Concatenation and Partition Files

Jacob L. Steenwyk

jlsteenwyk.github.io @JLSteenwyk

Adapted from Rosa Fernández, Lisa Pokorny & Marina Marcet-Houben

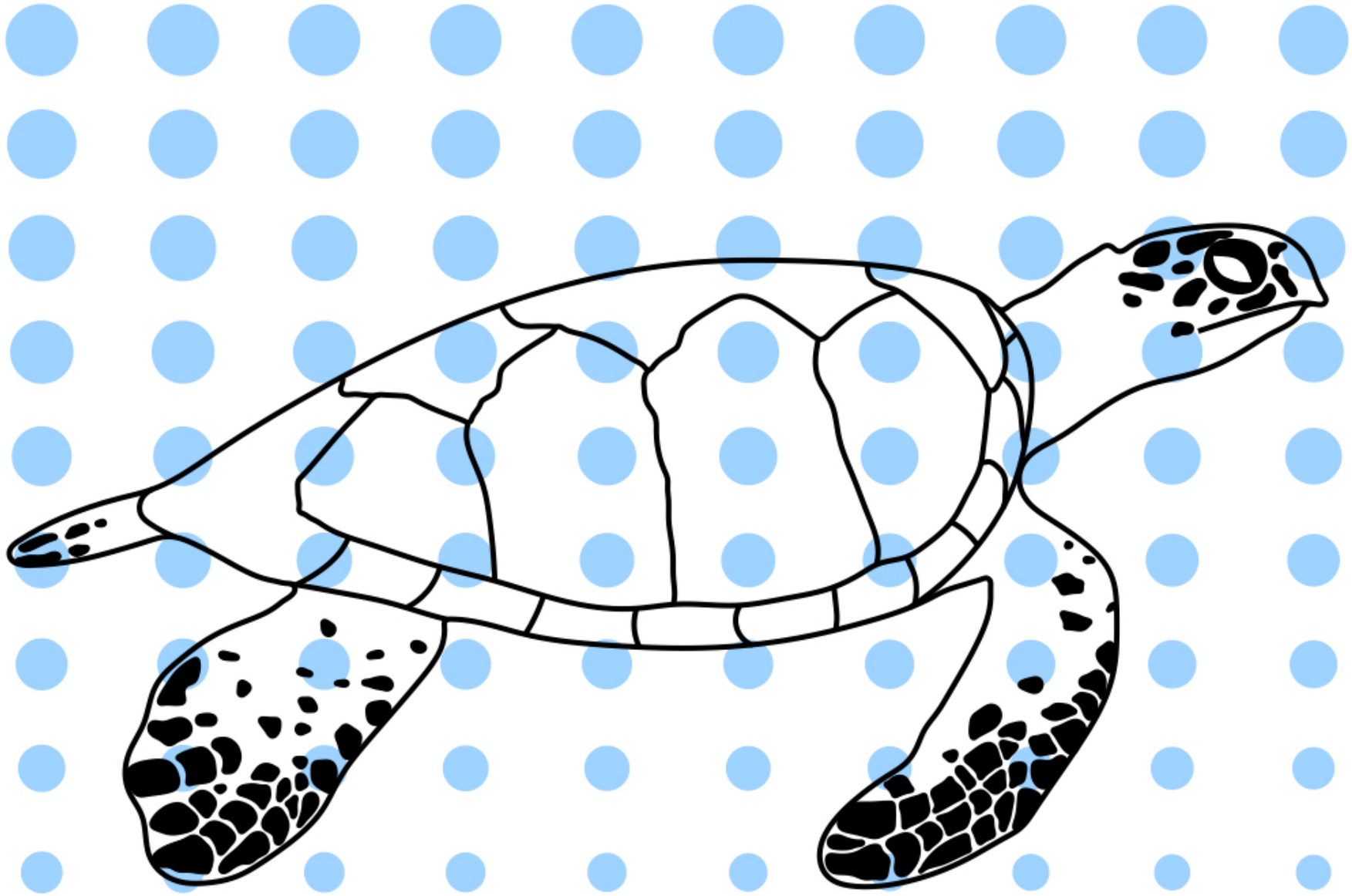

Who are you?


Who are you?
Dude,
I don't know

Who am I?

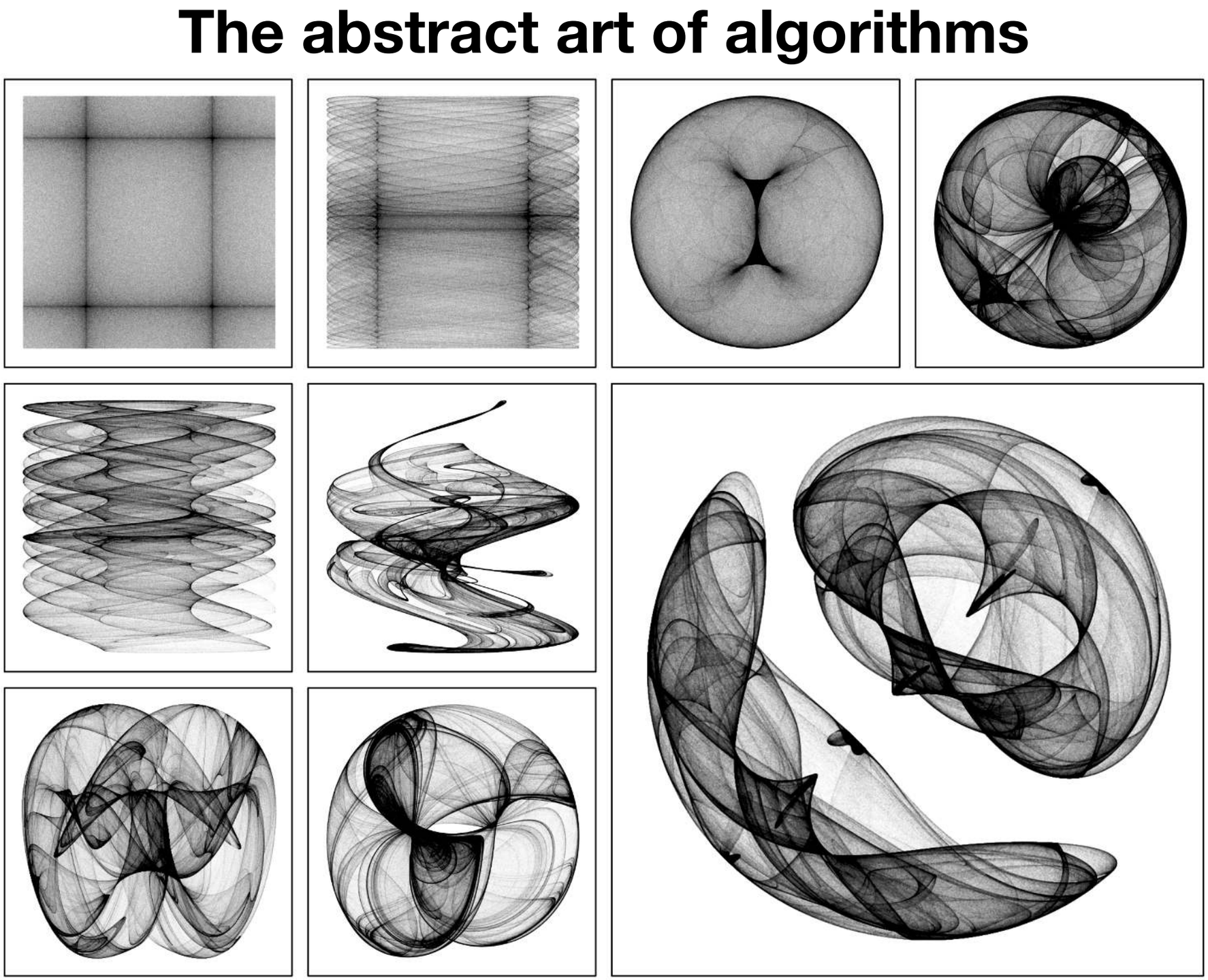
- Graphic artist

The critically endangered



Who am I?

- Graphic artist



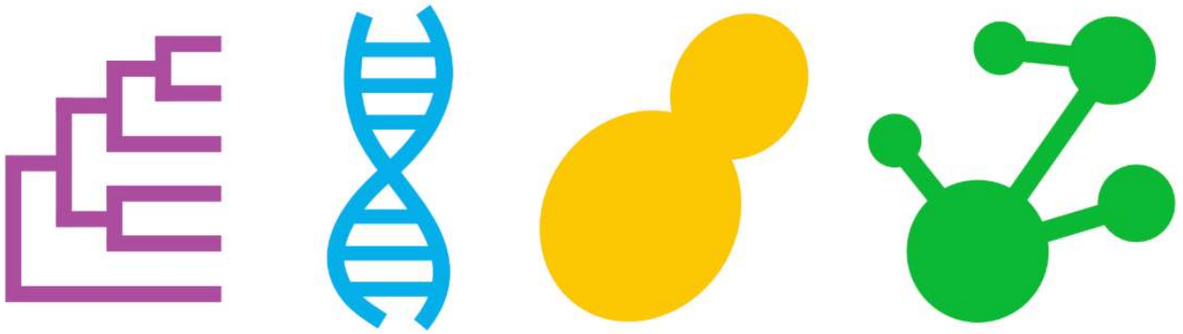
Who am I?

- Graphic artist
- Musician

Singer Songwriter

Disc Jockey

jlsteenwyk.github.io



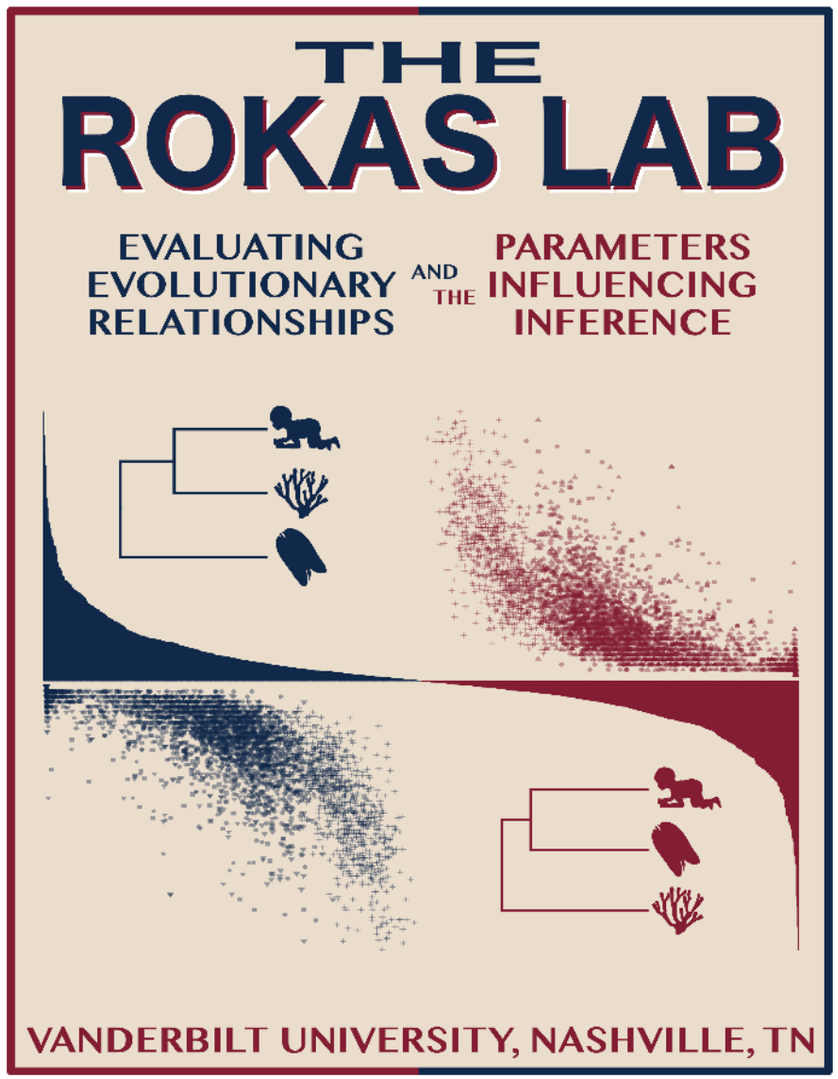
Biological scientist, educator, artist.

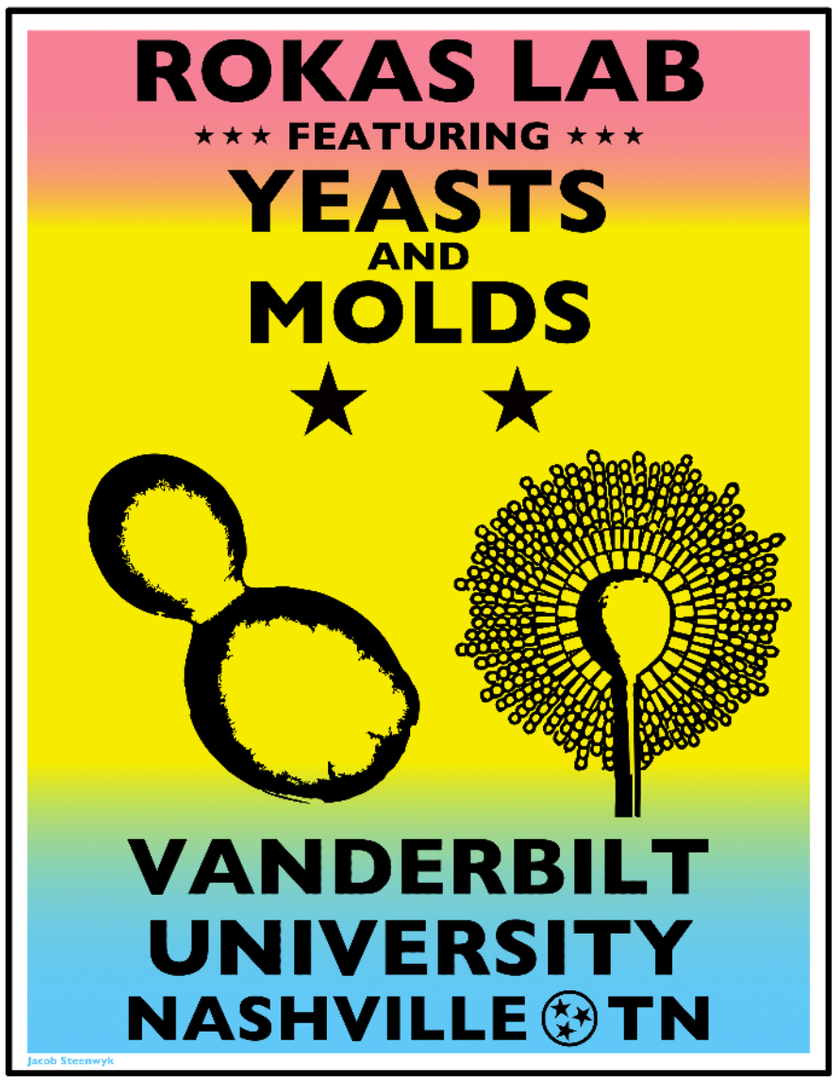
Genome evolution of medically and technologically important fungi.

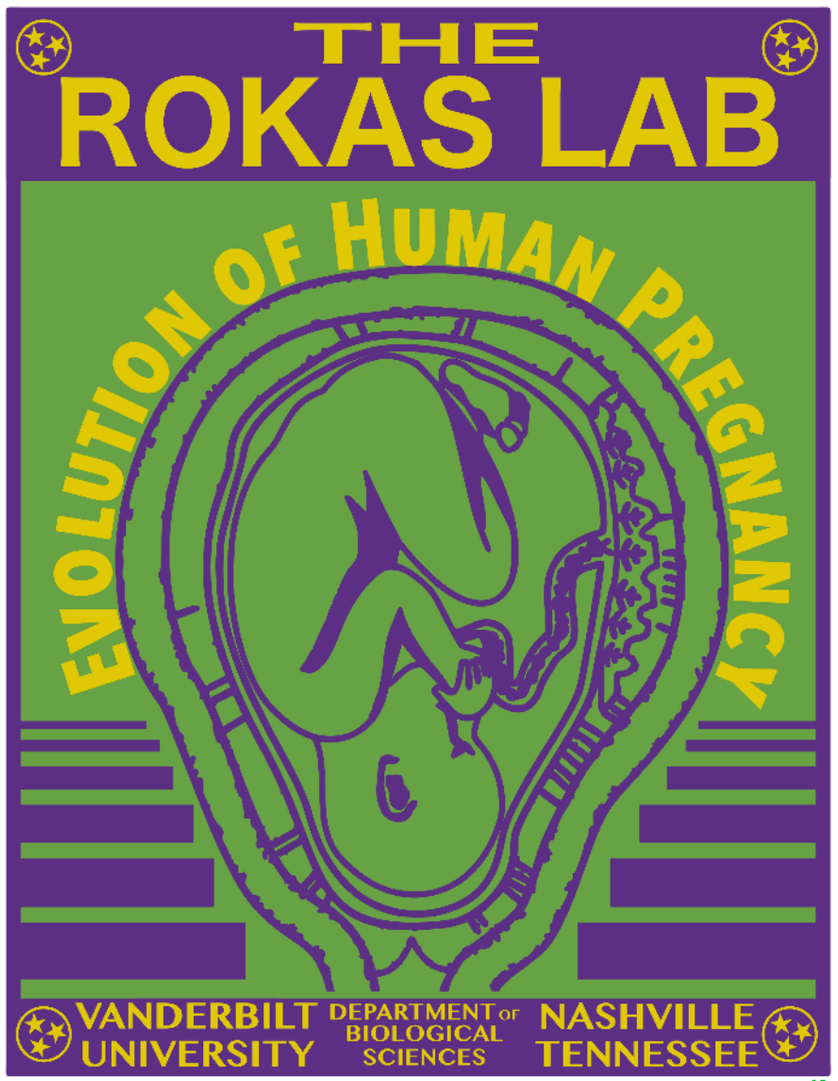
Education advocate aiming to make science more accessible to all.

Who am I?

- Graphic artist
- Musician
- Scientist



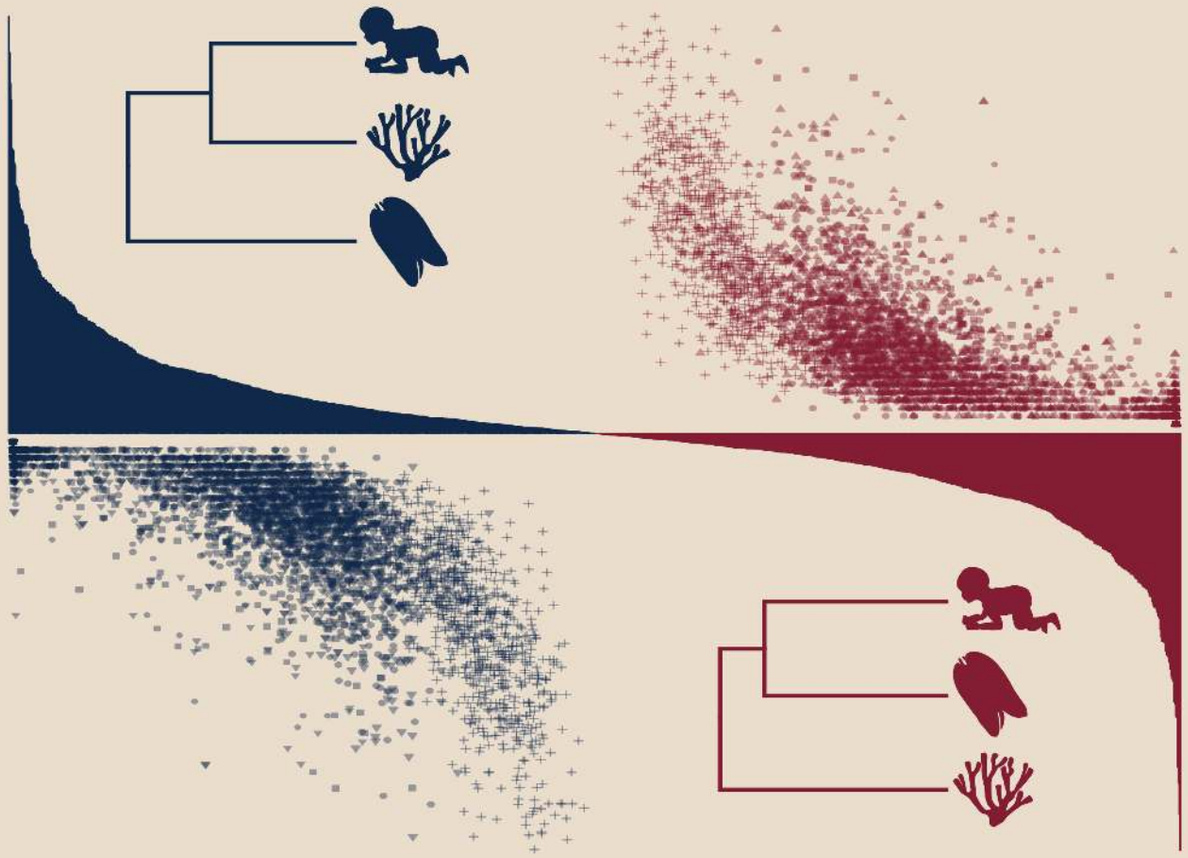




Designed by Jacob Steenwyk

THE ROKAS LAB

EVALUATING EVOLUTIONARY RELATIONSHIPS AND THE PARAMETERS INFLUENCING INFERENCE

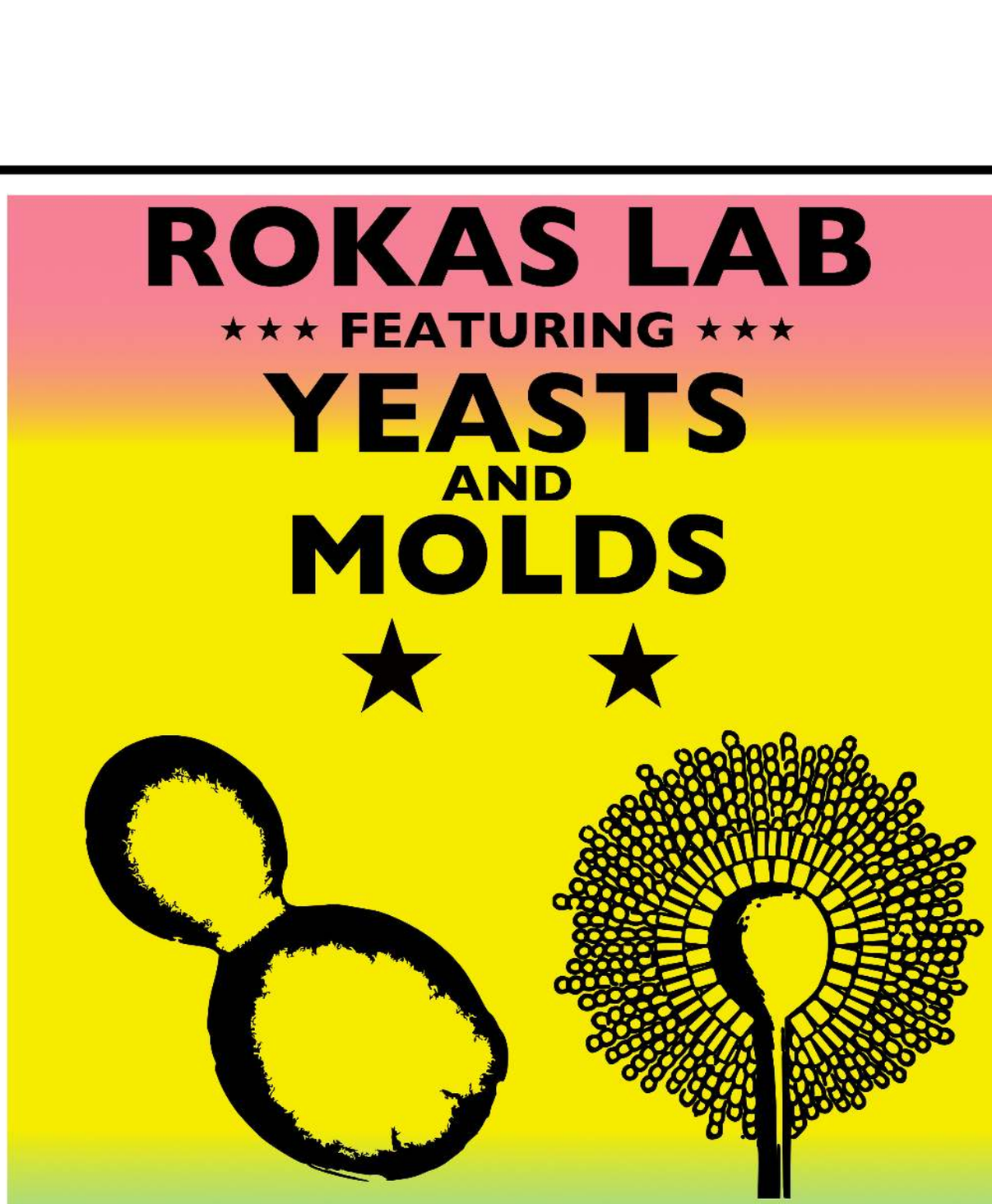


VANDERBILT UNIVERSITY, NASHVILLE, TN

ROKAS LAB

★★★ FEATURING ★★★

YEASTS AND MOLDS

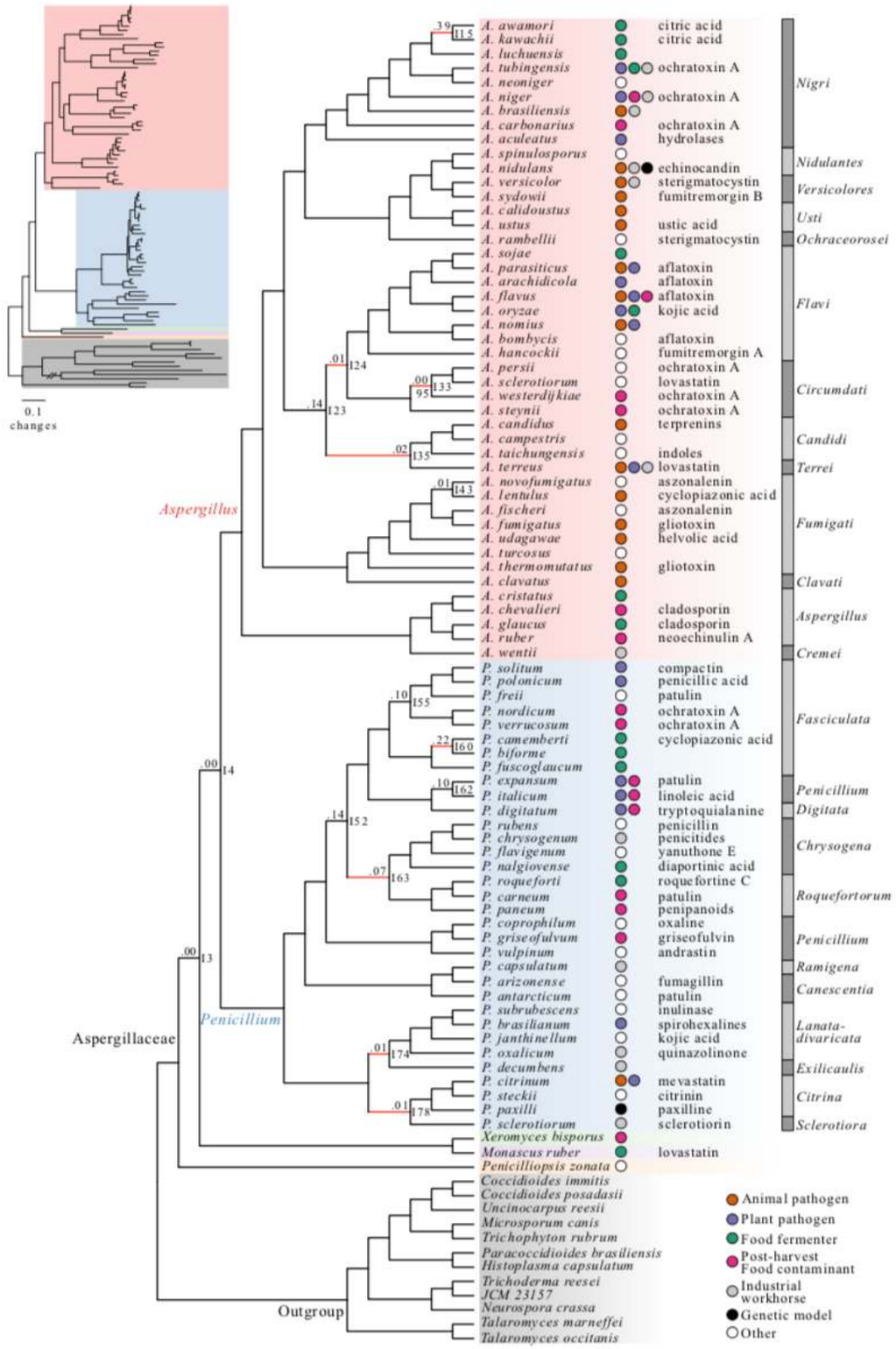


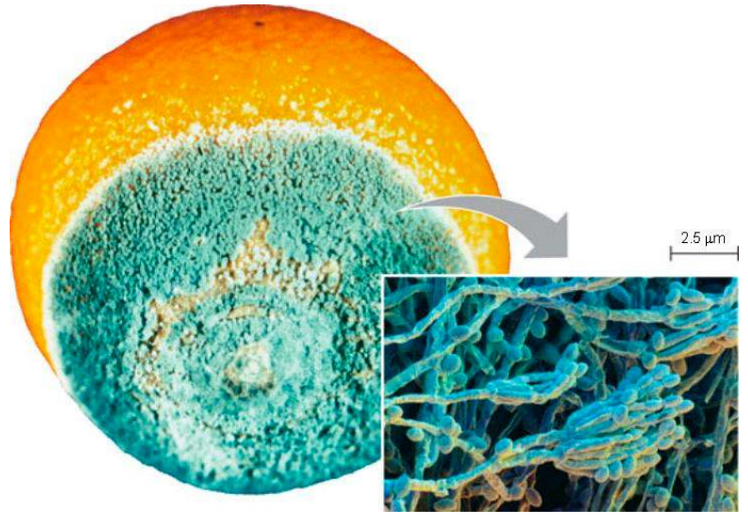
VANDERBILT UNIVERSITY

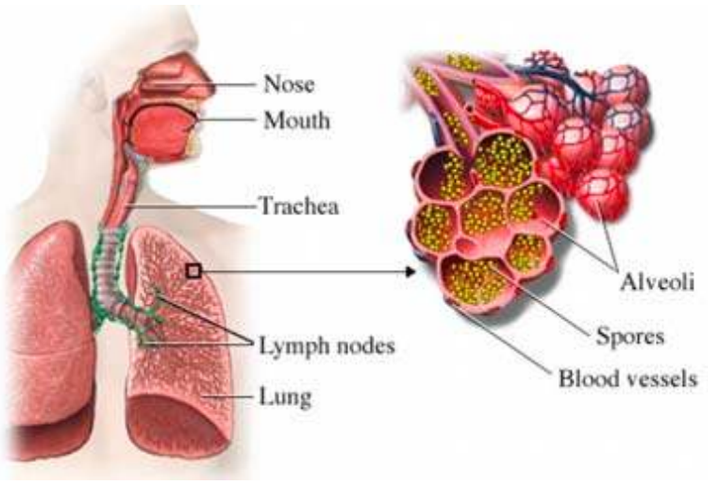
NASHVILLE TN

Jacob Steenwyk

81 genomes from mainly *Aspergillus and Penicillium*

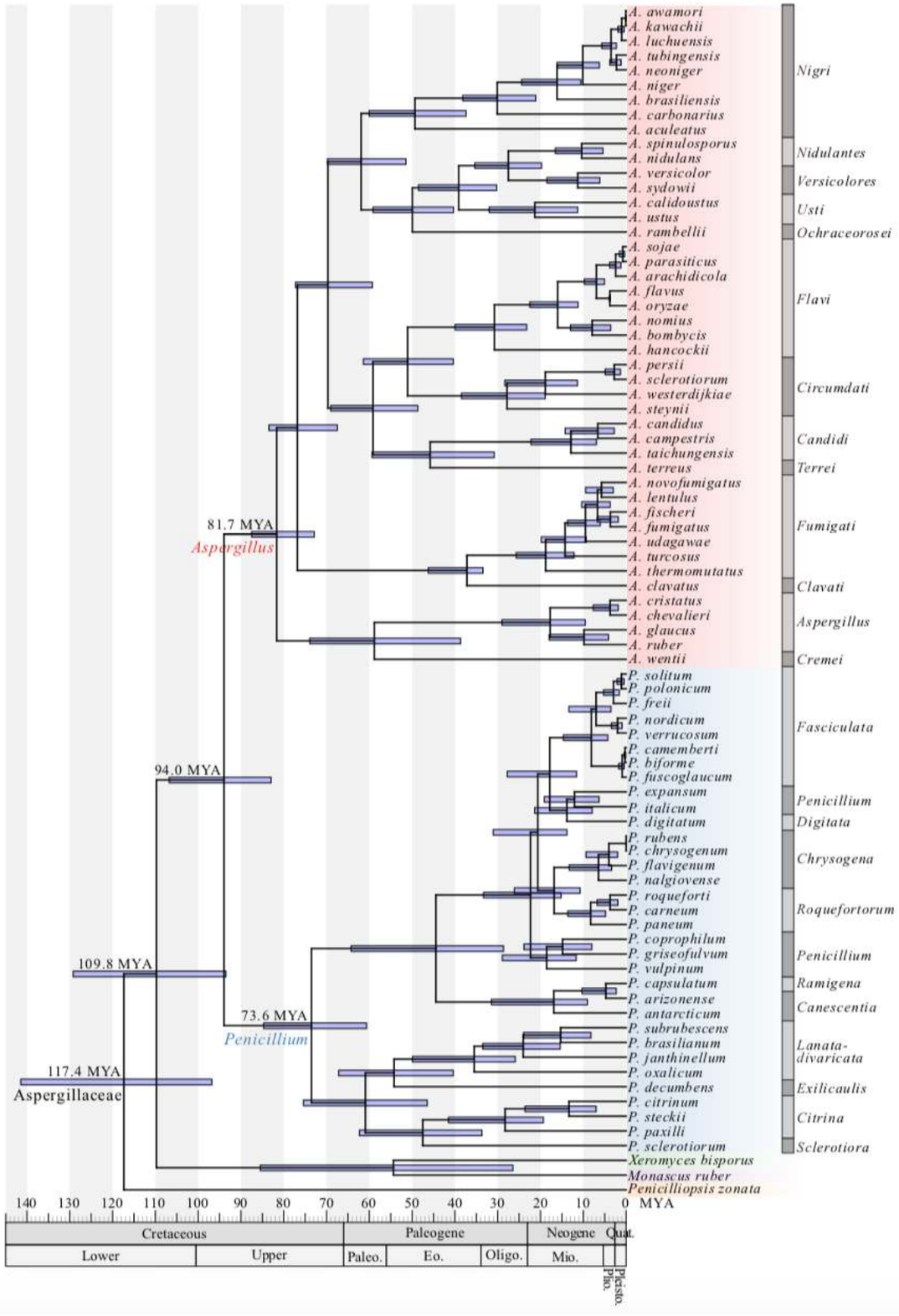


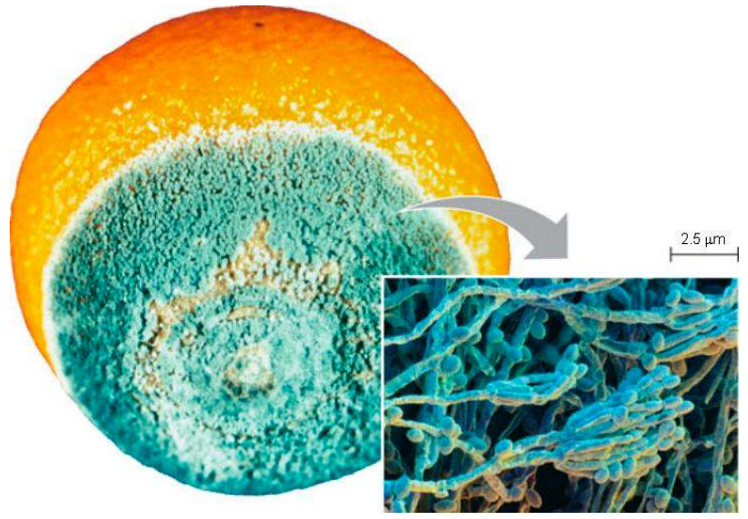


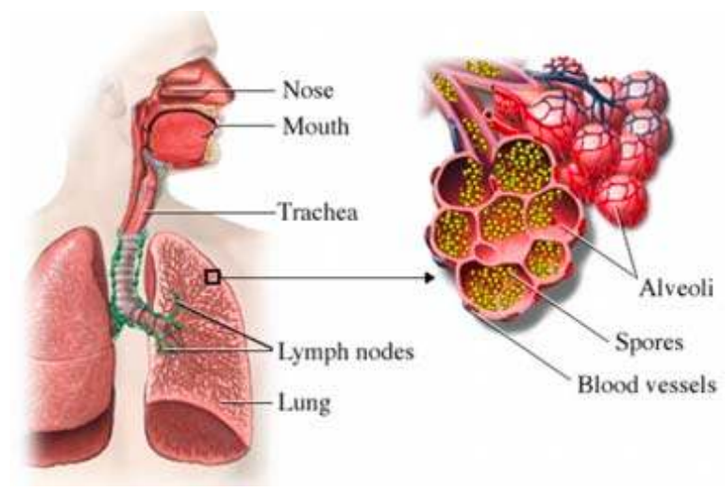


Steenwyk *et al.* (2018), *bioRxiv*

81 genomes from mainly *Aspergillus and Penicillium*



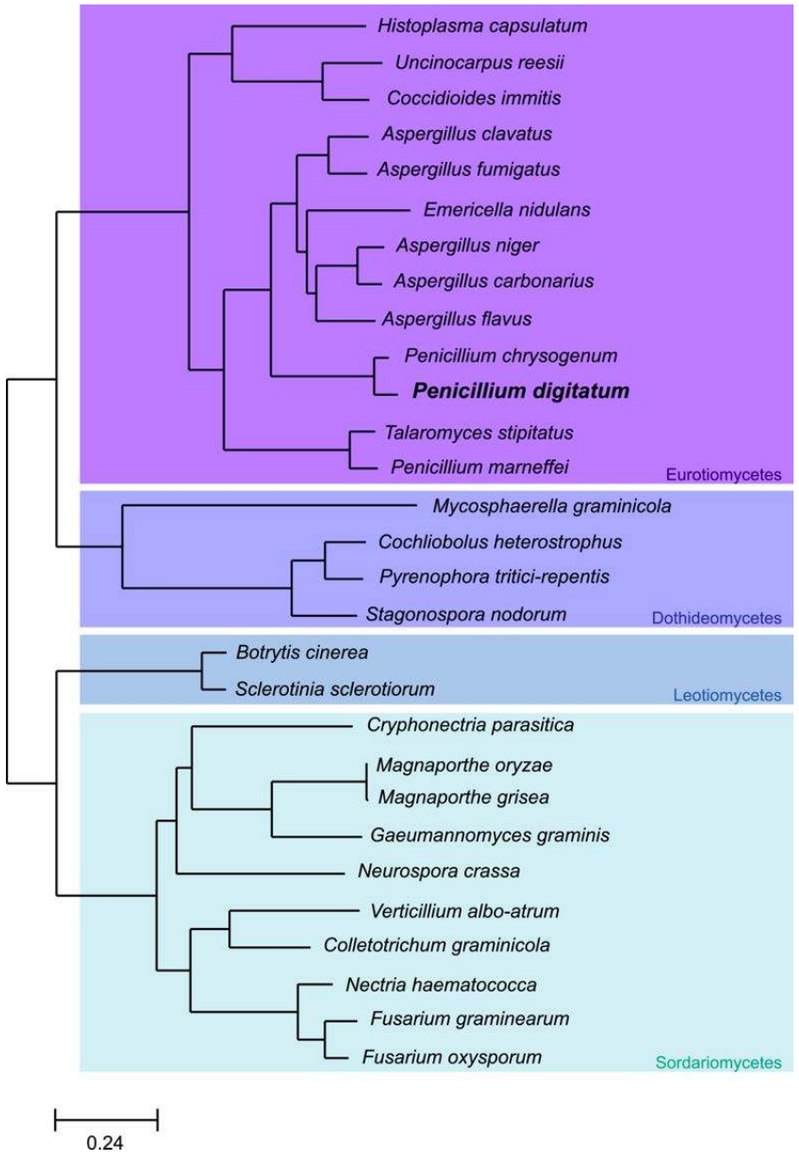




Steenwyk *et al.* (2018), *bioRxiv*

Utility of concatenation and coalescence

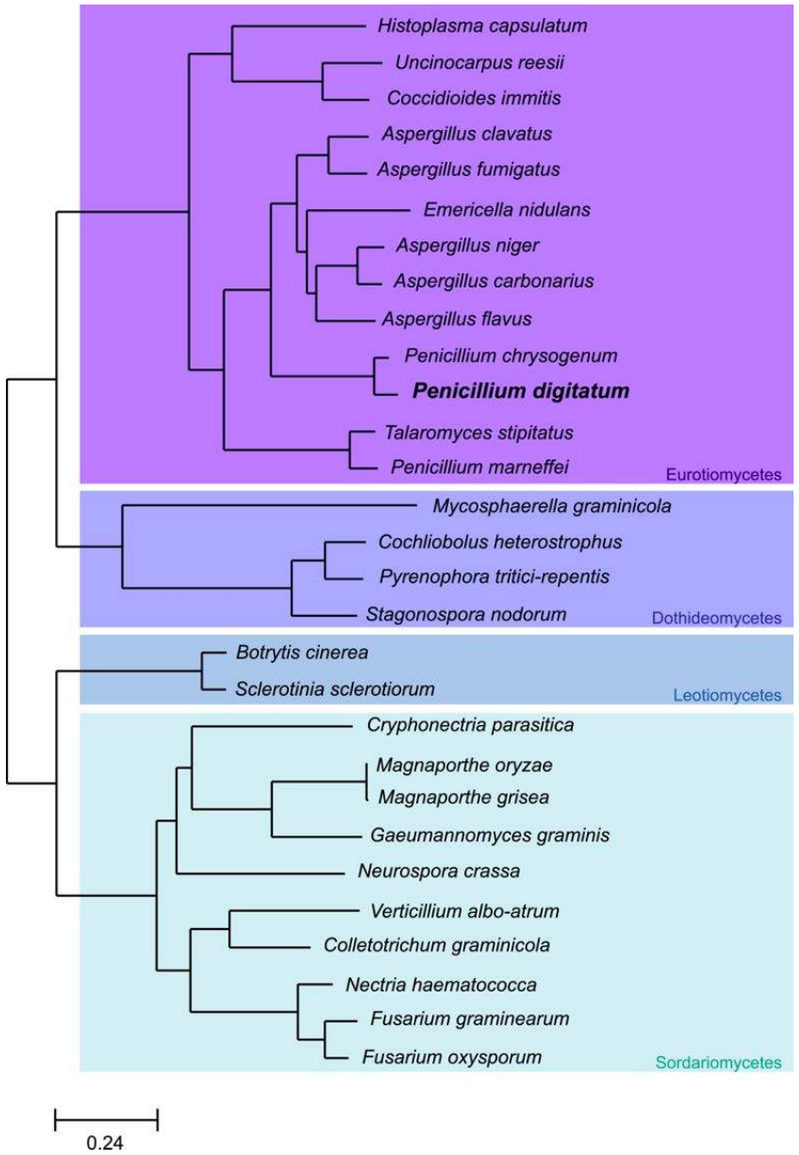
Marina — 29 fungi



Marcet-Houben, *et al.* 2012, BMC Genomics

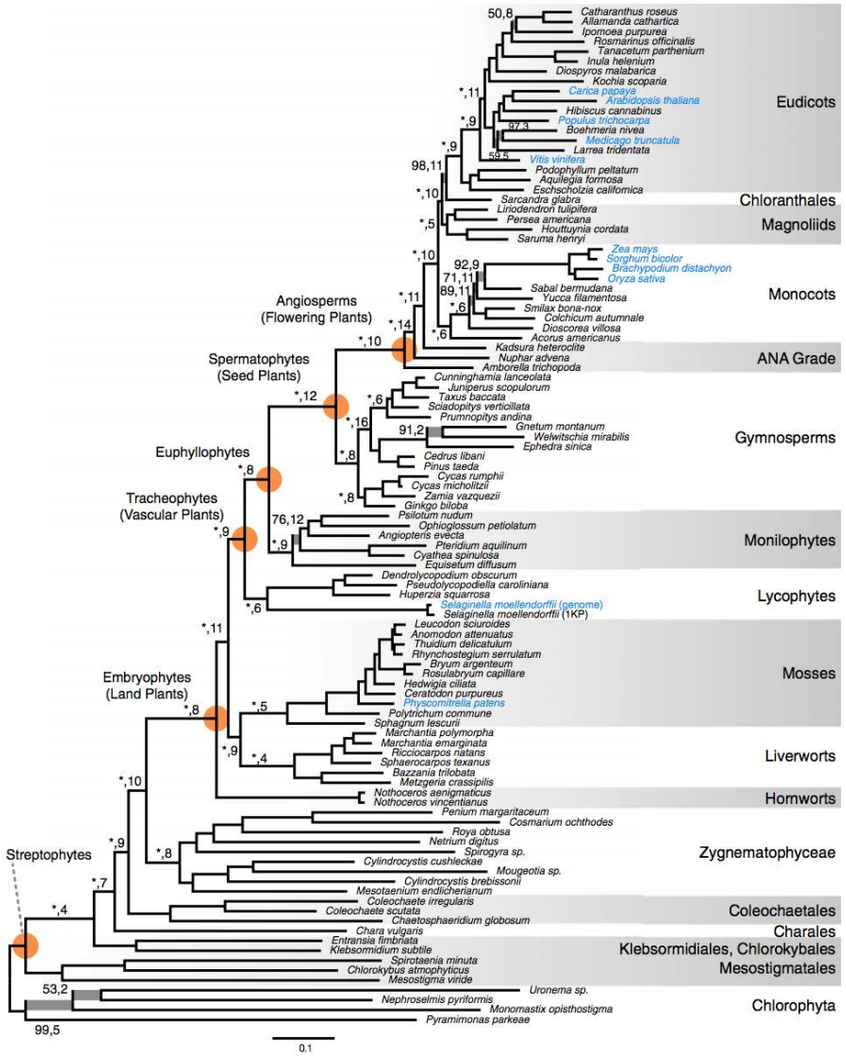
Utility of concatenation and coalescence

Marina — 29 fungi



Marcet-Houben, *et al.*
2012, BMC Genomics

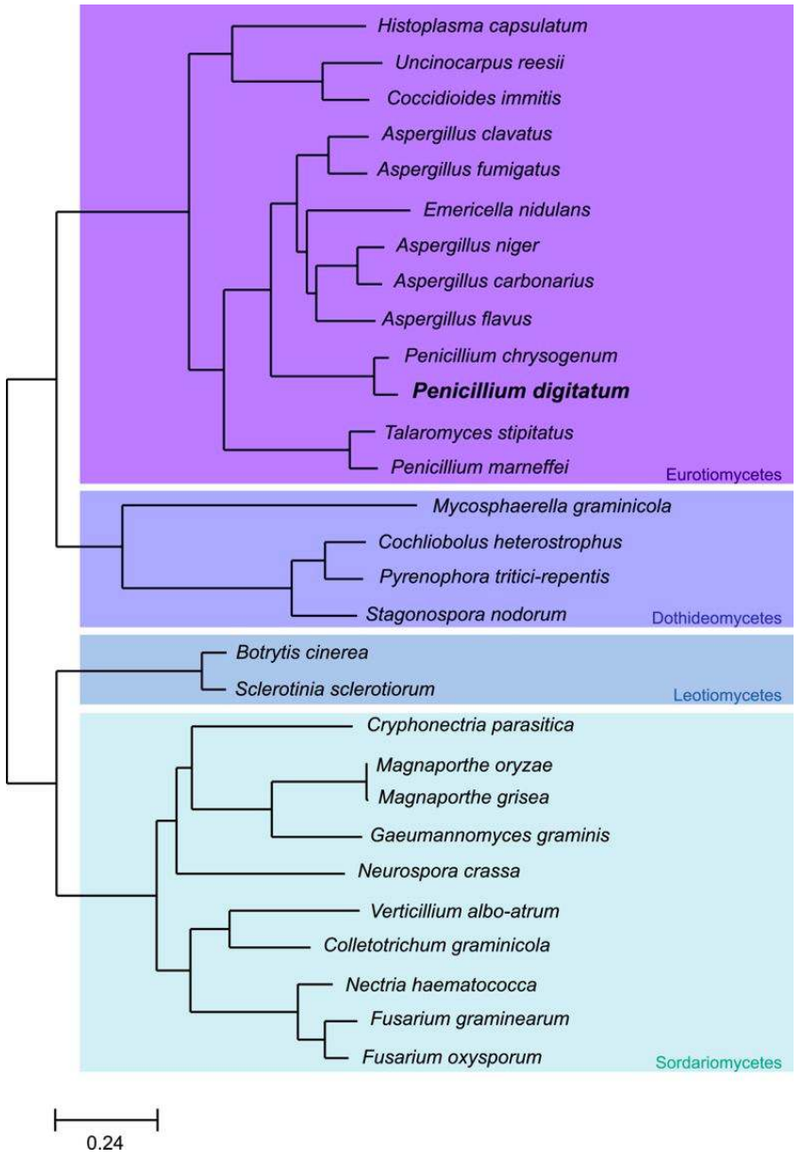
Lisa — 92 plants



Wickett, *et al.*
2014, PNAS

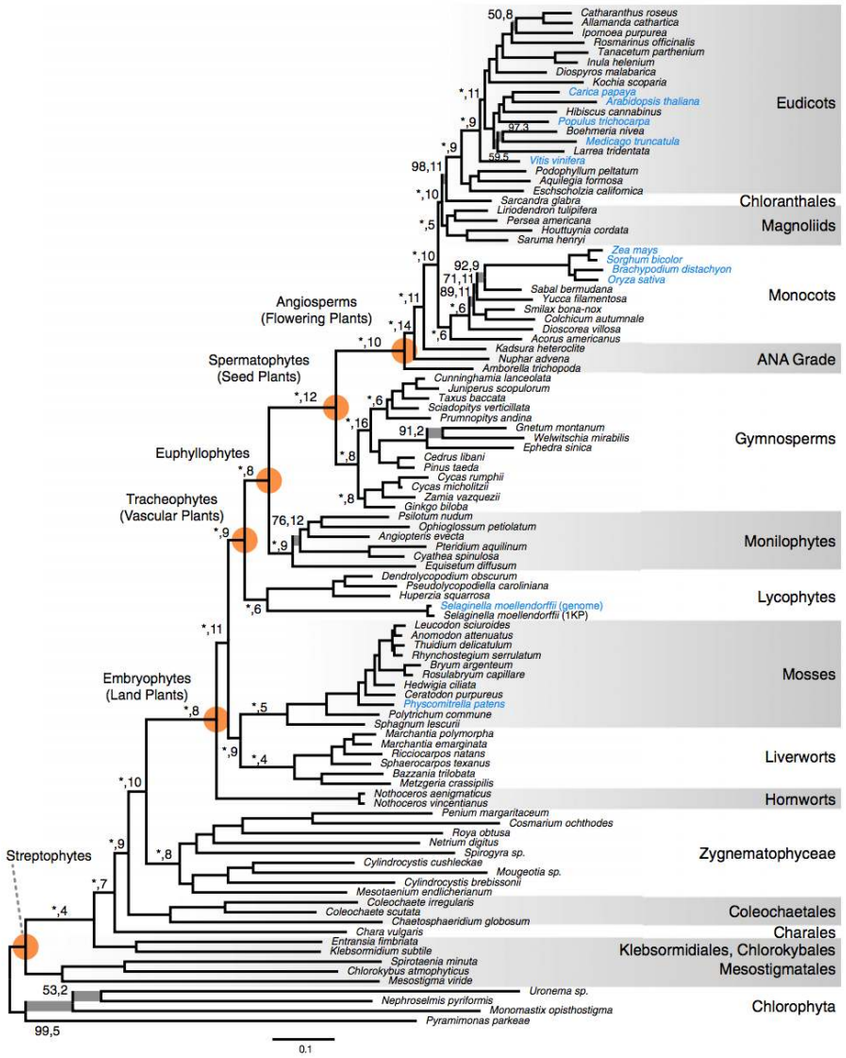
Utility of concatenation and coalescence

Marina — 29 fungi



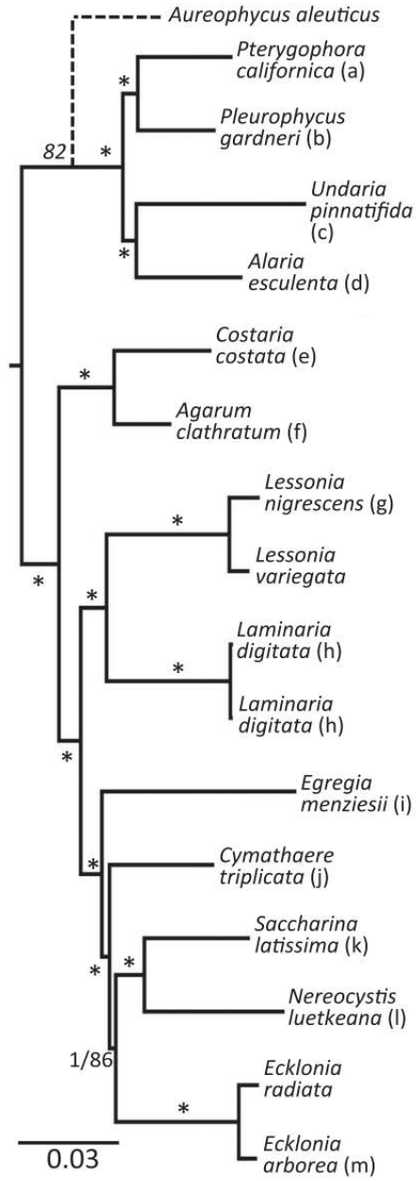
Marcet-Houben, *et al.* 2012, BMC Genomics

Lisa — 92 plants



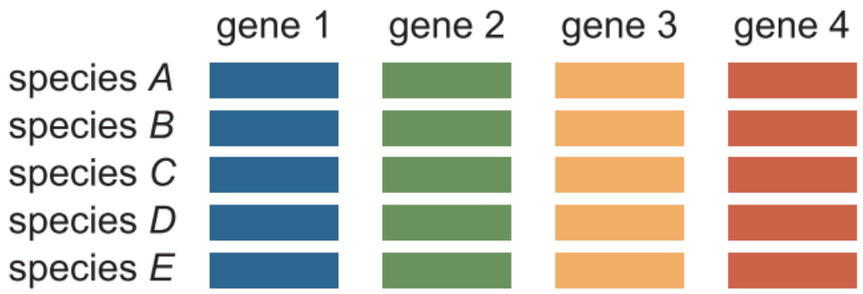
Wickett, *et al.* 2014, PNAS

Eric — 17 kelp



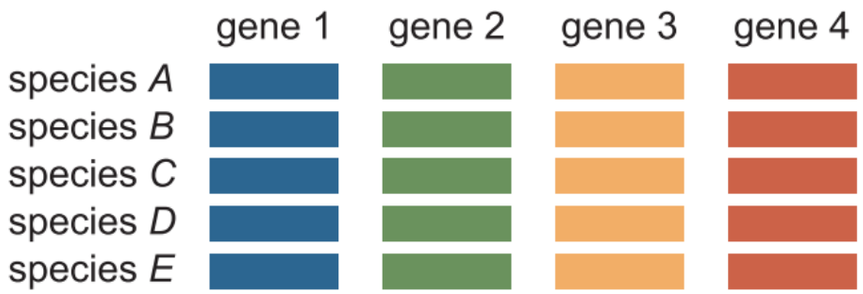
Jackson, *et al.* 2018, Journal of Phycology

Major methods in Phylogenomics

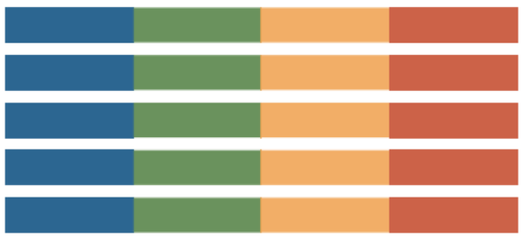


Liu *et al.* 2015, Ann. N.Y. Acad. Sci.

Major methods in Phylogenomics



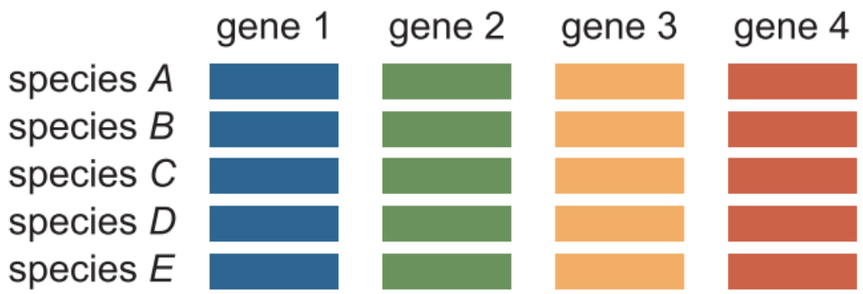
concatenation



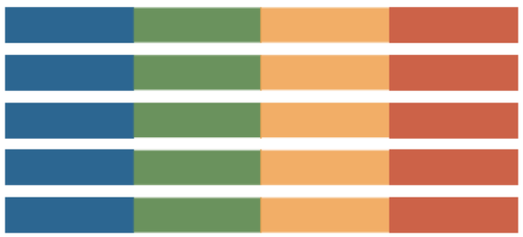
supermatrix

Liu *et al.* 2015, Ann. N.Y. Acad. Sci.

Major methods in Phylogenomics



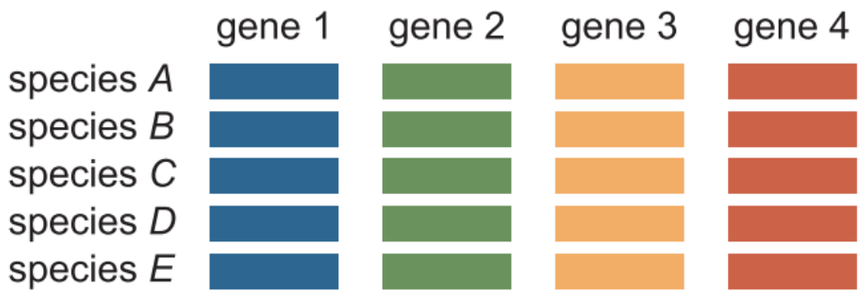
concatenation



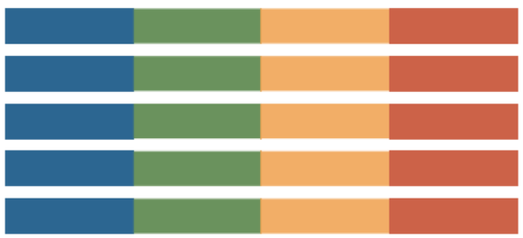
supermatrix

Liu *et al.* 2015, Ann. N.Y. Acad. Sci.

Major methods in Phylogenomics

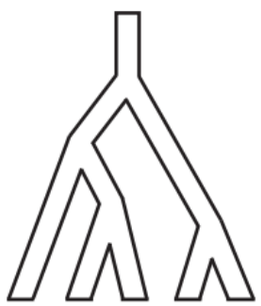


concatenation



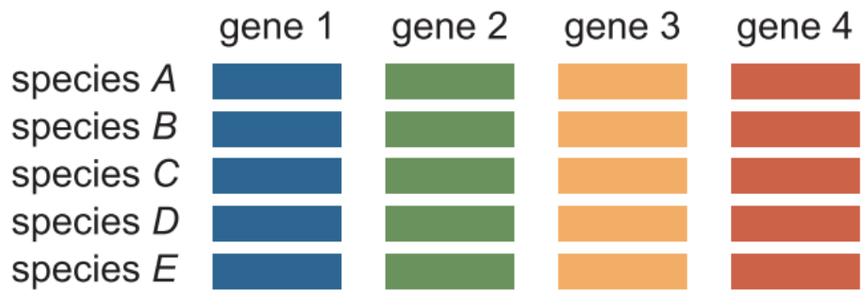
supermatrix

‘single step’ coalescent

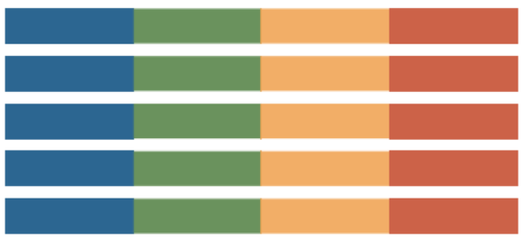


Liu *et al.* 2015, Ann. N.Y. Acad. Sci.

Major methods in Phylogenomics

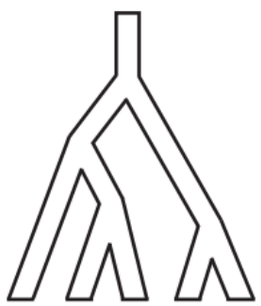


concatenation



supermatrix

‘single step’ coalescent

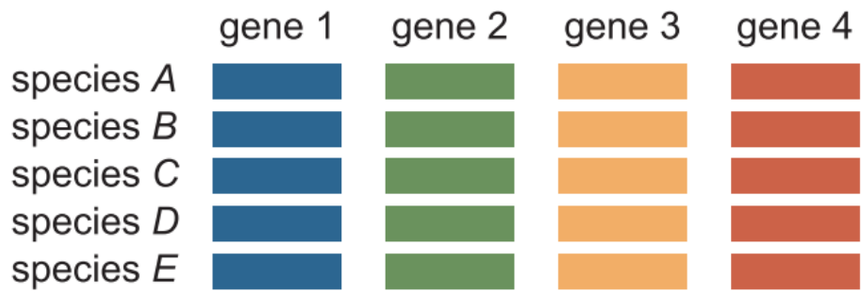


‘two-step’ coalescent

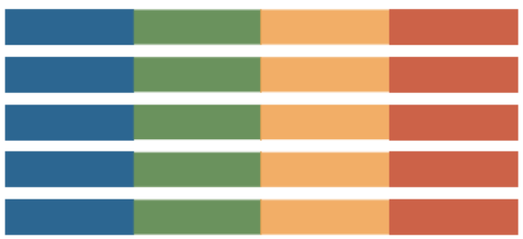
estimated gene trees

Liu *et al.* 2015, Ann. N.Y. Acad. Sci.

Major methods in Phylogenomics

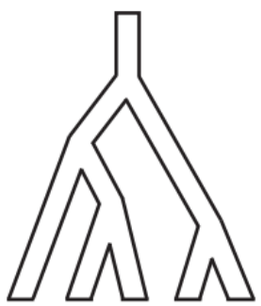


concatenation



supermatrix

‘single step’ coalescent

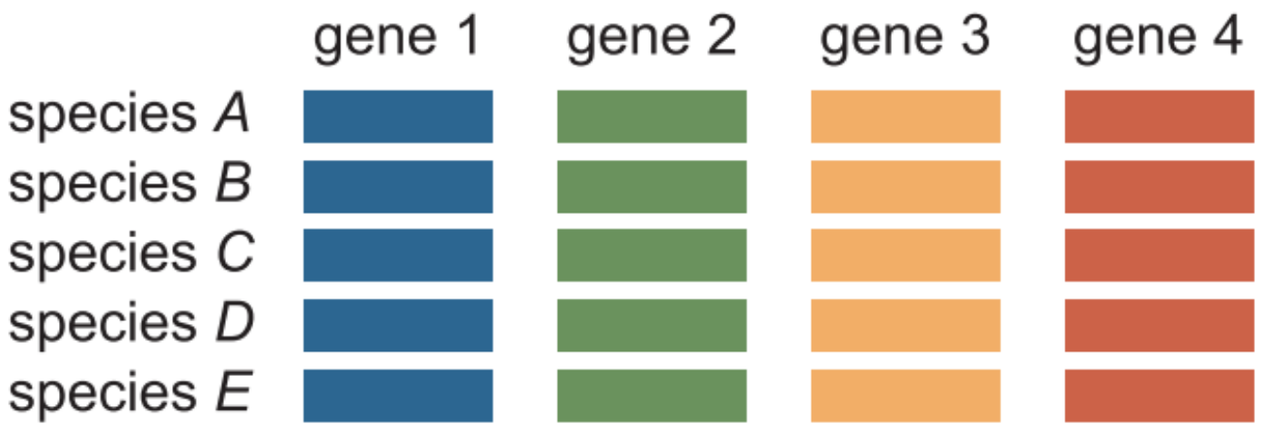


‘two-step’ coalescent

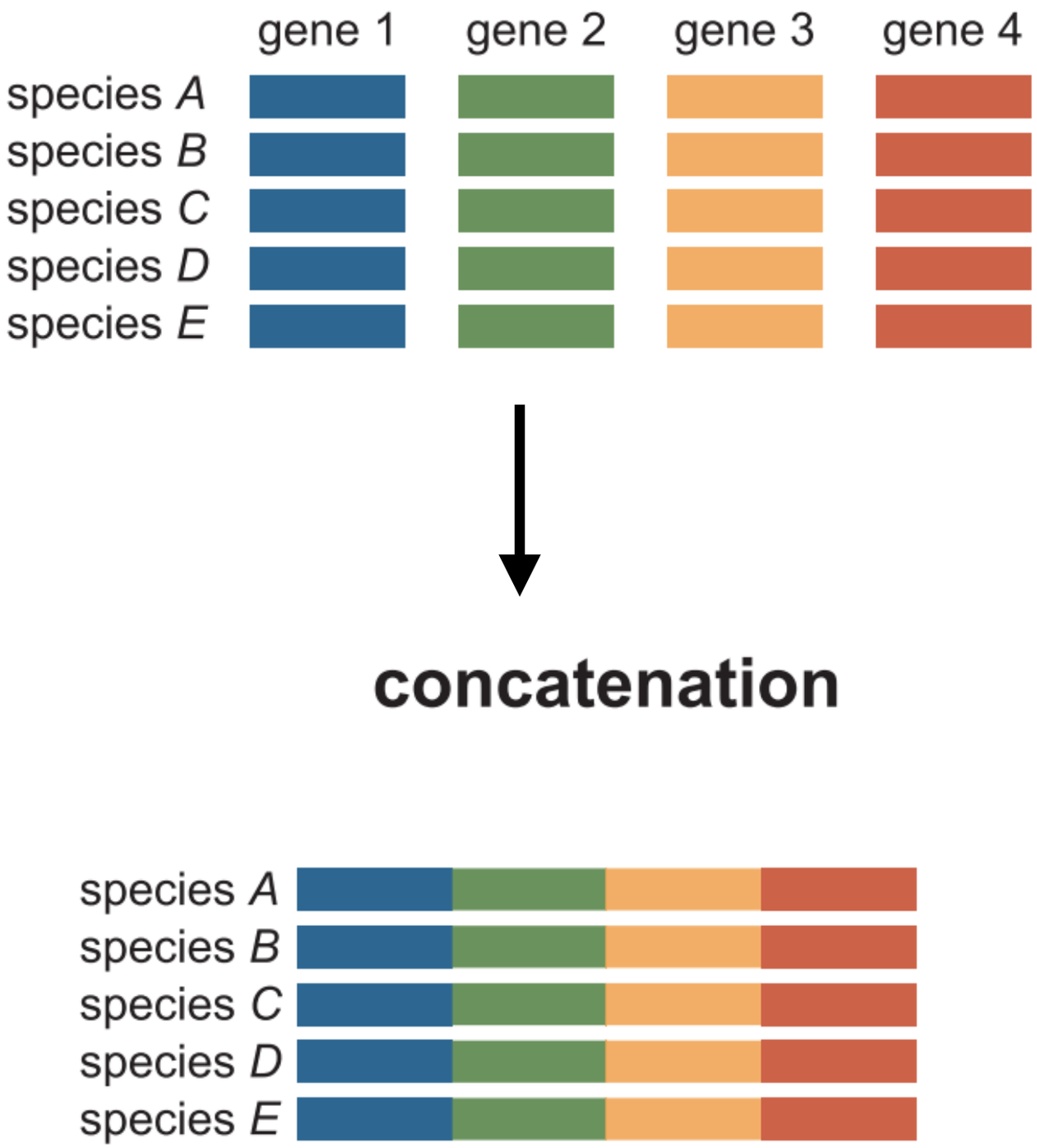
estimated gene trees

Liu *et al.* 2015, Ann. N.Y. Acad. Sci.

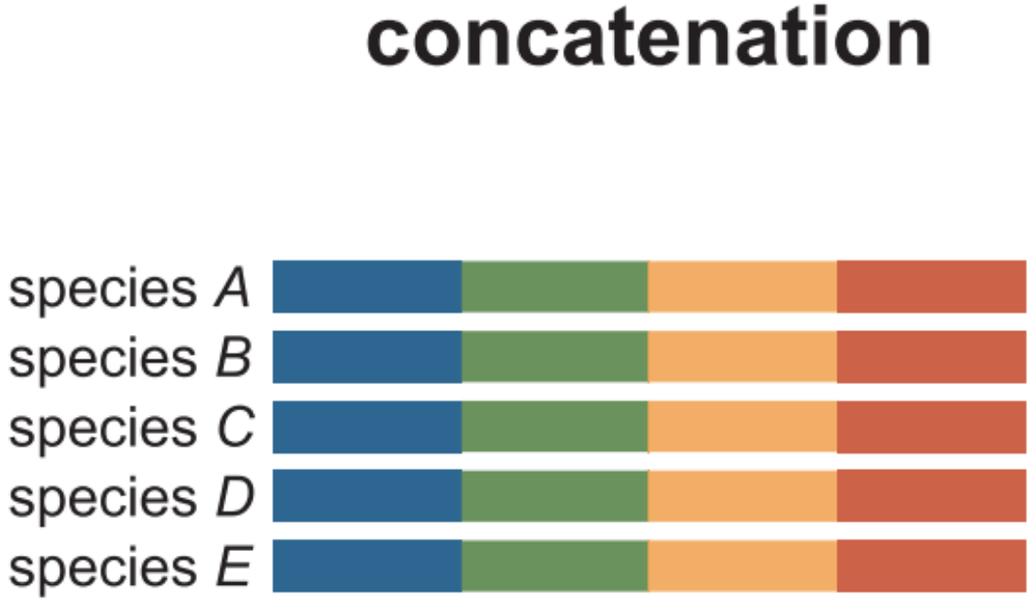
How do we concatenate sequences?



How do we concatenate sequences?



Partition file - providing boundary information



Model, Partition ID = start and stop boundaries

Model, Blue = 1-481
Model, Green = 482-1054
Model, Yellow = 1055-1492
Model, Red = 1493-1918

Methods to concatenate sequences

Methods to concatenate sequences

Manual

- That is, by hand

Methods to concatenate sequences

Manual

- That is, by hand….*but why???*

Methods to concatenate sequences

Manual

- That is, by hand….*but why???*

GUI (Graphical User Interface)

- SequenceMatrix

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1096-0031.2010.00329.x

- CONCATENATOR

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1755-0998.2008.02164.x

Methods to concatenate sequences

Manual

- That is, by hand….*but why???*

GUI (Graphical User Interface)

- SequenceMatrix

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1096-0031.2010.00329.x

- CONCATENATOR

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1755-0998.2008.02164.x

Command-line

- *catfasta2phyml*

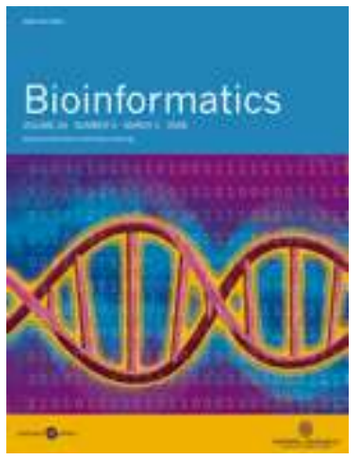
https://github.com/nylander/catfasta2phyml

- *FASconCAT-G*

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4243772/

Concatenation, partitioning, and model finding

1)



Phyutility: a phyloinformatics tool for trees, alignments and molecular data FREE

Stephen A. Smith ✉, Casey W. Dunn Author Notes

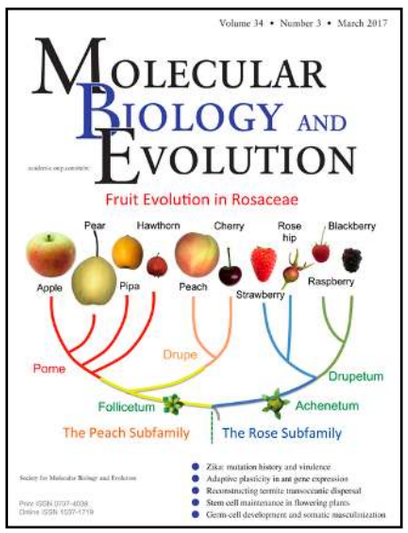
Bioinformatics, Volume 24, Issue 5, 1 March 2008, Pages 715–716,
https://doi.org/10.1093/bioinformatics/btm619
Published: 28 January 2008 **Article history** ▾

2)

A custom script I wrote just for you!

https://jlsteenwyk.github.io/resources.html

3)

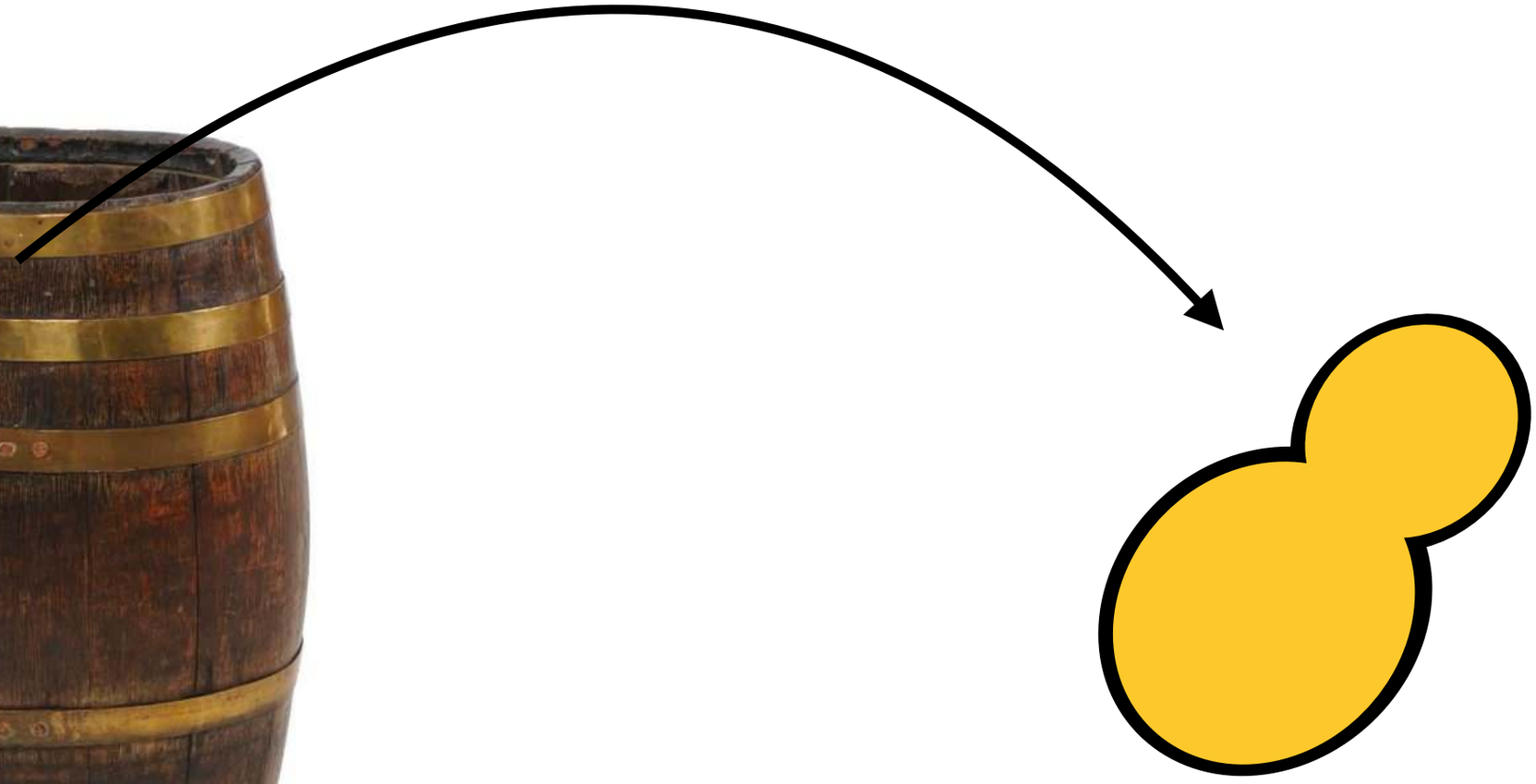


PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses FREE

Robert Lanfear ✉, Paul B. Frandsen, April M. Wright, Tereza Senfeld, Brett Calcott

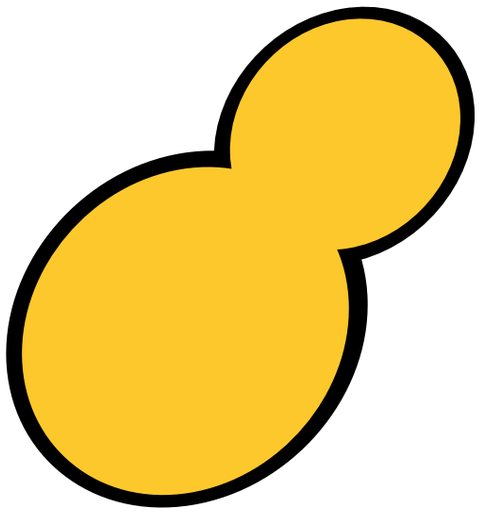
Molecular Biology and Evolution, Volume 34, Issue 3, 1 March 2017, Pages 772–773,
https://doi.org/10.1093/molbev/msw260
Published: 24 December 2016

Yeast from the brewmaster



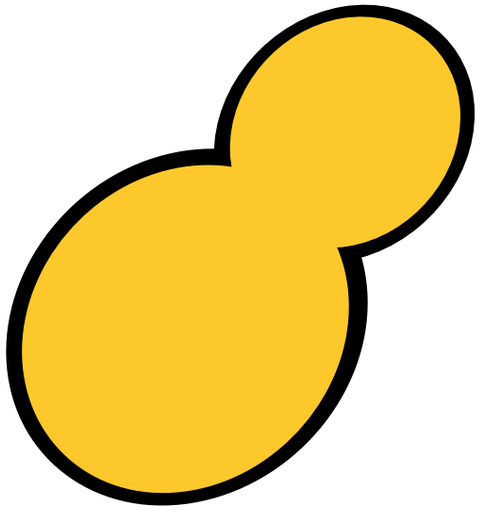
???

Steps taken before practical



???

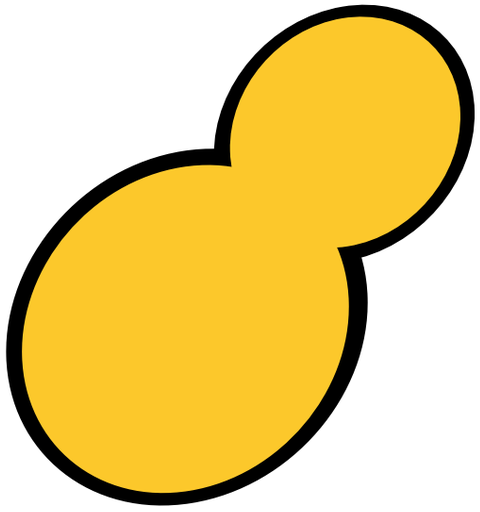
Steps taken before practical



???

Sequence
Genome

Steps taken before practical

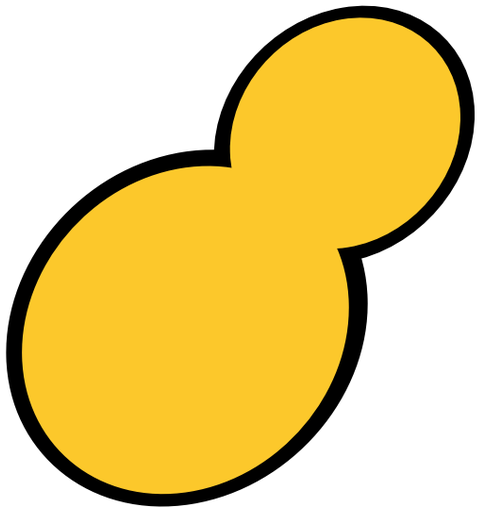


???

Sequence Genome

Add other yeast genomes

Steps taken before practical

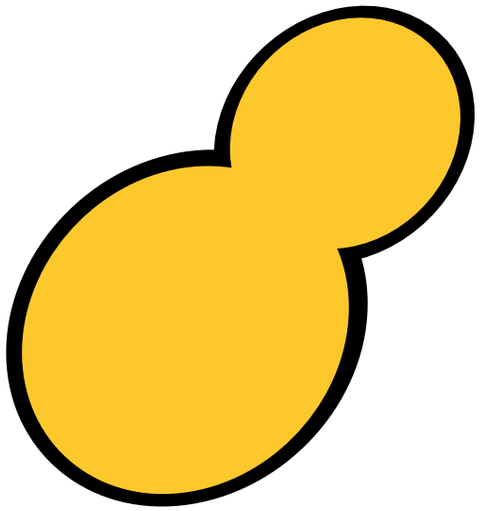


???

Sequence Genome

Add other yeast genomes

Steps taken before practical

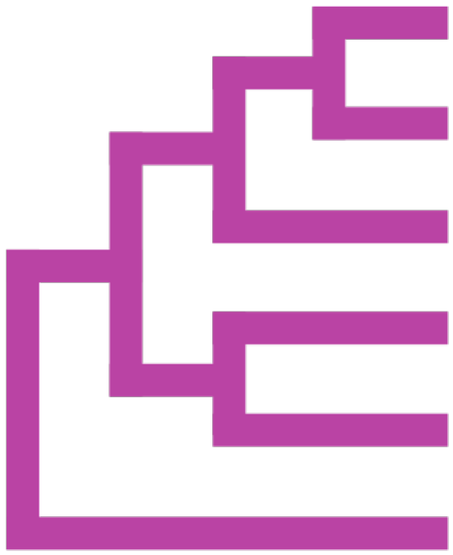


???

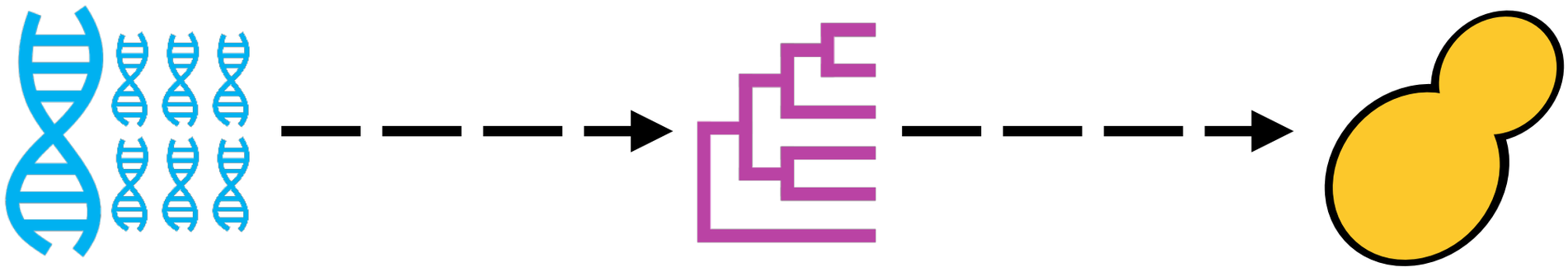
Sequence Genome

Add other yeast genomes

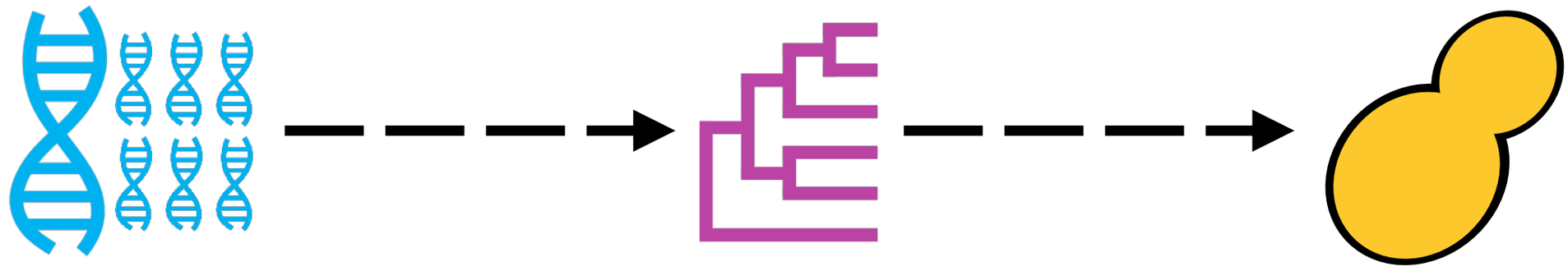
challenge



Challenge



Challenge

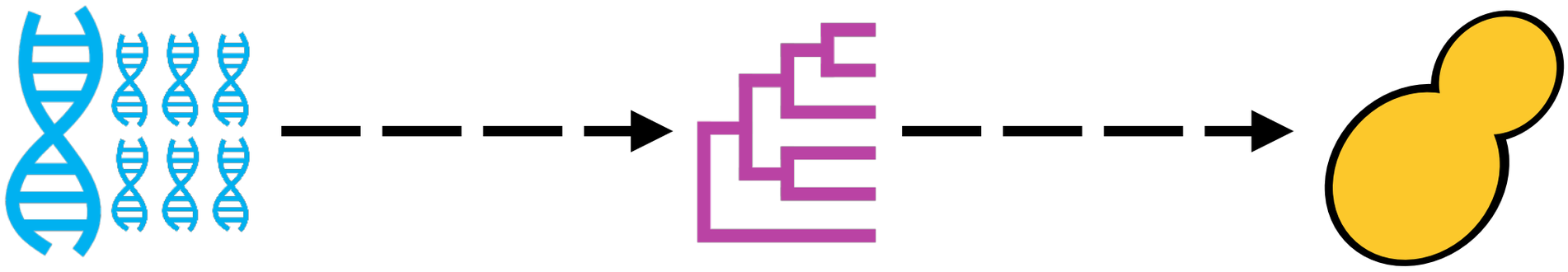


- Using a reduced set of protein sequences in FILES_Wed_challenge_fastas.tar.gz to determine what the yeast is

1) Call orthologs
2) Align and trim orthologs
3) Concatenate sequences
4) Infer putative species tree

Hint: outgroup taxa are
Starmerella apicola
Starmerella bombicola
Wickerhamiella versatilis

Challenge

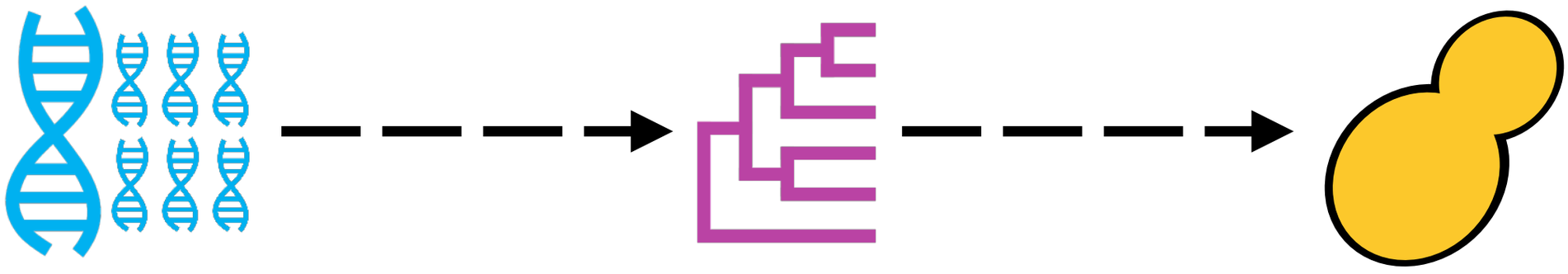


- Using a reduced set of protein sequences in FILES_Wed_challenge_fastas.tar.gz to determine what the yeast is

1) Call orthologs
2) Align and trim orthologs
3) Concatenate sequences
4) Infer putative species tree

Hint: outgroup taxa are
Starmerella apicola
Starmerella bombicola
Wickerhamiella versatilis

Challenge



- Using a reduced set of protein sequences in FILES_Wed_challenge_fastas.tar.gz to determine what the yeast is

1) Call orthologs
2) Align and trim orthologs
3) Concatenate sequences
4) Infer putative species tree

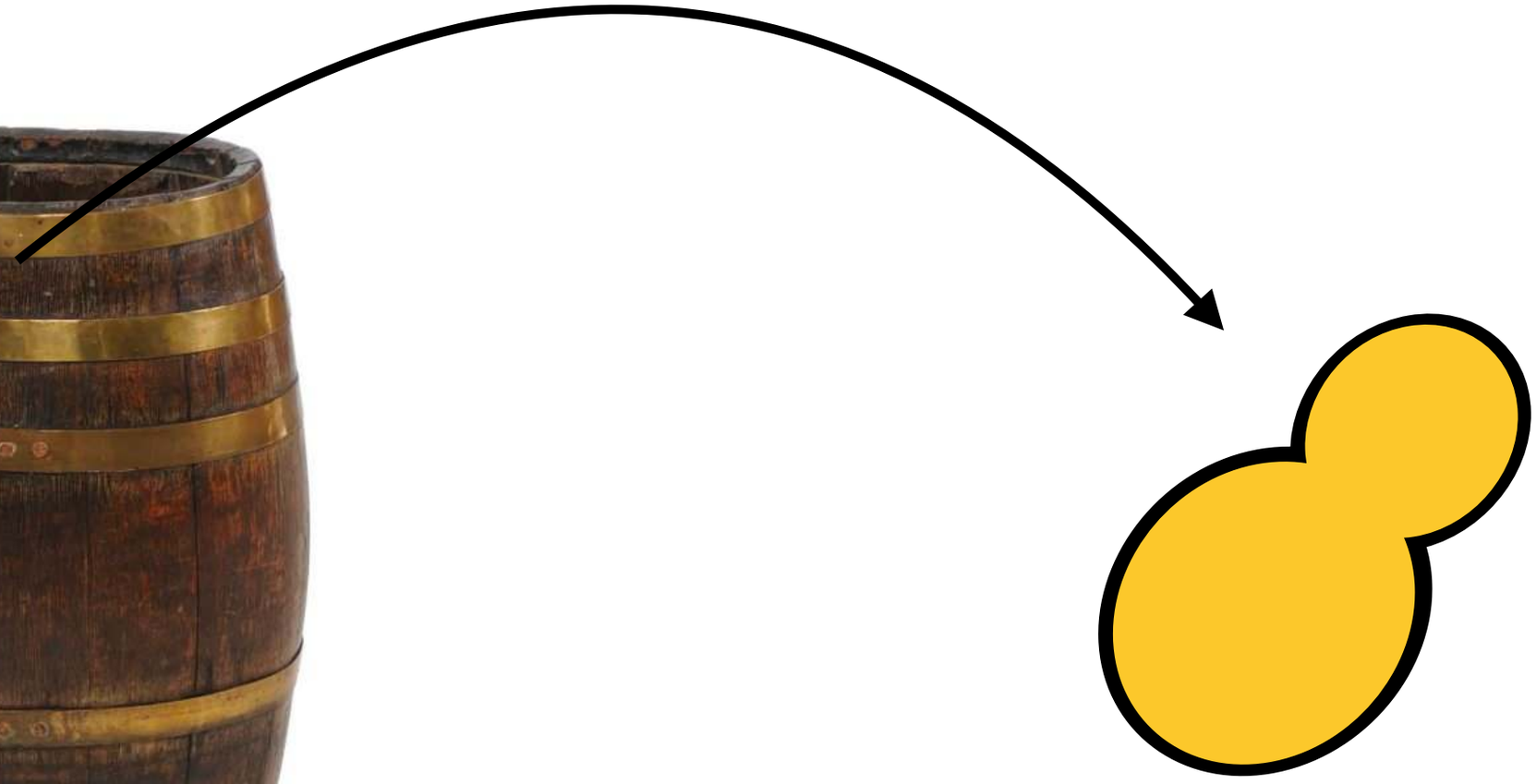
Hint: outgroup taxa are
Starmerella apicola
Starmerella bombicola
Wickerhamiella versatilis

Hint: You can extract a FASTA entry from a multi-FASTA file using samtools faidx function with the format:
Samtools faidx fasta.file fasta.entry

e.g. if I want to extract gene *Brewery_genome_1* from multi-FASTA file *Brewery_genome.fa* I would execute the command,

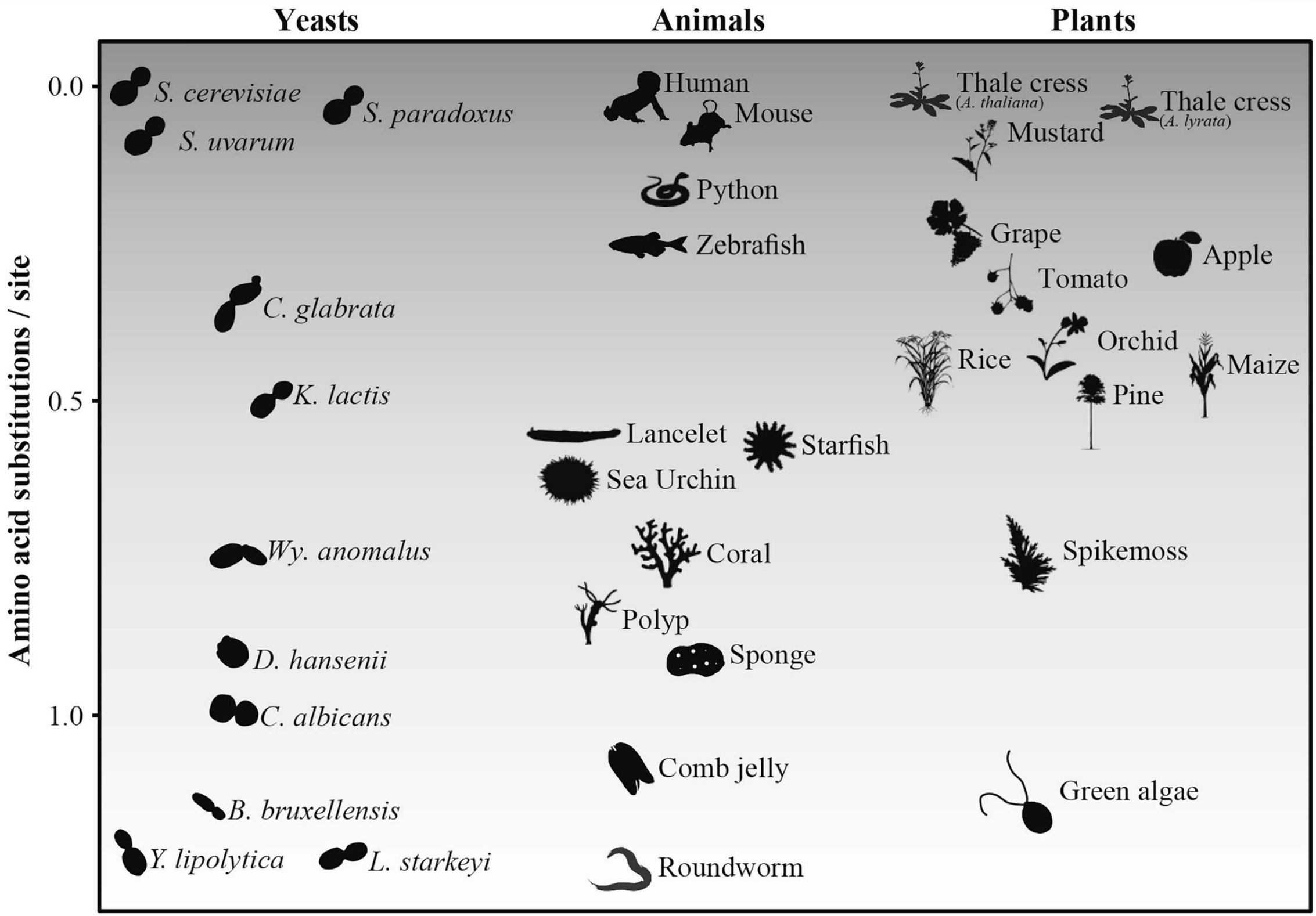
samtools faidx Brewery_genome.fa Brewery_genome_1

Yeast from the brewmaster



???

S. cerevisiae and *B. bruxellensis* are VERY distant



Shen X.-X. *et al.* (2018), *Cell*