genomics
aotearoa

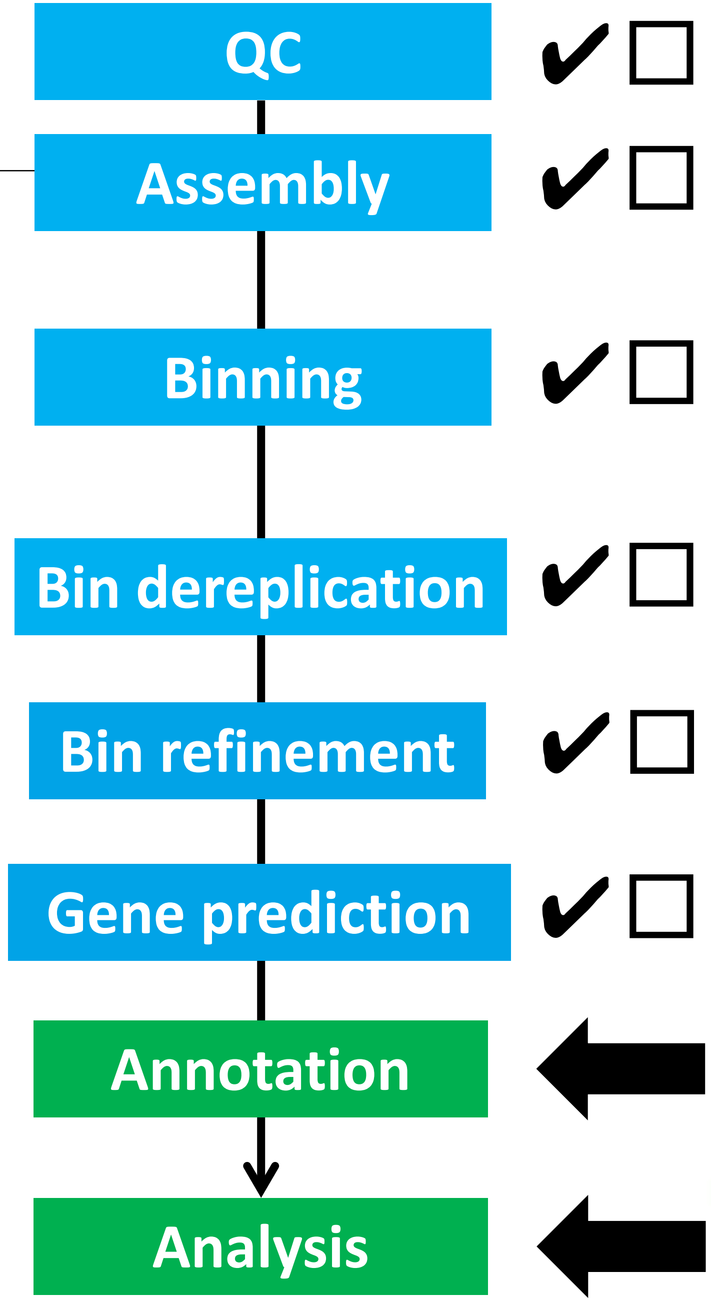
Metagenomics
Summer School 2019

Day 4

Genome annotation (continued)
Report findings
Presentation of data

Day overview

- **Goals:**
  - **Gene annotations (continued)**
  - **Presentation of group findings**
  - **Visualizing metagenomic data**



Genome annotation (continued)

Bin taxonomic classification

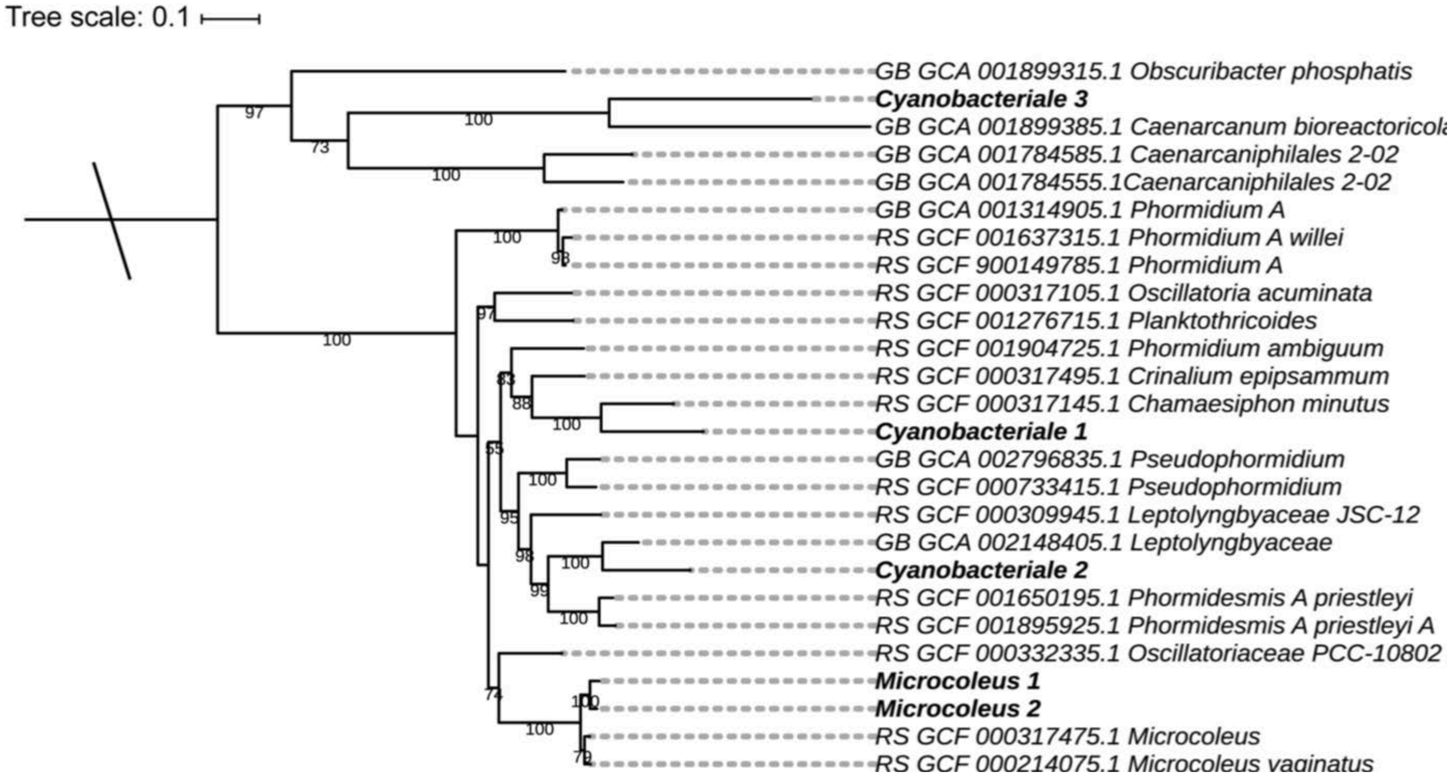
- **16S rRNA commonly not recovered by *de novo* assembly**

- **Can recover 16S and 18S using EMIRGE**
- **Caveat: can be difficult to assign to genomes in complex communities with many similar taxa**

Bin taxonomic classification

Solution:

- **Use one or more single copy core genes**
- **Concatenate protein sequences of multiple single copy core genes**



Tree scale: 0.1
GB GCA 001899315.1 Obscuribacter phosphatis
Cyanobacteriale 3
GB GCA 001899385.1 Caenarcanum bioreactoricola
GB GCA 001784585.1 Caenarcaniphilales 2-02
GB GCA 001784555.1Caenarcaniphilales 2-02
GB GCA 001314905.1 Phormidium A
RS GCF 001637315.1 Phormidium A willei
RS GCF 900149785.1 Phormidium A
RS GCF 000317105.1 Oscillatoria acuminata
RS GCF 001276715.1 Planktothricoides
RS GCF 001904725.1 Phormidium ambiguum
RS GCF 000317495.1 Crinalium epipsammum
RS GCF 000317145.1 Chamaesiphon minutus
Cyanobacteriale 1
GB GCA 002796835.1 Pseudophormidium
RS GCF 000733415.1 Pseudophormidium
RS GCF 000309945.1 Leptolyngbyaceae JSC-12
GB GCA 002148405.1 Leptolyngbyaceae
Cyanobacteriale 2
RS GCF 001650195.1 Phormidesmis A priestleyi
RS GCF 001895925.1 Phormidesmis A priestleyi A
RS GCF 000332335.1 Oscillatoriaceae PCC-10802
Microcoleus 1
Microcoleus 2
RS GCF 000317475.1 Microcoleus
RS GCF 000214075.1 Microcoleus vaginatus
97
73
100
100
100
98
100
97
83
88
100
55
100
95
98
100
99
100
74
100
100
79

Concatenated protein sequence tree: Phylogenetic placement of cyanobacterial genome bins (Wai-iti River, Nelson)

Bin taxonomic classification

- **Genome Taxonomy Database Toolkit (GTDB-Tk)**
- **Use to classify bins against reference genome trees (GTDB)**

Welcome to GTDB

GENOME TAXONOMY DATABASE

145,904 genomes

Release 04-RS89 (19th June 2019)

- **Uses set of 120 concatenated protein sequences (of single copy core genes)**

https://github.com/Ecogenomics/GtdbTk
https://gtdb.ecogenomic.org/

Rank assignment based on:

- Tree placement
- Relative Evolutionary Divergence (value between 0=root and 1=tip)
- Species assignment:
  - Average Nucleotide Identity

Discriminate species

Proxy for DNA-DNA hybridization

Pairwise genome comparisons:

- **Average Nucleotide Identities (ANI)**
  - **gene comparisons**
- **Average Amino Acid Identities (AAI)**
  - **predicted protein comparisons**
- **Alignable Fraction (AF)**
  - **proportion of genes that align**

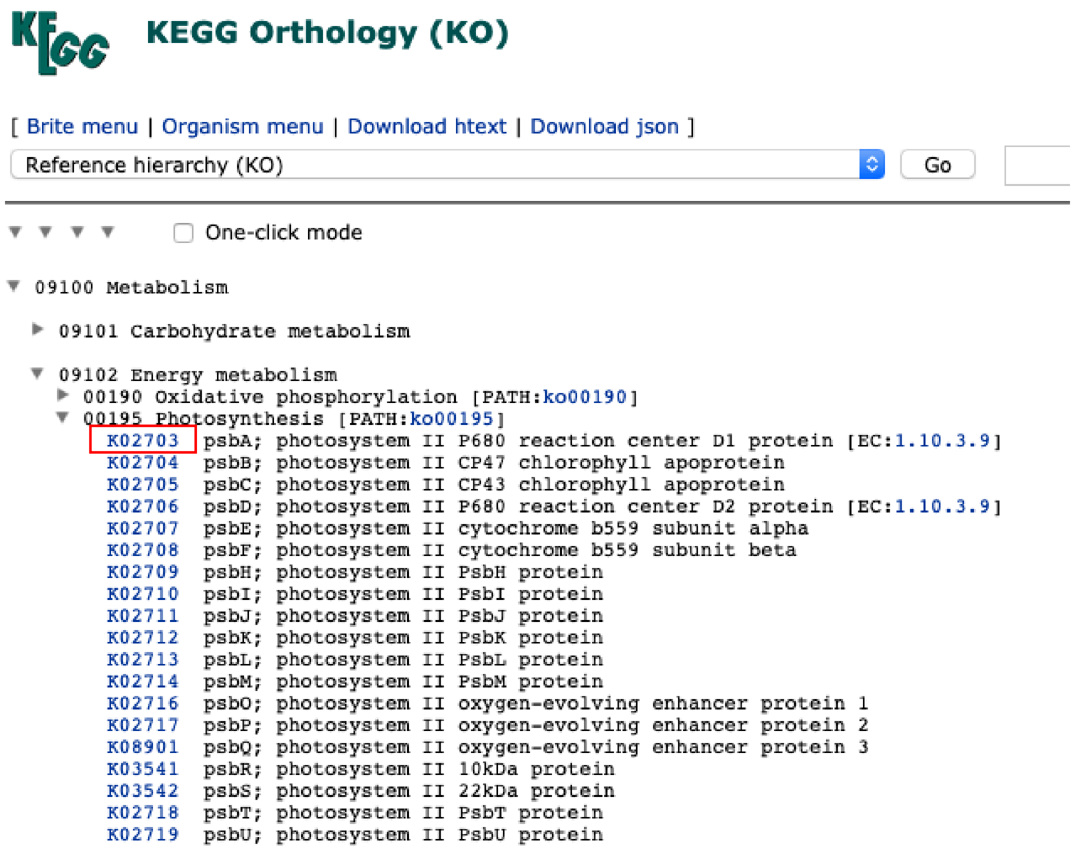
Determine via: Pairwise BLAST-like search

Phylogenetic trait distributions

- Interactive phylogenetic and trait based tree
- Annotree (http://annotree.uwaterloo.ca/app/)
- Trait searches by:
  - Taxonomic hierarchy
  - KEGG (KO number)
  - Pfam
  - TIGRFAM

Phylogenetic trait distributions

Get KEGG KO number from the KEGG website or your annotations



KEGG Orthology (KO)

[Brite menu | Organism menu | Download htext | Download json]

Reference hierarchy (KO) Go

□ One-click mode

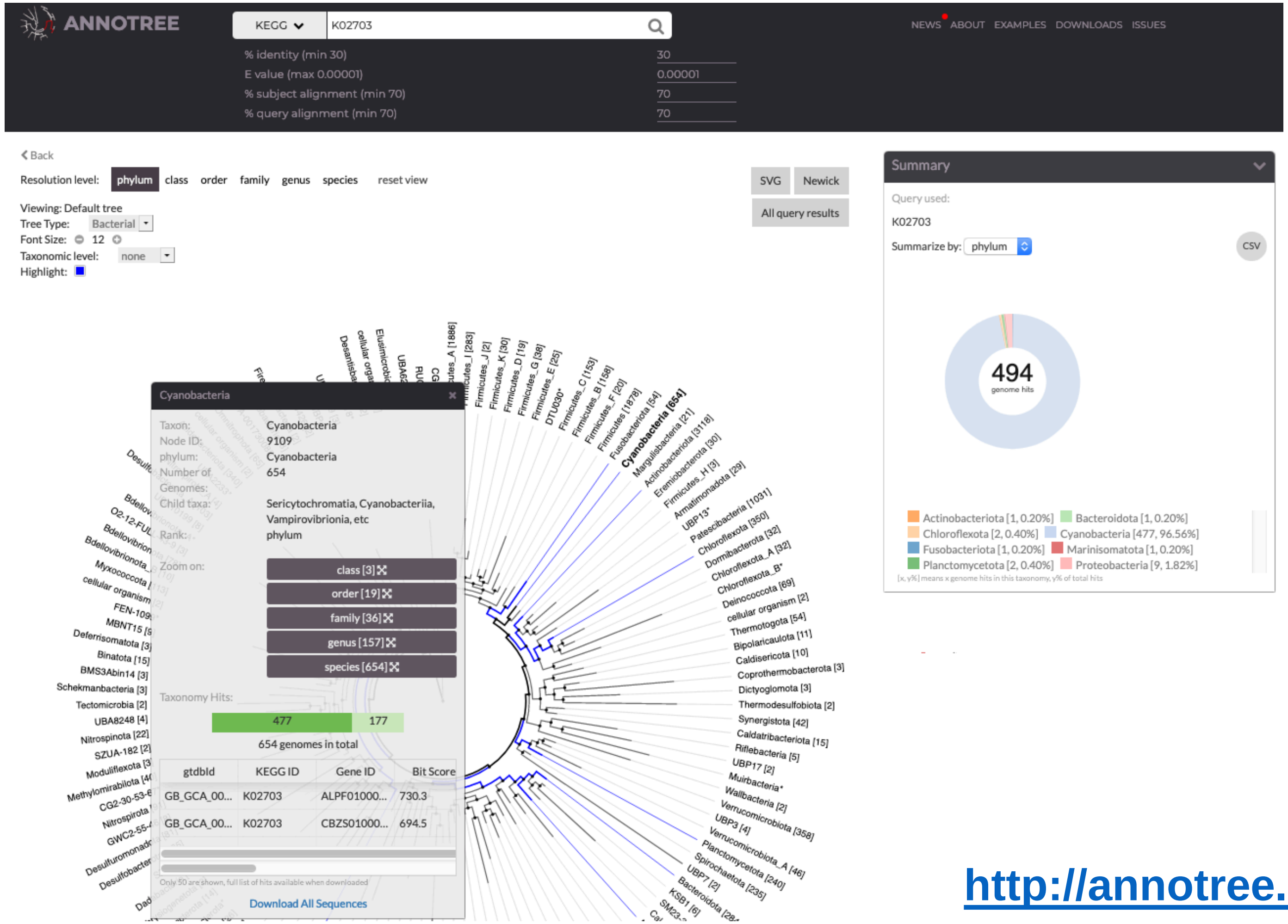
```
▼ 09100 Metabolism
  ▶ 09101 Carbohydrate metabolism
  ▼ 09102 Energy metabolism
    ▶ 00190 Oxidative phosphorylation [PATH:ko00190]
    ▼ 00195 Photosynthesis [PATH:ko00195]
        K02703  psbA; photosystem II P680 reaction center D1 protein [EC:1.10.3.9]
        K02704  psbB; photosystem II CP47 chlorophyll apoprotein
        K02705  psbC; photosystem II CP43 chlorophyll apoprotein
        K02706  psbD; photosystem II P680 reaction center D2 protein [EC:1.10.3.9]
        K02707  psbE; photosystem II cytochrome b559 subunit alpha
        K02708  psbF; photosystem II cytochrome b559 subunit beta
        K02709  psbH; photosystem II PsbH protein
        K02710  psbI; photosystem II PsbI protein
        K02711  psbJ; photosystem II PsbJ protein
        K02712  psbK; photosystem II PsbK protein
        K02713  psbL; photosystem II PsbL protein
        K02714  psbM; photosystem II PsbM protein
        K02716  psbO; photosystem II oxygen-evolving enhancer protein 1
        K02717  psbP; photosystem II oxygen-evolving enhancer protein 2
        K08901  psbQ; photosystem II oxygen-evolving enhancer protein 3
        K03541  psbR; photosystem II 10kDa protein
        K03542  psbS; photosystem II 22kDa protein
        K02718  psbT; photosystem II PsbT protein
        K02719  psbU; photosystem II PsbU protein
```

K02703

https://www.genome.jp/kegg-bin/get_htext#C17

Phylogenetic trait distributions

Add to ANNOTREE search box and select hierarchy



KEGG: K02703

% identity (min 30)	30
E value (max 0.00001)	0.00001
% subject alignment (min 70)	70
% query alignment (min 70)	70

Resolution level: phylum class order family genus species reset view

Viewing: Default tree

Tree Type: Bacterial

Font Size: 12

Taxonomic level: none

Highlight: 🟦

Cyanobacteria

Taxon:	Cyanobacteria
Node ID:	9109
phylum:	Cyanobacteria
Number of Genomes:	654
Child taxa:	Sericytochromatia, Cyanobacteriia, Vampirovibrionia, etc
Rank:	phylum
Zoom on:	class [3] order [19] family [36] genus [157] species [654]

Taxonomy Hits: 477 177

654 genomes in total

gtdbId	KEGG ID	Gene ID	Bit Score
GB_GCA_00...	K02703	ALPF01000...	730.3
GB_GCA_00...	K02703	CBZS01000...	694.5

Only 50 are shown, full list of hits available when downloaded

Download All Sequences

Summary

Query used:

K02703

Summarize by: phylum

🟧 Actinobacteriota [1, 0.20%]	🟩 Bacteroidota [1, 0.20%]
🟧 Chloroflexota [2, 0.40%]	🟦 Cyanobacteria [477, 96.56%]
🟦 Fusobacteriota [1, 0.20%]	🟥 Marinisomatota [1, 0.20%]
🟩 Planctomycetota [2, 0.40%]	🟥 Proteobacteria [9, 1.82%]

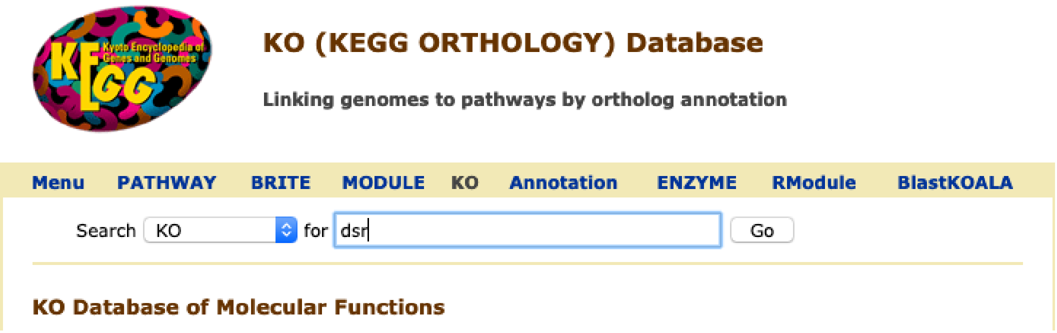
[x, y%] means x genome hits in this taxonomy, y% of total hits

http://annotree.uwaterloo.ca/app/

Use ANNOTREE
and
Prepare group presentations

Task 1: Use ANNOTREE

- Use ANNOTREE to explore the phylogenetic distribution of functions
- Try using attribute annotations for your group task
- You can use your KEGG Orthology (KO) numbers
- **Note:** You can also get KO numbers from the KEGG website (https://www.genome.jp/kegg/ko.html) by searching for gene names



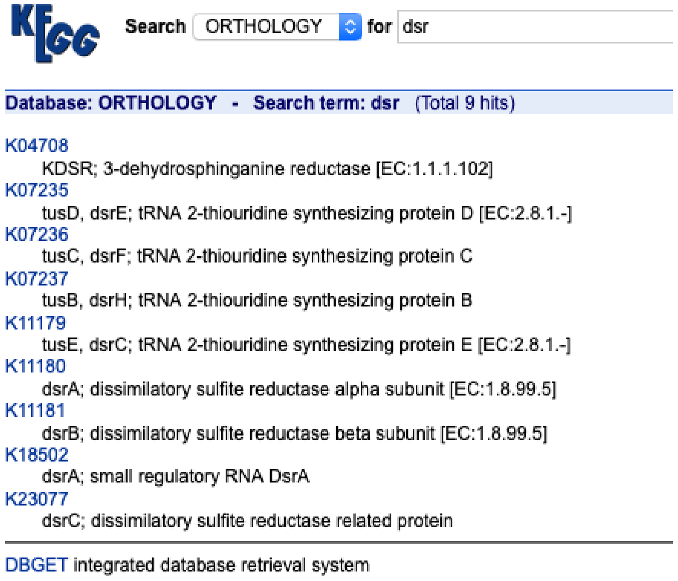
KO (KEGG ORTHOLOGY) Database

Linking genomes to pathways by ortholog annotation

Search KO for dsr

KO Database of Molecular Functions

→



Search ORTHOLOGY for dsr

Database: ORTHOLOGY - Search term: dsr (Total 9 hits)

K04708
KDSR; 3-dehydrosphinganine reductase [EC:1.1.1.102]
K07235
tusD, dsrE; tRNA 2-thiouridine synthesizing protein D [EC:2.8.1.-]
K07236
tusC, dsrF; tRNA 2-thiouridine synthesizing protein C
K07237
tusB, dsrH; tRNA 2-thiouridine synthesizing protein B
K11179
tusE, dsrC; tRNA 2-thiouridine synthesizing protein E [EC:2.8.1.-]
K11180
dsrA; dissimilatory sulfite reductase alpha subunit [EC:1.8.99.5]
K11181
dsrB; dissimilatory sulfite reductase beta subunit [EC:1.8.99.5]
K18502
dsrA; small regulatory RNA DsrA
K23077
dsrC; dissimilatory sulfite reductase related protein

DBGET integrated database retrieval system

Task 2: Prep for group presentations

- **Use the white board for illustrations**
- **Things to include:**
    - The attribute you found
    - Details about the attribute
    - The organism(s) you found it in
    - A brief explanation of biological relevance
    - The tools and annotations you used
    - Anything else?

Presentation of findings

Task: Report findings

Report your findings with regards to your objective

1. What did you find?
2. How did you make the discoveries?
   - Functional prediction only
   - Functional prediction and taxonomic context
3. Each group present for 5 mins each (max!)

Presentation of data

Presentation of data

Finding the answer to your question is only half of the issue

- How do we report/present our data?

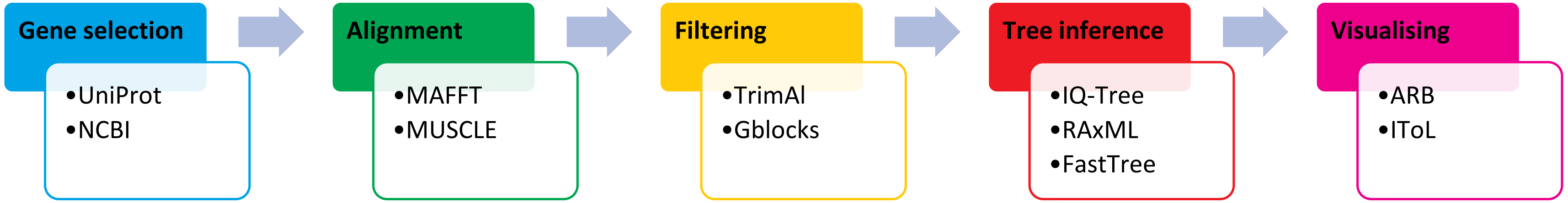
1. **Inference of gene trees**
2. **Heatmaps of genomic features**
3. **Gene synteny analysis**
4. **Creating metabolic schematics**

Presentation of data

1. Inference of gene trees

Gene trees are a great way to present

- Confirmation of annotation
- Novelty of detection / horizontal gene transfer
- Rate of evolution in the feature



Gene selection
- UniProt
- NCBI

Alignment
- MAFFT
- MUSCLE

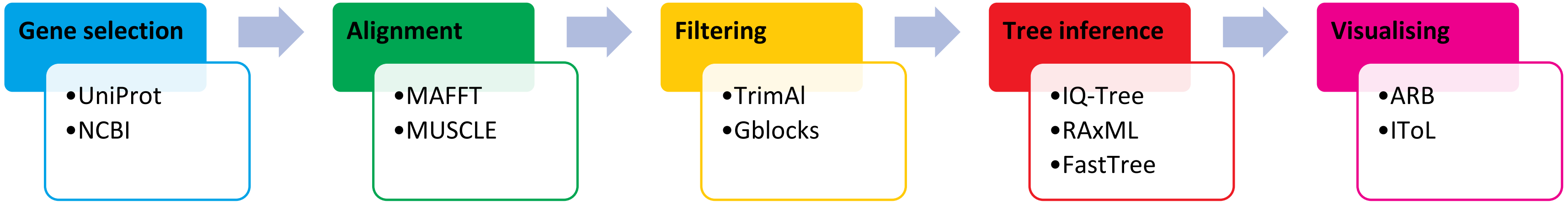
Filtering
- TrimAl
- Gblocks

Tree inference
- IQ-Tree
- RAxML
- FastTree

Visualising
- ARB
- IToL

Presentation of data

1. Inference of gene trees



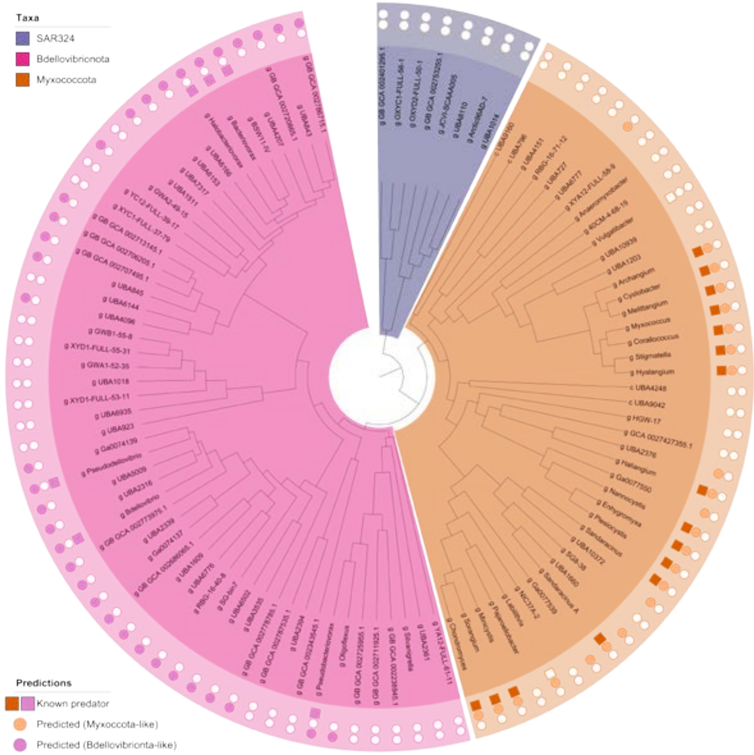
Gene selection
- UniProt
- NCBI

Alignment
- MAFFT
- MUSCLE

Filtering
- TrimAl
- Gblocks

Tree inference
- IQ-Tree
- RAxML
- FastTree

Visualising
- ARB
- IToL



Presentation of data

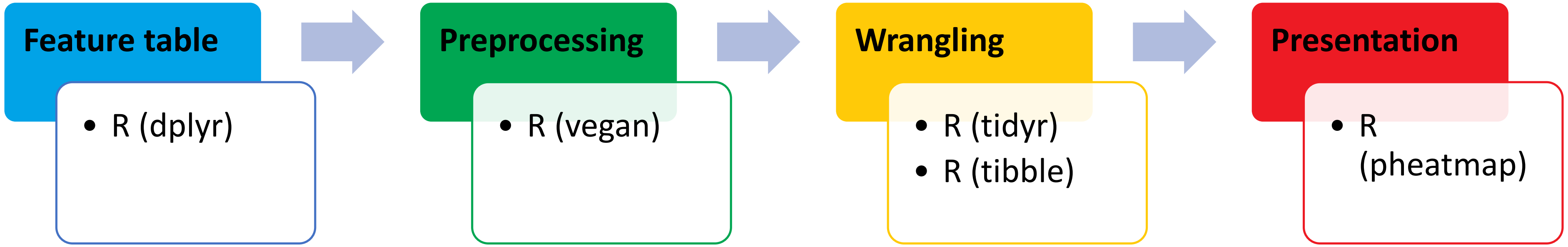
2. **Heatmaps of genomic features**

Simple figure to display complex data tables

- *M* genomes x *N* features in one place
- Presence/absence or relative abundance (multi-copy)
- Fixed layout, or clustering by patterns

Presentation of data

2. Heatmaps of genomic features

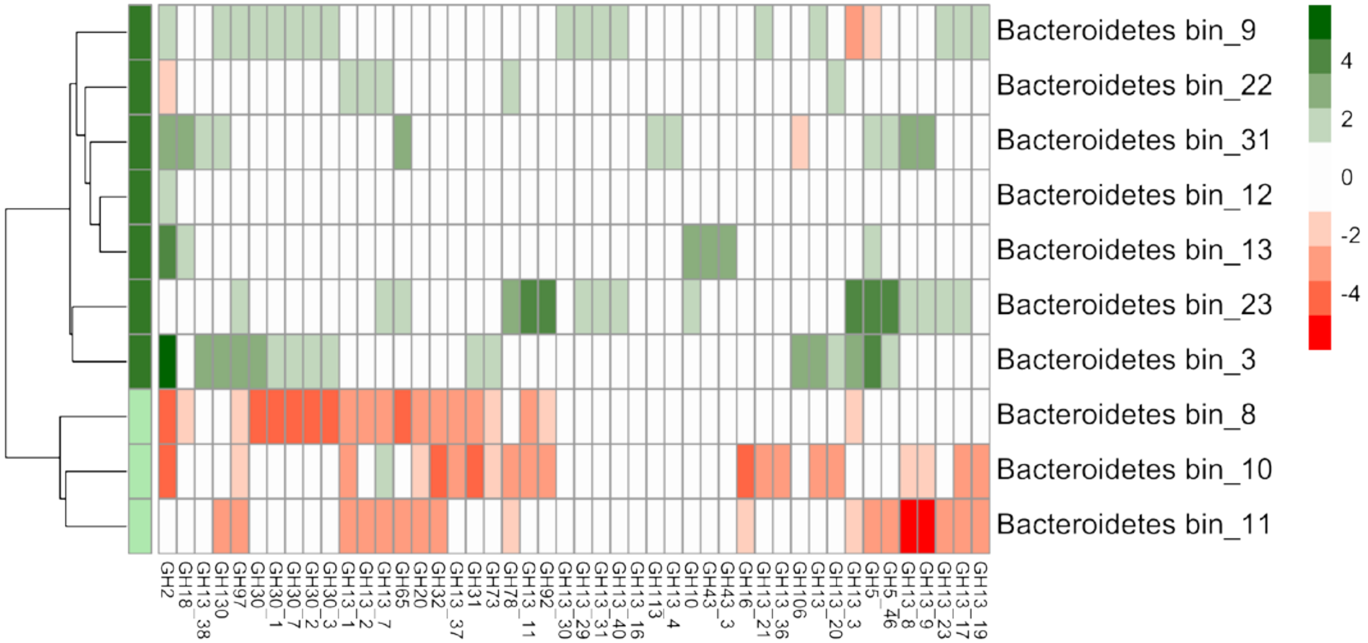


Feature table
- R (dplyr)

Preprocessing
- R (vegan)

Wrangling
- R (tidyr)
- R (tibble)

Presentation
- R (pheatmap)



Presentation of data

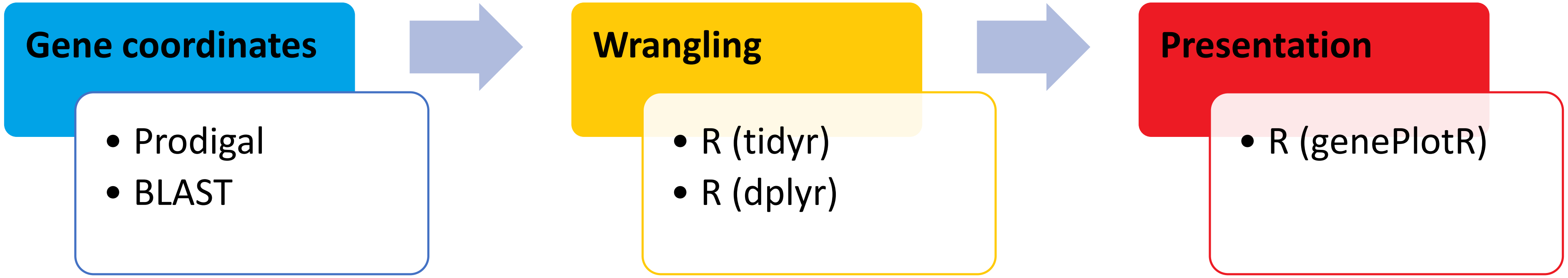
3. **Gene synteny analysis**

A more informative view of gene content

- Shows local gene context
- More detailed than reporting gene table
- Sometimes absence of genes from operon is biologically informative

Presentation of data

3. Gene synteny analysis



Gene coordinates

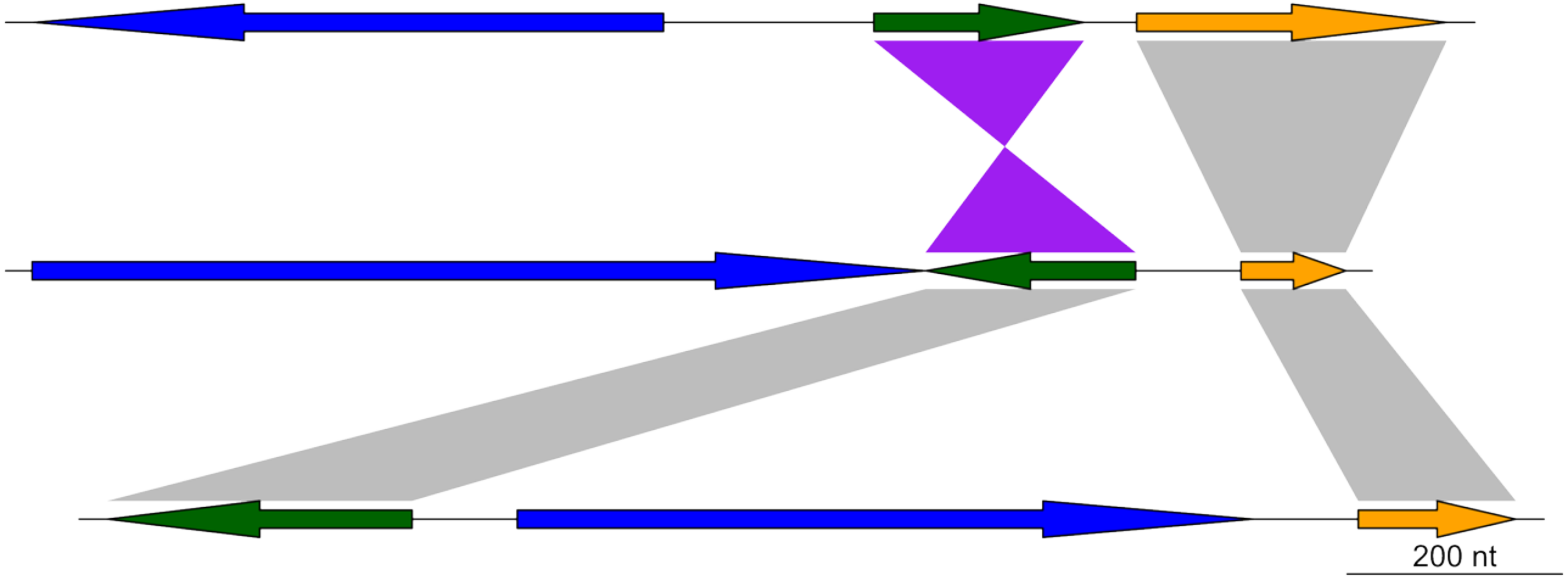
- Prodigal
- BLAST

Wrangling

- R (tidyr)
- R (dplyr)

Presentation

- R (genePlotR)



Presentation of data

4. Creating metabolic schematics

Summarise the entire core metabolism of an organism into a single figure

What’s the magic tool for producing these?

- Illustrator, Inkscape, GIMP (and a lot of time)
- Use tools for picking colour schemes
  - ColorBrewer2 (http://colorbrewer2.org)
  - IWantHue (https://medialab.github.io/iwanthue/)

Task: Presentation of data

- **Using Pheatmap to build CAZY plot**
- **Using genoPlot R to build comparative a gene map**

Wrap up and Q/A