



# Gene trees challenge



@JLSteenwyk



<https://jlsteenwyk.com/>

**SC-OGs are types of phylogenomic markers**

# Molecular evo. often relies on SC-OGs

Many molecular evolution studies *strictly* rely on single-copy orthologs (SC-OGs)

# Phylogenomics typically relies on SC-OGs

Many molecular evolution studies *strictly* rely on single-copy orthologs (SC-OGs)

- Phylogenomics,
- genome-wide surveys of (+) selection
- gene coevolution analysis
- others

# Phylogenomics typically relies on SC-OGs

Many molecular evolution studies *strictly* rely on single-copy orthologs (SC-OGs)

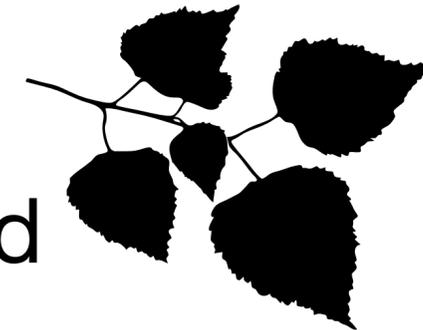
- Phylogenomics,
- genome-wide surveys of (+) selection
- gene coevolution analysis
- others

but SC-OGs are hard to find...

# The quest for SC-OGs

A dataset of 35 plants

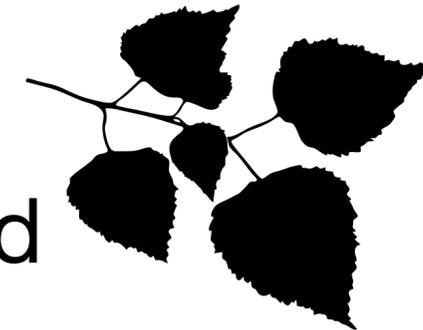
- only one single-copy orthogroup was identified



# The quest for SC-OGs

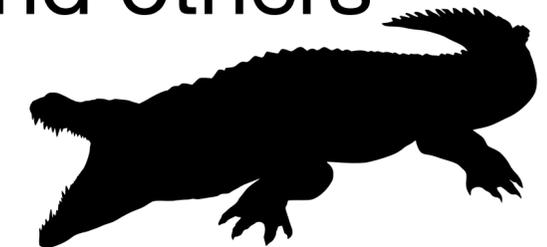
A dataset of 35 plants

- only one single-copy orthogroup was identified



A dataset of 30 turtles, tortoise, birds, crocodile, alligators, and others

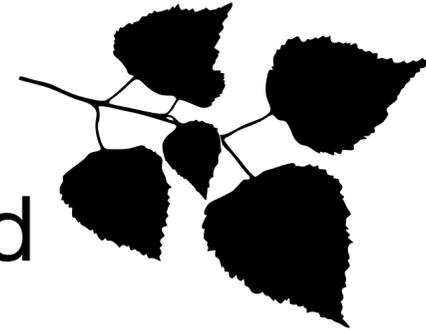
- only 27 single-copy orthogroups identified



# The quest for SC-OGs

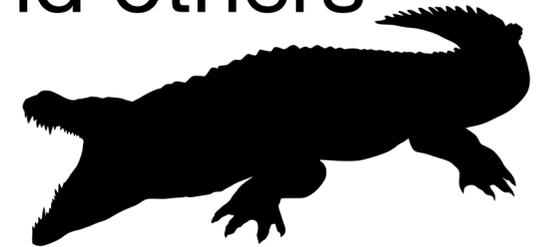
A dataset of 35 plants

- only one single-copy orthogroup was identified



A dataset of 30 turtles, tortoise, birds, crocodile, alligators, and others

- only 27 single-copy orthogroups identified



A dataset of 76 arthropods (Thomas et al. (2020), *Genome Biology*)

- Zero single-copy orthogroups with 100% occupancy



# Too few phylogenomic markers :(

A dataset of 35 plants  
• only one single-copy

A dataset of 30 turtles,  
• only 27 single-copy o

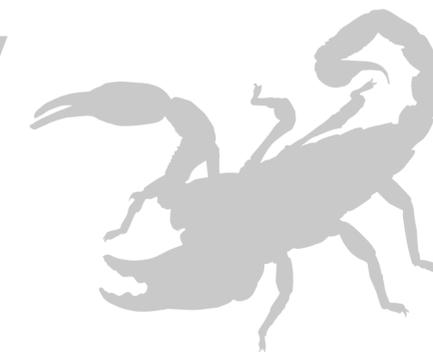
A dataset of 76 arthropods  
• Zero single-copy orthogroups with 100% occupancy



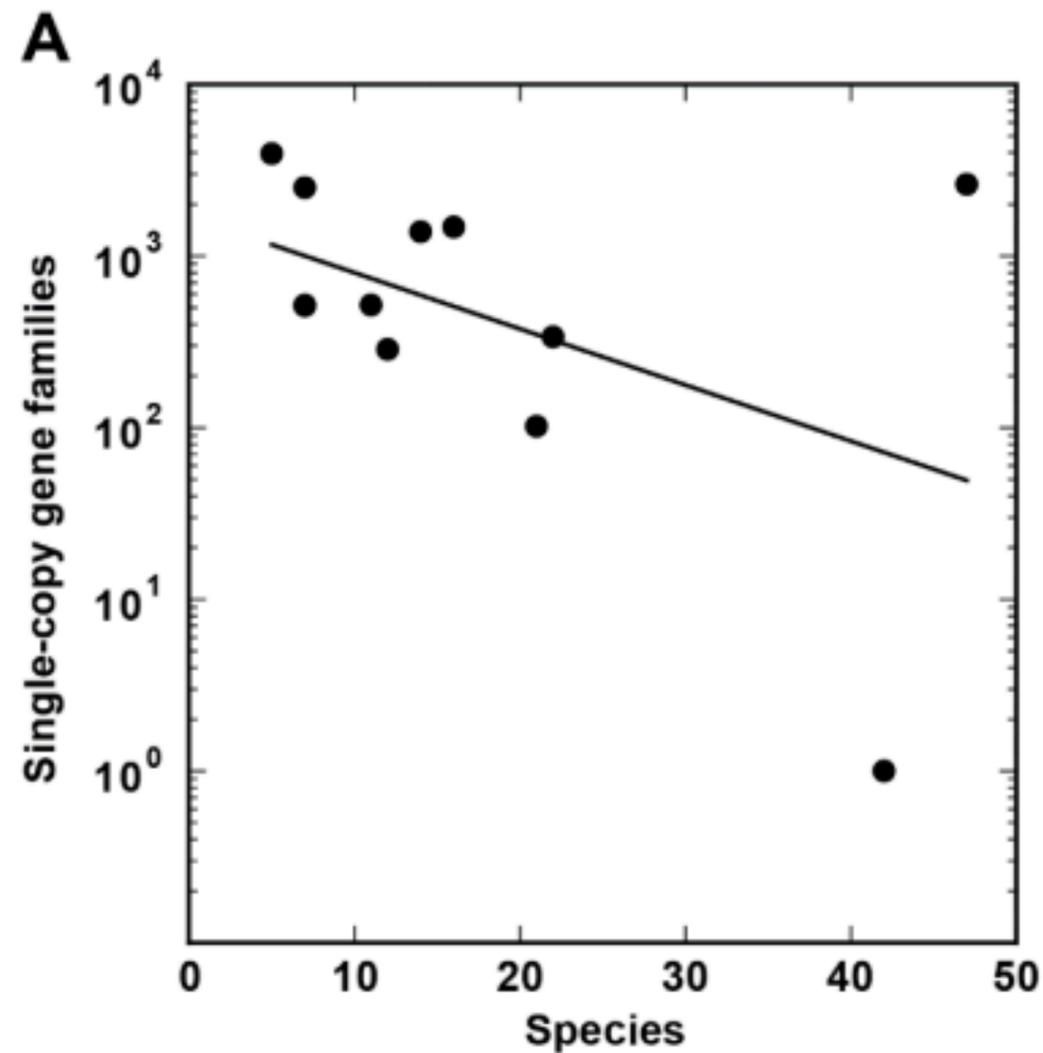
e, alligators, and others



!0), *Genome Biology*)

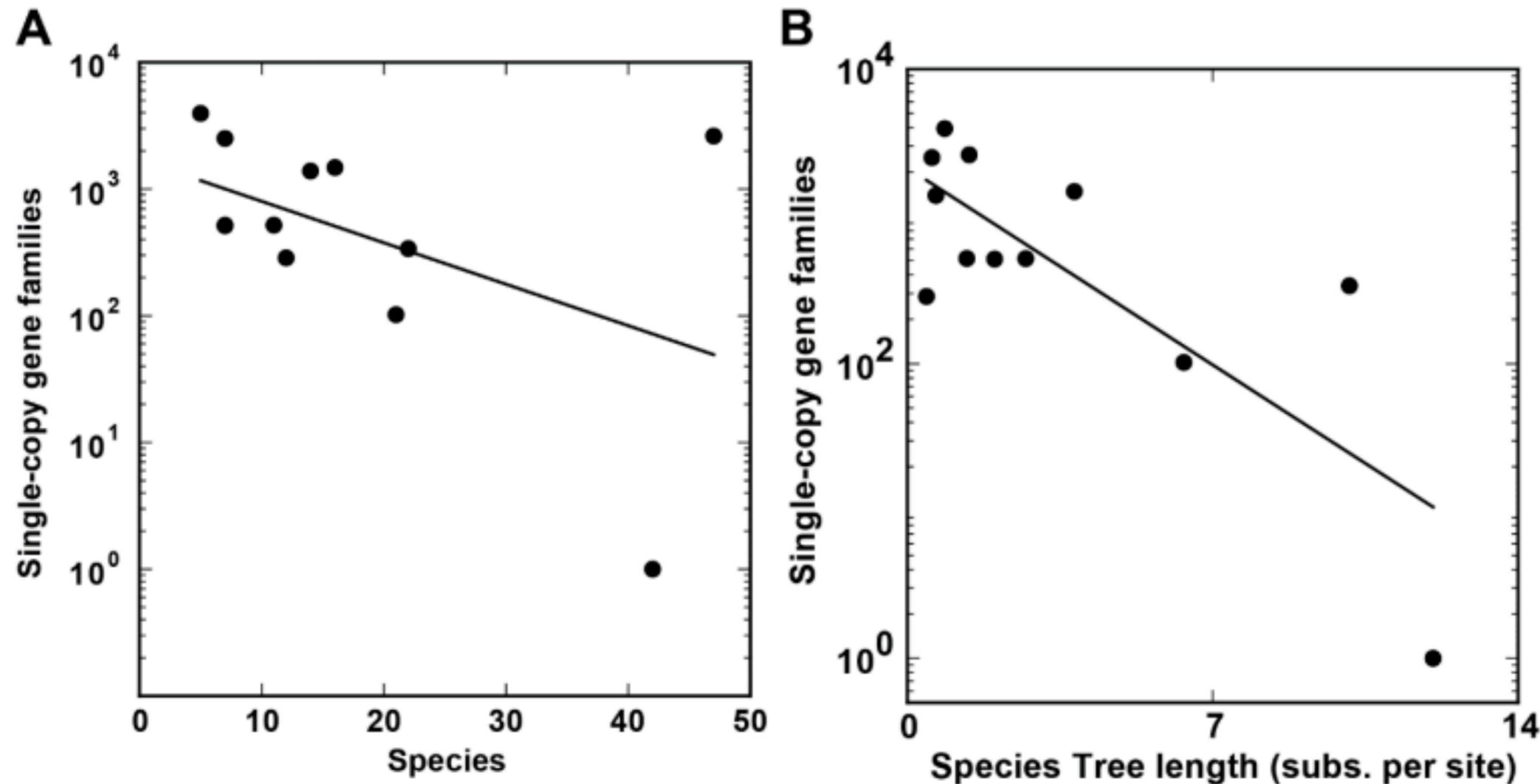


# Factors that impact SC-OG identification



The number of single-copy orthologs decreases as the number of species and evolutionary distance among species increases

# Factors that impact SC-OG identification

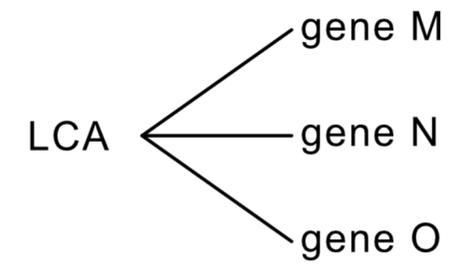


The number of single-copy orthologs decreases as the number of species and evolutionary distance among species increases

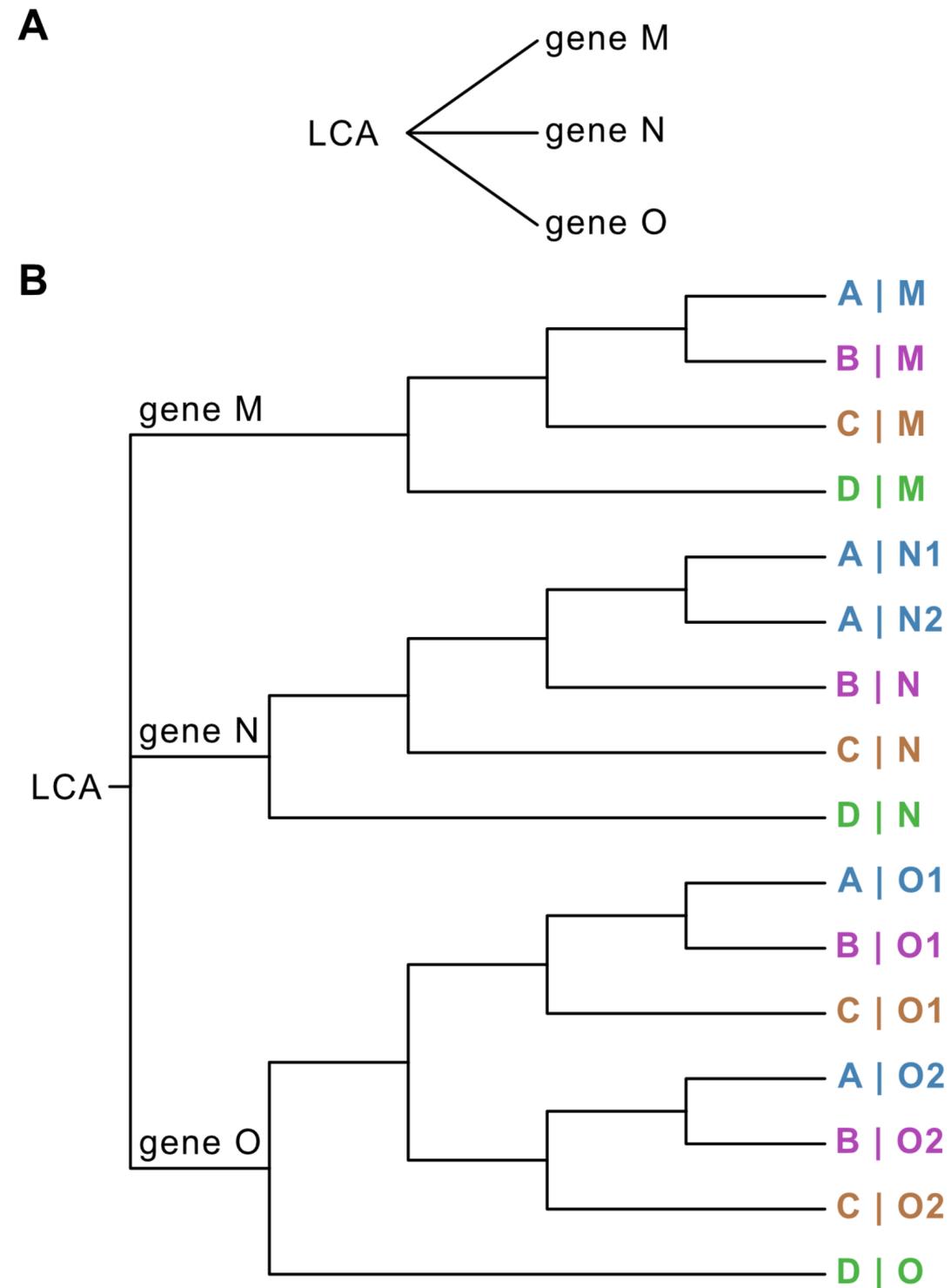
Can other types of orthologs be used for molecular evolution studies?

# Factors that impact SC-OG identification

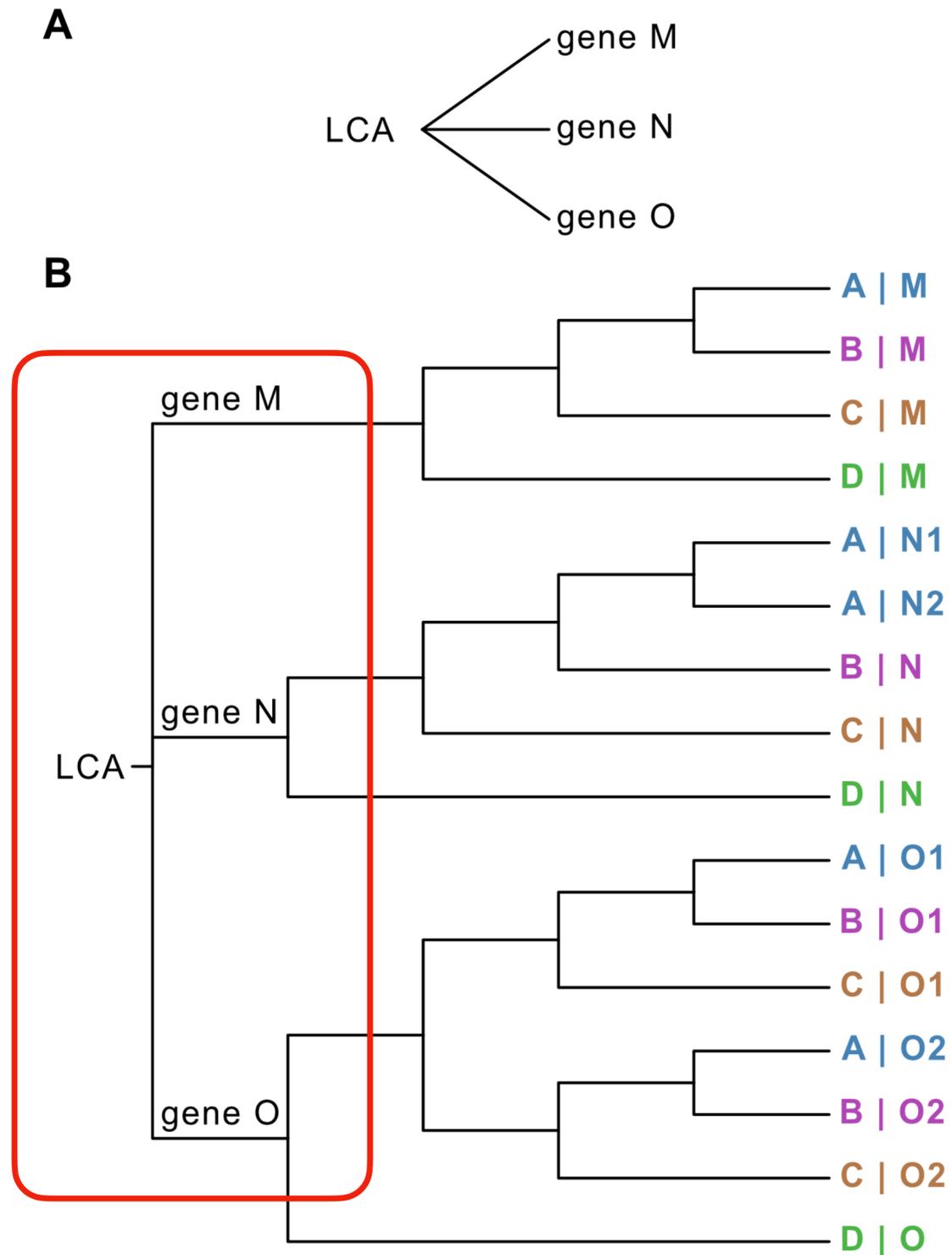
**A**



# Factors that impact SC-OG identification

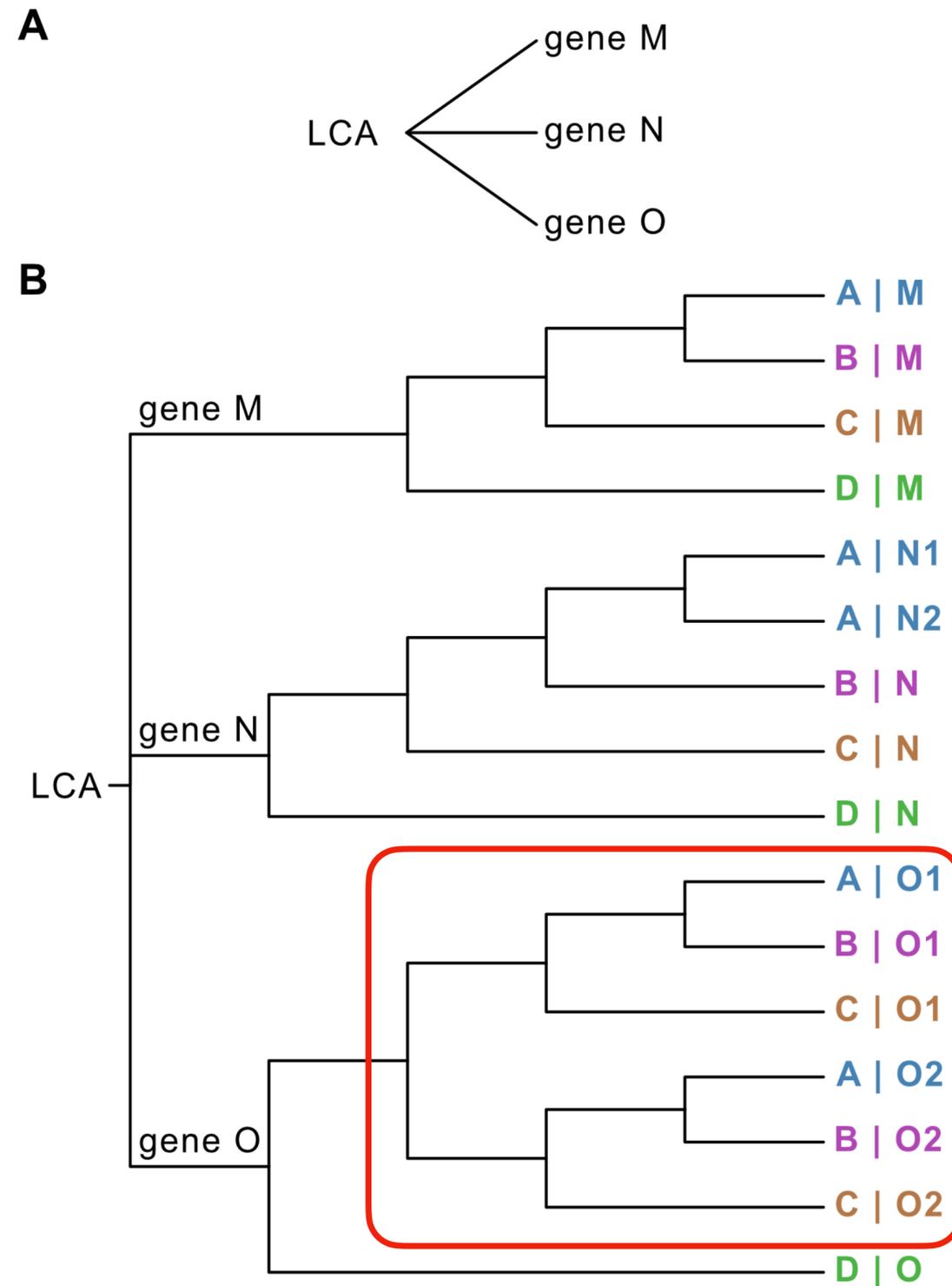


# Factors that impact SC-OG identification



- Gene M, N, and O are outparalogs – paralogous genes wherein duplication occurred prior to a speciation event

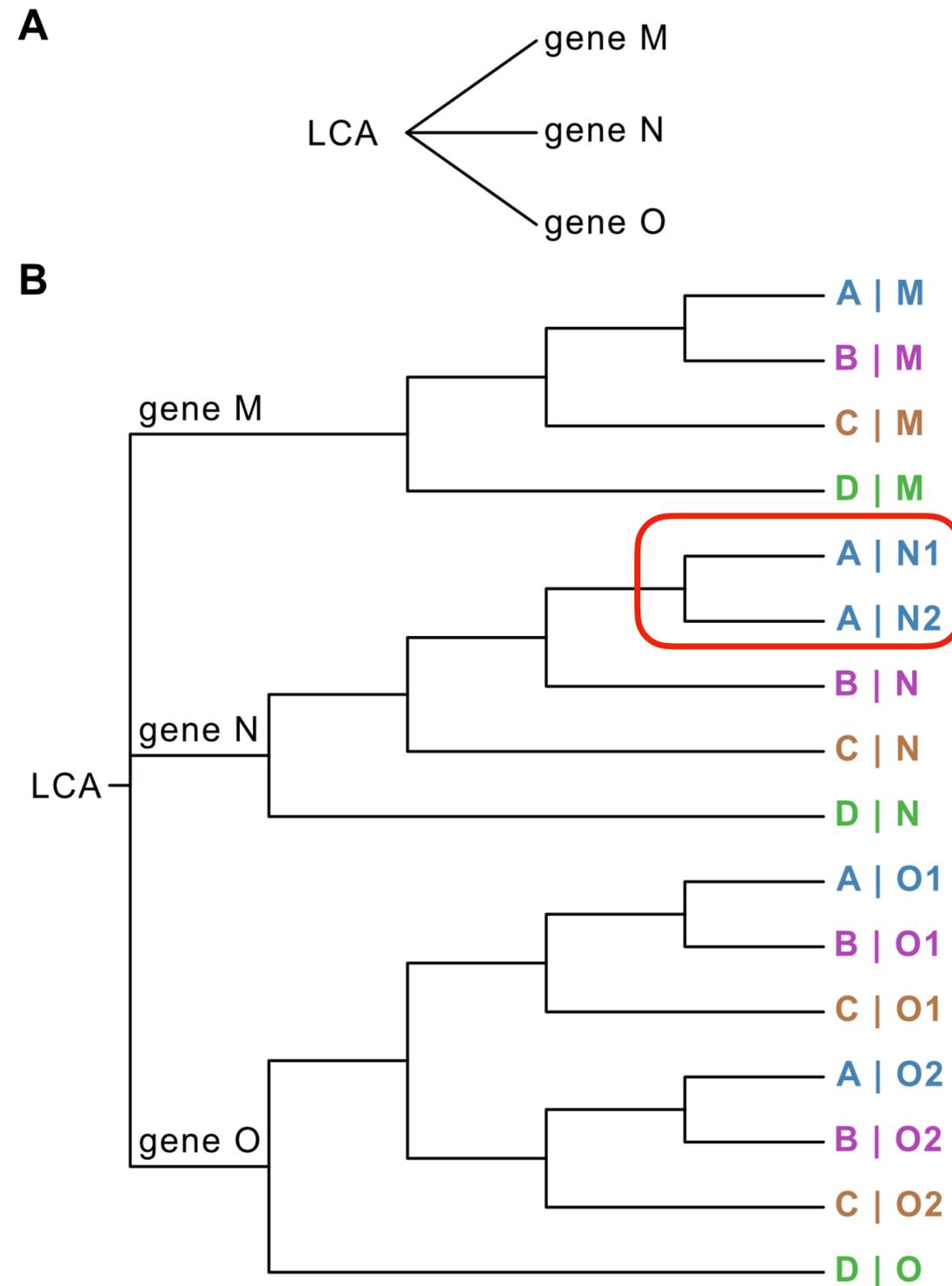
# Factors that impact SC-OG identification



- Gene M, N, and O are outparalogs – paralogous genes wherein duplication occurred prior to a speciation event

- A, B, and C O1 and O2 are inparalogs – paralogous genes wherein duplication occurred after a speciation event

# Factors that impact SC-OG identification

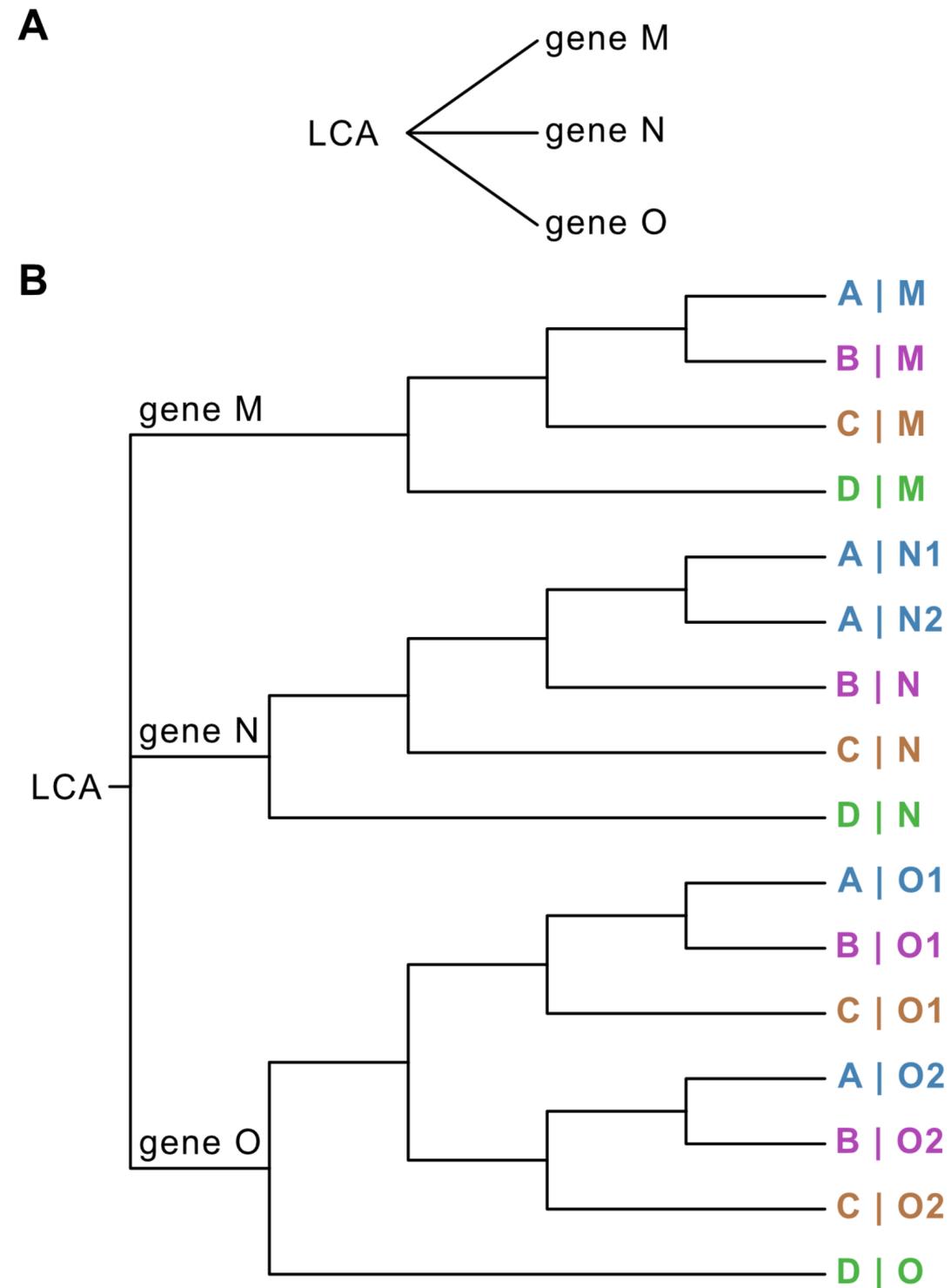


- Gene M, N, and O are outparalogs – paralogous genes wherein duplication occurred prior to a speciation event

- A, B, and C O1 and O2 are inparalogs – paralogous genes wherein duplication occurred after a speciation event

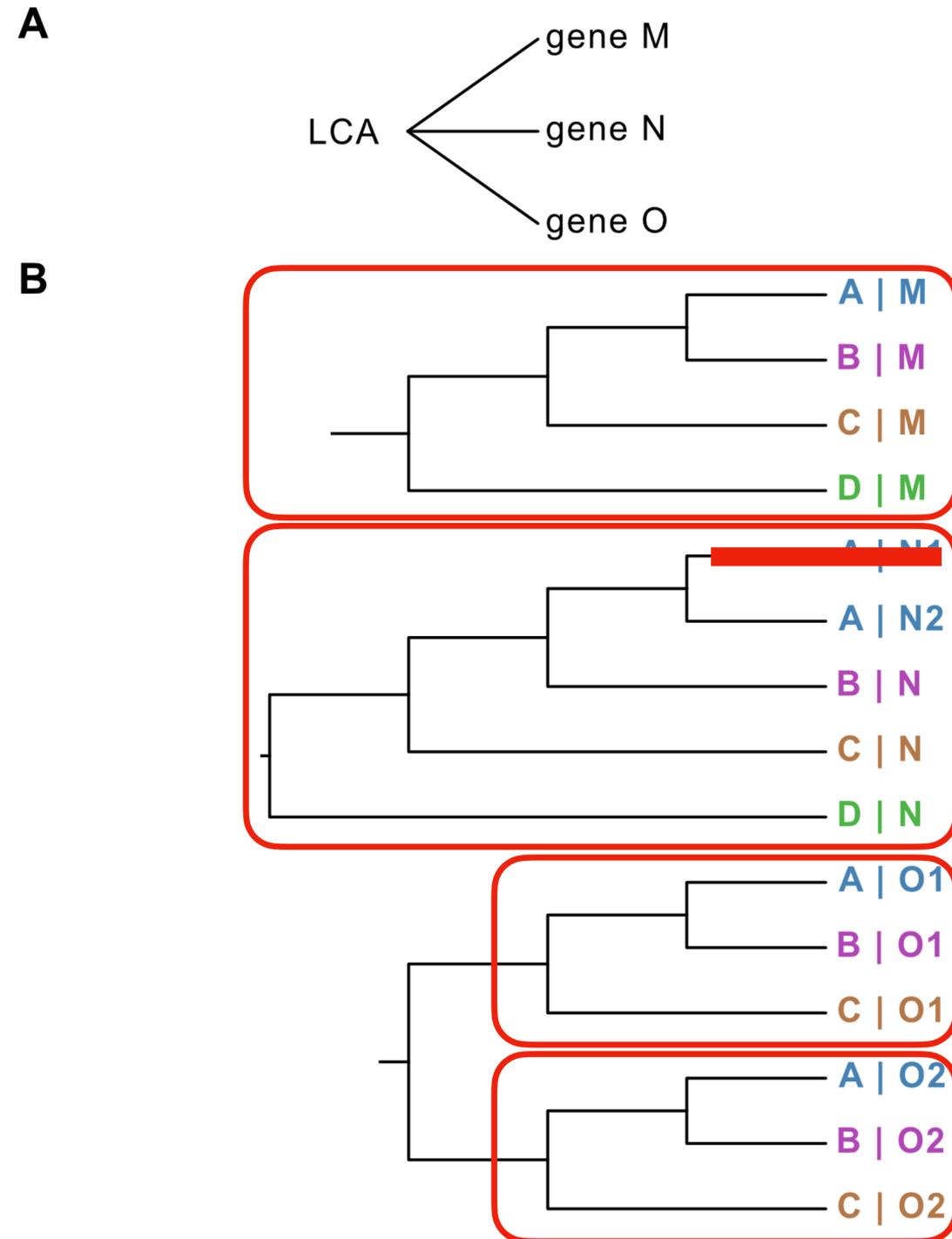
- A | N1 and A | N2 are within species inparalogs

# Factors that impact SC-OG identification



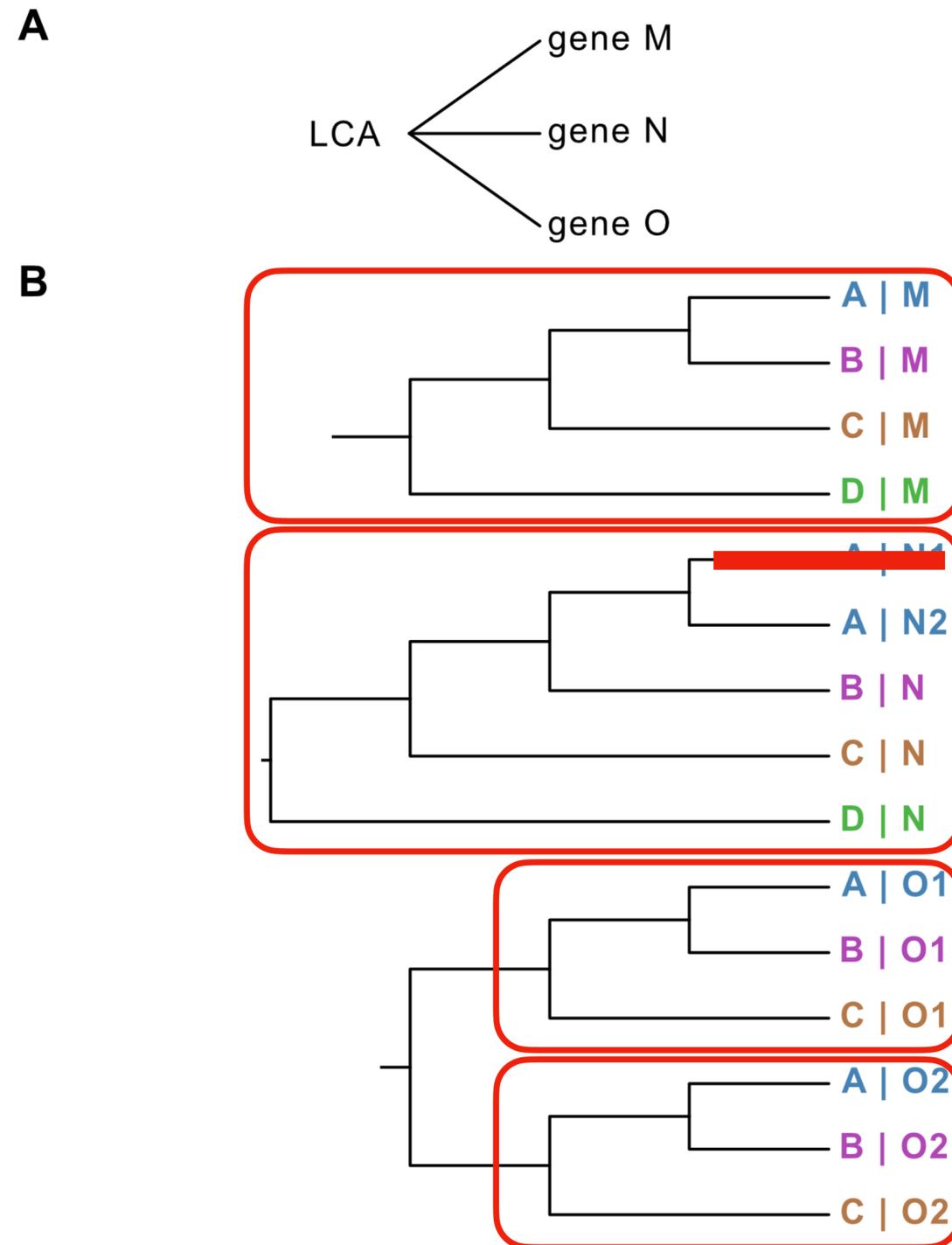
*Note, splitting this tree will result in multiple subgroups of single-copy orthologs*

# Factors that impact SC-OG identification



subgroups of  
single-copy in-  
and outparalogs  
from multi-copy  
orthologous  
groups of genes

# Factors that impact SC-OG identification



subgroups of  
single-copy in-  
and outparalogs  
from multi-copy  
orthologous  
groups of genes

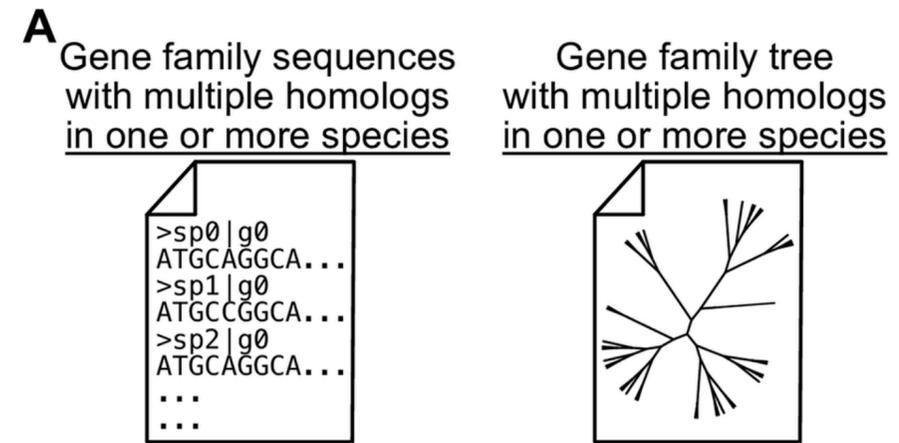
We term these SNAP-OGs  
because they are orthologs  
that have undergone a  
**s**plitting **a**nd **p**runing  
procedure

# OrthoSNAP

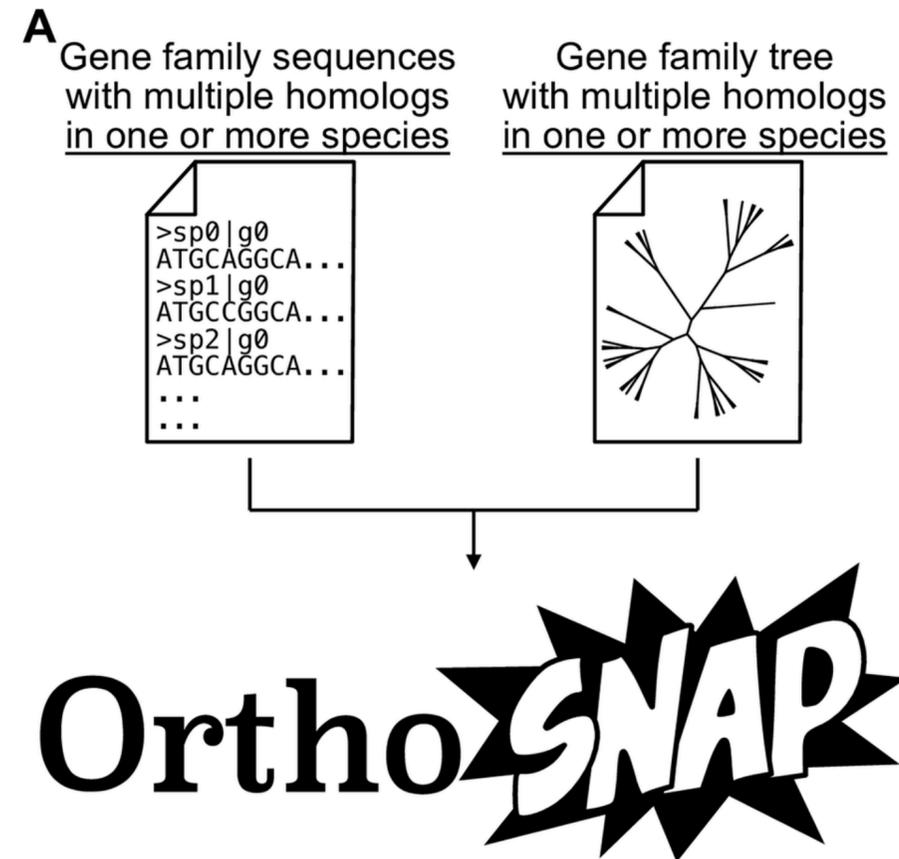
---

**identify single-copy orthologous genes  
nested within larger gene families**

# OrthoSNAP methods



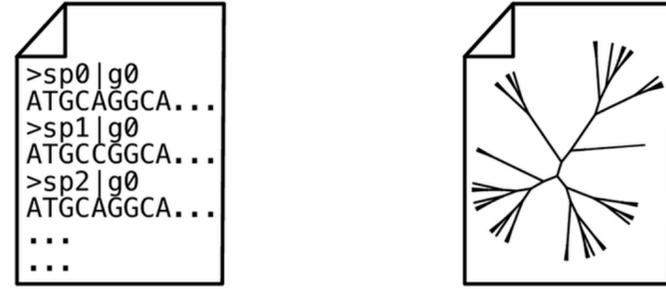
# OrthoSNAP methods





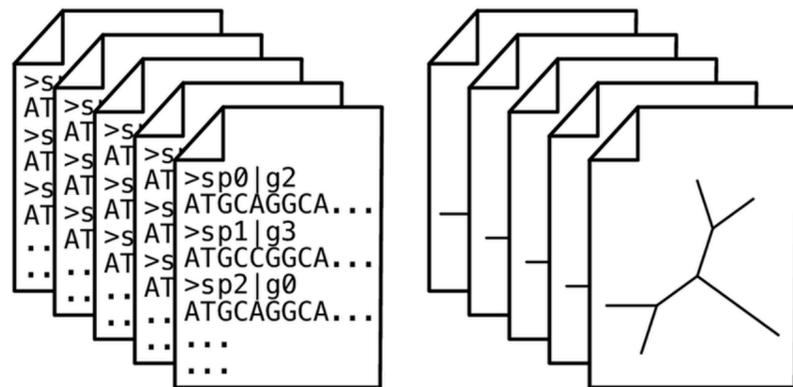
# OrthoSNAP methods

**A** Gene family sequences with multiple homologs in one or more species      Gene family tree with multiple homologs in one or more species

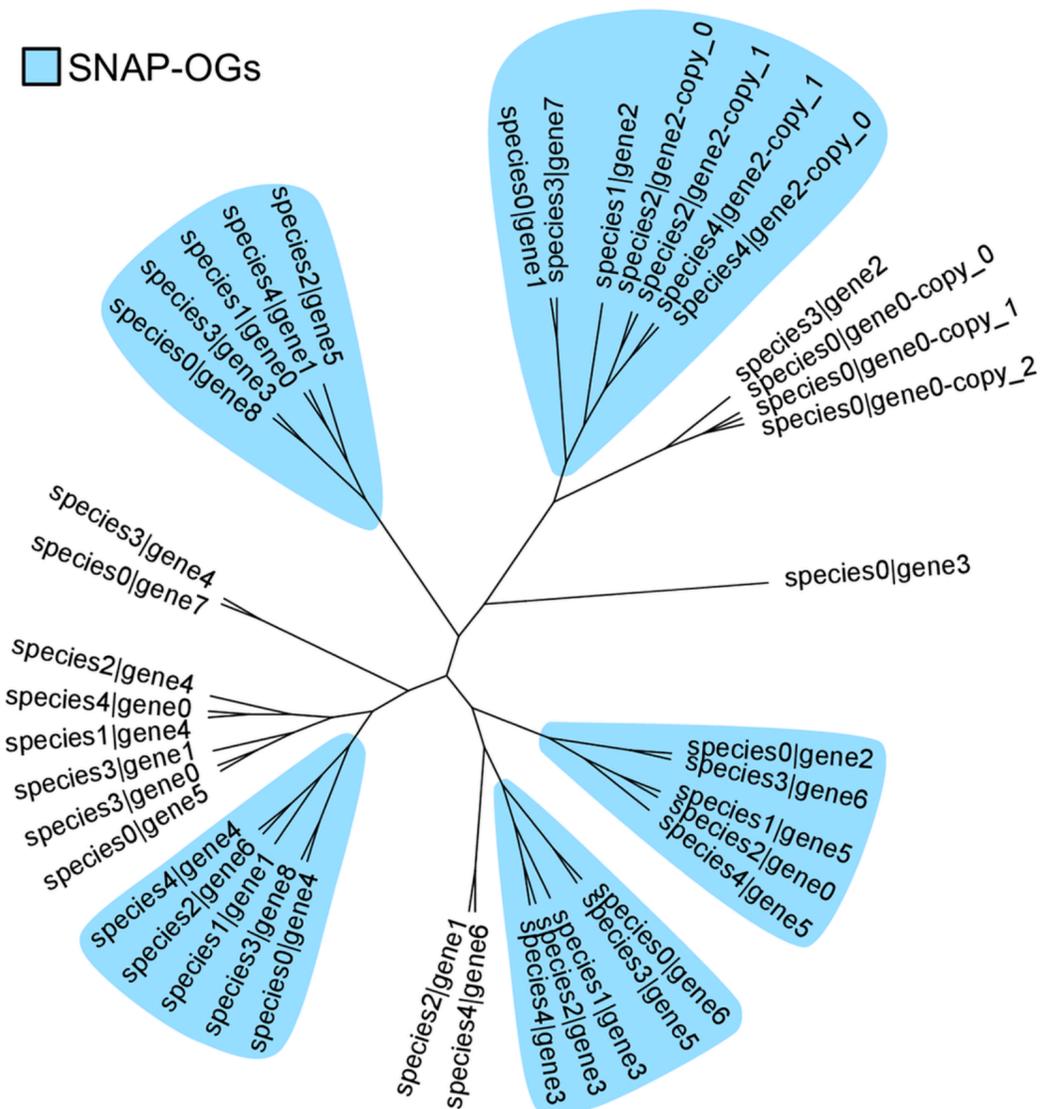


# OrthoSNAP

Individual FASTA and Newick tree files populated with subgroups of single-copy orthologous genes (SNAP-OGs)



**B** Cartoon depiction of OrthoSNAP tree-splitting



Naming convention for sequences

species1 | gene1  
 taxon identifier separator gene identifier

# SNAP-OGs can substantially increase datasets

	SC-OGs	SNAP-OGs	Fold difference
Budding yeast (No WGD)	1,668	1,392	0.83

# SNAP-OGs can substantially increase datasets

	SC-OGs	SNAP-OGs	Fold difference
Budding yeast (No WGD)	1,668	1,392	0.83
Budding yeast (WGD)	2,782	1,334	0.48

# SNAP-OGs can substantially increase datasets

	SC-OGs	SNAP-OGs	Fold difference
Budding yeast (No WGD)	1,668	1,392	0.83
Budding yeast (WGD)	2,782	1,334	0.48
Filamentous fungi ( <i>Aspergillus</i> and	4,393	2,035	0.46

# SNAP-OGs can substantially increase datasets

	SC-OGs	SNAP-OGs	Fold difference
Budding yeast (No WGD)	1,668	1,392	0.83
Budding yeast (WGD)	2,782	1,334	0.48
Filamentous fungi ( <i>Aspergillus</i> and	4,393	2,035	0.46
Mammals (Eutherians)	321	1,775	5.53

# SNAP-OGs can substantially increase datasets

	SC-OGs	SNAP-OGs	Fold difference
<b>Budding yeast</b> (No WGD)	1,668	1,392	0.83
<b>Budding yeast</b> (WGD)	2,782	1,334	0.48
<b>Filamentous fungi</b> ( <i>Aspergillus</i> and	4,393	2,035	0.46
<b>Mammals</b> (Eutherians)	321	1,775	5.53
<b>Plants</b> (Complex dup. And	15	653	43.53

# SNAP-OGs can substantially increase datasets

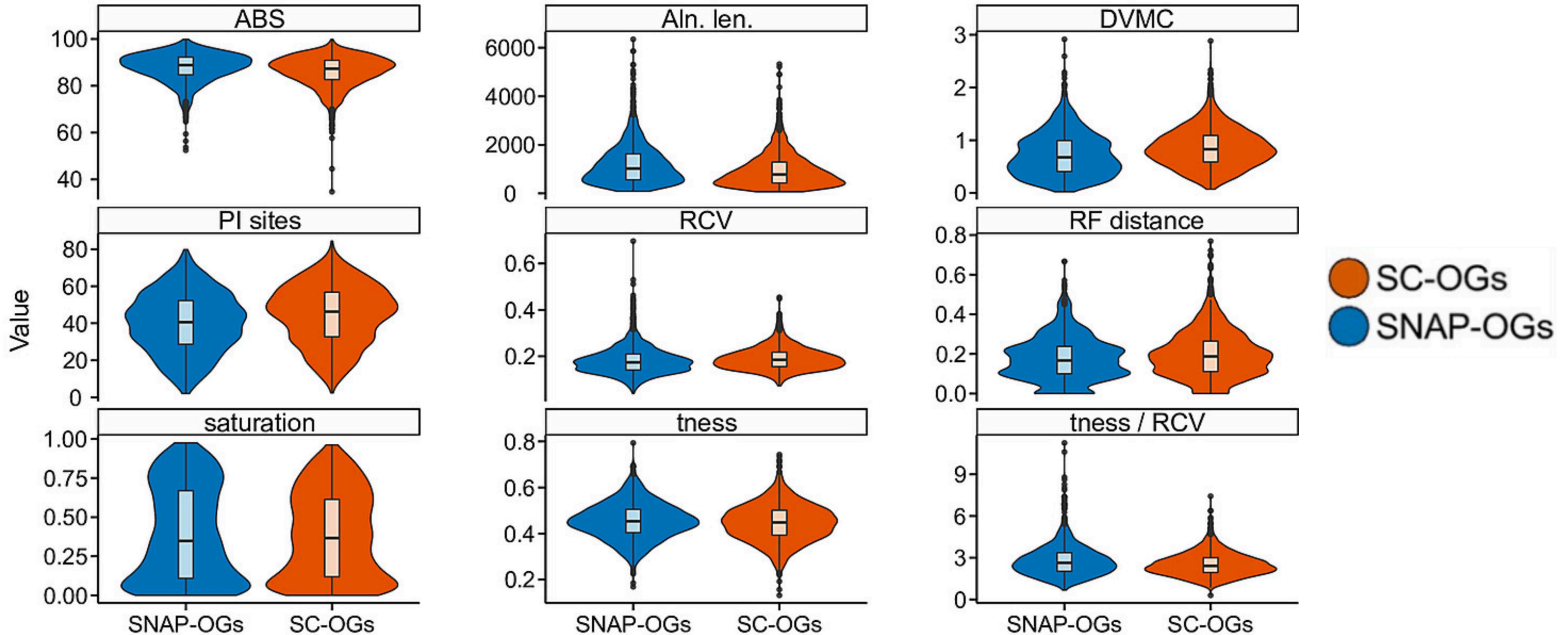
	SC-OGs	SNAP-OGs	Fold difference
<b>Budding yeast</b> (No WGD)	1,668	1,392	0.83
<b>Budding yeast</b> (WGD)	2,782	1,334	0.48
<b>Filamentous fungi</b> ( <i>Aspergillus</i> and	4,393	2,035	0.46
<b>Mammals</b> (Eutherians)	321	1,775	5.53
<b>Plants</b> (Complex dup. And	15	653	43.53
<b>Choanoflagellate</b> (Transcriptomes)	390	2,087	5.35

*But are SNAP-OGs bad markers?*

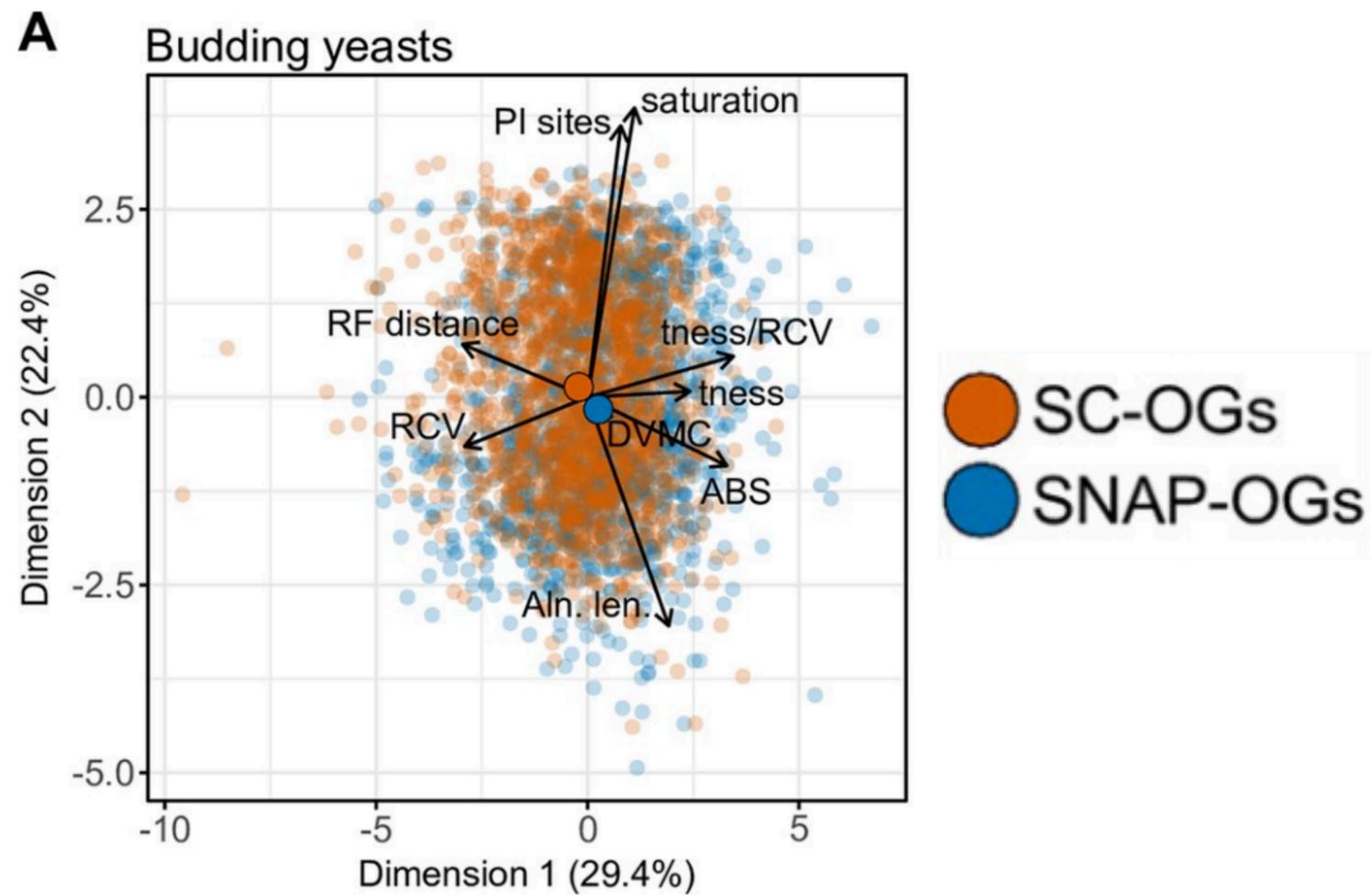
# SNAP- & SC-OGs are statistically indistinguishable

**A**

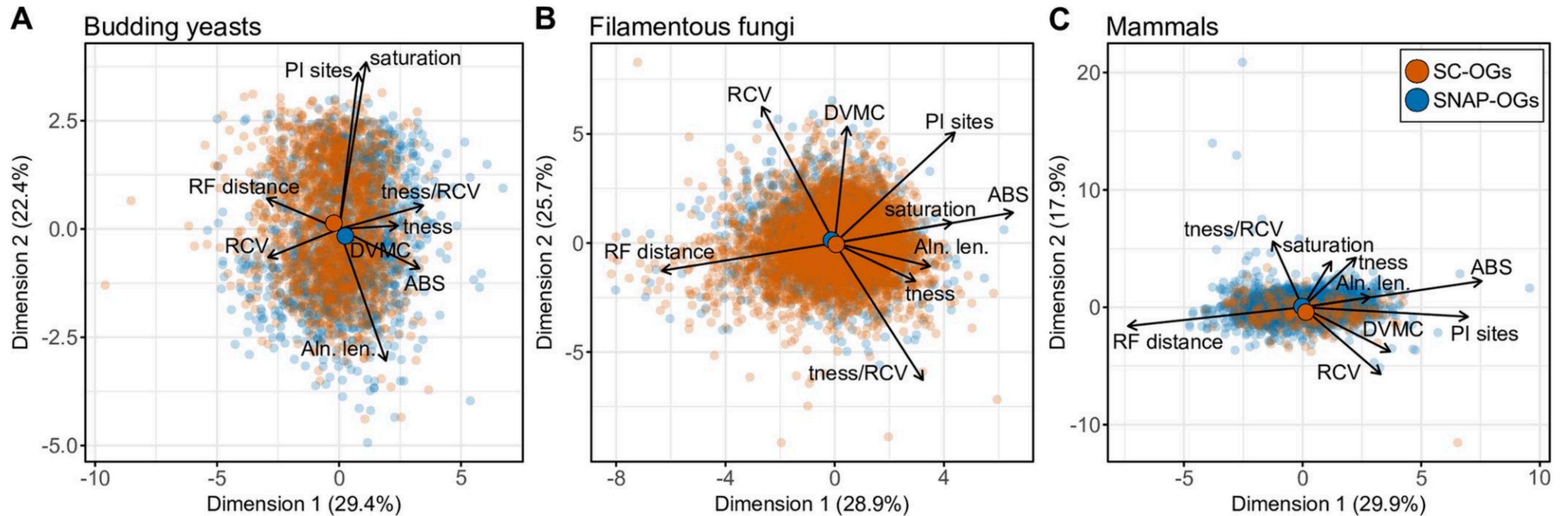
## Budding yeasts



# SNAP- & SC-OGs are statistically indistinguishable



# SNAP- & SC-OGs are statistically indistinguishable



# Phylogenomics typically relies on SC-OGs

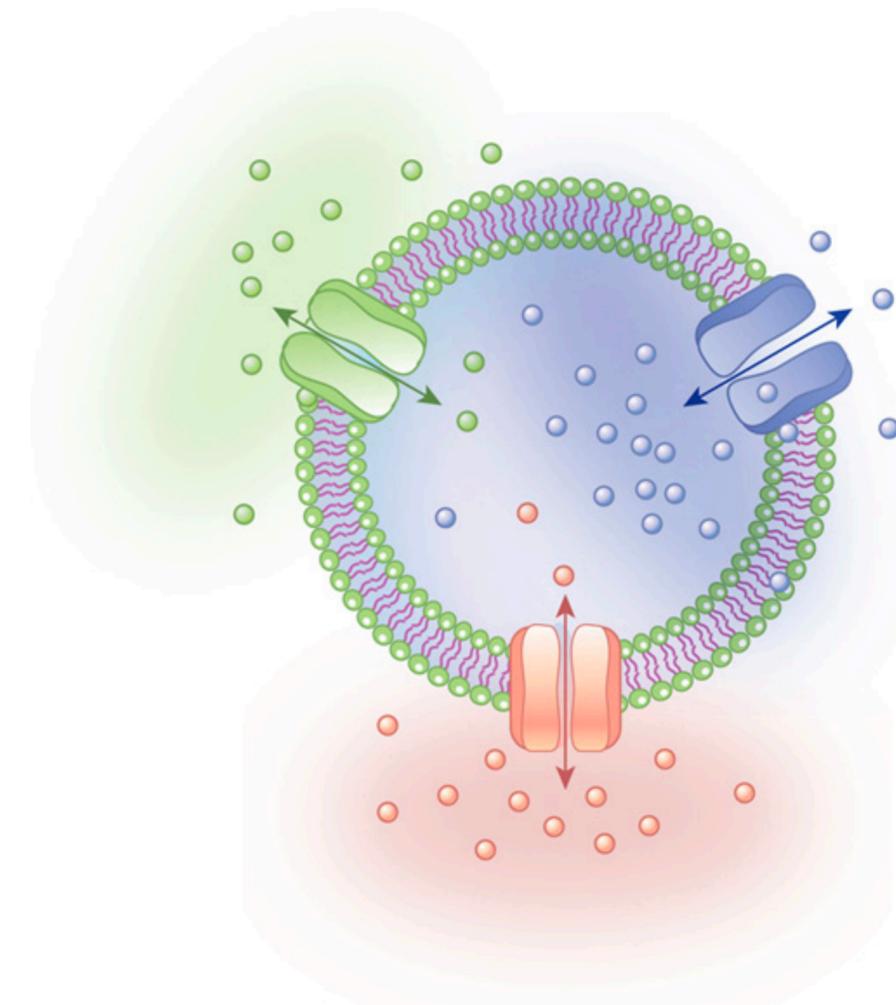
- High-throughput screens of (+) selection requires SC-OGs

# Phylogenomics typically relies on SC-OGs

- High-throughput screens of (+) selection requires SC-OGs
- What types of genes are not typically SC-OGs?
  - Receptors
  - Heat shock proteins
  - Transporters
  - Transcription factors
  - Kinases
  - Etc...

# Molecular evolution of *all* types of genes

- 5 SNAP-OGs were identified in OGs of transcription factors
- 5 SNAP-OGs were identified in OGs of MFS transporters
- 4 SNAP-OGs were identified in an OG of kinases



# Phylogenomics typically relies on SC-OGs

- High-throughput screens of (+) selection requires SC-OGs
- What types of genes are not typically SC-OGs?
  - **Receptors**
  - **Heat shock proteins**
  - **Transporters**
  - **Transcription factors**
  - **Kinases**
  - Etc...

# OrthoSNAP

---

**identify single-copy orthologous genes  
nested within larger gene families**

# Gene trees challenge