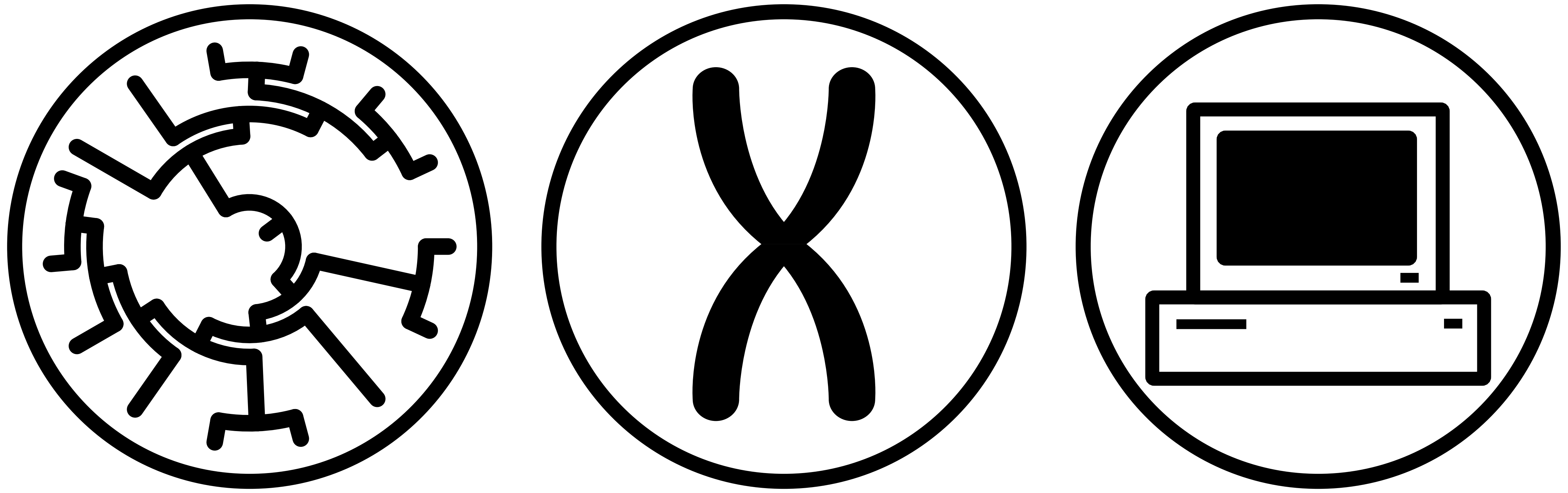
# Many thanks to Gemma!

@JLSteenwyk

# Outline



- Introduction
- Inferring genetic networks from phylogenies
- Phylogenomic subsampling
- Misc. notes before the tutorial


Jacob
Howard
Mercy
Emily
Josh
Nina
Gary

# Art for Earth

Purchase with purpose: 100% of profits go toward global conservation efforts

Shop by product type, conservation status, or buy a sticker of Sciart logo!

Vinyl Stickers

Poster Prints

Camper Mugs

Endangered Animals

No current concern

Have a question? Check out the Frequently Asked Questions (FAQ) section or get in touch via twitter!

# Art for Earth

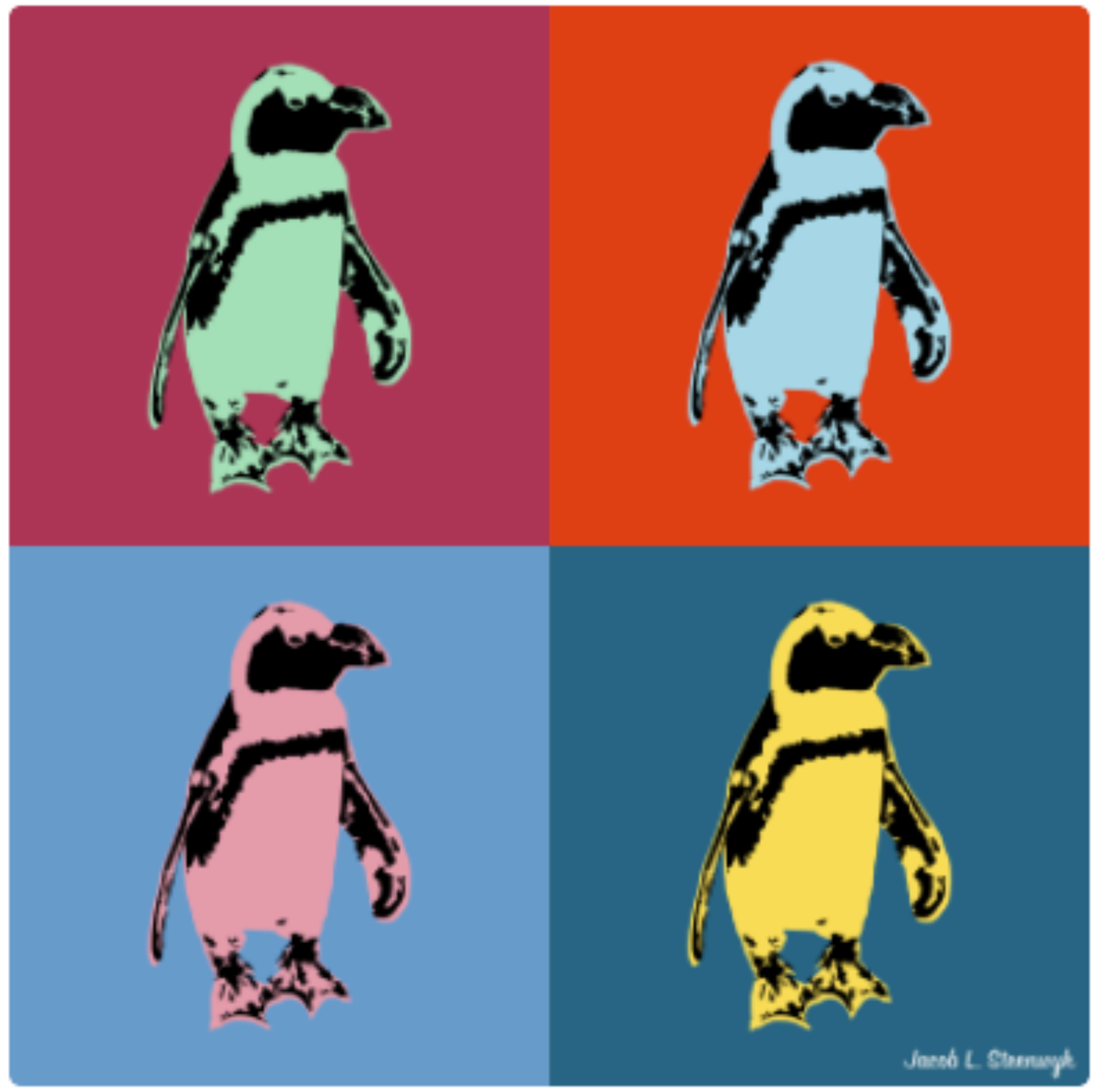
## Using art to raise awareness and immortalize endangered species

African wild dog (*Lycaon pictus*)

- Status: Endangered
- Population: 1,409

Blue whale (*Balaenoptera musculus*)

- Status: Endangered
- Population: 10,000 - 25,000



Galápagos penguin (*Spheniscus mendiculus*)

- Status: Endangered
- Population: fewer than 2,000

@JLSteenwyk

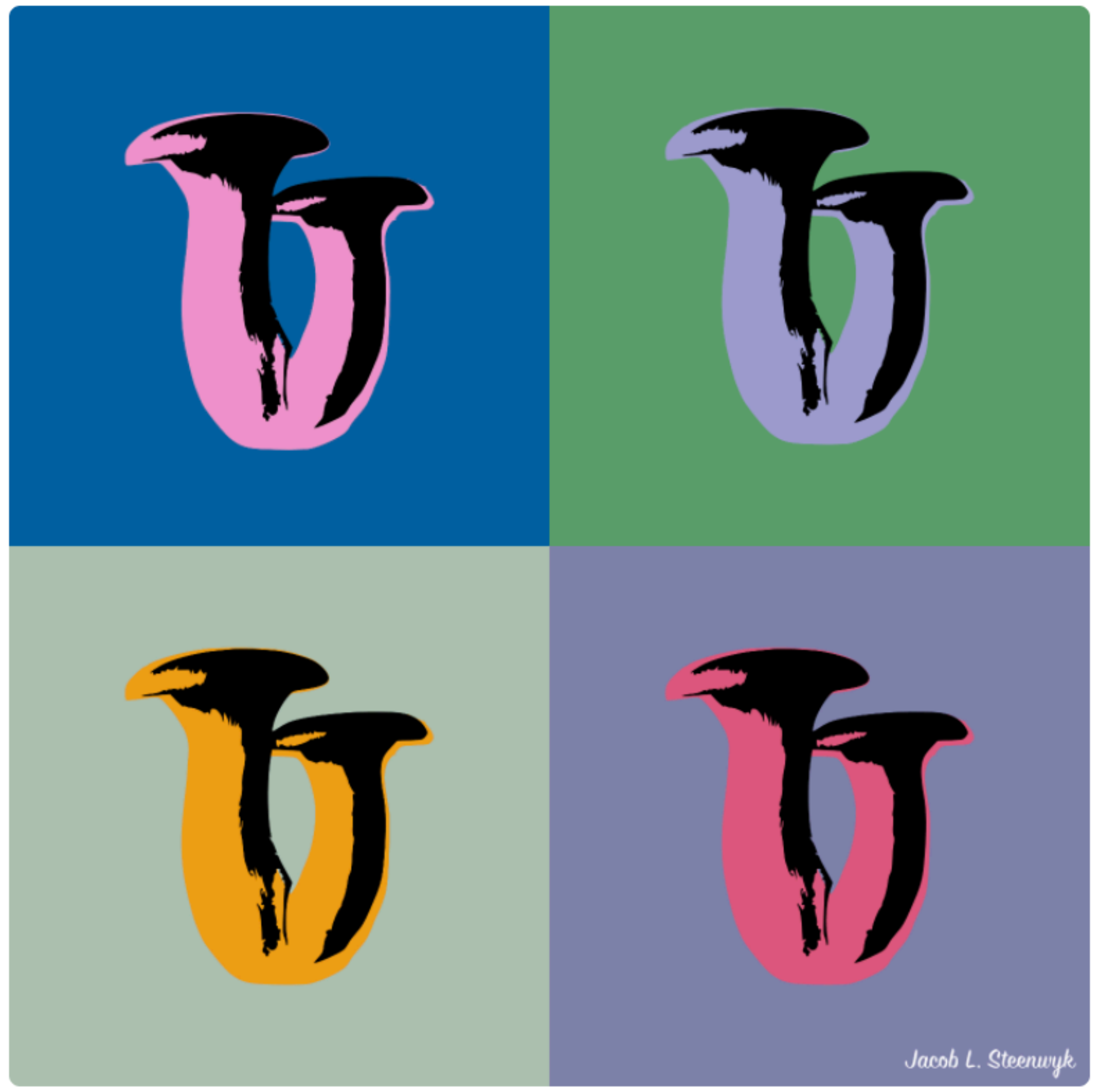
https://www.etsy.com/shop/JLSteenwyk

# Art for Earth

## Using art to raise awareness and immortalize endangered species or species I love


Fly agaric (*Amanita muscaria*)

- $8.99 vinyl sticker (FREE shipping)
- $22.99+ poster print (FREE shipping)
- $28.99 camper mug (FREE shipping)




Oyster mushroom (*Pleurotus ostreatus*)

- $8.99 vinyl sticker (FREE shipping)
- $22.99+ poster print (FREE shipping)
- $28.99 camper mug (FREE shipping)


Morel mushroom (*Morchella esculenta*)

- $8.99 vinyl sticker (FREE shipping)
- $22.99+ poster print (FREE shipping)
- $28.99 camper mug (FREE shipping)

@JLSteenwyk

# Featured on *Yeast* magazine

Received: 29 July 2020 | Accepted: 20 August 2020
DOI: 10.1002/yea.3518

SPECIAL ISSUE ARTICLE

Yeast WILEY

## A portrait of budding yeasts: A symbol of the arts, sciences and a whole greater than the sum of its parts

Jacob L. Steenwyk[1,2]

[1]Department of Biological Sciences, Vanderbilt University, Nashville, TN, USA

[2]Early Career Leadership Program Communication and Outreach Subcommittee, Genetics Society of America, Rockville, MD, USA

**Correspondence**
Jacob L. Steenwyk, Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA.
Email: jacob.steenwyk@vanderbilt.edu

**Funding information**
Vanderbilt University; Howard Hughes Medical Institute

KEYWORDS: art, budding yeast, cell cycle, Merian, naturalist, non-conventional yeasts, sciart, science, STEAM, Warhol

### 1 | INTRODUCTION

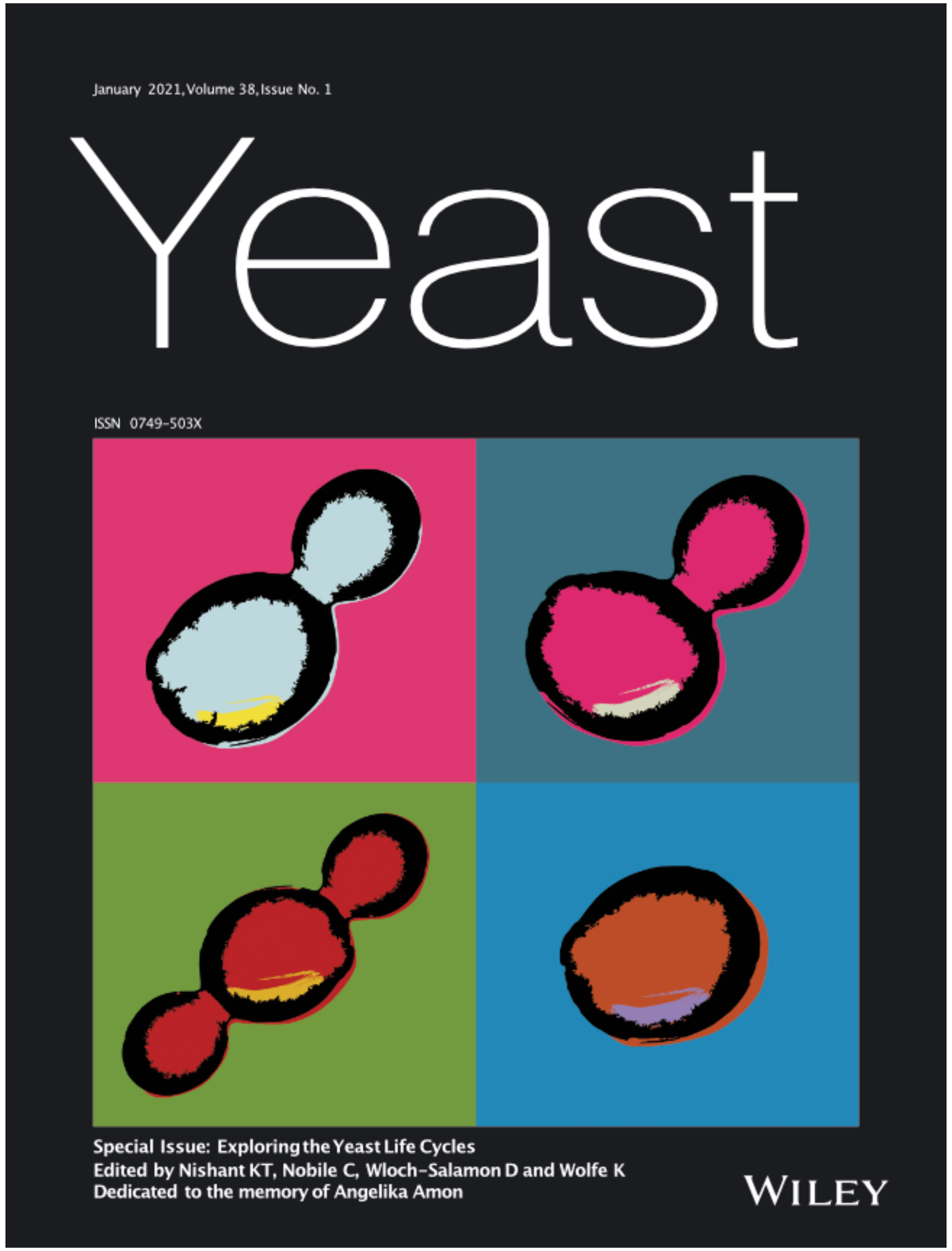
In the year 1660, 13-year-old Maria Sibylla Merian roamed the gardens and countryside of Germany taking detailed notes about caterpillars, moths, butterflies and their interactions with host plants, accompanying her notes were elaborate multimedia depictions of insect and plant life cycles (Figure 1). Merian's efforts in documenting interspecies relationships are regarded as early contributions to modern natural history and ecology, although the term 'ecology' was coined approximately two centuries later (Etheridge, 2011a, 2011b; Pieters & Winthagen, 1999). Her influence can be seen in the work of naturalists such as John James Audubon (Etheridge, 2015; Palmeri, 2017). Merian's success in part stems from her ability to use art to bolster her science and vice versa.

Merian is one of many scientists and artists who blended the arts and sciences over the centuries. In fact, scientist–artist polymaths like Aristotle and Leonardo da Vinci were more commonplace in part because of the common goal science and art share: interpreting and representing the natural world. The 'great divide' of the arts and sciences in Western cultures is thought to have started in the 19th century, coinciding with the term 'scientist' being coined (Braund & Reiss, 2019; Sumner, 1959; Zhu & Goyal, 2019). The division became reinforced. Schools for arts and sciences were separated as unfounded claims about brain differentiation formulated (Zhu & Goyal, 2019). For example, the right and left brain hemispheres were thought to be individually responsible for arts and science learning, respectively (Sperry, 1968). However, evidence from cognitive scientists favours a holistic view of the brain wherein a wide range of stimulation (e.g., arts and sciences) improves broad brain function and critical thinking skills (Braund & Reiss, 2019; Howes, Kaneva, Swanson, & Williams, 2013).

Today, the benefits of a holistic view of the arts and the sciences have been recognized by numerous institutions. For example, Science, Technology, Engineering, Arts and Mathematics (STEAM) inspired curriculum is used to help students build skills for broad problem solving in K–12 schools (Kim & Park, 2012; Peppler, 2013; Sochacka, Guyotte, & Walther, 2016). In higher education, artists, designers, researchers and inventors have formed forward-thinking coalitions such as the Center for Art, Science & Technology at Massachusetts Institute of Technology (https://arts.mit.edu/cast/) and ArtLab at Vanderbilt University (https://artlabvanderbilt.com/) to reunite the arts and sciences. These initiatives and many others have used the arts as an effective form of communication between scientists and the broader community (Illingworth, 2017), ultimately helping disseminate major scientific findings across society.

Perhaps one of the most important and recent scientific findings in the field of biological sciences is our understanding of the cellular life cycle. Seminal discoveries that unraveled the controls of the life cycle were made studying the model unipolar budding yeast *Saccharomyces cerevisiae* (Hartwell, Culotti, Pringle, & Reid, 1974). Comparative studies of *S. cerevisiae*, the fission yeast (*Schizosaccharomyces pombe*) and animals revealed striking similarities suggesting the life cycle is evolutionarily stable (Breeden & Nasmyth, 1987). Exploiting these similarities has enabled yeasts to be powerful models for cancer biology research and the development of anticancer therapeutics (Gao, Chen, & Huang, 2014; Guaragnella et al., 2014; Schwartz & Dickson, 2009). However, examination of non-conventional yeasts and their life cycles can provide novel insights important to the fields of cell biology, evolutionary biology and more. For example, species of the budding yeast genus *Hanseniaspora* have lost numerous cell cycle control genes, including *MAD1*, *MAD2* and *RAD9*, and components of the Anaphase Promoting Complex and display atypical bipolar budding patterns (Steenwyk




@JLSteenwyk

# Lineages of interest across my career

@JLSteenwyk

# Lineages of interest across my career

~18,000
genomes

~15,000
genomes

@JLSteenwyk

# Lineages of interest across my career

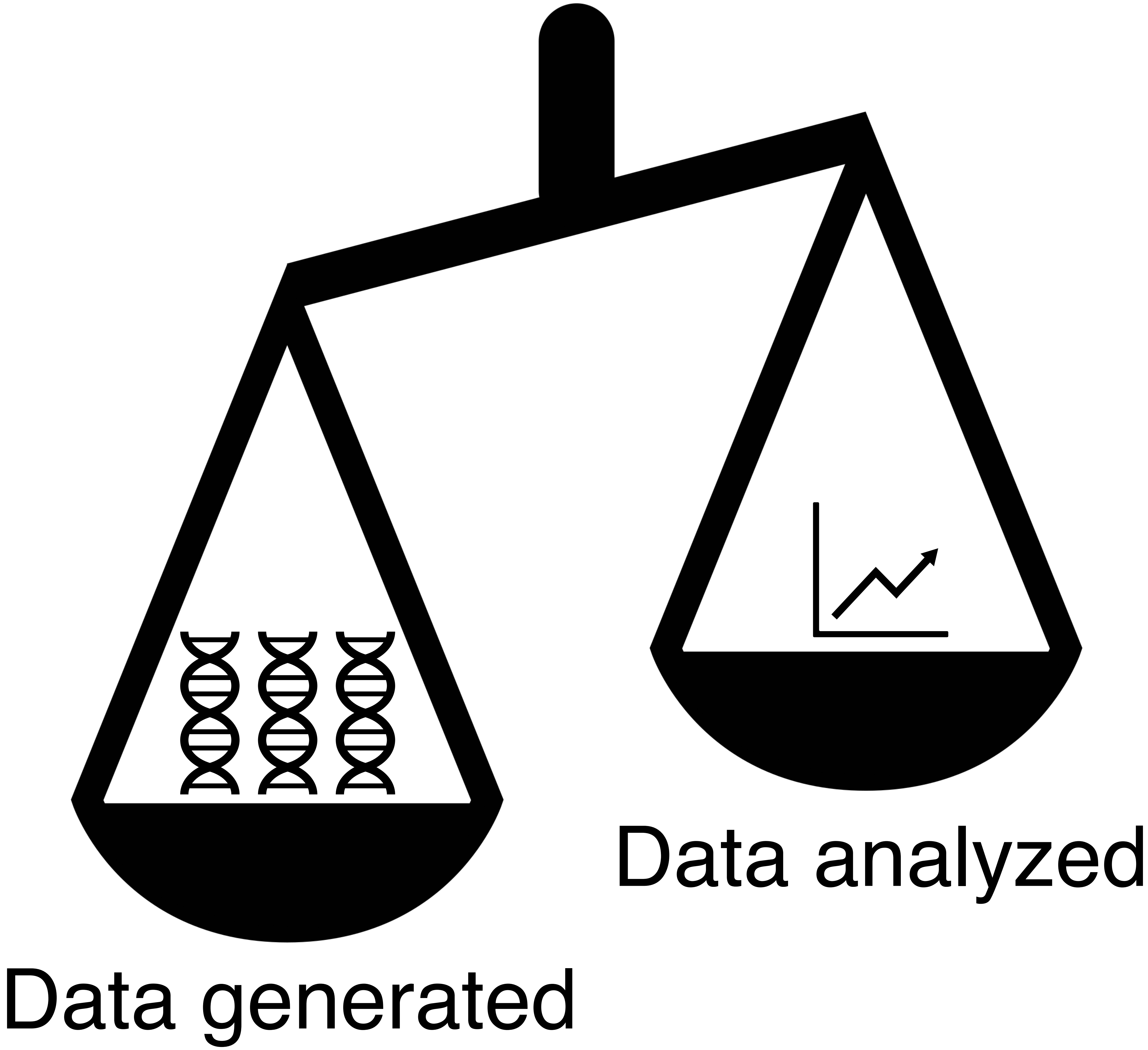
~18,000
genomes

~15,000
genomes

~4,000
plant genomes

@JLSteenwyk

# Data generation has outpaced data analysis



# Engineering software for 'omic inquiry

## Ortholog identification

OrthoSNAP

Steenwyk et al. (2022),
PLOS Biology

orthofisher

Steenwyk & Rokas (2021),
G3 Genes|Genomes|Genetics

## Phylogenomics

ClipKIT

Steenwyk et al. (2020),
PLOS Biology

PhyKIT

Steenwyk et al. (2021),
Bioinformatics

RemarKIT

Steenwyk et al. (in prep.)

## Genomics

BioKIT

Steenwyk et al. (2022),
Genetics

LVBRS

Le and Steenwyk et al. (2022),
bioRxiv

## Other

OVER UNDER

Steenwyk et al. (in prep.)

ggpubfigs

Steenwyk & Rokas (2021),
Micro. Resource Announcements

treehouse

Steenwyk & Rokas (2019),
BMC Research Notes

@JLSteenwyk

300,000+ downloads

# Engineering software for ‘omic inquiry

## Ortholog identification

OrthoSNAP

Steenwyk et al. (2022),
PLOS Biology

orthofisher

Steenwyk & Rokas (2021),
G3 Genes|Genomes|Genetics

## Phylogenomics

ClipKIT

Steenwyk et al. (2020),
PLOS Biology

PhyKIT

Steenwyk et al. (2021),
Bioinformatics

RemarKIT

Steenwyk et al. (in prep.)

## Genomics

BioKIT

Steenwyk et al. (2022),
Genetics

LVBRS

Le and Steenwyk et al. (2022),
bioRxiv

## Other

OVER UNDER

Steenwyk et al. (in prep.)

ggpubfigs

Steenwyk & Rokas (2021),
Micro. Resource Announcements

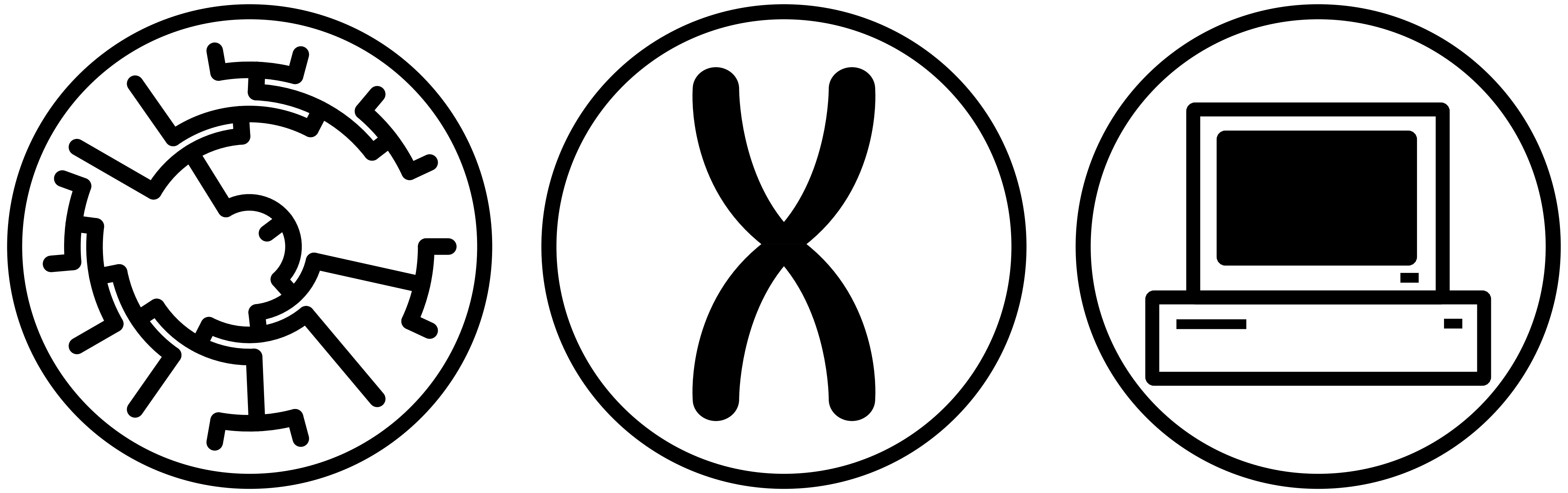
treehouse

Steenwyk & Rokas (2019),
BMC Research Notes

@JLSteenwyk

300,000+ downloads

# Outline



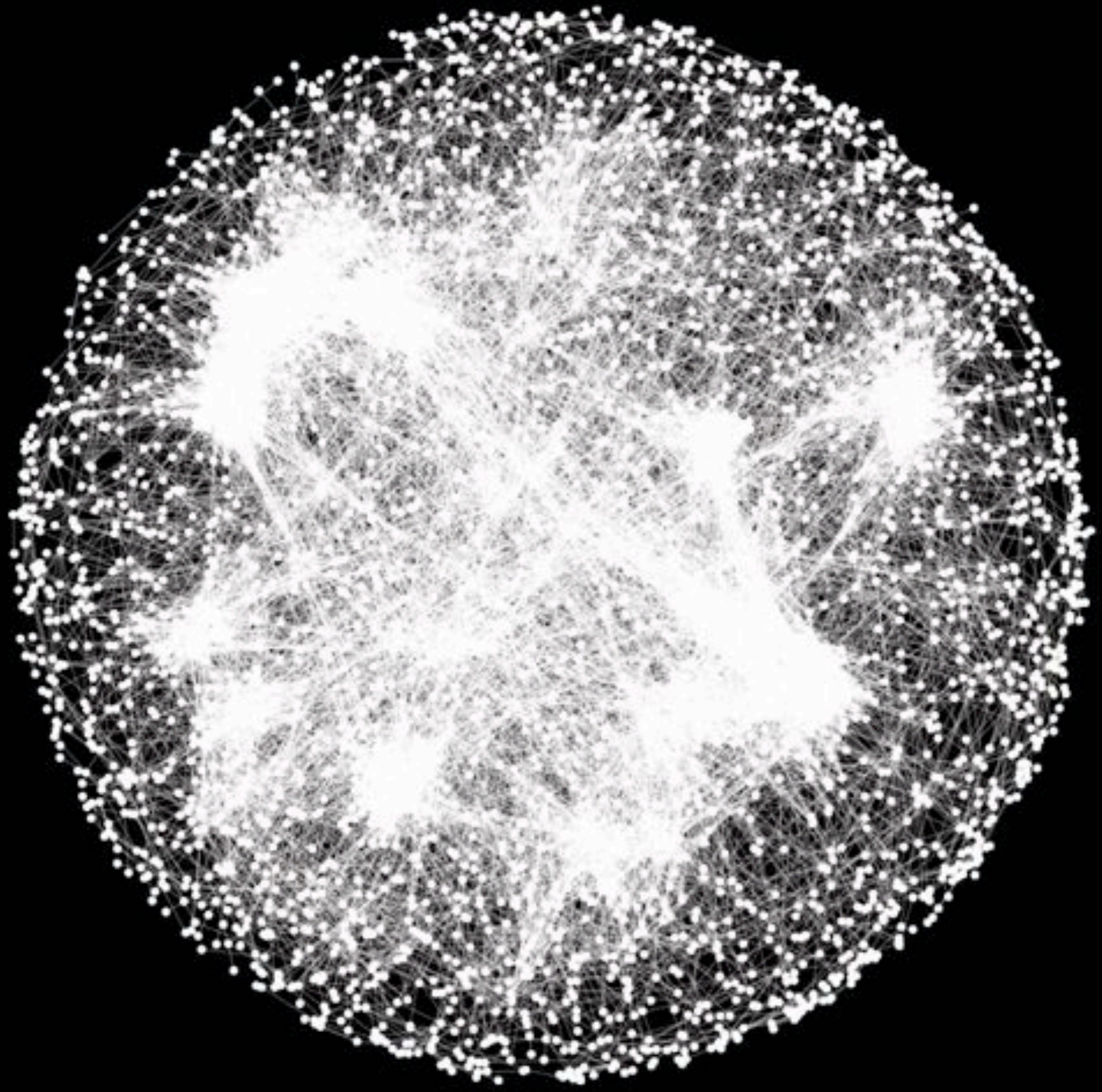
- Introduction
- **Inferring genetic networks from phylogenies**
- Phylogenomic subsampling
- Misc. notes before the tutorial

# Inferring genetic networks from phylogenies

Jacob L. Steenwyk

# Networks capture the complexity of genomic function

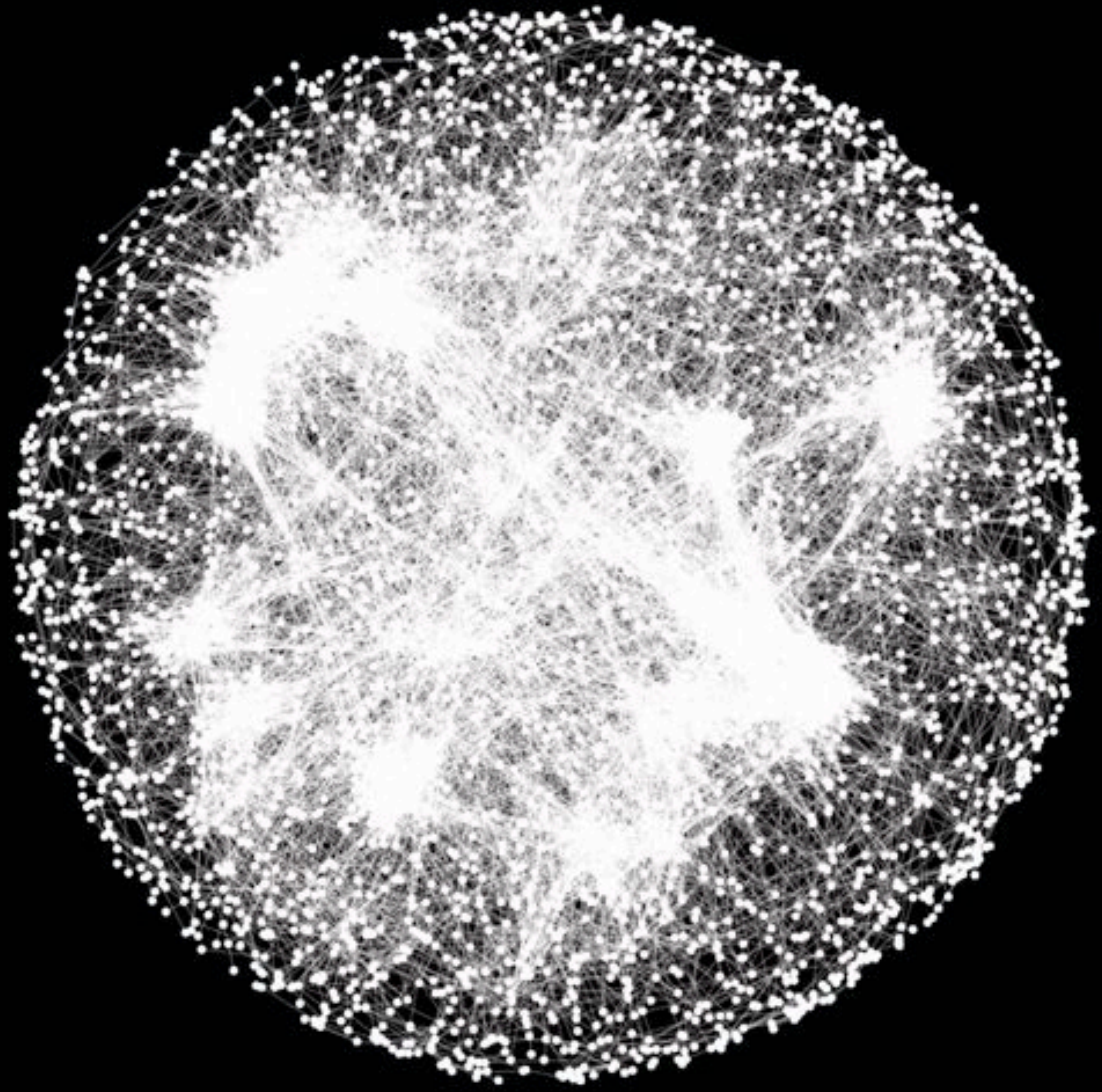
Baker's yeast

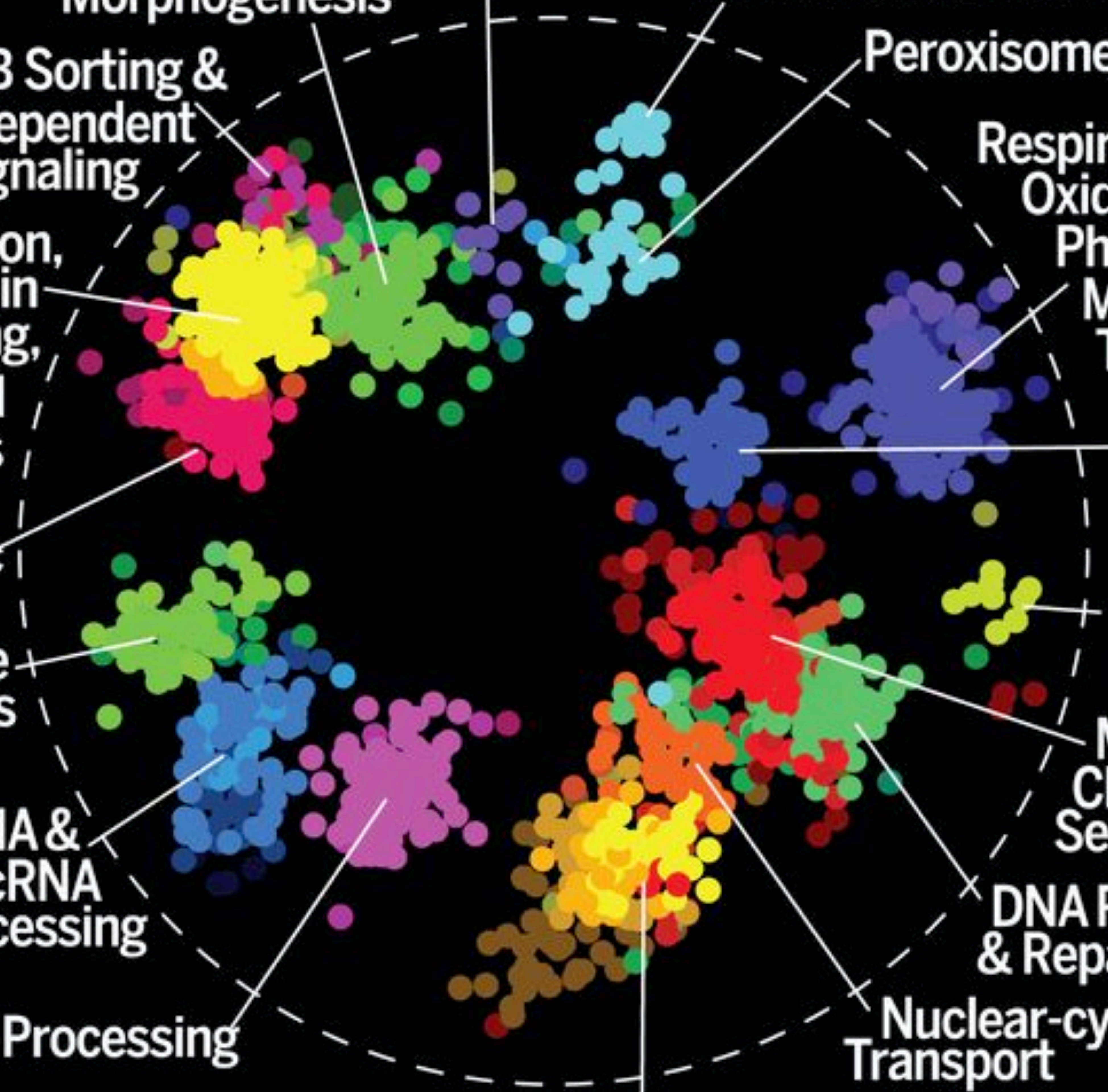


@JLSteenwyk

# Networks capture the complexity of genomic function

Baker's yeast






@JLSteenwyk

# PEX1 and PEX6 share function

Pex1p & Pex6p: forms a heterodimer involved in recycling peroxisomal signal receptor Pex5p

@JLSteenwyk

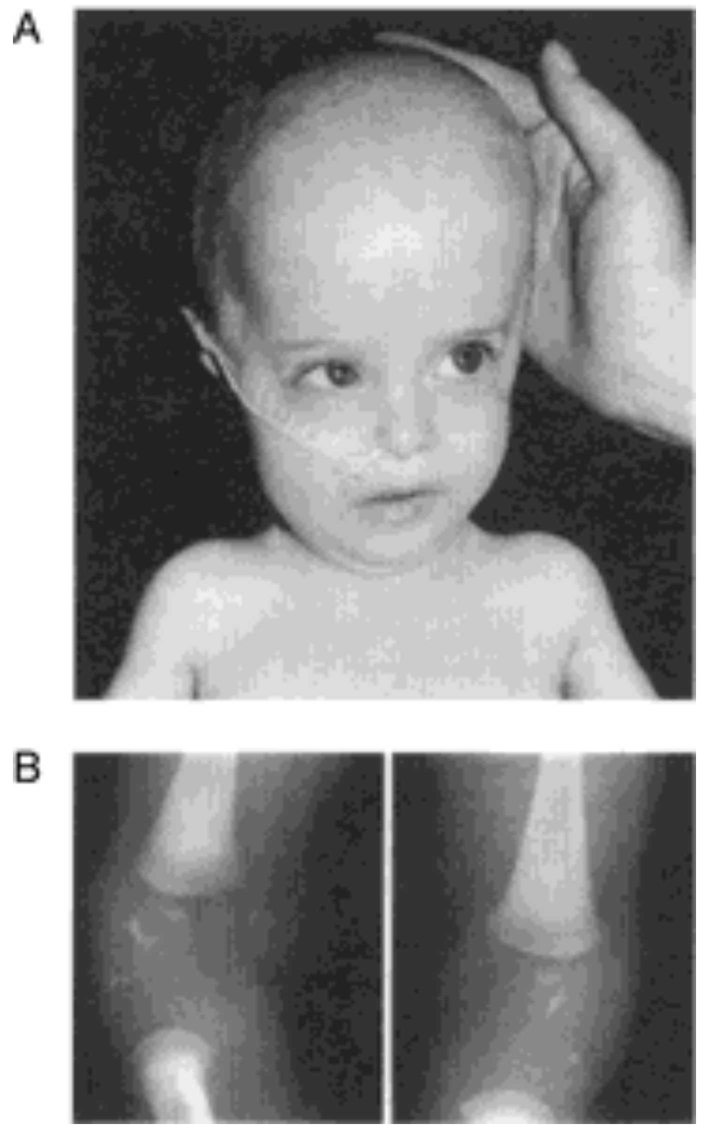
Geisbrecht et al. (1998), PMID: 9671729

# PEX1 and PEX6 share function

Pex1p & Pex6p: forms a heterodimer involved in recycling peroxisomal signal receptor Pex5p

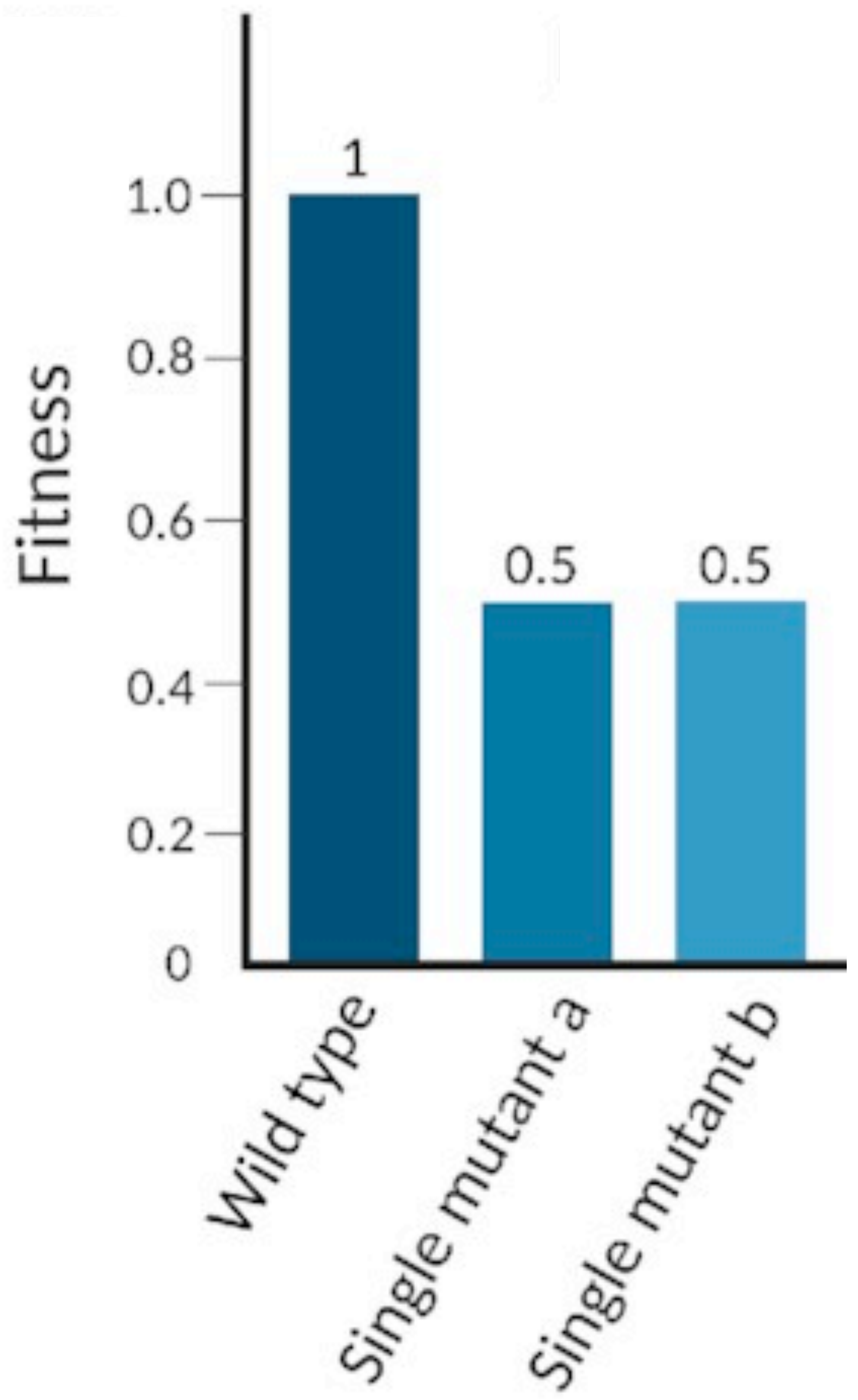
Mutations that disrupt protein interactions cause neurologic disorders including:

- Zellweger syndrome,
- neonatal adrenoleukodystrophy,
- infantile Refsum disease




@JLSteenwyk

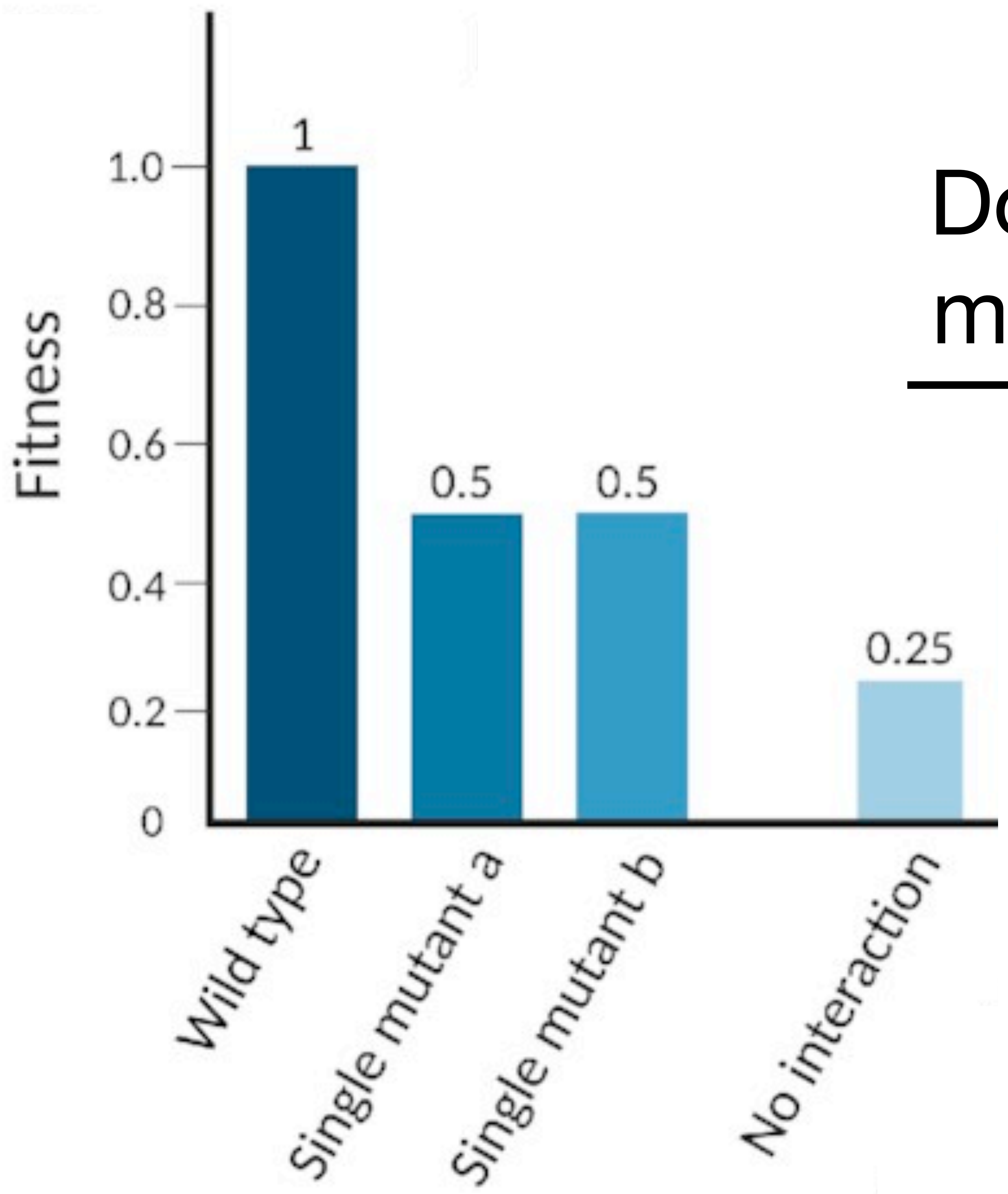
# Genetic interactions are gene-gene associations




@JLSteenwyk

Halder et al. (2021), PMID: 33145589

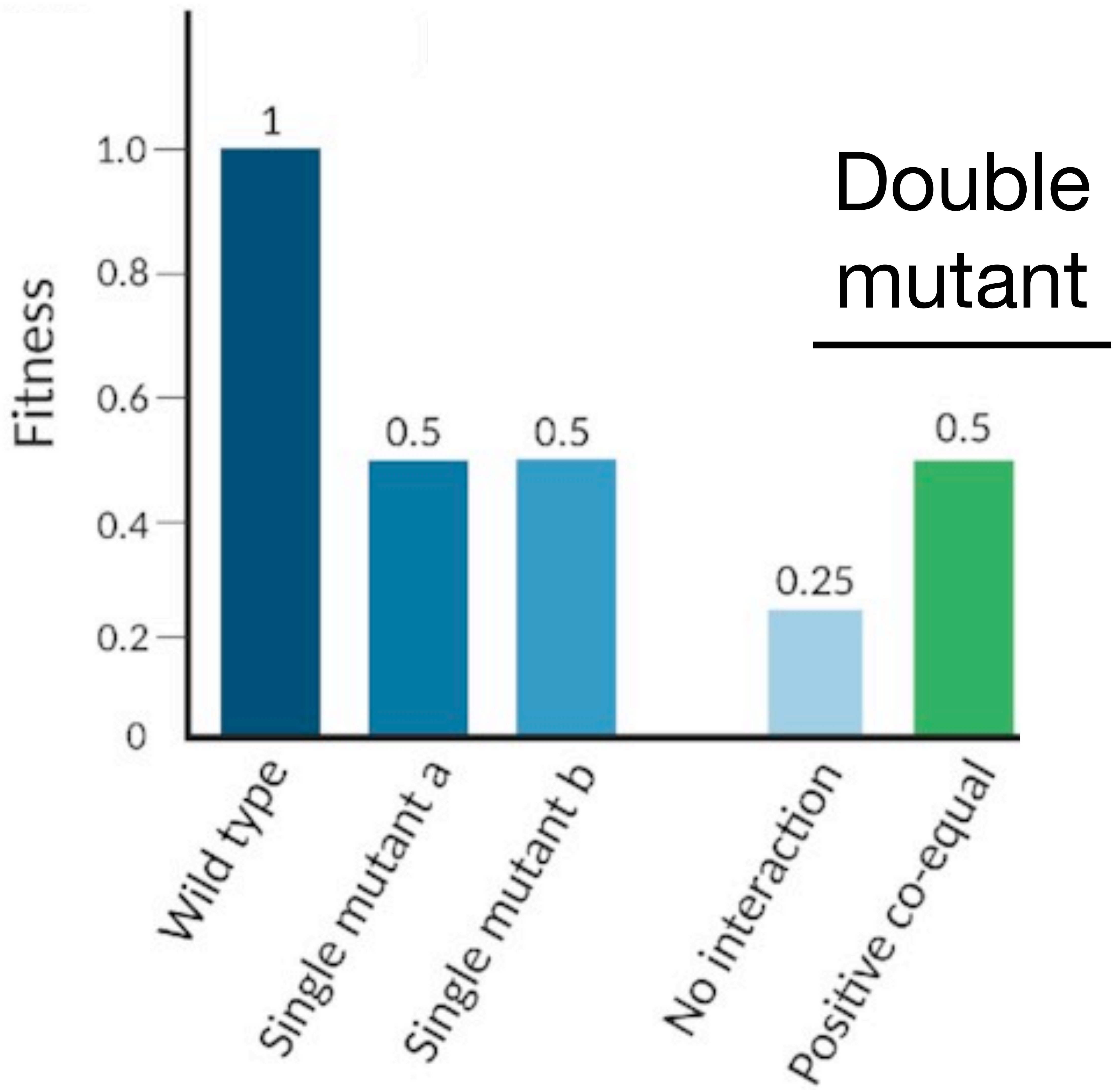
# Genetic interactions are gene-gene associations




Double mutant

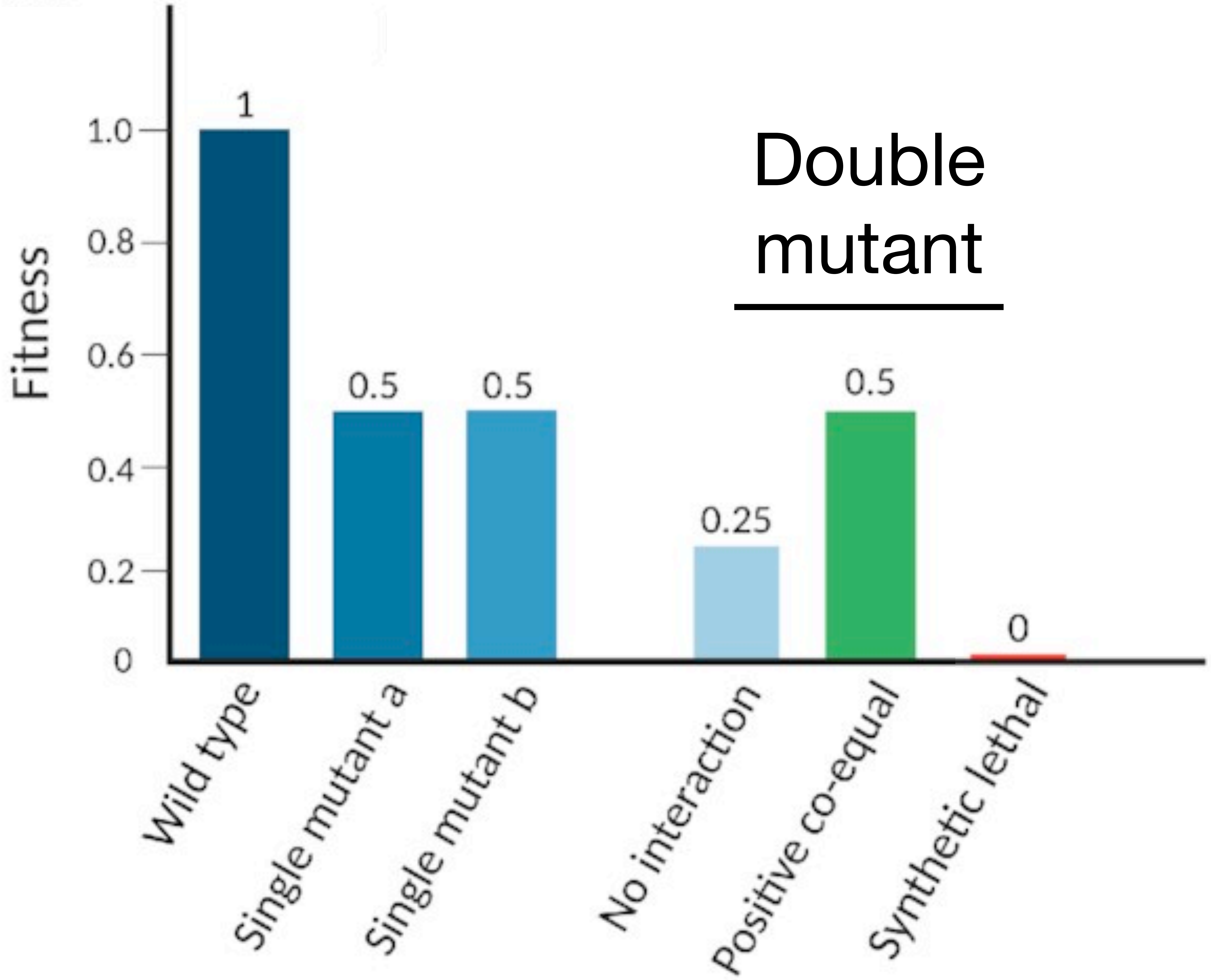
@JLSteenwyk

# Genetic interactions are gene-gene associations




@JLSteenwyk

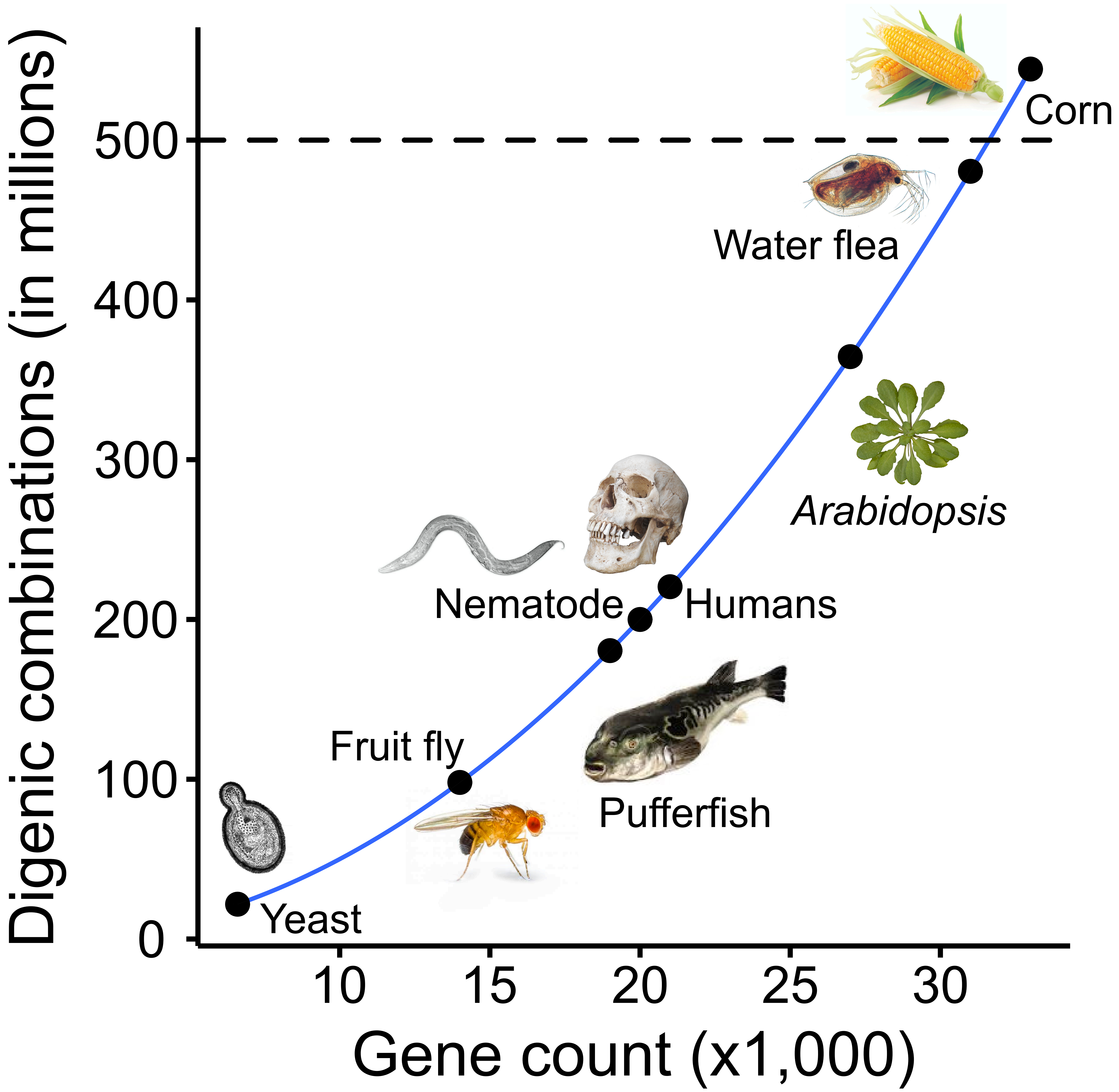
# Genetic interactions are gene-gene associations




@JLSteenwyk

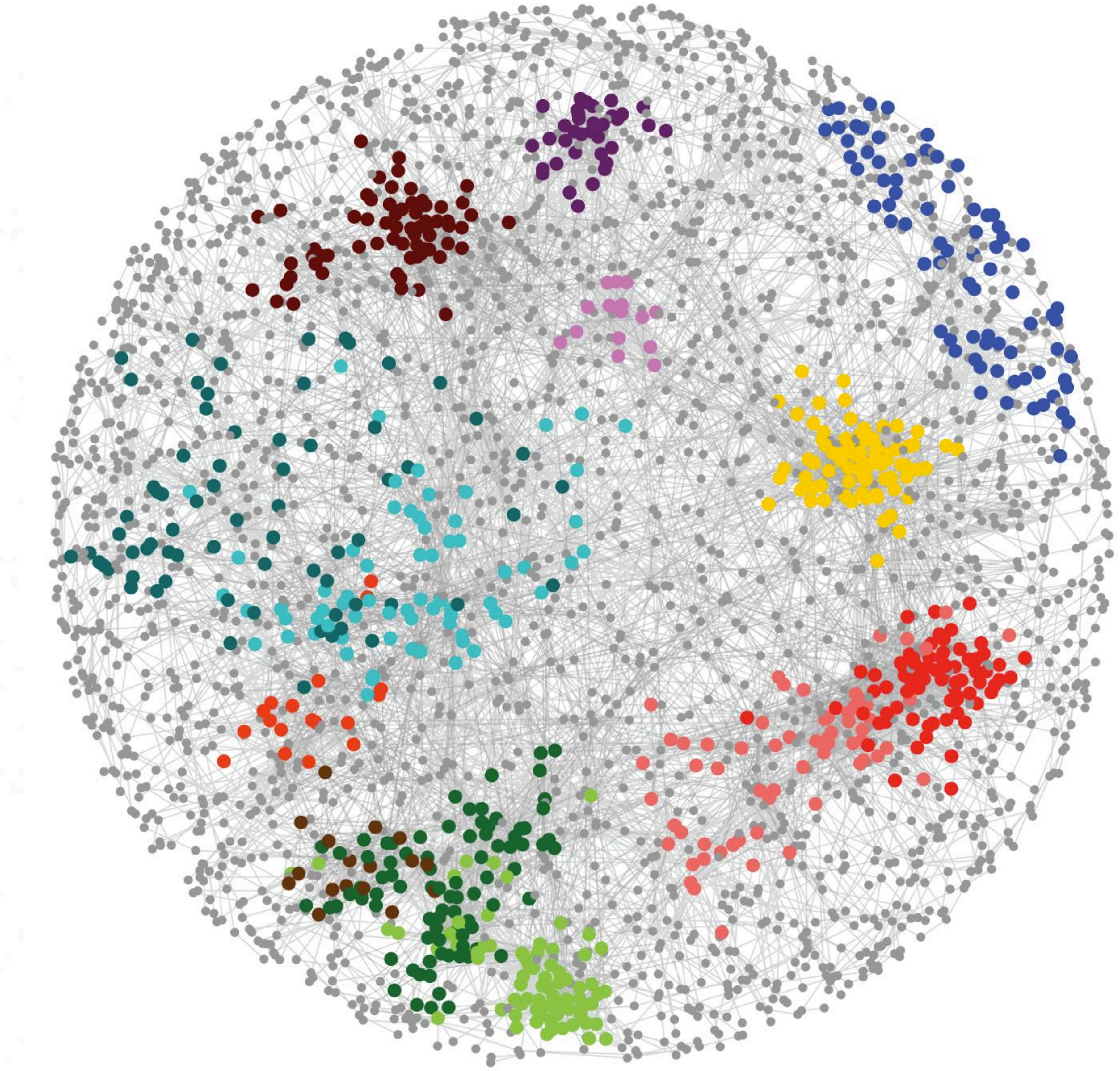
Halder et al. (2021), PMID: 33145589

# Too many pairwise combinations of genes



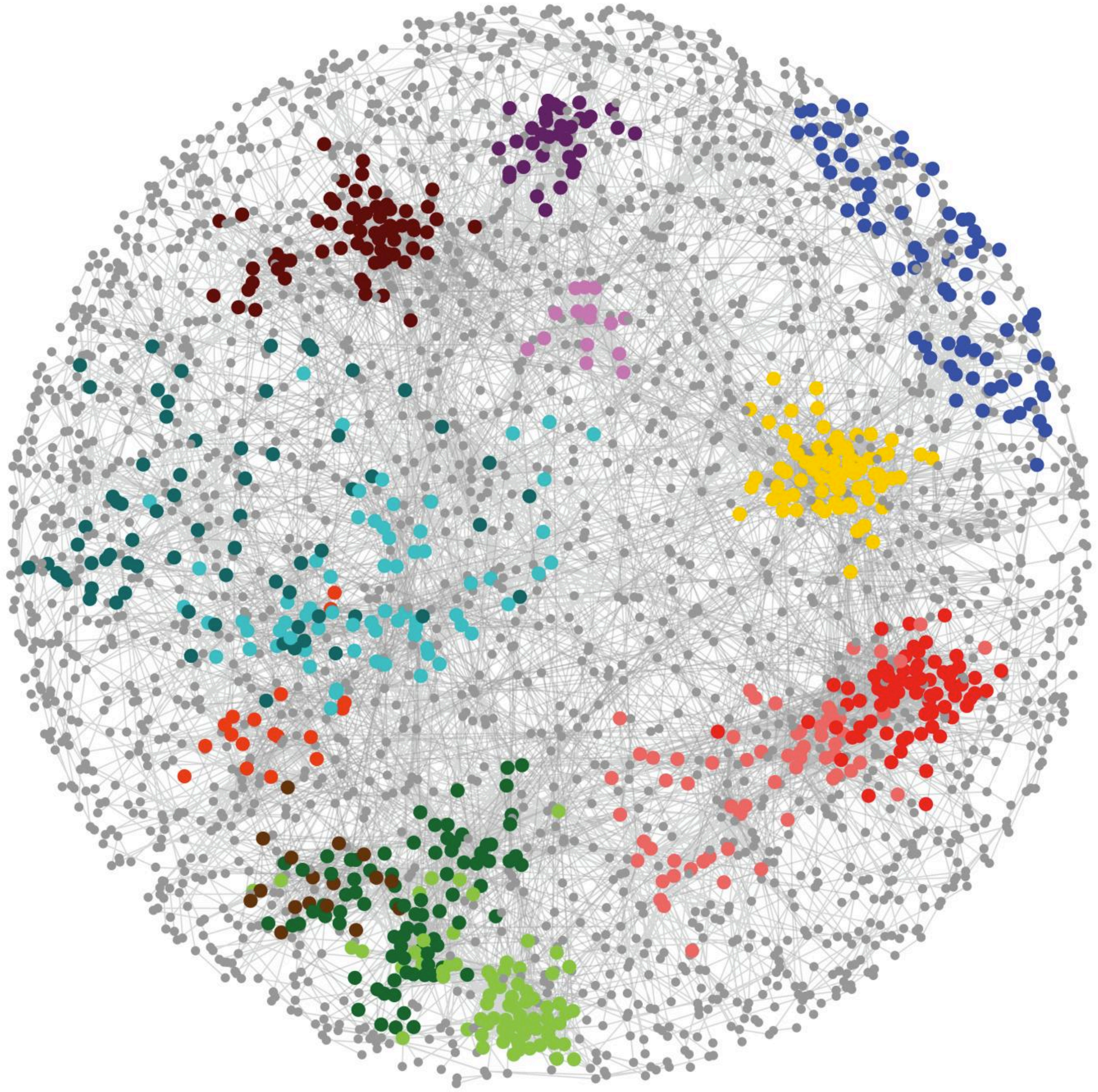

@JLSteenwyk

# Methods to infer gene-gene associations



- Coexpression
- Gene presence/ absence patterns

# Methods to infer gene-gene associations

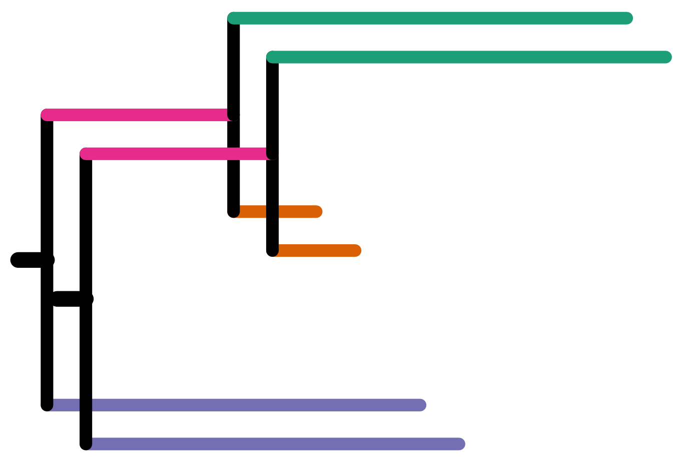


- Coexpression
- Gene presence/ absence patterns
- **Gene coevolution**

@JLSteenwyk

# Gene-gene coevolution predicts shared function

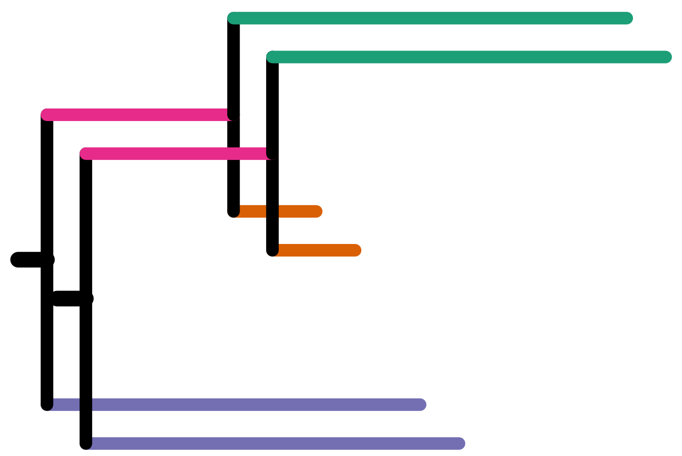
- gene coevolution refers to:
  - two genes that covary in parallel across speciation events
  - often observed among genes that share function, are coexpressed, or are part of the same multi-meric complexes



@JLSteenwyk

# Gene-gene coevolution predicts shared function

- gene coevolution refers to:
  - two genes that covary in parallel across speciation events
  - often observed among genes that share function, are coexpressed, or are part of the same multi-meric complexes



PhyKIT

a toolkit for examining multiple sequence alignments and trees

**PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data**

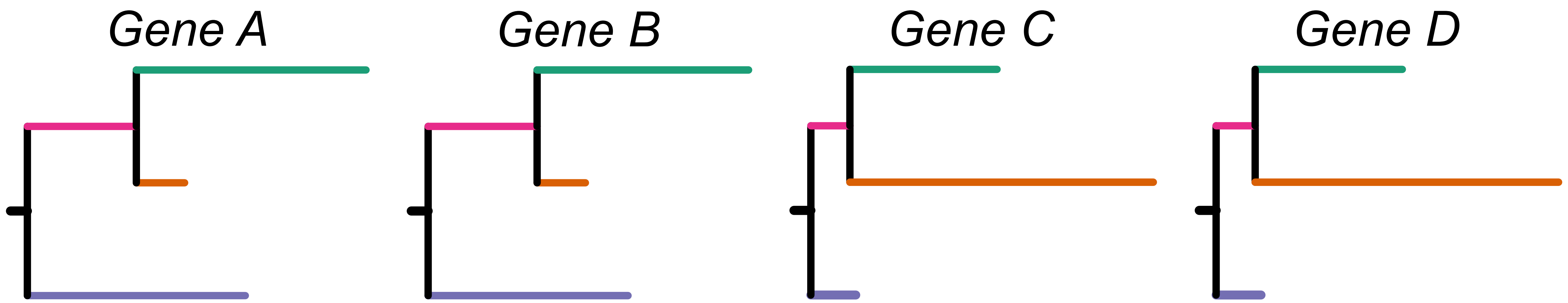
Jacob L Steenwyk, Thomas J Buida, III, Abigail L Labella, Yuanning Li, Xing-Xing Shen, Antonis Rokas

*Bioinformatics*, btab096, https://doi.org/10.1093/bioinformatics/btab096
**Published:** 09 February 2021 **Article history ▾**

@JLSteenwyk

# The mirror principle to detect gene coevolution

**Gene trees**



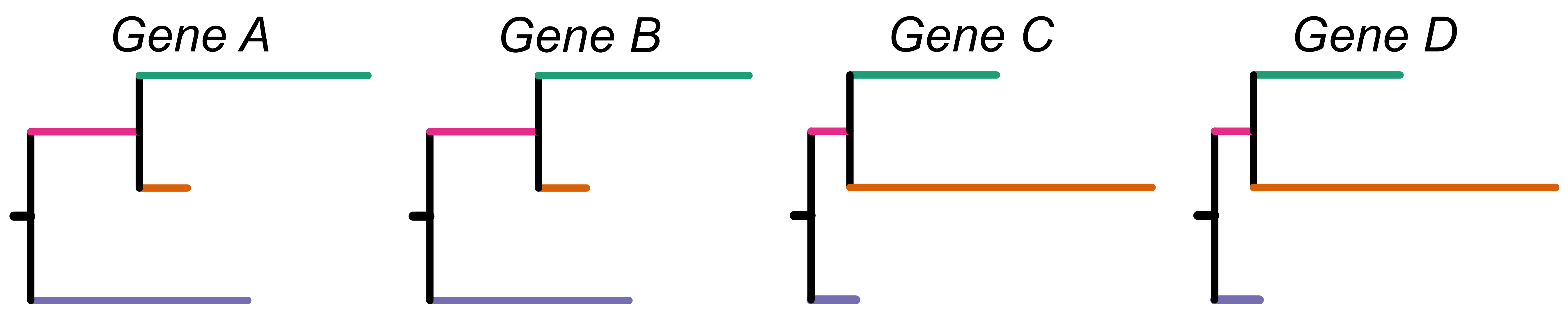

@JLSteenwyk

Steenwyk et al. (2021), PMID: 33560364

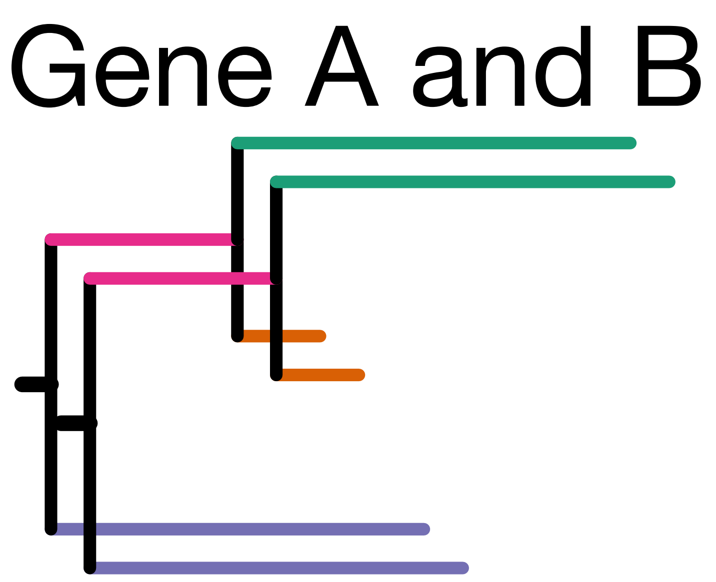
*gene trees are constrained to the species tree topology; mathematical corrections and transformations not shown

# The mirror principle to detect gene coevolution

**Gene trees**




**The 'mirror' principle**

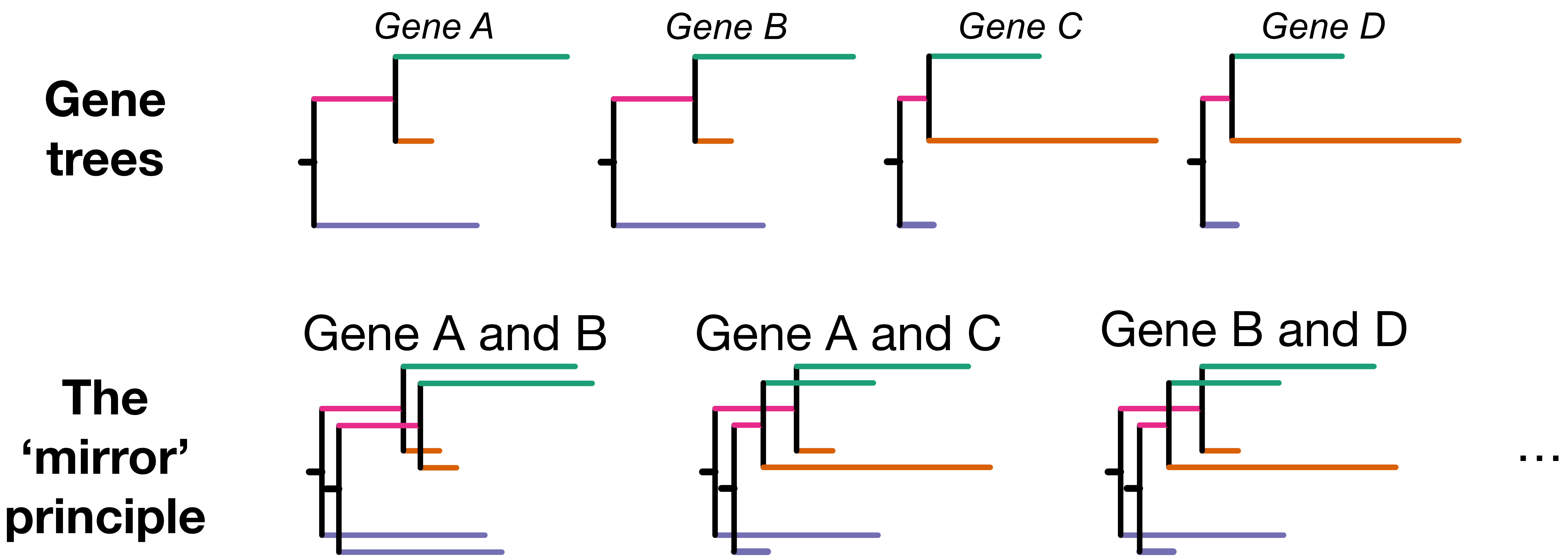



@JLSteenwyk

Steenwyk et al. (2021), PMID: 33560364

*gene trees are constrained to the species tree topology; mathematical corrections and transformations not shown

# The mirror principle to detect gene coevolution

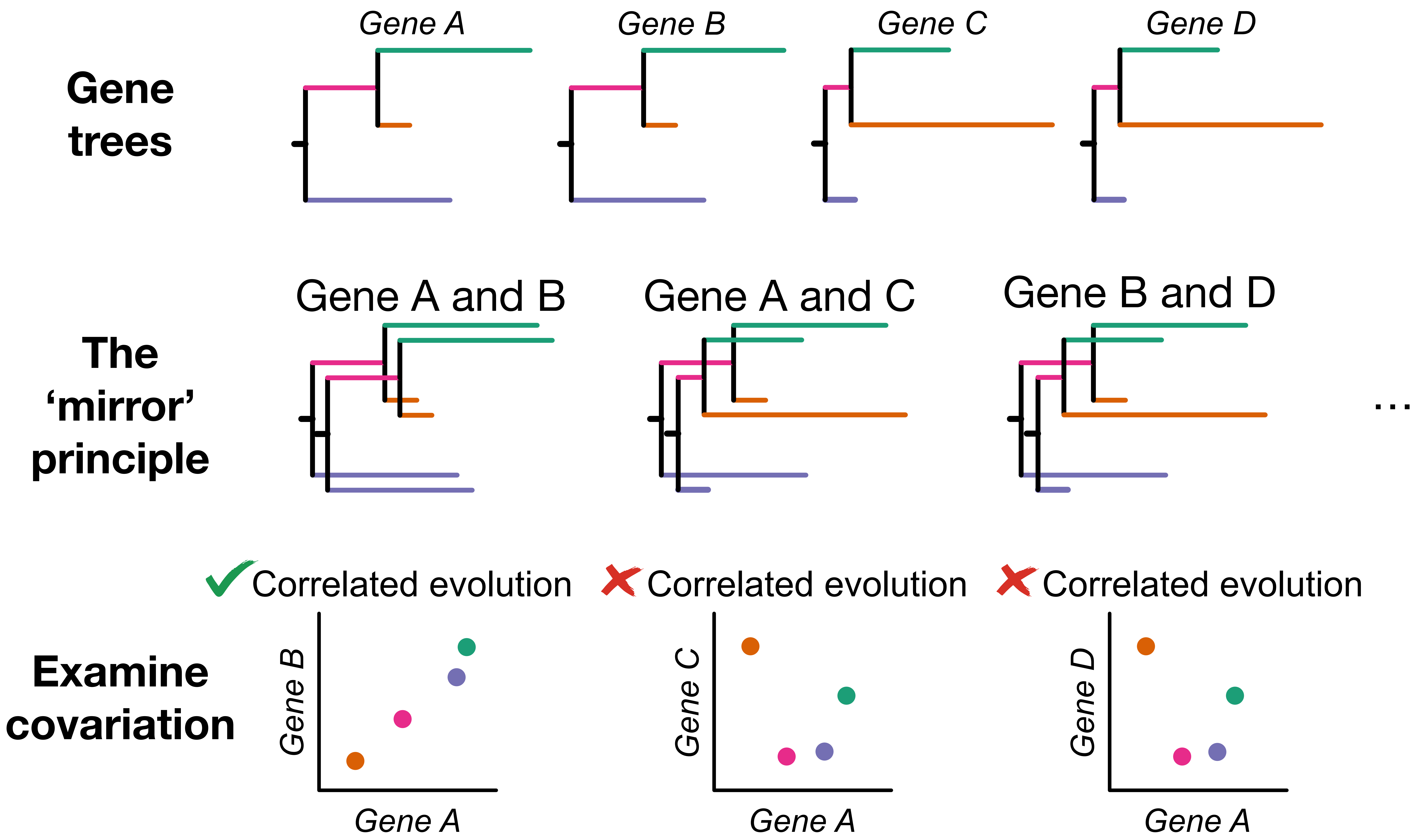



@JLSteenwyk



*gene trees are constrained to the species tree topology; mathematical corrections and transformations not shown

# The mirror principle to detect gene coevolution




@JLSteenwyk

Steenwyk et al. (2021), PMID: 33560364

*gene trees are constrained to the species tree topology; mathematical corrections and transformations not shown

# Genes of a feather evolve together

✓ Correlated evolution

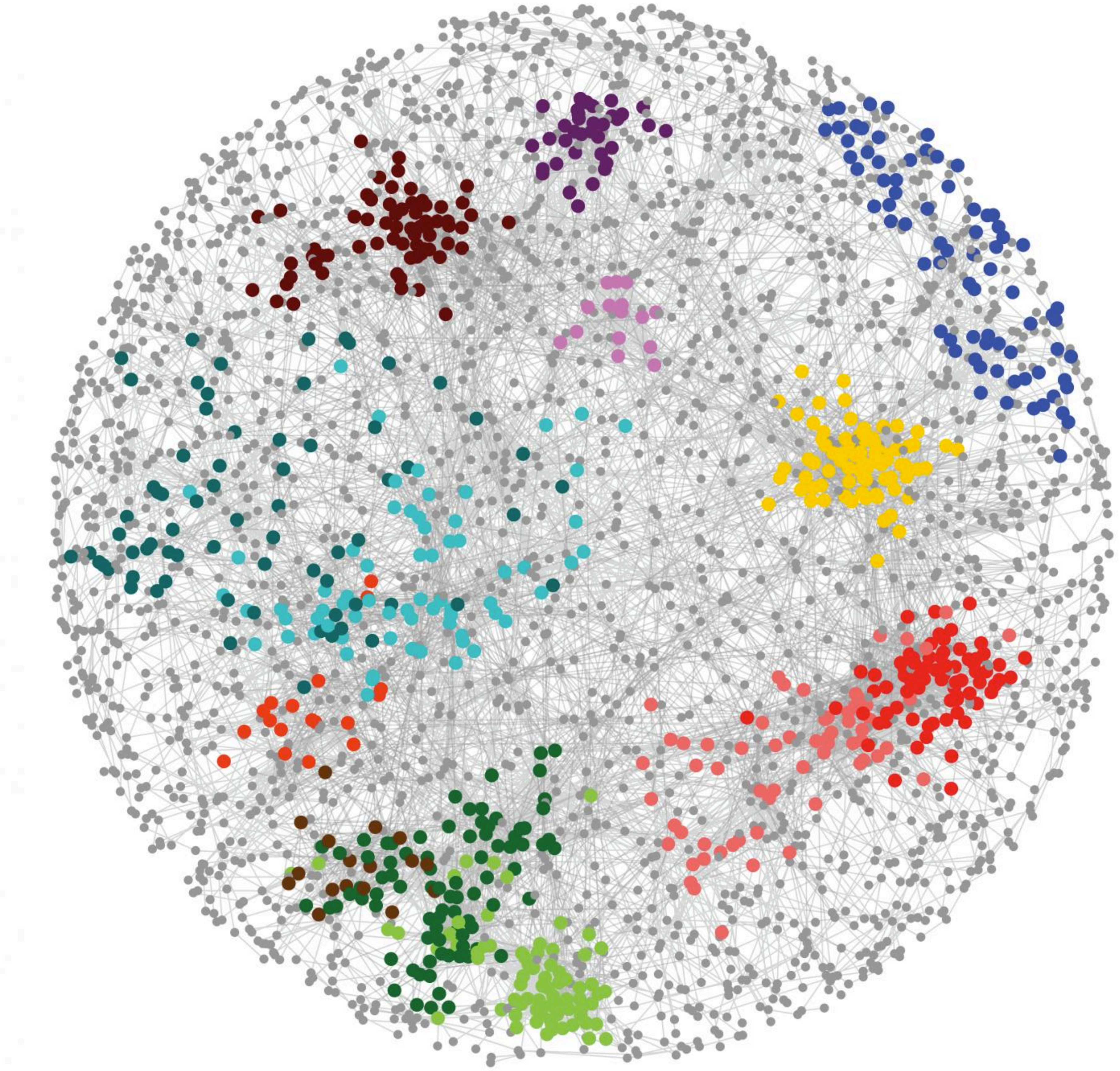

- Coevolving genes tend to share function, be coexpressed, or are part of the same multimeric complexes



Steenwyk et al. (2021), PMID: 33560364

*gene trees are constrained to the species tree topology; mathematical corrections and transformations not shown

# Genes of a feather evolve together



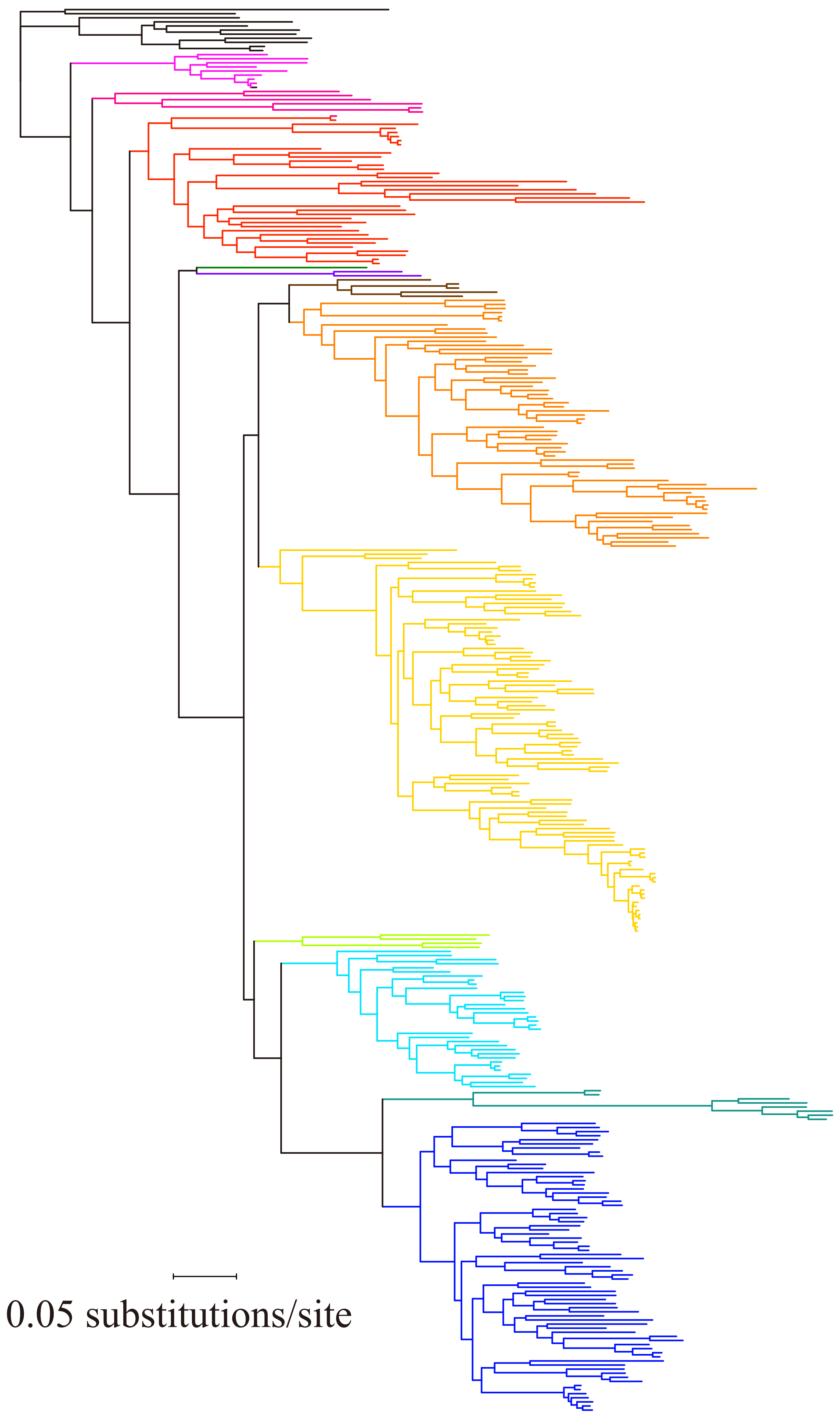
- Coevolving genes tend to share function, be coexpressed, or are part of the same multimeric complexes
- **But can we build a genetic network?**

@JLSteenwyk

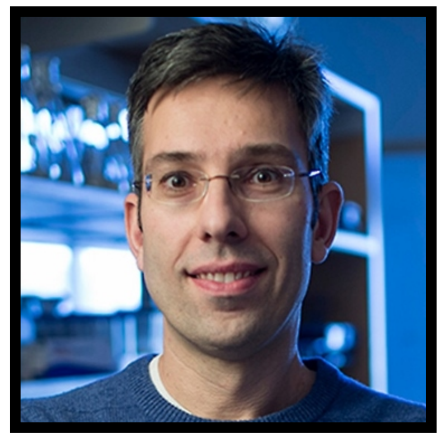


*gene trees are constrained to the species tree topology; mathematical corrections and transformations not shown

# Saccharomycotina yeast




- Saccharomycotina, a budding model subphylum

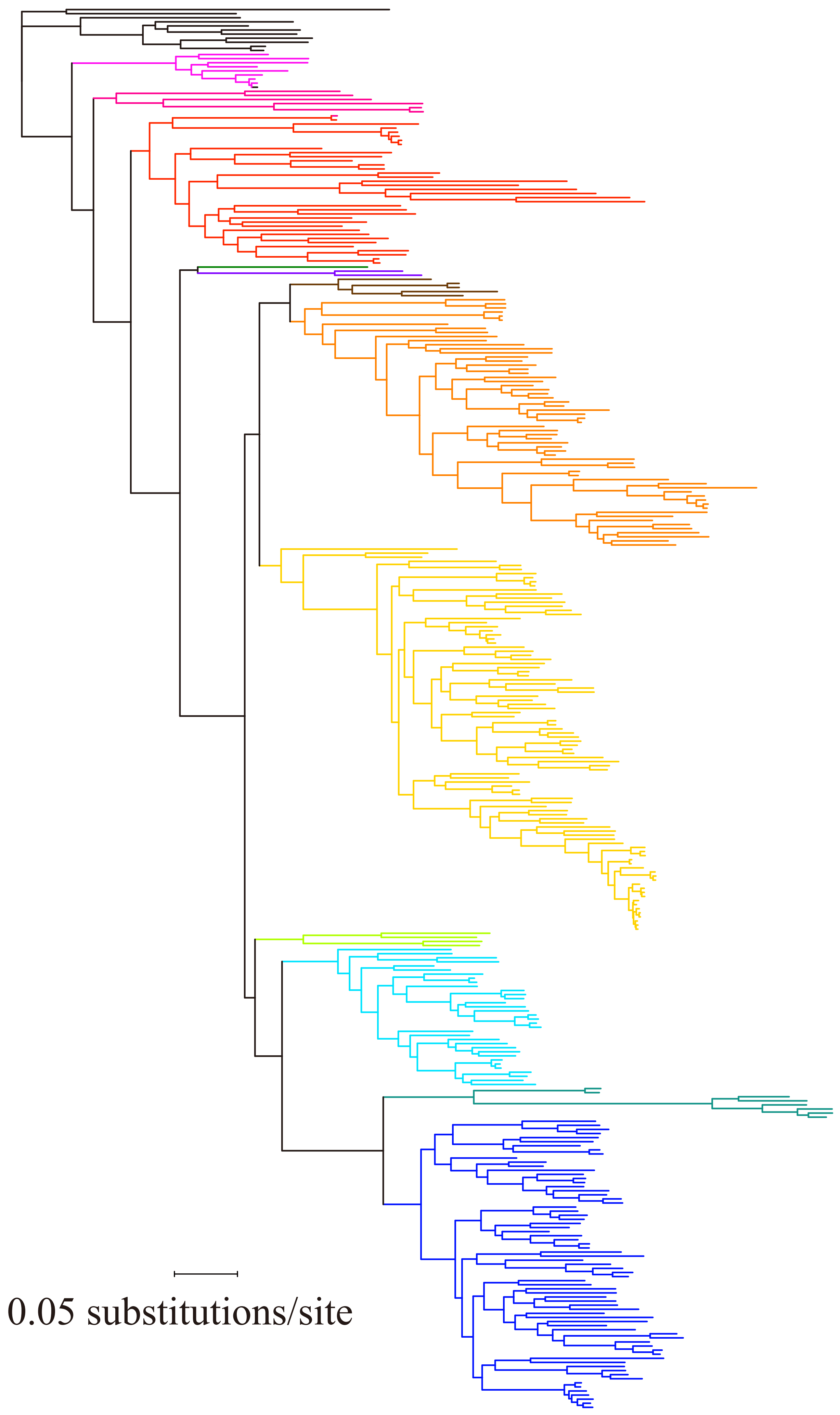


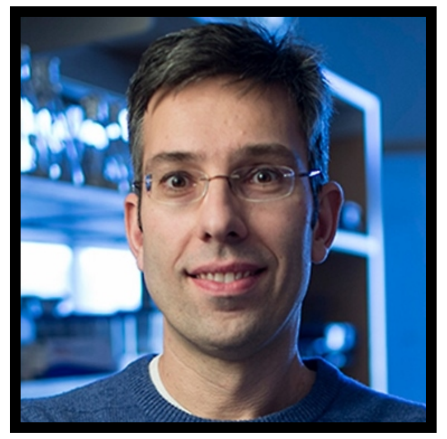
Antonis Rokas

Chris Hittinger

@JLSteenwyk

# Saccharomycotina yeast



Antonis Rokas

Chris Hittinger

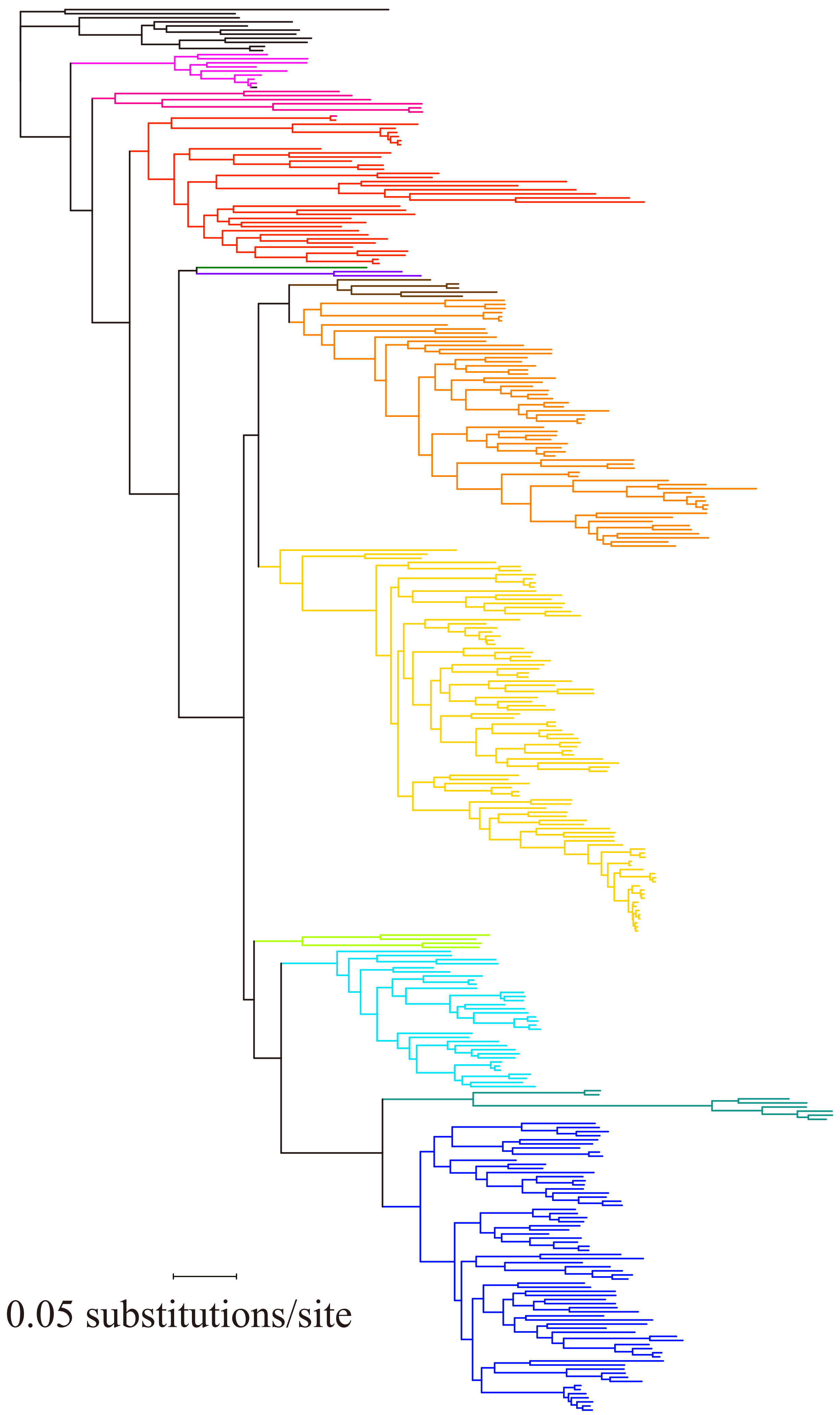
- Saccharomycotina, a budding model subphylum

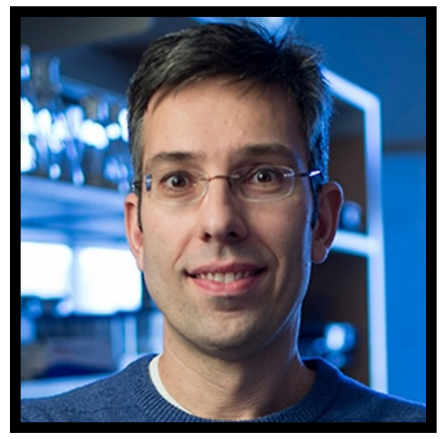
- Spans 332 species of budding yeast

@JLSteenwyk

Shen et al. (2018), PMID: 30415838

# Saccharomycotina yeast



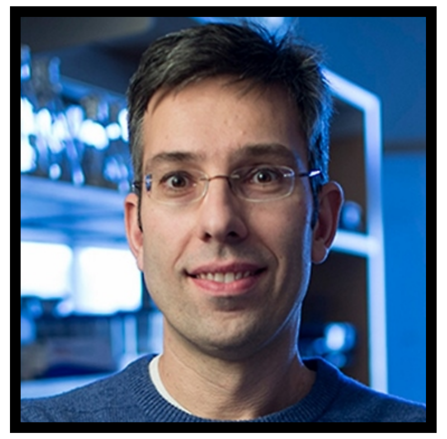
Antonis Rokas

Chris Hittinger

- Saccharomycotina, a budding model subphylum
- Spans 332 species of budding yeast
- 2,408 orthologous genes across all budding yeasts

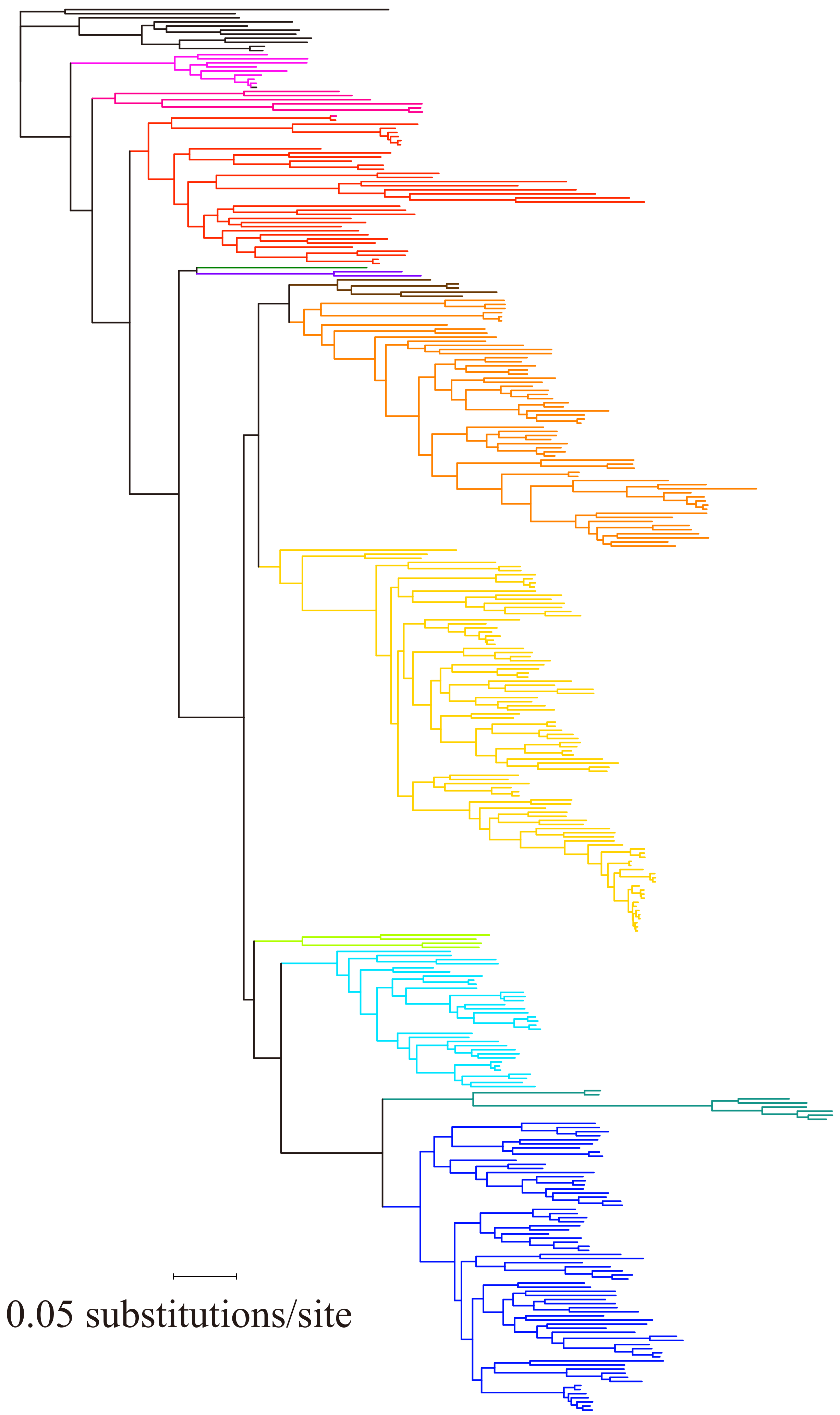
@JLSteenwyk

# Saccharomycotina yeast



Antonis Rokas

Chris Hittinger

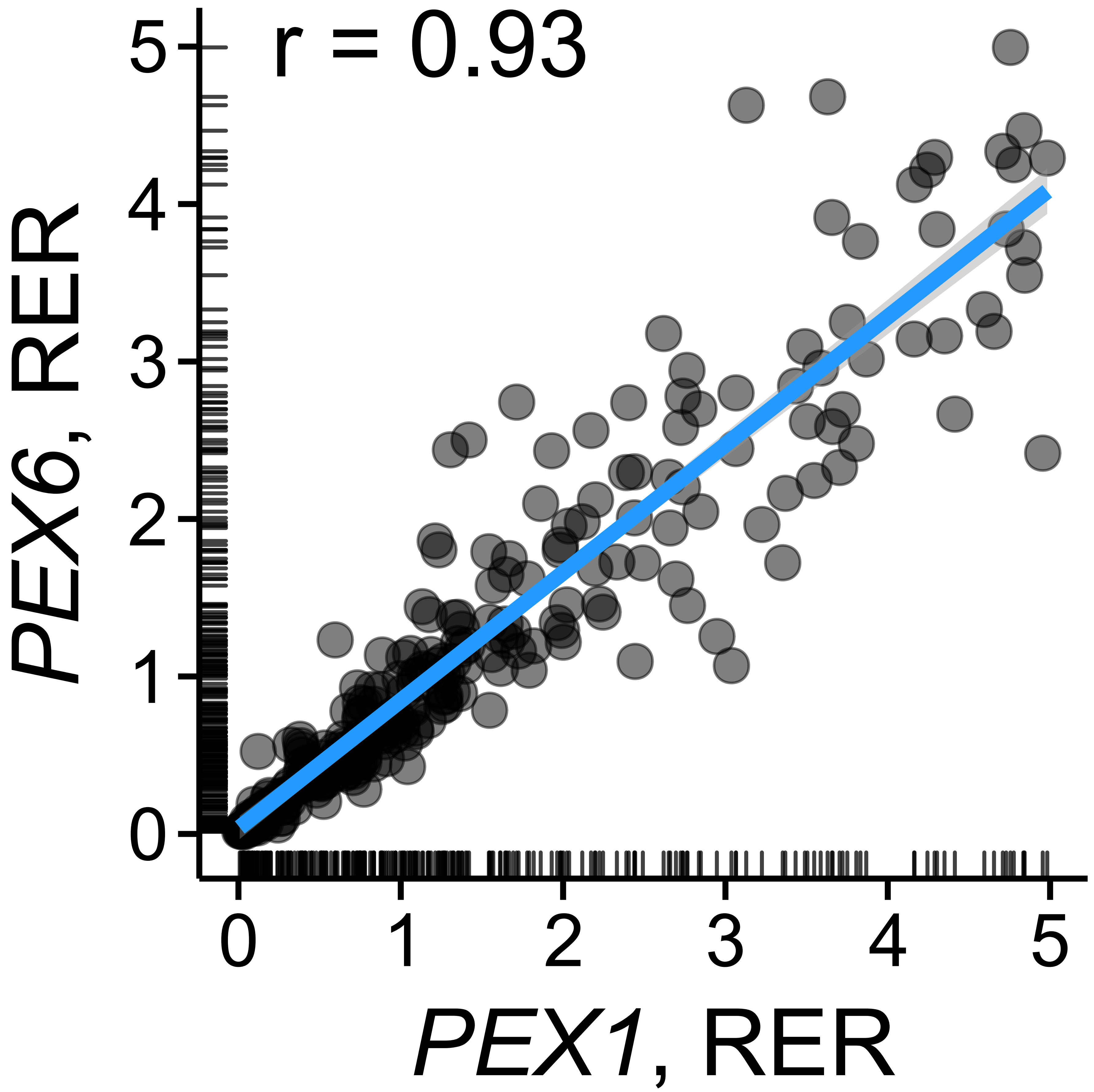



- Saccharomycotina, a budding model subphylum
- Spans 332 species of budding yeast
- 2,408 orthologous genes across all budding yeasts
- Calculate gene covariation across ~3 million pairwise combinations of genes

@JLSteenwyk

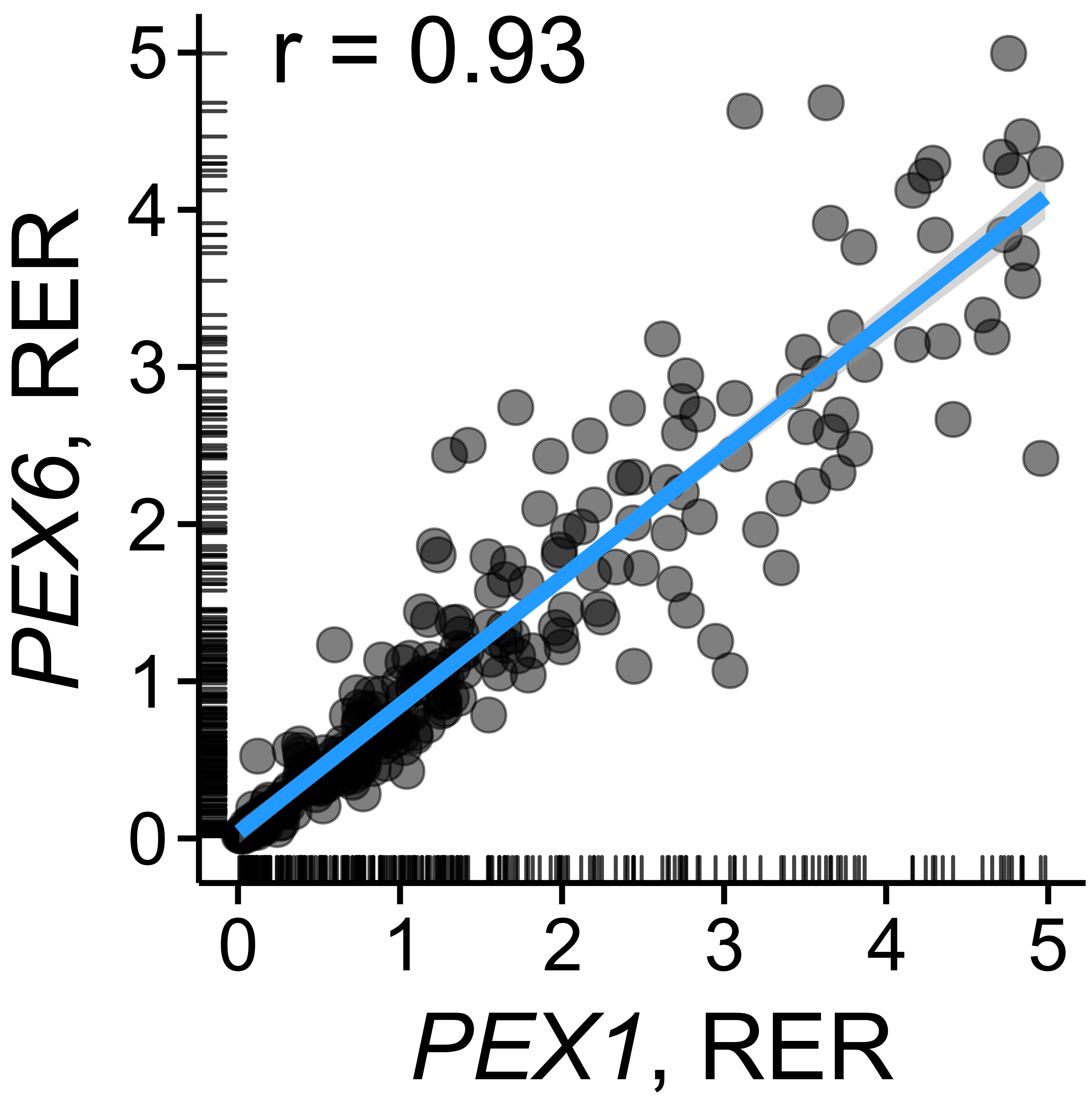
Shen et al. (2018), PMID: 30415838

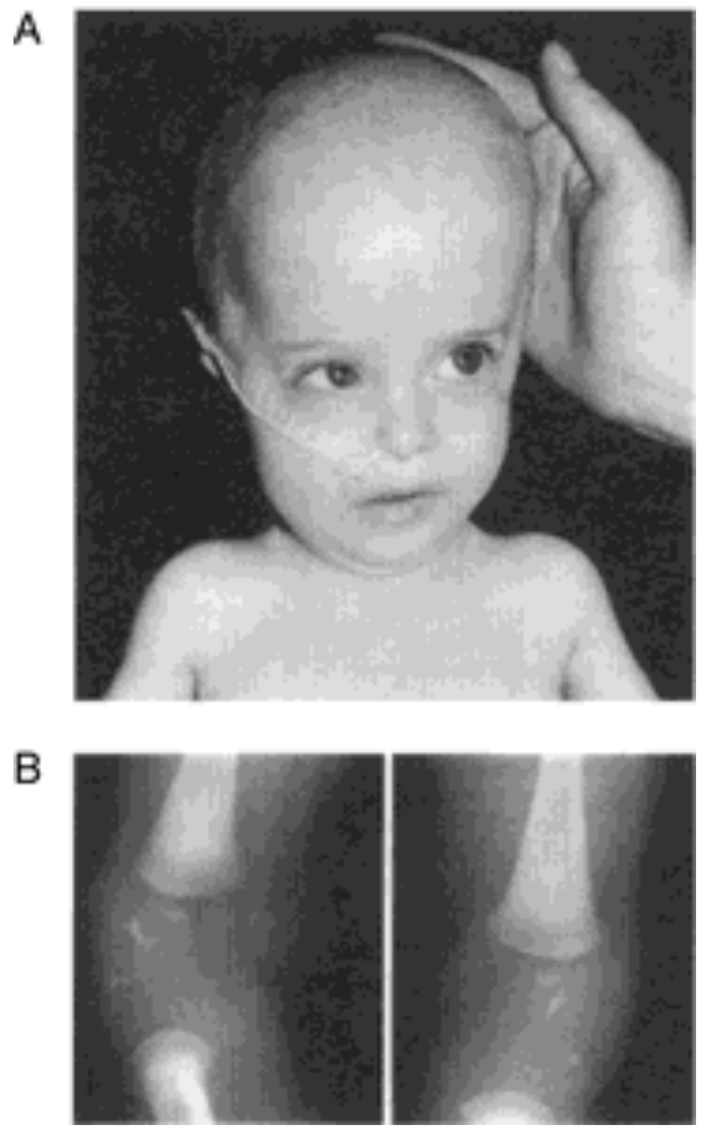
# PEX1 and PEX6 are coevolving




@JLSteenwyk

# PEX1 and PEX6 are coevolving

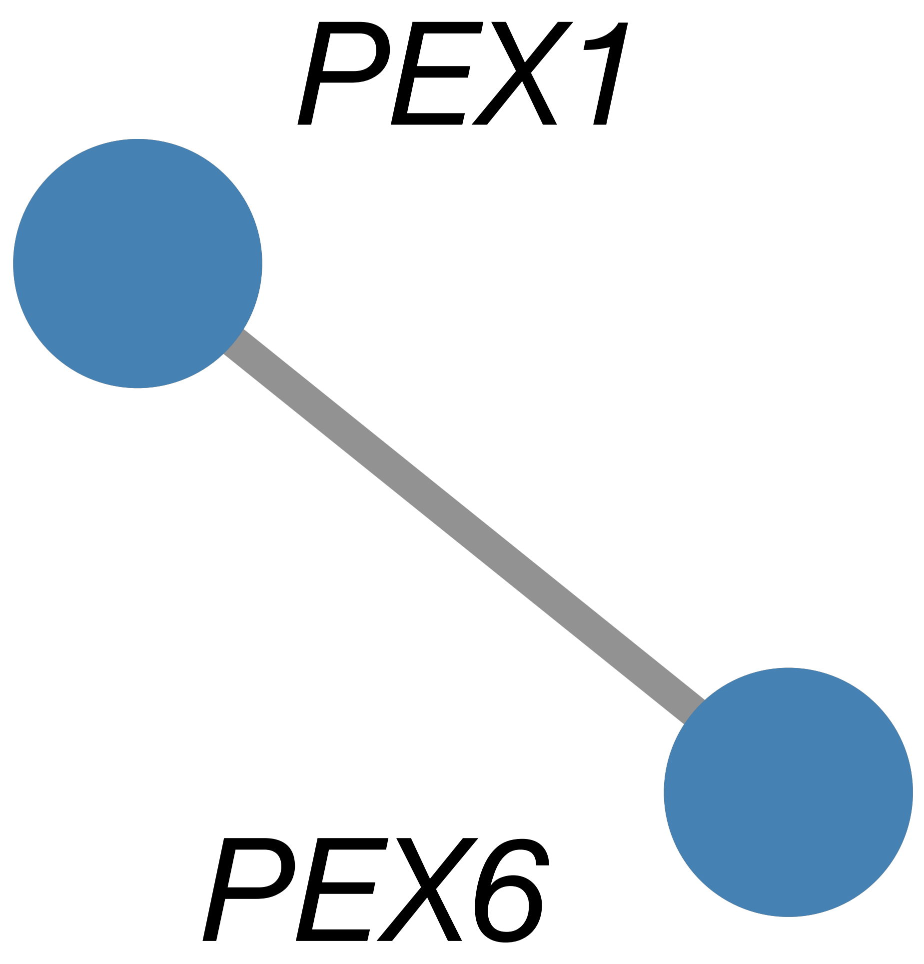



Pex1p & Pex6p: forms a heterodimer involved in recycling peroxisomal signal receptor Pex5p

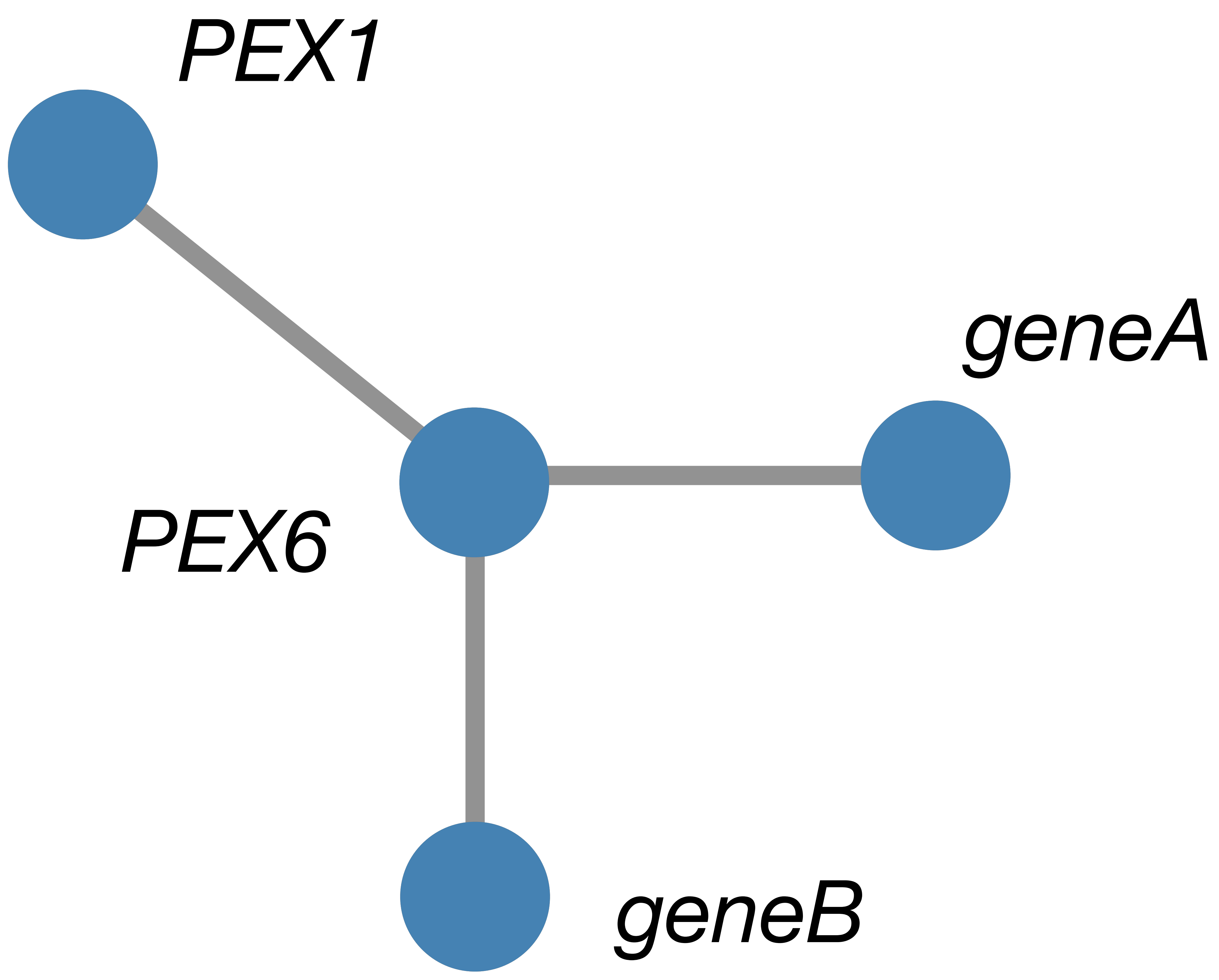
@JLSteenwyk

# Constructing a coevolutionary genetic network



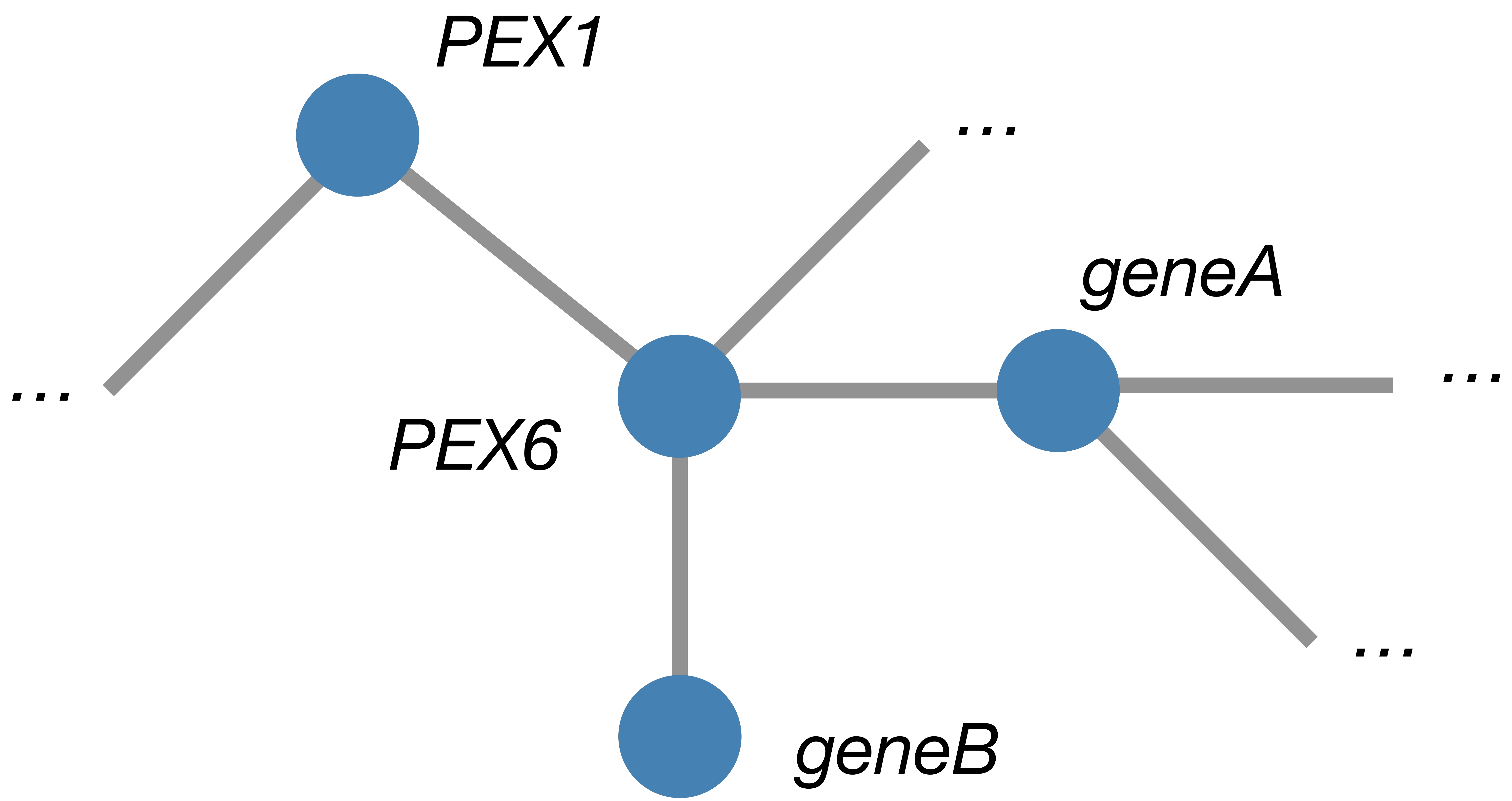

@JLSteenwyk

# Constructing a coevolutionary genetic network



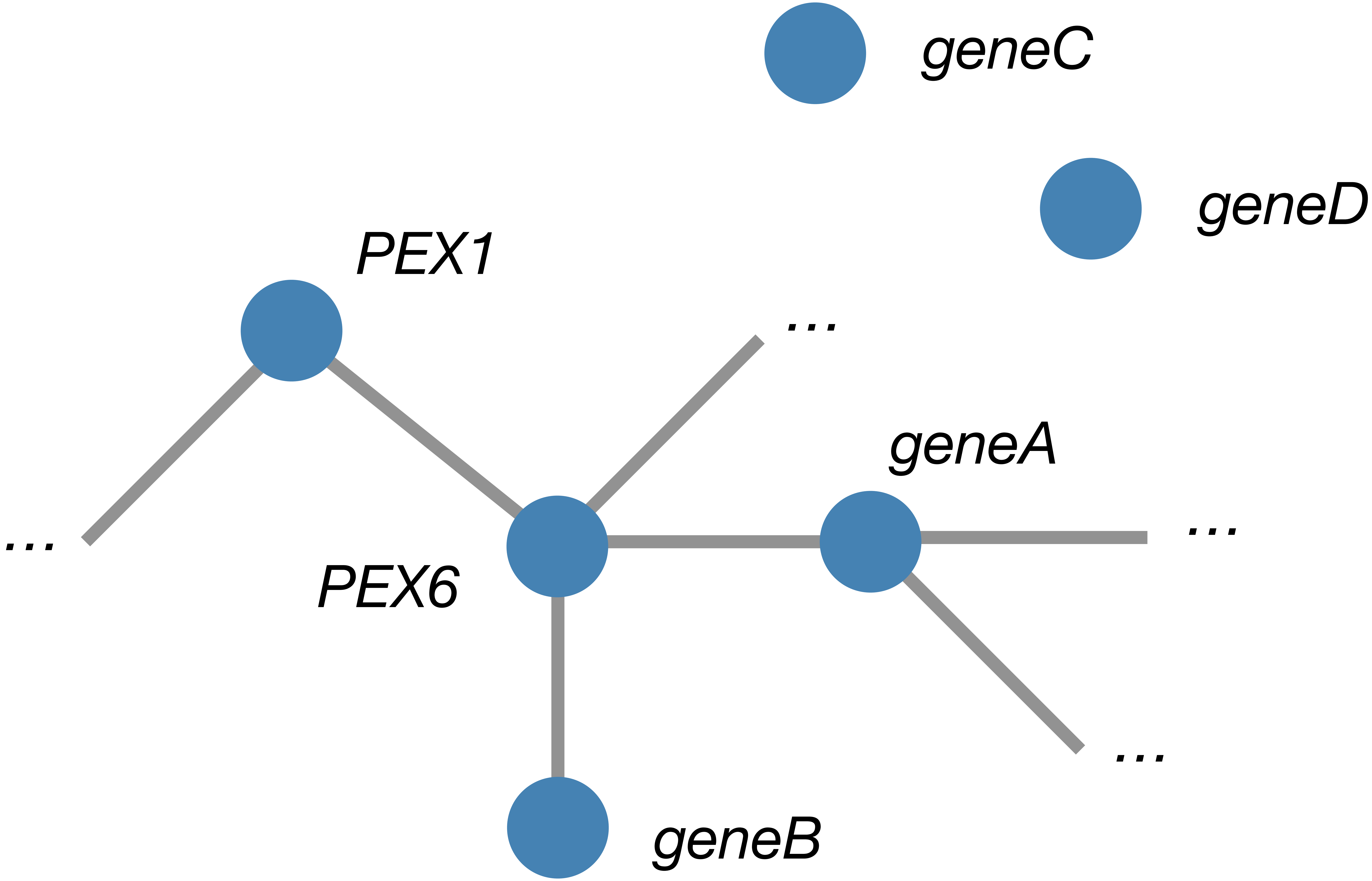

@JLSteenwyk

# Constructing a coevolutionary genetic network



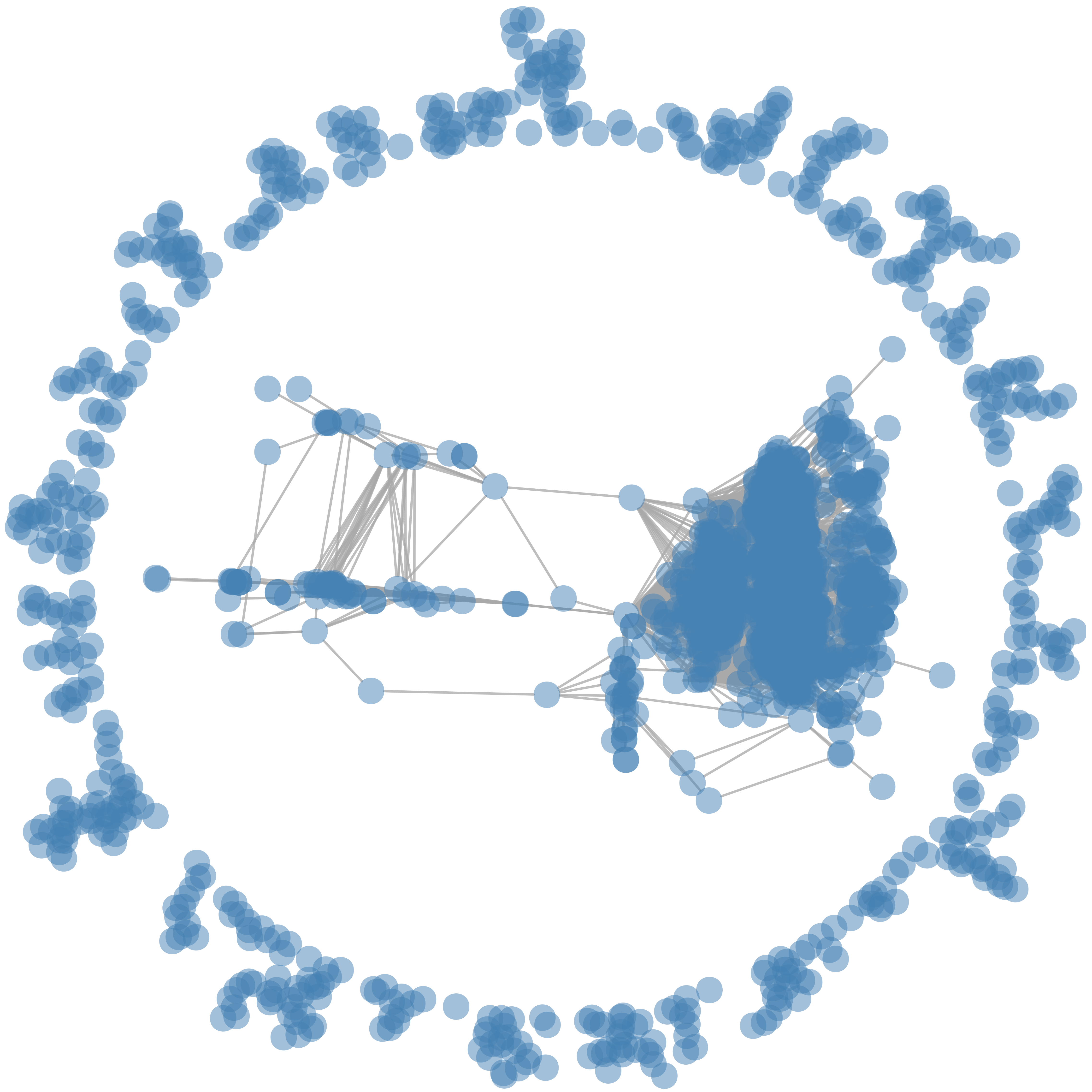

@JLSteenwyk

# Constructing a coevolutionary genetic network




@JLSteenwyk

# A global gene coevolutionary network

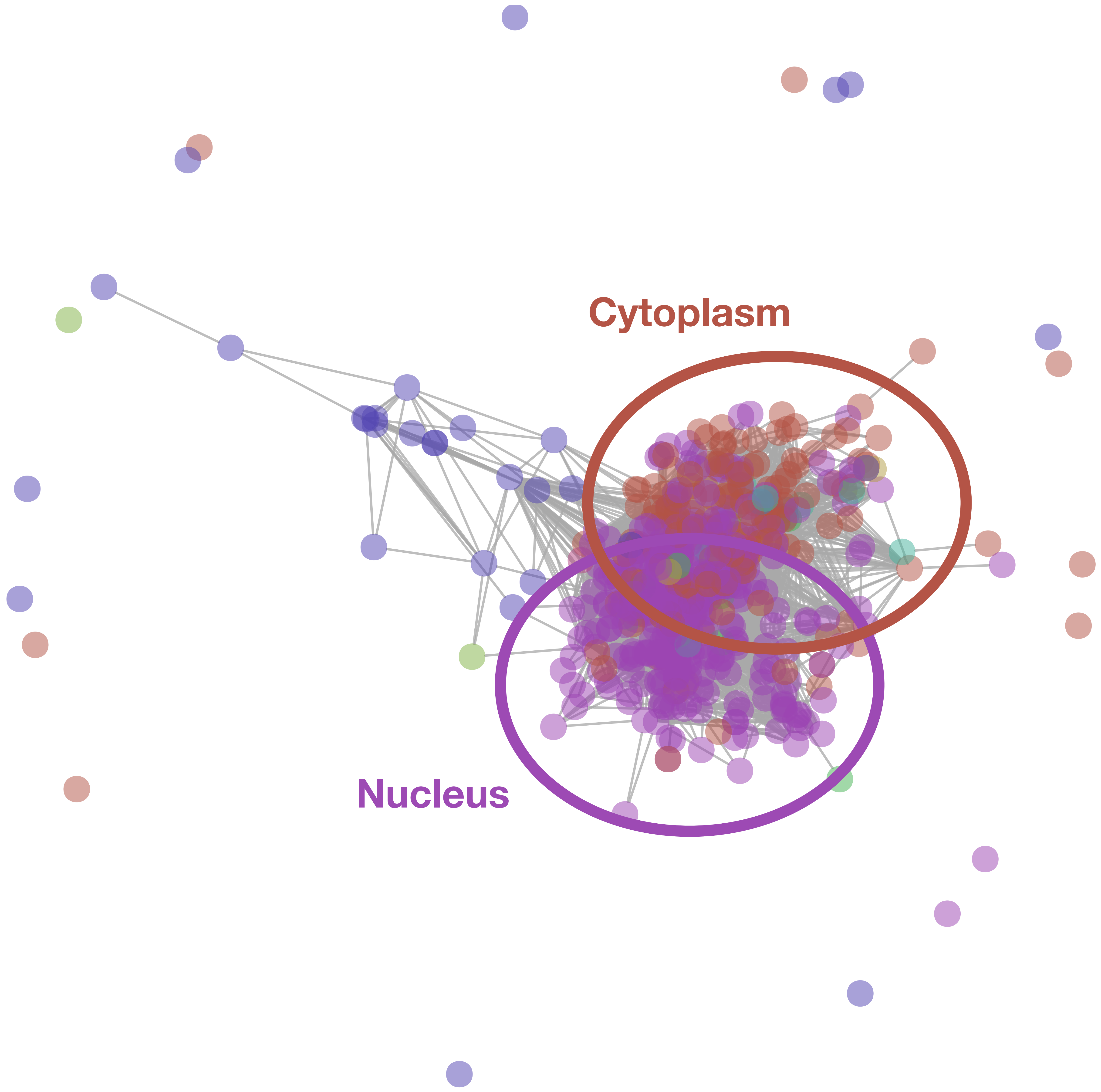


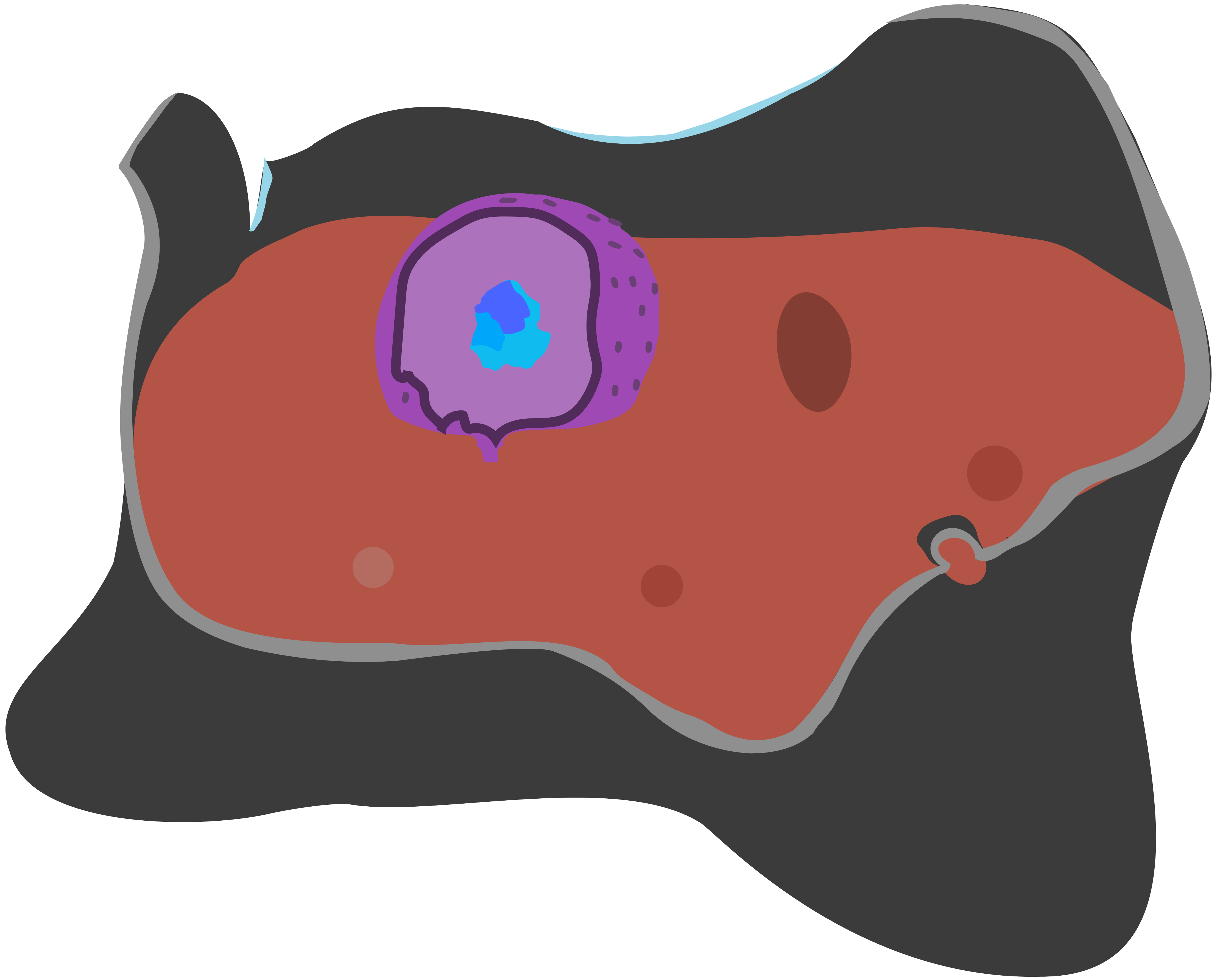
Nodes are genes

Edges connected coevolving genes

@JLSteenwyk

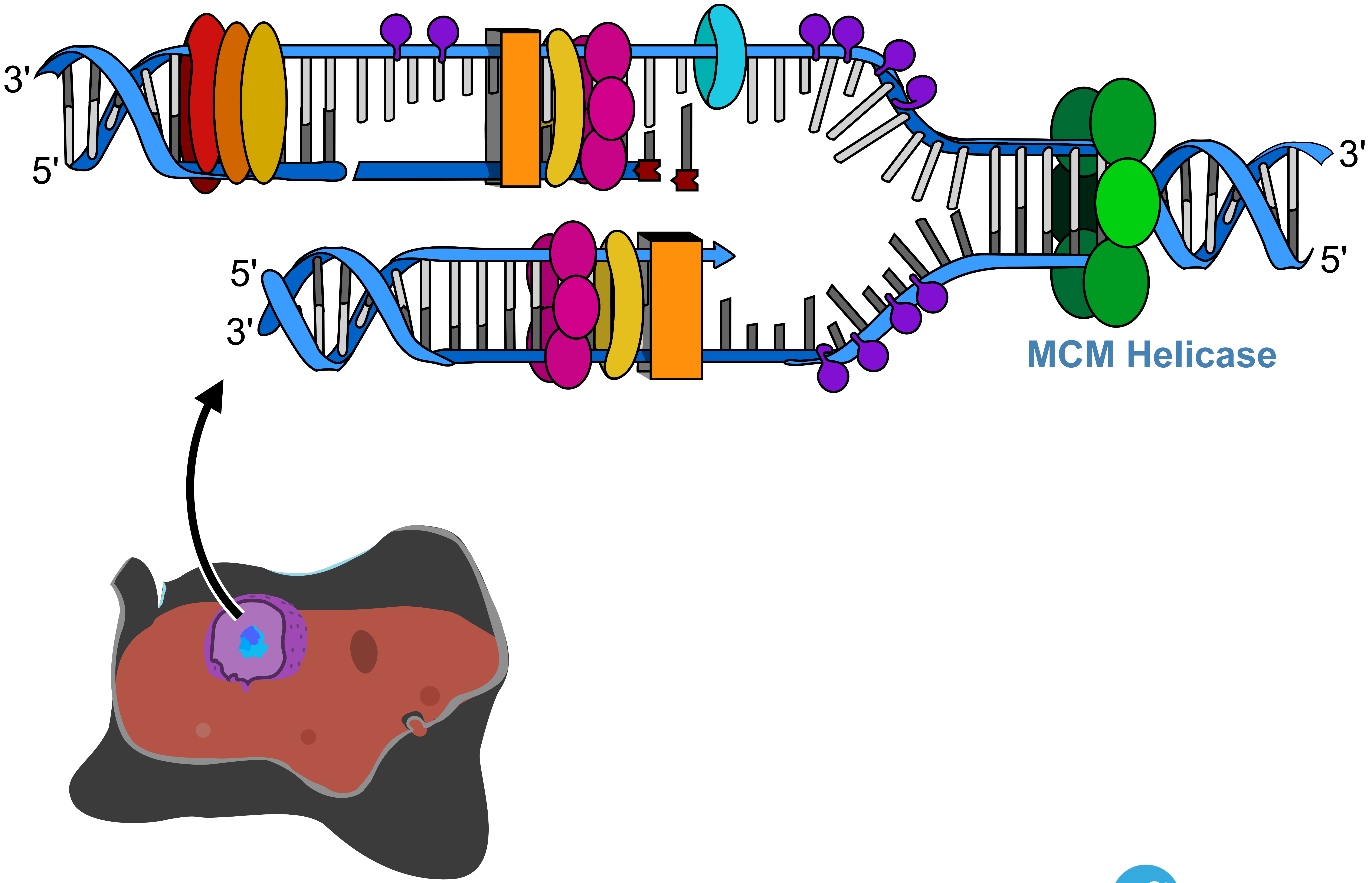
# Network reflections of cellular structure



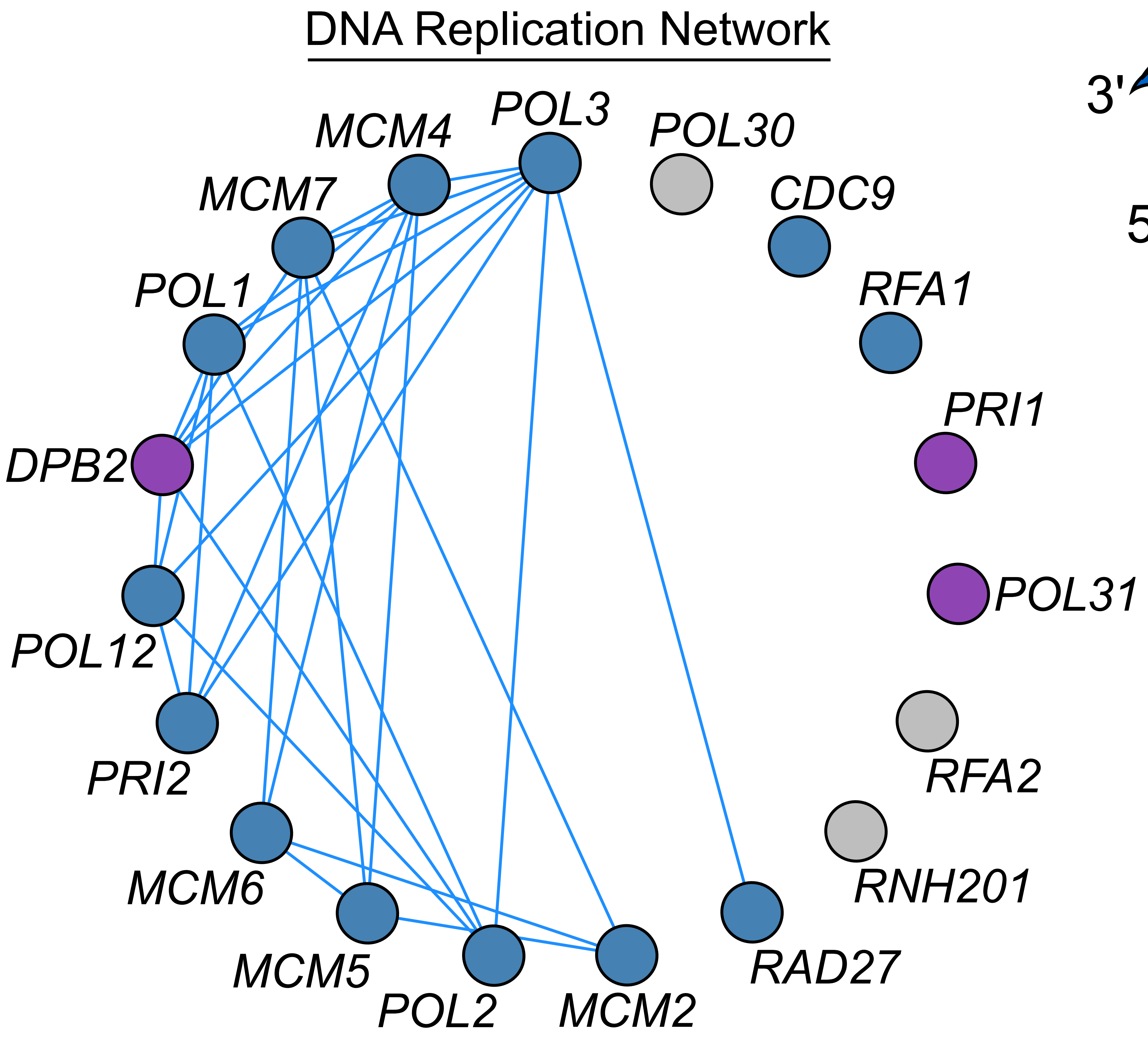
@JLSteenwyk

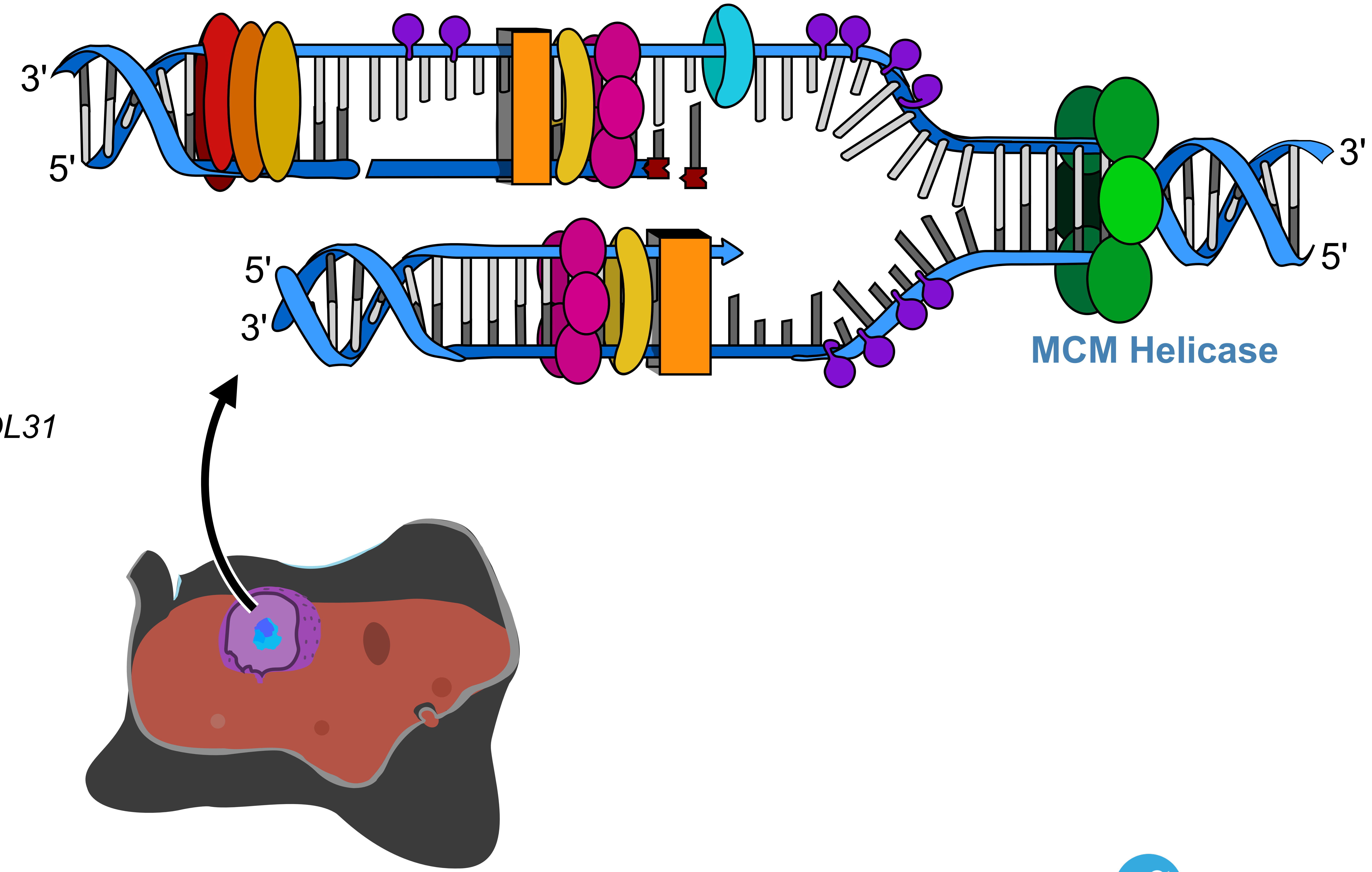
# Genes from pathways are coevolving




@JLSteenwyk

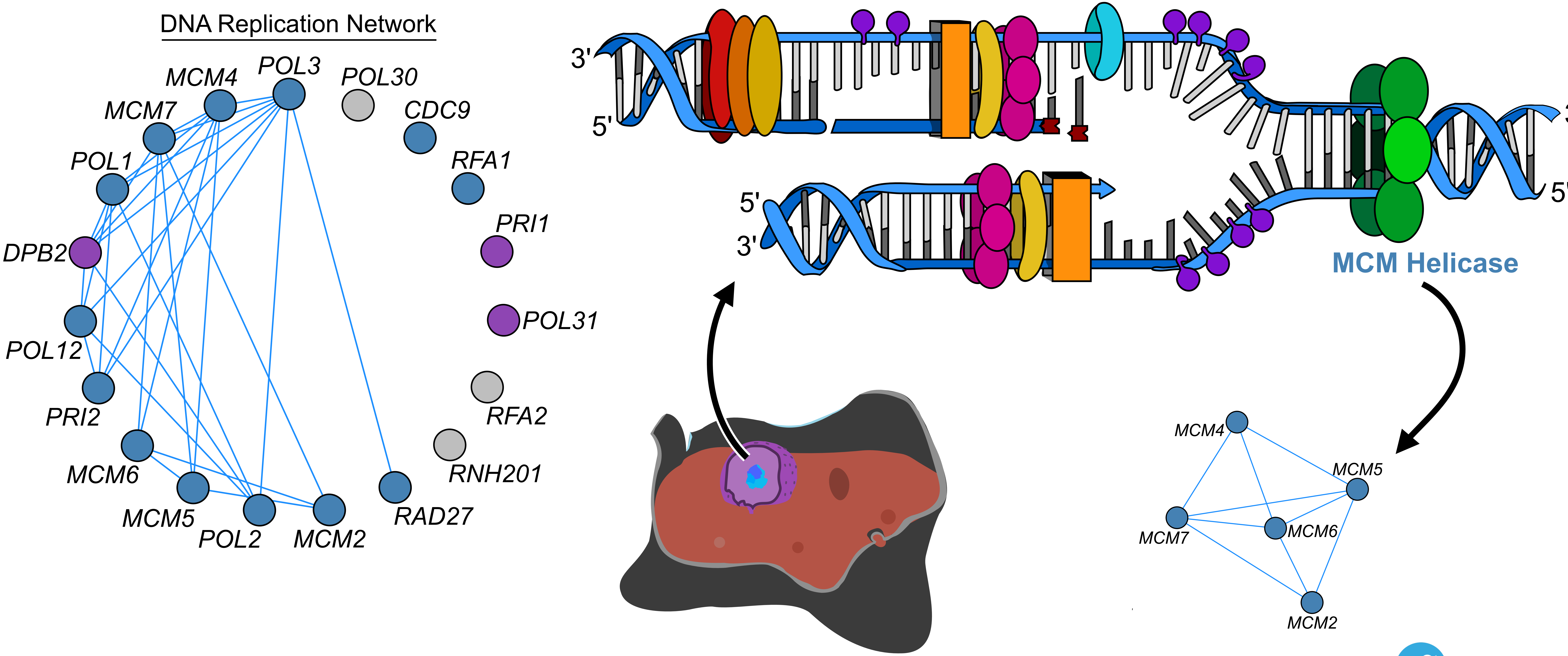
# Genes from pathways are coevolving




@JLSteenwyk

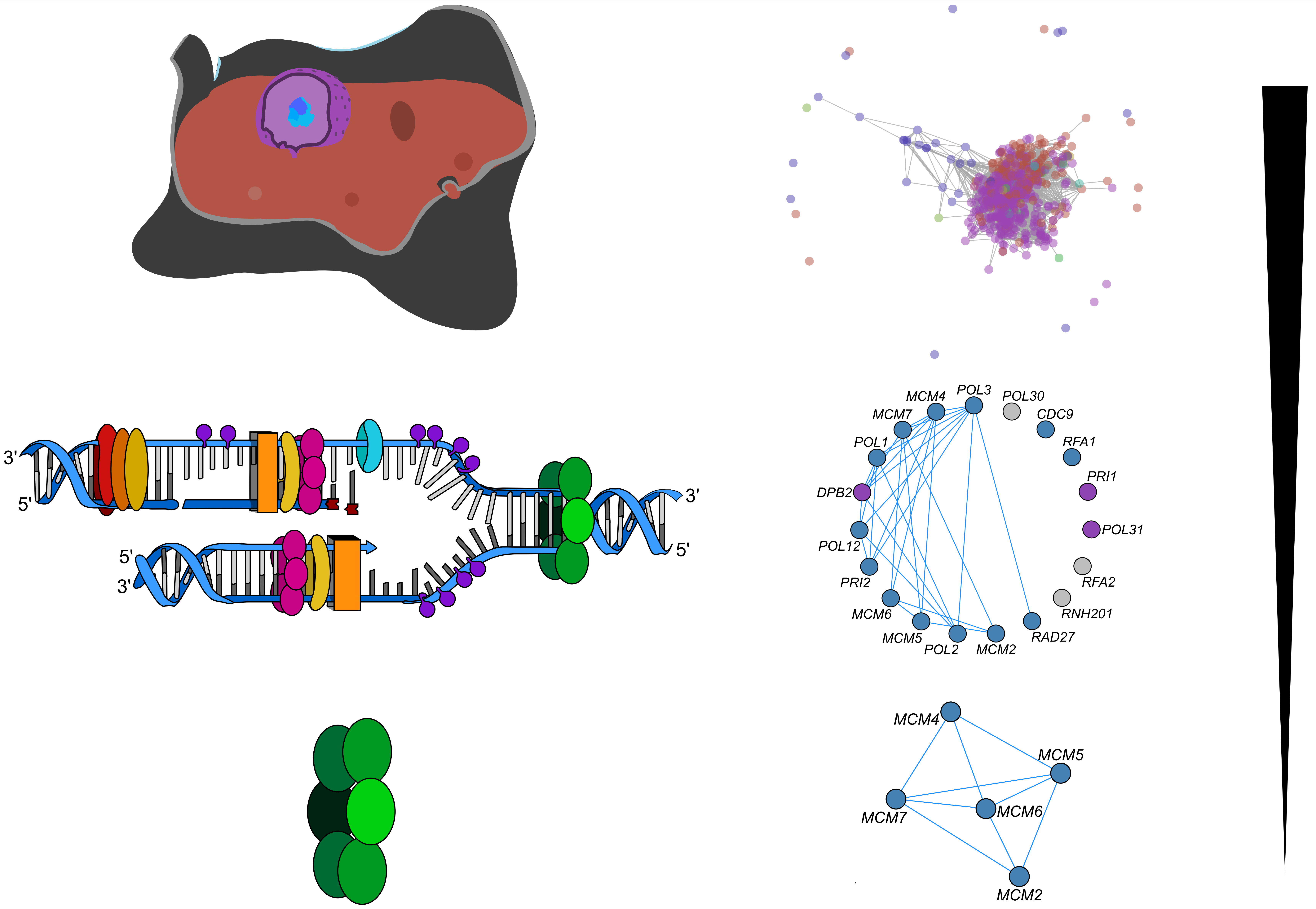
# Genes from multimeric proteins are coevolving




@JLSteenwyk

Steenwyk et al. (2022), PMID: 35507651

# A global network provides insight to a hierarchy of function




Cellular

Bioprocess

Protein complex

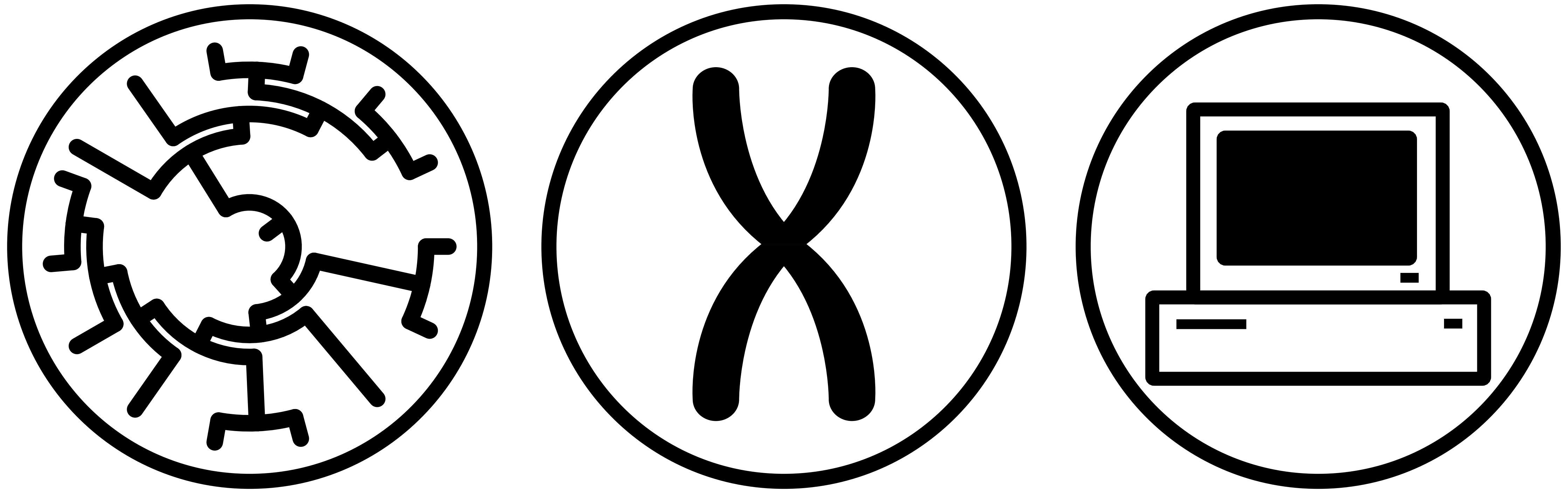
@JLSteenwyk

***Can signatures of gene coevolution provide insight to your genes of interest?***

***Can signatures of gene coevolution provide insight to your genes of interest?***

***YES!***

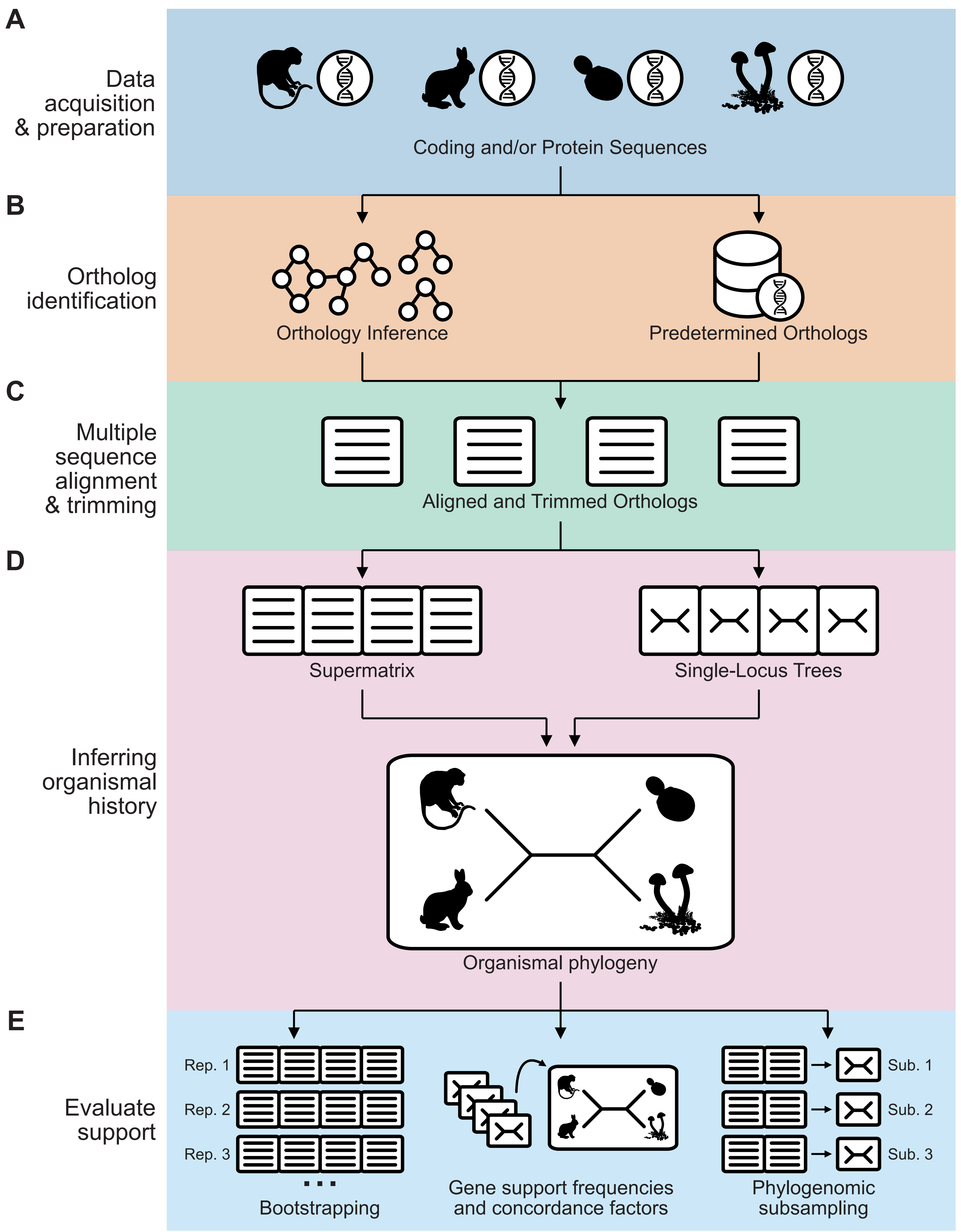
# Outline



- Introduction
- Inferring genetic networks from phylogenies
- **Phylogenomic subsampling**
- Misc. notes before the tutorial

@JLSteenwyk

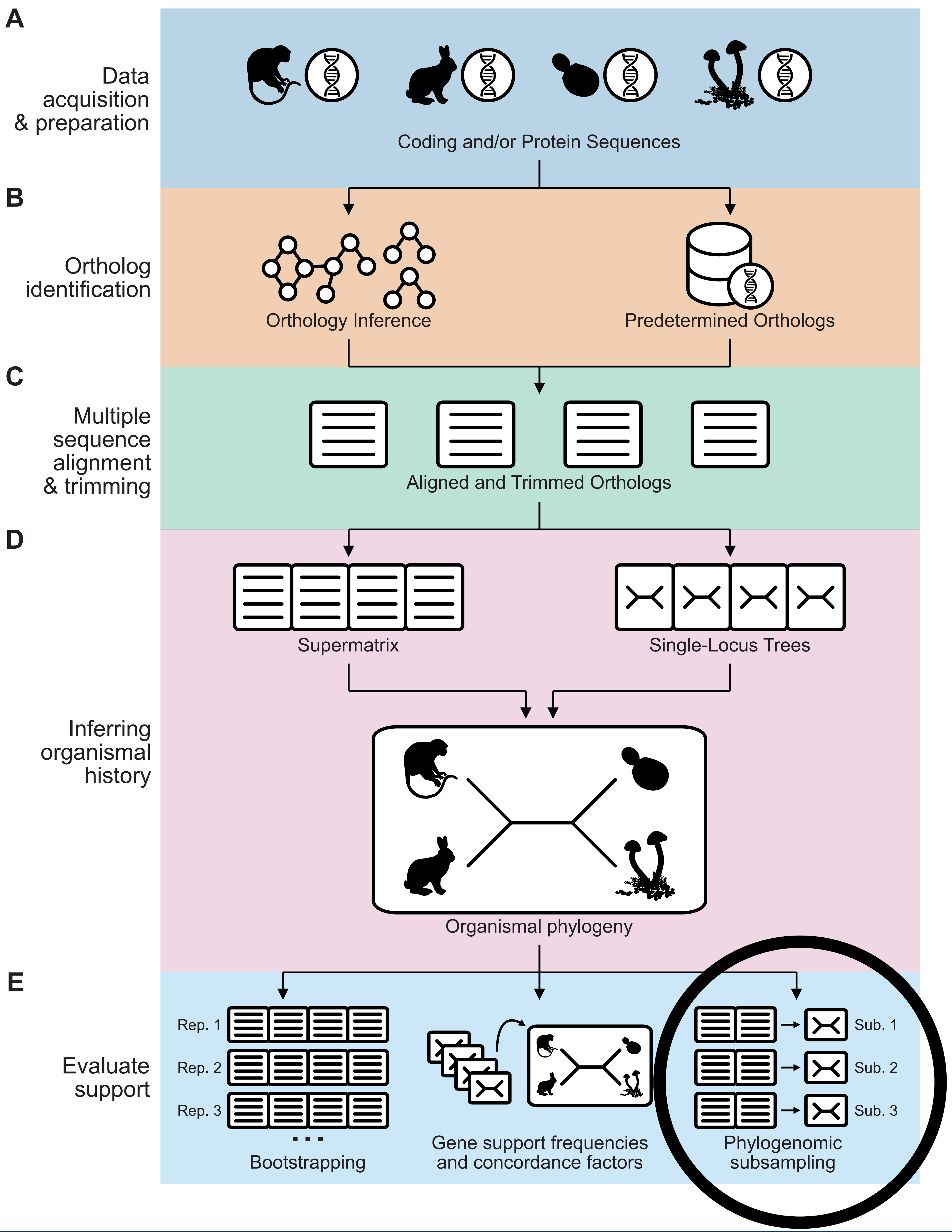
# Facilitating phylogenomic workflows and beyond




@JLSteenwyk

Bjornson et al. (2024), DOI: 10.20944/preprints202309.0905.v2

# Facilitating phylogenomic workflows and beyond




@JLSteenwyk

Bjornson et al. (2024), DOI: 10.20944/preprints202309.0905.v2

# Phylogenomics doesn’t solve everything

Review > Trends Genet. 2006 Apr;22(4):225-31. doi: 10.1016/j.tig.2006.02.003. Epub 2006 Feb 21.

## Phylogenomics: the beginning of incongruence?

Olivier Jeffroy [1], Henner Brinkmann, Frédéric Delsuc, Hervé Philippe

Affiliations + expand

PMID: 16490279 DOI: 10.1016/j.tig.2006.02.003

Free article

## Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough

Hervé Philippe, Henner Brinkmann, Dennis V. Lavrov, D. Timothy J. Littlewood, Michael Manuel, Gert Wörheide, Denis Baurain

@JLSteenwyk

# Incongruence is to be celebrated!

nature reviews genetics





Review Article | Published: 27 June 2023

## Incongruence in the phylogenomics era

Jacob L. Steenwyk, Yuanning Li, Xiaofan Zhou, Xing-Xing Shen & Antonis Rokas

*Nature Reviews Genetics* **24**, 834–850 (2023) | Cite this article

# Phylogenomic subsampling, in brief

# Phylogenomic subsampling, in brief

1. Unstable bipartitions will be sensitive to gene/taxon/site selection

# Phylogenomic subsampling, in brief

1. Unstable bipartitions will be sensitive to gene/taxon/site selection

2. Subsample the full data matrix and reinfer the species tree using fewer (but typically still several dozen to hundreds of genes)

# Phylogenomic subsampling, in brief

1. Unstable bipartitions will be sensitive to gene/taxon/site selection

2. Subsample the full data matrix and reinfer the species tree using fewer (but typically still several dozen to hundreds of genes)

3. Compare resulting phylogenies and determine which bipartition are unstable

# Phylogenomic subsampling, in brief

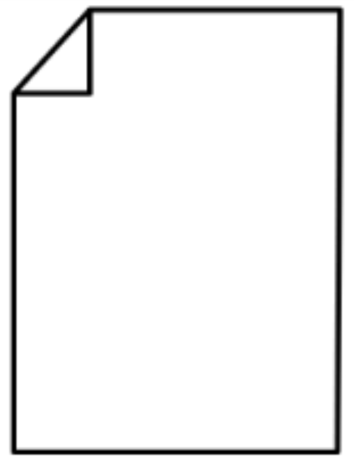
1. Unstable bipartitions will be sensitive to gene/taxon/site selection

2. Subsample the full data matrix and reinfer the species tree using fewer (but typically still several dozen to hundreds of genes)

3. Compare resulting phylogenies and determine which bipartition are unstable

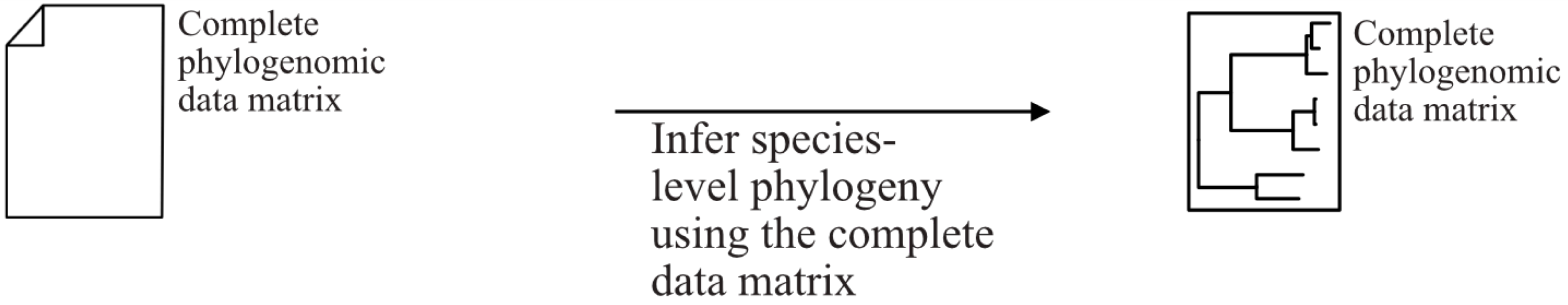
4. Examine potential drivers of incongruence thereafter. Incongruence will be examined in a later lab

# Phylogenetic subsampling



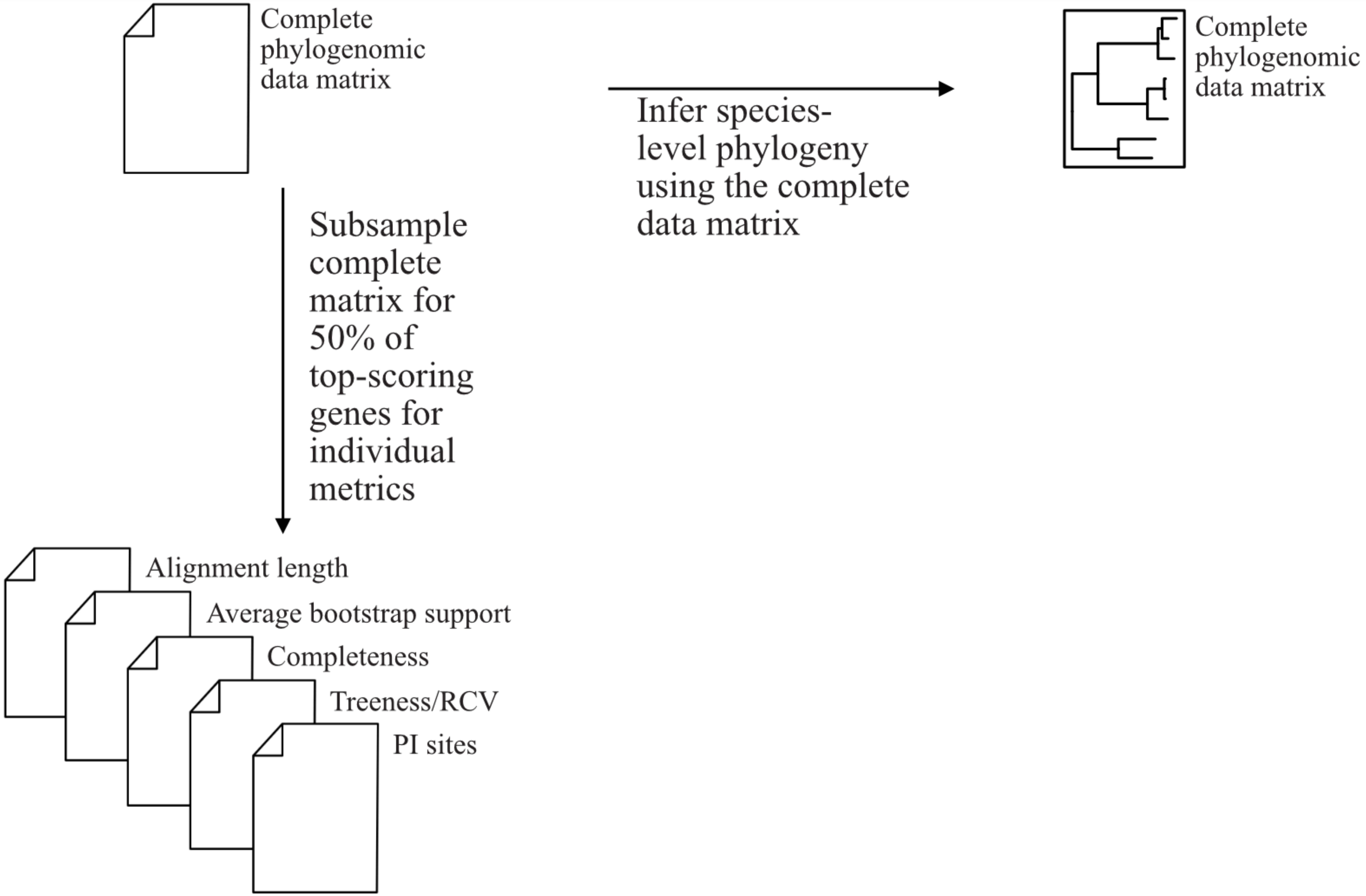
Complete
phylogenomic
data matrix

# Phylogenetic subsampling

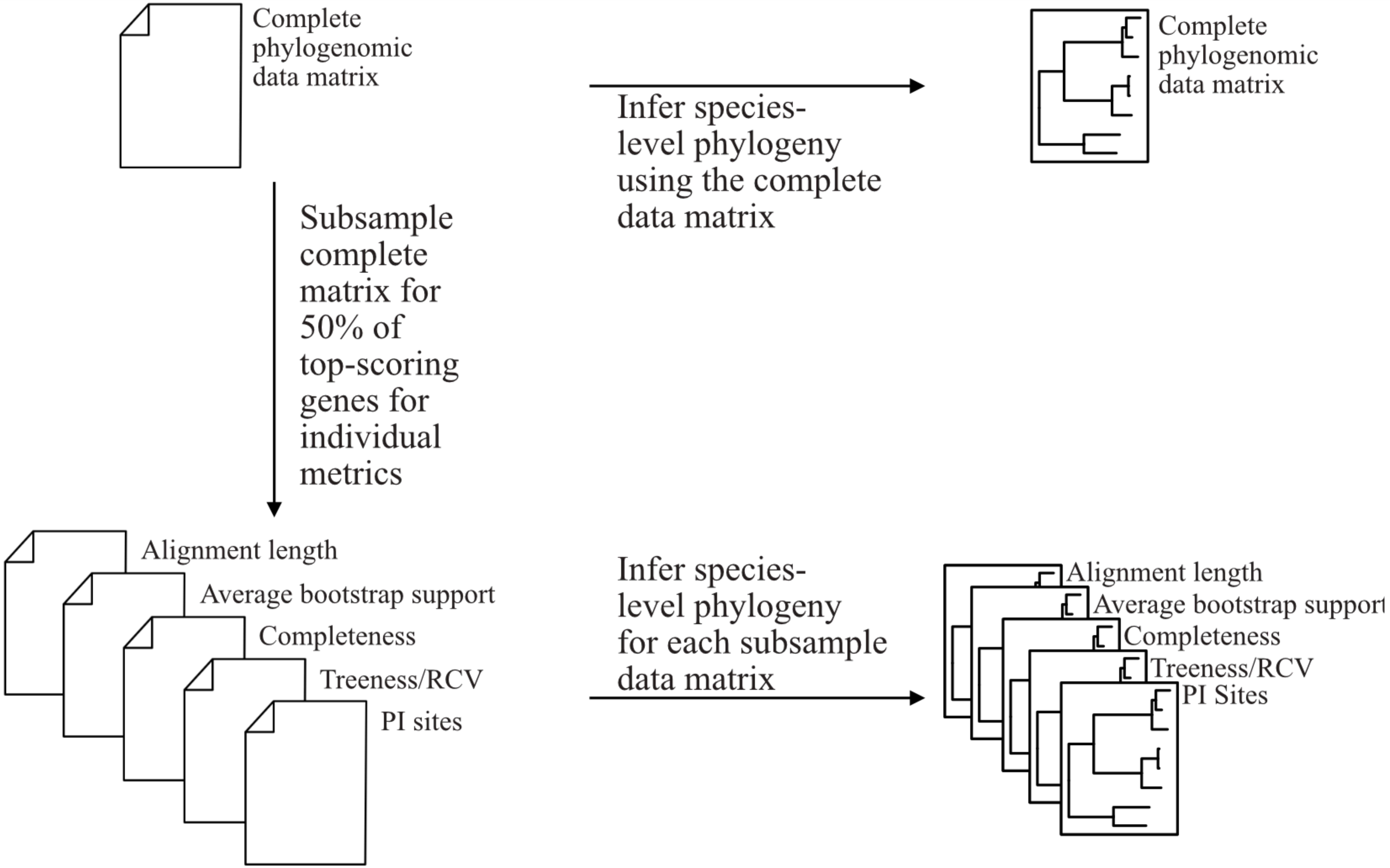



@JLSteenwyk

# Phylogenetic subsampling

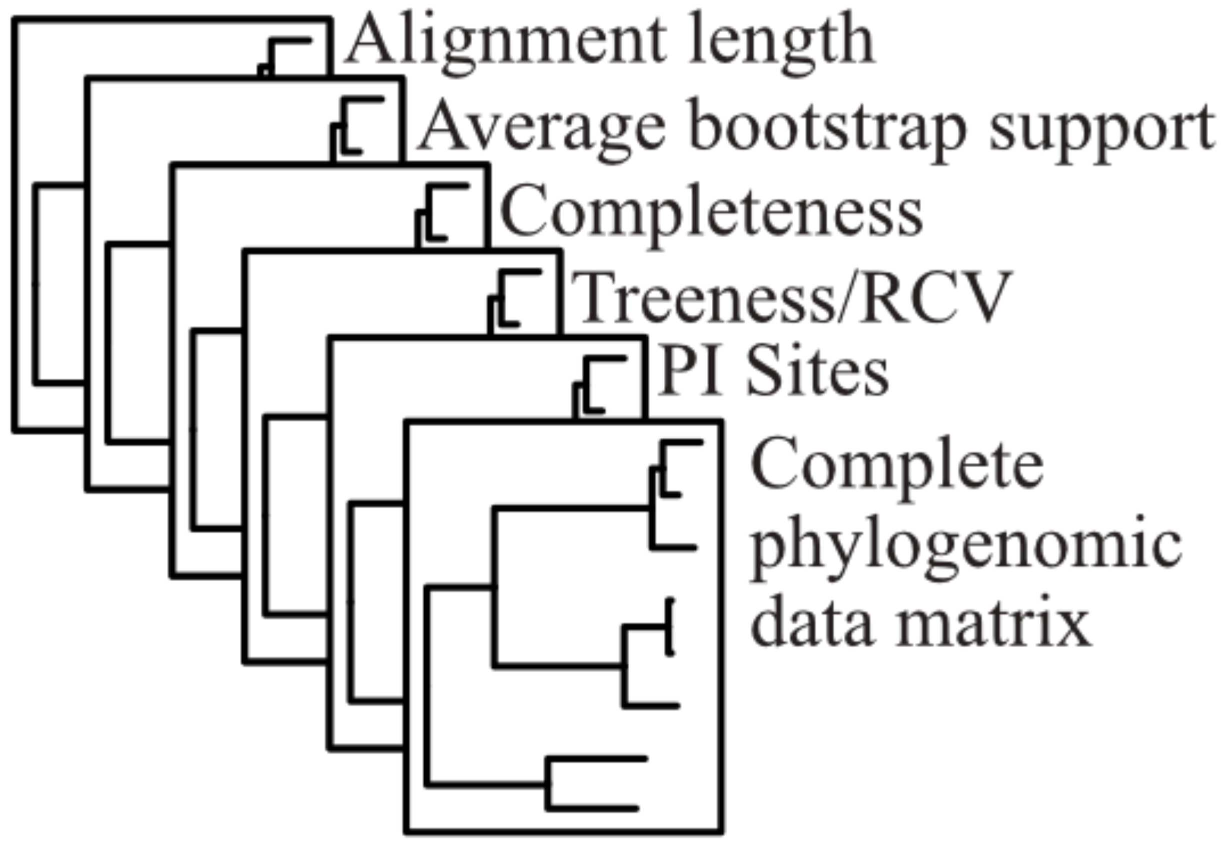


# Phylogenetic subsampling



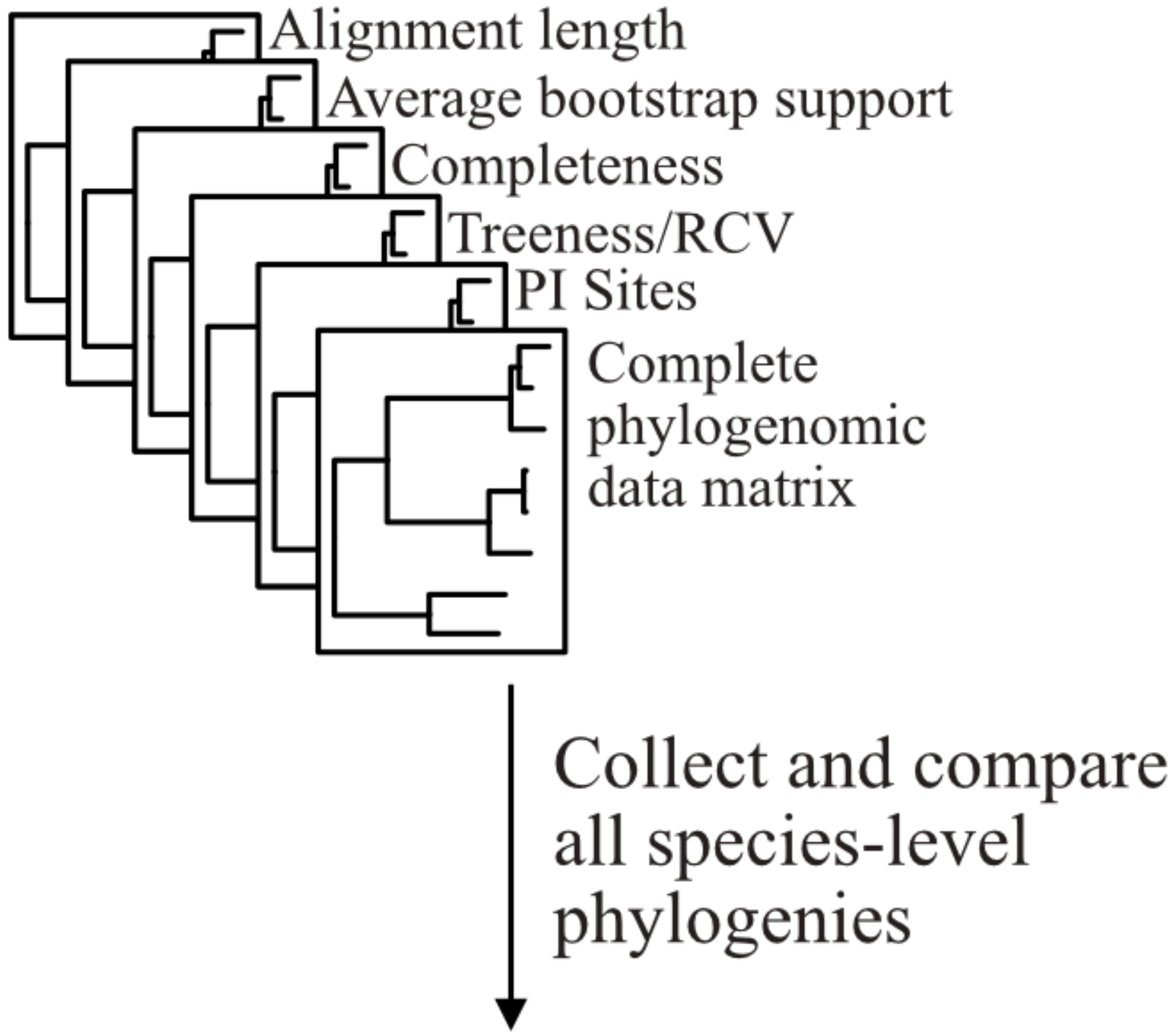

@JLSteenwyk

# Phylogenetic subsampling



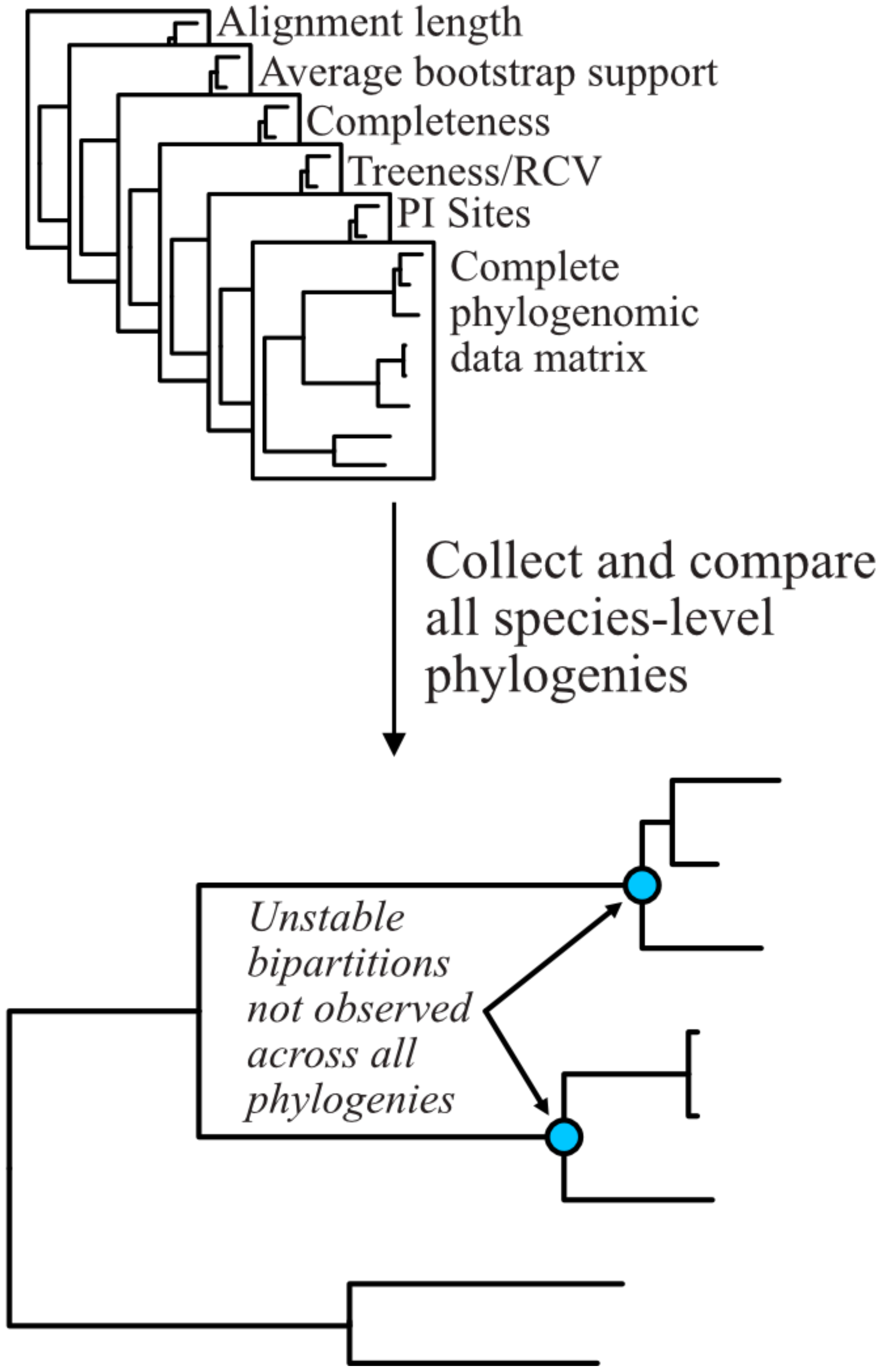

@JLSteenwyk

# Phylogenetic subsampling




@JLSteenwyk

# Phylogenetic subsampling




@JLSteenwyk

# Metrics that capture phylogenetic signal

1. Alignment length
2. Alignment length with no gaps
3. GC content (for NTs)
4. Pairwise identity
5. # of parsimony informative sites
6. # of variable sites
7. Relative composition variability
8. Average bootstrap support value
9. Degree of violation of a molecular clock
10. Evolutionary rate
11. Long branch score
12. Treeness
13. Saturation
14. Treeness / RCV
15. RCVT
16. Compositional bias per site
17. Evolutionary rate per site

# Metrics that capture phylogenetic signal

1. **Alignment length**
2. Alignment length with no gaps
3. GC content (for NTs)
4. Pairwise identity
5. # of parsimony informative sites
6. # of variable sites
7. **Relative composition variability**
8. Average bootstrap support value
9. Degree of violation of a molecular clock
10. Evolutionary rate
11. **Long branch score**
12. Treeness
13. **Saturation**
14. **Treeness / RCV**
15. **RCVT**
16. **Compositional bias per site**
17. **Evolutionary rate per site**

@JLSteenwyk

# Phylogenetic signal across genes

1. **Alignment length**
2. Alignment length with no gaps
3. GC content (for NTs)
4. Pairwise identity
5. # of parsimony informative sites
6. # of variable sites
7. **Relative composition variability**
8. Average bootstrap support value
9. Degree of violation of a molecular clock
10. Evolutionary rate
11. **Long branch score**
12. Treeness
13. **Saturation**
14. **Treeness / RCV**

# Alignment length

```
>sp1
ACGTAGCG-TCGATC
>sp2
ACGT-GCGATCGATC
>sp3
ACGTAGC-ATCGATC
>sp4
ACGTAGCGATCGATG
>sp5
AC--AGCGATCGATC
>sp6
ACGTAGCGA---ATC
```

The length of this alignment is 15 sites

@JLSteenwyk

# Alignment length

```
>sp1
ACGTAGCG-TCGATC
>sp2
ACGT-GCGATCGATC
>sp3
ACGTAGC-ATCGATC
>sp4
ACGTAGCGATCGATG
>sp5
AC--AGCGATCGATC
>sp6
ACGTAGCGA---ATC
```

***Higher values are better!***

The length of this alignment is 15 sites

# Relative composition variability

- Average variability in the sequence composition among taxa in an MSA

@JLSteenwyk

# Relative composition variability

- Average variability in the sequence composition among taxa in an MSA
- Evaluates potential composition biases
    - violate assumptions of site composition homogeneity in standard substitution models

# Relative composition variability

- Average variability in the sequence composition among taxa in an MSA
- Evaluates potential composition biases
  - violate assumptions of site composition homogeneity in standard substitution models

$$\sum_{i=1}^{c} \sum_{j=1}^{n} \frac{\left| c_{ij} - \overline{c_i} \right|}{s \times n}$$

# Relative composition variability

- Average variability in the sequence composition among taxa in an MSA
- Evaluates potential composition biases
  - violate assumptions of site composition homogeneity in standard substitution models

$$\sum_{i=1}^{c} \sum_{j=1}^{n} \frac{\left| c_{ij} - \overline{c_i} \right|}{s \times n}$$

- *c* is the number of different character states per sequence type
- *n* is the number of taxa in an MSA
- *s* is the number of sites in an MSA

# Relative composition variability

```
>Seq 1
MKGATTLAK

>Seq 2
MK-AITLAK

>Seq 3
MKGATT--K

>Seq 4
MK-AITLA-
```

RCV = 0.375

# Relative composition variability

```
>Seq 1
MKGATTLAK

>Seq 2
MK-AITLAK

>Seq 3
MKGATT--K

>Seq 4
MK-AITLA-
```

RCV = 0.375

→

```
>Seq_1
MKTTTTTTT

>Seq_2
MKQQQQQQQ

>Seq_3
MKKKKKKKK

>Seq_4
MKLLLLLLL
```

RCV = 1.1667

# Relative composition variability

Lower compositional bias

```
>Seq 1
MKGATTLAK

>Seq 2
MK-AITLAK

>Seq 3
MKGATT--K

>Seq 4
MK-AITLA-
```

RCV = 0.375

Higher compositional bias

```
>Seq_1
MKTTTTTTT

>Seq_2
MKQQQQQQQ

>Seq_3
MKKKKKKKK

>Seq_4
MKLLLLLLL
```

RCV = 1.1667

@JLSteenwyk

# Relative composition variability

***Lower RCV values are better***

Lower compositional bias

```
>Seq 1
MKGATTLAK

>Seq 2
MK-AITLAK

>Seq 3
MKGATT--K

>Seq 4
MK-AITLA-
```

RCV = 0.375

Higher compositional bias

```
>Seq_1
MKTTTTTTT

>Seq_2
MKQQQQQQQ

>Seq_3
MKKKKKKKK

>Seq_4
MKLLLLLLL
```

RCV = 1.1667

@JLSteenwyk

# Relative composition variability

***Lower RCV values are better***

Lower compositional bias

```
>Seq 1
MKGATTLAK

>Seq 2
MK-AITLAK

>Seq 3
MKGATT--K

>Seq 4
MK-AITLA-
```

RCV = 0.375

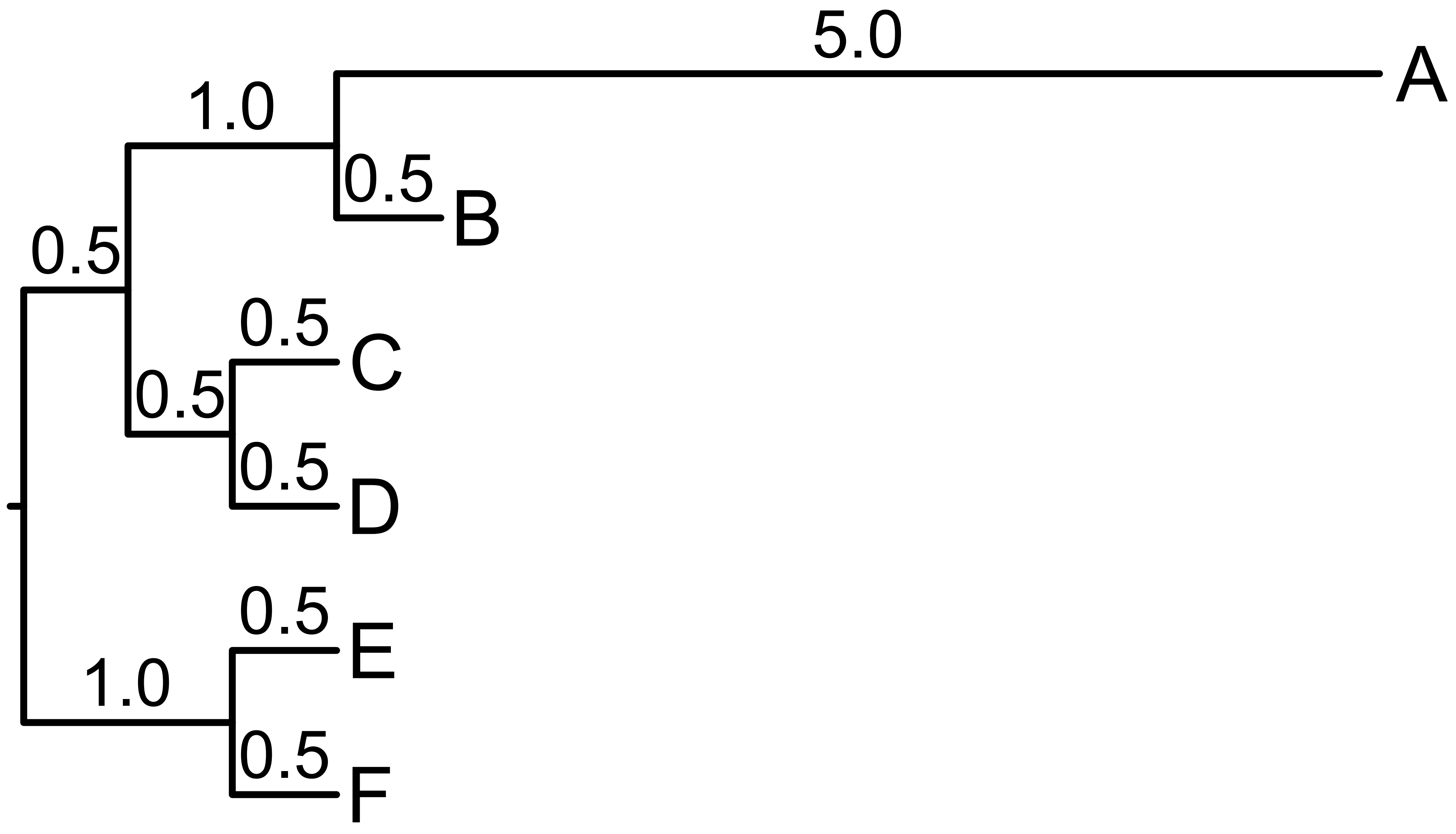
Higher compositional bias

```
>Seq_1
MKTTTTTTT

>Seq_2
MKQQQQQQQ

>Seq_3
MKKKKKKKK

>Seq_4
MKLLLLLLL
```

RCV = 1.1667

@JLSteenwyk

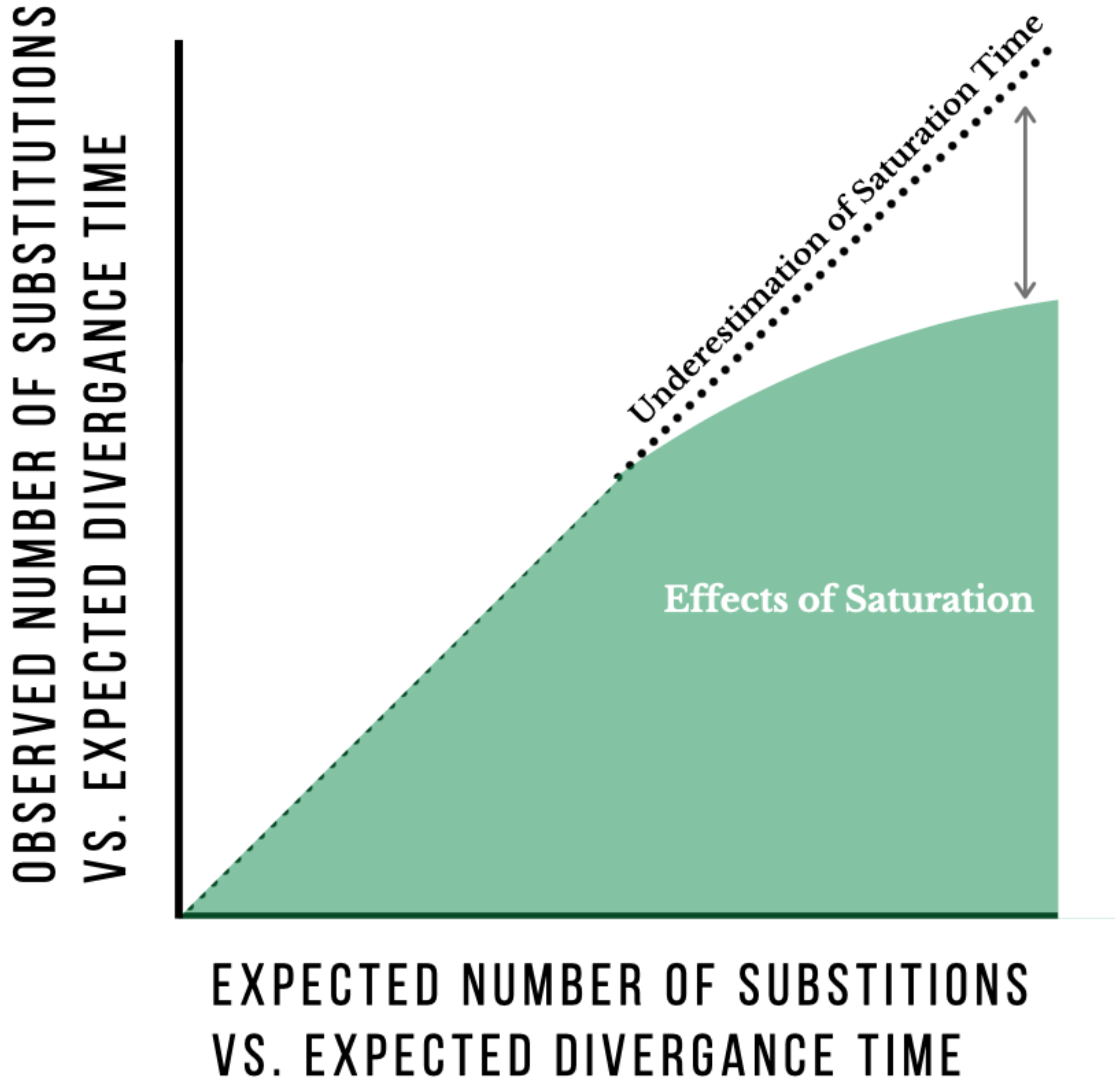
# Long branch score




Long branch scores

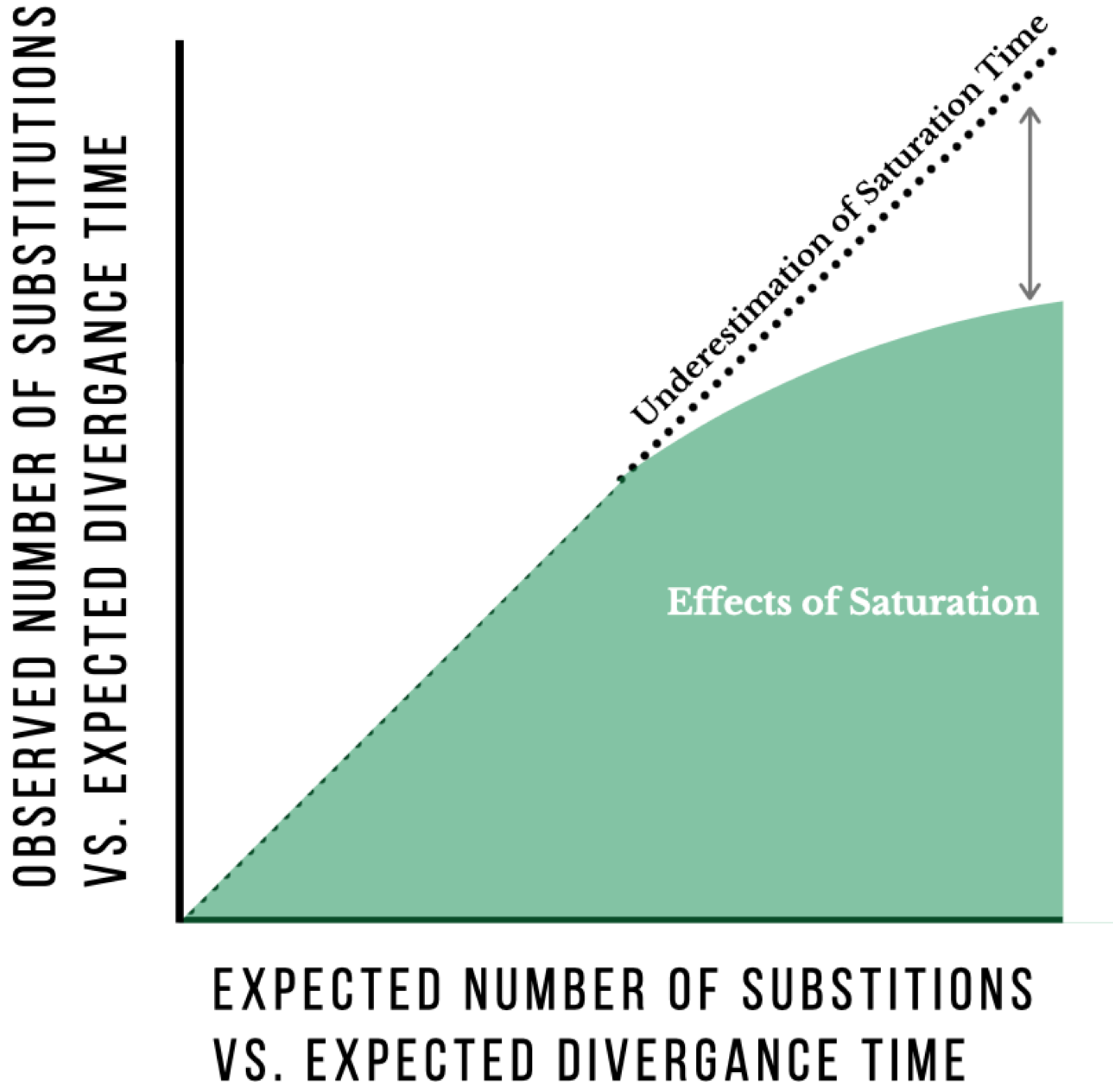
A: 73.17
B: -14.63
C: -19.51
D: -19.51
E: -9.76
F: -9.76

# Saturation by multiple substitutions



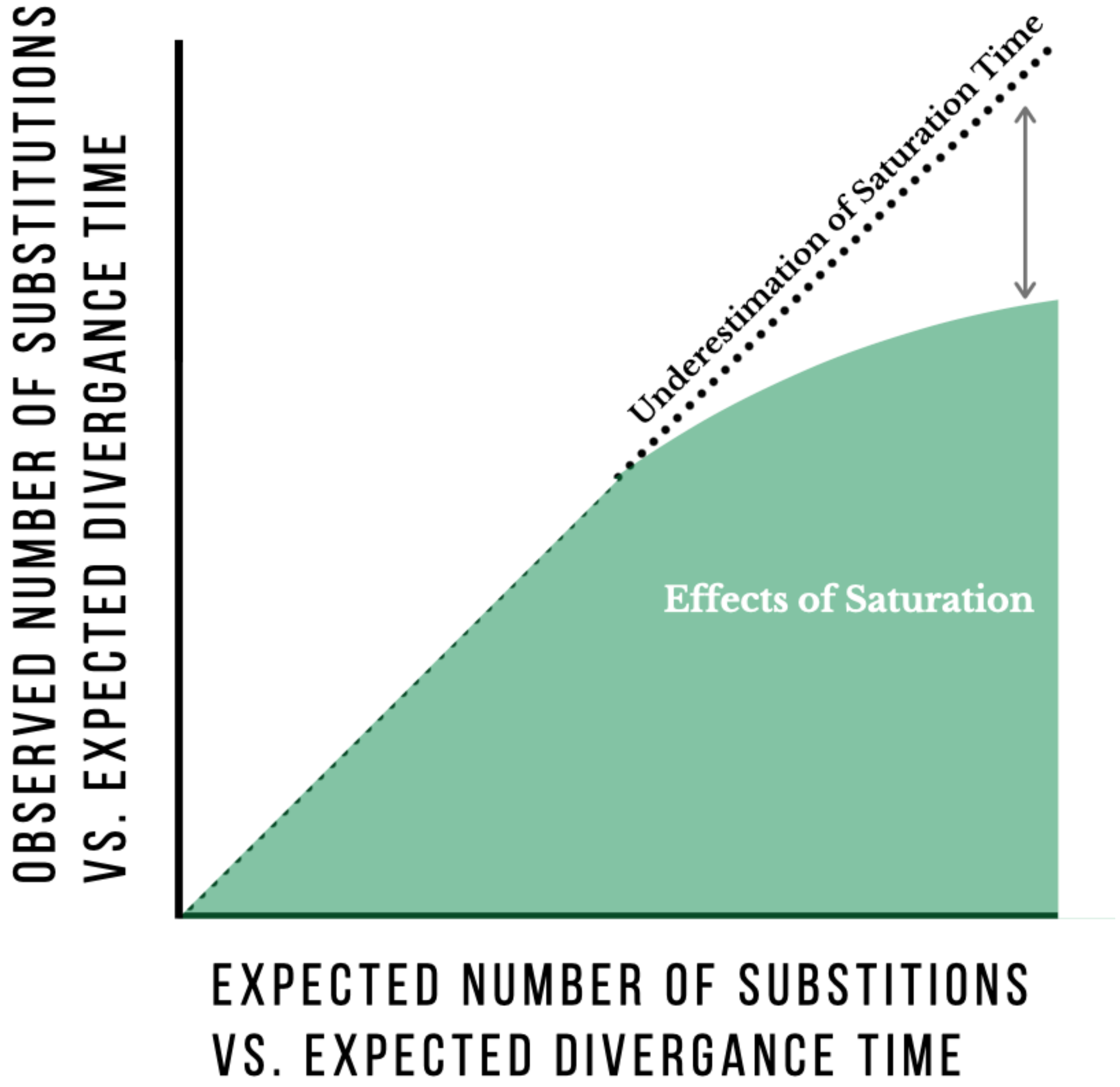

@JLSteenwyk

# Saturation by multiple substitutions



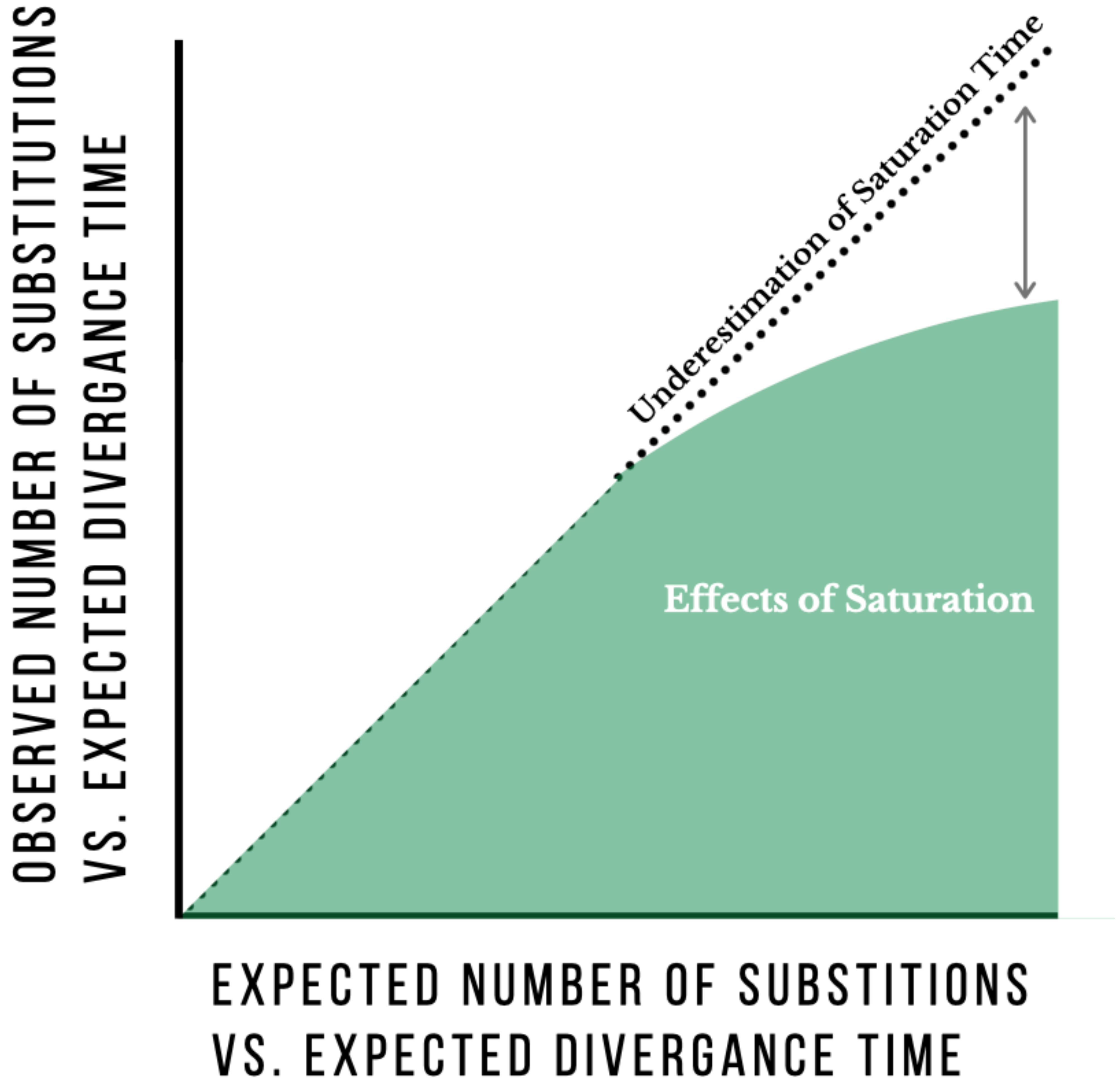

- X-axis can be approximated using phylogenetic distances
    - Tip-to-tip distances in a tree

# Saturation by multiple substitutions




- X-axis can be approximated using phylogenetic distances
  - Tip-to-tip distances in a tree
- Y-axis can be approximated using pairwise identity
  - Distance in an MSA

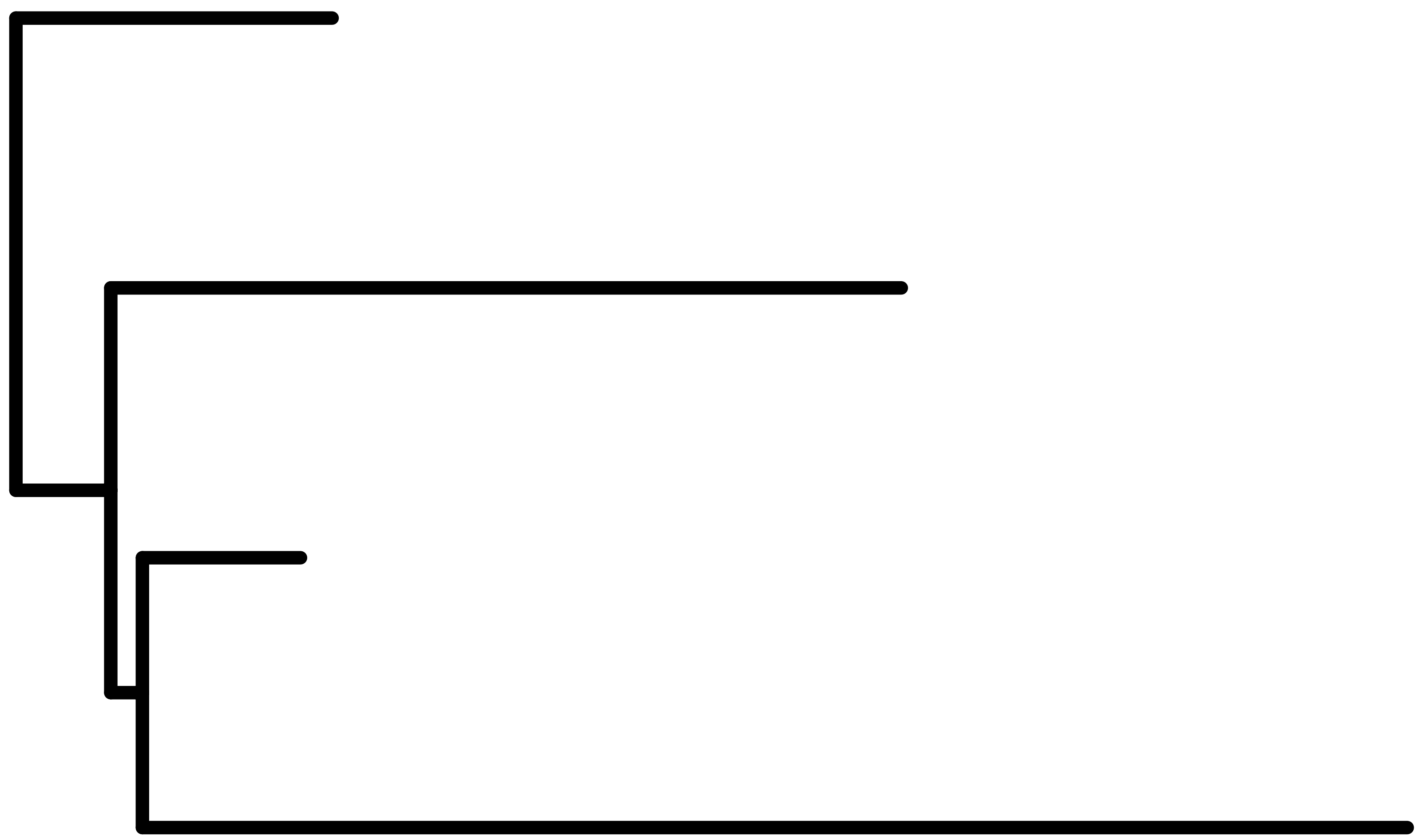
# Saturation by multiple substitutions




- The closer the slope is to 1, the better.
- PhyKIT reports the slope
- PhyKIT also reports the absolute difference between the slope and 1
    - Thus, the lower the value the better

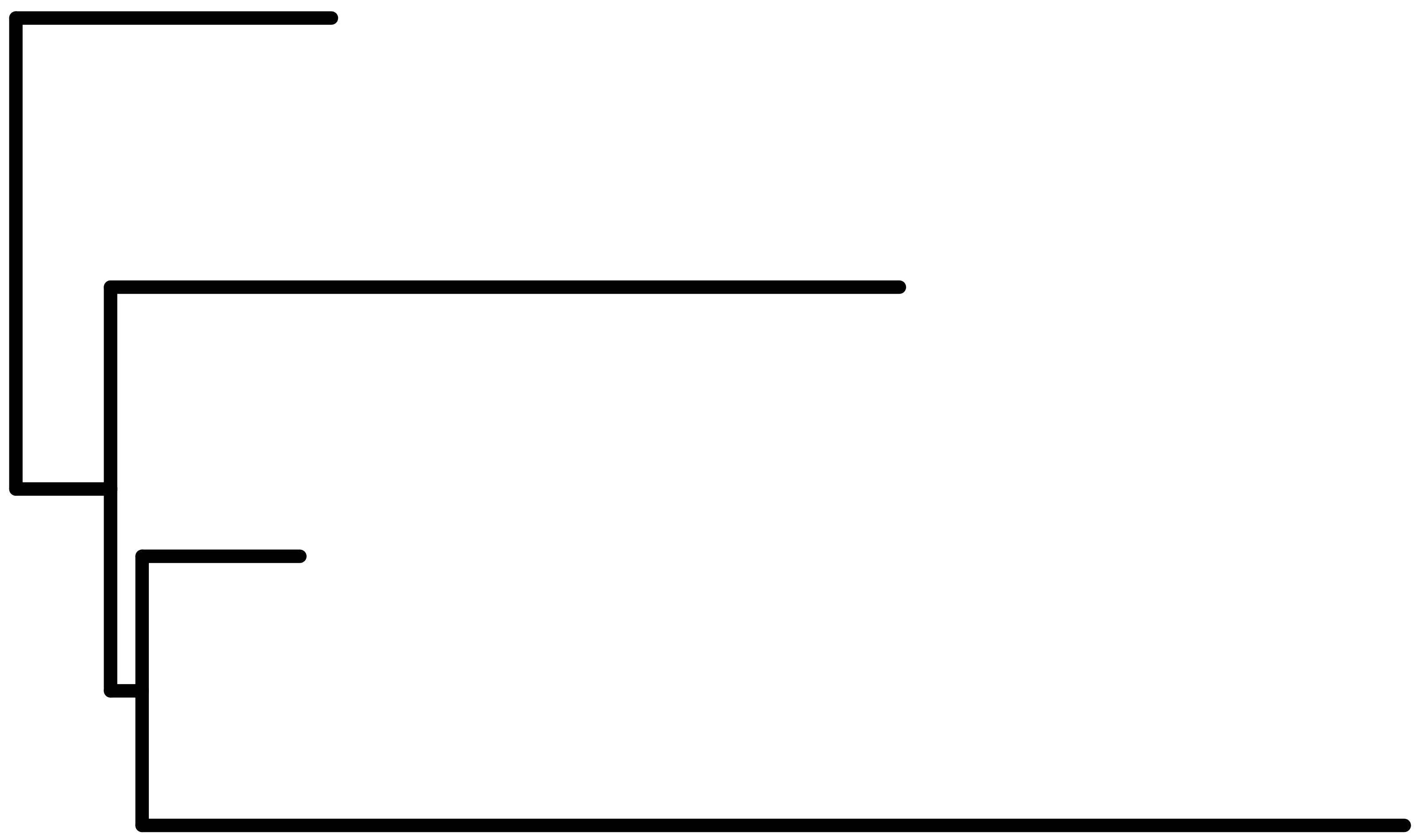
@JLSteenwyk

# Treeness

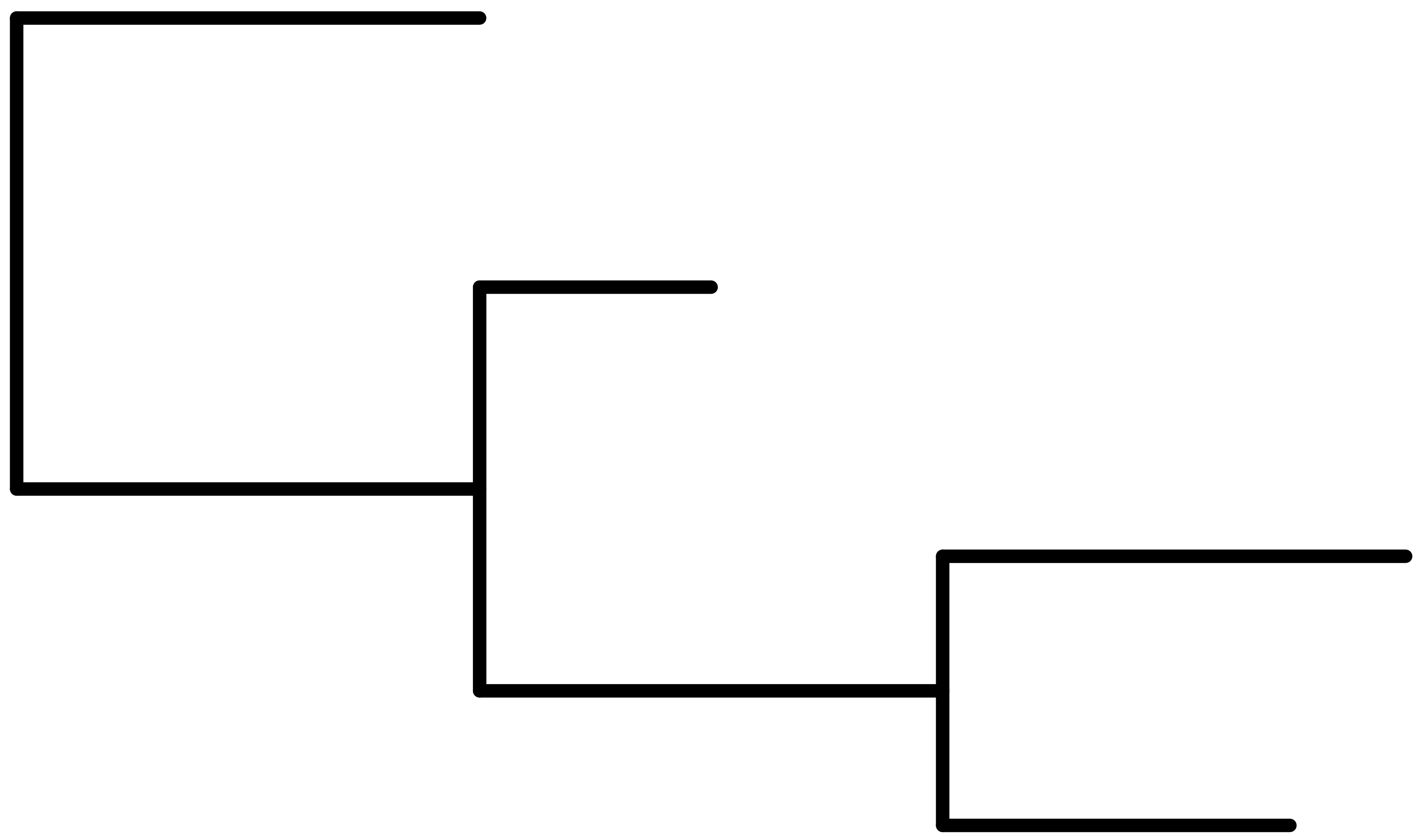


treeness = 0.0476

# Treeness



treeness = 0.0476

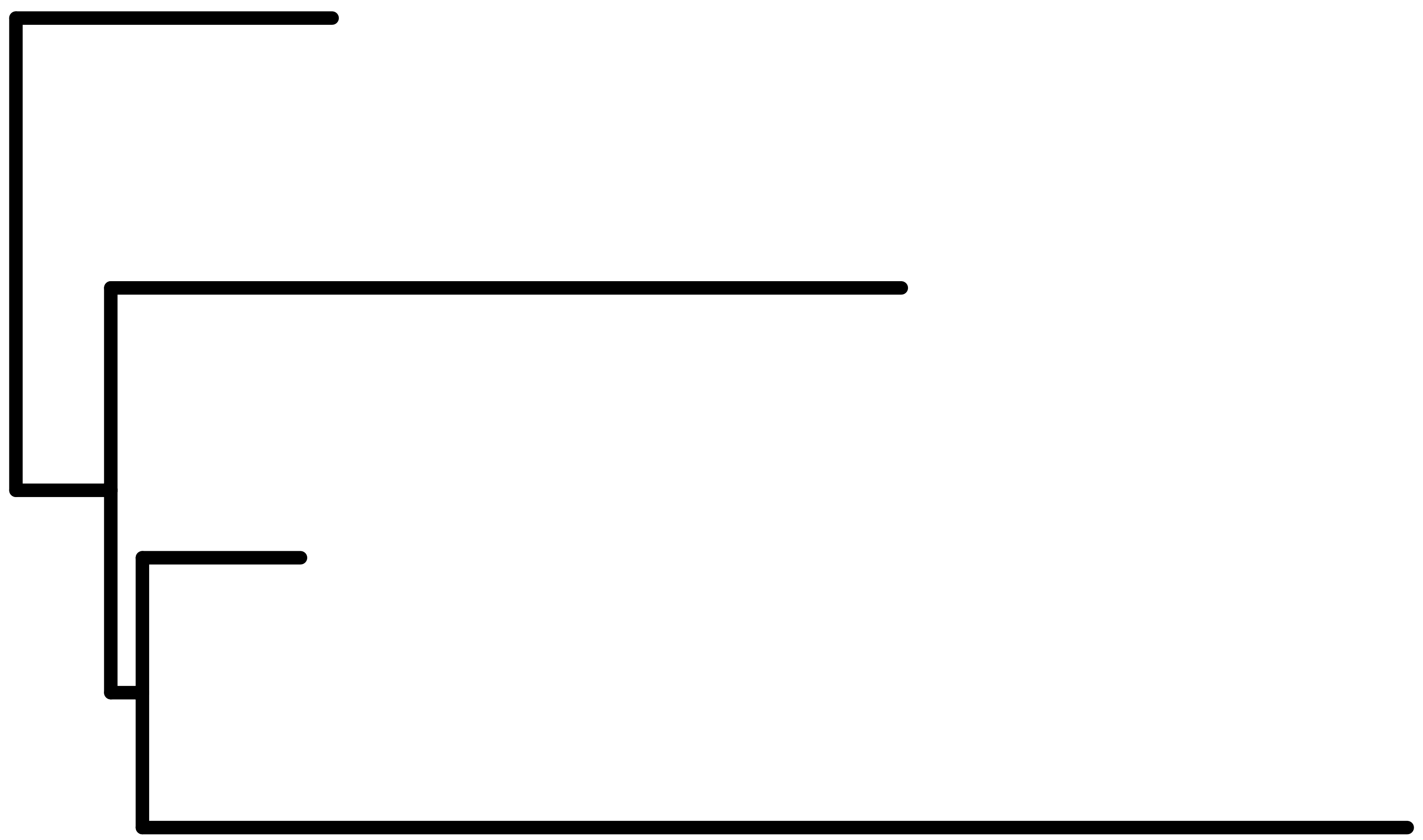


treeness = 0.381

@JLSteenwyk

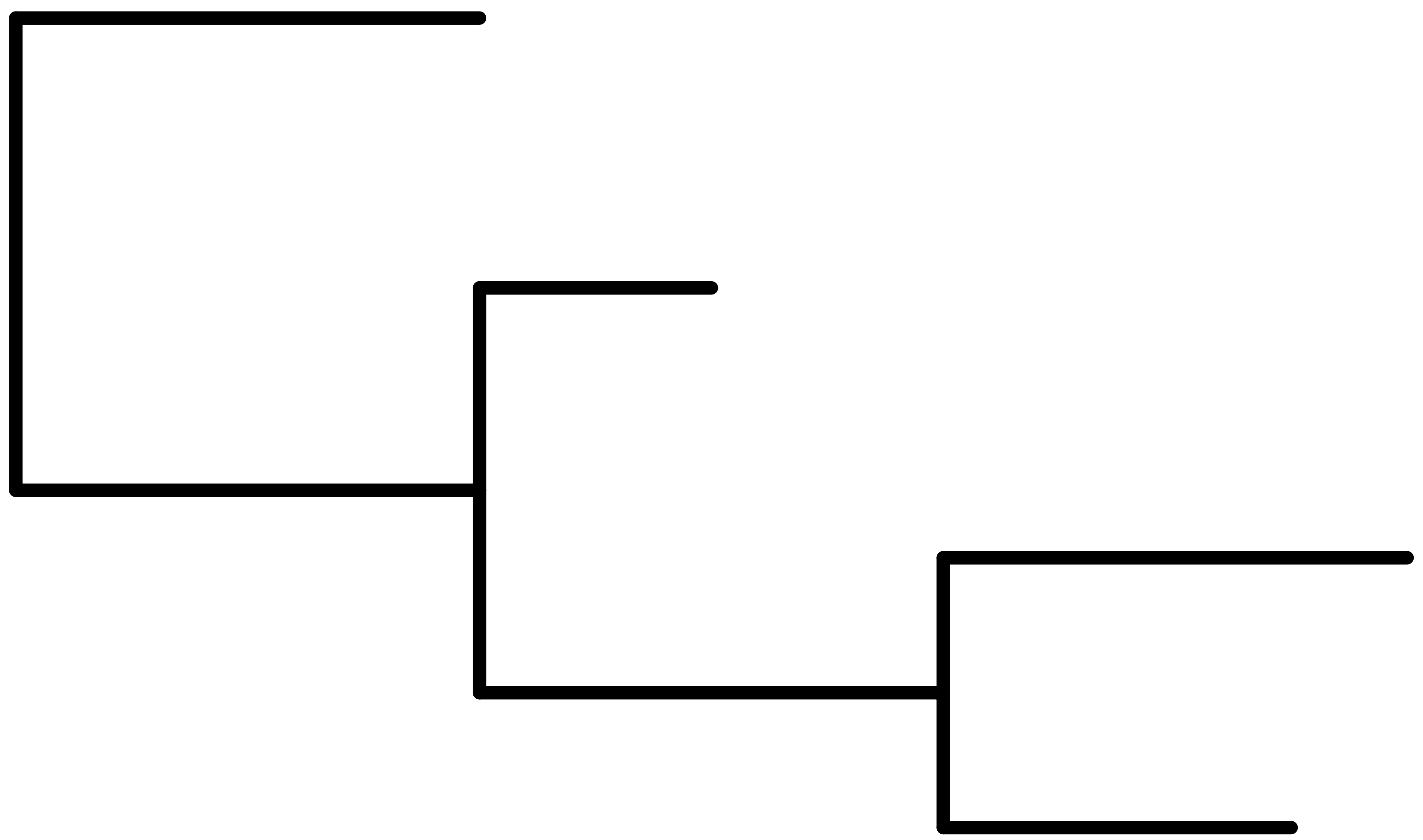
# Treeness

Low treeness



treeness = 0.0476

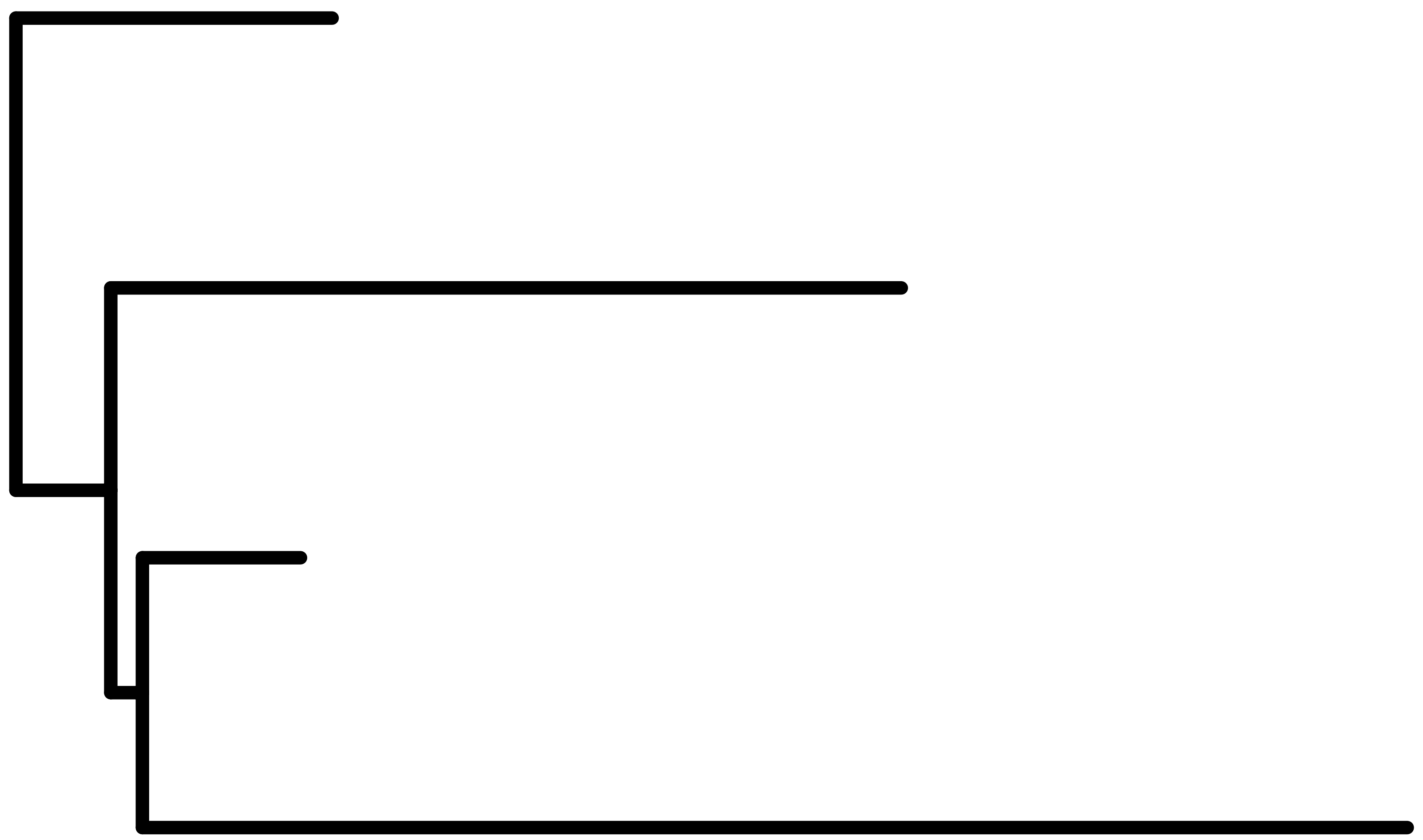
High treeness



treeness = 0.381

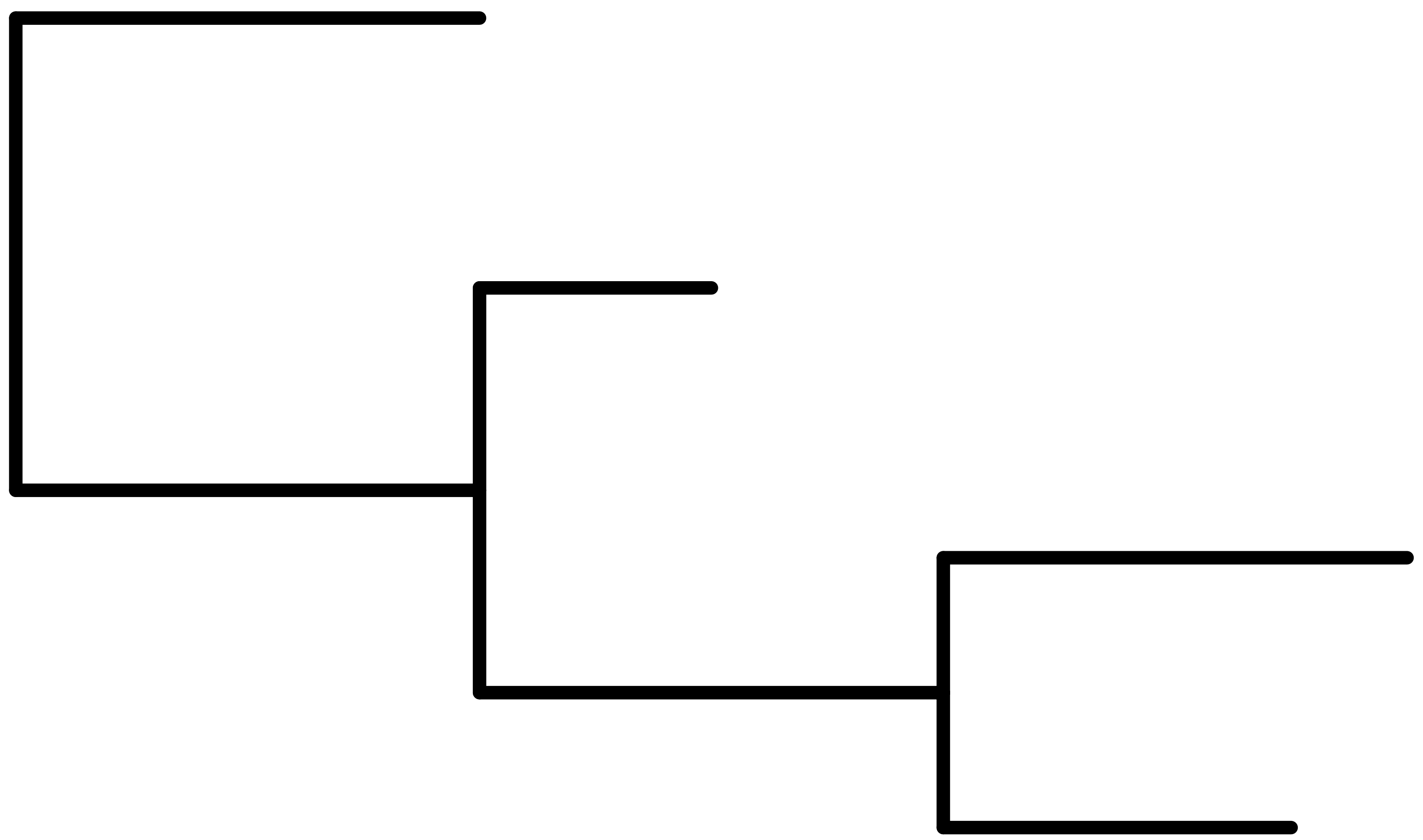
# Treeness

Low treeness



treeness = 0.0476

High treeness

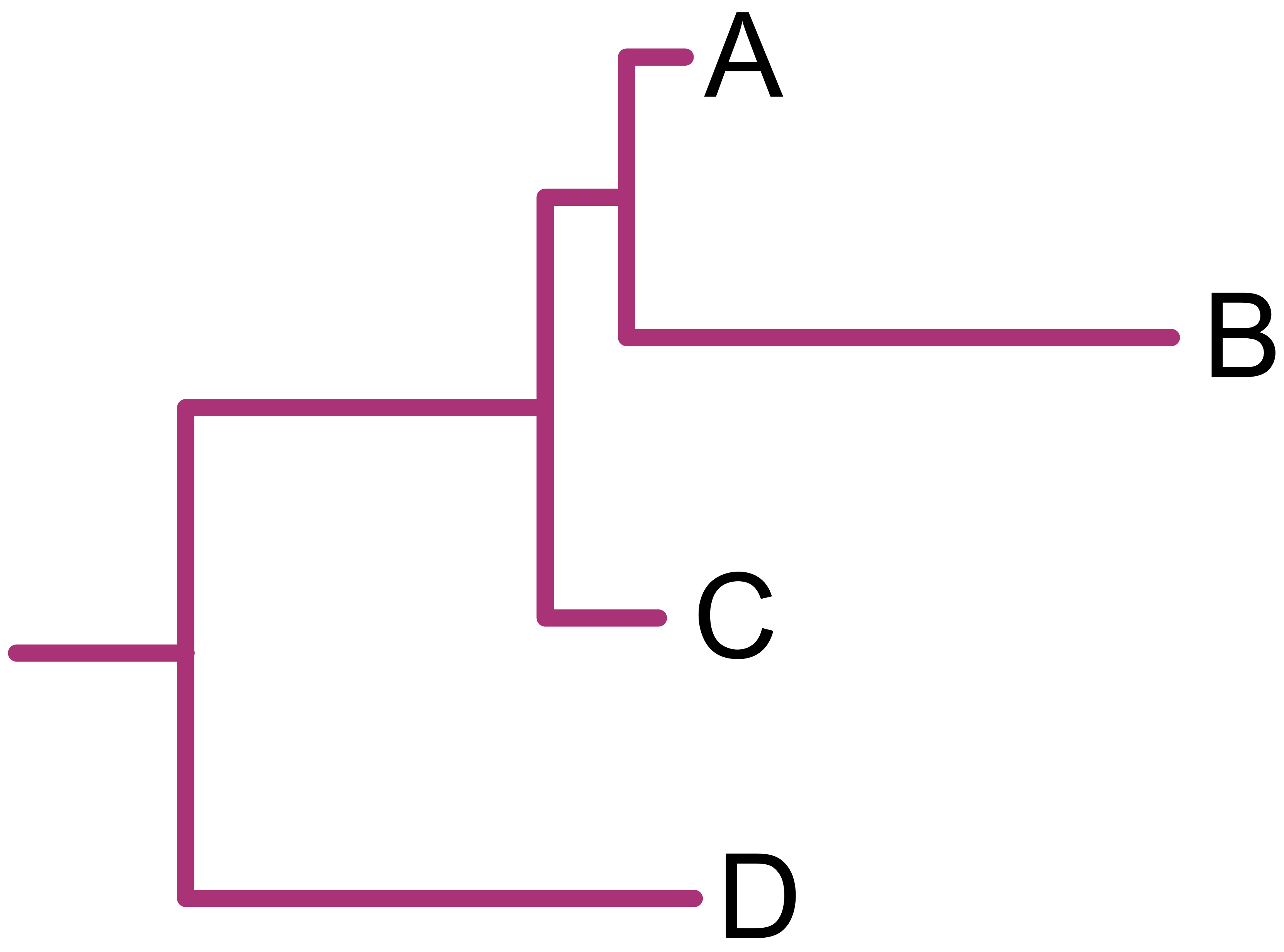


treeness = 0.381

***Higher treeness values are better***

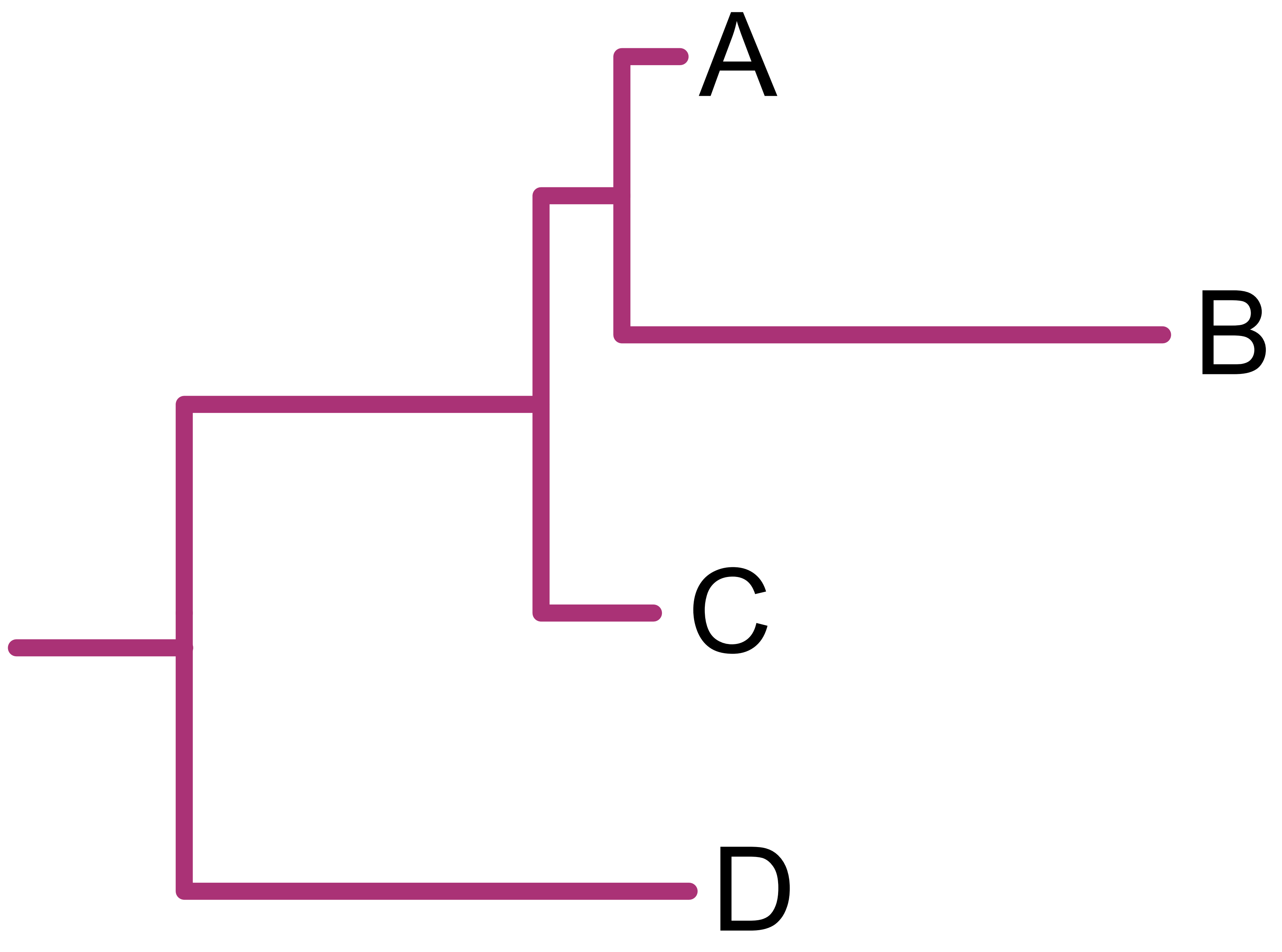
# Degree of violation of a molecular clock

Non-clock like



# Degree of violation of a molecular clock

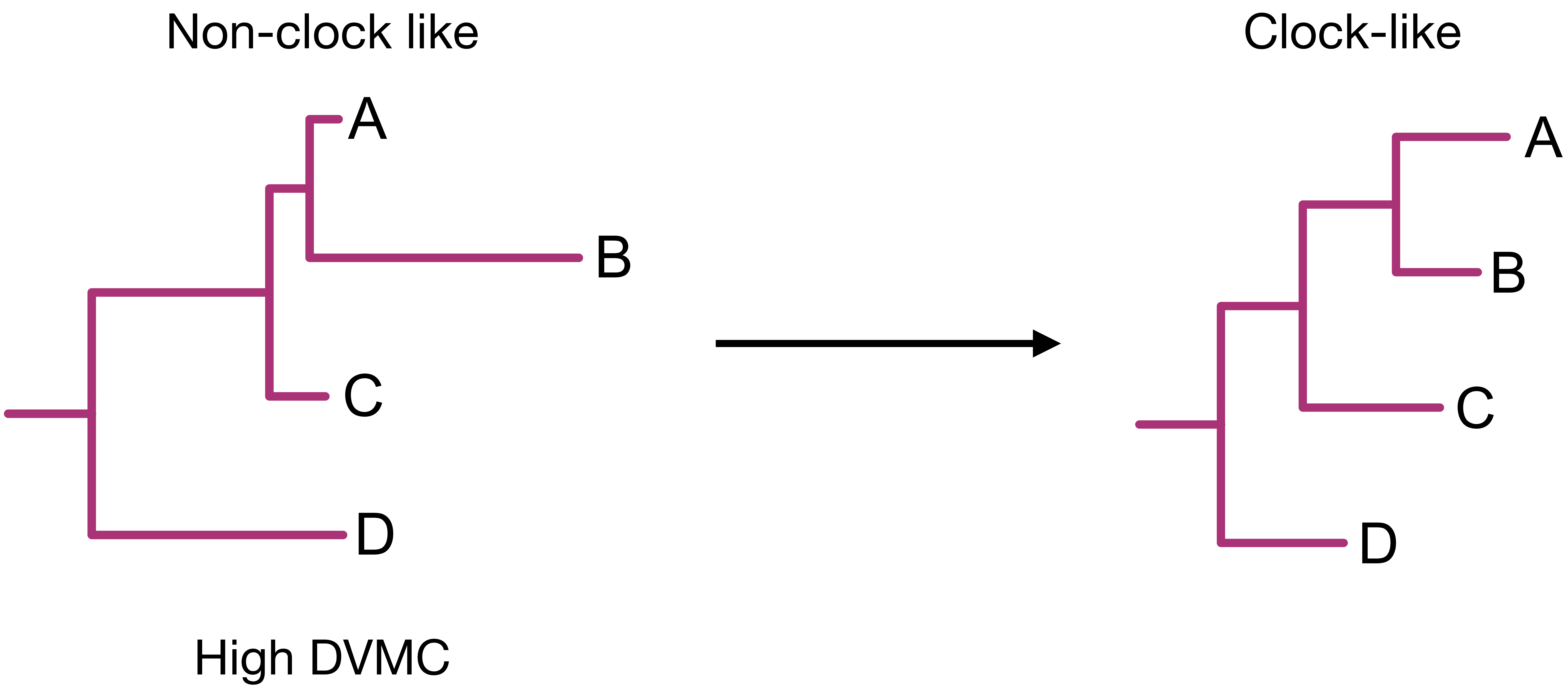
Non-clock like




High DVMC

@JLSteenwyk

# Degree of violation of a molecular clock

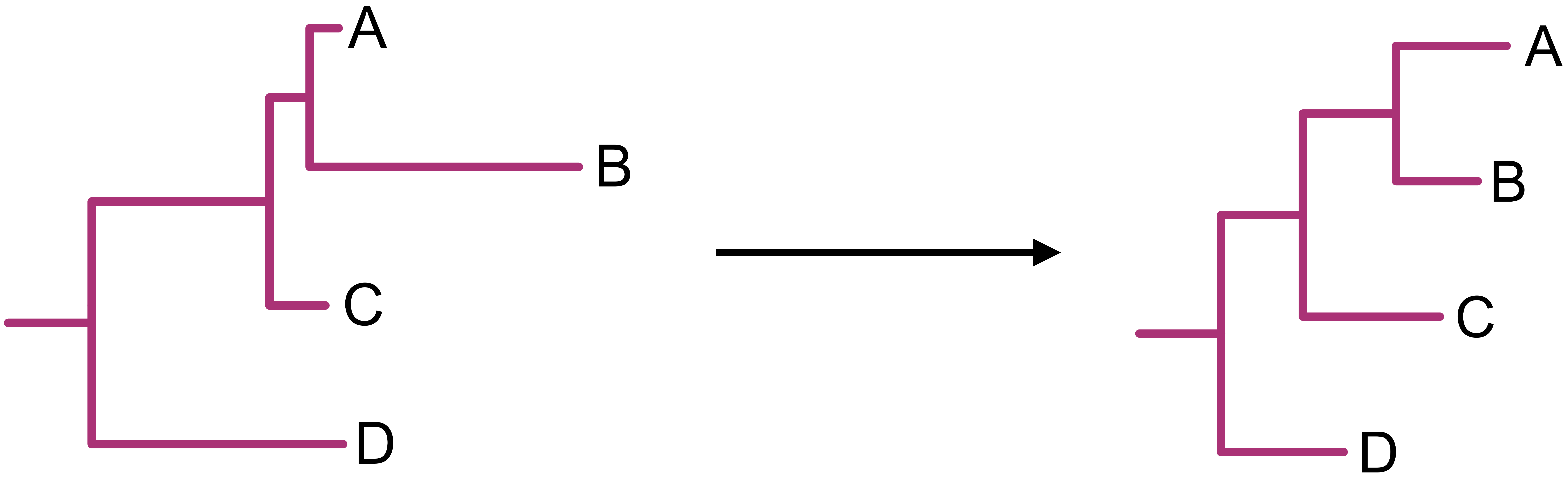



@JLSteenwyk

# Degree of violation of a molecular clock

Non-clock like

Clock-like




High DVMC

Low DVMC

@JLSteenwyk

# Degree of violation of a molecular clock

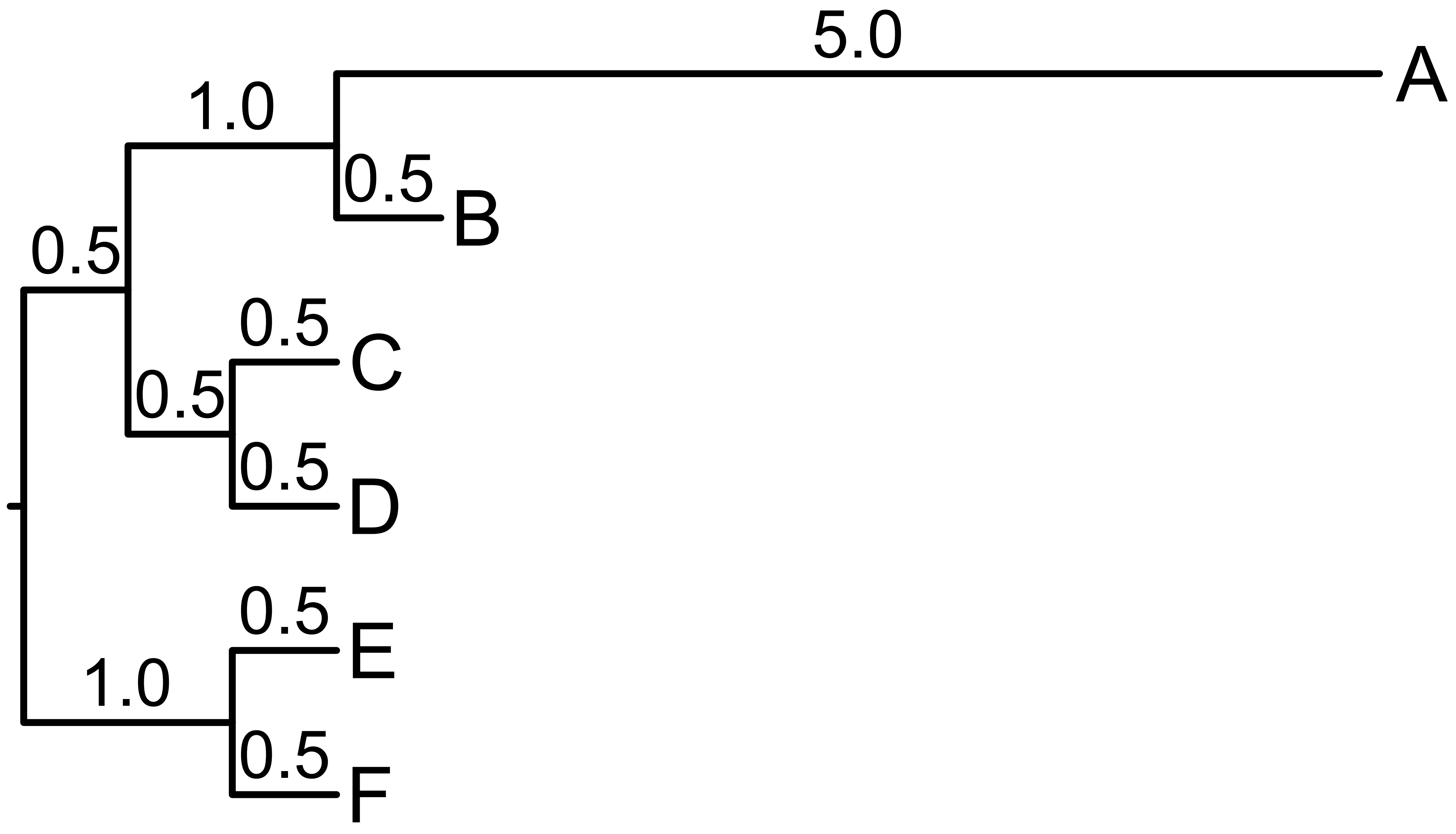
Genes with low DVMC may be more useful for divergence time analysis

# Phylogenetic signal across taxa

1. Long branch score
2. RCVT

@JLSteenwyk

# Long branch score




<u>Long branch scores</u>

A: 73.17
B: -14.63
C: -19.51
D: -19.51
E: -9.76
F: -9.76

# Relative composition variability per taxon

$$RCVT_j = \sum_{i=1}^{c} \frac{c_{ij} - \bar{c}_i}{s \times n}$$

- $RCVT_j$: relative composition variability of $j$th taxon
- $c$: the number of different character states per sequence type in an alignment
- $c_{ij}$: number of occurrences of the $i$th character state for the $j$th taxon
- $\bar{c}_i$: the average number of the $i$th $c$ character state across $n$ taxa
- $s$: total number of sites
- $n$: number of taxa

# Relative composition variability per taxon

```
>1
GGGGGCCC

>2
ATGCATGC

>3
ATGCATGC

>4
ATGCATGC

>5
GGGGGGGG
```

# Relative composition variability per taxon

Sequences 1 and 5 are GC rich

```
>1
GGGGGCCC

>2
ATGCATGC

>3
ATGCATGC

>4
ATGCATGC

>5
GGGGGGGG
```

# Relative composition variability per taxon

Sequences 1 and 5 are GC rich

```
>1
GGGGGCCC

>2
ATGCATGC

>3
ATGCATGC

>4
ATGCATGC

>5
GGGGGGGG
```

Calc RCVT

# Relative composition variability per taxon

Sequences 1 and 5 are GC rich

```
>1
GGGGGCCC

>2
ATGCATGC

>3
ATGCATGC

>4
ATGCATGC

>5
GGGGGGGG
```

Calc RCVT →

```
1
0.12

2
0.09

3
0.09

4
0.09

5
0.21
```

# Relative composition variability per taxon

Sequences 1 and 5 are GC rich

```
>1
GGGGGCCC

>2
ATGCATGC

>3
ATGCATGC

>4
ATGCATGC

>5
GGGGGGGG
```

Calc RCVT →

```
1
0.12

2
0.09

3
0.09

4
0.09

5
0.21
```

Lower values indicate lower biases

# Phylogenetic signal across sites

1. Compositional bias
2. Evolutionary rate

@JLSteenwyk

# Compositional bias per site

```
>1
GGGGGCCC

>2
ATGCATGC

>3
ATGCATGC

>4
ATGCATGC

>5
GGGGGGGG
```

# Compositional bias per site

```
>1
GGGGGCCC
>2
ATGCATGC
>3
ATGCATGC
>4
ATGCATGC
>5
GGGGGGGG
```

→ Calc comp bias per site

| | | | |
|---|---|---|---|
| 1 | 0.2 | 0.6547 | 0.6547 |
| 2 | 0.2 | 0.6547 | 0.6547 |
| 3 | 0.0 | nan | nan |
| 4 | 0.2 | 0.6547 | 0.6547 |
| 5 | 0.2 | 0.6547 | 0.6547 |
| 6 | 1.6 | 0.6547 | 0.4493 |
| 7 | 1.8 | 0.6290 | 0.1797 |
| 8 | 1.8 | 0.6290 | 0.1797 |

# Compositional bias per site

```
>1
GGGGGCCC

>2
ATGCATGC

>3
ATGCATGC

>4
ATGCATGC

>5
GGGGGGGG
```

Calc comp bias per site

| | chi-square | p-val | Multi-test corrected p-val |
|---|---|---|---|
| 1 | 0.2 | 0.6547 | 0.6547 |
| 2 | 0.2 | 0.6547 | 0.6547 |
| 3 | 0.0 | nan | nan |
| 4 | 0.2 | 0.6547 | 0.6547 |
| 5 | 0.2 | 0.6547 | 0.6547 |
| 6 | 1.6 | 0.6547 | 0.4493 |
| 7 | 1.8 | 0.6290 | 0.1797 |
| 8 | 1.8 | 0.6290 | 0.1797 |

# Compositional bias per site

```
>1
GGGGGCCC
>2
ATGCATGC
>3
ATGCATGC
>4
ATGCATGC
>5
GGGGGGGG
```

Calc comp bias per site

| | chi-square | p-val | Multi-test corrected p-val |
|---|---|---|---|
| 1 | 0.2 | 0.6547 | 0.6547 |
| 2 | 0.2 | 0.6547 | 0.6547 |
| 3 | 0.0 | nan | nan |
| 4 | 0.2 | 0.6547 | 0.6547 |
| 5 | 0.2 | 0.6547 | 0.6547 |
| 6 | 1.6 | 0.6547 | 0.4493 |
| 7 | 1.8 | 0.6290 | 0.1797 |
| 8 | 1.8 | 0.6290 | 0.1797 |

# Compositional bias per site

```
>1
GGGGGCCC
>2
ATGCATCC
>3
ATGCATCC
>4
ATGCATCC
>5
GGGGGGGG
```

Calc comp bias per site →

| | chi-square | p-val | Multi-test corrected p-val |
|---|---|---|---|
| 1 | 0.2 | 0.6547 | 0.6547 |
| 2 | 0.2 | 0.6547 | 0.6547 |
| 3 | 0.0 | nan | nan |
| 4 | 0.2 | 0.6547 | 0.6547 |
| 5 | 0.2 | 0.6547 | 0.6547 |
| 6 | 1.6 | 0.6547 | 0.4493 |
| 7 | 1.8 | 0.6290 | 0.1797 |
| 8 | 1.8 | 0.6290 | 0.1797 |

# Evolutionary rate per site

```
>1
GGGGGCCC

>2
ATGCATGC

>3
ATGCATGC

>4
ATGCATGC

>5
GGGGGGGG
```

# Evolutionary rate per site

```
>1
GGGGGCCC

>2
ATGCATGC

>3
ATGCATGC

>4
ATGCATGC

>5
GGGGGGGG
```

Calc evo rate per site

| | |
|---|---|
| 1 | 0.48 |
| 2 | 0.48 |
| 3 | 0.0 |
| 4 | 0.48 |
| 5 | 0.48 |
| 6 | 0.56 |
| 7 | 0.32 |
| 8 | 0.32 |

# Evolutionary rate per site

```
>1
GGGGGCCC

>2
ATGCATGC

>3
ATGCATGC

>4
ATGCATGC

>5
GGGGGGGG
```

Calc evo rate per site

| | |
|---|---|
| 1 | 0.48 |
| 2 | 0.48 |
| **3** | **0.0** |
| 4 | 0.48 |
| 5 | 0.48 |
| 6 | 0.56 |
| 7 | 0.32 |
| 8 | 0.32 |

# Evolutionary rate per site

```
>1
GGGGGCCC
>2
ATGCATCC
>3
ATGCATCC
>4
ATGCATCC
>5
GGGGGGGG
```

Calc evo rate per site

| | |
|---|---|
| 1 | 0.48 |
| 2 | 0.48 |
| 3 | 0.0 |
| 4 | 0.48 |
| 5 | 0.48 |
| 6 | 0.56 |
| 7 | 0.32 |
| 8 | 0.32 |

# Evolutionary rate per site

```
>1
GGGGGCCC
>2
ATGCATCC
>3
ATGCATCC
>4
ATGCATCC
>5
GGGGGGCG
```

→ Calc evo rate per site

| | |
|---|---|
| 1 | 0.48 |
| 2 | 0.48 |
| 3 | 0.0 |
| 4 | 0.48 |
| 5 | 0.48 |
| 6 | 0.56 |
| 7 | 0.32 |
| 8 | 0.32 |

# So many metrics, so many details

1. Alignment length - **higher better**
2. Alignment length with no gaps - **higher better**
3. GC content (for NTs) - **lower better**
4. Pairwise identity - **depends**
5. # of parsimony informative sites - **higher better**
6. # of variable sites - **higher better**
7. Relative composition variability - **lower better**
8. Average bootstrap support value - **higher better**
9. Degree of violation of a molecular clock - **lower better**
10. Evolutionary rate - **depends**
11. Long branch score - **lower better**
12. Treeness - **higher better**
13. Saturation - **lower better**
14. Treeness / RCV - **higher better**
15. RCVT - **lower better**
16. Compositional bias per site - **lower better**
17. Evolutionary rate per site - **depends**

@JLSteenwyk

# Where known, PhyKIT documentation will say

PhyKIT

Search docs

About

⊟ Usage

⊞ General usage

⊞ Alignment-based functions

⊟ **Tree-based functions**

- Bipartition support statistics
- Branch length multiplier
- Collapse bipartitions
- Covarying evolutionary rates
- **Degree of violation of the molecular clock**
- Evolutionary rate

## Degree of violation of the molecular clock

Function names: degree_of_violation_of_a_molecular_clock, dvmc
Command line interface: pk_degree_of_violation_of_a_molecular_clock, pk_dvmc

Calculate degree of violation of a molecular clock (or DVMC) in a phylogeny.

Lower DVMC values are thought to be desirable because they are indicative of a lower degree of violation in the molecular clock assumption.

Typically, outgroup taxa are not included in molecular clock analysis. Thus, prior to calculating DVMC from a single gene tree, users may want to prune outgroup taxa from the phylogeny. To prune tips from a phylogeny, see the prune_tree function.

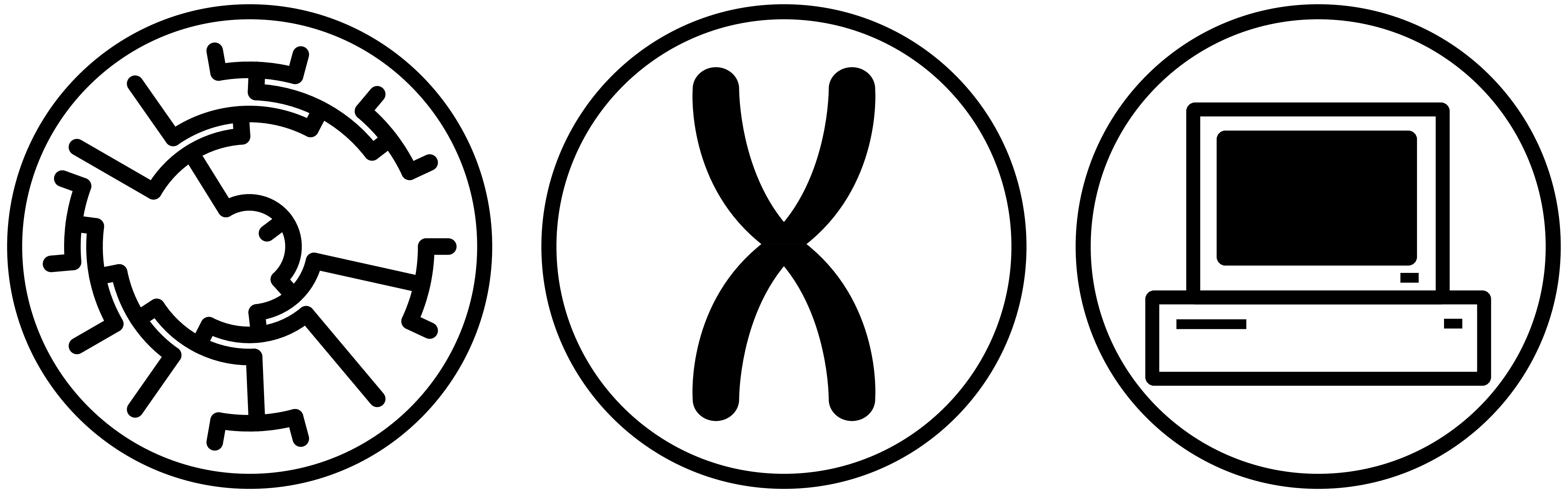
Calculate DVMC in a tree following Liu et al., PNAS (2017), doi: 10.1073/pnas.1616744114.

```
phykit degree_of_violation_of_a_molecular_clock <tree>
```

Options:
<*tree*>: input file tree name

https://jlsteenwyk.com/PhyKIT

# Outline



- Introduction
- Inferring genetic networks from phylogenies
- Phylogenomic subsampling
- **Misc. notes before the tutorial**

@JLSteenwyk

# Misc. notes on the tutorial

- There are steps in the tutorial for plotting
  - These steps are for the sake of completeness
  - But exporting figures in the container is a little complicated
  - Feel free to skip executing these steps
    - But please read and understand them

# Misc. notes on the tutorial

- There are steps in the tutorial for plotting
  - These steps are for the sake of completeness
  - But exporting figures in the container is a little complicated
  - Feel free to skip executing these steps
    - But please read and understand them

- Gemma will have an easier time helping you then me

# Misc. notes on the tutorial

- There are steps in the tutorial for plotting
    - These steps are for the sake of completeness
    - But exporting figures in the container is a little complicated
    - Feel free to skip executing these steps
        - But please read and understand them
- Gemma will have an easier time helping you then me
- Curious about career or something not related related to the workshop?
    - Feel free to ask!

# Thank you for your time and attention!

**King Lab**

Becca Arruda
Chrisa Staikou
Alain G. De Las Bayonas
Maxwell C. Coyle
Josean Reyes-Rivera
Michael Carver
Stefany Gonzalez

hhmi
Howard Hughes Medical Institute

Life Sciences
RESEARCH FOUNDATION

**King, N**

**Coyle, M**

**Buida, J**

**Li, Y**