EXCELENCIA SEVERO OCHOA

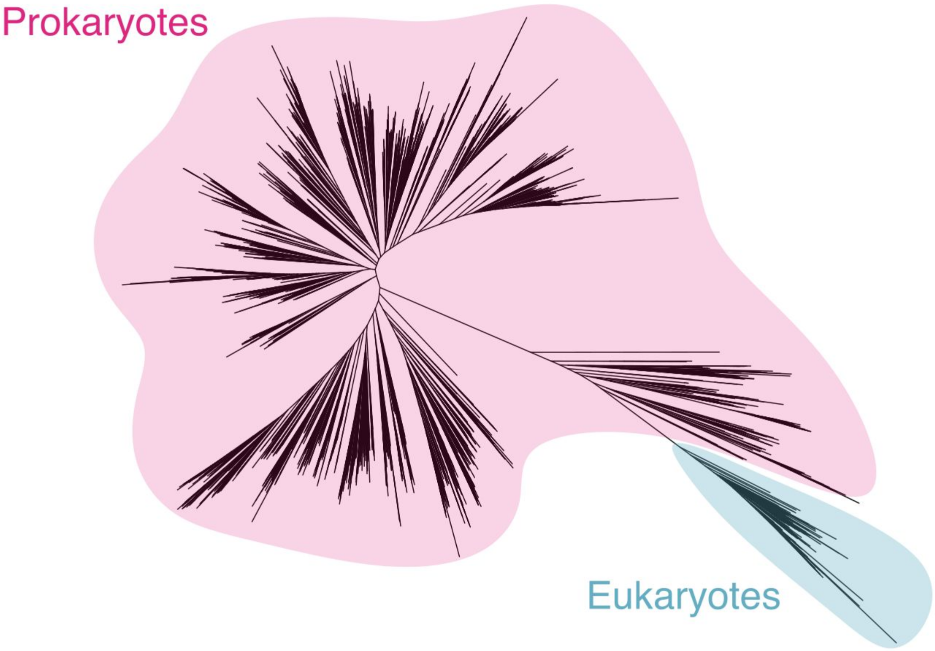
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

11th BSC PhD Symposium

Reconstructing prokaryotic metabolic contributions to LECA

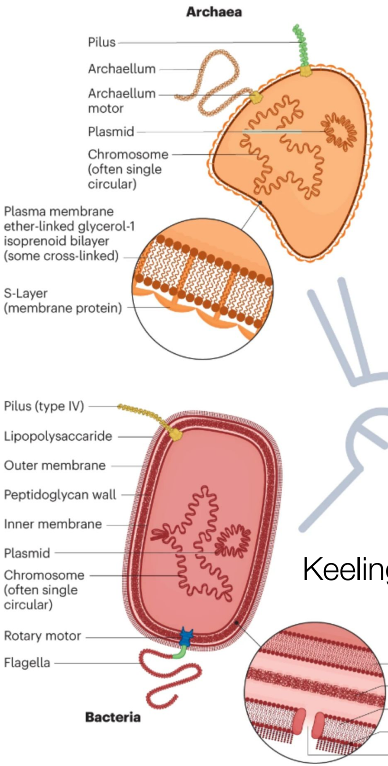
Saioa Manzano-Morales
Comparative Genomics - Life Sciences

Prokaryotes and eukaryotes, the great divide

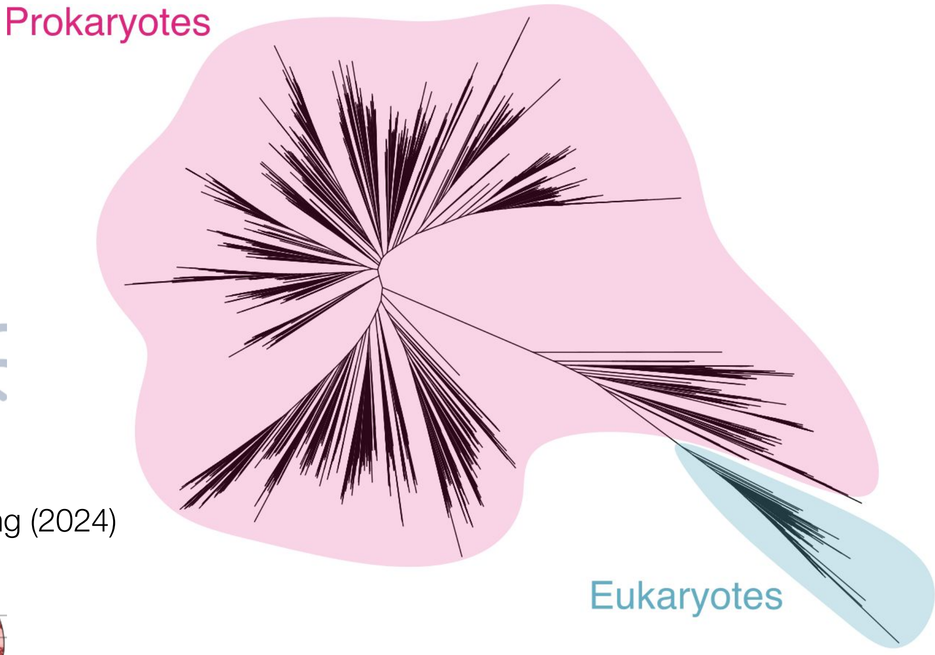


Adapted from Hug *et al.* (2016)

Prokaryotes and eukaryotes, the great divide

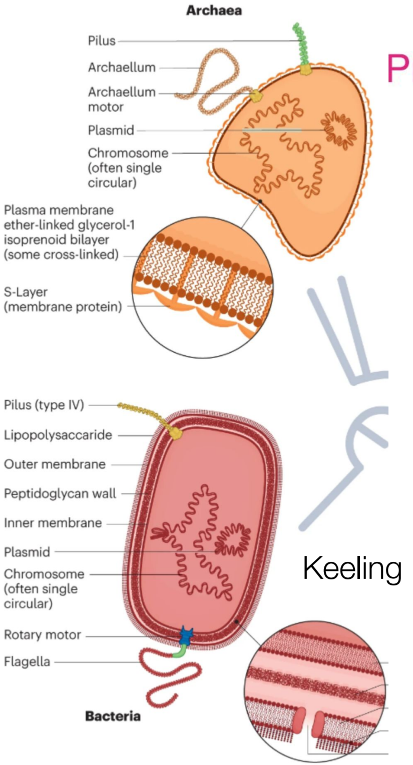

Archaea
Pilus
Archaellum
Archaellum motor
Plasmid
Chromosome (often single circular)
Plasma membrane ether-linked glycerol-1 isoprenoid bilayer (some cross-linked)
S-Layer (membrane protein)
Pilus (type IV)
Lipopolysaccaride
Outer membrane
Peptidoglycan wall
Inner membrane
Plasmid
Chromosome (often single circular)
Rotary motor
Flagella
Bacteria

Keeling (2024)

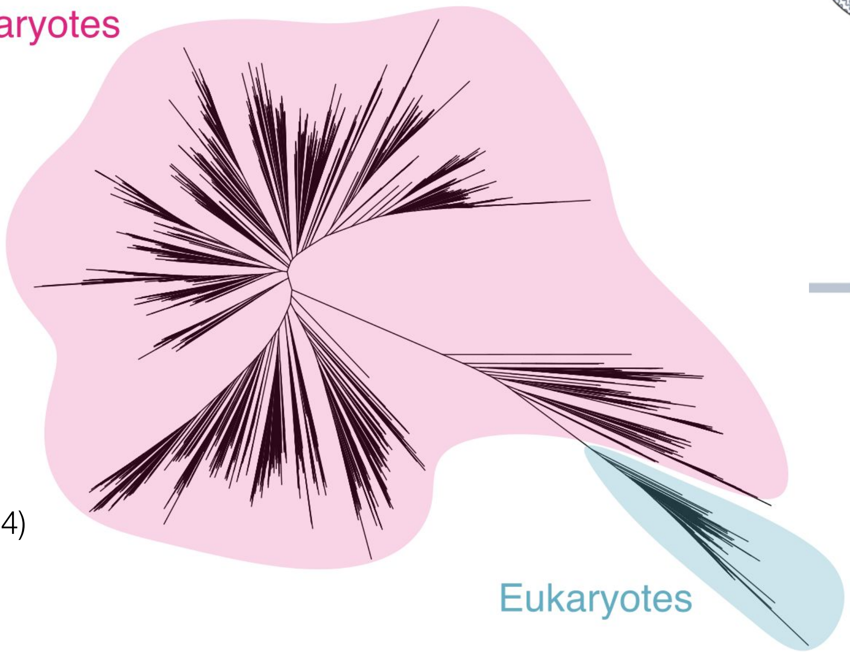


Adapted from Hug *et al.* (2016)

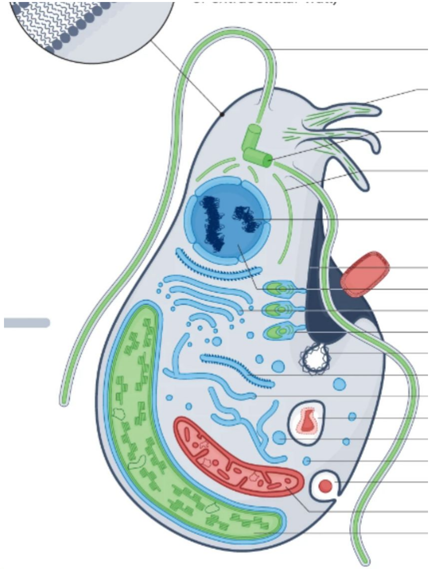
Prokaryotes and eukaryotes, the great divide


Archaea
Pilus
Archaellum
Archaellum motor
Plasmid
Chromosome (often single circular)
Plasma membrane ether-linked glycerol-1 isoprenoid bilayer (some cross-linked)
S-Layer (membrane protein)
Pilus (type IV)
Lipopolysaccaride
Outer membrane
Peptidoglycan wall
Inner membrane
Plasmid
Chromosome (often single circular)
Rotary motor
Flagella
Bacteria

Keeling (2024)

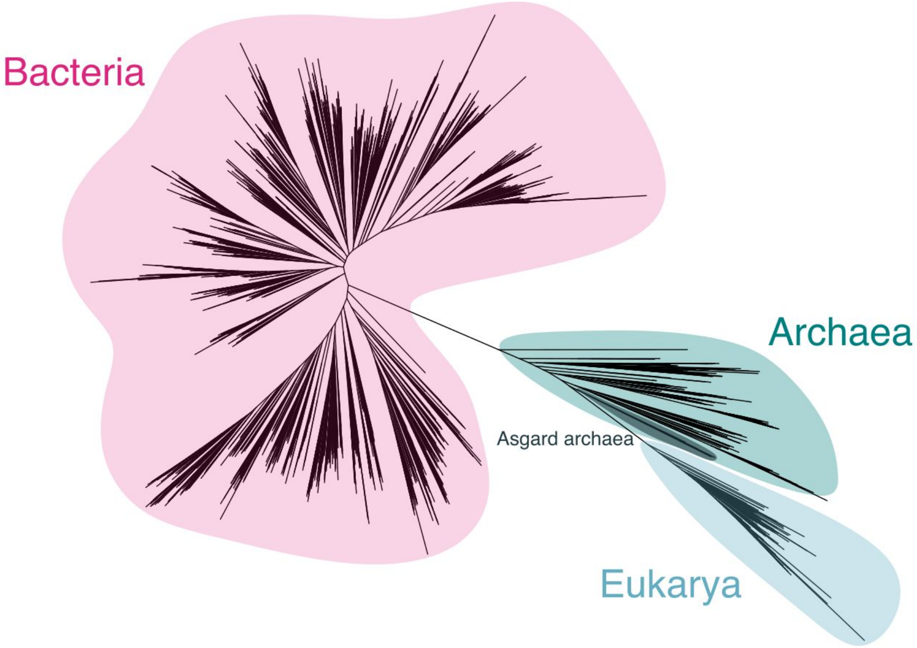


Adapted from Hug *et al.* (2016)



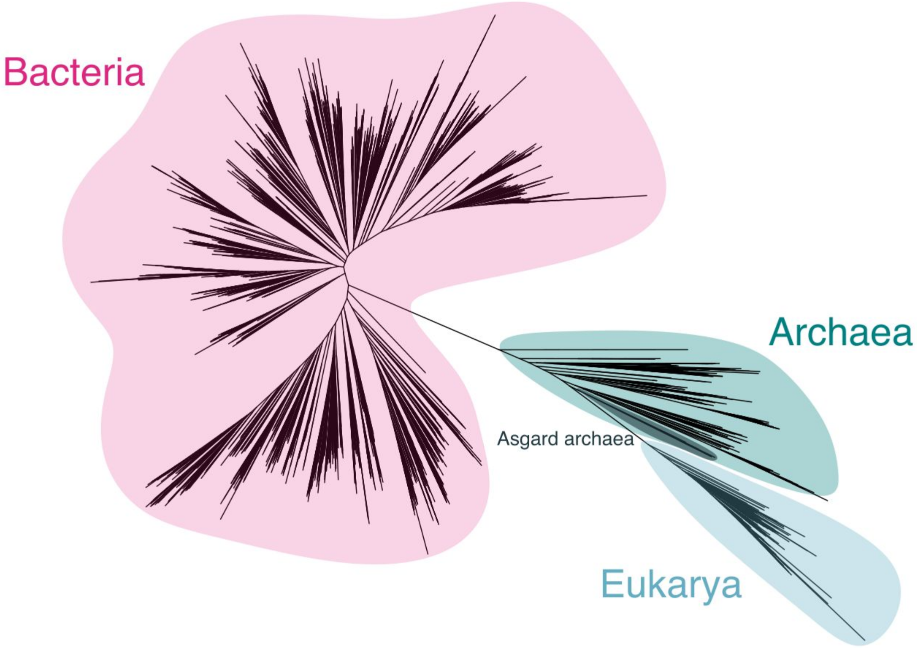
Keeling (2024)

Eukaryotes: the “love child” of a prokaryotic affair

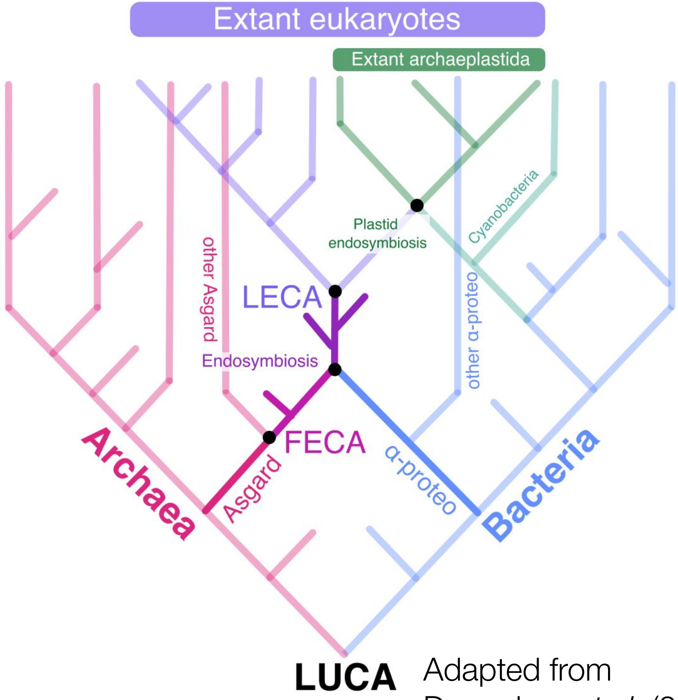


Adapted from Hug *et al.* (2016)

Eukaryotes: the “love child” of a prokaryotic affair



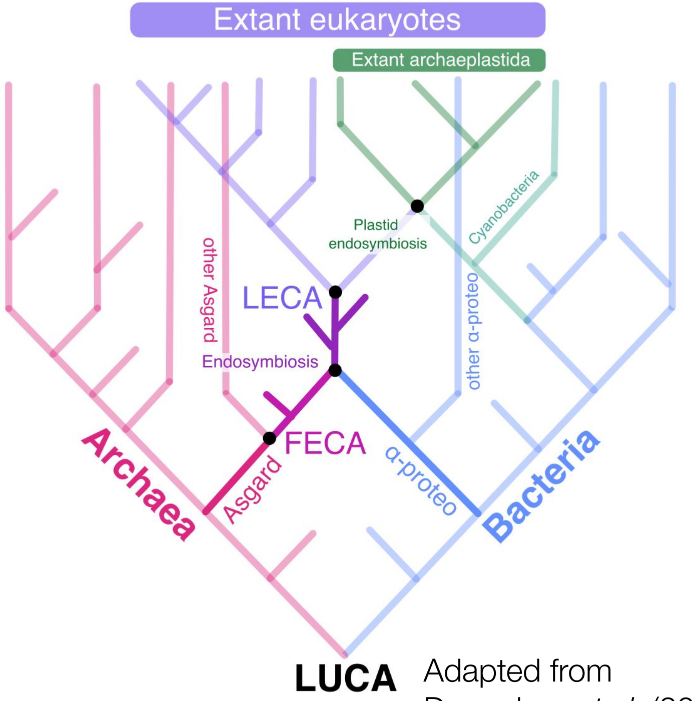
Adapted from Hug *et al.* (2016)



Adapted from Donoghue *et al.* (2023)

Eukaryotes: the “love child” of a prokaryotic affair

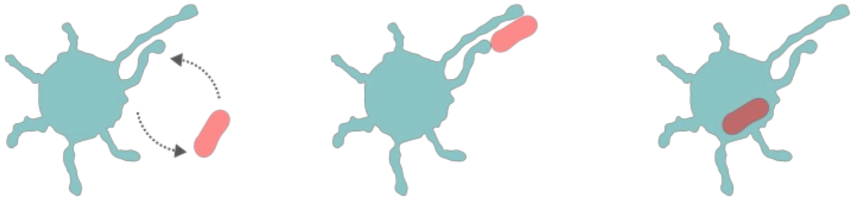
Eukaryotes stem from an **endosymbiotic event** between an Asgard archaeon (the host) and an alpha-proteobacterial endosymbiont

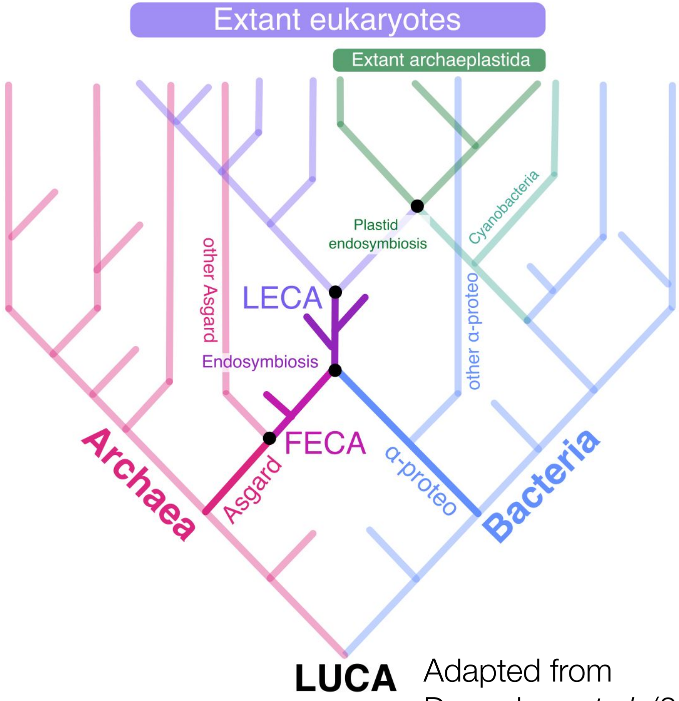


Adapted from Donoghue *et al.* (2023)

Eukaryotes: the "love child" of a prokaryotic affair

Eukaryotes stem from an **endosymbiotic event** between an Asgard archaeon (the host) and an alpha-proteobacterial endosymbiont



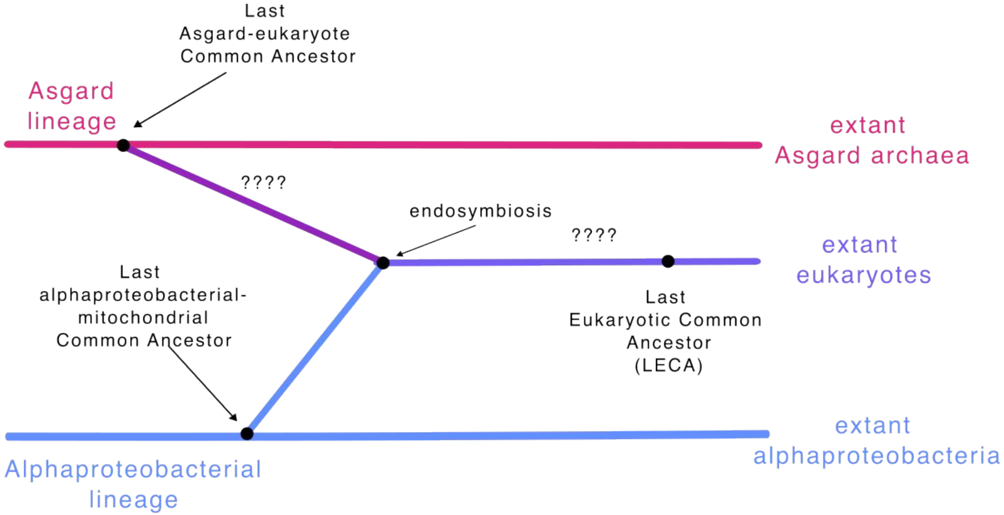


Adapted from Donoghue *et al.* (2023)

Eukaryotes: the “love child” of a prokaryotic affair

But this story does not explain all:

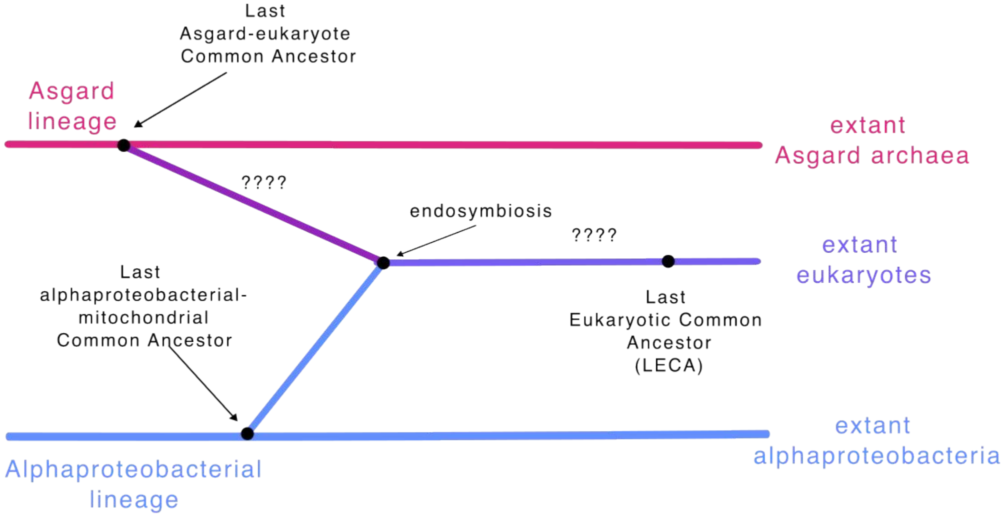
- How did **eukaryotic traits** emerge and evolve?
    - nucleus
    - introns
    - phagocytosis
    - endomembrane system
    - …
- Why does a **two-partner scenario** not fit **all** data?

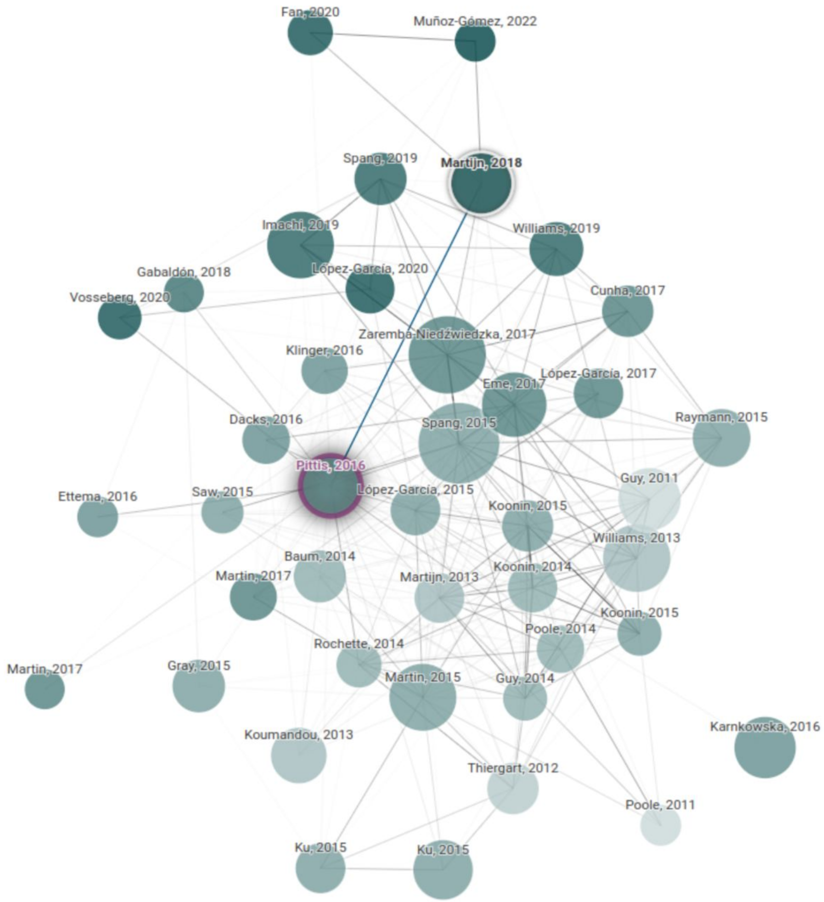

Last Asgard-eukaryote Common Ancestor
Asgard lineage
extant Asgard archaea
????
endosymbiosis
????
extant eukaryotes
Last alphaproteobacterial-mitochondrial Common Ancestor
Last Eukaryotic Common Ancestor (LECA)
extant alphaproteobacteria
Alphaproteobacterial lineage

Eukaryotes: the "love child" of a prokaryotic affair

But this story does not explain all:

- How did **eukaryotic traits** emerge and evolve?
- Why does a **two-partner scenario** not fit **all** data?

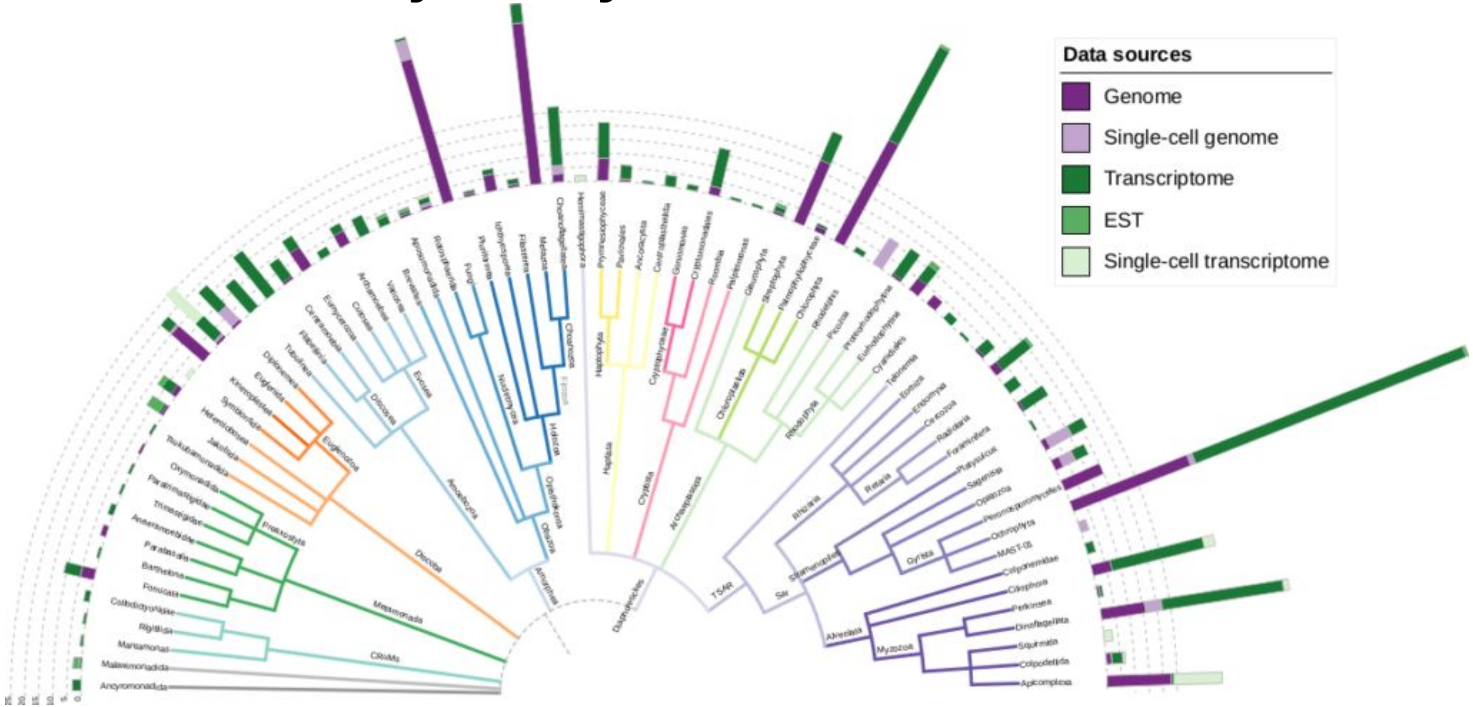

Last Asgard-eukaryote Common Ancestor
Asgard lineage
extant Asgard archaea
????
endosymbiosis
????
extant eukaryotes
Last alphaproteobacterial-mitochondrial Common Ancestor
Last Eukaryotic Common Ancestor (LECA)
extant alphaproteobacteria
Alphaproteobacterial lineage


Fan, 2020
Muñoz-Gómez, 2022
Spang, 2019
Martijn, 2018
Imachi, 2019
Williams, 2019
Gabaldón, 2018
López-García, 2020
Vosseberg, 2020
Cunha, 2017
Zaremba-Niedźwiedzka, 2017
Klinger, 2016
Eme, 2017
López-García, 2017
Dacks, 2016
Spang, 2015
Raymann, 2015
Pittis, 2016
Guy, 2011
Ettema, 2016
Saw, 2015
López-García, 2015
Koonin, 2015
Williams, 2013
Baum, 2014
Martin, 2017
Martijn, 2013
Koonin, 2014
Koonin, 2015
Poole, 2014
Rochette, 2014
Martin, 2017
Gray, 2015
Martin, 2015
Guy, 2014
Karnkowska, 2016
Koumandou, 2013
Thiergart, 2012
Poole, 2011
Ku, 2015
Ku, 2015

Why now?

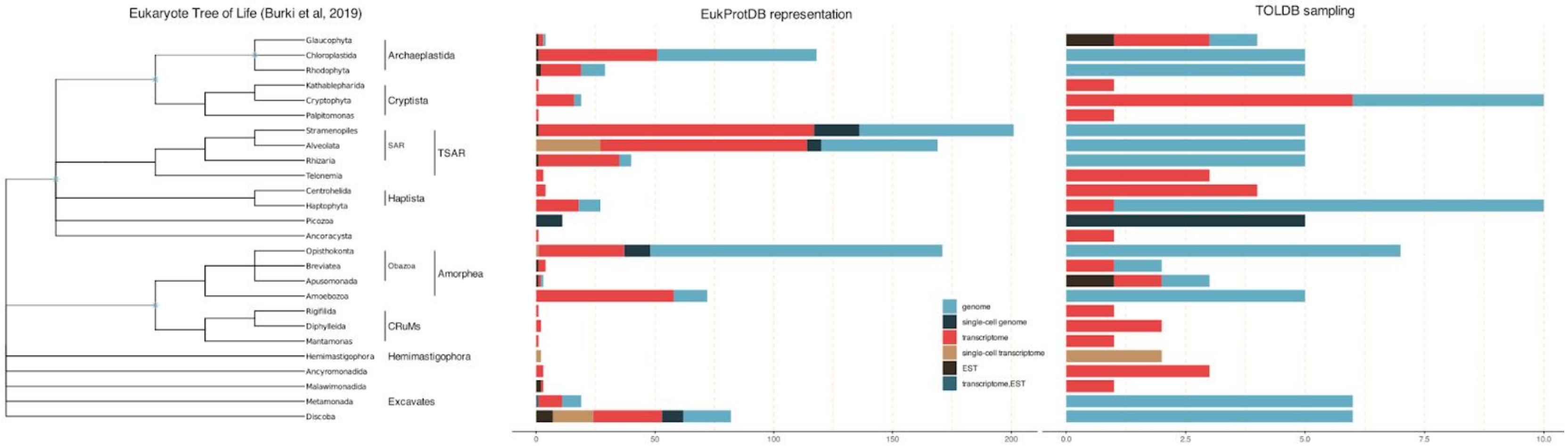
2011
2022

Sequencing boom allows us to gain information on underrepresented and key eukaryotic clades



(Eukprot V3)

Better taxon sampling allows for a more representative image of eukaryotes



TOLDBA

TOLDBB

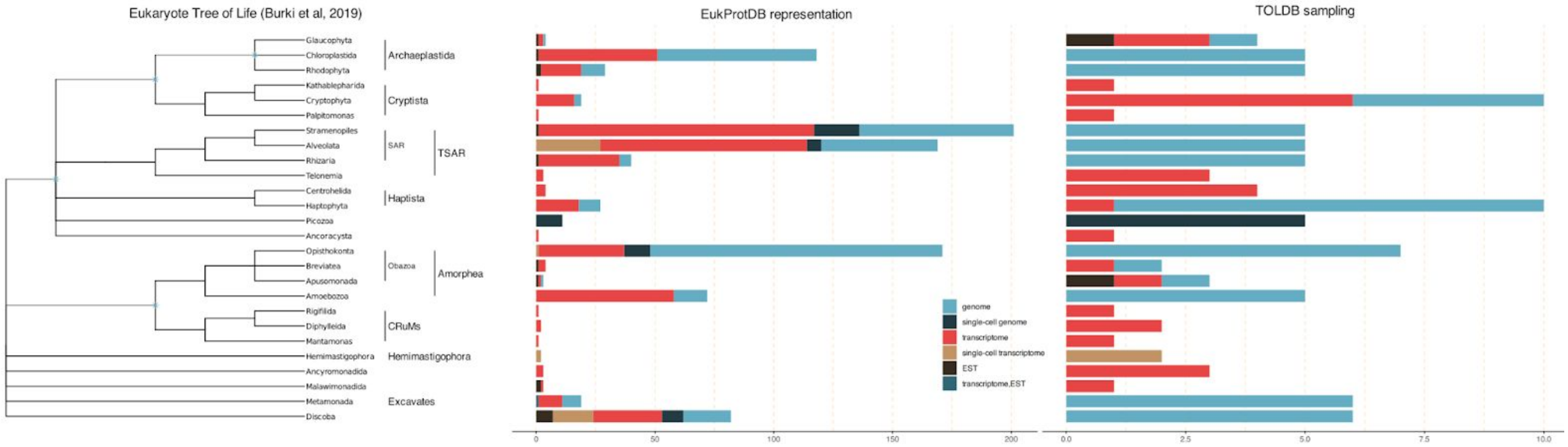
TOLDBC

2.93M proteins

2.91M proteins

2.89M proteins

Better taxon sampling allows for a more representative image of eukaryotes



HPC allows us unprecedented scale in the analysis of eukaryotic gene families

TOLDBA	TOLDBB	TOLDBC
2.93M proteins	2.91M proteins	2.89M proteins

Taxonomically balanced, focus on underrepresented groups

HPC allows us unprecedented scale in the analysis of eukaryotic gene families

TOLDBA	TOLDBB	TOLDBC
2.93M proteins	2.91M proteins	2.89M proteins

↓ gene family inference

	Prok+	Prok-	NonLECA
TOLDBA	**5152**	7567	70204
TOLDBB	**4458**	6470	73382
TOLDBC	**4371**	6181	74319

HPC allows us unprecedented scale in the analysis of eukaryotic gene families

TOLDBA	TOLDBB	TOLDBC
2.93M proteins	2.91M proteins	2.89M proteins

↓ gene family inference

	Prok+	Prok-	NonLECA
TOLDBA	**5152**	7567	70204
TOLDBB	**4458**	6470	73382
TOLDBC	**4371**	6181	74319

For each gene family:

1. Search against dataset of prokaryotic proteins (BROAD-DB) - 232M proteins → Known prokaryotic protein universe

HPC allows us unprecedented scale in the analysis of eukaryotic gene families

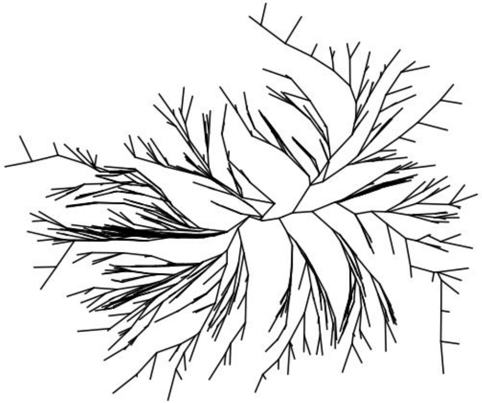
TOLDBA	TOLDBB	TOLDBC
2.93M proteins	2.91M proteins	2.89M proteins

↓ gene family inference

	Prok+	Prok-	NonLECA
TOLDBA	**5152**	7567	70204
TOLDBB	**4458**	6470	73382
TOLDBC	**4371**	6181	74319

For each gene family:

1. Search against dataset of prokaryotic proteins (BROAD-DB) - 232M proteins
2. Gene tree inference (FastTree)



HPC allows us unprecedented scale in the analysis of eukaryotic gene families

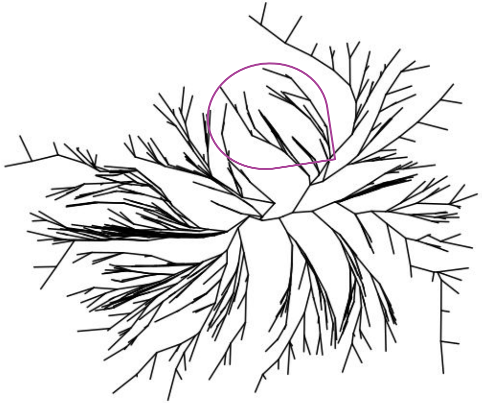
TOLDBA	TOLDBB	TOLDBC
2.93M proteins	2.91M proteins	2.89M proteins

gene family inference

	Prok+	Prok-	NonLECA
TOLDBA	**5152**	7567	70204
TOLDBB	**4458**	6470	73382
TOLDBC	**4371**	6181	74319

For each gene family:

1. Search against dataset of prokaryotic proteins (BROAD-DB) - 232M proteins
2. Gene tree inference (FastTree)
3. Tree pruning



HPC allows us unprecedented scale in the analysis of eukaryotic gene families

TOLDBA	TOLDBB	TOLDBC
2.93M proteins	2.91M proteins	2.89M proteins

↓ gene family inference

	Prok+	Prok-	NonLECA
TOLDBA	**5152**	7567	70204
TOLDBB	**4458**	6470	73382
TOLDBC	**4371**	6181	74319

For each gene family:

1. Search against dataset of prokaryotic proteins (BROAD-DB) - 232M proteins
2. Gene tree inference (FastTree)
3. Tree pruning
4. Final gene tree inference (IQ-TREE)

HPC allows us unprecedented scale in the analysis of eukaryotic gene families

TOLDBA	TOLDBB	TOLDBC
2.93M proteins	2.91M proteins	2.89M proteins

↓ gene family inference

	Prok+	Prok-	NonLECA
TOLDBA	**5152**	7567	70204
TOLDBB	**4458**	6470	73382
TOLDBC	**4371**	6181	74319

For each gene family:

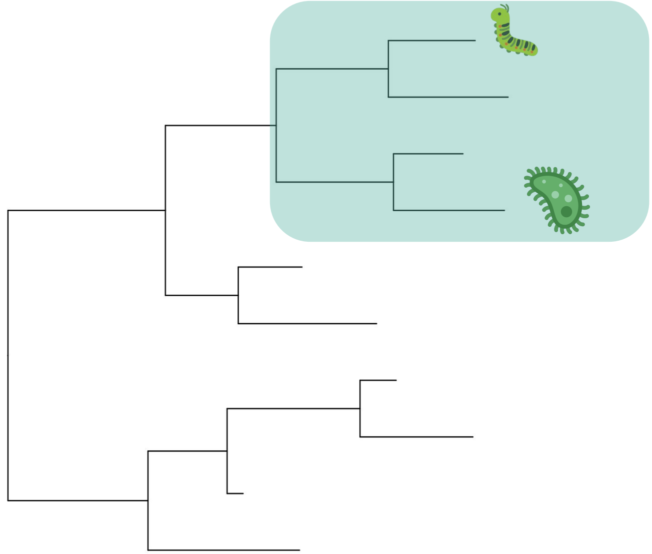
1. Search against dataset of prokaryotic proteins (BROAD-DB) - 232M proteins
2. Gene tree inference (FastTree)
3. Tree pruning
4. Final gene tree inference (IQ-TREE)

1 single tree for step 4, in 4 threads - 4h (optimistically)

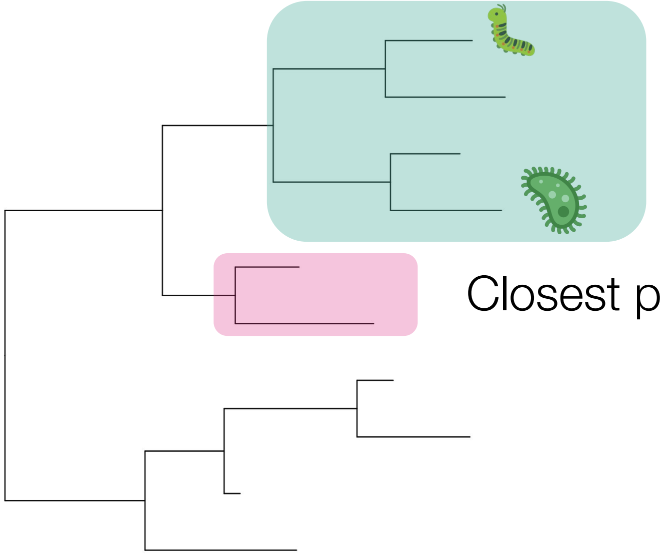
All trees on 1 thread:
547184h - **62 years**

Marina Marcet-Houben

Improved methods in phylogenetic trees allow us to make better inferences on the origin and evolution of gene families

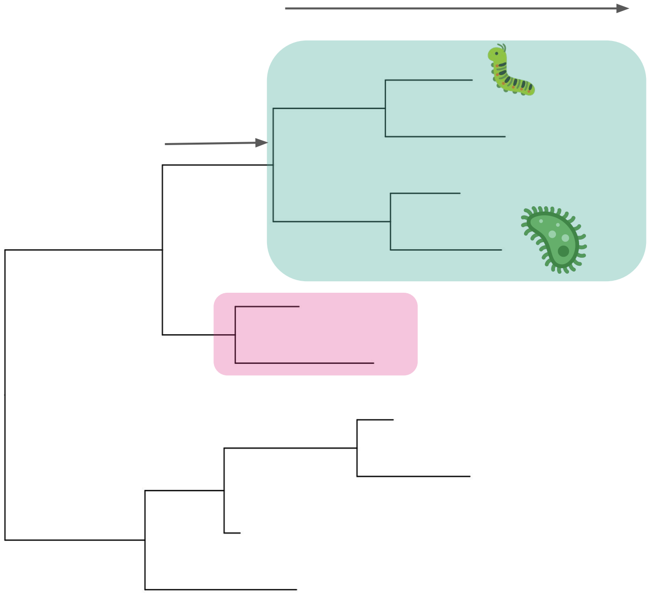


Improved methods in phylogenetic trees allow us to make better inferences on the origin and evolution of gene families



Closest prokaryotic group - Likely donor

Improved methods in phylogenetic trees allow us to make better inferences on the origin and evolution of gene families

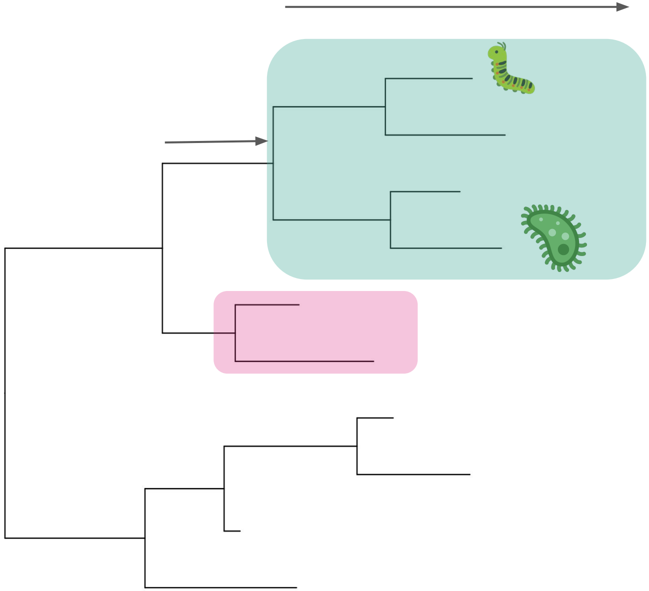


Branch lengths - relative timing

$$sl = \frac{rsl}{\mathrm{Median}\ (ebl)}$$

(Pittis & Gabaldón, 2016)

Improved methods in phylogenetic trees allow us to make better inferences on the origin and evolution of gene families



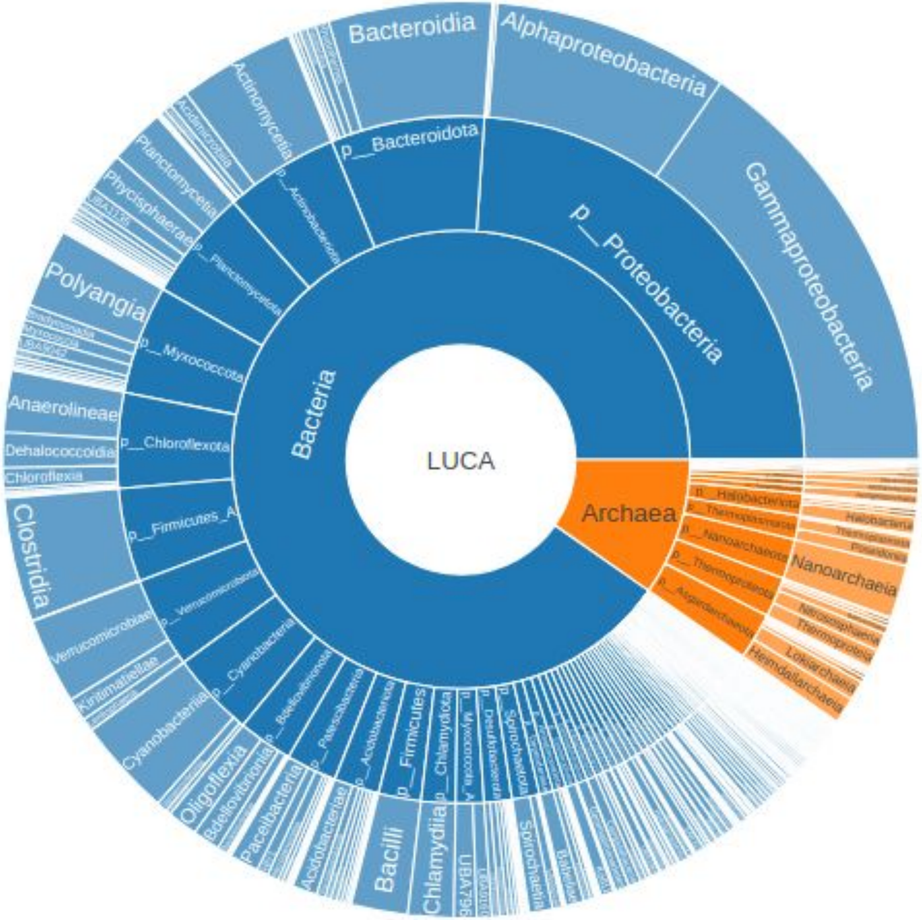
Branch lengths - relative timing

$$sl = \frac{rsl}{\text{Median (ebl)}}$$

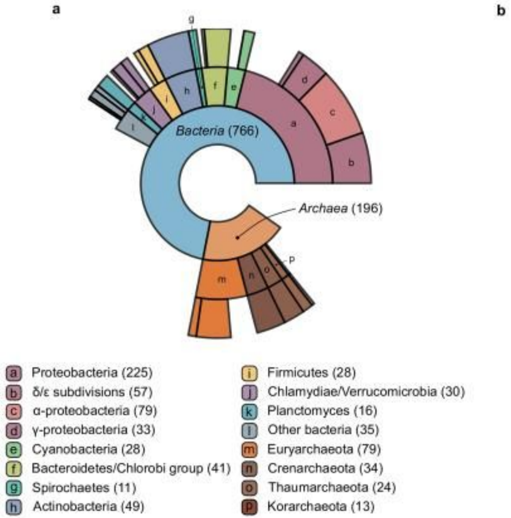
(Pittis & Gabaldón, 2016)

Moisès Bernabeu

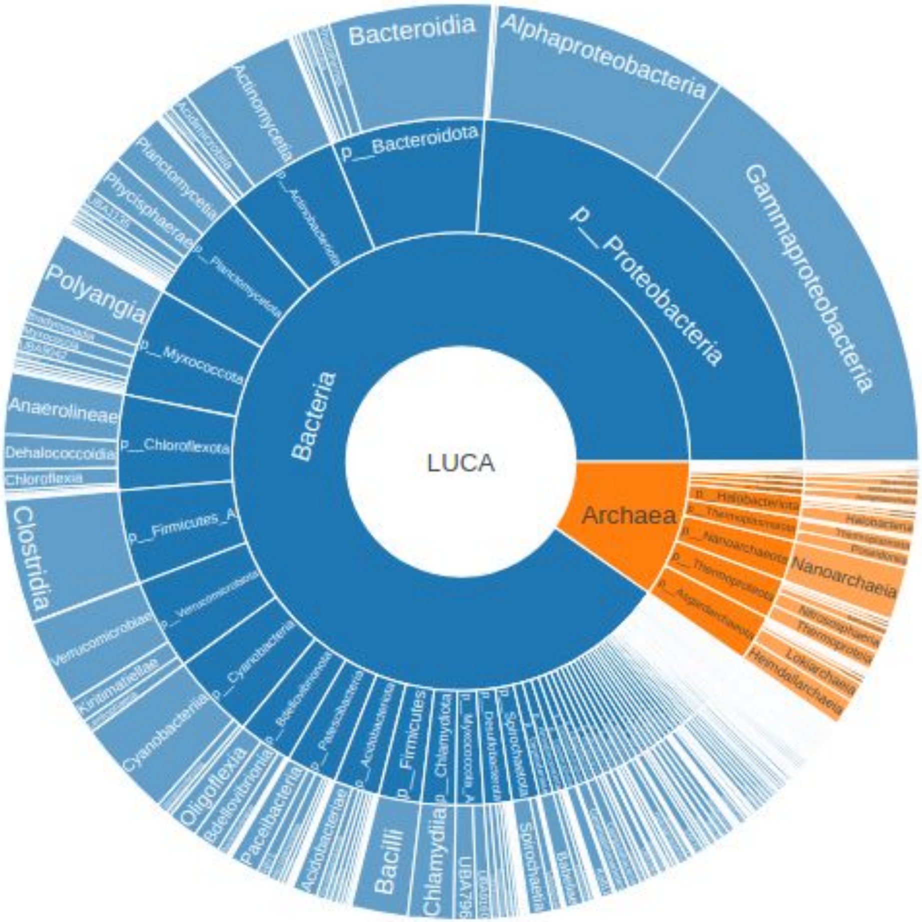
LECA is a mosaic of proteins of vastly different origins



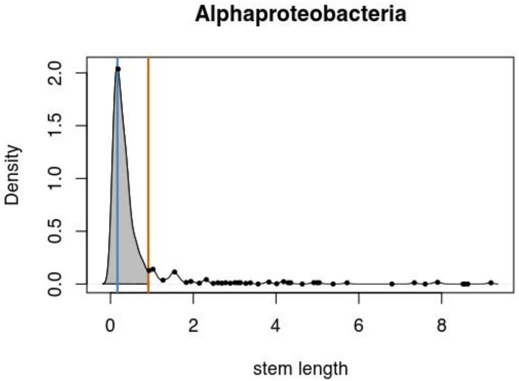
LECA is a mosaic of proteins of vastly different origins



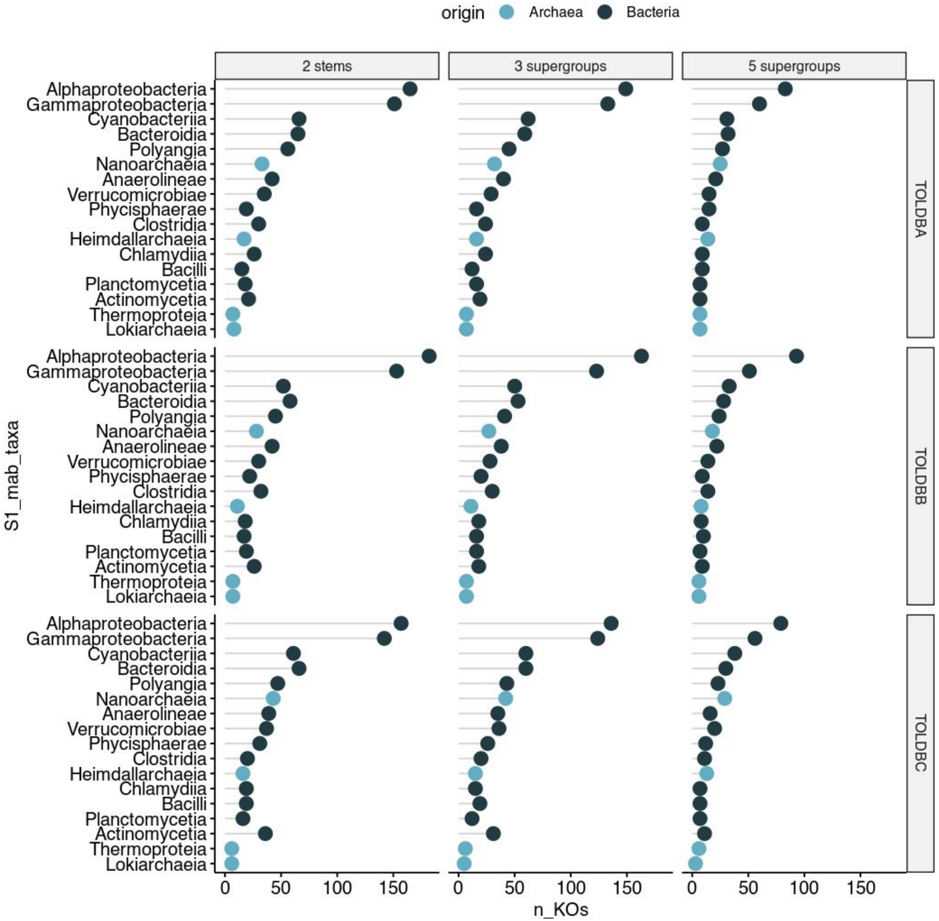
(Pittis & Gabaldón, 2016)



We can classify these into roughly 17 “modules” - major waves of gene acquisition from same donor at roughly the same time



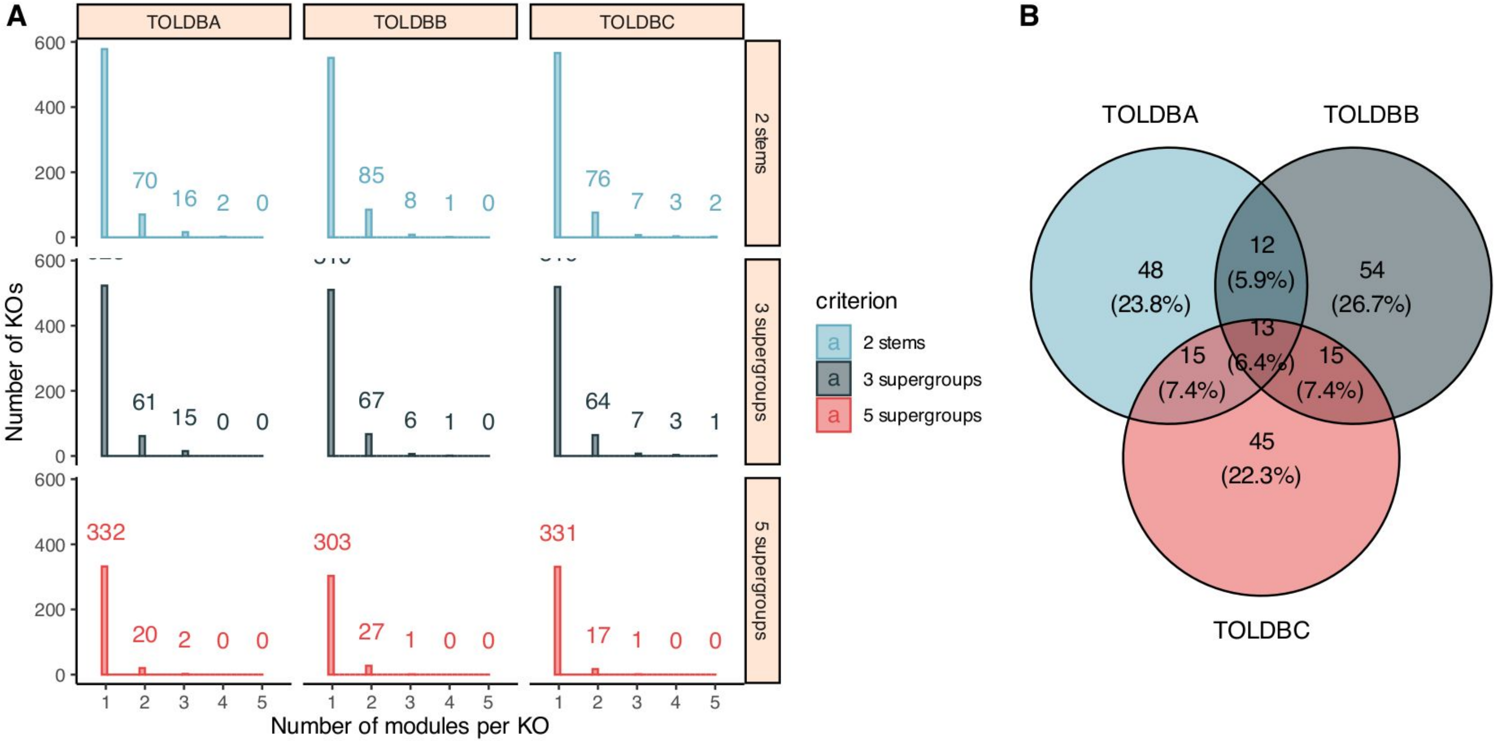
Moisès Bernabeu



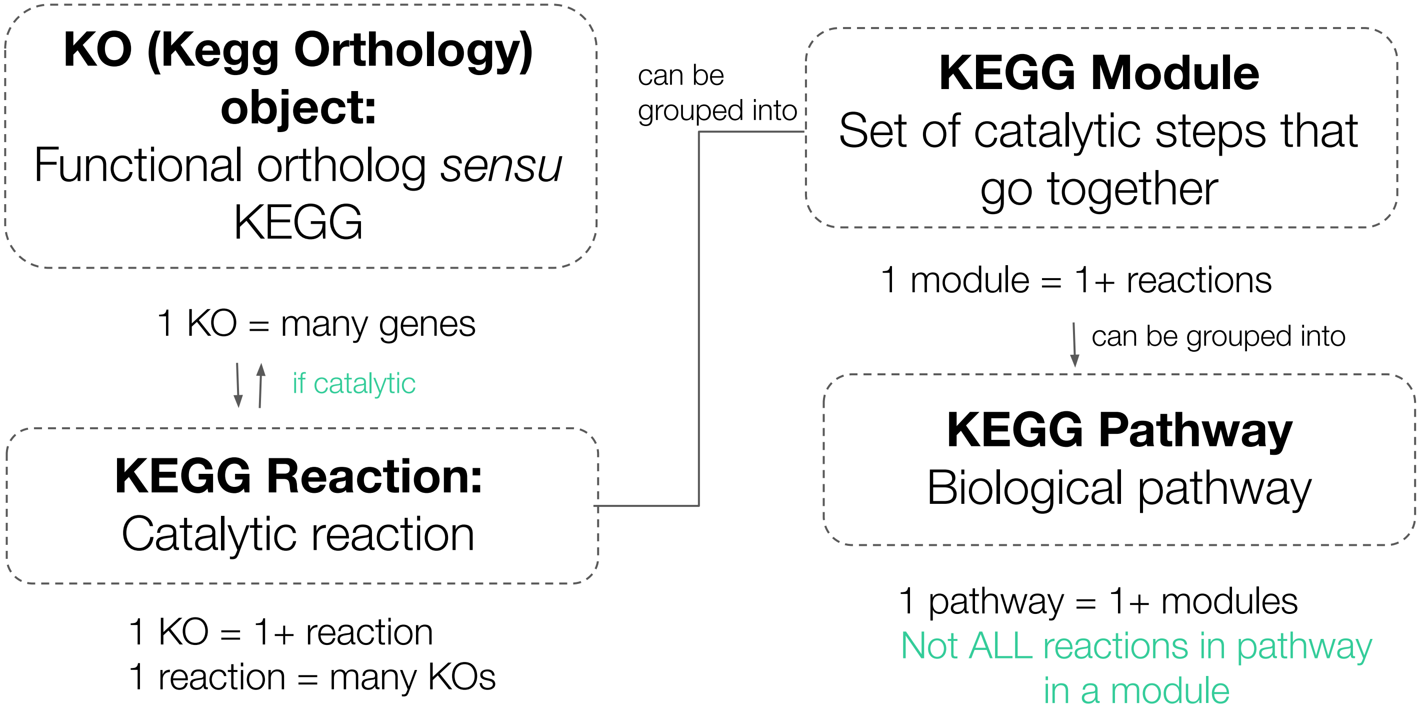
We can classify these gene families into units of functional "equivalence" and map them to known metabolic functions

KO (Kegg Orthology) object:
Functional ortholog *sensu* KEGG

Each module donates a distinct set of gene families with different functionalities



We can classify these gene families into units of functional “equivalence” and map them to known metabolic functions



KO (Kegg Orthology) object:
Functional ortholog *sensu* KEGG

1 KO = many genes

if catalytic

KEGG Reaction:
Catalytic reaction

1 KO = 1+ reaction
1 reaction = many KOs

can be grouped into

KEGG Module
Set of catalytic steps that go together

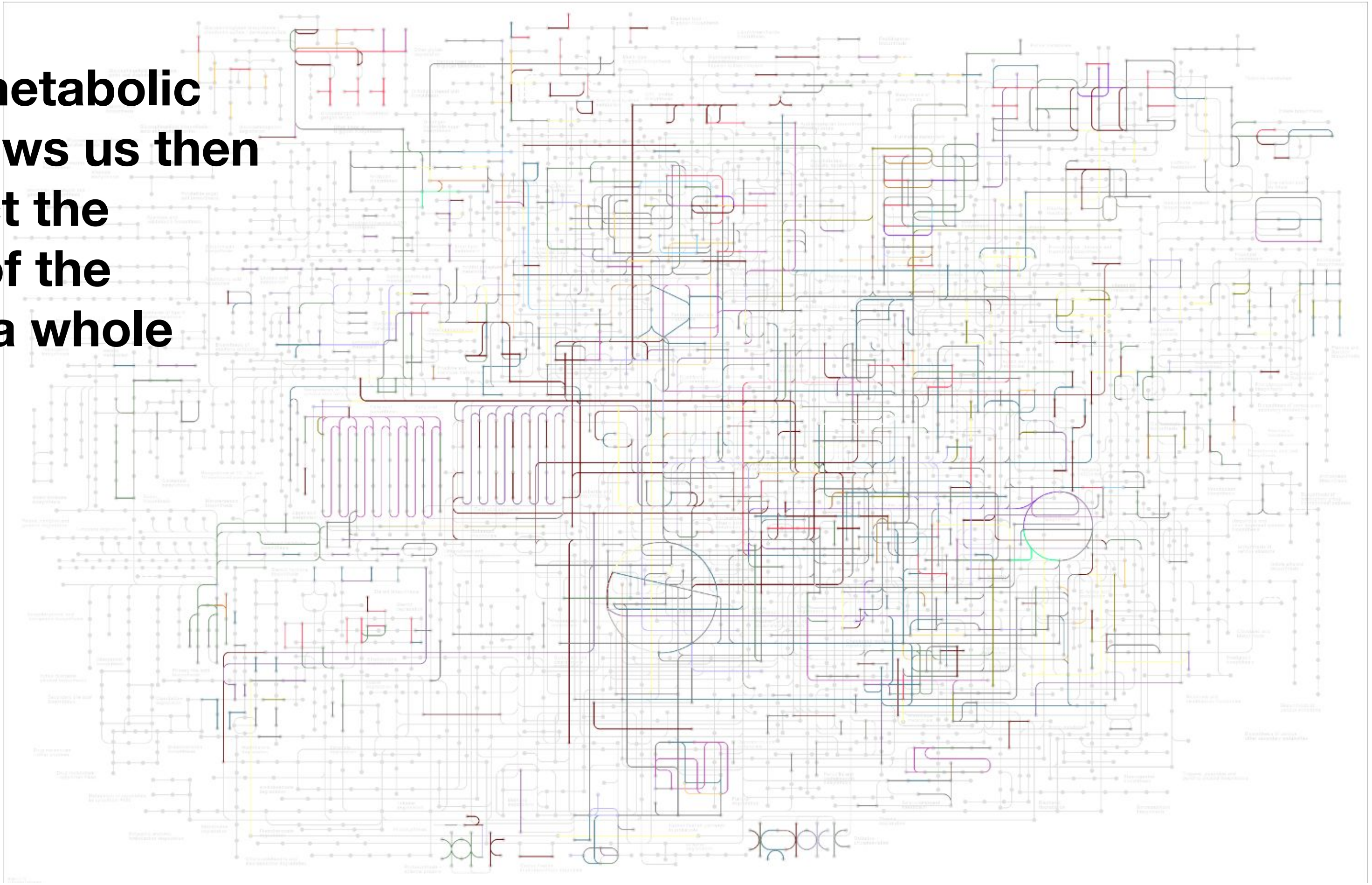
1 module = 1+ reactions

can be grouped into

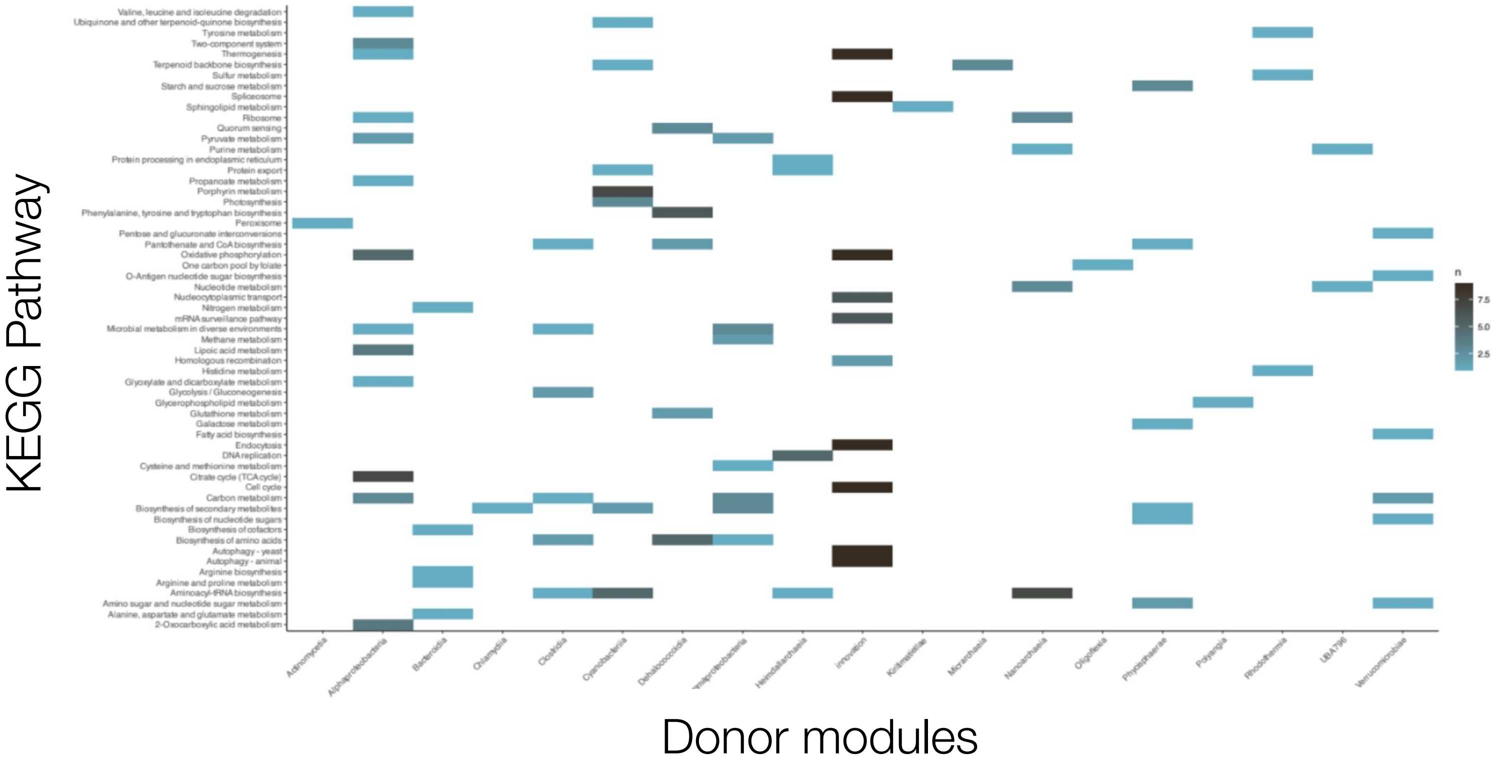
KEGG Pathway
Biological pathway

1 pathway = 1+ modules
Not ALL reactions in pathway in a module

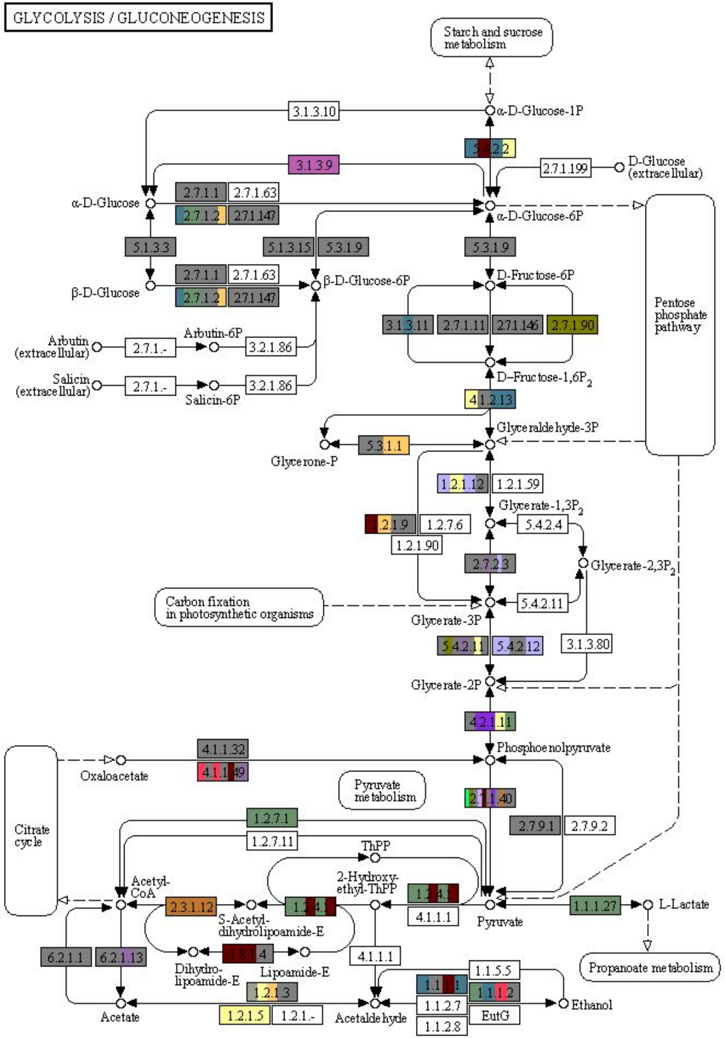
Mapping to metabolic functions allows us then to reconstruct the metabolism of the organism as a whole



Each donor module is enriched in different pathways and contributed to the proto-eukaryote in distinct metabolic abilities



Mapping to metabolic functions allows us then to reconstruct the metabolism of the organism as a whole



- 🟦 Gammaproteobacteria
- 🟫 Bacilli
- 🟥 Alphaproteobacteria
- 🟪 Verrucomicrobiae
- 🟩 Bacteroidia
- 🟨 Phycisphaerae
- 🟪 Clostridia
- 🟥 Planctomycetia
- 🟪 Polyangia
- 🟦 Thermoproteia
- 🟩 Cyanobacteriia
- 🟫 Chlamydiia
- 🟨 Actinomycetia
- 🟧 Heimdallarchaeia
- ⬜ Anaerolineae
- 🟩 Lokiarchaeia
- 🟪 Nanoarchaeia
- ⬛ Others

HPC allows us to automatize analysis of metabolic networks

KEGG Kyoto Encyclopedia of Genes and Genomes

```
download.kegg()
(pathview)
```

>

.xml

```
parseKGML()
(KEGGgraph)
```

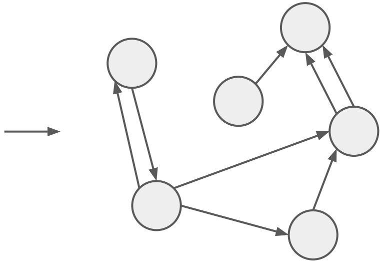
KGML Files

KO2Reaction

Reaction	EC	KO
R00014	1.2.4.1 2.2.1.6 4.1.1.1	K00161 pyruvate dehydrogenase E1 component subunit alpha K00162 pyruvate dehydrogenase E1 component subunit beta K00163 pyruvate dehydrogenase E1 component [EC:1.2.4.1] K01568 pyruvate decarboxylase [EC:4.1.1.1] K01652 acetolactate synthase I/II/III large subunit [EC:2.2.1.6] K01653 acetolactate synthase I/III small subunit [EC:2.2.1.6] K11258 acetolactate synthase II small subunit [EC:2.2.1.6]

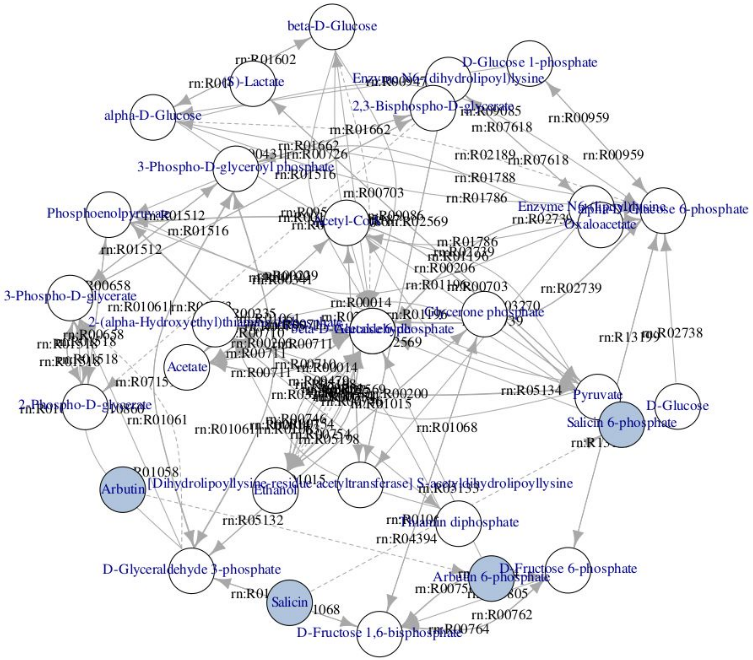
Reaction2 Metabolite

Reaction	Substate	Product
R00014	C00022 C00068	C05125 C00011



Pathway reaction Network

KO00010 (GLYCOLYSIS/GLUCONEOGENESIS)



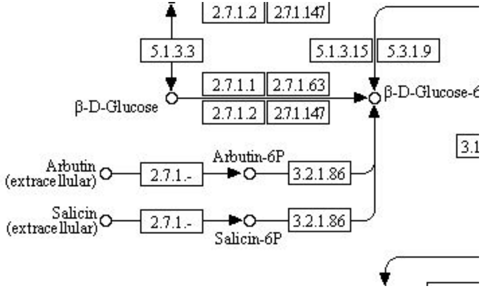
Missing reactions:

- rn:R02189
- rn:R02187
- rn:R05132
- rn:R04394
- rn:R09127
- rn:R07159
- rn:R09532
- rn:R09479
- rn:R09127

Completeness: 91.08%

Missing metabolites:

- Salicin
- Salicin-6P
- Arbutin
- Arbutin-6P



Take-home messages

1. Eukaryotes are a **mosaic** of genes of many prokaryotic donors, outside a simple alphaproteobacteria-Asgard endosymbiotic scenario

Take-home messages

1. Eukaryotes are a **mosaic** of genes of many prokaryotic donors, outside a simple alphaproteobacteria-Asgard endosymbiotic scenario
2. These **gene acquisitions** can be grouped into waves, discernible from the **topology** of the **gene trees**

Take-home messages

1. Eukaryotes are a **mosaic** of genes of many prokaryotic donors, outside a simple alphaproteobacteria-Asgard endosymbiotic scenario
2. These **gene acquisitions** can be grouped into waves, discernible from the **topology** of the **gene trees**
3. Each donor left **unique functionalities** into the proto-eukaryote, that correspond to **distinct metabolic pathways**

Take-home messages

1. Eukaryotes are a **mosaic** of genes of many prokaryotic donors, outside a simple alphaproteobacteria-Asgard endosymbiotic scenario
2. These **gene acquisitions** can be grouped into waves, discernible from the **topology** of the **gene trees**
3. Each donor left **unique functionalities** into the proto-eukaryote, that correspond to **distinct metabolic pathways**
4. This will allow us to relatively time the origin of the complex features that differentiate prokaryotes and eukaryotes

Thank you

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

GORDON AND BETTY MOORE FOUNDATION

EXCELENCIA
SEVERO
OCHOA

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

11th BSC PhD Symposium

Reconstructing prokaryotic metabolic contributions to LECA

Saioa Manzano-Morales
Comparative Genomics - Life Sciences