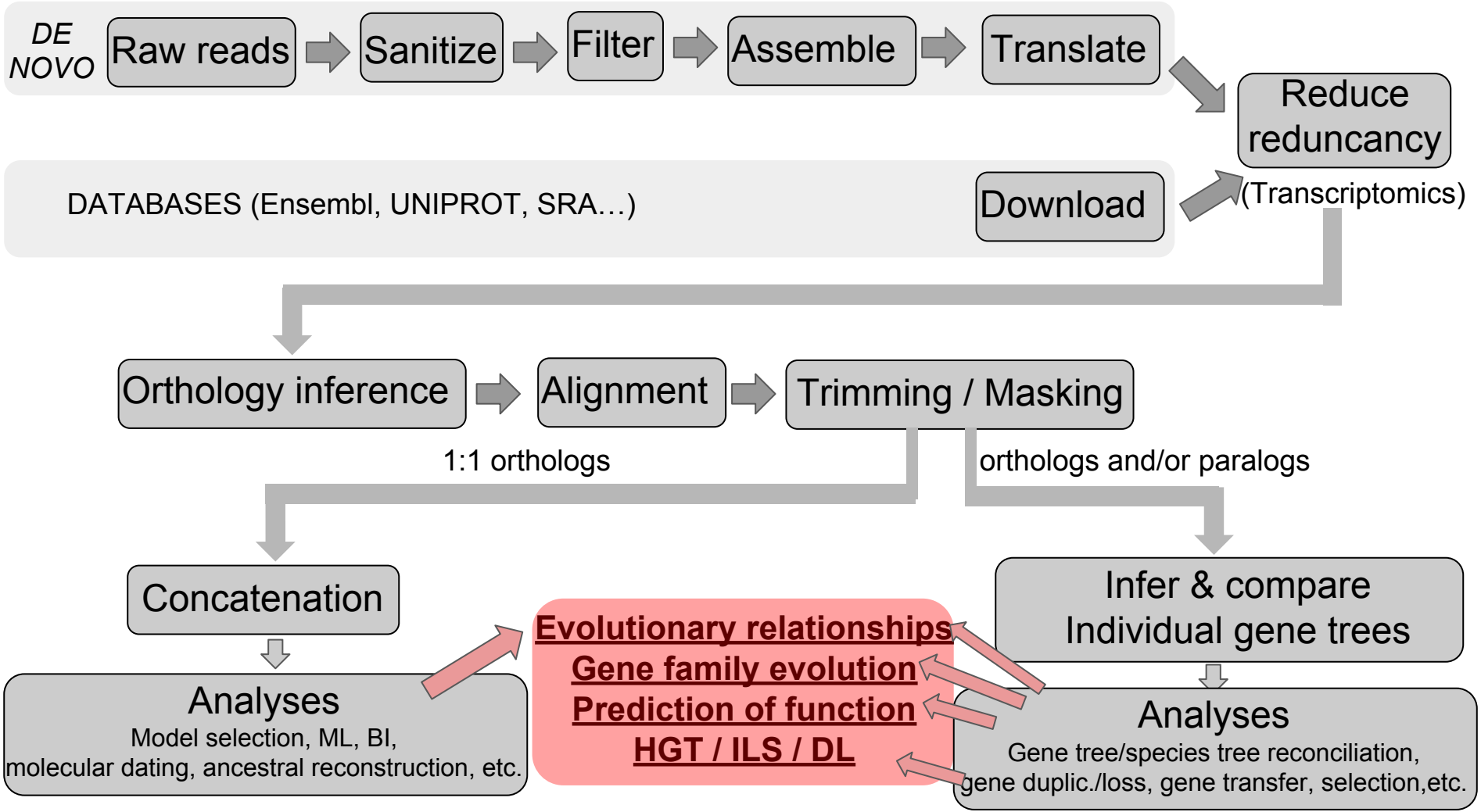
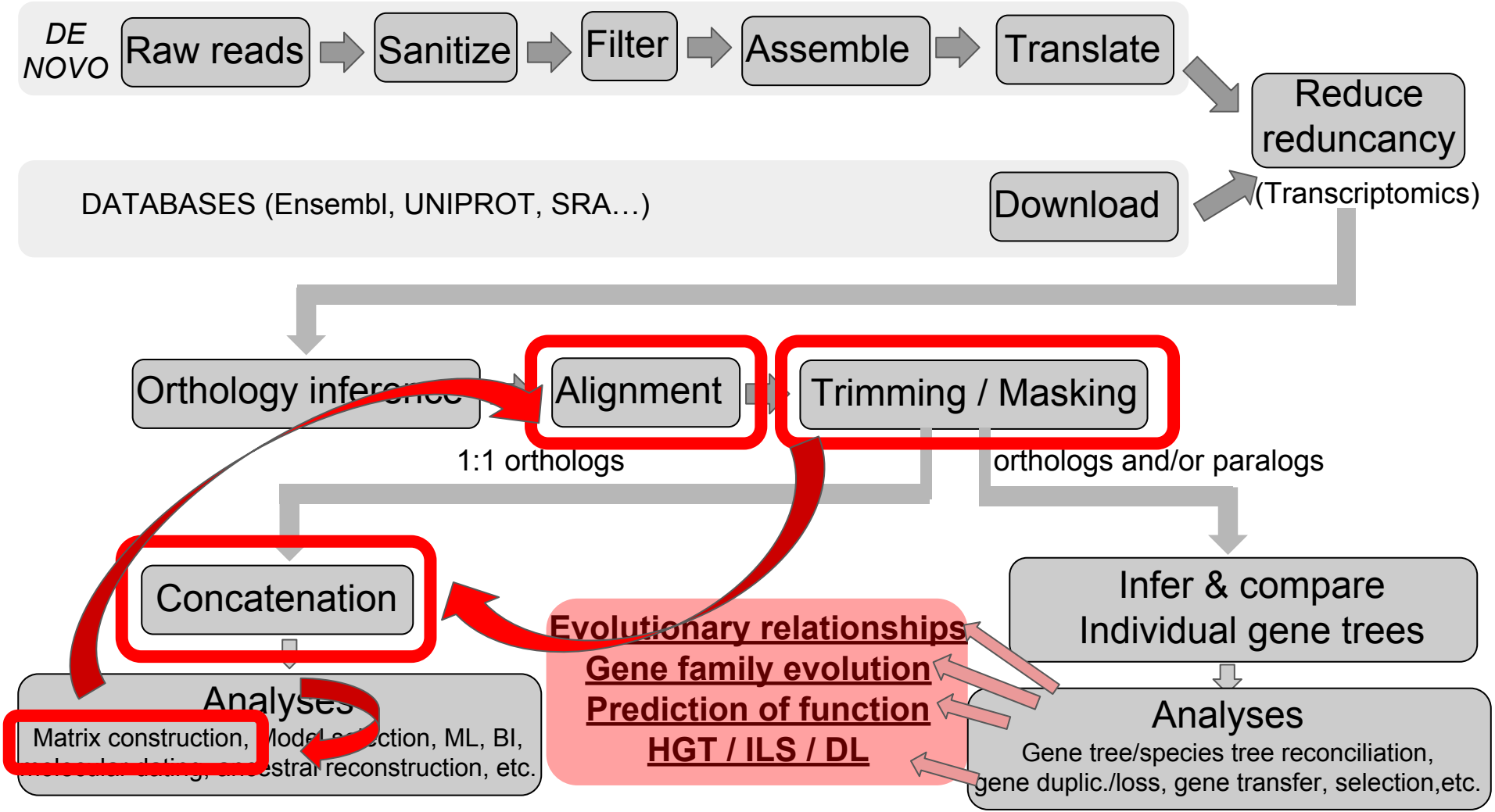
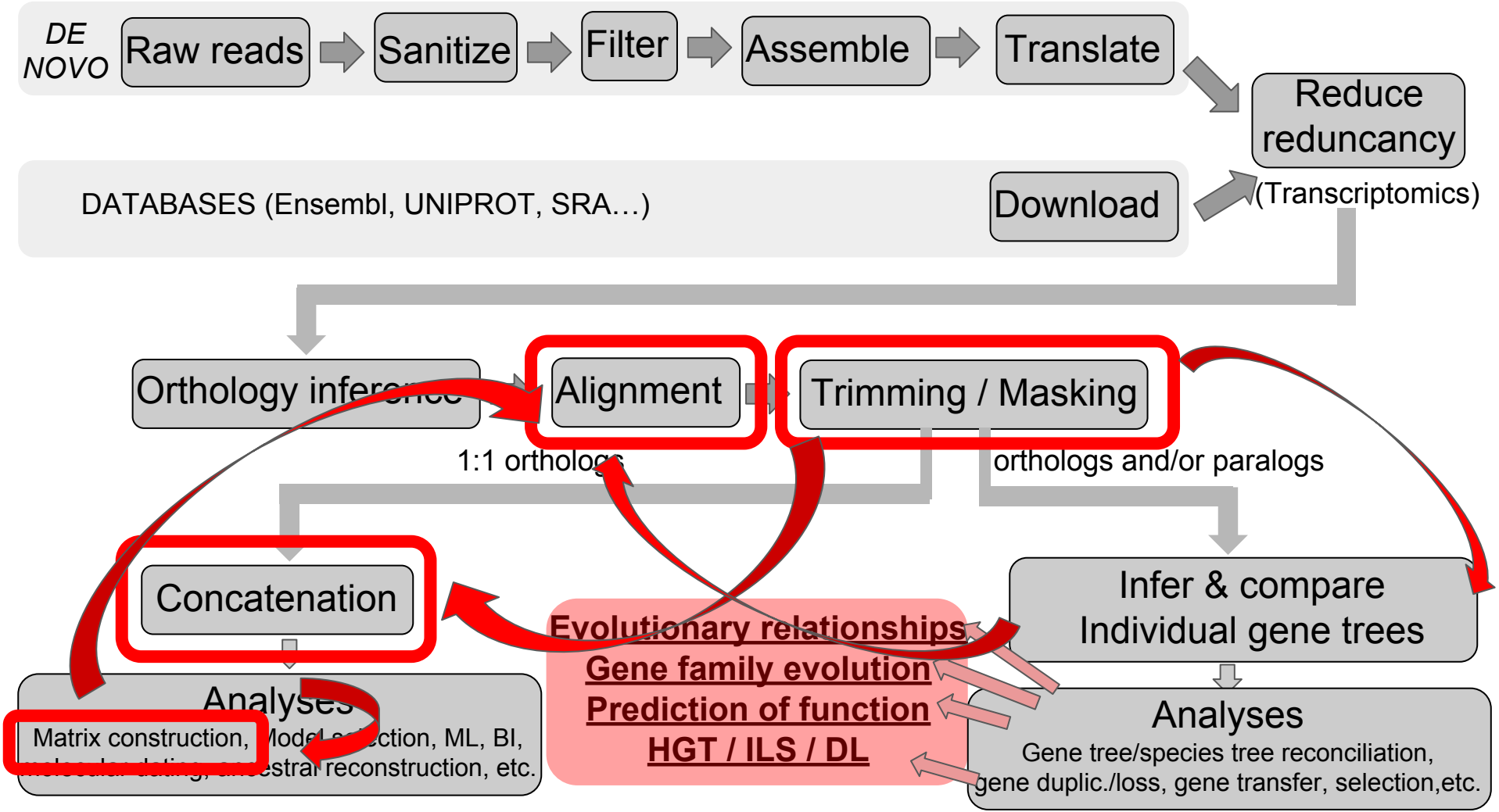
Techniques for generating phylogenomic data matrices: transcriptomics vs genomics

Rosa Fernández & Marina Marcet-Houben


DE NOVO
Raw reads
Sanitize
Filter
Assemble
Translate
Reduce reduncancy
(Transcriptomics)
DATABASES (Ensembl, UNIPROT, SRA…)
Download
Orthology inference
Alignment
Trimming / Masking
1:1 orthologs
orthologs and/or paralogs
Concatenation
Infer & compare Individual gene trees
Evolutionary relationships
Gene family evolution
Prediction of function
HGT / ILS / DL
Analyses
Model selection, ML, BI, molecular dating, ancestral reconstruction, etc.
Analyses
Gene tree/species tree reconciliation, gene duplic./loss, gene transfer, selection,etc.


DE NOVO
Raw reads
Sanitize
Filter
Assemble
Translate
Reduce reduncancy
(Transcriptomics)
DATABASES (Ensembl, UNIPROT, SRA…)
Download
Orthology inference
Alignment
Trimming / Masking
1:1 orthologs
orthologs and/or paralogs
Concatenation
Infer & compare Individual gene trees
Evolutionary relationships
Gene family evolution
Prediction of function
HGT / ILS / DL
Analyses
Matrix construction,
ML, BI,
reconstruction, etc.
Analyses
Gene tree/species tree reconciliation,
gene duplic./loss, gene transfer, selection,etc.


DE NOVO
Raw reads
Sanitize
Filter
Assemble
Translate
Reduce reduncancy
(Transcriptomics)
DATABASES (Ensembl, UNIPROT, SRA…)
Download
Orthology inference
Alignment
Trimming / Masking
1:1 orthologs
orthologs and/or paralogs
Concatenation
Infer & compare Individual gene trees
Evolutionary relationships
Gene family evolution
Prediction of function
HGT / ILS / DL
Analyses
Matrix construction, Model selection, ML, BI, reconstruction, etc.
Analyses
Gene tree/species tree reconciliation, gene duplic./loss, gene transfer, selection,etc.

TECHNIQUES FOR GENERATING PHYLOGENOMIC DATA MATRICES

INHERENT TO THE TECHNIQUE

FACTORS	GENOMICS	TRANSCRIPTOMICS	TARGET ENRICHMENT	RADseq
MISSING DATA	LOW	HIGH	HIGH	HIGH
BALANCED LINEAGE REPRESENTATION	LOW	HIGH	HIGH	HIGH (ALLELE DROP OUT)
GENE DUPLICATION	REAL DUPLICATIONS	ISOFORMS	DEPENDS ON PROBE DESIGN	-
HORIZONTAL GENE TRANSFER	YES	NO	?	NO

TECHNIQUES FOR GENERATING PHYLOGENOMIC DATA MATRICES

		WHOLE GENOME	GENOME REDUCTION		
	FACTORS	**GENOMICS**	**TRANSCRIPTOMICS**	**TARGET ENRICHMENT**	**RADseq**
INHERENT TO THE TECHNIQUE	MISSING DATA	LOW	HIGH	HIGH	HIGH
	BALANCED LINEAGE REPRESENTATION	LOW	HIGH	HIGH	HIGH (ALLELE DROP OUT)
	GENE DUPLICATION	REAL DUPLICATIONS	ISOFORMS	DEPENDS ON PROBE DESIGN	CAN'T BE ASSESSED
	HORIZONTAL GENE TRANSFER	CAN BE INFERRED	NOT RECOMMENDED	NO	NO

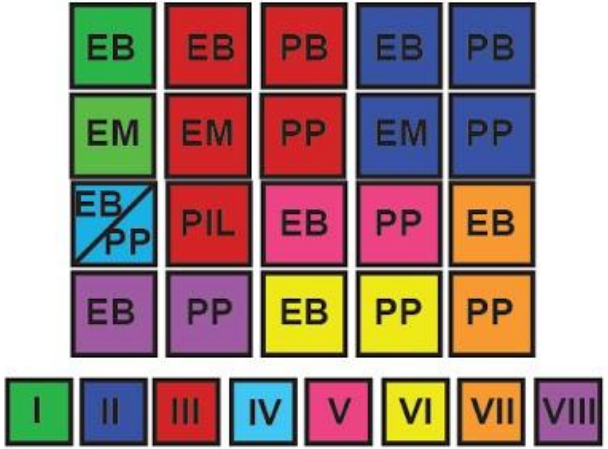
TECHNIQUES FOR GENERATING PHYLOGENOMIC DATA MATRICES

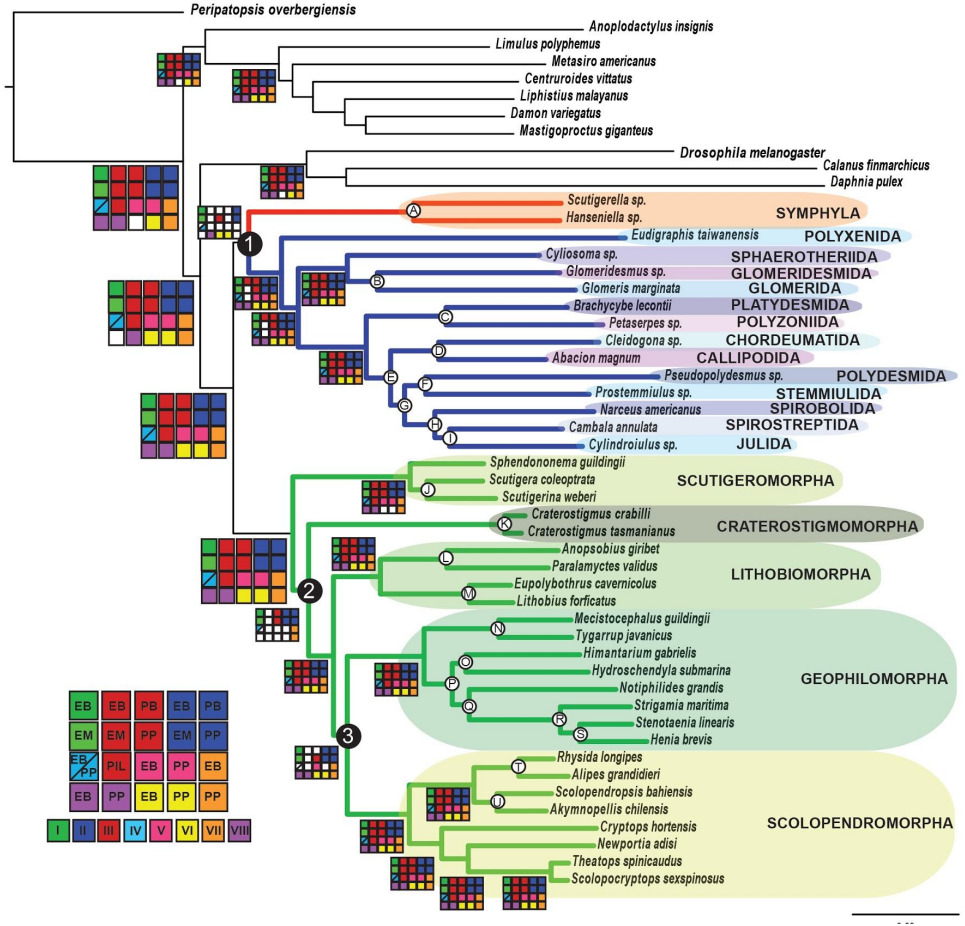
FACTORS	GENOMICS	TRANSCRIPTOMICS	TARGET ENRICHMENT	RADseq
MISSING DATA	LOW	HIGH	HIGH	HIGH
BALANCED LINEAGE REPRESENTATION	LOW	HIGH	HIGH	HIGH (ALLELE DROP OUT)
GENE DUPLICATION	REAL DUPLICATIONS	ISOFORMS	DEPENDS ON PROBE DESIGN	CAN'T BE ASSESSED
HORIZONTAL GENE TRANSFER	CAN BE INFERRED	NOT RECOMMENDED	NO	NO
EVOLUTIONARY RATE	YES	YES	YES	-
COMPOSITIONAL HETEROGENEITY	YES	YES	YES	-
HETEROTACHY	YES	YES	YES	-
LONG BRANCH ATTRACTION (LBA)	YES	YES	YES	-

INHERENT TO THE TECHNIQUE

INHERENT TO THE GENE PROPERTIES

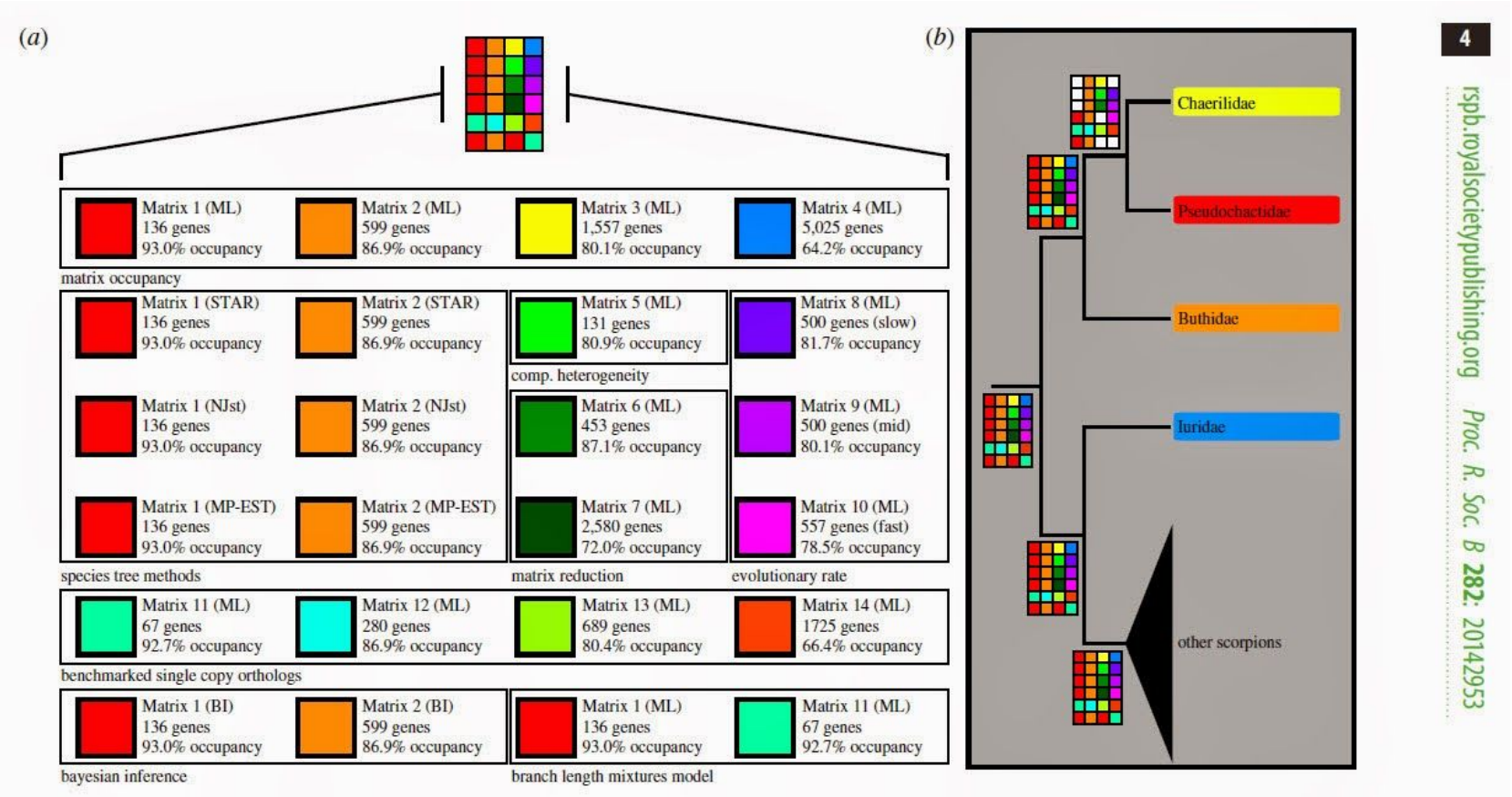
SO… HOW MANY DATA MATRICES SHOULD I ANALYZE?





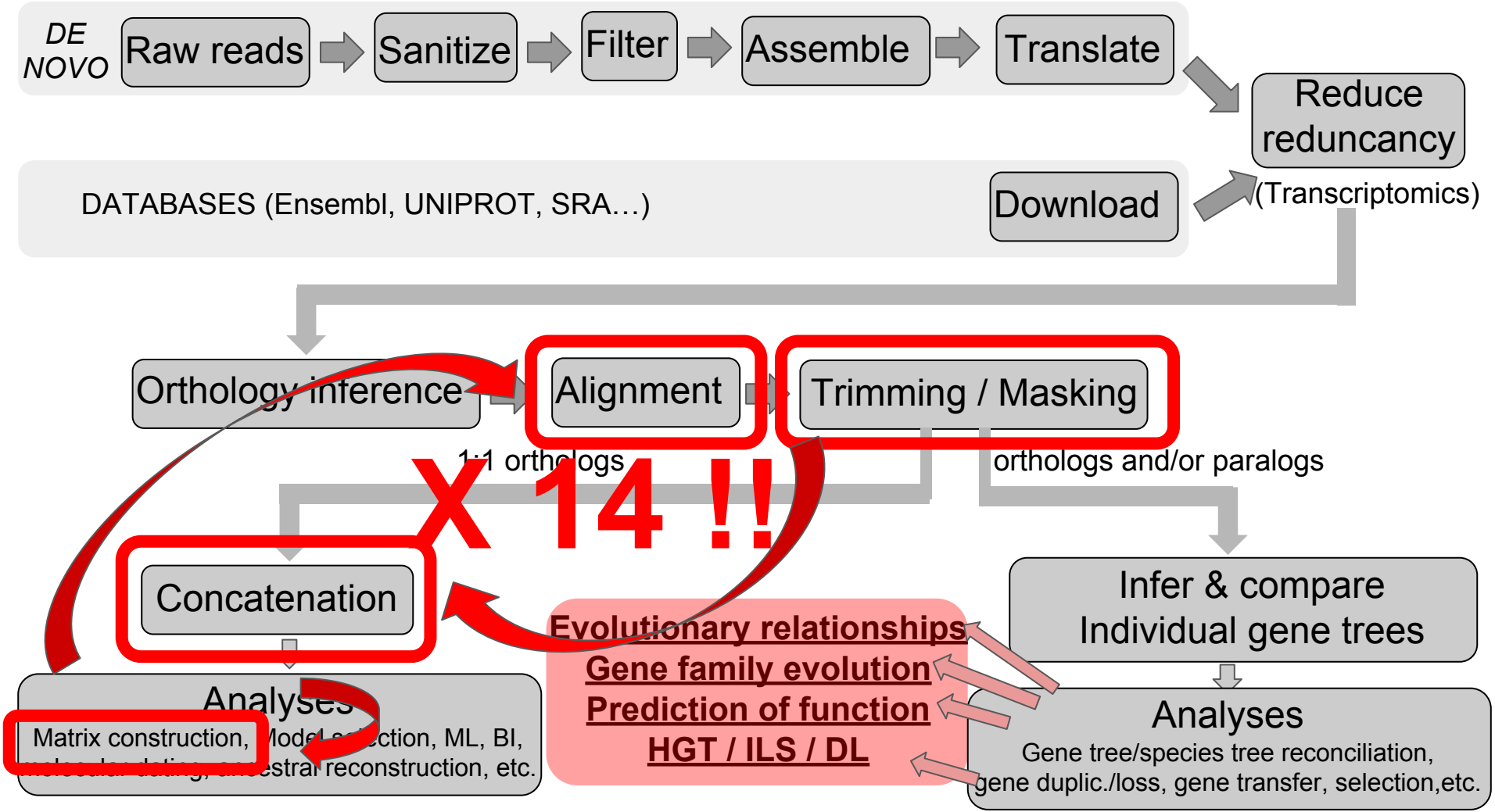
Fernández, Edgecombe & Giribet (2016) Syst Biol

SO… HOW MANY DATA MATRICES SHOULD I ANALYZE?



4
rspb.royalsocietypublishing.org Proc. R. Soc. B 282: 20142953

Sharma, Fernández et al. (2015) Proc Royal Soc B


DE NOVO
Raw reads
Sanitize
Filter
Assemble
Translate
Reduce reduncancy
(Transcriptomics)
DATABASES (Ensembl, UNIPROT, SRA…)
Download
Orthology inference
Alignment
Trimming / Masking
1:1 orthologs
orthologs and/or paralogs
X 14 !!
Concatenation
Evolutionary relationships
Gene family evolution
Prediction of function
HGT / ILS / DL
Infer & compare Individual gene trees
Analyses
Matrix construction, Model selection, ML, BI,
molecular dating, ancestral reconstruction, etc.
Analyses
Gene tree/species tree reconciliation,
gene duplic./loss, gene transfer, selection,etc.

DE NOVO Raw reads → Sanitize → Filter → Assemble → Translate → Reduce reduncancy (Transcriptomics)

DATABASES (Ensembl, UNIPROT, SRA…) → Download → (Transcriptomics)

Orthology inference → Alignment → Trimming / M

1:1 orthologs

ort

X 14 !!

Concatenation

Analyses

Matrix construction, Model selection, ML, BI, molecular dating, ancestral reconstruction, etc.

Evolutionary relationships
Gene family evolution
Prediction of function
HGT / ILS / DL

gene

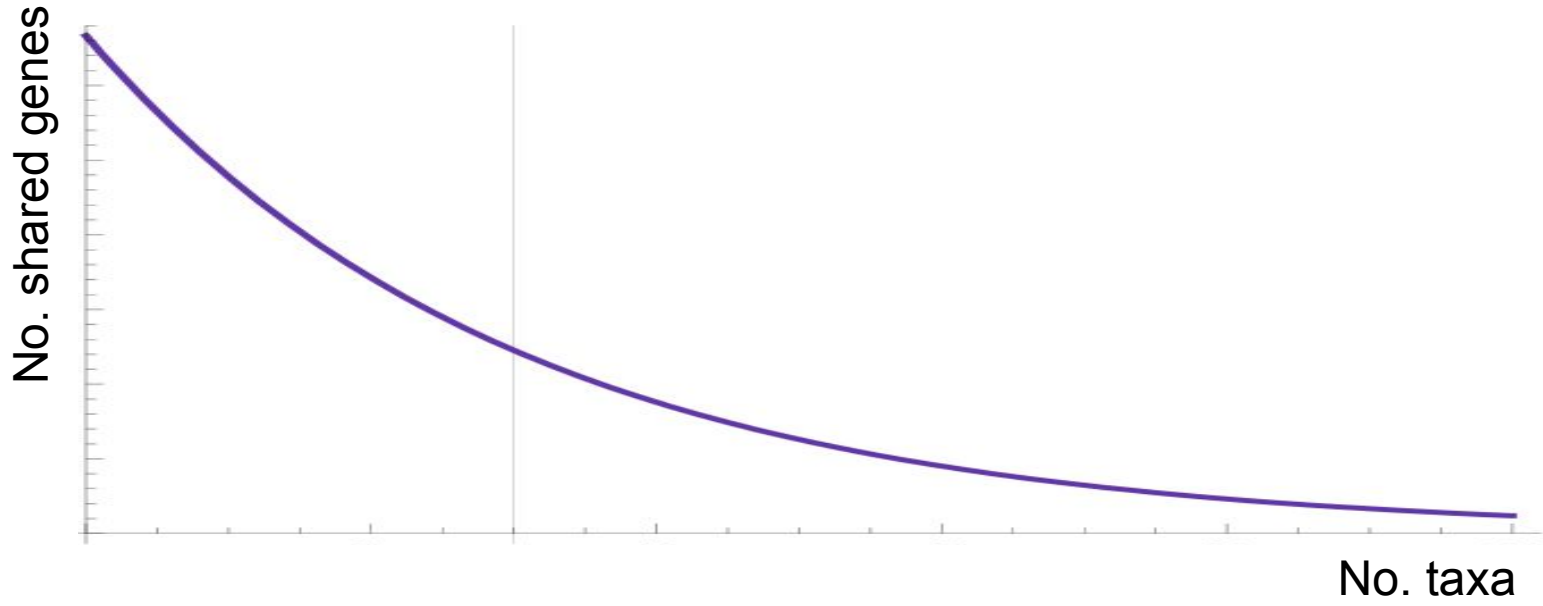
GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

- **MISSING DATA:** depends on the type of data (transcriptomes/target enr. > genomes)

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

- **MISSING DATA:** depends on the type of data (transcriptomes/target enr. > genomes)

Empirical evidence from different datasets in transcriptomics

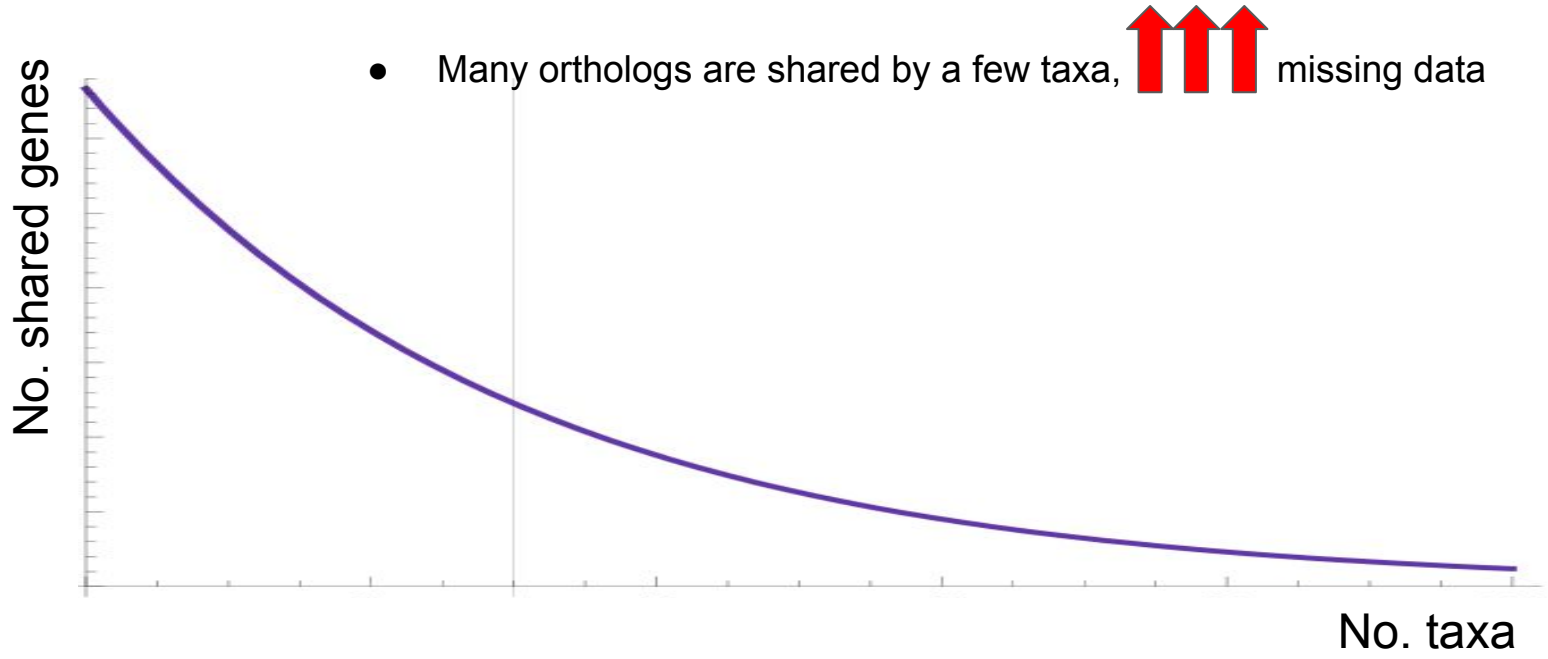


GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

- **MISSING DATA: depends on the type of data (transcriptomes/target enr. > genomes)**

Empirical evidence from different datasets in transcriptomics

- Many orthologs are shared by a few taxa, missing data

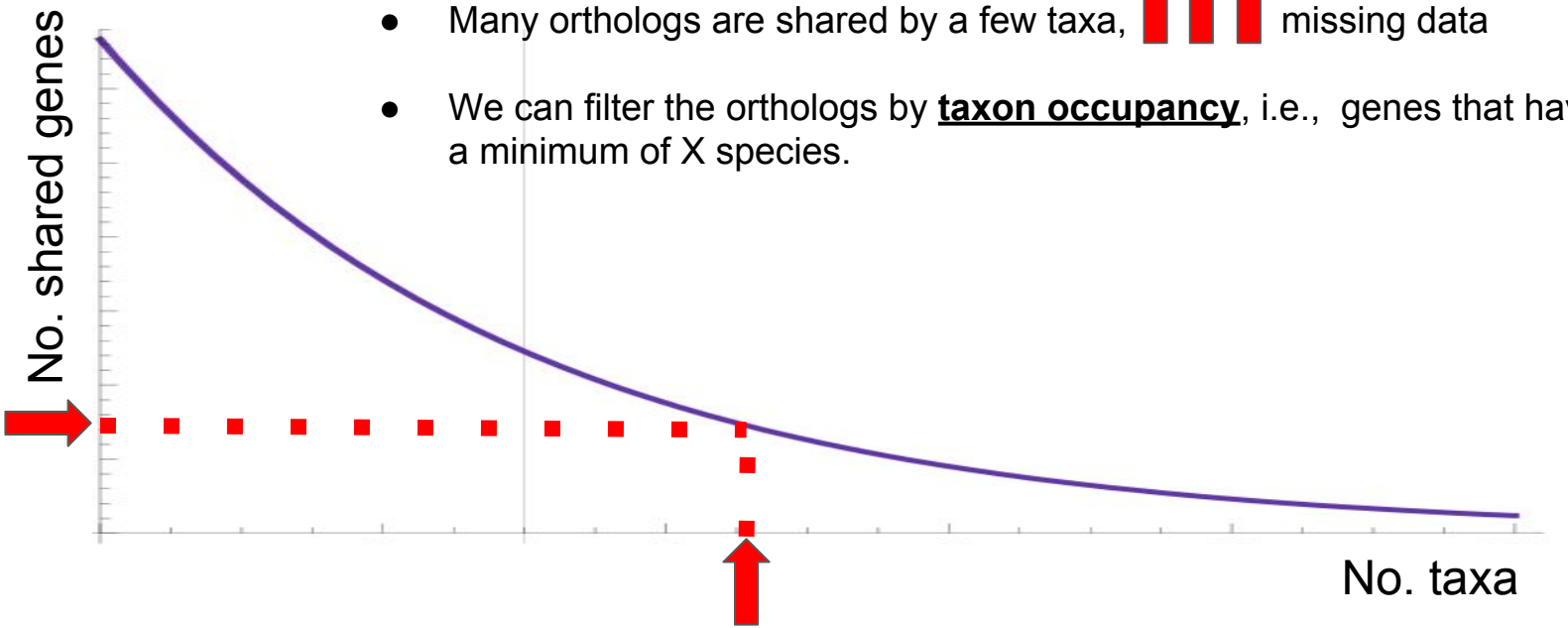


GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

- **MISSING DATA: depends on the type of data (transcriptomes/target enr. > genomes)**

Empirical evidence from different datasets in transcriptomics

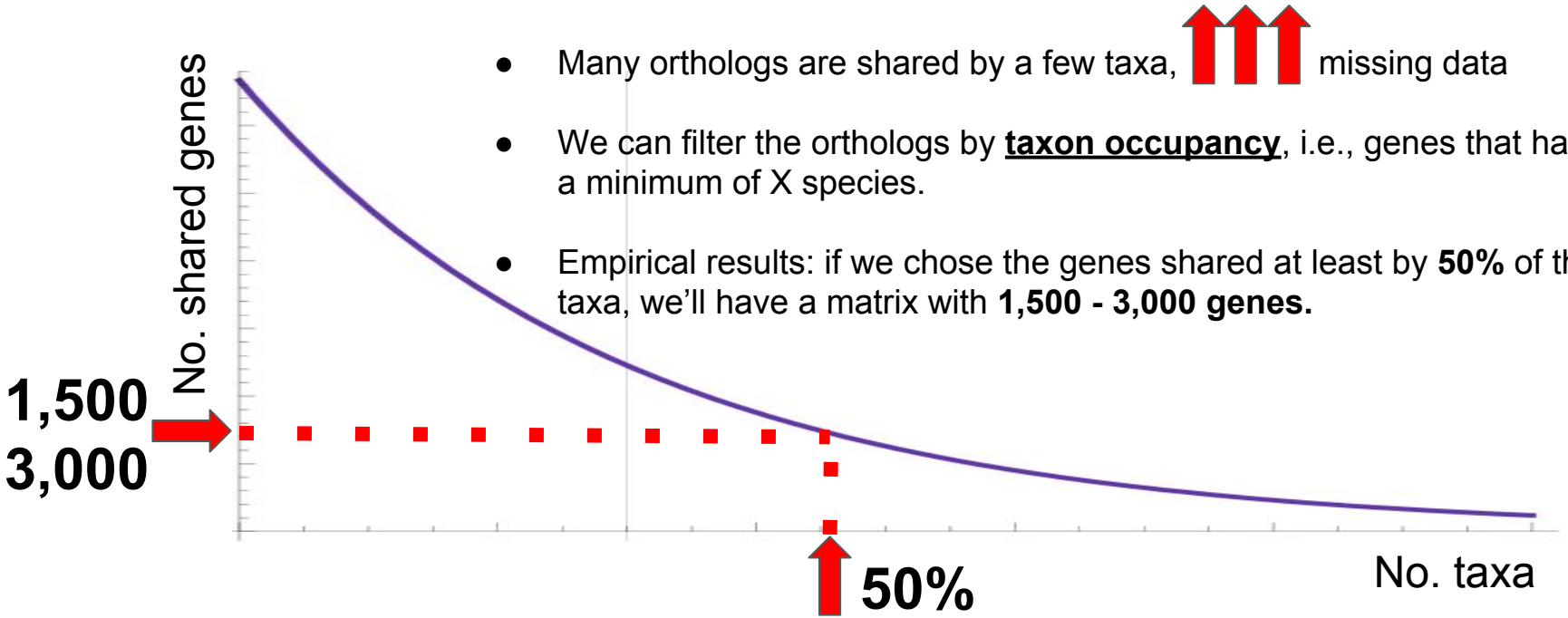
- Many orthologs are shared by a few taxa, ⇑⇑⇑ missing data
- We can filter the orthologs by **taxon occupancy**, i.e., genes that have a minimum of X species.



GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

- **MISSING DATA: depends on the type of data (transcriptomes/target enr. > genomes)**

Empirical evidence from different datasets in transcriptomics

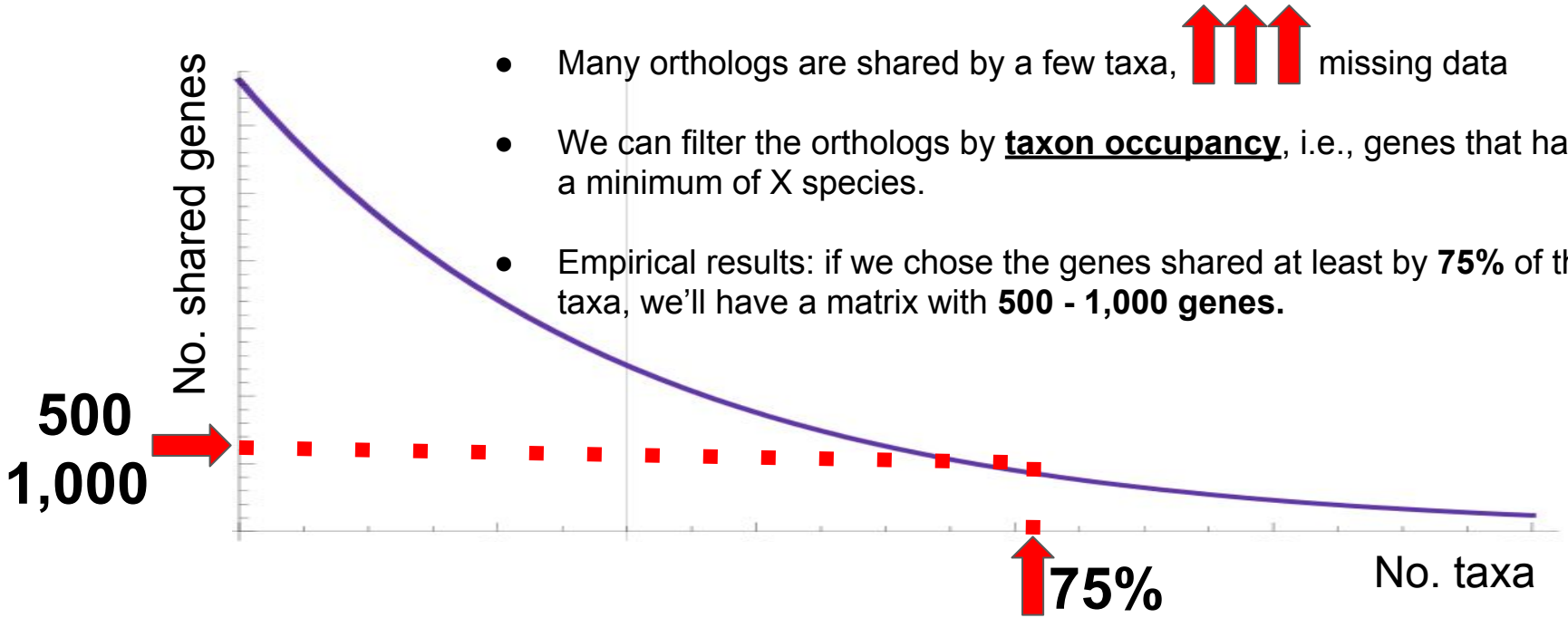


- Many orthologs are shared by a few taxa, missing data
- We can filter the orthologs by **taxon occupancy**, i.e., genes that have a minimum of X species.
- Empirical results: if we chose the genes shared at least by **50%** of the taxa, we'll have a matrix with **1,500 - 3,000 genes.**

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

- **MISSING DATA: depends on the type of data (transcriptomes/target enr. > genomes)**

Empirical evidence from different datasets in transcriptomics



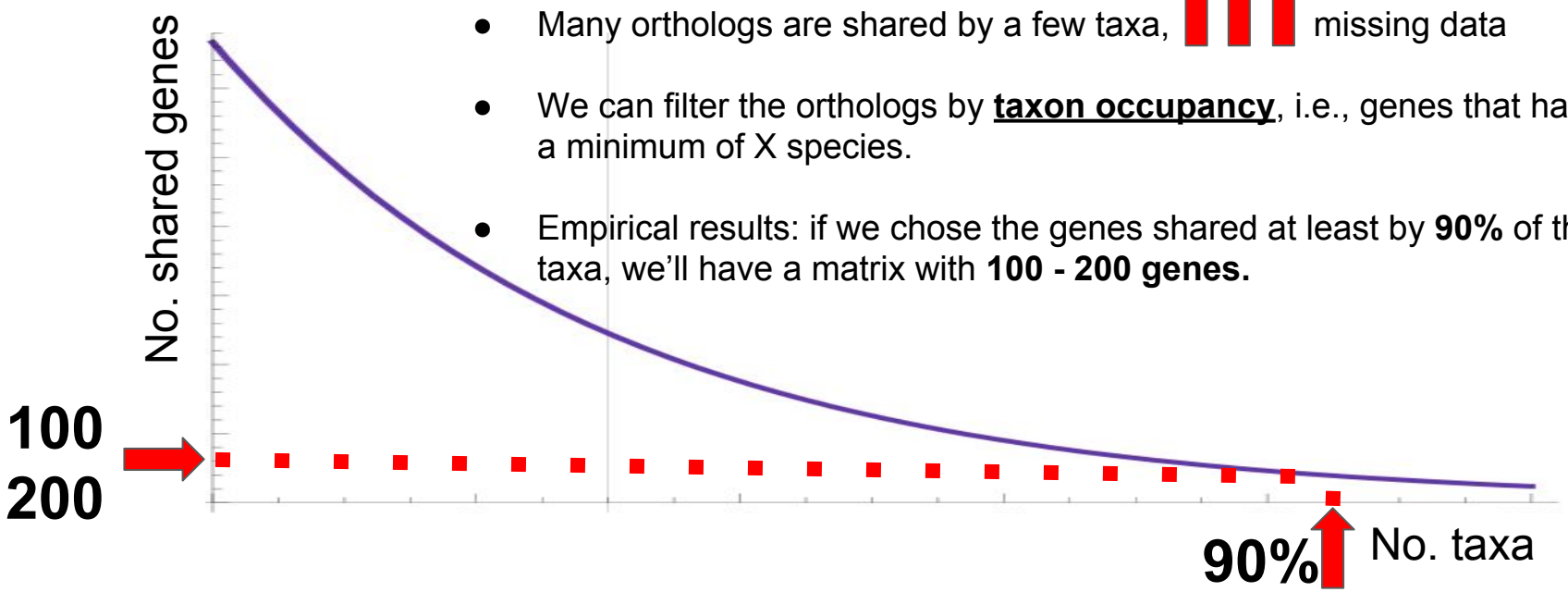
- Many orthologs are shared by a few taxa, ⇑⇑⇑ missing data
- We can filter the orthologs by **taxon occupancy**, i.e., genes that have a minimum of X species.
- Empirical results: if we chose the genes shared at least by **75%** of the taxa, we'll have a matrix with **500 - 1,000 genes.**

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

- **MISSING DATA: depends on the type of data (transcriptomes/target enr. > genomes)**

Empirical evidence from different datasets in transcriptomics

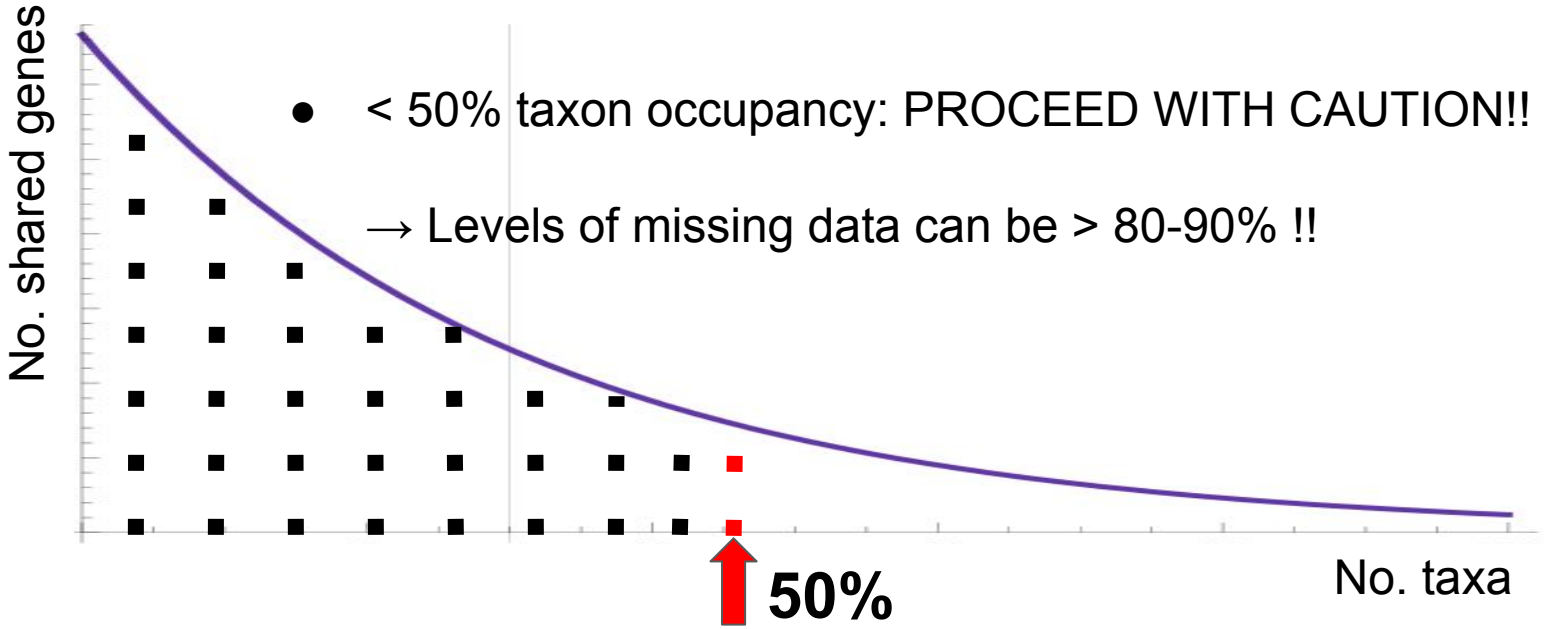
- Many orthologs are shared by a few taxa, ⇑⇑⇑ missing data
- We can filter the orthologs by **taxon occupancy**, i.e., genes that have a minimum of X species.
- Empirical results: if we chose the genes shared at least by **90%** of the taxa, we'll have a matrix with **100 - 200 genes.**



GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

- **MISSING DATA:** depends on the type of data (transcriptomes/target enr. > genomes)

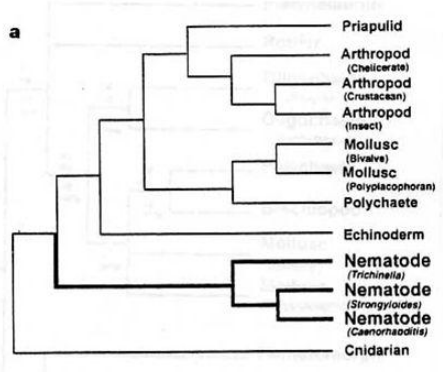
Empirical evidence from different datasets in transcriptomics



GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

- **LONG BRANCH ATTRACTION (LBA)**

Long-Branch Attraction Problem



Some taxa have **fast-evolving DNA**

Often drop out at **base** of tree, clustered with:

(a) basal lineages / outgroups

(b) other fast-evolvers, whom they may not be related to

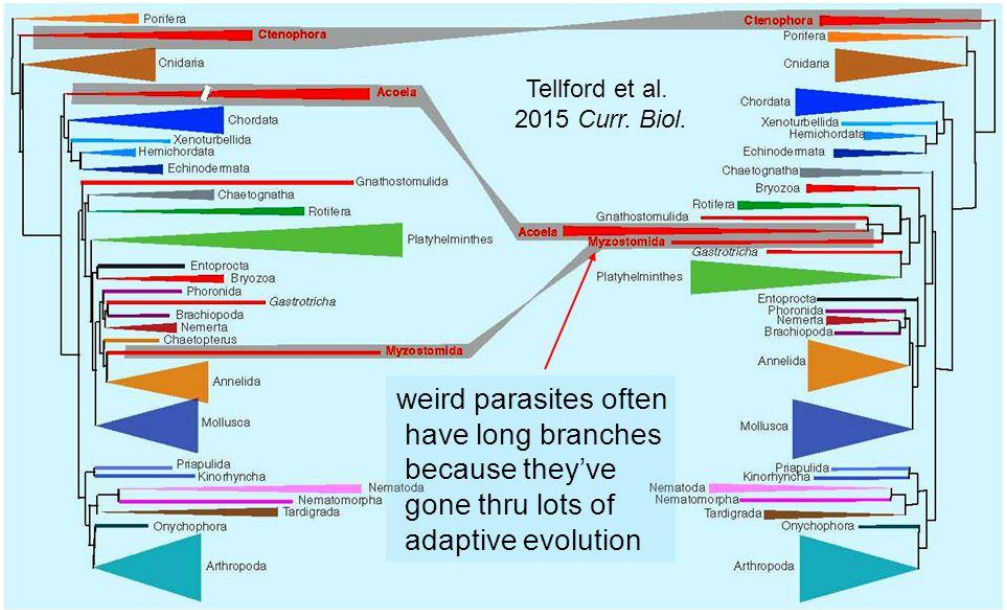
This is an artifact (a false result) of how computer programs analyze DNA sequences, called **long-branch attraction**

- sequences that are fast-evolving give very long branches on trees, which tend to "attract" other long branches

→ sequences that are very different (fast mutating) get lumped together with other fast-evolving sequences

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

- **LONG BRANCH ATTRACTION (LBA)**



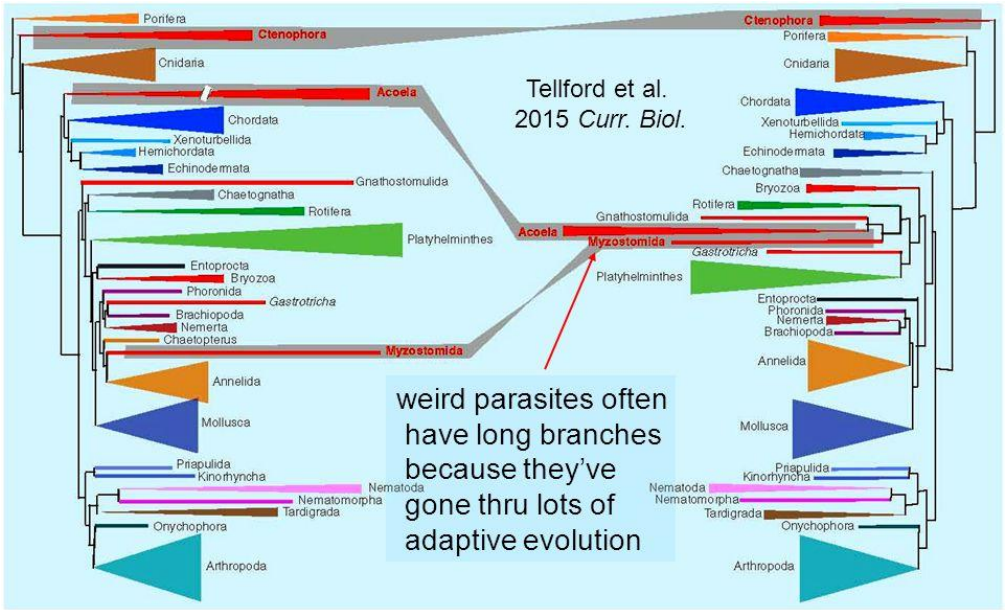
correcting for long-branch attraction

not correcting for long-branch attraction

- **Correcting for LBA:**
  - Compositional heterogeneity

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

- **LONG BRANCH ATTRACTION (LBA)**



correcting for long-branch attraction

not correcting for long-branch attraction

- **Correcting for LBA:**
  - Compositional heterogeneity
  - Evolutionary rate (also related to resolving deep nodes vs shallow ones)

A Genome-Scale Investigation of How Sequence, Function, and Tree-Based Gene Properties Influence Phylogenetic Inference

Xing-Xing Shen[1], Leonidas Salichos[1,2], and Antonis Rokas[1,*]

[1]Department of Biological Sciences, Vanderbilt University

[2]Department of Molecular Biophysics and Biochemistry, Yale University

*Corresponding author: E-mail: antonis.rokas@vanderbilt.edu.

Accepted: July 25, 2016

Shen et al. (2016)
Genome Biol. Evol. 8(8):2565–2580.

A Genome-Scale Investigation of How Sequence, Function, and Tree-Based Gene Properties Influence Phylogenetic Inference

Shen et al. (2016)
Genome Biol. Evol. 8(8):2565–2580.

Xing-Xing Shen[1], Leonidas Salichos[1,2], and Antonis Rokas[1,*]

[1]Department of Biological Sciences, Vanderbilt University

[2]Department of Molecular Biophysics and Biochemistry, Yale University

*Corresponding author: E-mail: antonis.rokas@vanderbilt.edu.

Accepted: July 25, 2016

Abstract

Molecular phylogenetic inference is inherently dependent on choices in both methodology and data. Many insightful studies have shown how choices in methodology, such as the model of sequence evolution or optimality criterion used, can strongly influence inference. In contrast, much less is known about the impact of choices in the properties of the data, typically genes, on phylogenetic inference. We investigated the relationships between 52 gene properties (24 sequence-based, 19 function-based, and 9 tree-based) with each other and with three measures of phylogenetic signal in two assembled data sets of 2,832 yeast and 2,002 mammalian genes. We found that most gene properties, such as evolutionary rate (measured through the percent average of pairwise identity across taxa) and total tree length, were highly correlated with each other. Similarly, several gene properties, such as gene alignment length, Guanine-Cytosine content, and the proportion of tree distance on internal branches divided by relative composition variability (treeness/RCV), were strongly correlated with phylogenetic signal. Analysis of partial correlations between gene properties and phylogenetic signal in which gene evolutionary rate and alignment length were simultaneously controlled, showed similar patterns of correlations, albeit weaker in strength. Examination of the relative importance of each gene property on phylogenetic signal identified gene alignment length, alongside with number of parsimony-informative sites and variable sites, as the most important predictors. Interestingly, the subsets of gene properties that optimally predicted phylogenetic signal differed considerably across our three phylogenetic measures and two data sets; however, gene alignment length and RCV were consistently included as predictors of all three phylogenetic measures in both yeasts and mammals. These results suggest that a handful of sequence-based gene properties are reliable predictors of phylogenetic signal and could be useful in guiding the choice of phylogenetic markers.

Key words: phylogenetic signal, nuclear gene, correlation, prediction, gene function, gene tree.

A Genome-Scale Investigation of How Sequence, Function, and Tree-Based Gene Properties Influence Phylogenetic Inference

Shen et al. (2016)
Genome Biol. Evol. 8(8):2565–2580.

Xing-Xing Shen[1], Leonidas Salichos[1,2], and Antonis Rokas[1,*]

[1]Department of Biological Sciences, Vanderbilt University

[2]Department of Molecular Biophysics and Biochemistry, Yale University

*Corresponding author: E-mail: antonis.rokas@vanderbilt.edu.

Accepted: July 25, 2016

Abstract

Molecular phylogenetic inference is inherently dependent on choices in both methodology and data. Many insightful studies have shown how choices in methodology, such as the model of sequence evolution or optimality criterion used, can strongly influence

across taxa) and total tree length, were highly correlated with each other. Similarly, several gene properties, such as gene alignment length, Guanine-Cytosine content, and the proportion of tree distance on internal branches divided by relative composition variability (treeness/RCV), were strongly correlated with phylogenetic signal. Analysis of partial correlations between gene properties and

across taxa) and total tree length, were highly correlated with each other. Similarly, several gene properties, such as gene alignment length, Guanine-Cytosine content, and the proportion of tree distance on internal branches divided by relative composition variability (treeness/RCV), were strongly correlated with phylogenetic signal. Analysis of partial correlations between gene properties and phylogenetic signal in which gene evolutionary rate and alignment length were simultaneously controlled, showed similar patterns of correlations, albeit weaker in strength. Examination of the relative importance of each gene property on phylogenetic signal identified gene alignment length, alongside with number of parsimony-informative sites and variable sites, as the most important predictors. Interestingly, the subsets of gene properties that optimally predicted phylogenetic signal differed considerably across our three phylogenetic measures and two data sets; however, gene alignment length and RCV were consistently included as predictors of all three phylogenetic measures in both yeasts and mammals. These results suggest that a handful of sequence-based gene properties are reliable predictors of phylogenetic signal and could be useful in guiding the choice of phylogenetic markers.

Key words: phylogenetic signal, nuclear gene, correlation, prediction, gene function, gene tree.

A Genome-Scale Investigation of How Sequence, Function, and Tree-Based Gene Properties Influence Phylogenetic Inference

Shen et al. (2016)
Genome Biol. Evol. 8(8):2565–2580.

Xing-Xing Shen[1], Leonidas Salichos[1,2], and Antonis Rokas[1,*]

[1]Department of Biological Sciences, Vanderbilt University

[2]Department of Molecular Biophysics and Biochemistry, Yale University

*Corresponding author: E-mail: antonis.rokas@vanderbilt.edu.

Accepted: July 25, 2016

Abstract

Molecular phylogenetic inference is inherently dependent on choices in both methodology and data. Many insightful studies have shown how choices in methodology, such as the model of sequence evolution or optimality criterion used, can strongly influence across taxa) and total tree length, were highly correlated with each other. Similarly, several gene properties, such as gene alignment length, Guanine-Cytosine content, and the proportion of tree distance on internal branches divided by relative composition variability (treeness/RCV), were strongly correlated with phylogenetic signal. Analysis of partial correlations between gene properties and of correlations, albeit weaker in strength. Examination of the relative importance of each gene property on phylogenetic signal identified gene alignment length, alongside with number of parsimony-informative sites and variable sites, as the most important predictors. Interestingly, the subsets of gene properties that optimally predicted phylogenetic signal differed considerably across our three phylogenetic measures and two data sets; however, gene alignment length and RCV were consistently included as predictors of all three phylogenetic measures in both yeasts and mammals. These results suggest that a handful of sequence-based gene properties are reliable predictors of phylogenetic signal and could be useful in guiding the choice of phylogenetic markers.

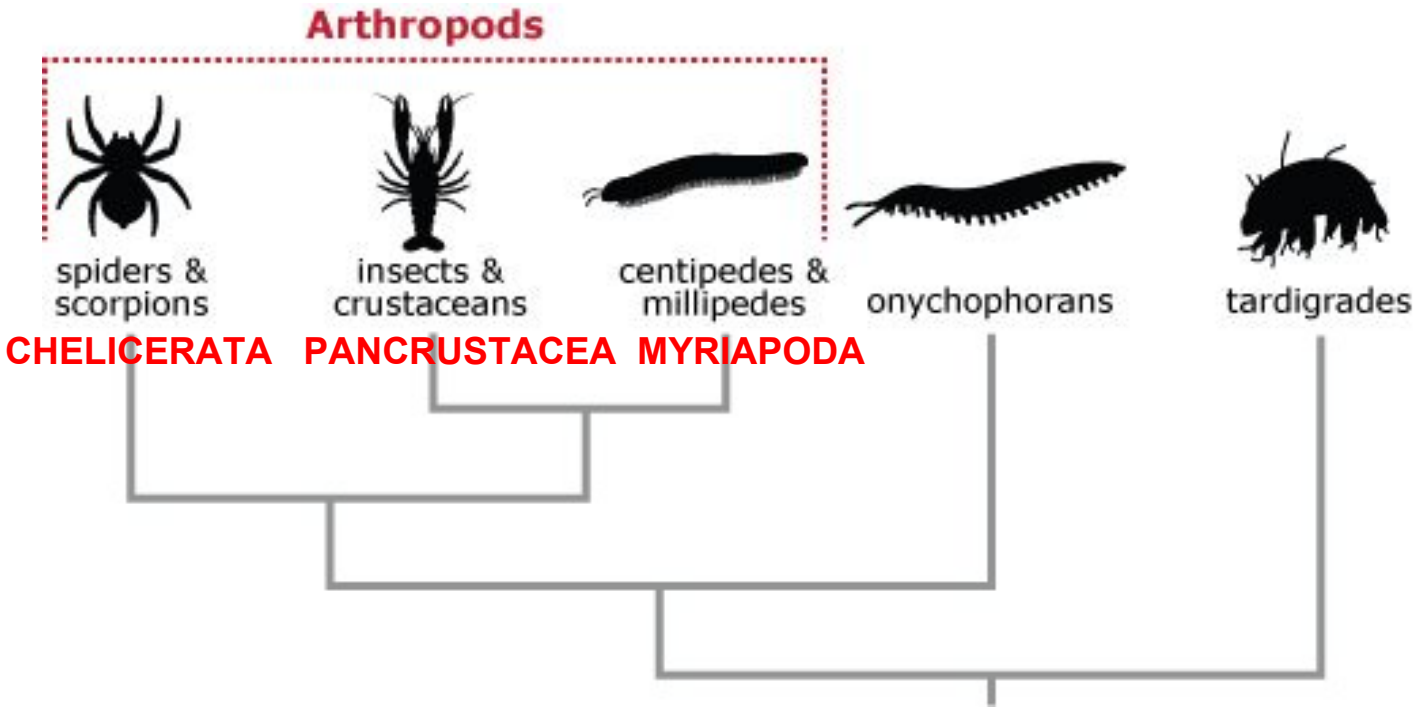
Key words: phylogenetic signal, nuclear gene, correlation, prediction, gene function, gene tree.

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life

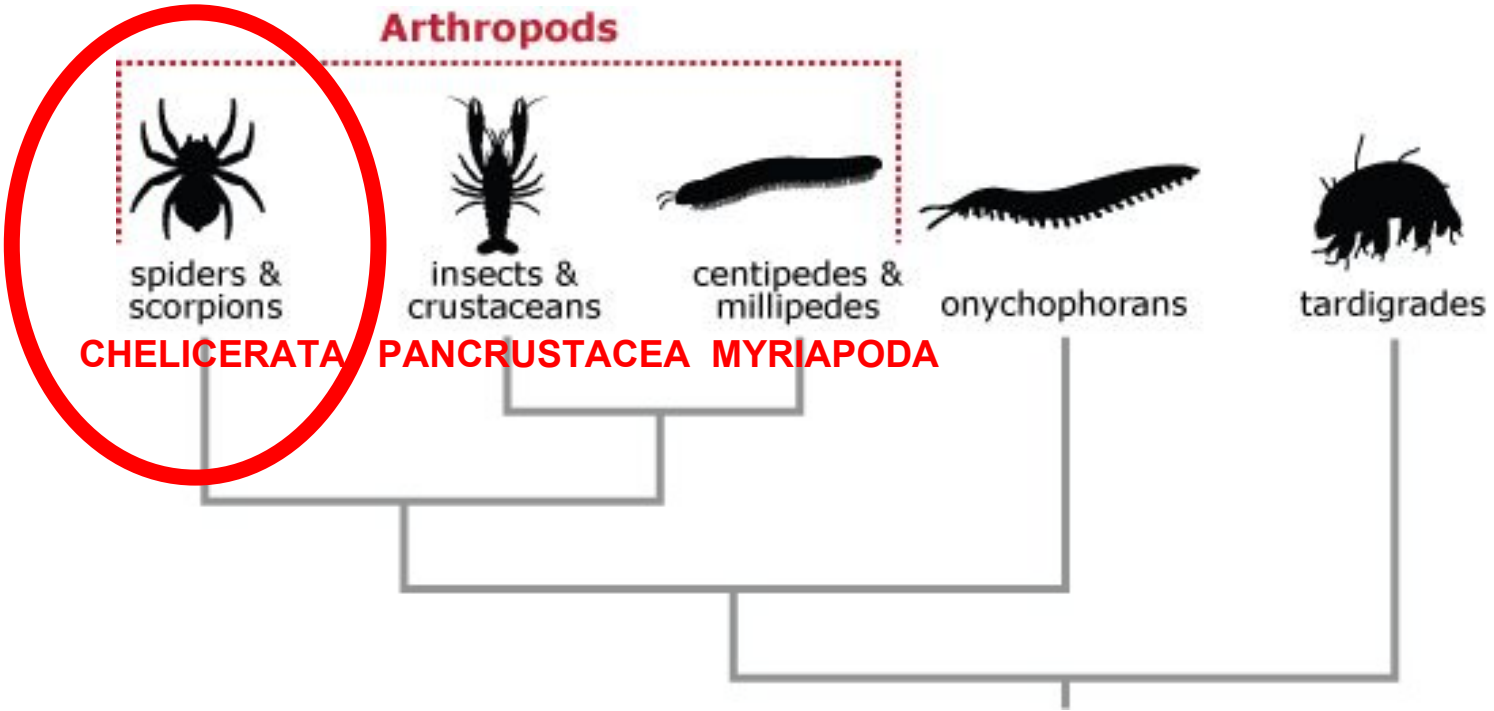
GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life



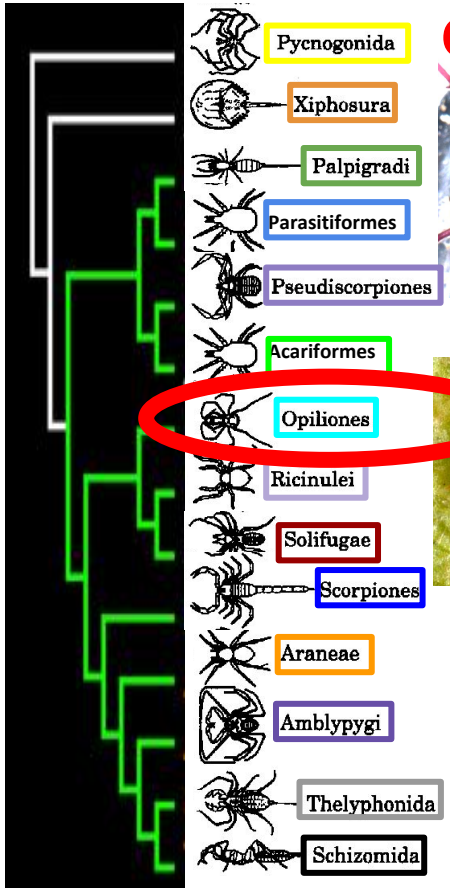
GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life



GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case Study: The Opiliones Tree of Life



sea spiders

horseshoe crabs

microwhip scorpions

ticks

false scorpions

mites

harvestmen

hooded tickspiders

camel spiders

scorpions

spiders

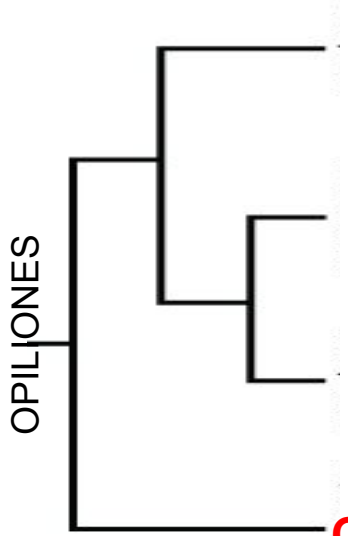
whip spiders

whip scorpions

shorttailed whipscorpions

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

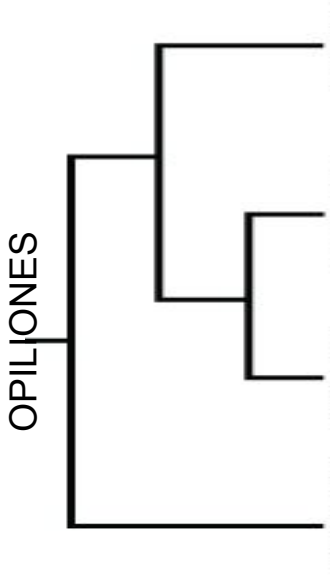
Case study: The Opiliones Tree of Life



CYPHOPHTHALMI
(mite harvestmen)

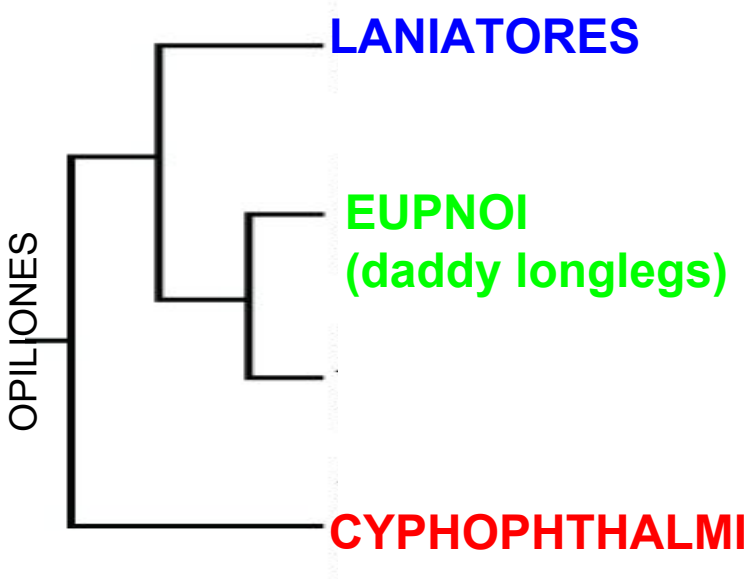
GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life



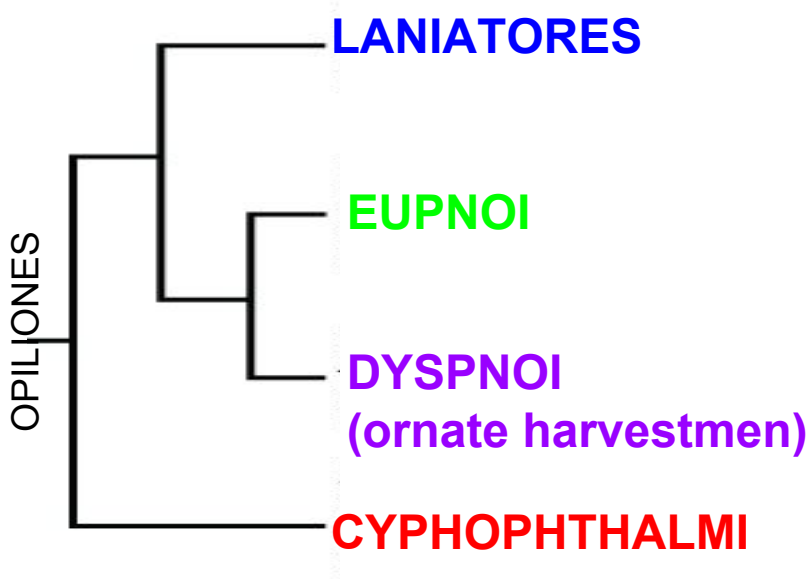
GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life



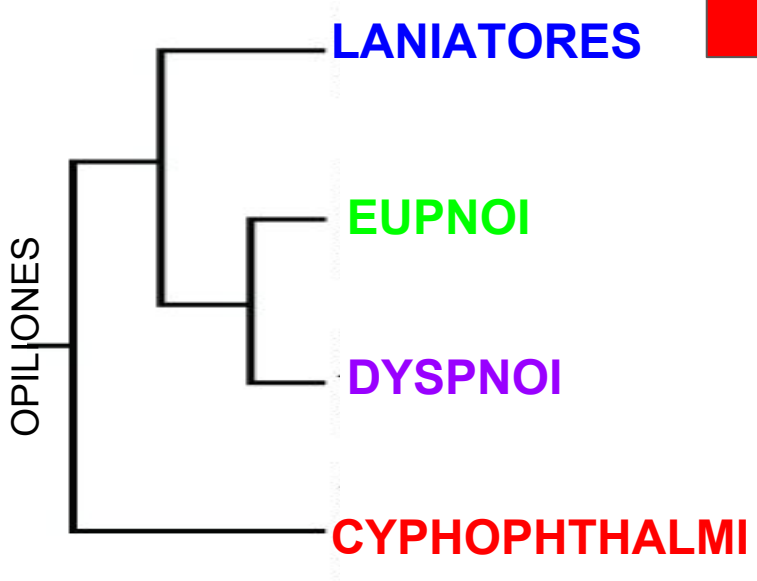
GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life



GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

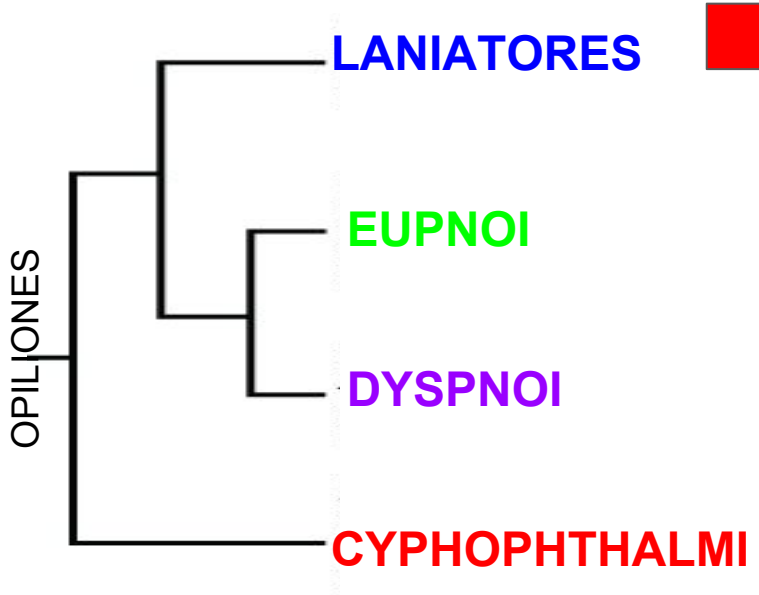
Case study: The Opiliones Tree of Life



FIELD EXPEDITION TO EXOTIC PLACES

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life

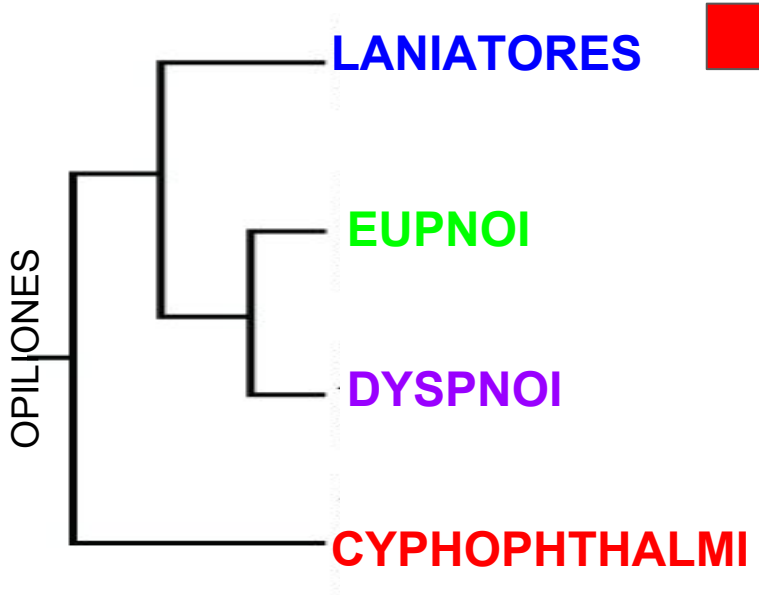


FIELD EXPEDITION TO EXOTIC PLACES

SEQUENCING

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life



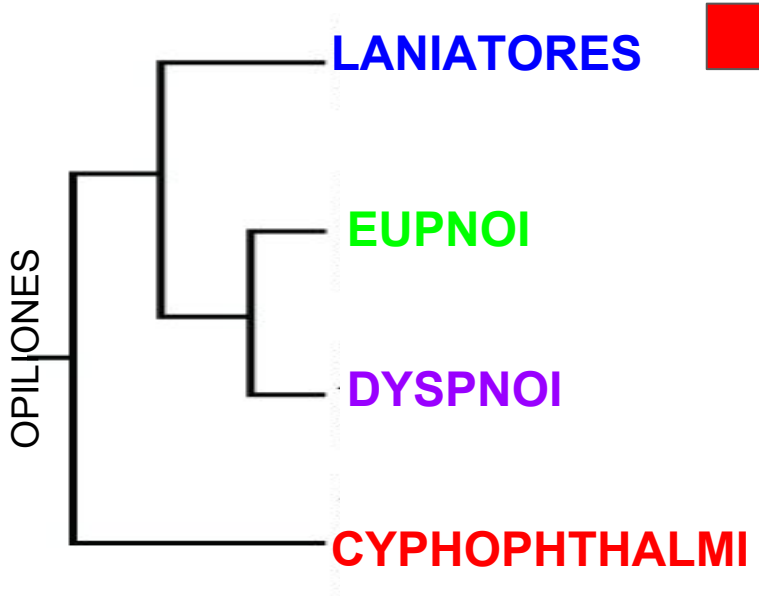
FIELD EXPEDITION TO EXOTIC PLACES

SEQUENCING

ASSEMBLY

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life



FIELD EXPEDITION TO EXOTIC PLACES

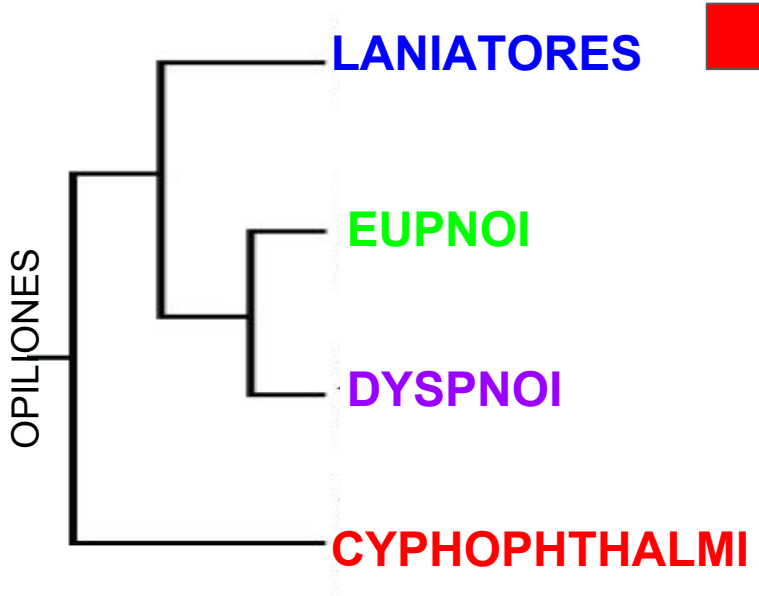
SEQUENCING

ASSEMBLY

ORTHOLOGY INFERENCE

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life



FIELD EXPEDITION TO EXOTIC PLACES

SEQUENCING

ASSEMBLY

ORTHOLOGY INFERENCE

READY TO CREATE A MATRIX… BUT… WHICH GENES SHOULD WE INCLUDE??

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life

- **MISSING DATA**

Let’s create different matrices with different taxon occupancy to account for the effect of missing data.

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life

- **MISSING DATA**

Let's create different matrices with different taxon occupancy to account for the effect of missing data.

1) Download the folder **lab_matrices_Cesky** from the webpage of the workshop (click on 'Transcriptomics'). Extract the files. In the folder, you'll see a subset of 1:1 orthologs from Fernández et al. (2017) Proc Royal Soc B. (1,508 fasta files, already aligned) and 3 python scripts.

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life

- **MISSING DATA**

Let's create different matrices with different taxon occupancy to account for the effect of missing data.

1) Download the folder **lab_matrices_Cesky** from the webpage of the workshop (click on 'Transcriptomics'). Extract the files. In the folder, you'll see a subset of 1:1 orthologs from Fernández et al. (2017) Proc Royal Soc B. (1,508 fasta files, already aligned) and 3 python scripts.

2) We'll need to install a couple of python libraries that are not in the instance (NumPy and PyCogent). For that, open the terminal and type:

   pip install numpy

   pip install cogent

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life

- **MISSING DATA**

3) Let's first explore the amount of missing data that we have in each taxon. Let's run the script:

   python count_species.py

   Explore the amount of missing data in each taxon. Which taxa are poorly represented?

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life

- **MISSING DATA**

3) Let's first explore the amount of missing data that we have in each taxon. Let's run the script:

python count_species.py

Explore the amount of missing data in each taxon.

Which taxa are poorly represented?

You can decide on a minimum percentage of genes that you'd like the taxa to have, and delete the rest (it should be easy to do in any text editor). Try this several times with different percentages, and see if the support for some clades changes. Does the inclusion of poorly represented taxa affect the support of the clades where they fall?

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life

- **MISSING DATA**

4) Now select the genes that have a taxon occupancy above a certain threshold (ie, we want to create a matrix only with the genes that have a minimum of, let's say, 50 species). Open the script and have a look at it. Then run:

python select_slide.py

It will ask you to select the minimum taxon occupancy. Let's start by 50. It will create a folder called 'orthologs_min_[number]_taxa'. Open it and check how many genes were selected with this threshold.

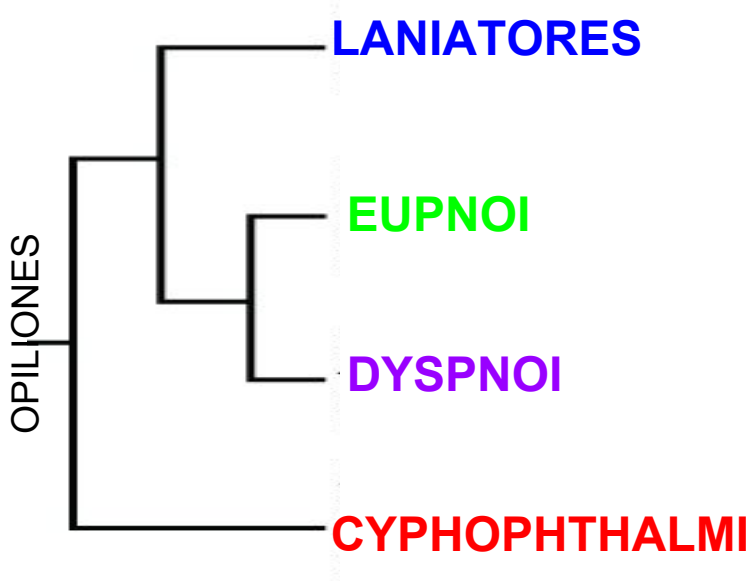
Run the script with different thresholds and check how the number of genes varies.

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life

- **MISSING DATA**

5) Do you remember that Opiliones had four main groups: Cyphophthalmi, Eupnoi, Laniatores and Dyspnoi?

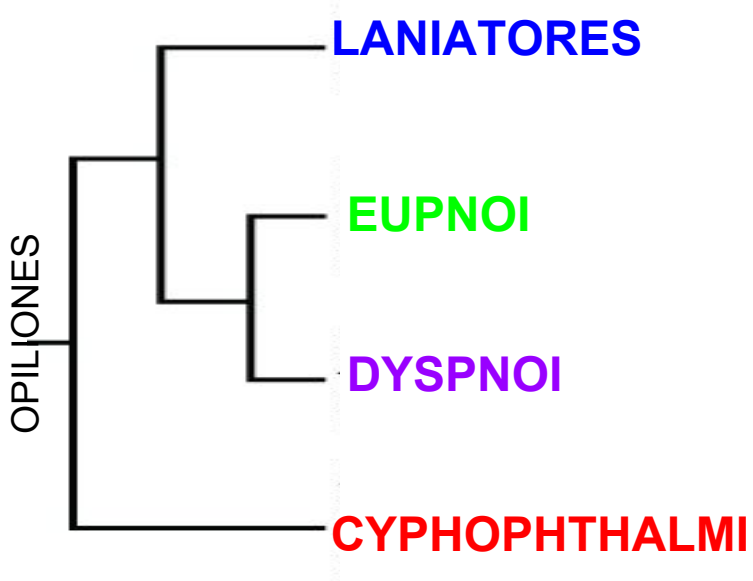


GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life

- **MISSING DATA**

5) Do you remember that Opiliones had four main groups: Cyphophthalmi, Eupnoi, Laniatores and Dyspnoi?



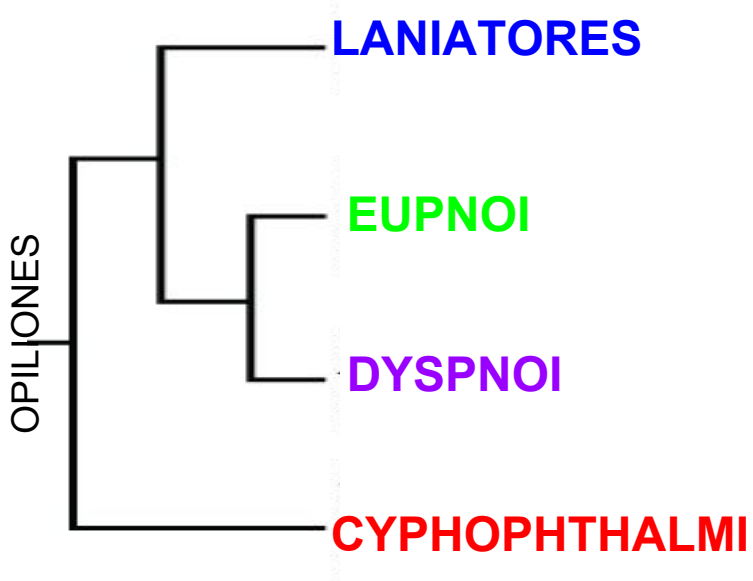
If we select genes just based on taxon occupancy, we may select some that do not include representatives of one or more of these lineages.

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life

- **MISSING DATA**

5) Do you remember that Opiliones had four main groups: Cyphophthalmi, Eupnoi, Laniatores and Dyspnoi?



If we select genes just based on taxon occupancy, we may select some that do not include representatives of one or more of these lineages.

6) Let's try to select genes that have an homogeneous representation of all our lineages of interest. Let's open the **decisive_genes.py** script and inspect it together.

Notice that at the end of the script we're defining our four lineages and choosing a minimum number of taxa representing each lineage in the genes that will be selected (4 in this case). Run the script: **python decisive_genes.py**

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life

- **MISSING DATA**

7) We now have 2 folders called **‘Decisive’** and **‘Non Decisive’**. Check how many genes you have in the **‘Decisive’** one. Change the threshold in the script, rerun it and check how the selected (=decisive) genes change. Do you think this may affect phylogenetic relationships?

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

Case study: The Opiliones Tree of Life

- **MISSING DATA**

7) We now have 2 folders called **‘Decisive’** and **‘Non Decisive’**. Check how many genes you have in the **‘Decisive’** one. Change the threshold in the script, rerun it and check how the selected (=decisive) genes change. Do you think this may affect phylogenetic relationships?

8) Now (or in the open labs) you can play with these scripts to create different matrices, run some trees and see how the topology and the support for each node/lineage changes.

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

- **COMPOSITIONAL HETEROGENEITY**

1) We can check compositional heterogeneity at three levels: at the level of **gene**, at the level of **taxon** or at the level of **site**.

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

- **COMPOSITIONAL HETEROGENEITY**

1) We can check compositional heterogeneity at three levels: at the level of **gene**, at the level of **taxon** or at the level of **site**.

2) Let's explore compositional bias at the level of **site.** We're going to use a program called BMGE (Block Mapping and Gathering with Entropy).

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

- ## COMPOSITIONAL HETEROGENEITY

1) We can check compositional heterogeneity at three levels: at the level of **gene**, at the level of **taxon** or at the level of **site**.

2) Let's explore compositional bias at the level of **site.** We're going to use a program called BMGE (Block Mapping and Gathering with Entropy).

3) As it is not in the instance, let's download it from here. Let's extract it by double-clicking on it (or from the terminal with **tar -xcvf BMGE-1.12.tar.gz**).

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

- ## COMPOSITIONAL HETEROGENEITY

1) We can check compositional heterogeneity at three levels: at the level of **gene**, at the level of **taxon** or at the level of **site**.

2) Let's explore compositional bias at the level of **site.** We're going to use a program called BMGE (Block Mapping and Gathering with Entropy).

3) As it is not in the instance, let's download it from here. Let's extract it by double-clicking on it (or from the terminal with **tar -xcvf BMGE-1.12.tar.gz**).

4) We can do different types of trimming with BMGE. You can check the user guide here. Today we're going to focus on detecting and trimming compositionally-heterogeneous positions. In the terminal, type:

java -jar BMGE.jar -i examples/prmA.pam150.phy -t AA -s YES -on prmABMGE.nex

What's the difference in size before and after trimming compositionally-heterogeneous sites?

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

- **ADVANCED EXERCISES**

1) You can check compositional heterogeneity at the level of gene and taxon with BaCoCa. You can download it from here and play with the toy data set. You'll also need to download the Statistics::R module.

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

- **ADVANCED EXERCISES**

1) You can check compositional heterogeneity at the level of gene and taxon with BaCoCa. You can download it from here and play with the toy data set. You'll also need to download the Statistics::R module.

2) You can use TIGER to bin the sites in your matrix depending on their evolutionary rate, and then create submatrices eliminating the fast evolving ones, the slowest evolving ones, etc. (note that they detected a bug and are working on it. The following version of the software will be called bioTIGER but they only have a beta version of it so far).

GENERATING PHYLOGENOMIC DATA MATRICES IN TRANSCRIPTOMICS

- **ADVANCED EXERCISES**

1) You can check compositional heterogeneity at the level of gene and taxon with BaCoCa. You can download it from here and play with the toy data set. You'll also need to download the Statistics::R module.

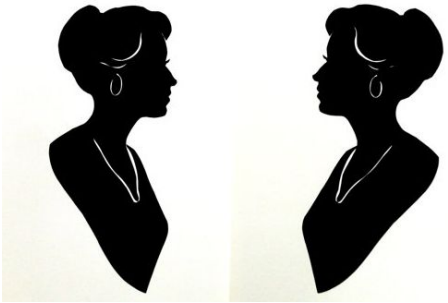
2) You can use TIGER to bin the sites in your matrix depending on their evolutionary rate, and then create submatrices eliminating the fast evolving ones, the slowest evolving ones, etc. (note that they detected a bug and are working on it. The following version of the software will be called bioTIGER but they only have a beta version of it so far).

3) With TreSpEx (from the same developers that created BaCoCa) you can do a bunch of analyses to detect misleading signal in phylogenomic inference, such as calculating saturation indices, genes prone to LBA artifacts, paralogy detection, etc.

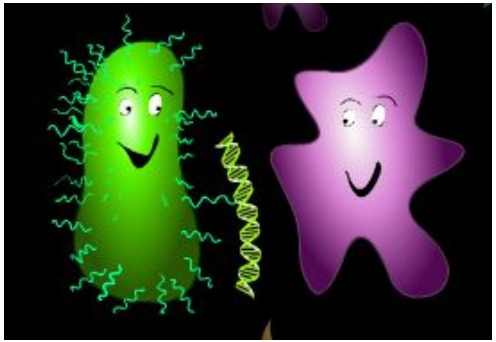
Exercise for the break...

Exercise for the break...

GENERATING PHYLOGENOMIC DATA MATRICES IN GENOMICS



Eukaryotes - Gene duplication

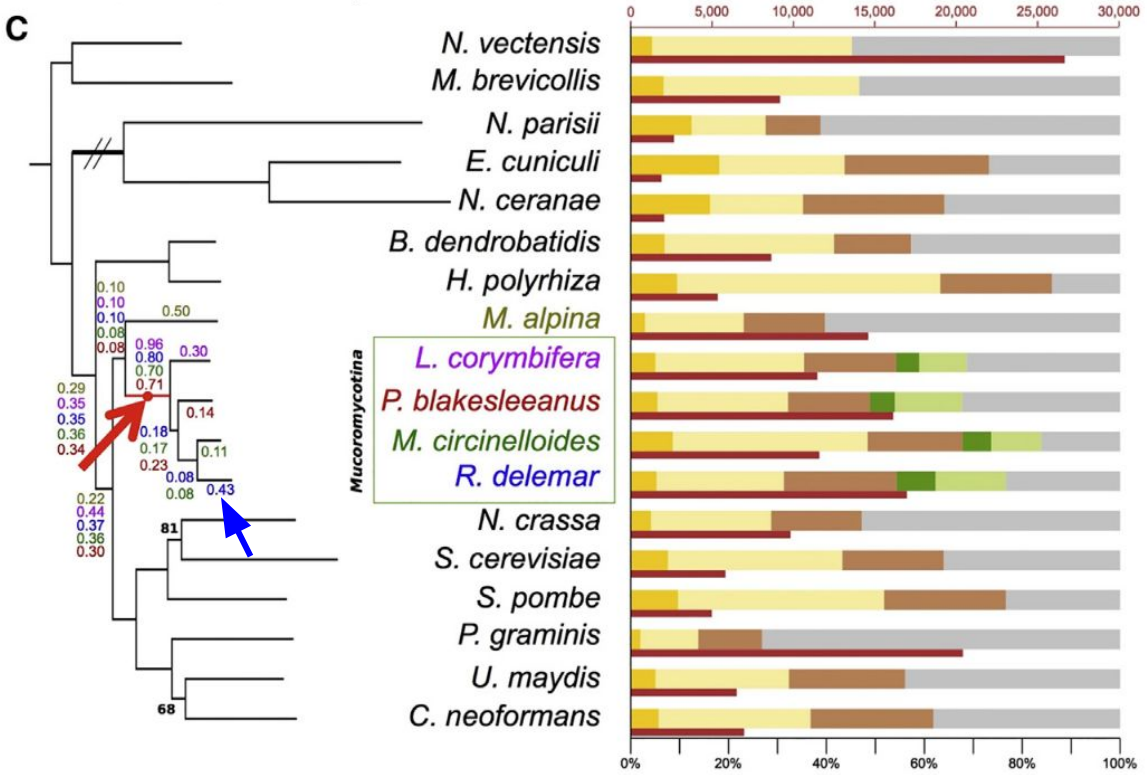


Prokaryotes - Horizontal gene transfer

To create a data matrix you need one-to-one orthology relationships between all the genes that are going to form the matrix. The reconstruction methods are very sensitive to the inclusion of paralogs and of genes that have undergone HGT processes.

Expansion of Signal Transduction Pathways in Fungi by Extensive Genome Duplication

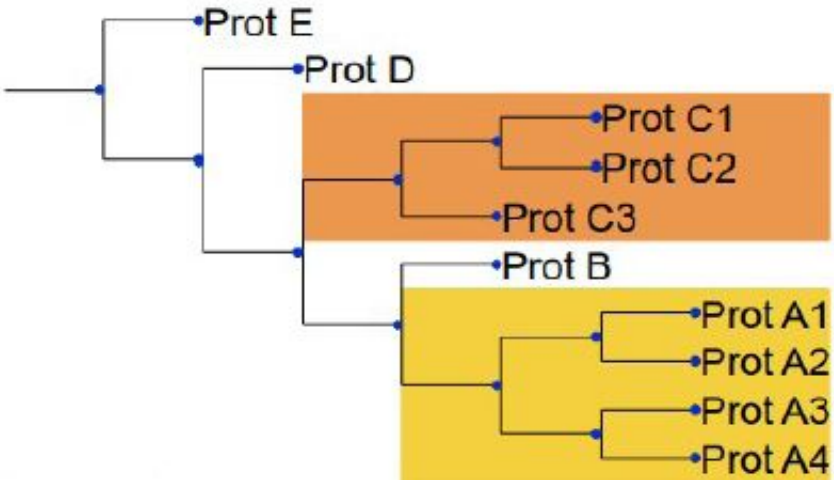
Luis M. Corrochano,[1,*] Alan Kuo,[2] Marina Marcet-Houben,[3,4] Silvia Polaino,[5] Asaf Salamov,[2]
José M. Villalobos-Escobedo,[6] Jane Grimwood,[2,7] M. Isabel Álvarez,[8] Javier Avalos,[1] Diane Bauer,[2] Ernesto P. Benito,[8,9]
Isabelle Benoit,[10] Gertraud Burger,[11] Lola P. Camino,[1] David Cánovas,[1] Enrique Cerdá-Olmedo,[1] Jan-Fang Cheng,[2]
Angel Domínguez,[8] Marek Eliás,[12] Arturo P. Eslava,[8] Fabian Glaser,[13] Gabriel Gutiérrez,[1] Joseph Heitman,[14]
Bernard Henrissat,[15,16] Enrique A. Iturriaga,[8] B. Franz Lang,[11] José L. Lavín,[17] Soo Chan Lee,[14] Wenjun Li,[14]
Erika Lindquist,[2] Sergio López-García,[18] Eva M. Luque,[1] Ana T. Marcos,[1] Joel Martin,[2] Kevin McCluskey,[19]

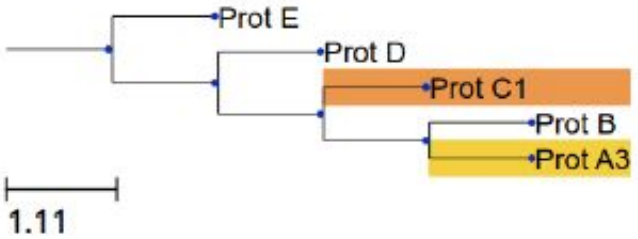


Number of genes you can select to reconstruct your matrix according to species.
Less species means we'll introduce missing data

Number of species	Strict one-to-one		
18	11		
17	27		
16	50		
15	92		
14	165		
13	243		
12	311		
11	384		

Tricks to expand your dataset to reconstruct a data matrix: delete species specific expansions





Number of genes you can select to reconstruct your matrix according to species.
Less species means we'll introduce missing data

Number of species	Strict one-to-one	Delete species specific duplications	
18	11	54	
17	27	117	
16	50	183	
15	92	285	
14	165	441	
13	243	599	
12	311	725	
11	384	857	

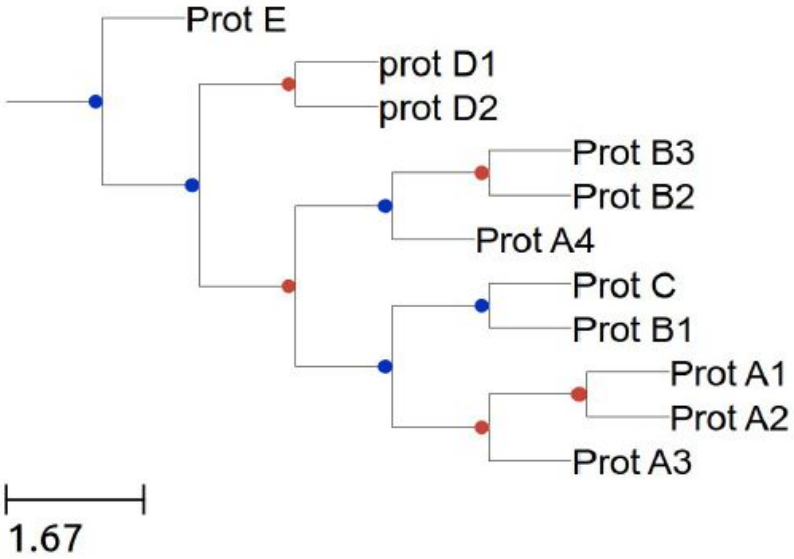
TreeKO: a duplication-aware algorithm for the comparison of phylogenetic trees

Marina Marcet-Houben and Toni Gabaldón*

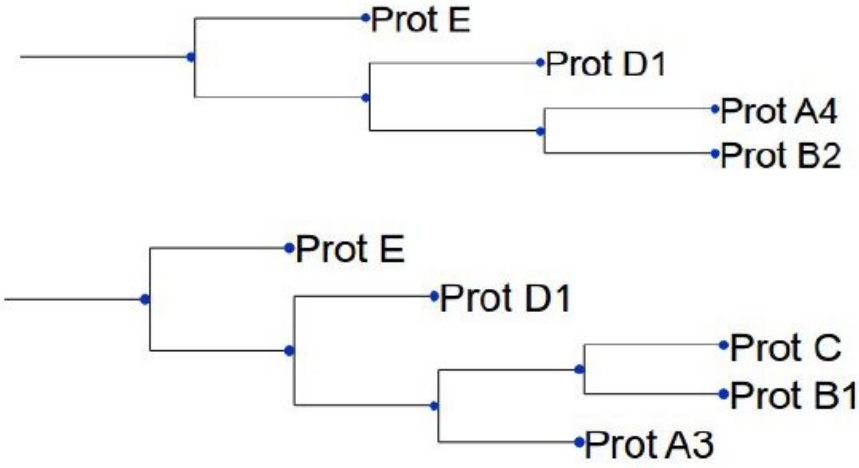
ete-compare

calculate distances and compare trees

Phylogenetic tree that contains duplications → Obtain the list of orthologous trees



Gene tree with duplications



Obtain orthologous trees and keep only those that contain the seed protein

Number of genes you can select to reconstruct your matrix according to species.
Less species means we'll introduce missing data

Number of species	Strict one-to-one	Delete species specific duplications	Obtain orthologous trees
18	11	54	198
17	27	117	444
16	50	183	687
15	92	285	1122
14	165	441	1633
13	243	599	2176
12	311	725	2662
11	384	857	3197

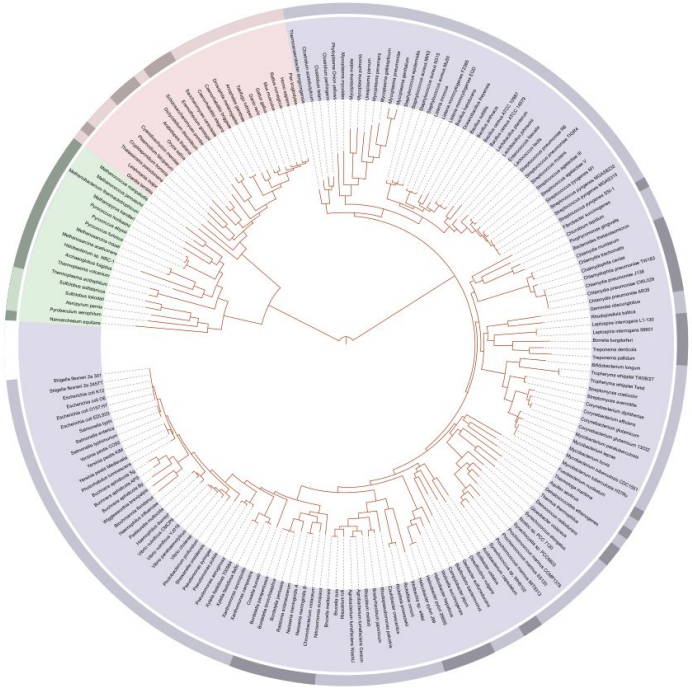
Toward Automatic Reconstruction of a Highly Resolved Tree of Life

Francesca D. Ciccarelli[1,2,3,*], Tobias Doerks[1,*], Christian von Mering[1], Christopher J. Creevey[1], Berend Snel[4], Peer Bork[1,5,†]

Based on very few genes.

The tree of one percent

Tal Dagan ✉ and William Martin



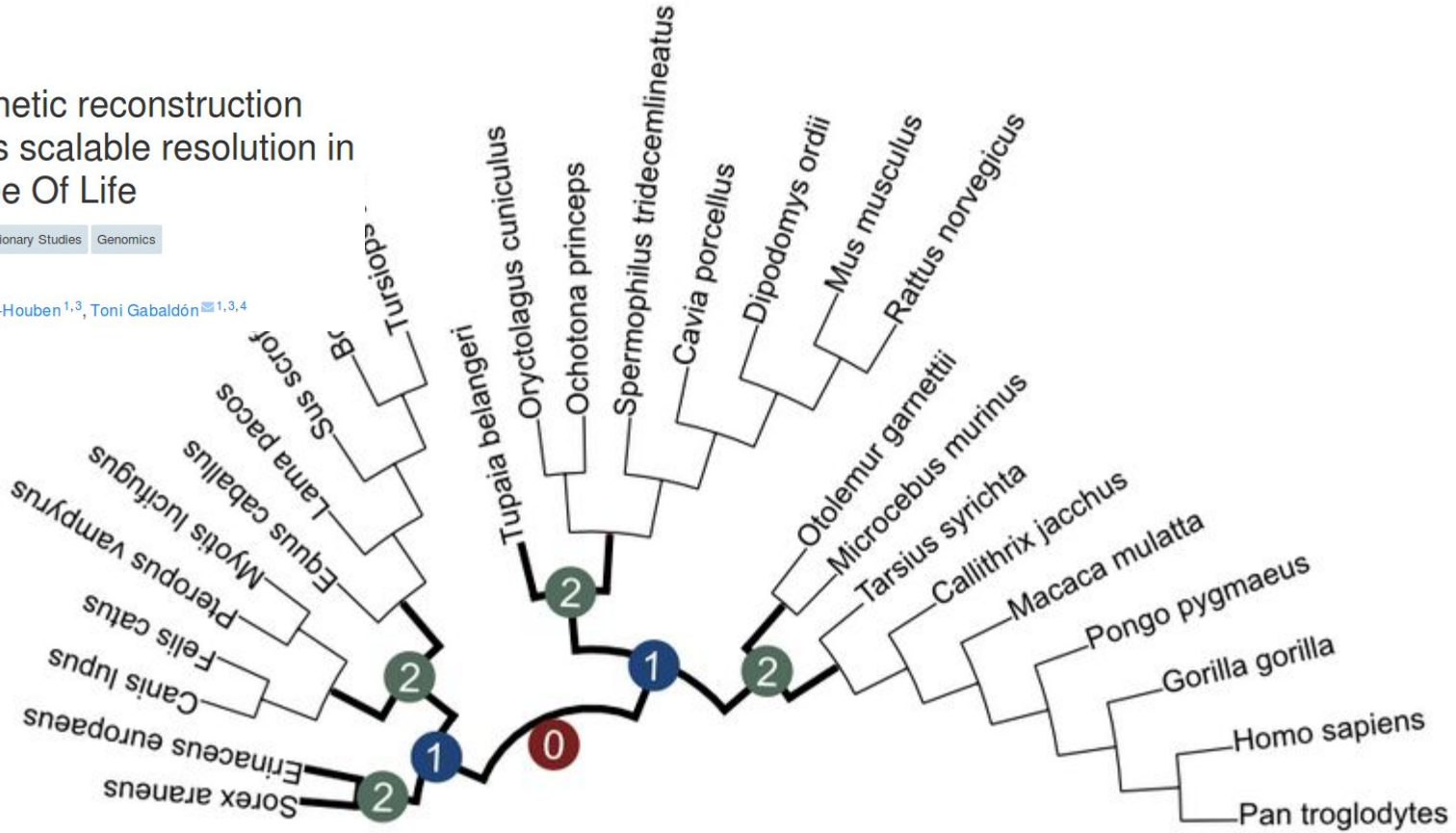
To build such a broad tree it is difficult to obtain a set of gene trees that are not affected by duplications of HGT events and that are still present in the majority of the species

Divide & Conquer approach to reconstruct nested phylogenetic trees.

A nested phylogenetic reconstruction approach provides scalable resolution in the eukaryotic Tree Of Life

Bioinformatics | Computational Biology | Evolutionary Studies | Genomics

Jaime Huerta-Cepas [1,2], Marina Marcet-Houben [1,3], Toni Gabaldón [1,3,4]

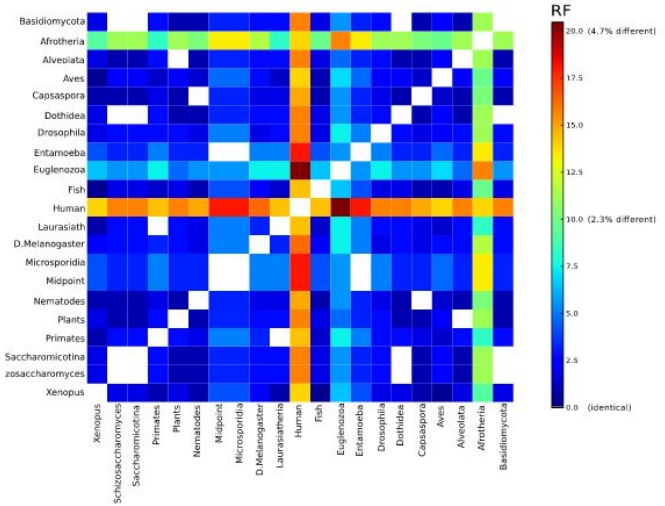


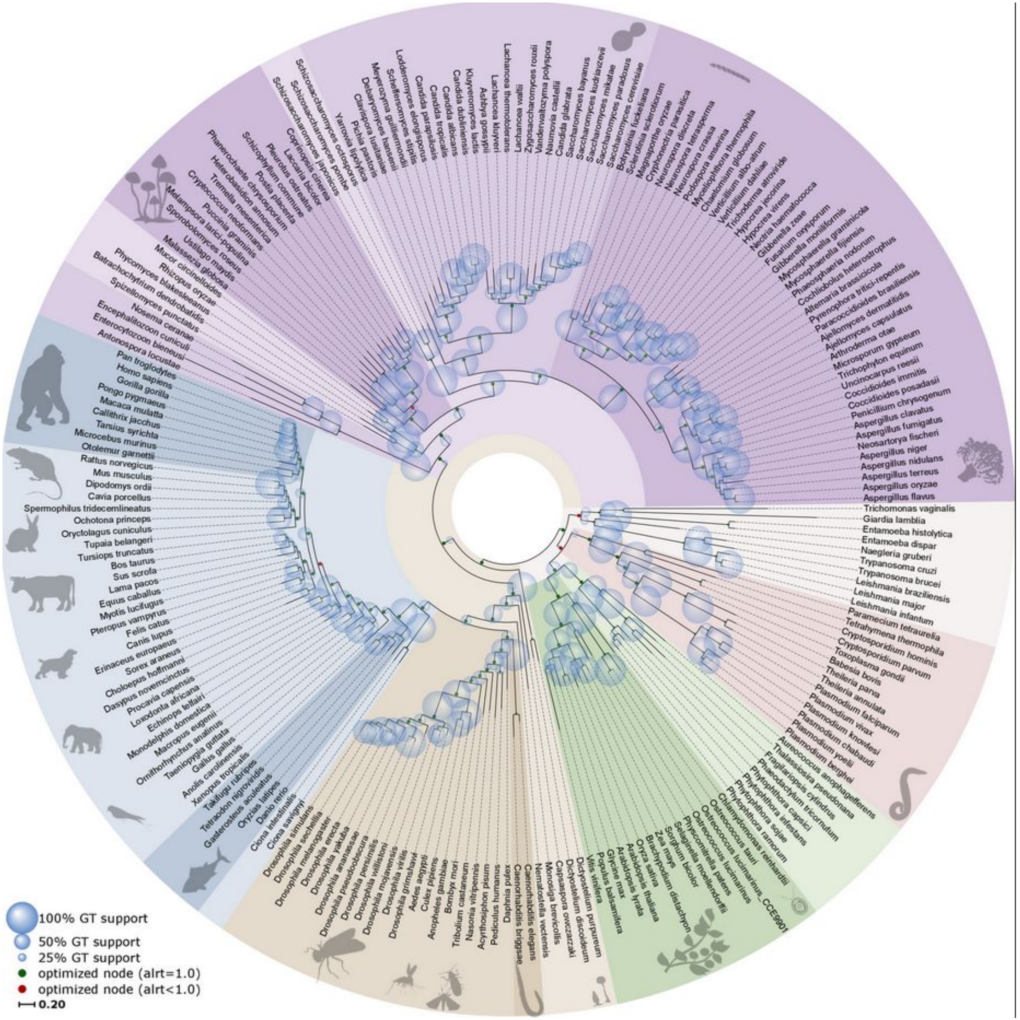
There are several drawback to this method:

1.- Branch lengths will need to be calculated afterwards since they will be changing every time a node is re-calculated.

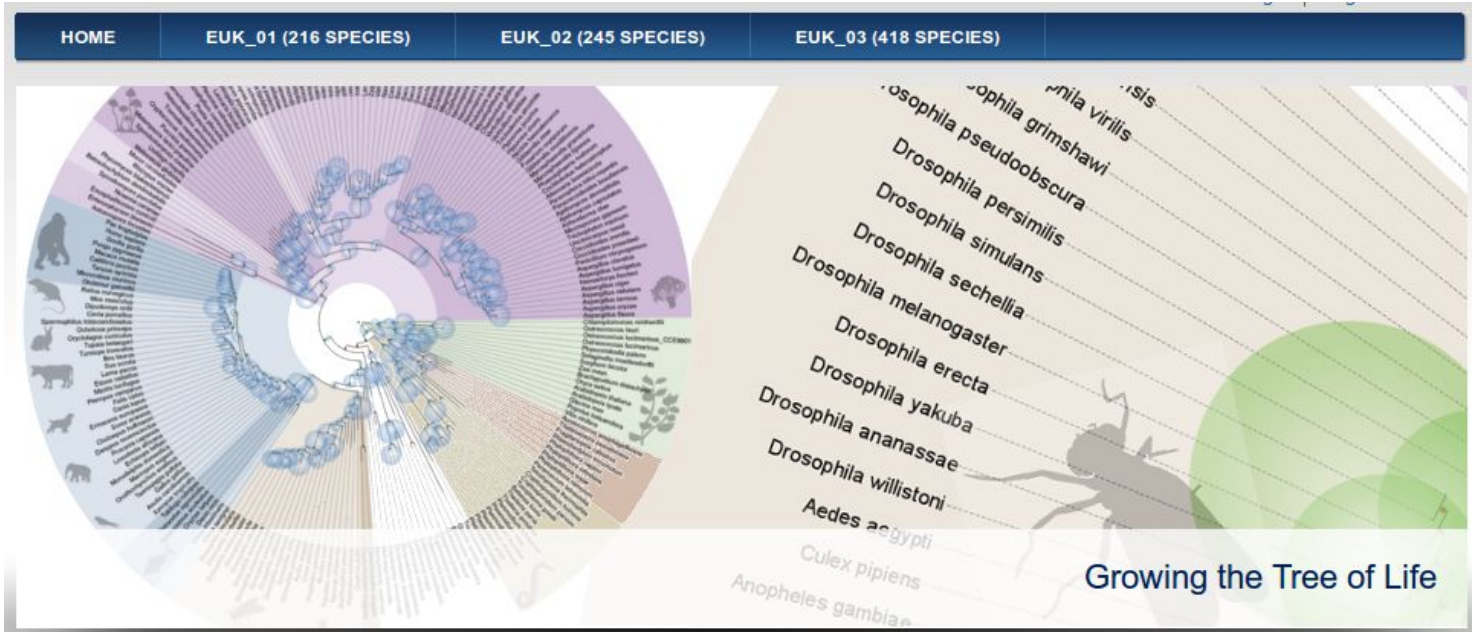
2.- It strictly depends on the consecutives trees being correct, a bad decision will destroy the whole analysis. Errors tend to happen more often in early branches where less data can be used and therefore tend to have a bigger effect.

3.- It is highly dependant on having a good outgroup.





Trichomonas vaginalis
Giardia lamblia
Entamoeba histolytica
Entamoeba dispar
Naegleria gruberi
Trypanosoma cruzi
Trypanosoma brucei
Leishmania braziliensis
Leishmania major
Leishmania infantum
Paramecium tetraurelia
Tetrahymena thermophila
Cryptosporidium hominis
Cryptosporidium parvum
Toxoplasma gondii
Babesia bovis
Theileria parva
Theileria annulata
Plasmodium falciparum
Plasmodium vivax
Plasmodium knowlesi
Plasmodium chabaudi
Plasmodium yoelii
Plasmodium berghei
Aureococcus anophagefferens
Thalassiosira pseudonana
Fragilariopsis cylindrus
Phaeodactylum tricornutum
Phytophthora capsici
Phytophthora infestans
Phytophthora sojae
Phytophthora ramorum
Chlamydomonas reinhardtii
Aspergillus flavus
Aspergillus oryzae
Aspergillus terreus
Aspergillus nidulans
Aspergillus niger
Neosartorya fischeri
Aspergillus fumigatus
Aspergillus clavatus
Penicillium chrysogenum
Coccidioides posadasii
Coccidioides immitis
Uncinocarpus reesii
Trichophyton equinum
Microsporum gypseum
Arthroderma otae
Ajellomyces capsulatus
Ajellomyces dermatitidis
Paracoccidioides brasiliensis
Pyrenophora tritici-repentis
Alternaria brassicicola
Cochliobolus heterostrophus
Phaeosphaeria nodorum
Mycosphaerella fijiensis
Mycosphaerella graminicola
Gibberella moniliformis
Fusarium oxysporum
Gibberella zeae
Nectria haematococca
Pan troglodytes
Homo sapiens
Gorilla gorilla
Pongo pygmaeus
Macaca mulatta
Callithrix jacchus
Tarsius syrichta
Microcebus murinus
Otolemur garnettii
Rattus norvegicus
Mus musculus
Dipodomys ordii
Cavia porcellus
Spermophilus tridecemlineatus
Ochotona princeps
Oryctolagus cuniculus
Tupaia belangeri
Tursiops truncatus
Bos taurus
Sus scrofa
Lama pacos
Equus caballus
Myotis lucifugus
Pteropus vampyrus
Felis catus
Canis lupus
Erinaceus europaeus
Sorex araneus
Choloepus hoffmanni
Dasypus novemcinctus
Procavia capensis
Loxodonta africana
Echinops telfairi
Monodelphis domestica
Macropus eugenii
Ornithorhynchus anatinus
Taeniopygia guttata
Gallus gallus
Anolis carolinensis
Xenopus tropicalis
Takifugu rubripes
Tetraodon nigroviridis
Gasterosteus aculeatus
Oryzias latipes
Danio rerio
Ciona intestinalis
Ciona savignyi
100% GT support
50% GT support
25% GT support
optimized node (alrt=1.0)
optimized node (alrt<1.0)
0.20



A nested phylogenetic reconstruction approach provides scalable resolution in the eukaryotic Tree Of Life

http://tol.cgenomics.org/

1.- Exercice: How to get trees that have only one-to-one orthologs given a list of gene trees.

Download the list of phylogenetic trees from the evomics website.

Open an ipython terminal. We will first obtain the list of trees that only contain one-to-one orthologs.

```
import ete3

def get_species_name(node):
    return node.split("_")[1]

outfile = open("one2one_trees.txt","w")
for line in open("trees.1000.txt"):
    line = line.strip()
    code,newick = line.split("\t")
    t = ete3.PhyloTree(newick,sp_naming_function=get_species_name)
    species_size = len(t.get_species())
    leaf_size = len(t.get_leaves())
    if species_size == leaf_size:
        print >>outfile,code,len(t.get_species()),t.write()
outfile.close()
```

2.- Exercice: Filter out species specific duplications

```
outfile = open("filtered_trees.txt","w")
for line in open("trees.1000.txt"):
    line = line.strip()
    code,newick = line.split("\t")
    t = ete3.PhyloTree(newick,sp_naming_function=get_species_name)
    t = t.collapse_lineage_specific_expansions()
    if len(t.get_leaves()) == len(t.get_species()):
        print >>outfile,code,len(t.get_species()),t.write()
outfile.close()
```

3.- Exercice: Obtain orthologous trees

```
outfile = open("list_orthologous_trees.txt","w")
for line in open("fileName.txt"):
    line = line.strip()
    code,newick = line.split("\t")
    t = ete3.PhyloTree(newick,sp_naming_function=get_species_name)
    t = t.collapse_lineage_specific_expansions()
    orthoTrees = t.get_speciation_trees()[2]
    best_ortho_tree = None
    for ot in orthoTrees:
        tree_size = len(ot.get_leaf_names())
        if not best_ortho_tree:
            best_ortho_tree = ot
        else:
            saved_tree_size = len(best_ortho_tree.get_leaf_names())
            if saved_tree_size < tree_size:
                best_ortho_tree = ot
    print code,len(best_ortho_tree.get_leaf_names()),best_ortho_tree.write()
outfile.close()
```

4.- Exercice: Obtain a tree using a super-tree approach

```
outfile = open("duptree_trees.txt","w")
for line in open("fileName.txt"):
    line = line.strip()
    code,newick = line.split("\t")
    t = ete3.PhyloTree(newick,sp_naming_function=get_species_name)
    for leaf in t.iter_leaves():
        leaf.name = leaf.species
    print t.write(format=9)
outfile.close()
```

Download duptree from their website: http://genome.cs.iastate.edu/CBL/DupTree/
Uncompress the file `tar -zxvf linux.i386.tar.gz`
Now use duptree to obtain the super-tree:

```
./linux-i386/duptree -i duptree_trees.txt -o species_tree.nw
```

Genome-Scale Phylogenetics: Inferring the Plant Tree of Life from 18,896 Gene Trees

J. Gordon Burleigh,[1,2,*] Mukul S. Bansal,[3,4] Oliver Eulenstein,[3] Stefanie Hartmann,[5,6] André Wehe,[3] and Todd J. Vision[2,5]