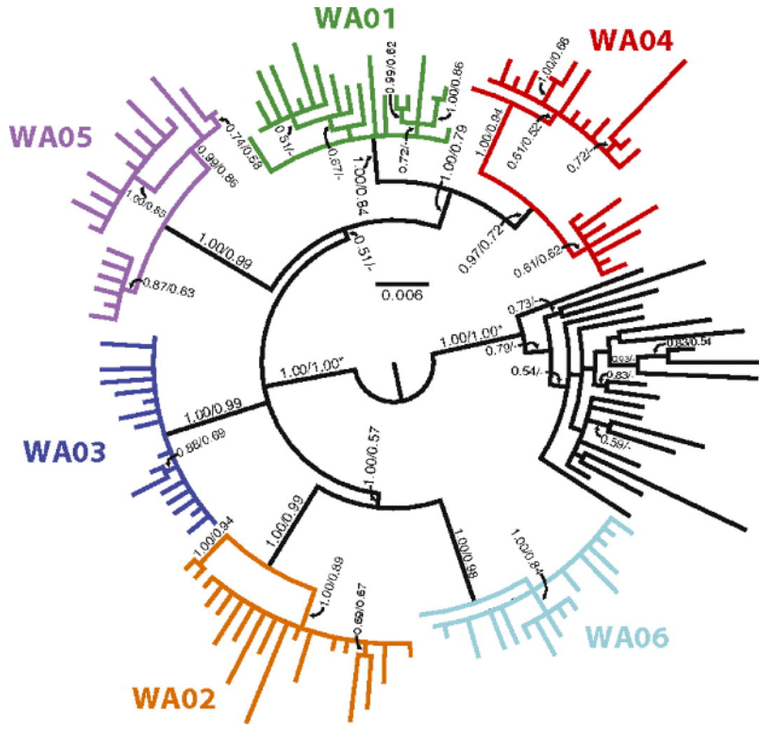
Biomedical Applications

"***Nothing in evolution makes sense except in the light of phylogeny***." multiple coinage



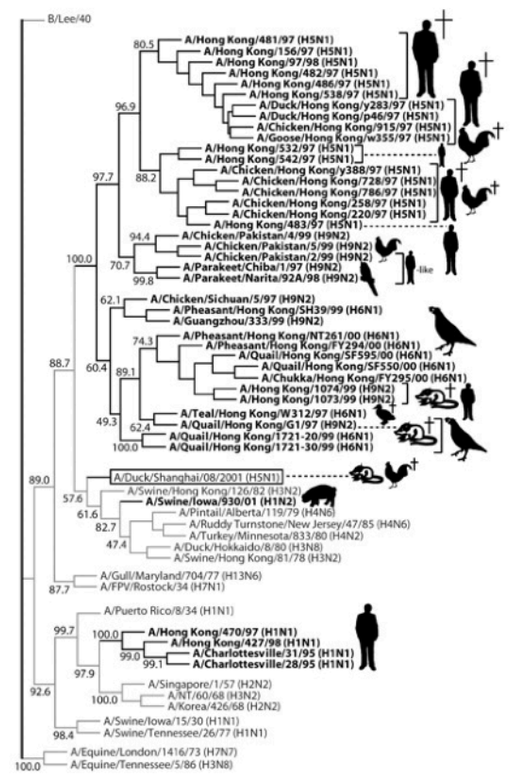
source: Langille et al., Nature, 2013

Microbiome



source: Scaduto et al., PNAS, 2010

Forensic



Avian influenze

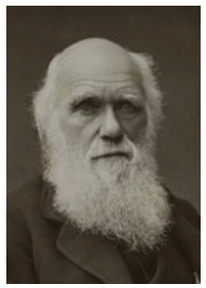
Why related to engineering?

Computational challenges?

Phylogenetic reconstruction from data

Gorilla

ACTGCACACCG



Human

ACTGCCCCCG

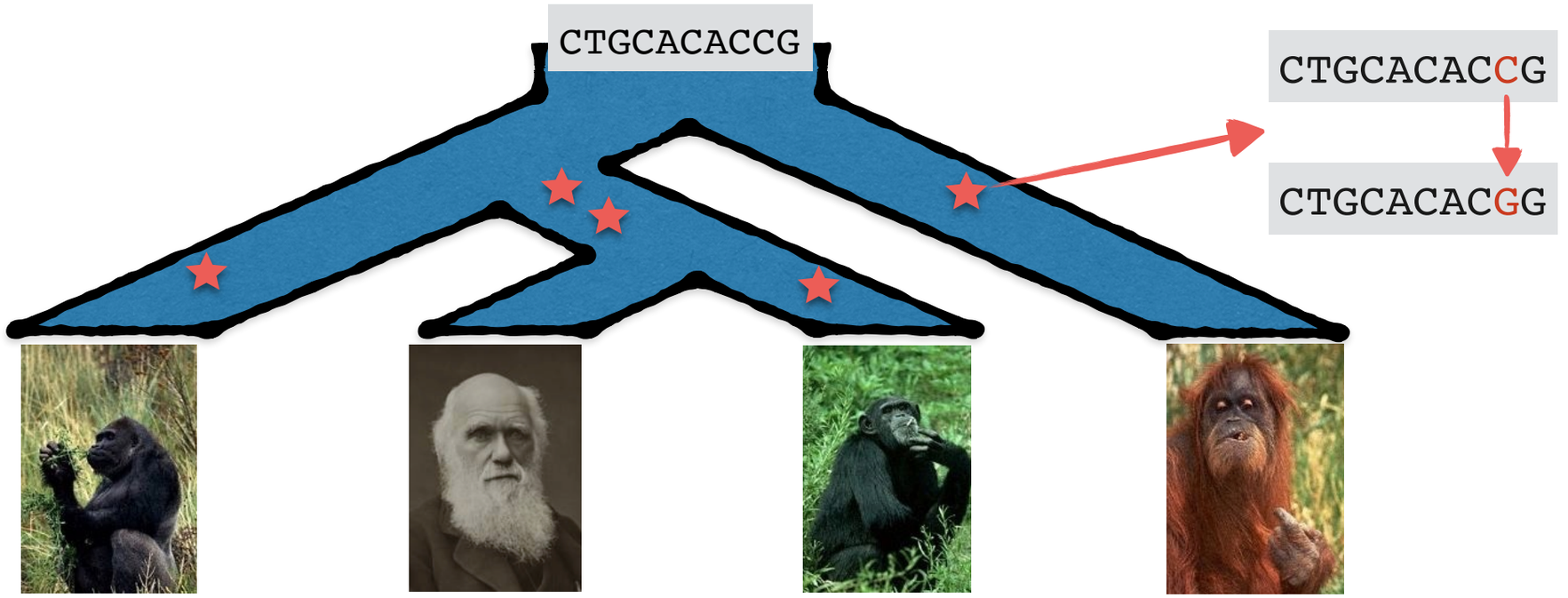
Chimpanzee

AATGCCCCCG

Orangutan

CTGCACACGG

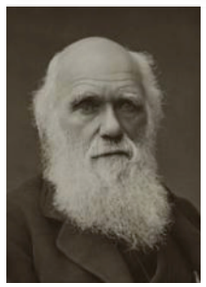
Phylogenetic reconstruction from data



Phylogenetic reconstruction from data

Gorilla

ACTGCACACCG



Human

ACTGCCCCCG

Chimpanzee

AATGCCCCCG

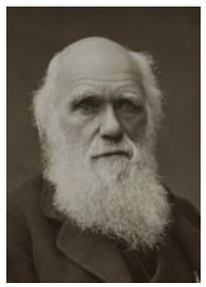
Orangutan

CTGCACACGG

Phylogenetic reconstruction from data

Gorilla

ACTGCACACCG



Human

ACTGCCCCCG

Chimpanzee

AATGCCCCCG

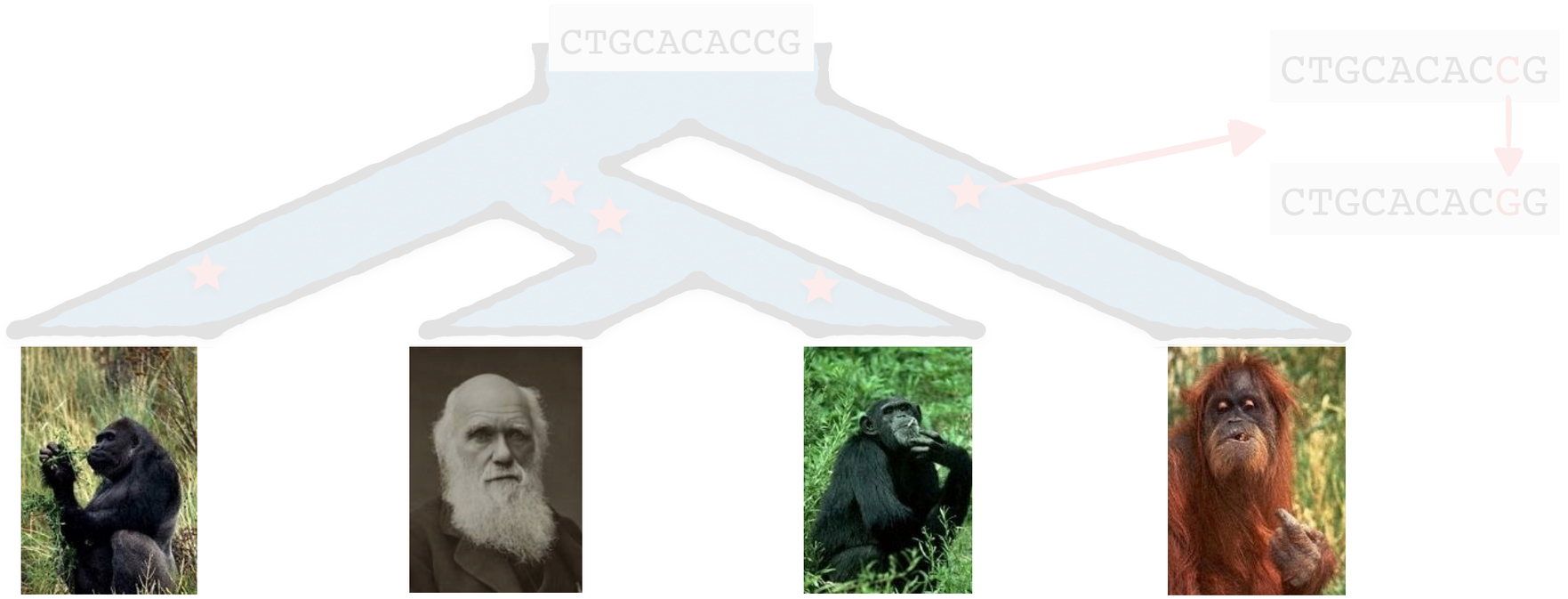
Orangutan

CTGCACACGG

Gorilla	ACTGCACACCG
Human	ACTGC-CCCCG
Chimpanzee	AATGC-CCCCG
Orangutan	-CTGCACACGG

$$D$$

Phylogenetic reconstruction from data



Gorilla
ACTGCACACCG

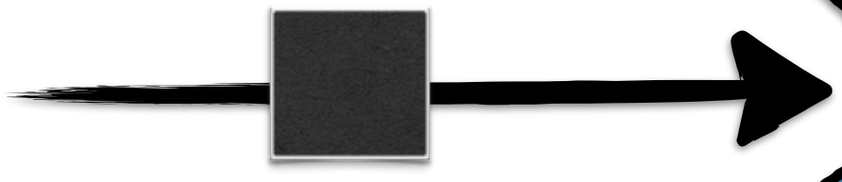
Human
ACTGCCCCCG

Chimpanzee
AATGCCCCCG

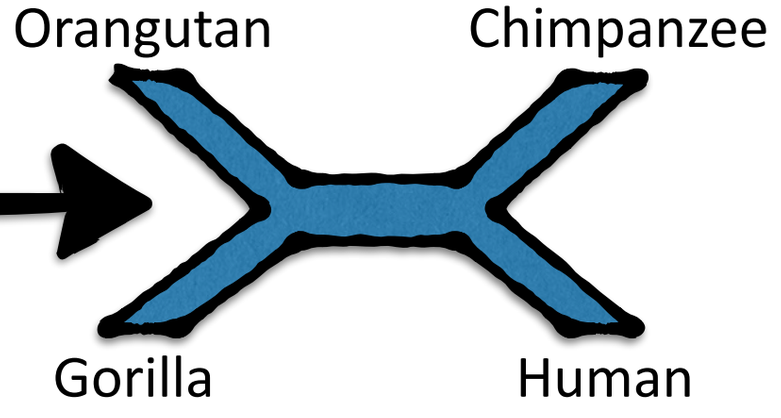
Orangutan
CTGCACACGG

Gorilla ACTGCACACCG
Human ACTGC-CCCCG
Chimpanzee AATGC-CCCCG
Orangutan -CTGCACACGG

D



$P(D|T)$



T

Computational inference

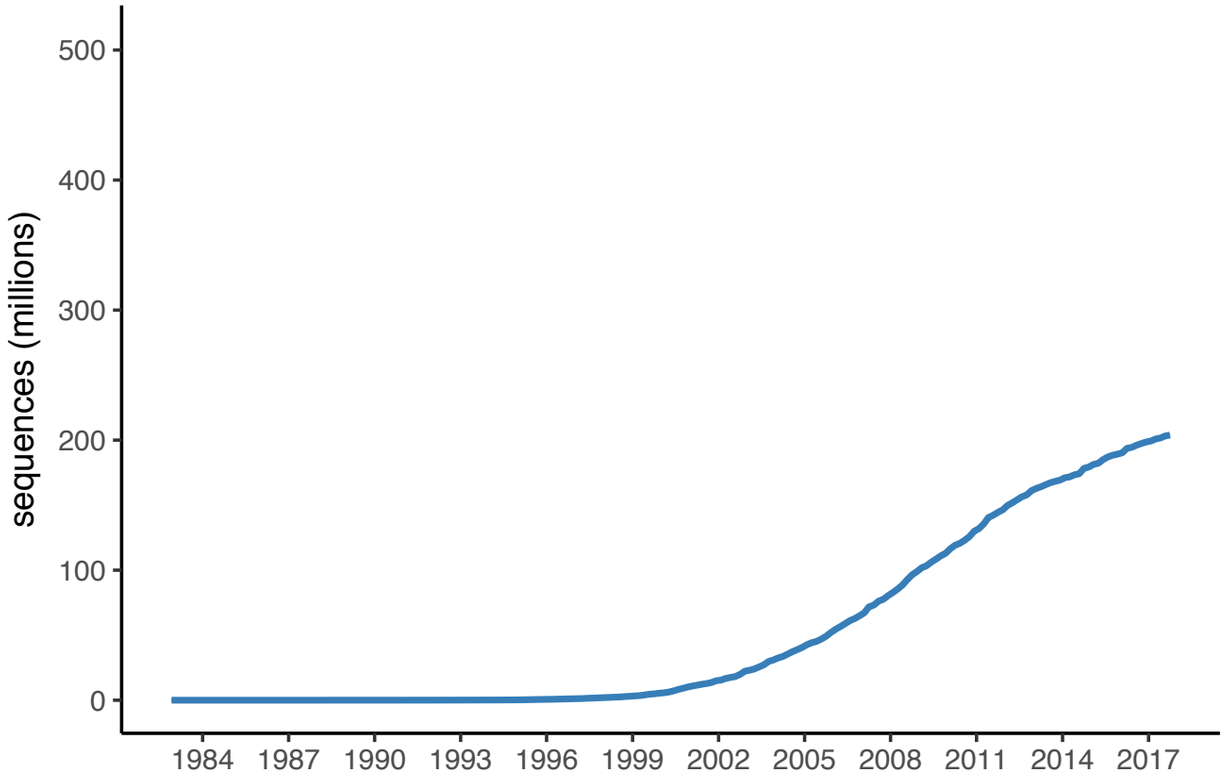
- Really no “experimental” way to validate results.
    - Phylogenetics is ENABLEd only due to computation
    - We use computation-heavy procedures to calculate statistical support

Computational inference

- Really no “experimental” way to validate results.
  - Phylogenetics is ENABLEd only due to computation
    - We use computation-heavy procedures to calculate statistical support
- Genome evolution is complex. We need complex statistical models for accuracy
  - And need large data to infer under complex models

Sequence data growth

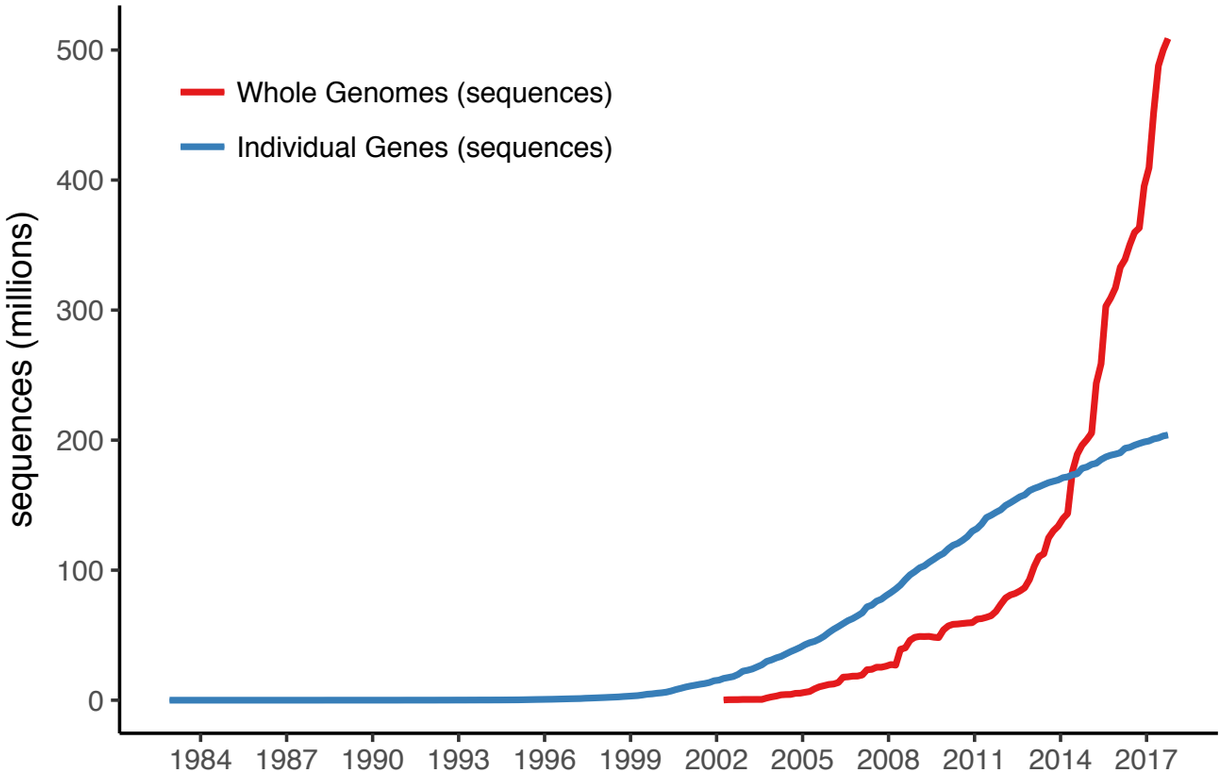
data source: NCBI (http://www.ncbi.nlm.nih.gov/genbank/statistics)



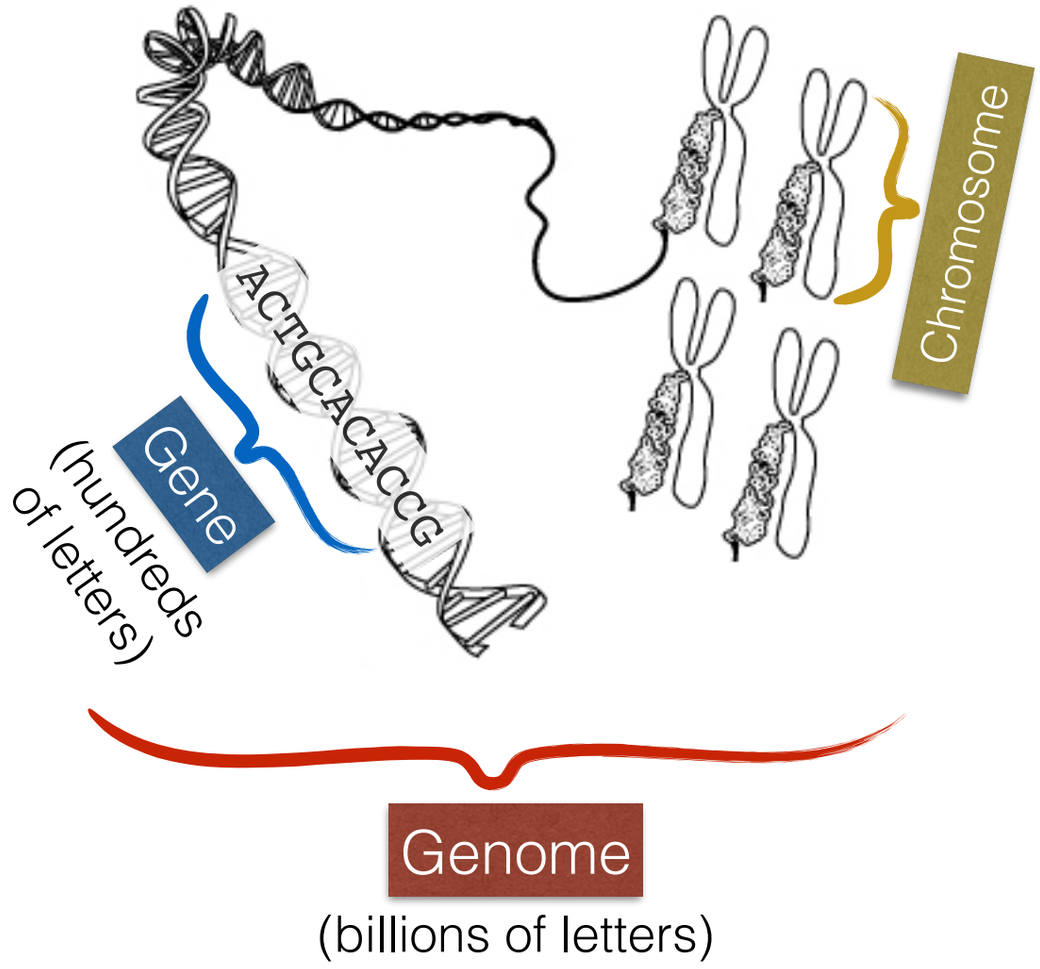
- Rapid growth in the available sequences

Sequence data growth

data source: NCBI (http://www.ncbi.nlm.nih.gov/genbank/statistics)



- Rapid growth in the available sequences

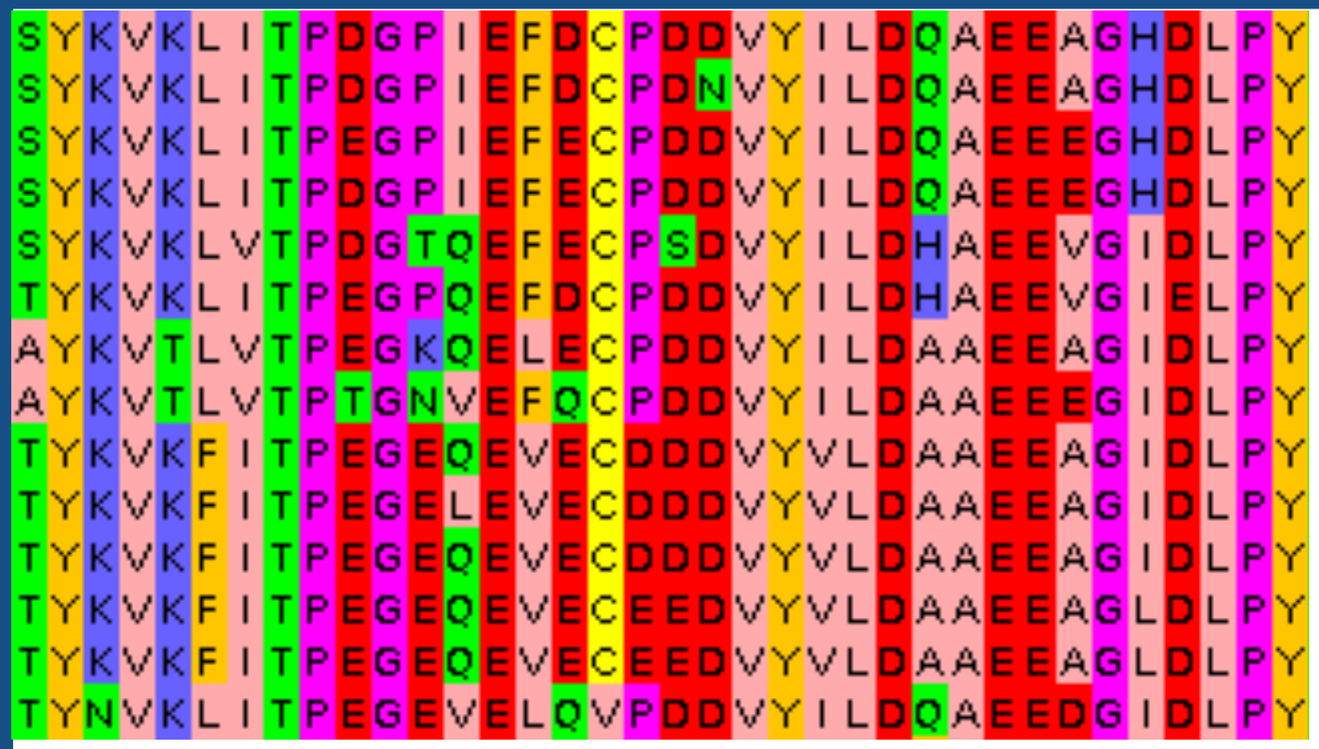


- Our focus has shifted to “whole genomes”

More genomic regions

Billions of columns

More sequences



A million rows

Phylogenetic reconstruction : a computational nightmare

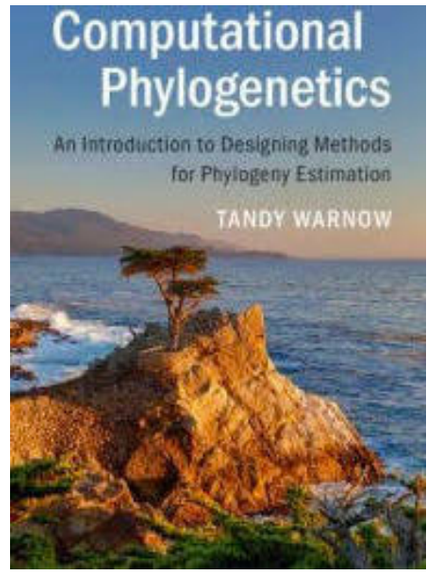
- Almost all problems are NP-Hard!

- The search space is huge.

  - Focusing on topology alone, there are $(2n\text{-}5)!!$ trees with **n** leaves

  - We also care about branch lengths: $\mathbb{R}^{2\text{n-}3}$

- We are interested in n in $10^1 \ldots 10^6$ range



Tandy Warnow

To scale to large datasets …

- Approximate and heuristic solutions

To scale to large datasets …

- Approximate and heuristic solutions
- Make the problem easier
- Divide-and-conquer
- Constrained search

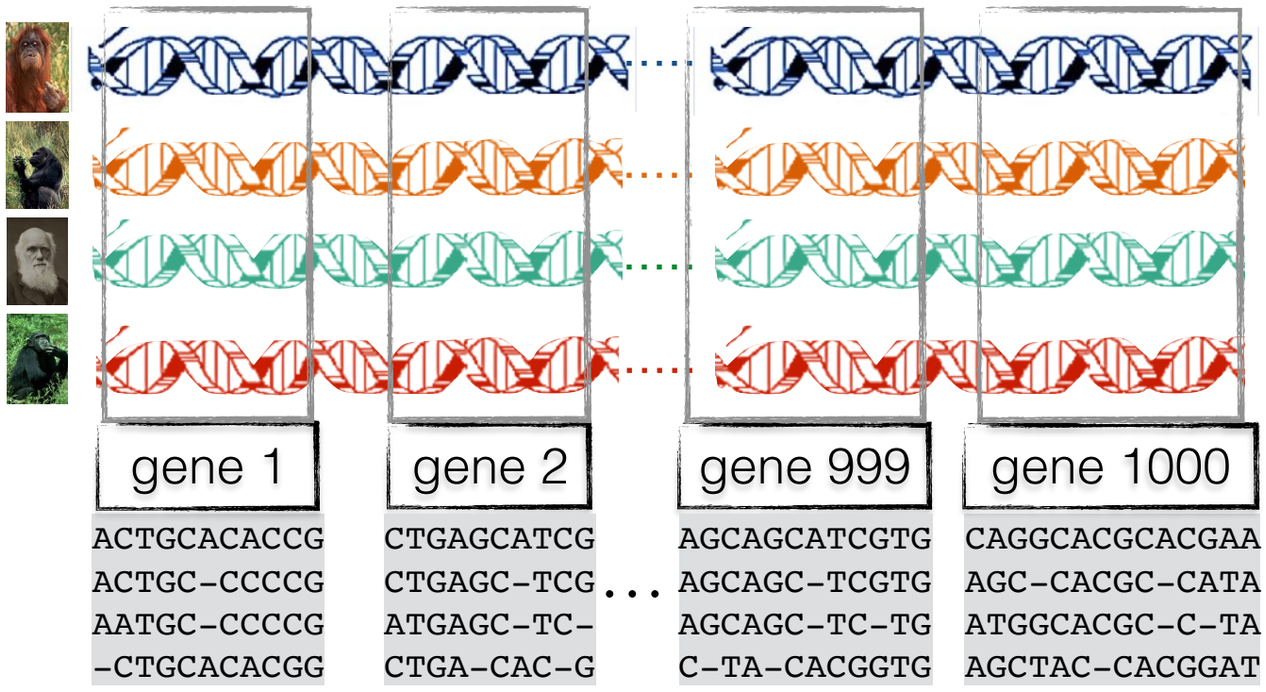
To scale to large datasets …

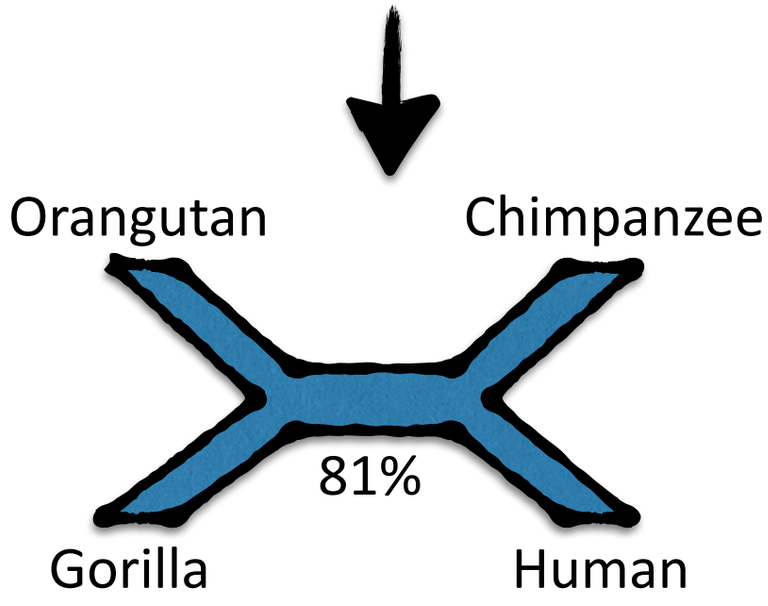
- Approximate and heuristic solutions
- Make the problem easier
    - Divide-and-conquer
    - Constrained search
- Develop optimized code.
    - Take advantage of High Performance Computing

To scale to large datasets …

- Approximate and heuristic solutions
- Make the problem easier
- But how about accuracy?
- Do large data help?
- Develop optimized code.
  - Take advantage of High Performance Computing

Phylogenomics:
more data ⟶ better inference





Phylogenomics:
more data ⟶ better inference

Species tree error

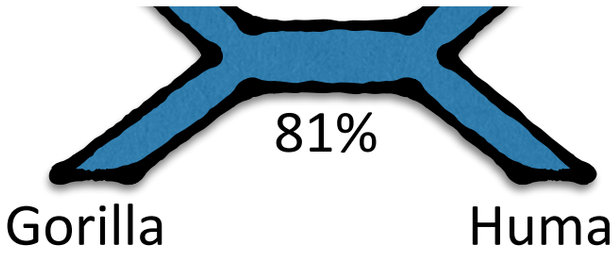
nature
International journal of science

Evolution

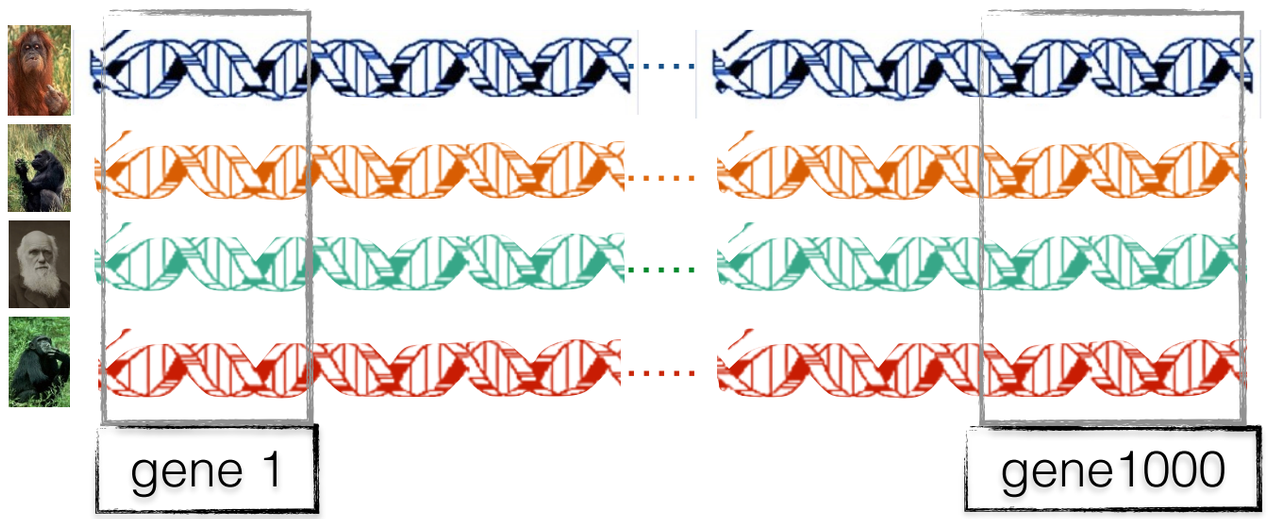
Ending incongruence

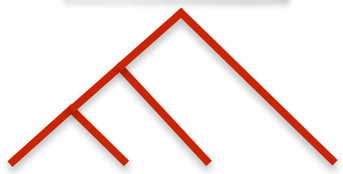
Henry Gee

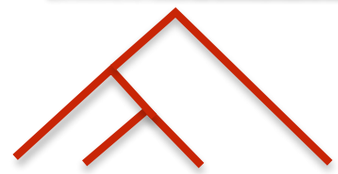
Recovering the true evolutionary history of any group of organisms has seemed impossible. The availability of large amounts of genomic data promises an era in which the uncertainties are better constrained.



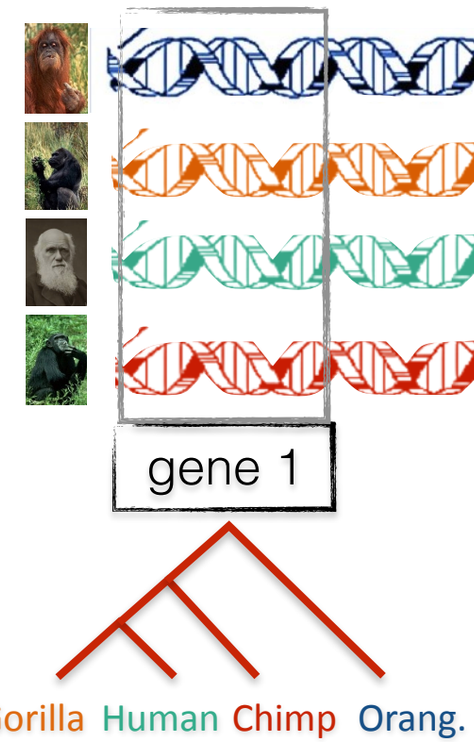
Gene tree discordance

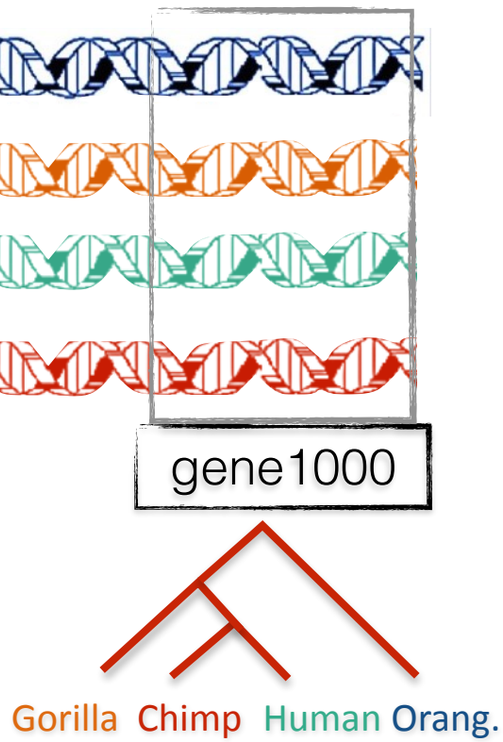


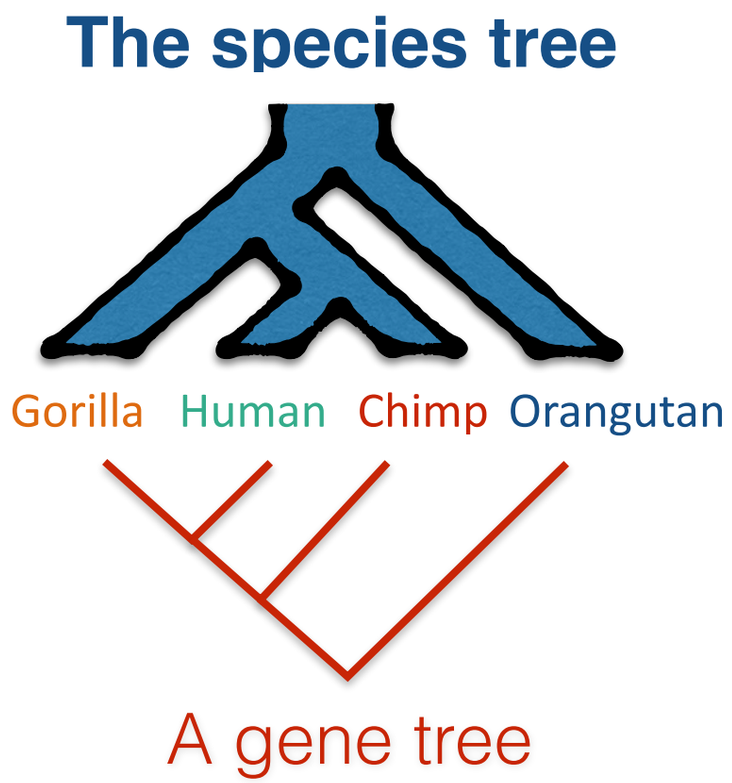




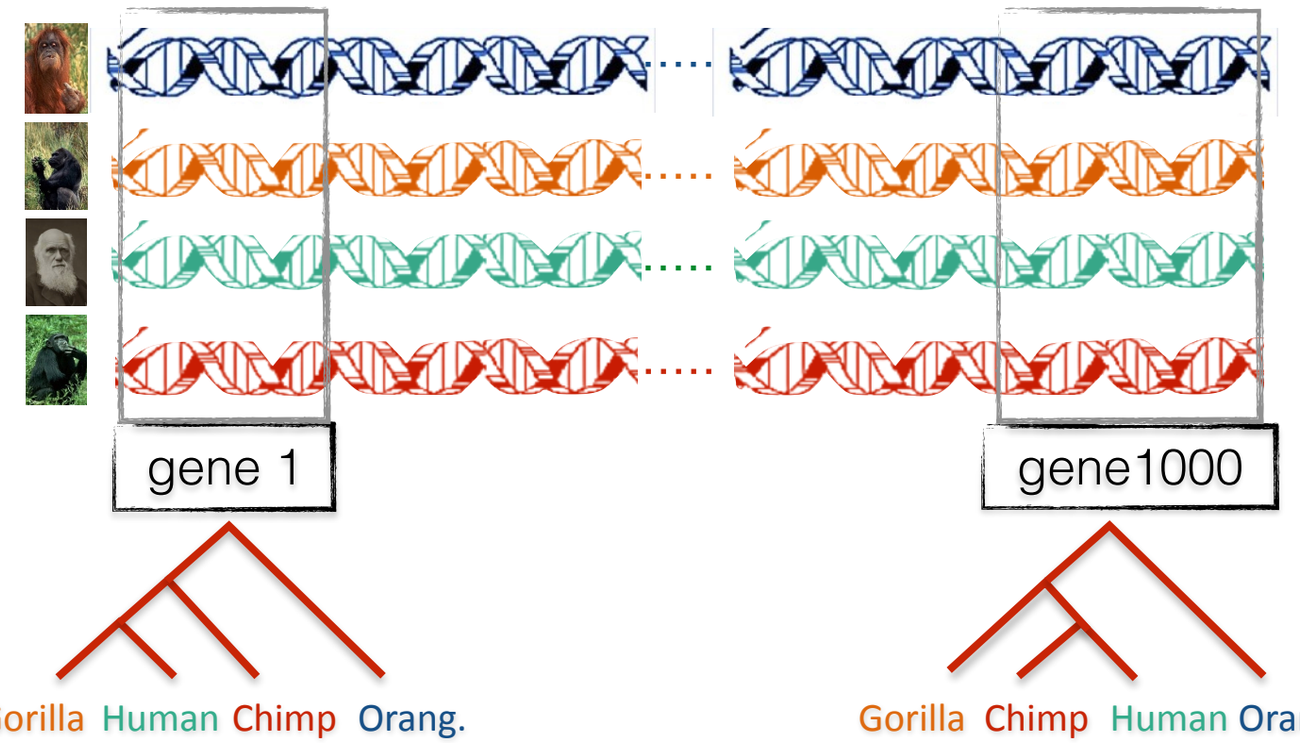
Gene tree discordance

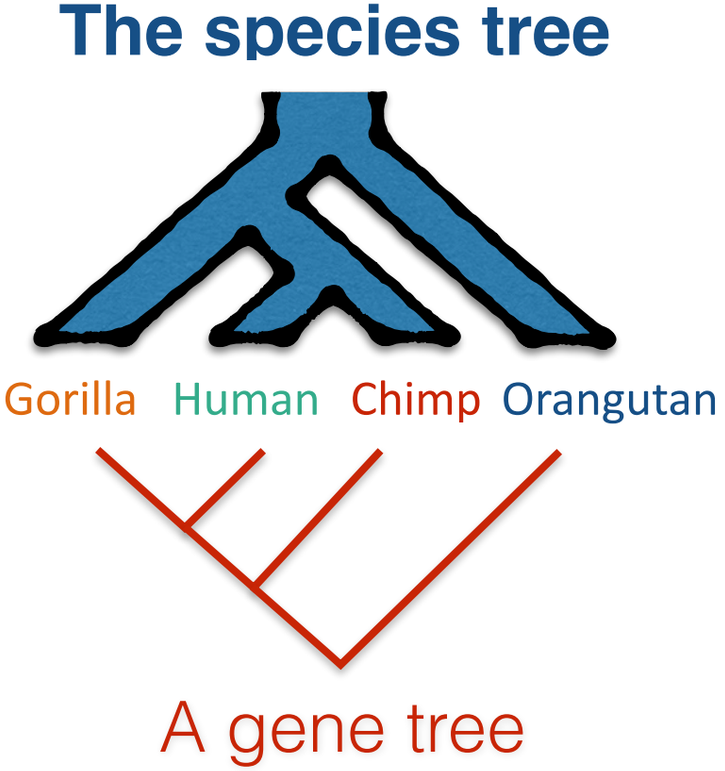






Gene tree discordance

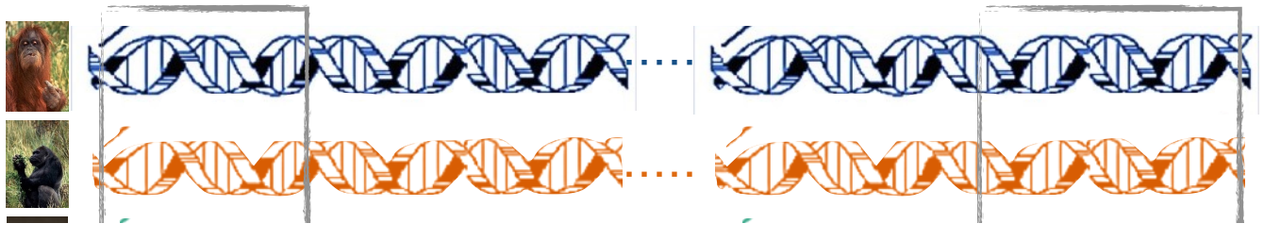




Causes of gene tree discordance include:

- Incomplete Lineage Sorting (ILS)
- Duplication and loss
- Horizontal Gene Transfer (HGT)

Gene tree discordance



The species tree

Trends in Genetics

CellPress

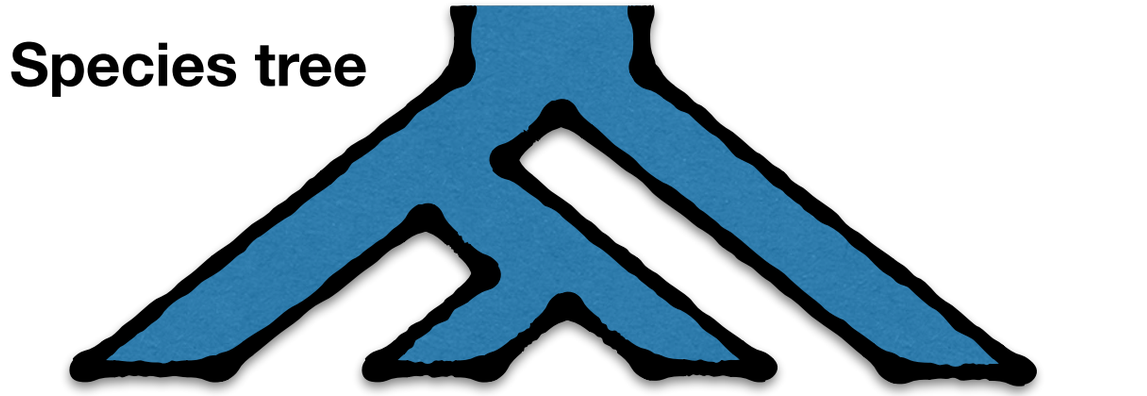
Volume 22, Issue 4, April 2006, Pages 225–231

Phylogenomics: the beginning of incongruence?

Olivier Jeffroy, Henner Brinkmann, Frédéric Delsuc, Hervé Philippe

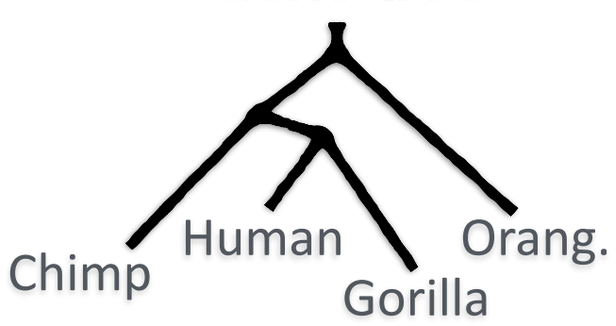
Show more

https://doi.org/10.1016/j.tig.2006.02.003

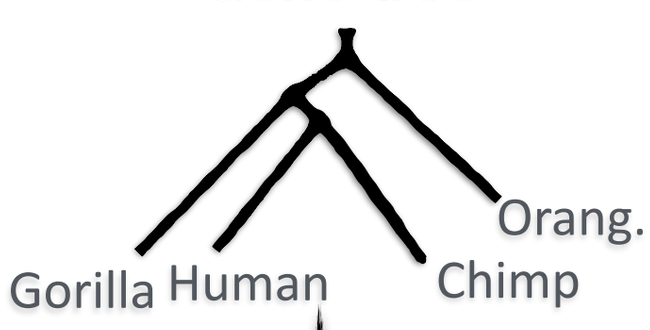


Gene evolution model

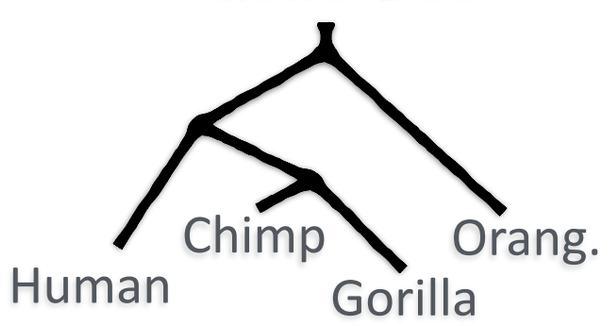
Gene tree



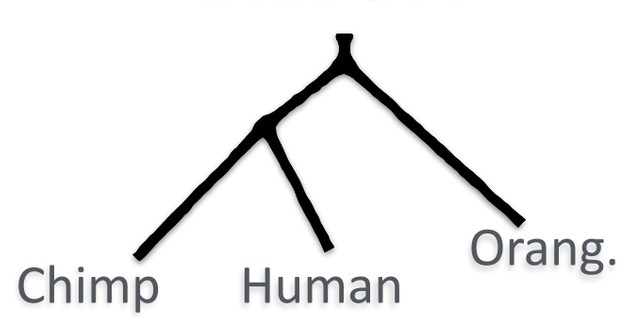
Gene tree



Gene tree



Gene tree



Sequence evolution model

Sequence data (Alignments)

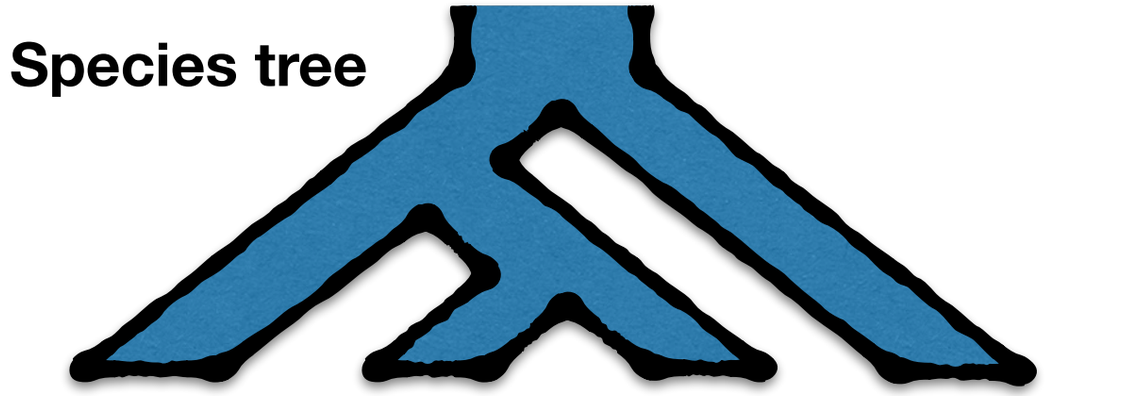
```
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
```

```
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
```

Sequence data (Alignments)

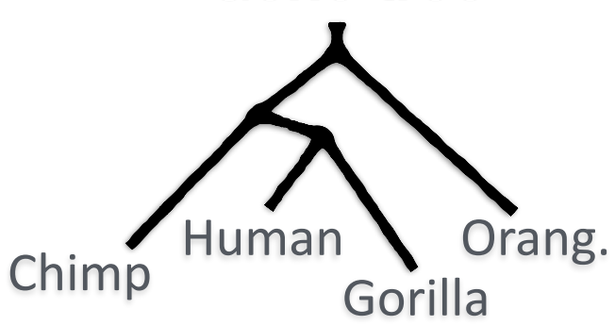
```
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
```

```
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT
```

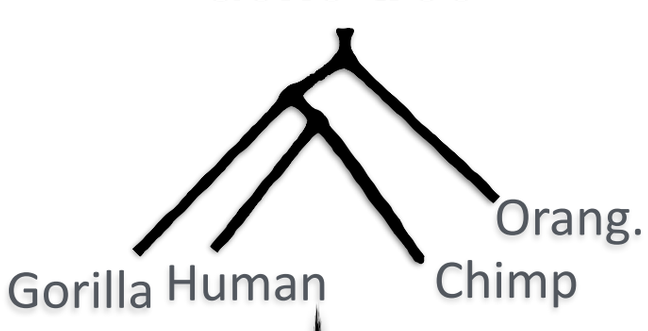


Gene evolution model $\mathbf{f(G|S)}$

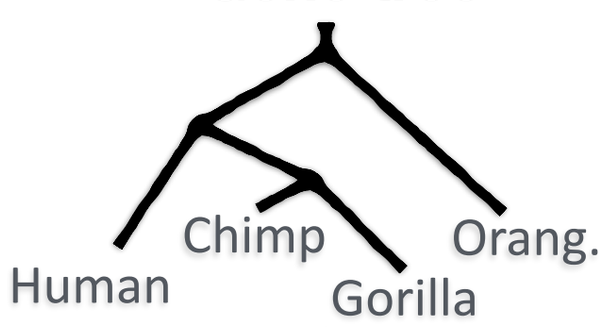
Gene tree



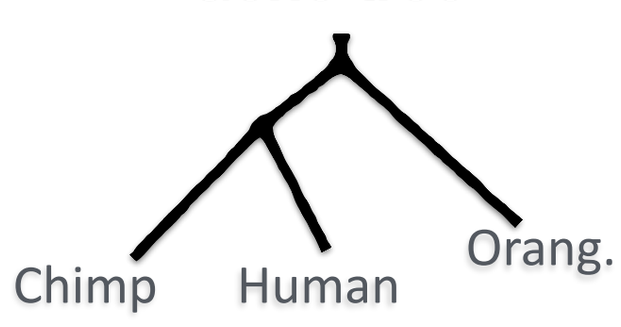
Gene tree



Gene tree



Gene tree



Sequence evolution model $\mathbf{f(D|G,\lambda)f(\lambda)}$

Sequence data (Alignments)

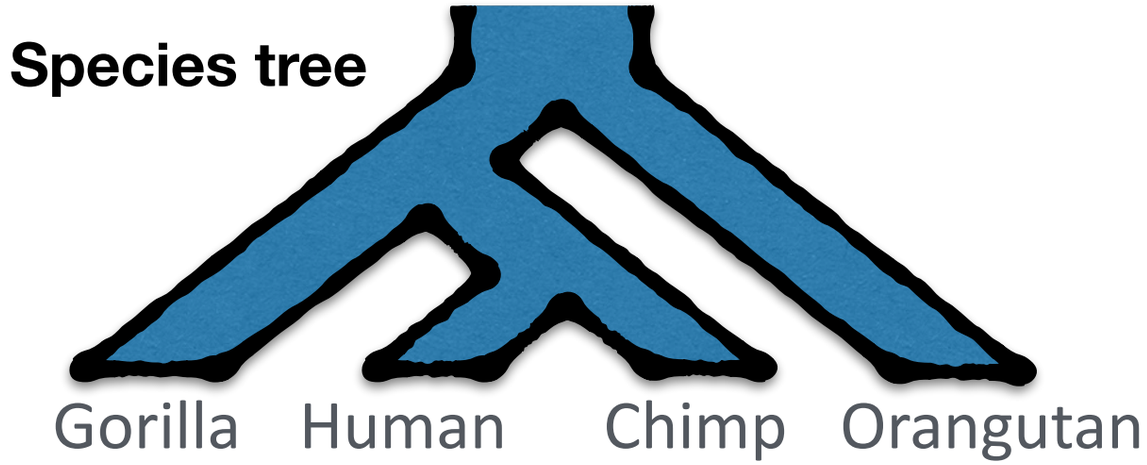
```
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
```

```
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
```

Sequence data (Alignments)

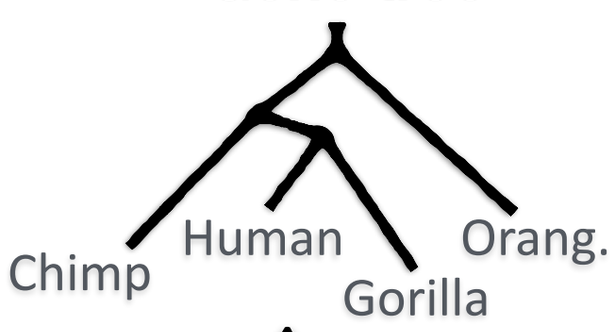
```
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
```

```
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT
```

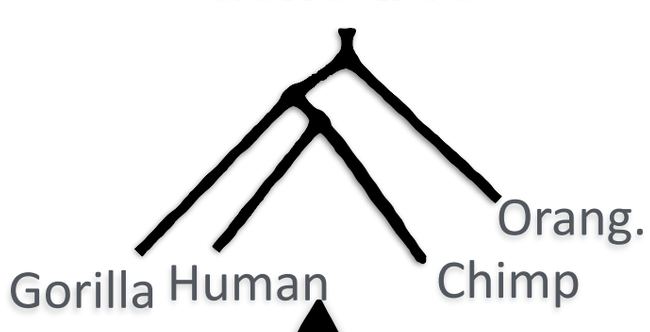


Gene evolution model $\mathbf{f(G|S)}$

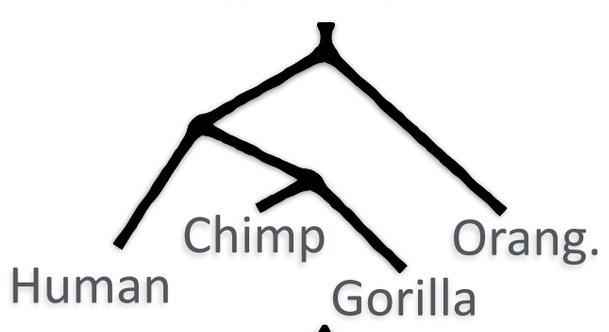
Gene tree



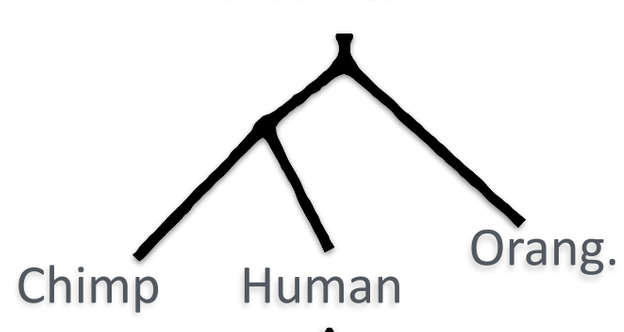
Gene tree



Gene tree



Gene tree



Sequence evolution model $\mathbf{f(D|G,\lambda)f(\lambda)}$

Sequence data (Alignments)

```
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
```

```
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
```

Sequence data (Alignments)

```
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
```

```
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT
```

1KP: Plant whole transcriptomes

[Wickett, Mirarab, *et al.*, PNAS, 2014]

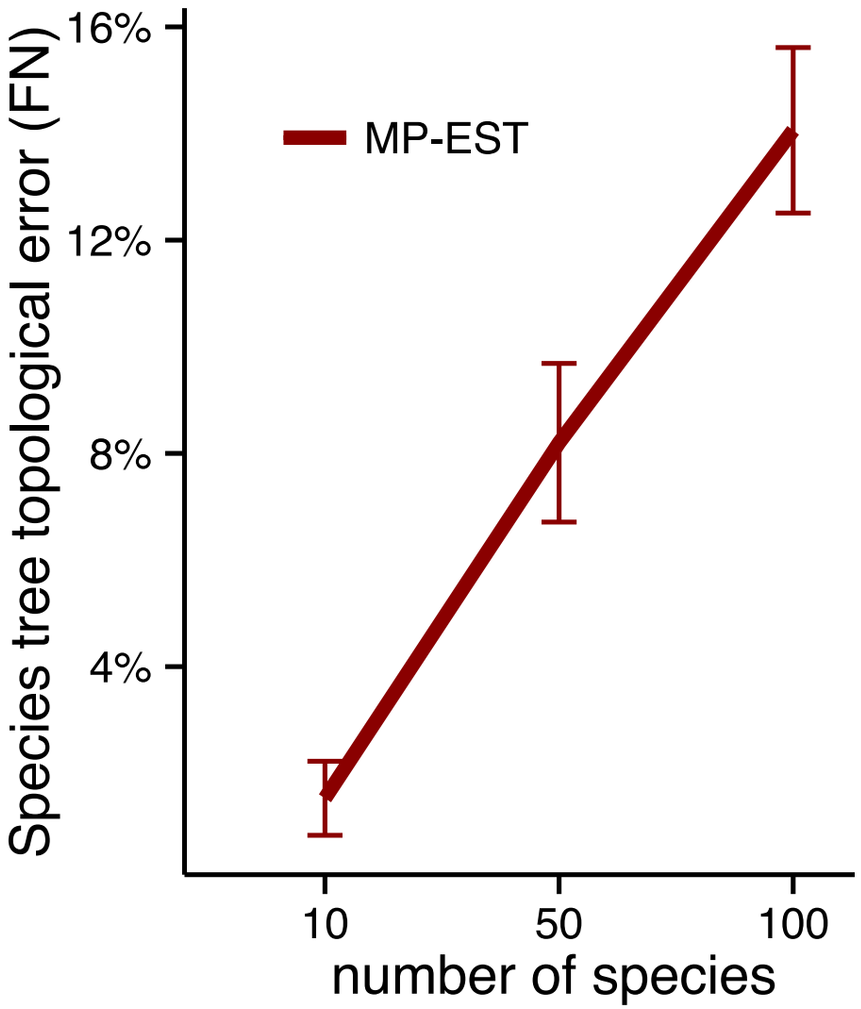
gane wong

华大基因 BGI

- Two phases
  - Pilot: Whole transcriptomes for 103 plant species and ~800 single copy genes
  - Capstone: Whole transcriptomes for 1178 plant species and ~410 single copy genes
- Spans all plants and algae: ~1 billion years of evolution
- Many unanswered questions about plant evolution

Jim leebens-mack

Number of species impacts error

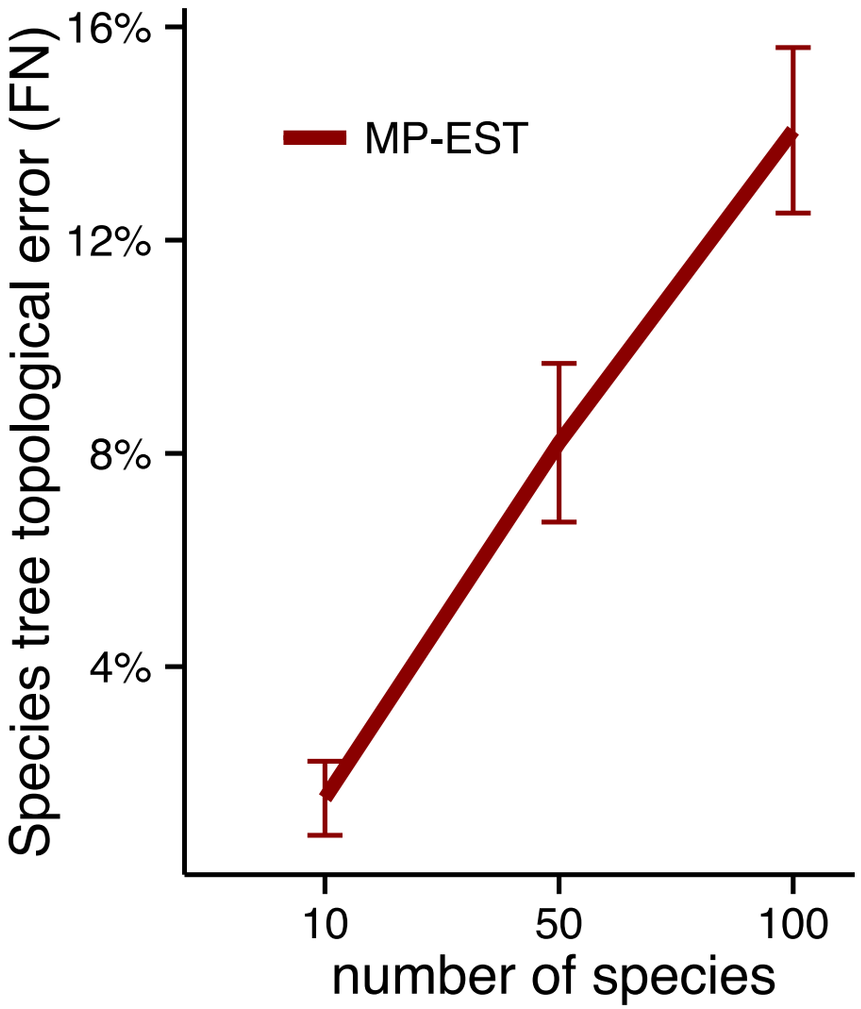


1000 genes, “medium” levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]

Our response: a new method called ASTRAL

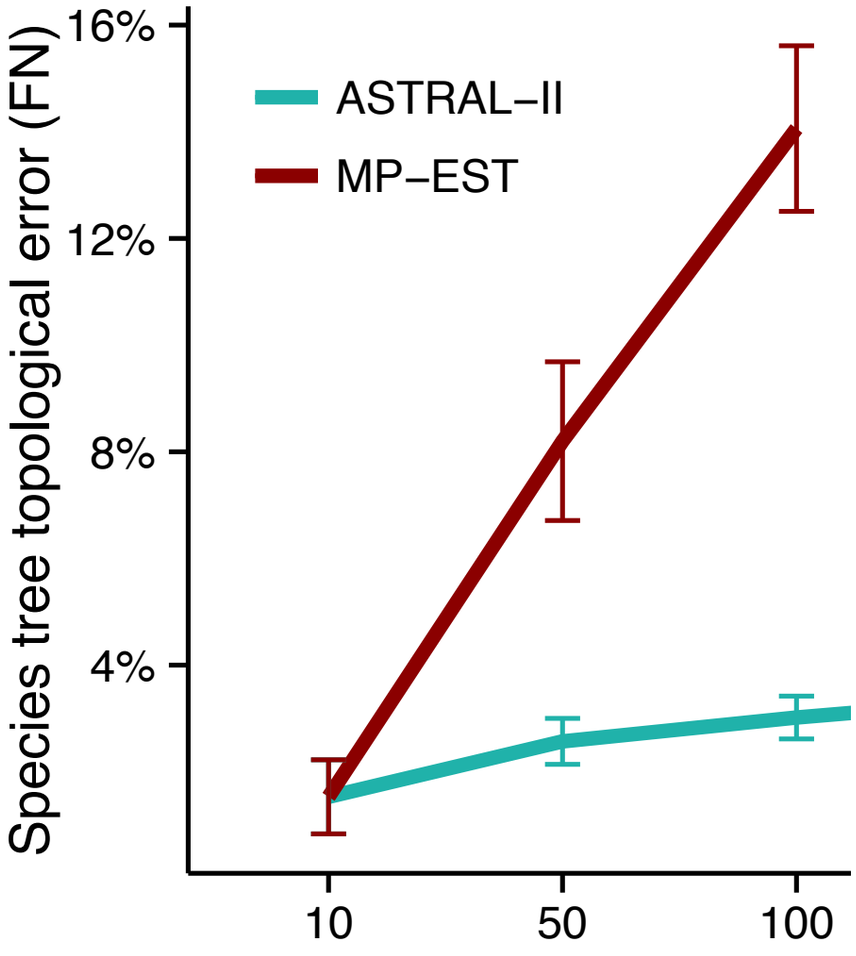
- Solves a simpler problem:
  - Instead of computing likelihood (difficult), it introduces a distance-based consistent estimator
- Uses divide-and-conquer
  - Each tree on n species can be broken into $\binom{n}{4}$ "quartet trees"

Number of species impacts error



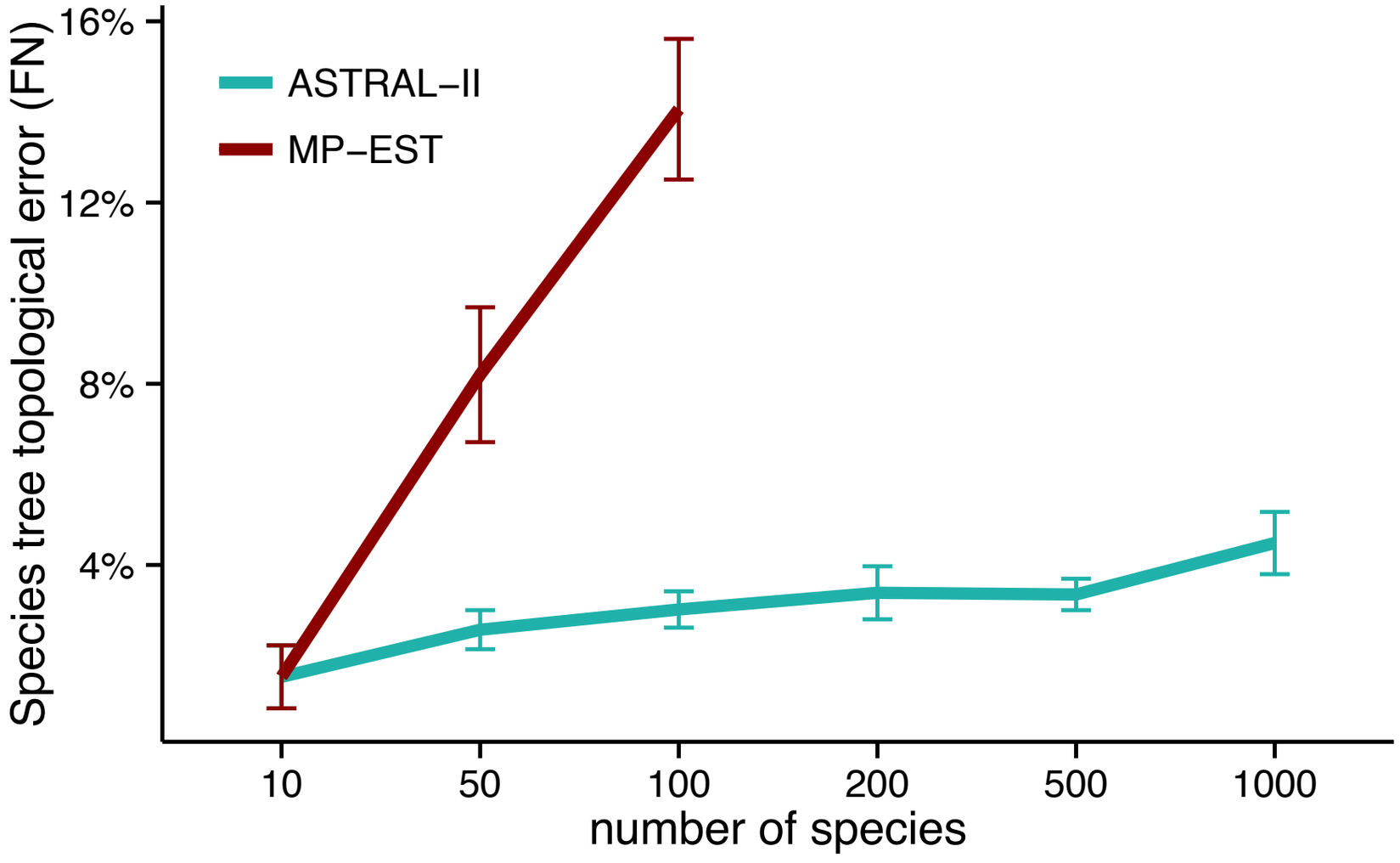
1000 genes, “medium” levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]

ASTRAL: accurate and scalable



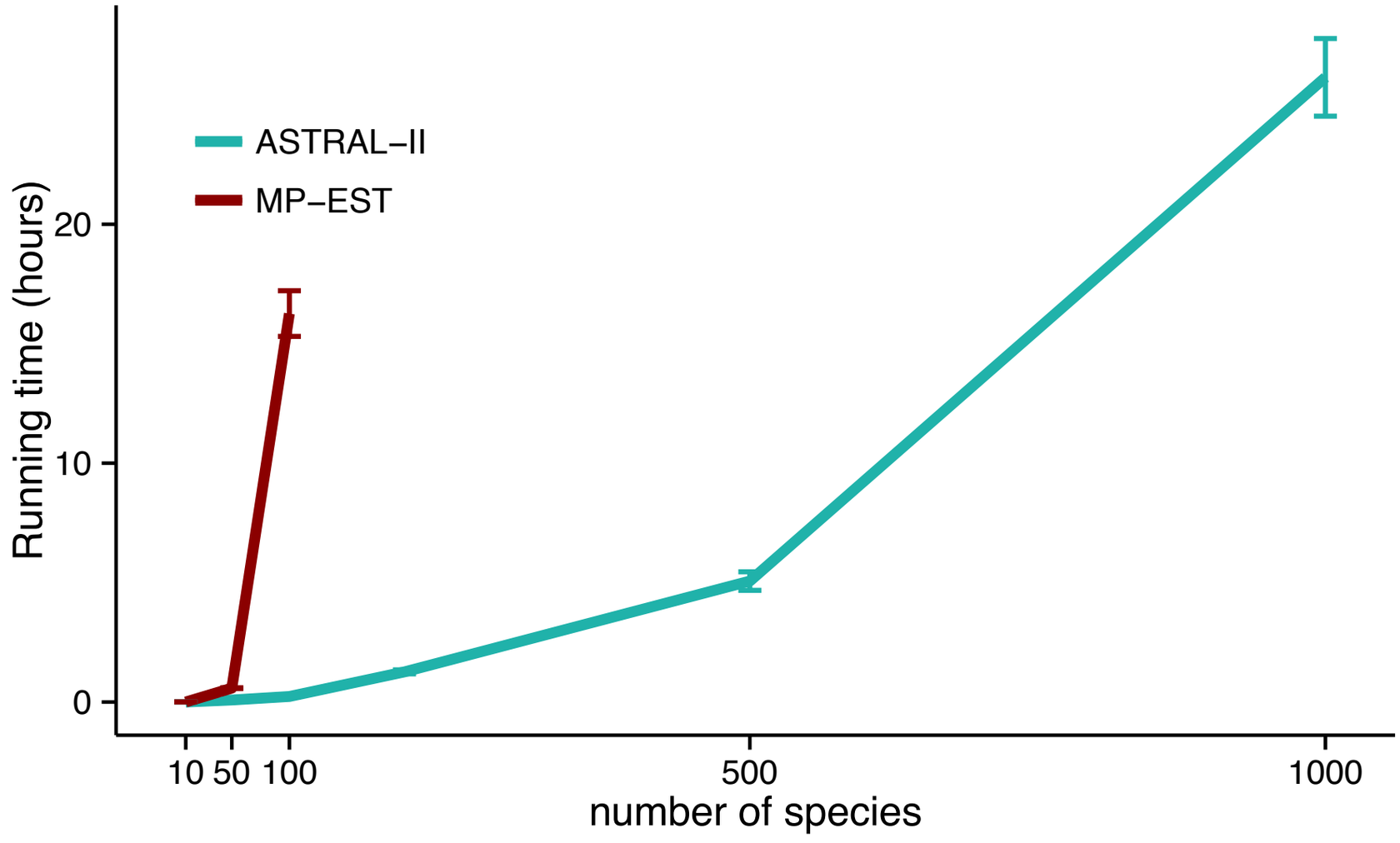
1000 genes, “medium” levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]

ASTRAL: accurate and scalable



1000 genes, “medium” levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]

Running time as function of # species



1000 genes, “medium” levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]

Maximum Quartet Support Species Tree

- Median tree problem:

Find the species tree with the maximum number of induced quartet trees shared with the collection of input gene trees

$$Score(T) = \sum_{1}^{m} |Q(T) \cap Q(t_i)|$$

the set of quartet trees induced by T

a gene tree

Maximum Quartet Support Species Tree

- Median tree problem:

Find the species tree with the maximum number of induced quartet trees shared with the collection of input gene trees

$$Score(T) = \sum_{1}^{m} |Q(T) \cap Q(t_i)|$$

the set of quartet trees induced by T

a gene tree

- **Theorem**: Statistically consistent under the multi-species coalescent model when solved exactly

ASTRAL

[Mirarab, et al., 2014] [Mirarab and Warnow, 2015] [Zhang et. al., 2018]

- Solve the Maximum Quartet Support problem exactly using dynamic programming

ASTRAL

[Mirarab, et al., 2014] [Mirarab and Warnow, 2015] [Zhang et. al., 2018]

- Solve the Maximum Quartet Support problem exactly using dynamic programming
- Constrains the search space using gene trees
- The constrained version remains statistically consistent
- Running time of the constrained version increases polynomially with the input size: $O(\mathcal{D}(nk)^{1.73})$

  $\mathcal{D} = O(nk)$ is the sum of degrees of all unique gene tree nodes

ASTRAL

[Mirarab, et al., 2014] [Mirarab and Warnow, 2015] [Zhang et. al., 2018]

- Solve the Maximum Quartet Support problem exactly using dynamic programming
- Constrains the search space using gene trees
  - The constrained version remains statistically consistent
  - Running time of the constrained version increases polynomially with the input size: $O(\mathcal{D}(nk)^{1.73})$

    $\mathcal{D} = O(nk)$ is the sum of degrees of all unique gene tree nodes
- Sample complexity: Can find the correct species tree with arbitrarily high probability given $O(\log(n) f^{-2})$ gene trees where f is the length of the shortest branch in the tree [Shekhar, Roch, Mirarab, TCBB, 2017]

ASTRAL on the “pilot” 1KP

[Wickett, Mirarab, *et al.*, PNAS, 2014]

- The ASTRAL tree had high support and was one of the accepted trees in the resulting publication
- ASTRAL took only 10 minutes on 103 taxa and 400 genes
- Solved many but not all questions of interest

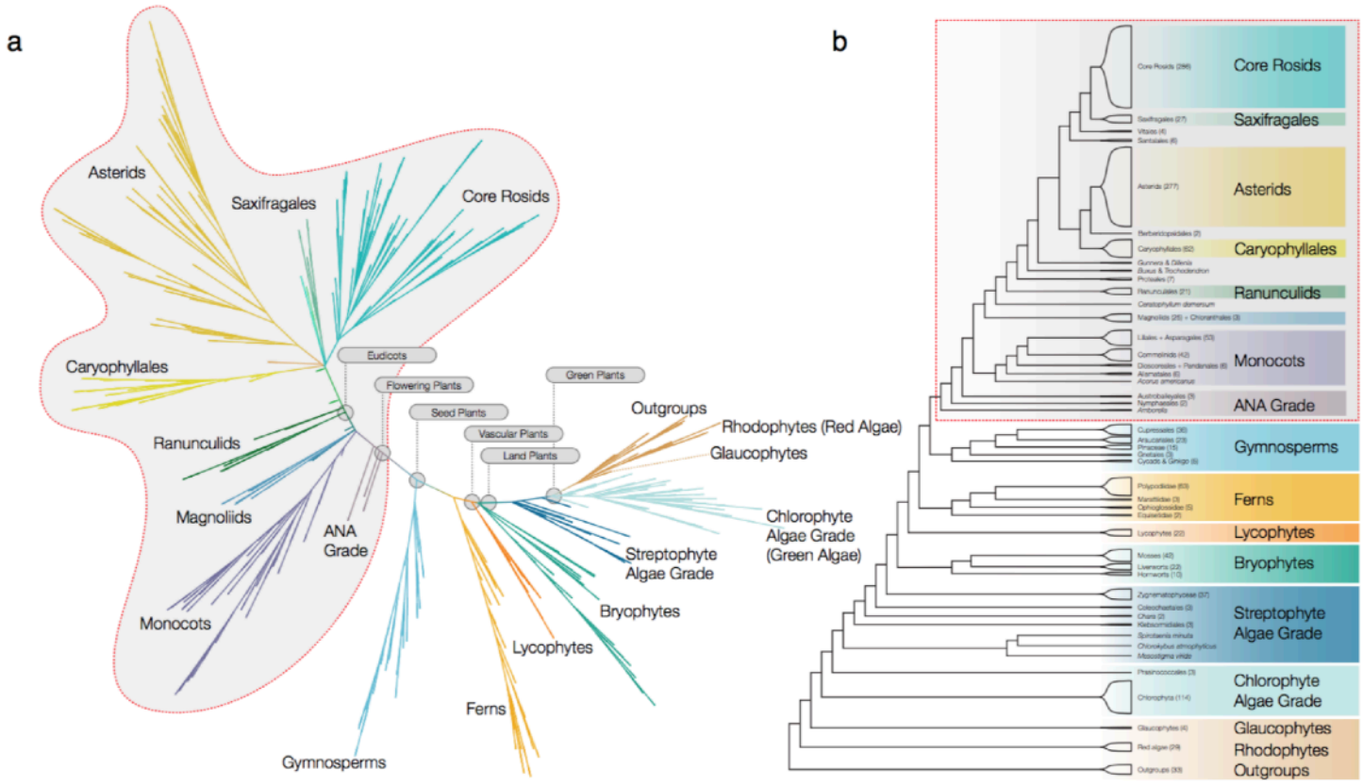
Phylotranscriptomic analysis of the origin and early diversification of land plants

Norman J. Wickett[a,b,1,2], Siavash Mirarab[c,1], Nam Nguyen[c], Tandy Warnow[c], Eric Carpenter[d], Naim Matasci[e,f], Saravanaraj Ayyampalayam[g], Michael S. Barker[f], J. Gordon Burleigh[h], Matthew A. Gitzendanner[h,i], Brad R. Ruhfel[h,j,k], Eric Wafula[l], Joshua P. Der[l], Sean W. Graham[m], Sarah Mathews[n], Michael Melkonian[o], Douglas E. Soltis[h,i,k], Pamela S. Soltis[h,i,k], Nicholas W. Miles[k], Carl J. Rothfels[p,q], Lisa Pokorny[p,r], A. Jonathan Shaw[p], Lisa DeGironimo[s], Dennis W. Stevenson[s], Barbara Surek[o], Juan Carlos Villarreal[t], Béatrice Roure[u], Hervé Philippe[u,v], Claude W. dePamphilis[l], Tao Chen[w], Michael K. Deyholos[d], Regina S. Baucom[x], Toni M. Kutchan[y], Megan M. Augustin[y], Jun Wang[z], Yong Zhang[v], Zhijian Tian[z], Zhixiang Yan[z], Xiaolei Wu[z], Xiao Sun[z], Gane Ka-Shu Wong[d,z,aa,2], and James Leebens-Mack[g,2]

ASTRAL-III on the capstone 1KP

[unpublished, 2018, 2019]

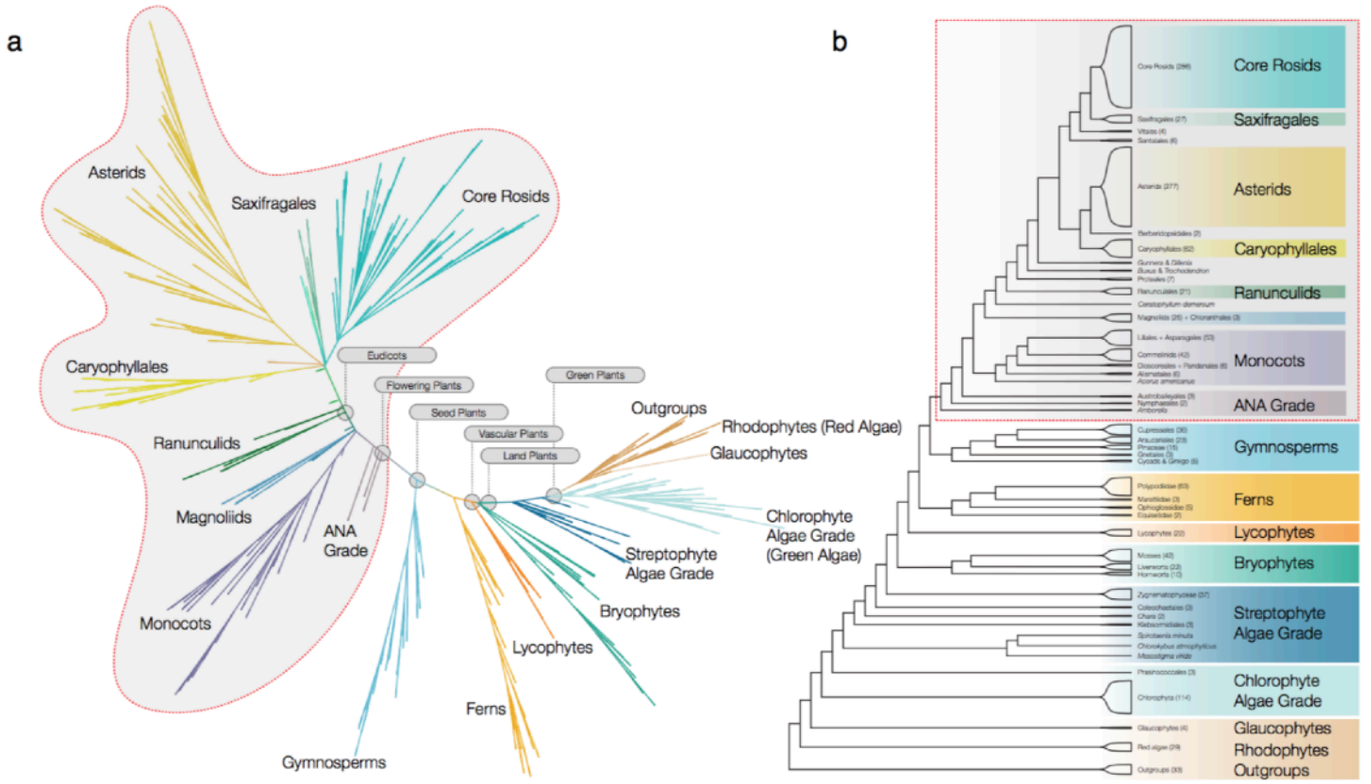
- ASTRAL is the main tree used
  - Took eight hours for 1178 species and 410 genes

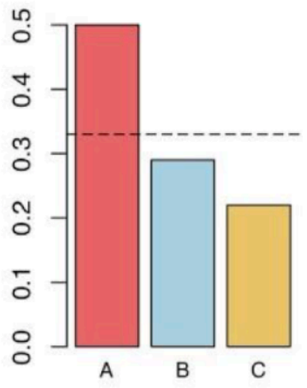


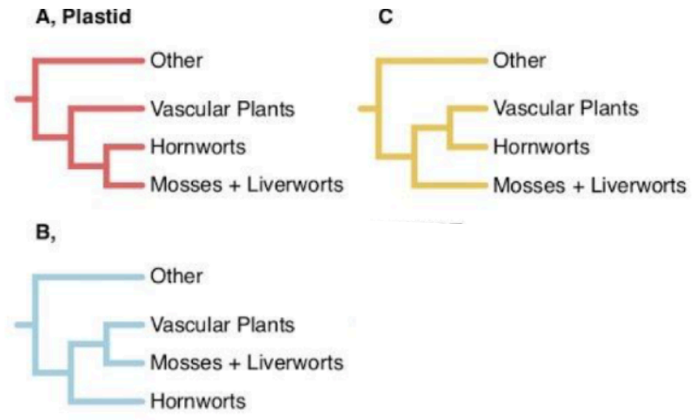
ASTRAL-III on the capstone 1KP

[unpublished, 2018, 2019]

- ASTRAL is the main tree used
  - Took eight hours for 1178 species and 410 genes
- Sheds light on trait evolution and differences among genes





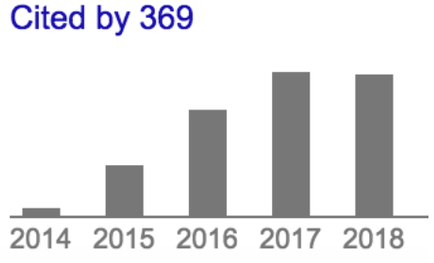


ASTRAL used widely

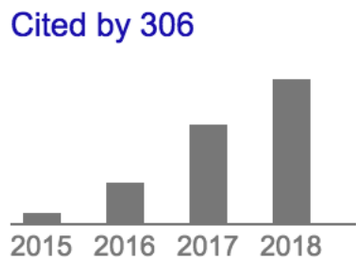
Early use:

- Plants: Wickett, et al., 2014, PNAS
- Birds: Prum, et al., 2015, Nature
- Xenoturbella, Cannon et al., 2016, Nature
- Xenoturbella, Rouse et al., 2016, Nature
- Flatworms: Laumer, et al., 2015, eLife
- Shrews: Giarla, et al., 2015, Syst. Bio.
- Frogs: Yuan et al., 2016, Syst. Bio.
- Tomatoes: Pease, et al., 2016, PLoS Bio.
- Angiosperms: Huang et al., 2016, MBE
- Worms: Andrade, et al., 2015, MBE

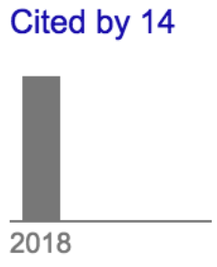
ASTRAL

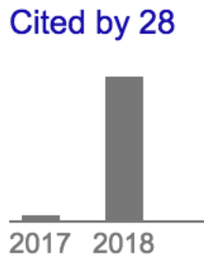


ASTRAL-II



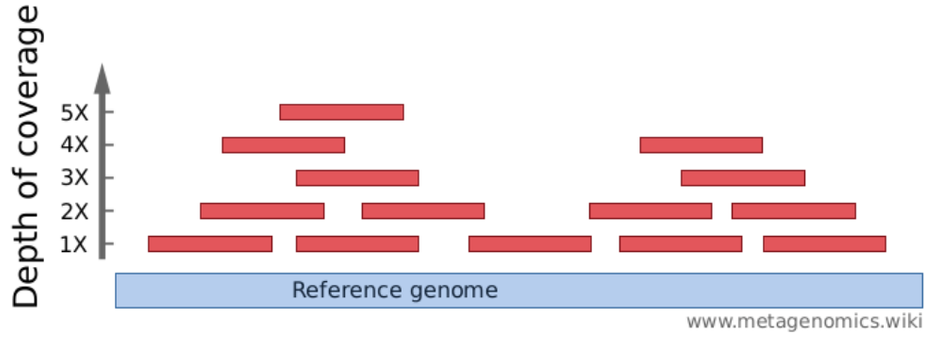
ASTRAL-III





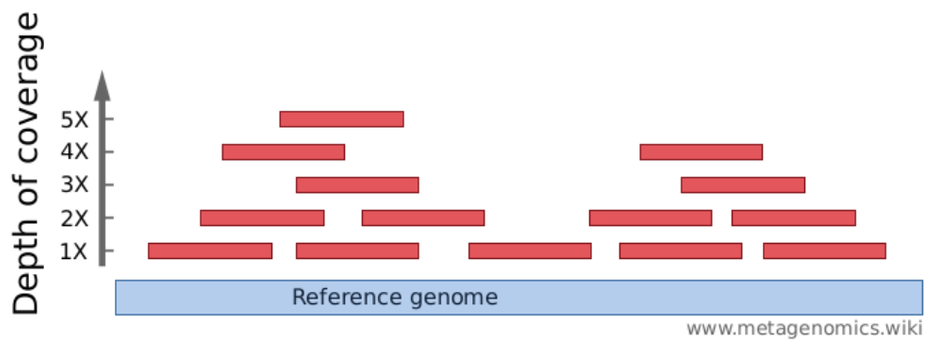
How do we get genome-wide data?

- **Whole-genome** sequencing: very expensive but complete



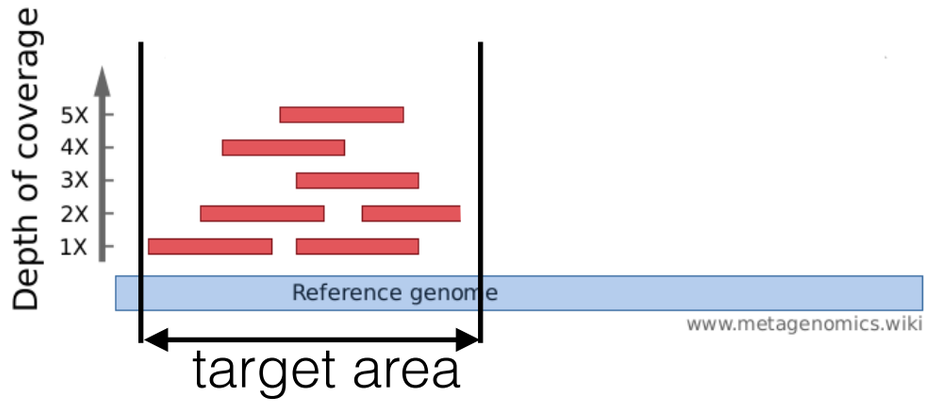
How do we get genome-wide data?

- **Whole-genome** sequencing: very expensive but complete



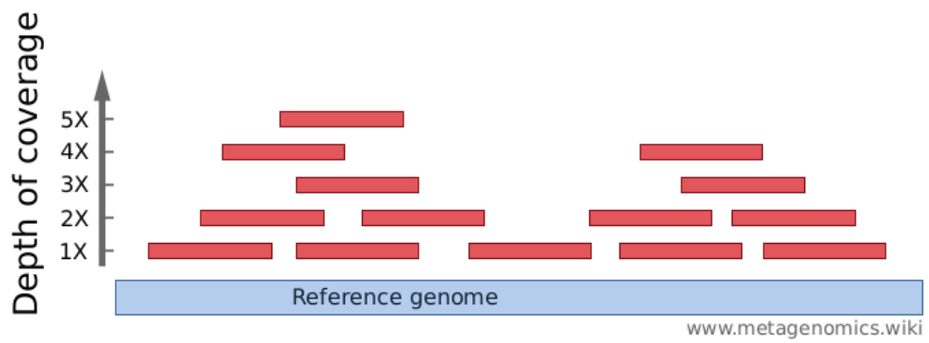
- **Transcriptome** sequencing: still expensive but much less data

- **Targeted capture**: a bit less expensive but less useful data



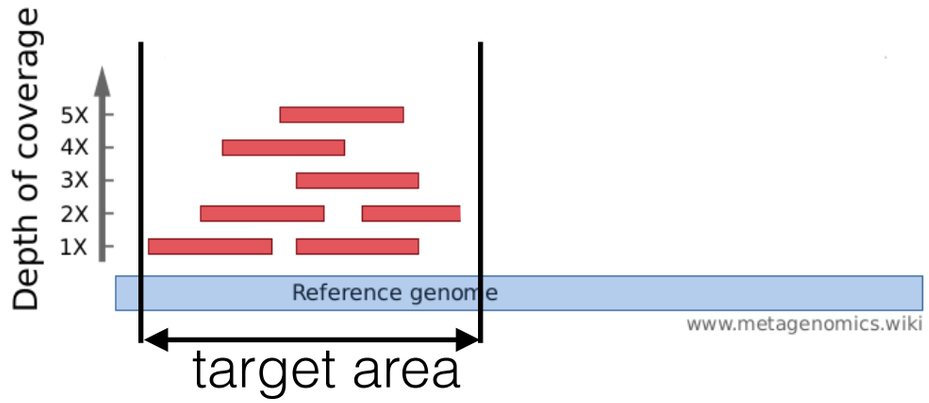
How do we get genome-wide data?

- **Whole-genome** sequencing: very expensive but complete

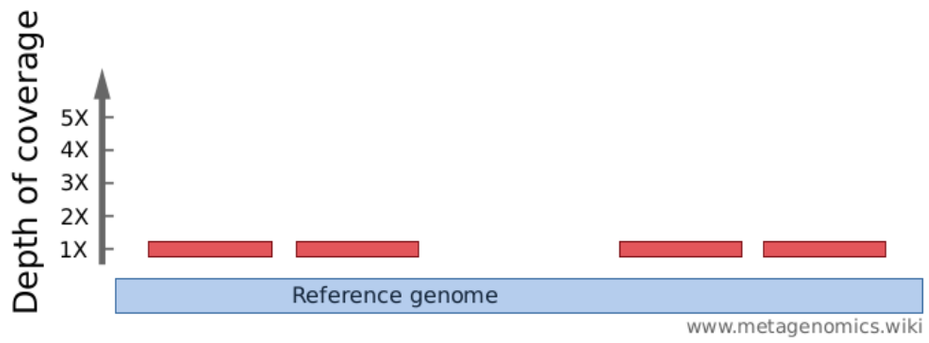


- **Transcriptome** sequencing: still expensive but much less data

- **Targeted capture**: a bit less expensive but less useful data



- **Genome skimming**: cheap ($50) but little data and hard to use



Genome skimming

- Sequence genomes at very low coverage
  - Not enough for assemblies
  - May be enough for sample identification: find the closest species to a query skim
  - Build phylogenetic trees with reduced cost

Tom Gilbert, Copenhagen

guojie zhang

华大基因 BGI

Genome skimming

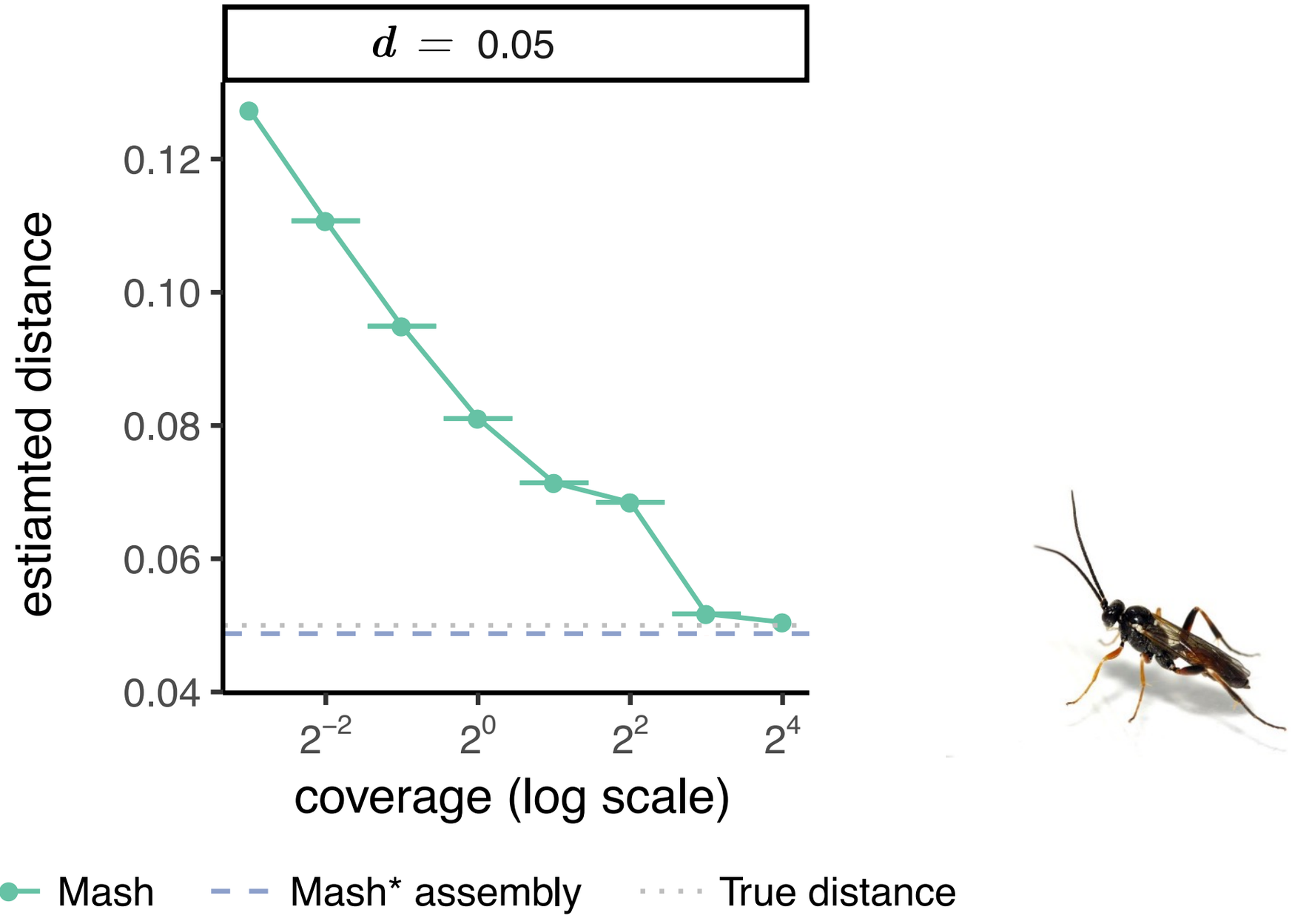
- Sequence genomes at very low coverage
  - Not enough for assemblies
  - May be enough for sample identification: find the closest species to a query skim
  - Build phylogenetic trees with reduced cost
- Both applications boil down to computing distances between genomes

Tom Gilbert, Copenhagen

guojie zhang

华大基因 BGI

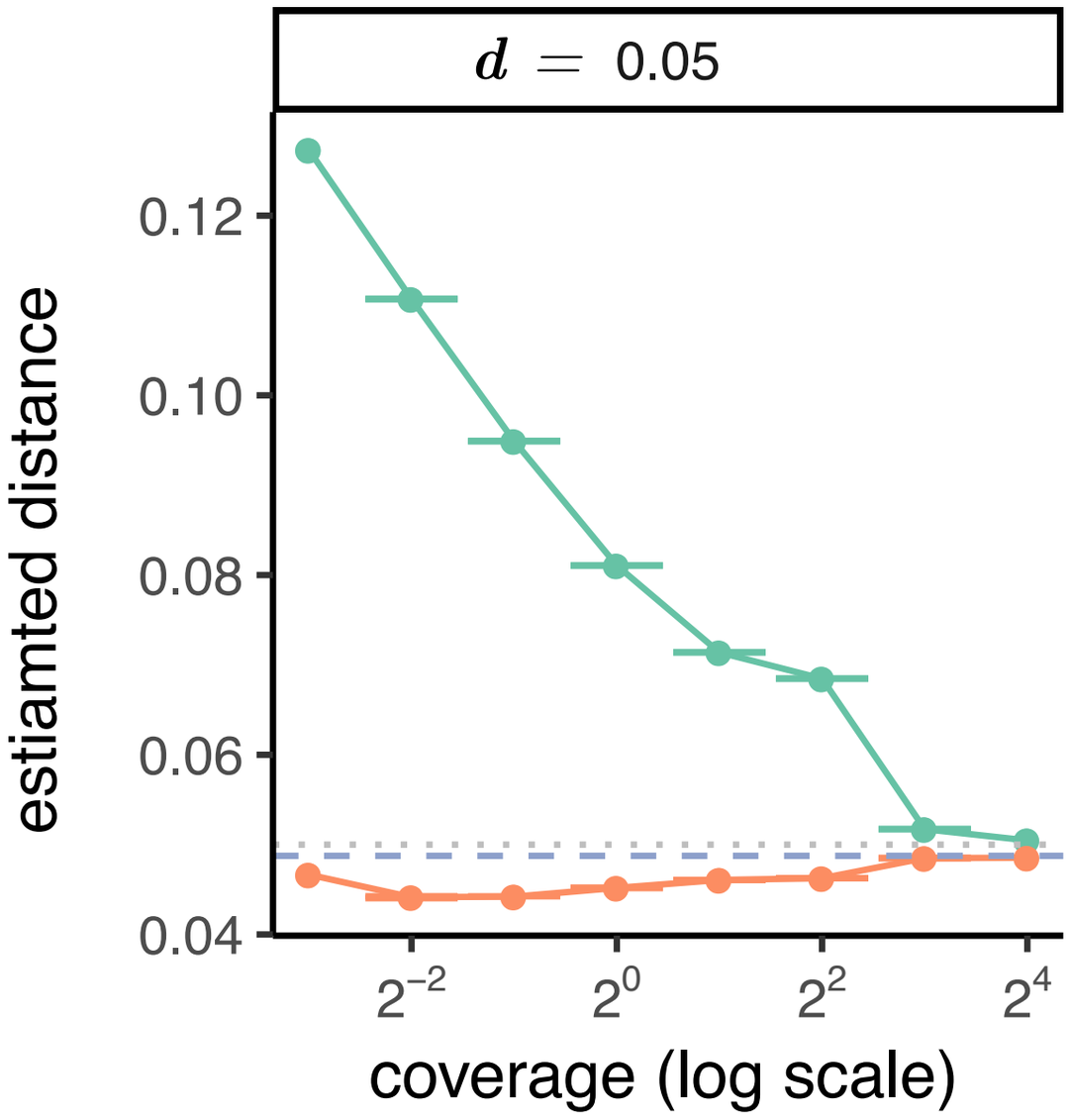
What coverage is needed?

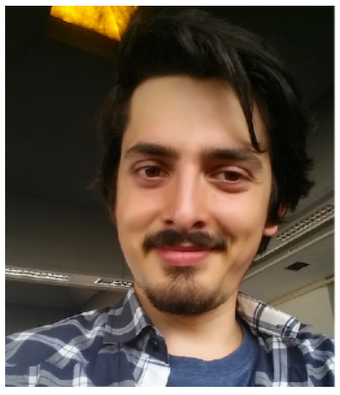


[Sarmashghi, 2018]

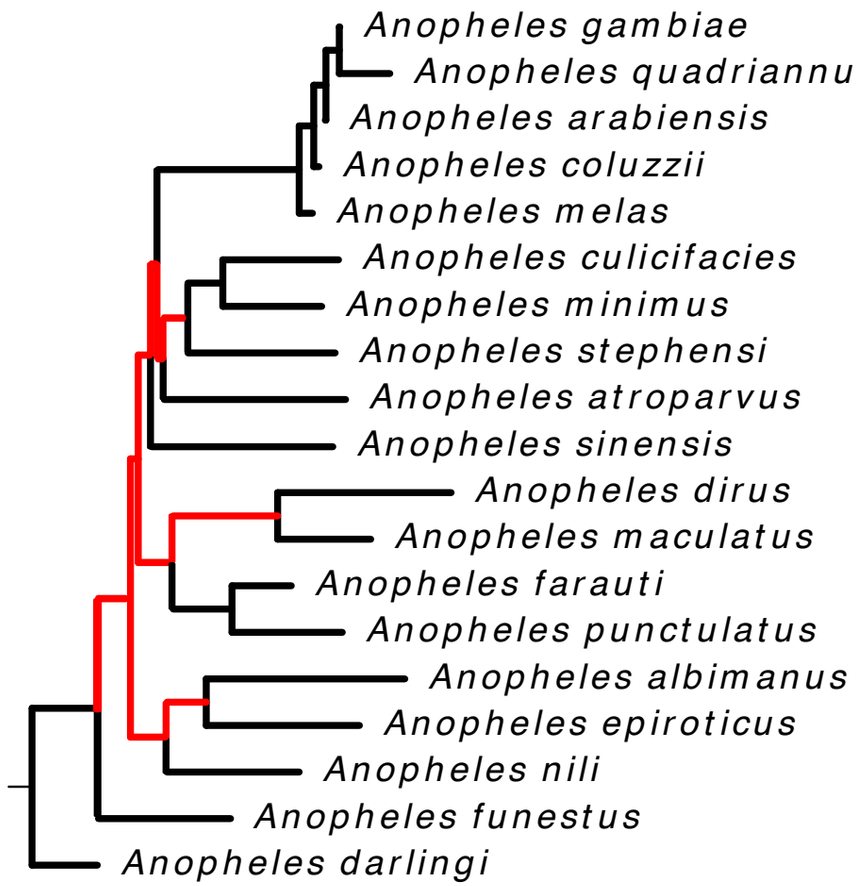
Skmer: a better translation from J to D

[Sarmashghi, *et al,* Genome biolog, 2018]

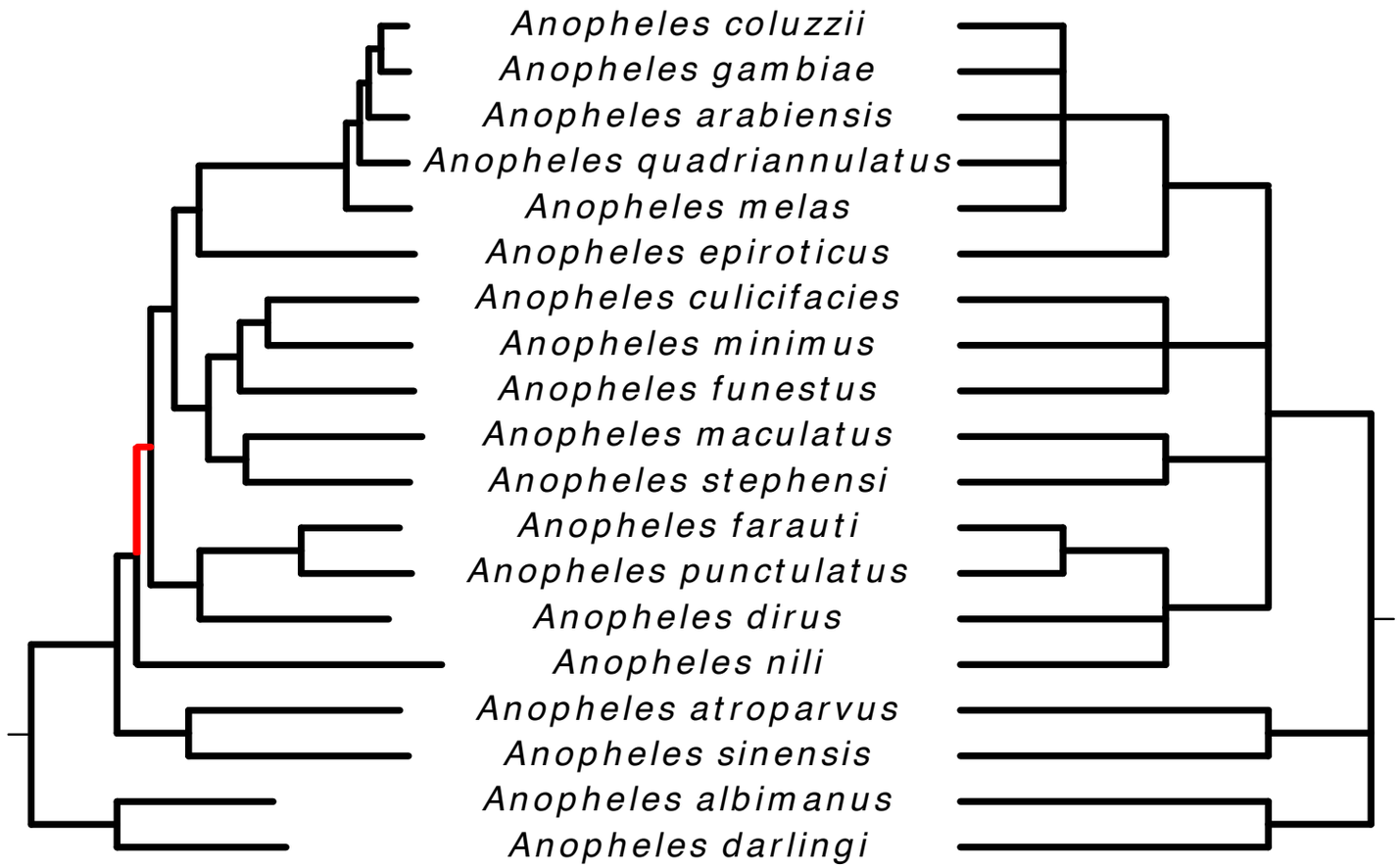




Building Phylogenies using Skimmed data



COI tree



Skmer tree (0.1Gb – 1Gb)

Reference tree (oTL)

[Sarmashghi, 2018]

Conclusions

- Evolution can be reconstructed using genomic data
  - It ENABLEs answering basic science questions in biology and has many downstream applications
- We constantly get bigger and bigger data, but
  - With more data, we often want to solve harder problems
  - Large data increase the running time but *can* reduce accuracy of methods not designed for large size
  - There is a need for developing new algorithms that can handle large dataset

Acknowledgments

Tandy Warnow

Siavash Mirarab

S.M. Bayzid

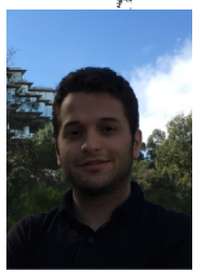
Nam Nguyen (now at UIUC)

Jim Leebens-mack (UGA)

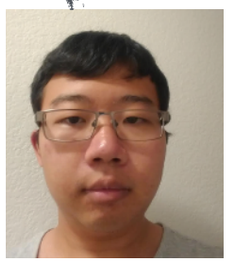
Norman Wickett (U Chicago)

Gane Wong (U of Alberta)

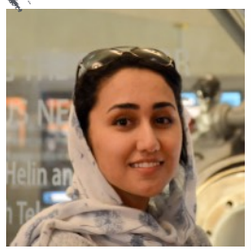
. . .



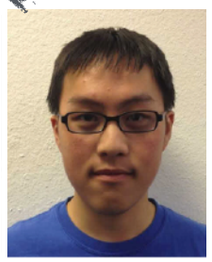
Erfan Sayyari



Chao Zhang



Maryam Hashemi



John Yin

Uyen Mai

Guojie Zhang (BGI, China)

Tom Gilbert (U Copenhagen)

Erich Jarvis (Duke, HMMI)

Bastien Boussau (Université Lyon)

Ed Braun (U Florida)

HHMI HOWARD HUGHES MEDICAL INSTITUTE

NSERC CRSNG

ALFRED P. SLOAN FOUNDATION

NSF

华大基因 BGI