

Systematics, Evolution & Genomics

Learning From Genomes: Comparative Genomic Approaches to Investigate Thylakoid Emergence in Cyanobacteria

Louise Hambücken, Denis Baurain and Luc Cornet
University of Liège

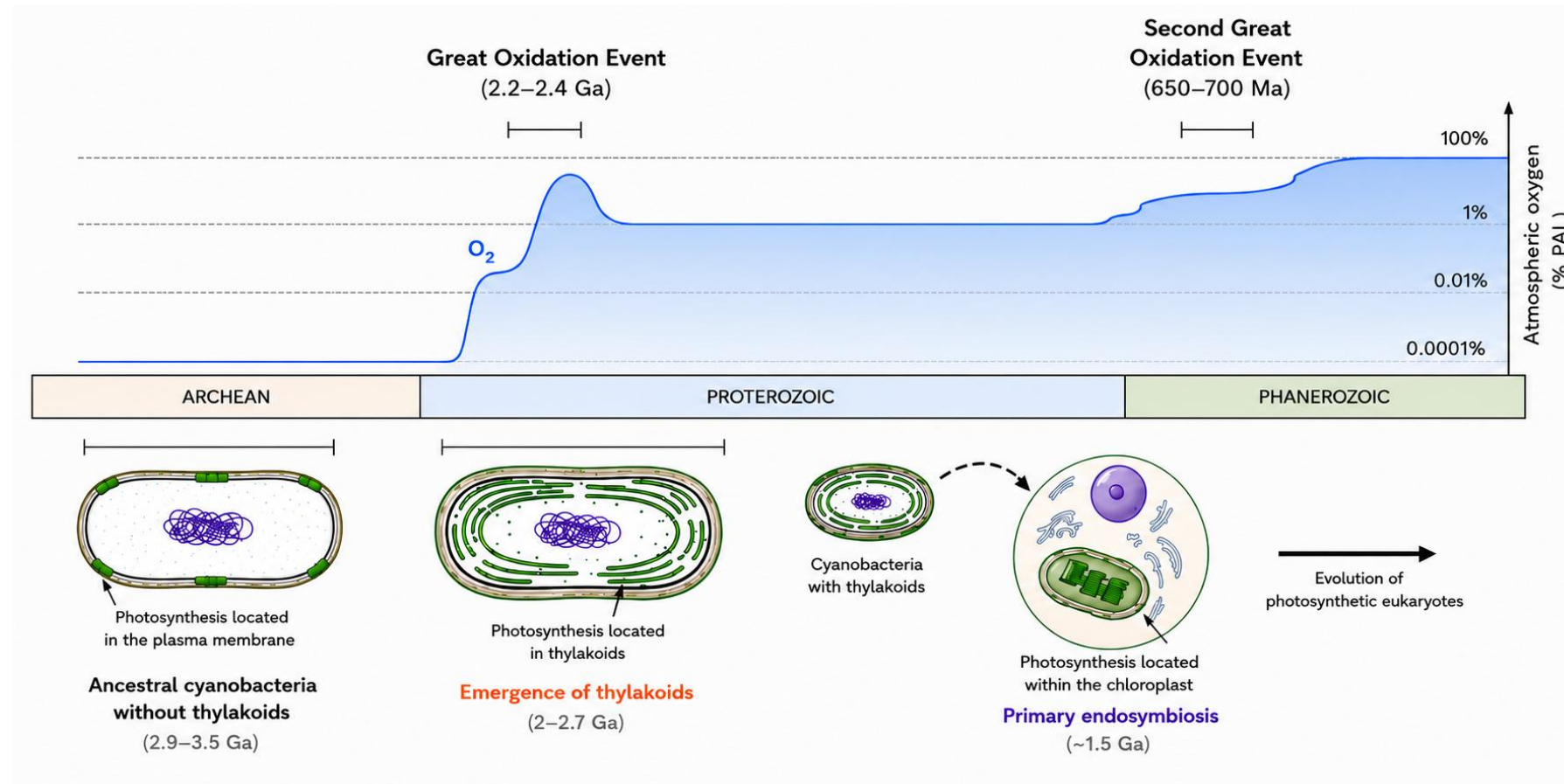
August 3rd, 2026

23rd Association for Cyanophyte/Cyanobacteria Research Symposium



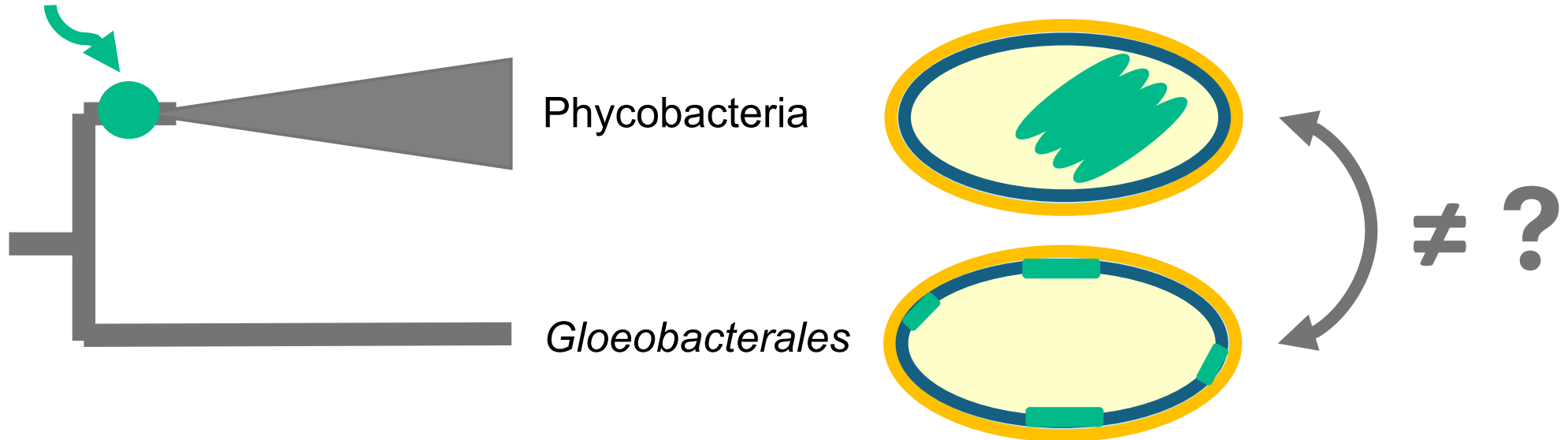
Why do thylakoids matter?

The **emergence of thylakoids** in cyanobacteria likely contributed to the **Great Oxidation Event**



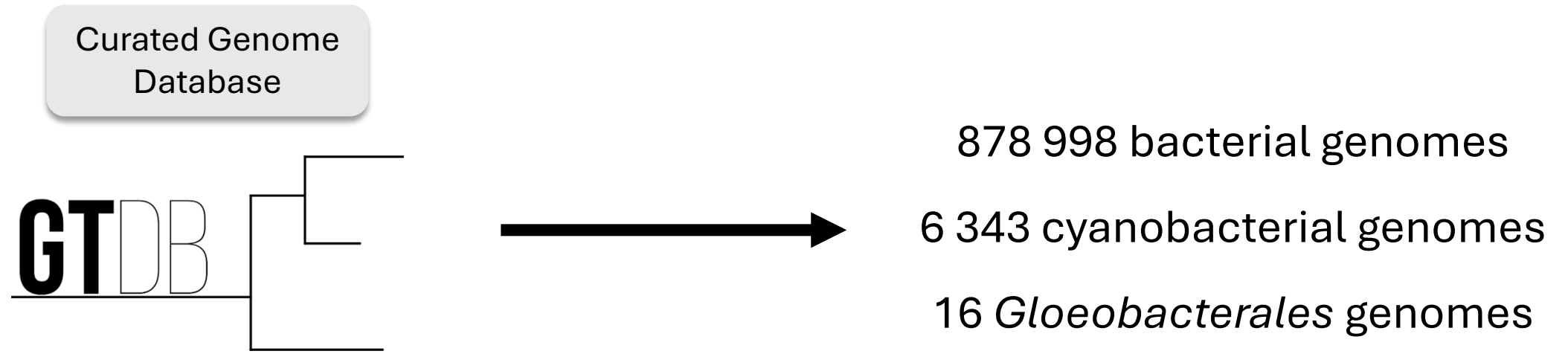
Cyanobacteria provide a natural comparative framework

Thylakoid Emergence
(~2 -2.7 Ga)



What are the key differences in their genomes?

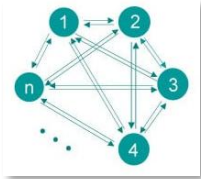
Public genomes offer an unprecedented sampling



Public genome databases allow us to investigate thylakoid evolution across cyanobacterial diversity.

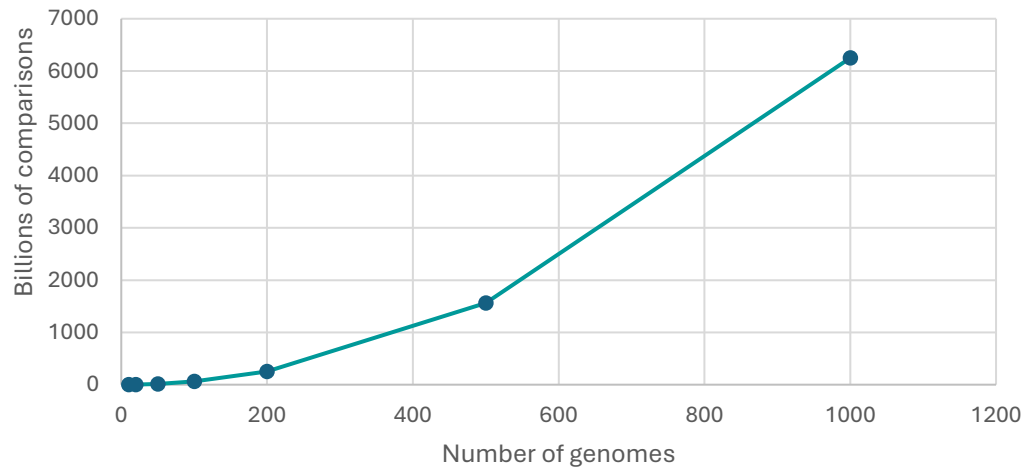
Why not analyse all genomes directly?

Too Many Genomes, Too Little Power



Comparison
All vs All protein sequences
(assuming ~2500 proteins per genome)

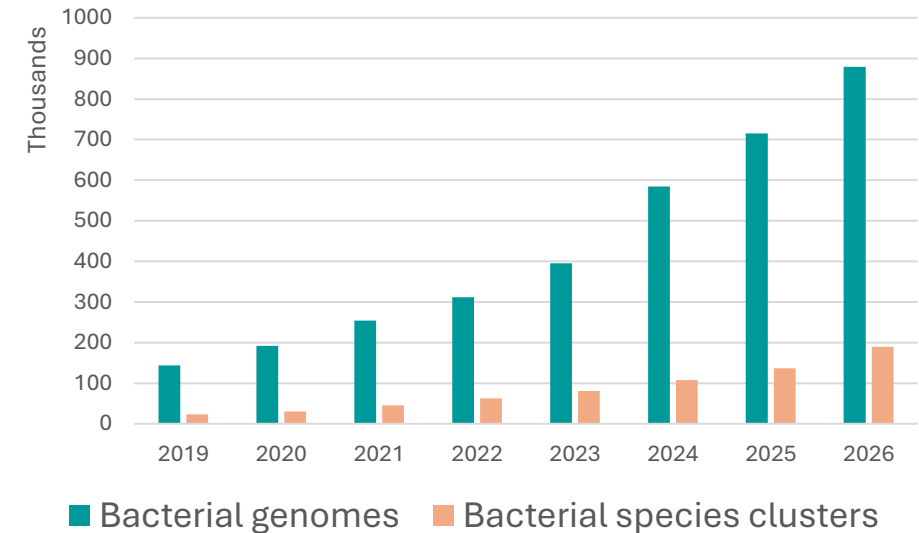
In order to use OrthoFinder:



N Genomes $\rightarrow (N \times 2500)^2$ Comparisons

The Illusion of Diversity

Evolution of bacterial genomes and bacterial species clusters between 2019 and 2026

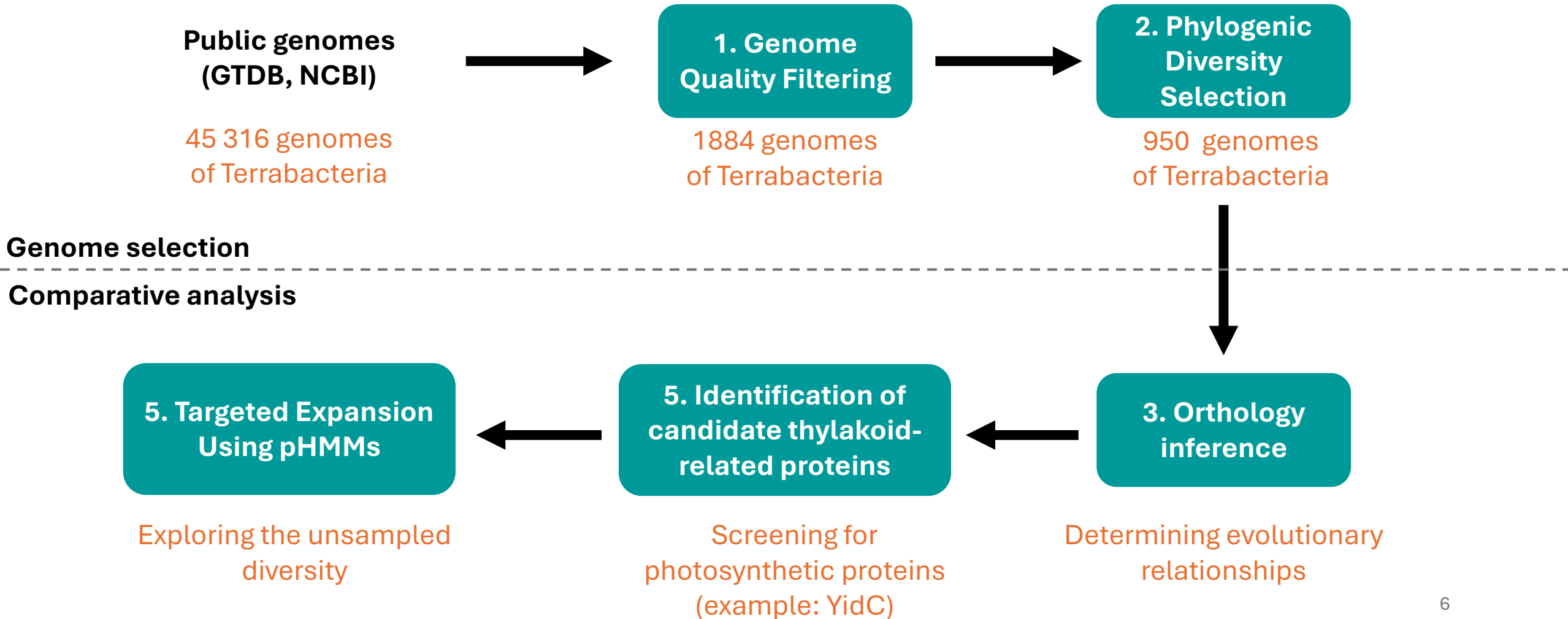


Source: <https://gtadb.ecogenomic.org/>

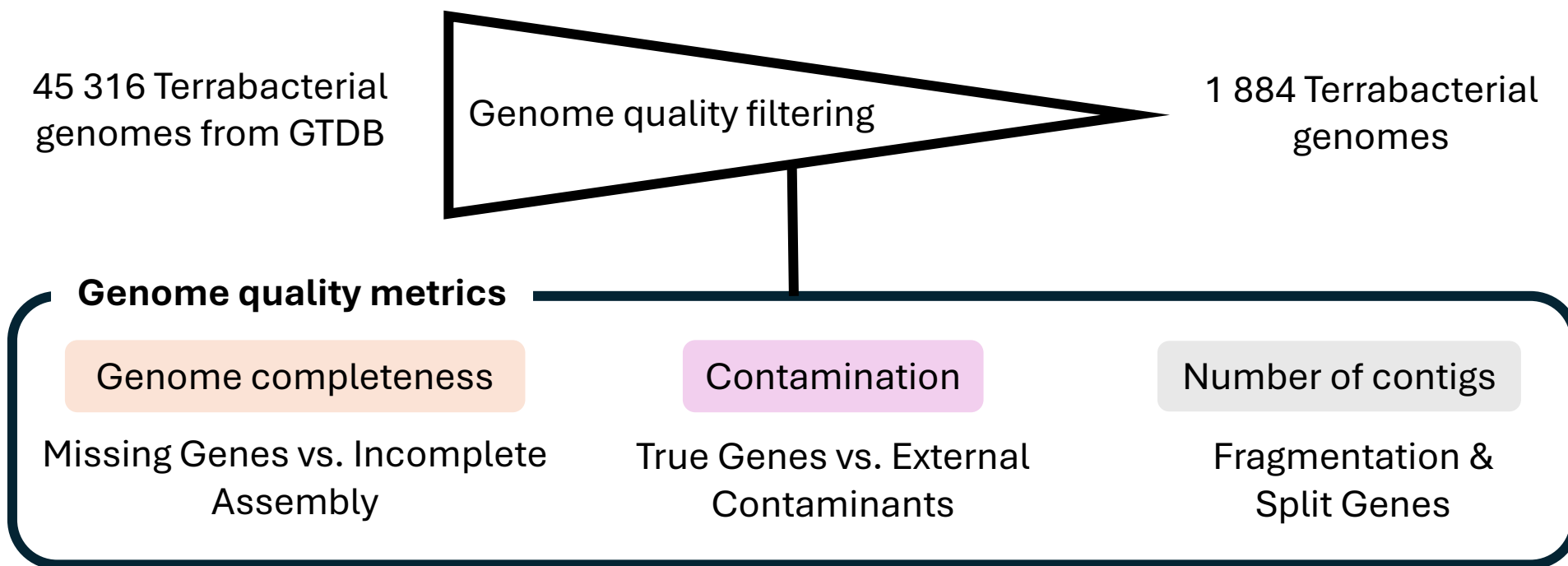
Adding Genomes, Repeating the Data

The challenge is not simply dataset size, but the ratio between phylogenetic information and redundancy.

From genomic data curation to downstream analysis



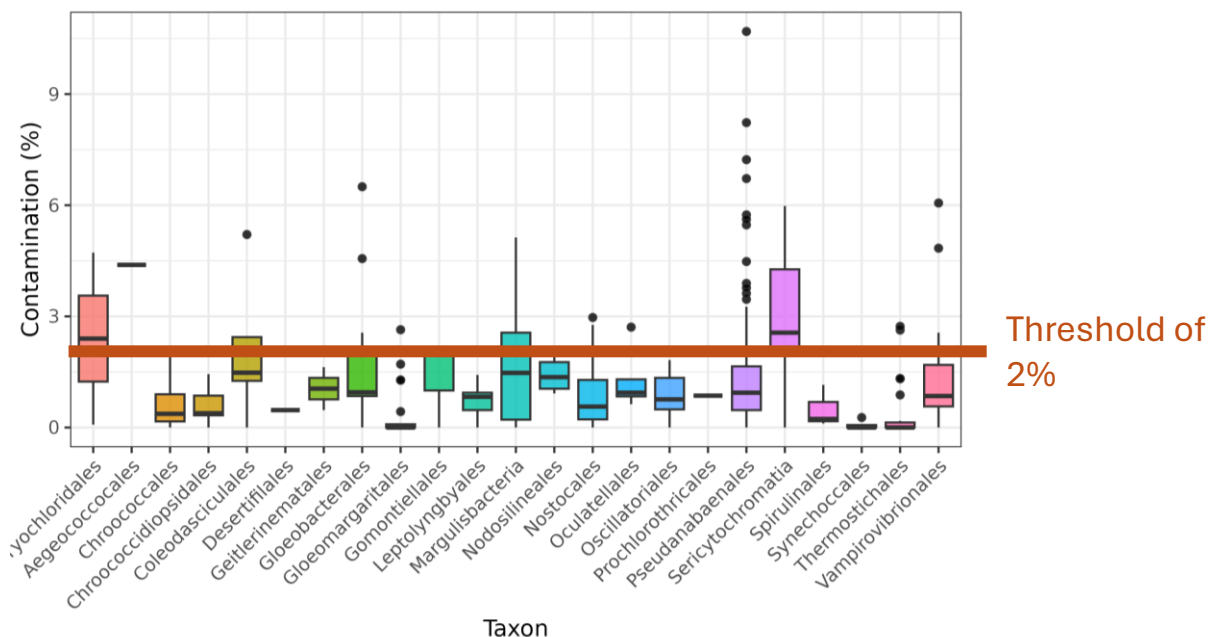
Filtering genomes without filtering evolutionary novelty



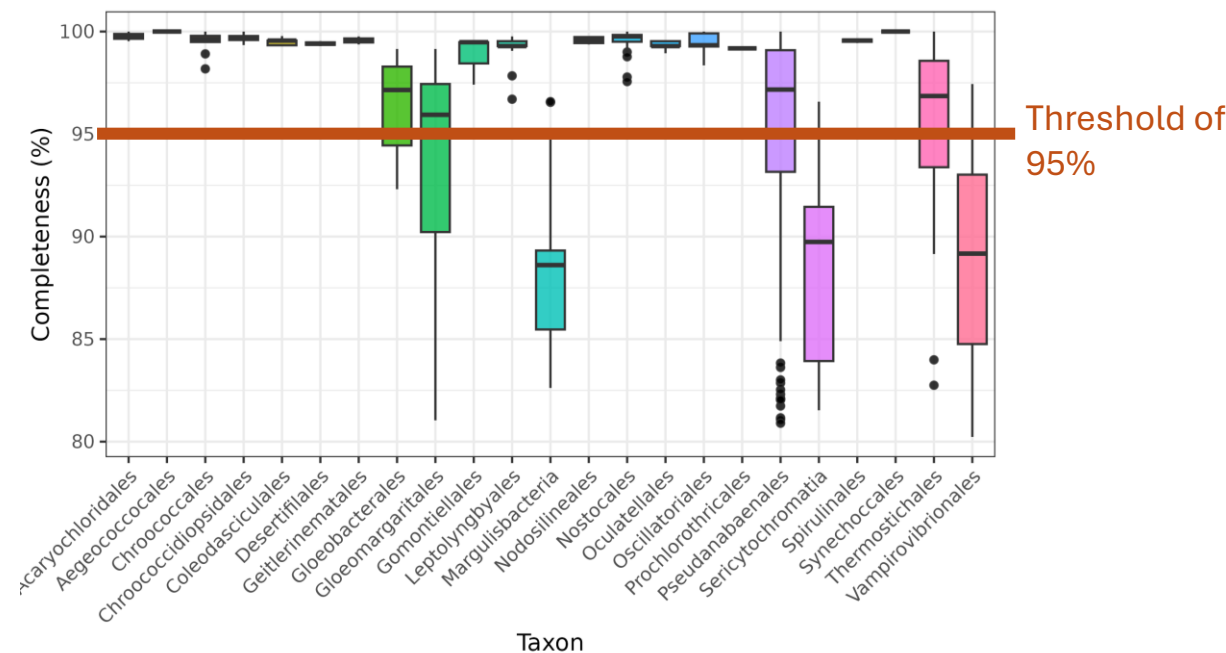
$$\text{Quality Score} = \text{Genome Completeness} - 5 \times \text{Genome contamination}$$

Genome quality is heterogeneous

Genome contamination (CheckM2) per cyanobacteriota order



Genome completeness (CheckM2) per cyanobacteriota order

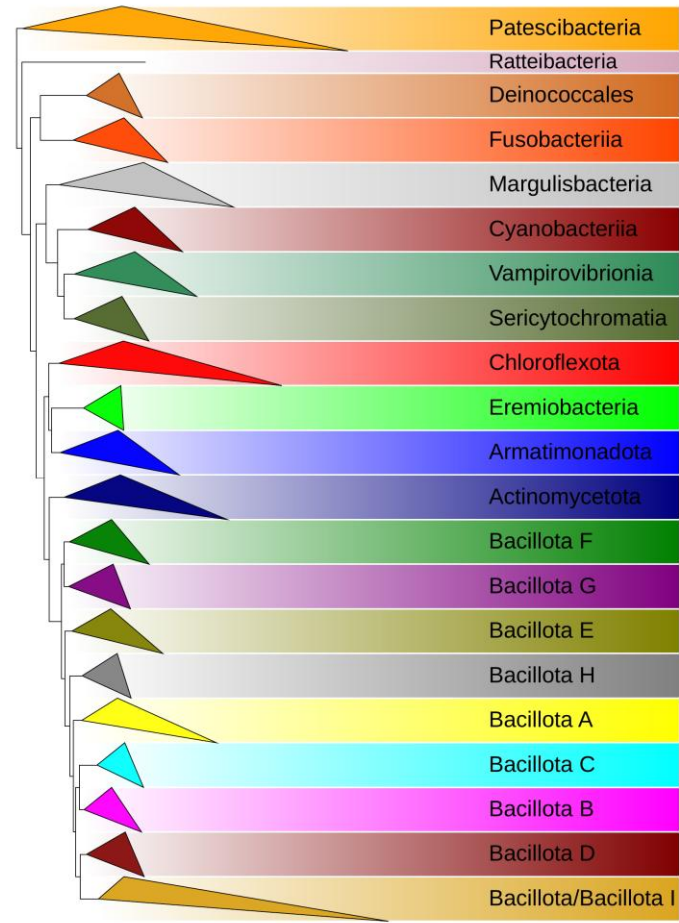


Selecting one representative genome per order based on a quality score

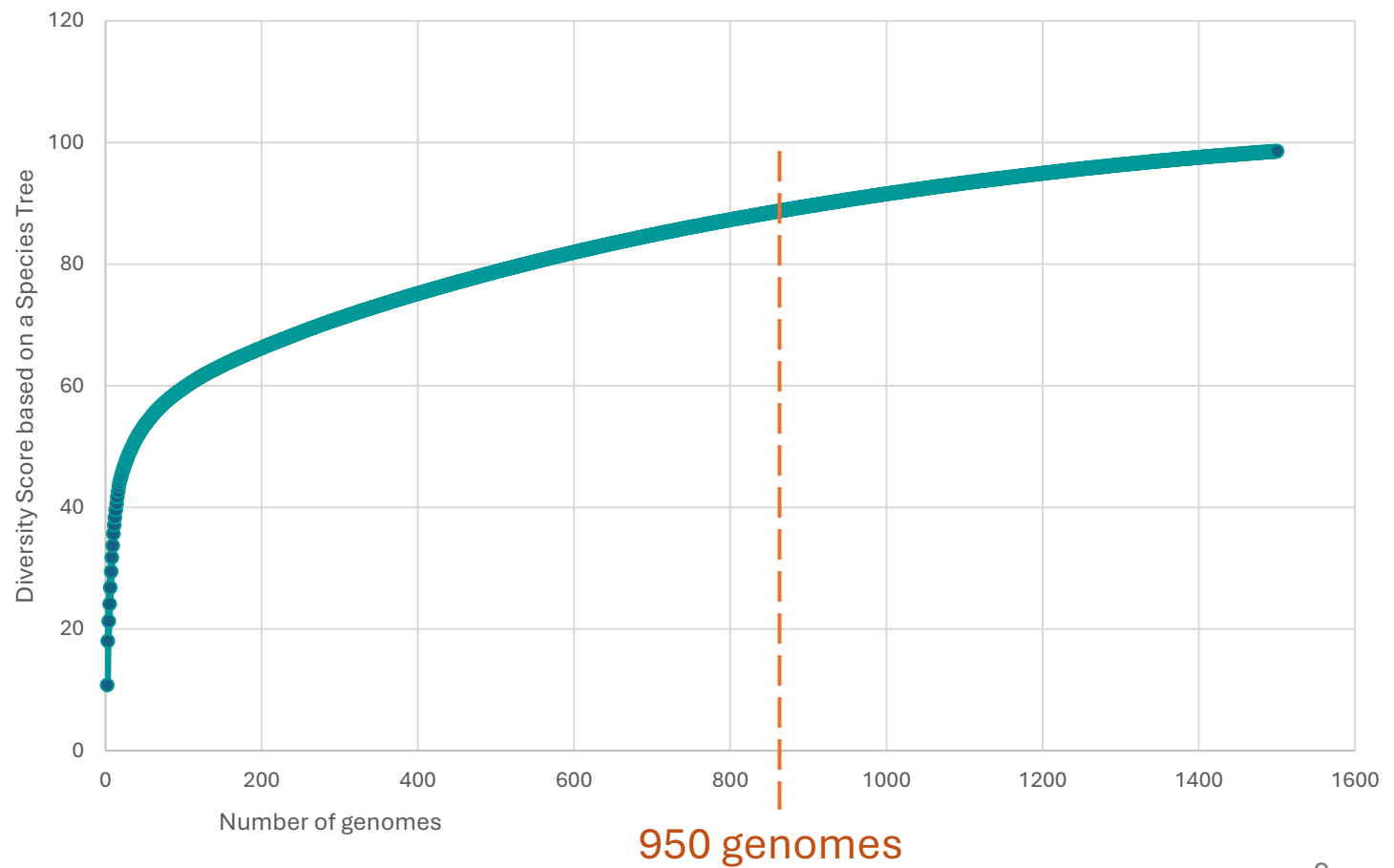
$$\text{Quality Score} = \text{Genome Completeness} - 5 \times \text{Genome contamination}$$

Retaining phylogenetic diversity

Ribosomal species tree using the
filtered 1884 Terrabacterial genomes

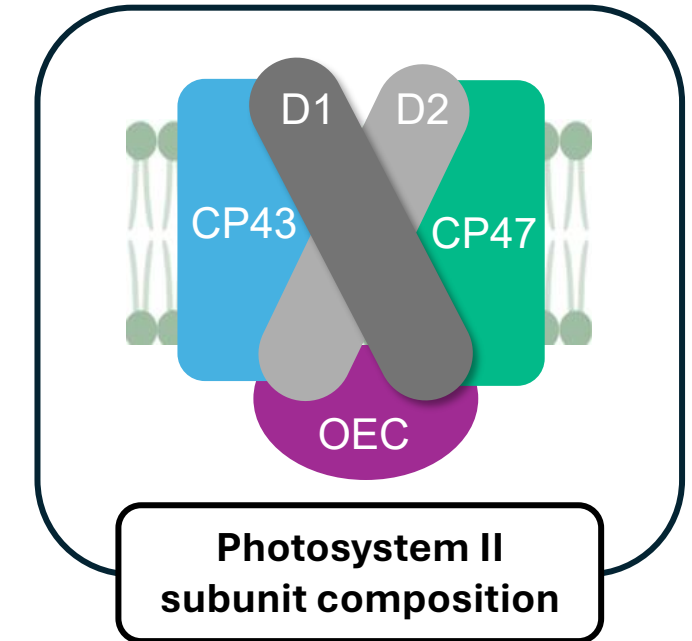
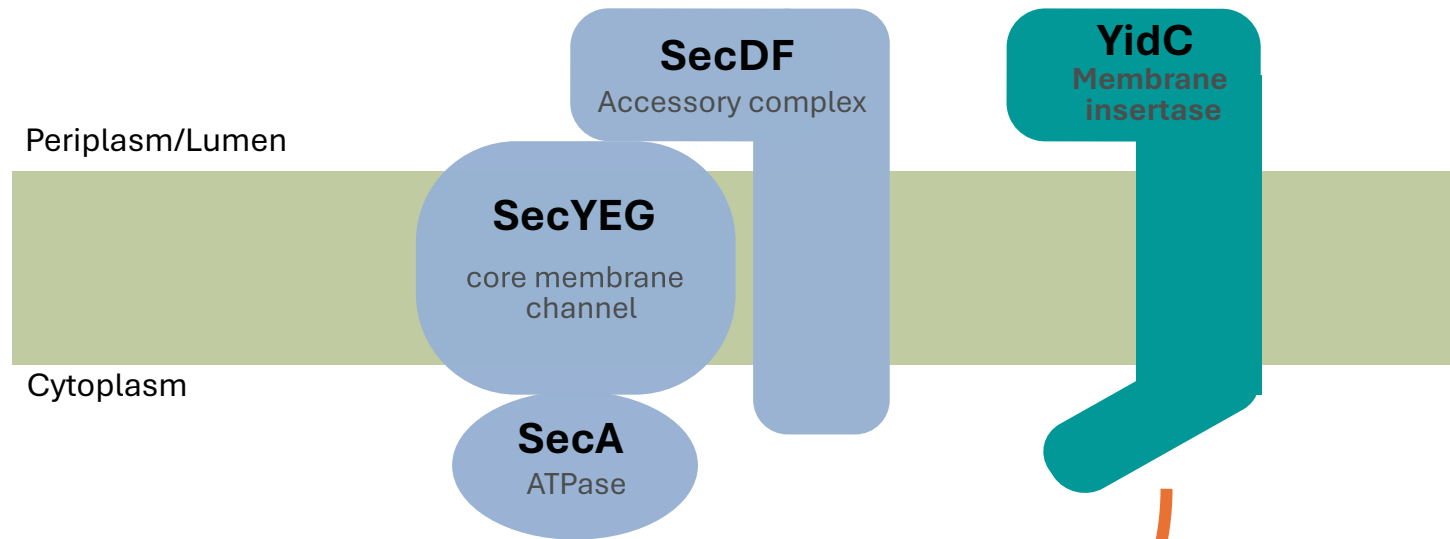


Diversity curve for the species tree of 1800
Terrabacterial genomes using PARNAS



Orthologous Inference

950 genomes → Filtering Orthologous Groups (OGs) for downstream analysis

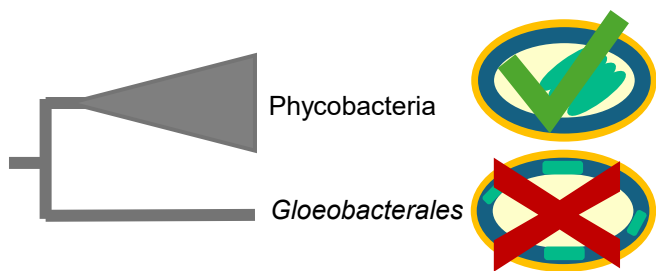


A case study:

YidC in the protein insertion of D1 into the membrane

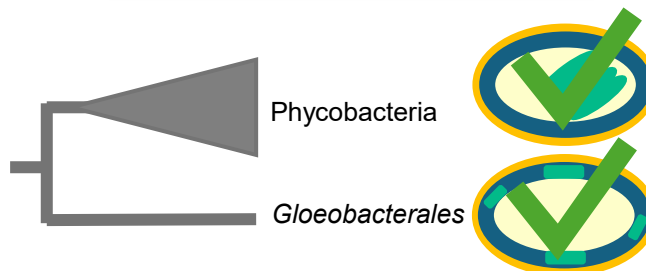
Comparative genomics identifies candidate proteins in thylakoid emergence

What's missing?



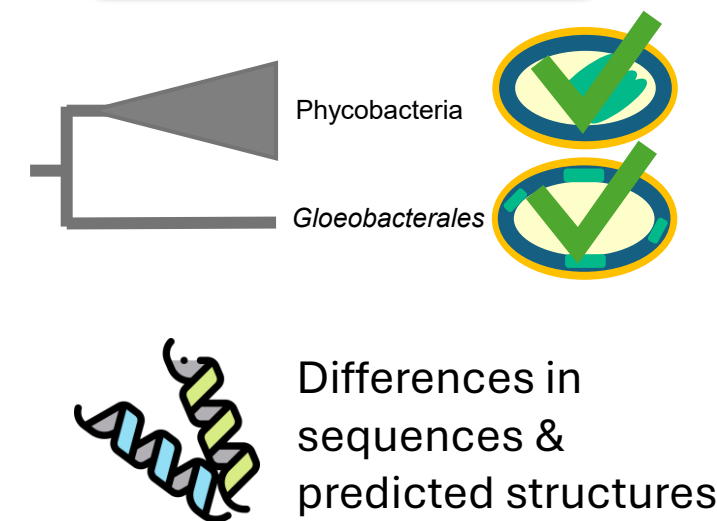
Some proteins involved in
membrane dynamics

What's conserved?



Most early proteins
mediating PSII assembly

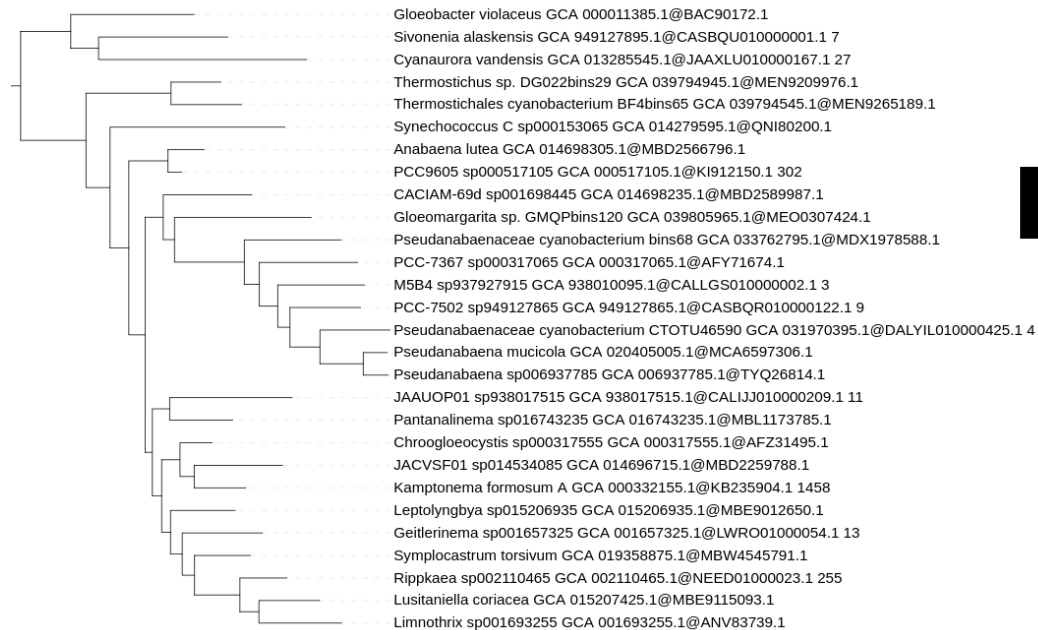
What changes?



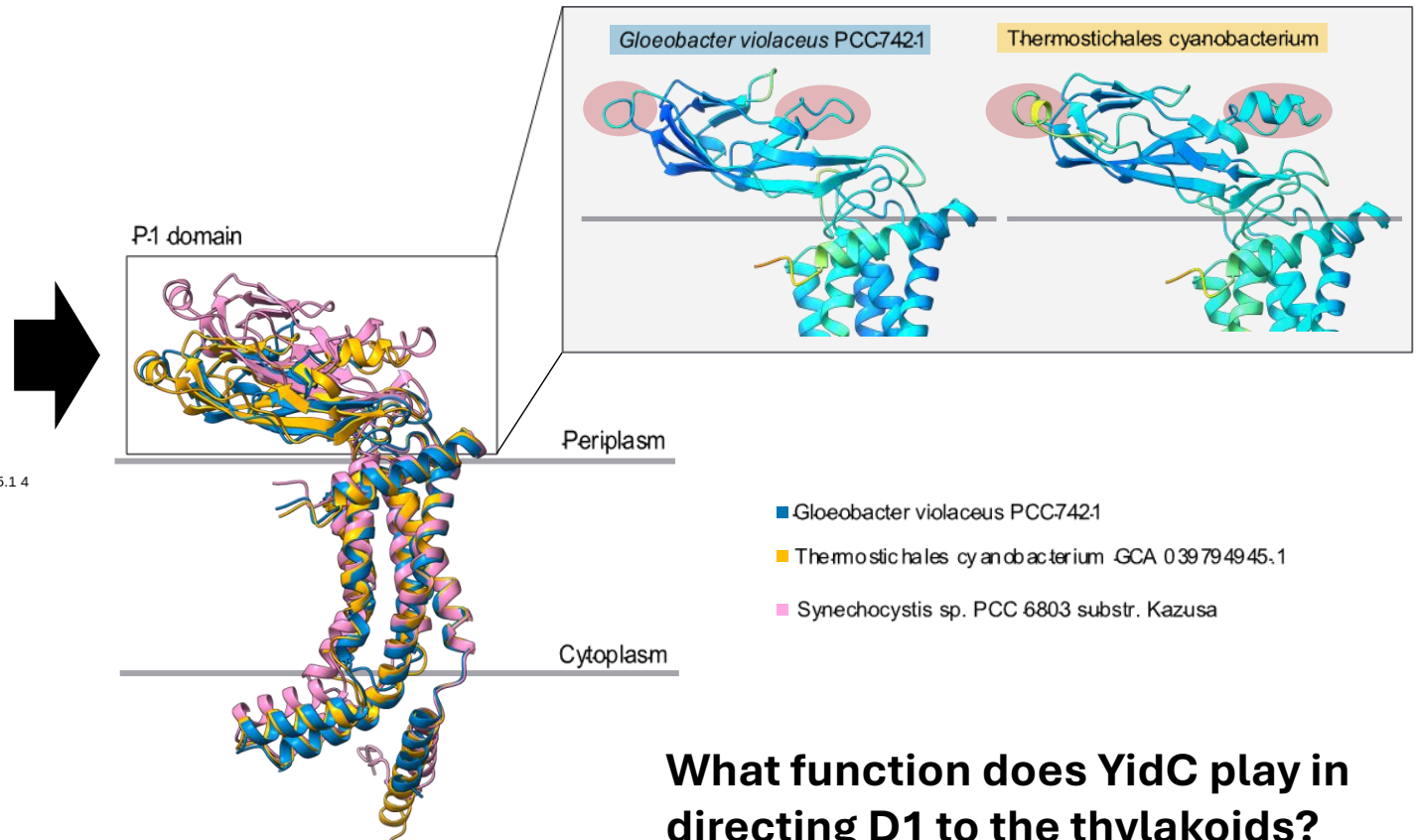
YidC

Comparative genomics identifies candidate proteins in thylakoid emergence

YidC is universally conserved across the cyanobacterial phylum



Structural changes in the periplasmic domain of YidC in *Gloeobacterales*



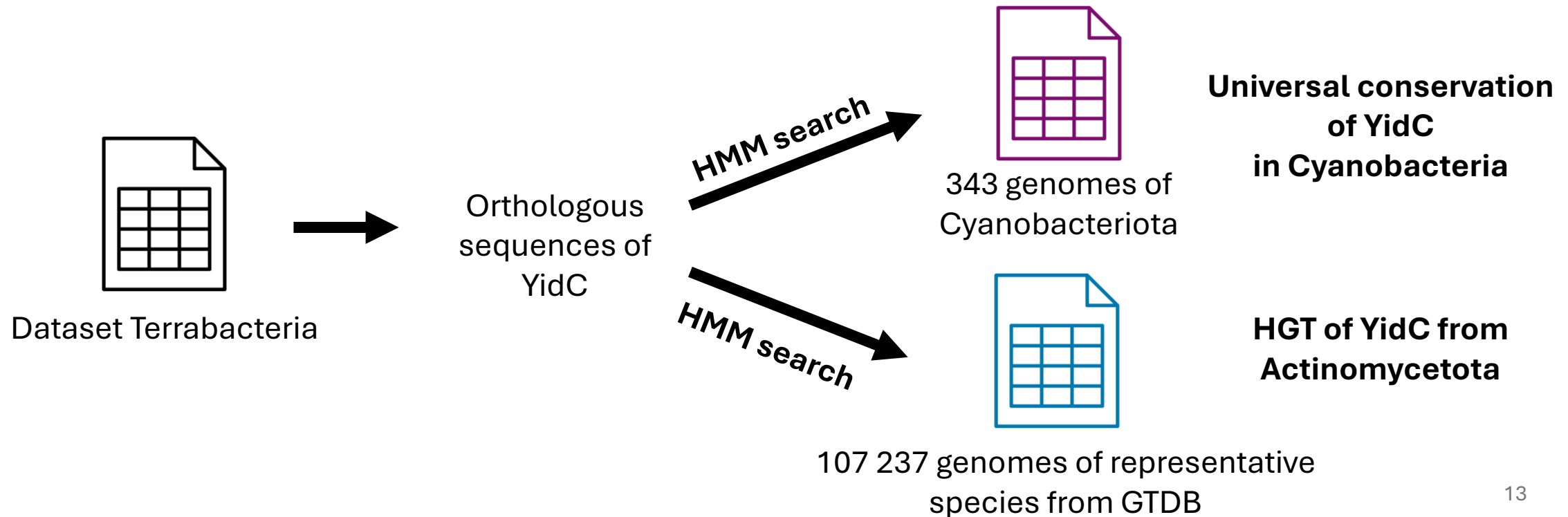
What function does YidC play in directing D1 to the thylakoids?

Expanding the sampling with profile HMMs

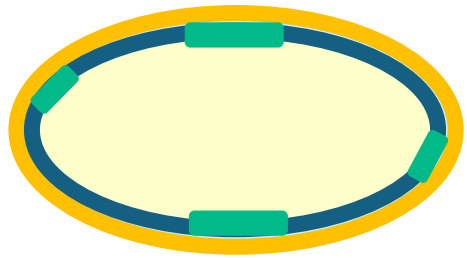
950 Terrabacterial genomes including 30 cyanobacterial genomes



Are my conclusions true for the whole cyanobacterial diversity ?
How to detect horizontal gene transfer (HGT)?

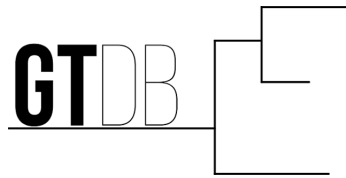


Learning from genomes: from data volume to evolutionary insight



Tracking Evolutionary
Events Through
Comparative Patterns

1 Public Genomes

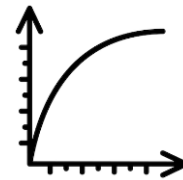


A Window into
Thylakoid Origins

2 Quality Over Quantity



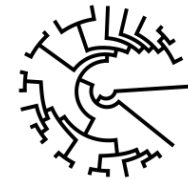
Quality
filtering



Redundancy
elimination

The Crucial Cleanup
Phase

3 A Scalable Framework



Orthologous
sequences



Data enrichment
with HMM
searches

Cyanobacteriota

Orthology & Hidden Markov
Models

Genome discovery relies on biologically informed selection first, followed by HMM-based expansion to large-scale datasets.

Acknowledgements



Eukaryotic
Phylogenomics
Laboratory

Supervisors:

Luc Cornet
Denis Baurain

BASAL Team:

Benoit Durieu
Jeanne Matisse
Valentina Savaglia
Edi Sudianto

Other lab members:

Mathieu Bruyneel
Marie Harmel
Valérien Lupo
Mick Van Vlierberghe

Eukaryotic Phylogenomics Lab
University of Liège

Financements:

FRIA – F.R.S.-FNRS funding