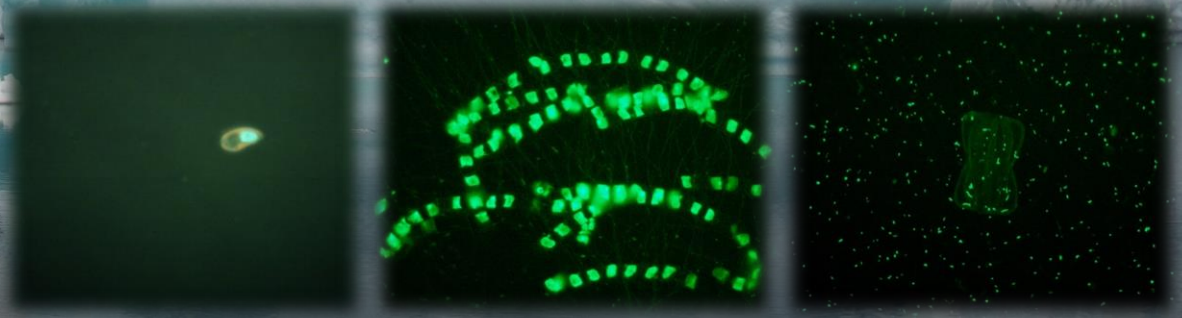
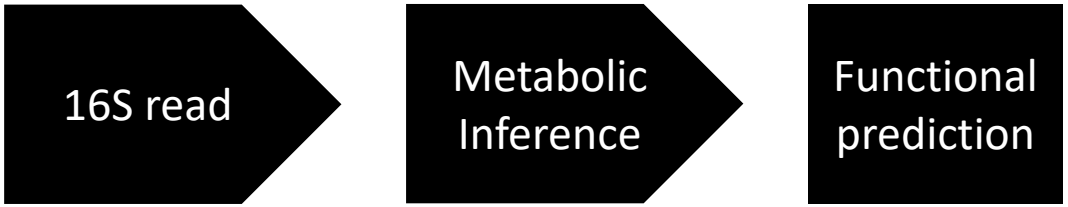
OTU-free community structure, metabolic inference, and meta-'omic analysis with paprica (PAthway PRedIction by phylogenetIC plAcement)

Jeff S. Bowman
jsbowman@ucsd.edu
www.polarmicrobes.org

Introduction
Part 1: paprica – overview of method
Part 2: A practical tutorial
Part 3: Some fun applications



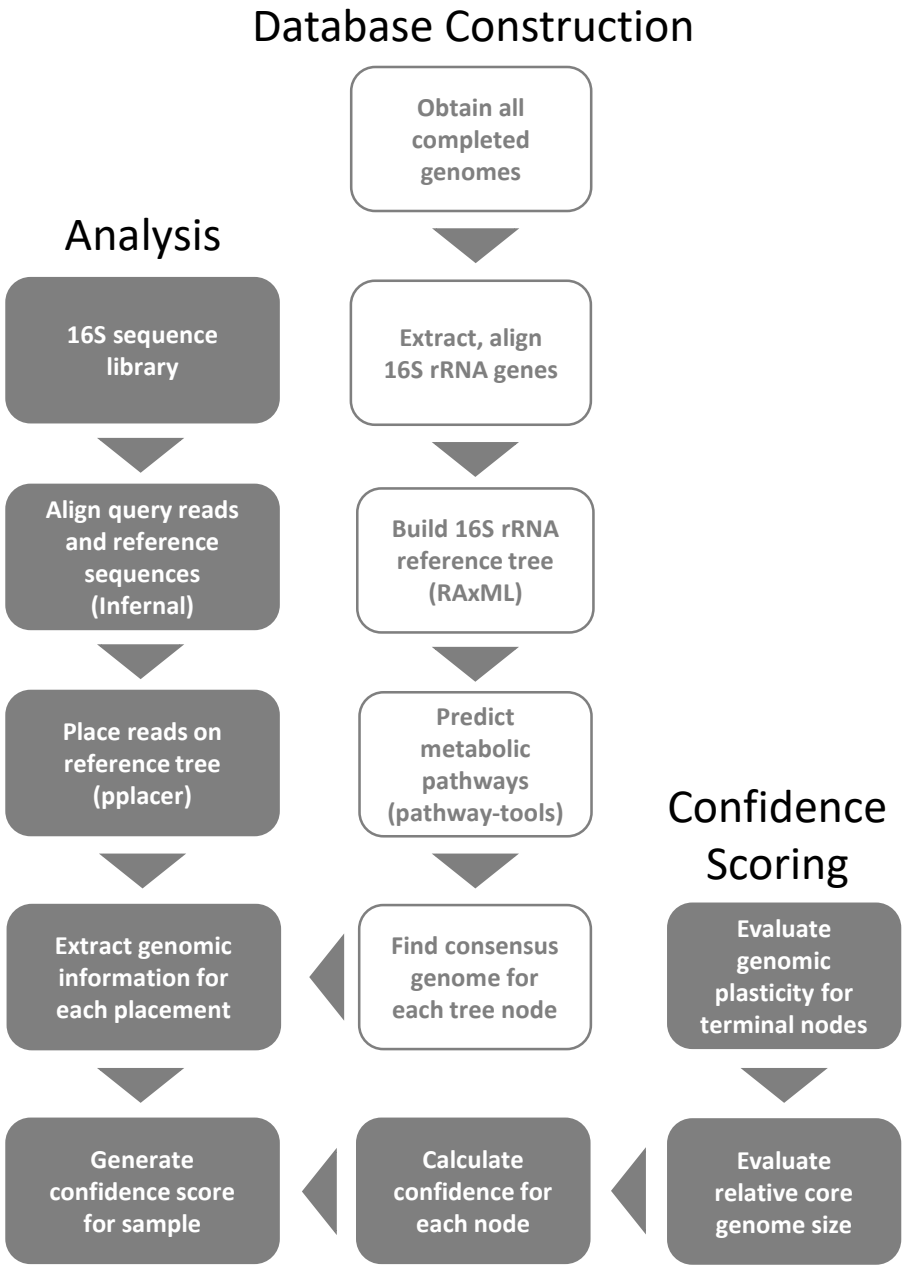


Why metabolic inference?

- Sequencing has gotten cheap, but still impractical to sequence metagenomes for large sample sets
- Even deep metagenomes still miss rare genes
- Useful to have a “model” of the distribution of functions, as we currently know them

Other software

- PICRUSt: tried and true, but underlying model is complex and requires closed OTU-based description of community structure
- tax4fun: method not clear from manuscript/documentation, also requires closed OTU-based description of community structure



Three components to paprica:

1. Database construction
2. Analysis
3. Confidence scoring

Caveats:

Metabolic inference is only as good as…

- Our genome annotations
- The diversity of *completed* genomes
- Our knowledge of enzymes and metabolic pathways

And is further limited by…

- Genomic plasticity

Database Construction

Obtain all completed genomes

During last database build there were 6,103 bacterial and 227 archaeal genomes, add to this 649 transcriptomes in MMETSP

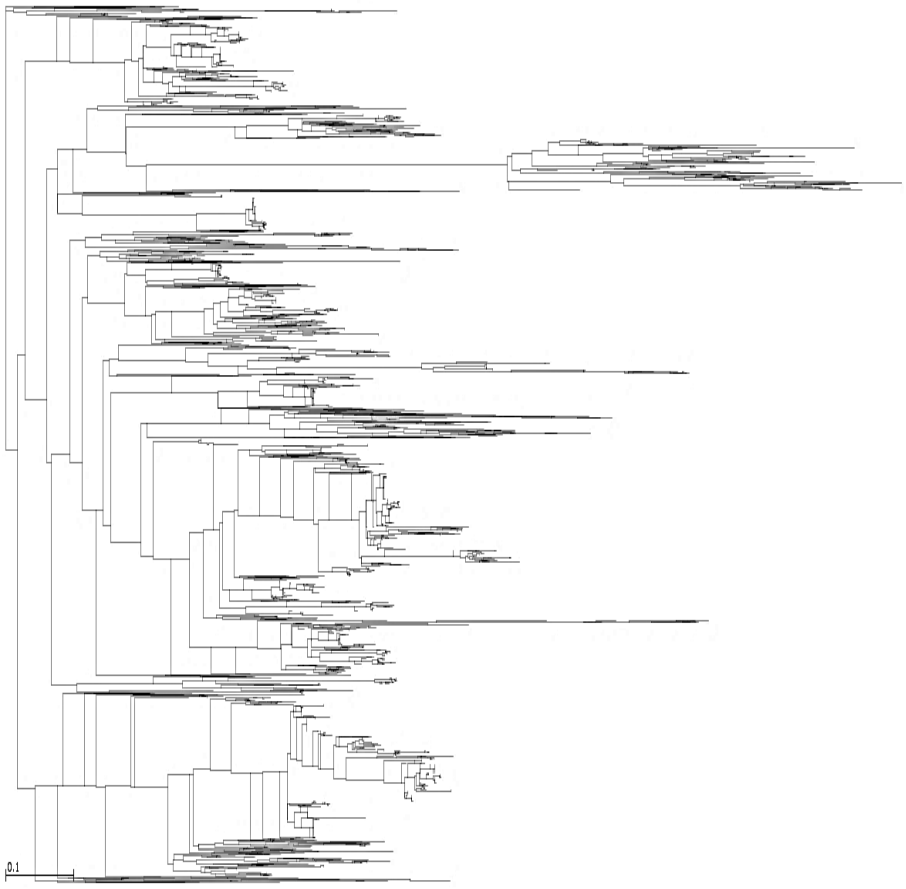
Database Construction

Obtain all completed genomes

Extract, align 16S rRNA genes

Build 16S rRNA reference tree (RAxML)

Build reference tree using the Infernal aligner and RAxML)



Database Construction

Obtain all completed genomes

↓

Extract, align 16S rRNA genes

↓

Build 16S rRNA reference tree (RAxML)

↓

Predict metabolic pathways (pathway-tools)

Predict metabolic pathways using pathway-tools

Database Construction

Obtain all completed genomes

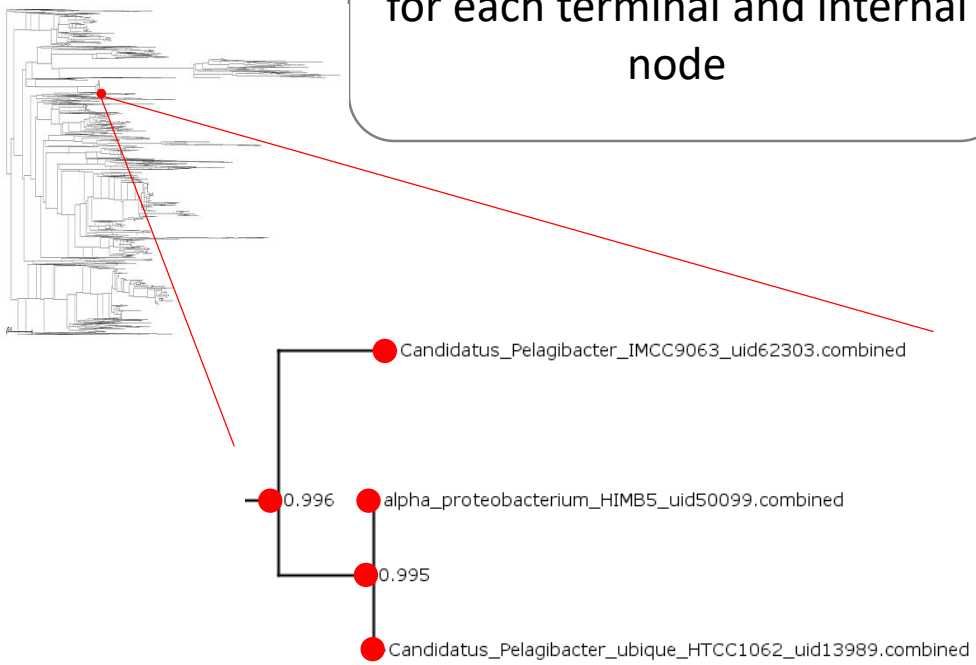
Extract, align 16S rRNA genes

Build 16S rRNA reference tree (RAxML)

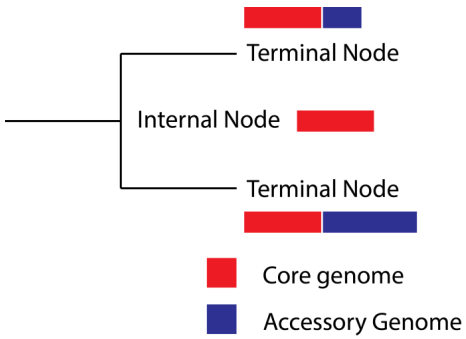
Predict metabolic pathways (pathway-tools)

Find consensus genome for each tree node

Collect relevant information for each terminal and internal node

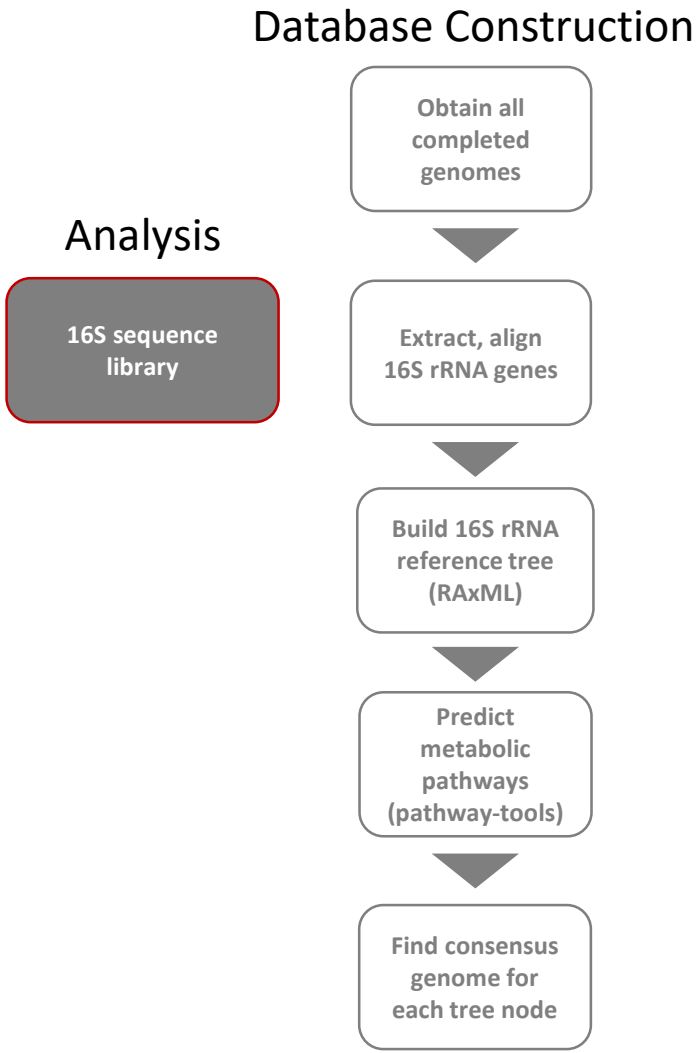


Pelagibacter genus

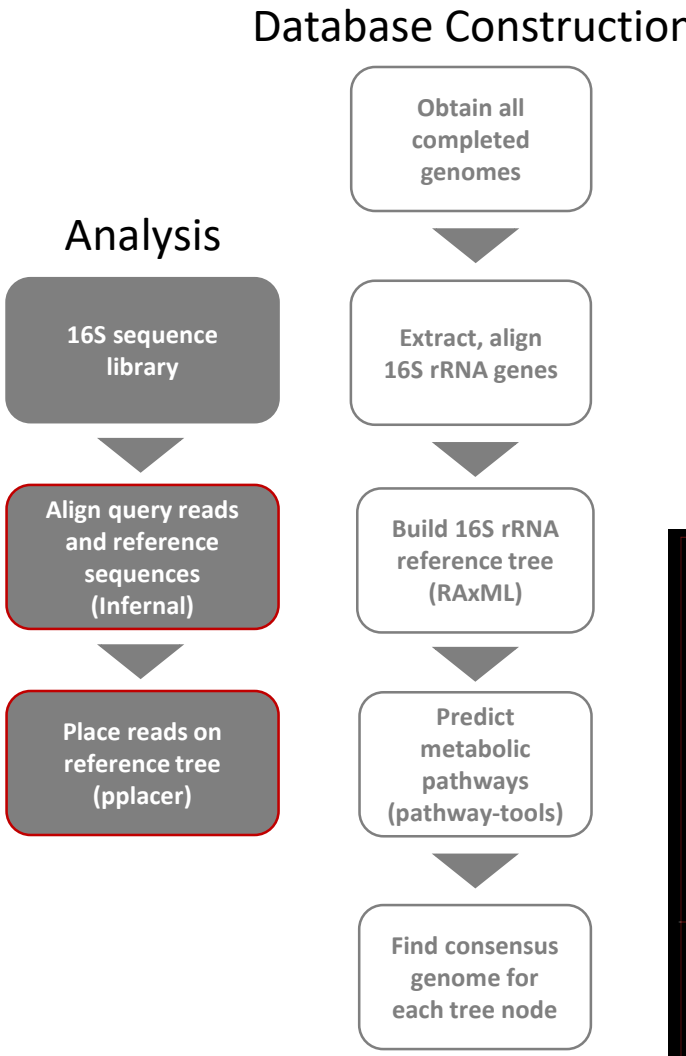


Information collected

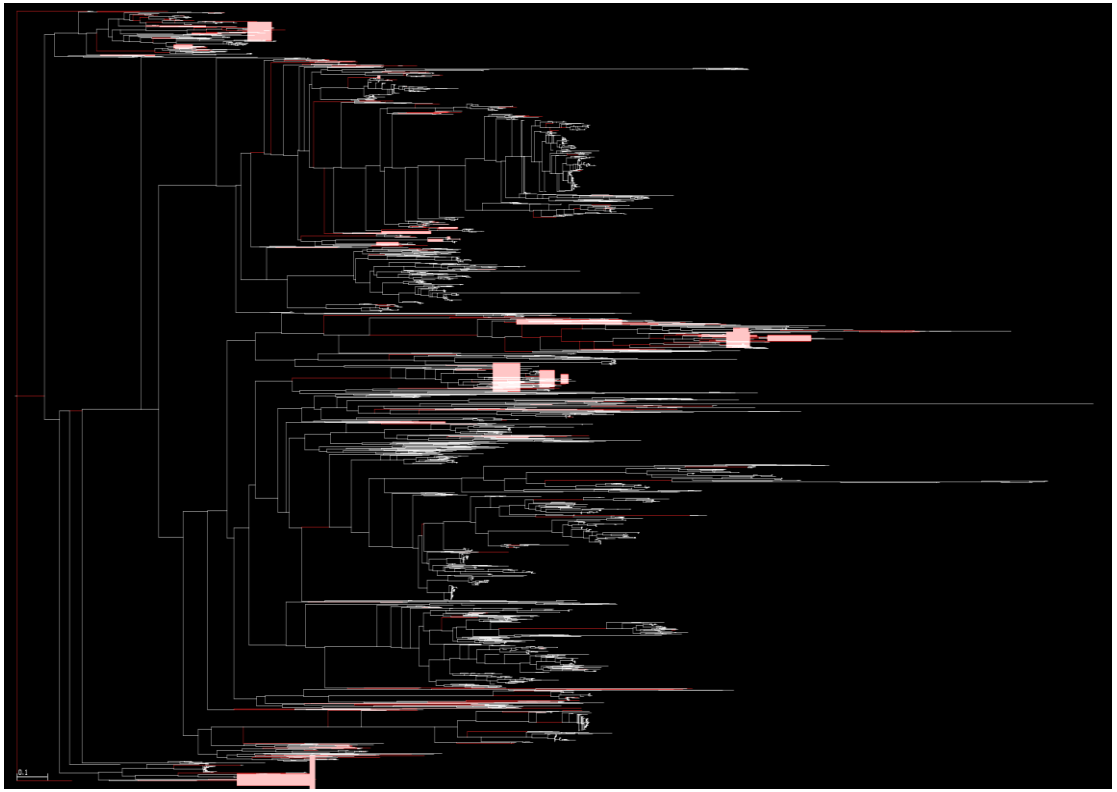
- Metabolic pathways
- Enzymes
- GC content
- n16S genes
- Genome length
- nCDS
- Phi parameter
- Number of genetic elements

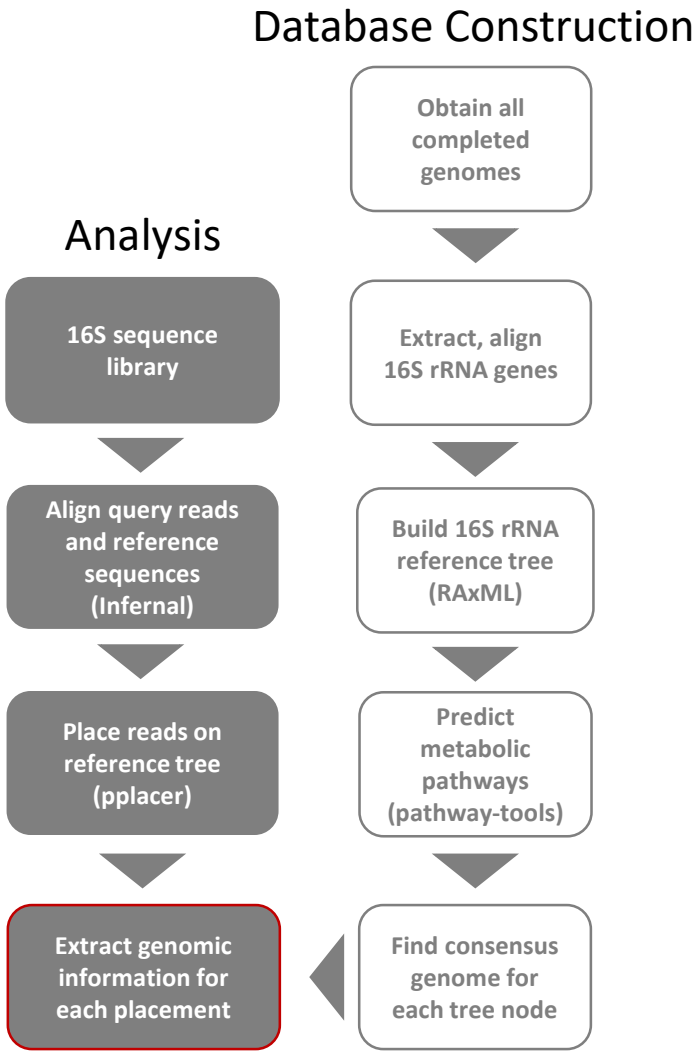


For analysis, perform normal quality controls on 16S dataset (I like Seqmagick for this).

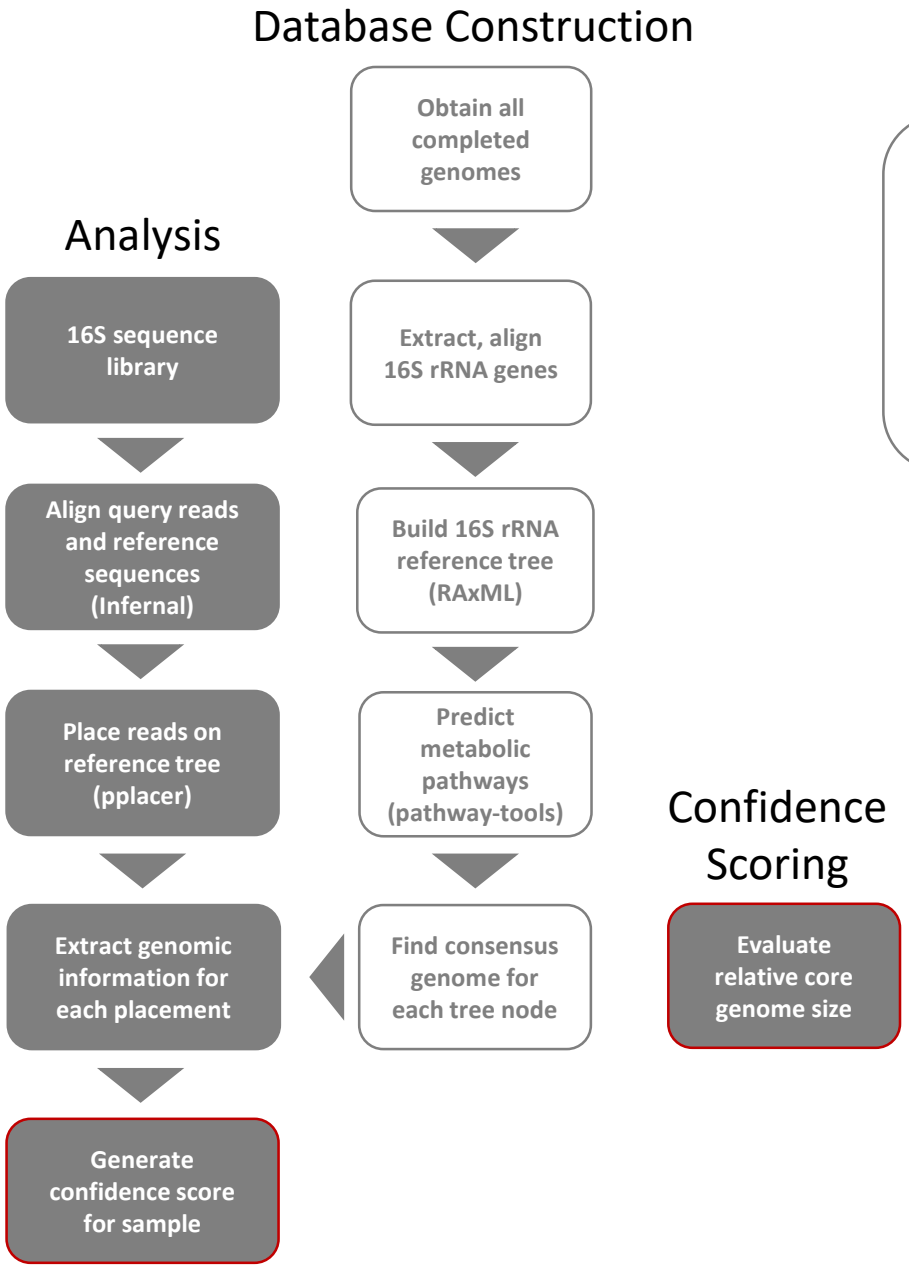


Use pplacer to carry out phylogenetic placement: place query reads at best spot on reference tree





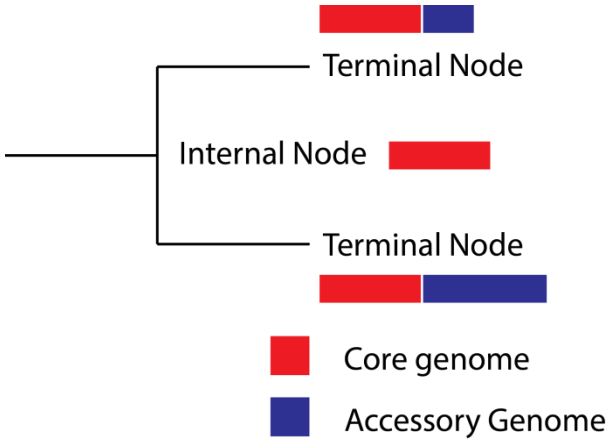
Once reads are placed, tally the pathways, enzymes, and other features associated with each placement.

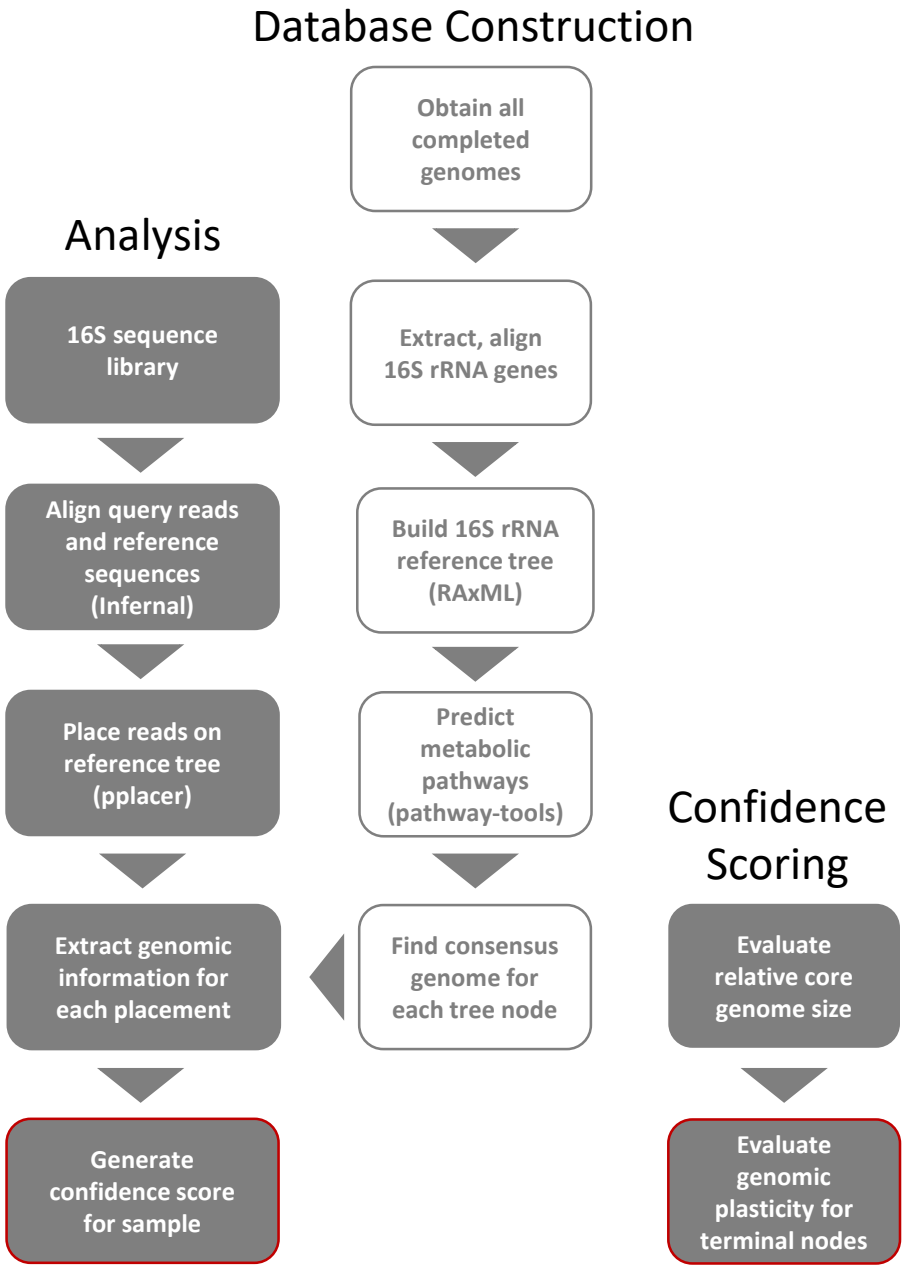


For internal nodes, the confidence score takes into account the relative size of the core genomes compared to the mean genome size of all clade members.

$$c = \frac{S_{core}}{mean(S_{clade})} * (1 - mean(\phi))$$

S_{core} = size core genome
S_{clade} = mean clade genome size
ϕ = genomic plasticity of all clade members



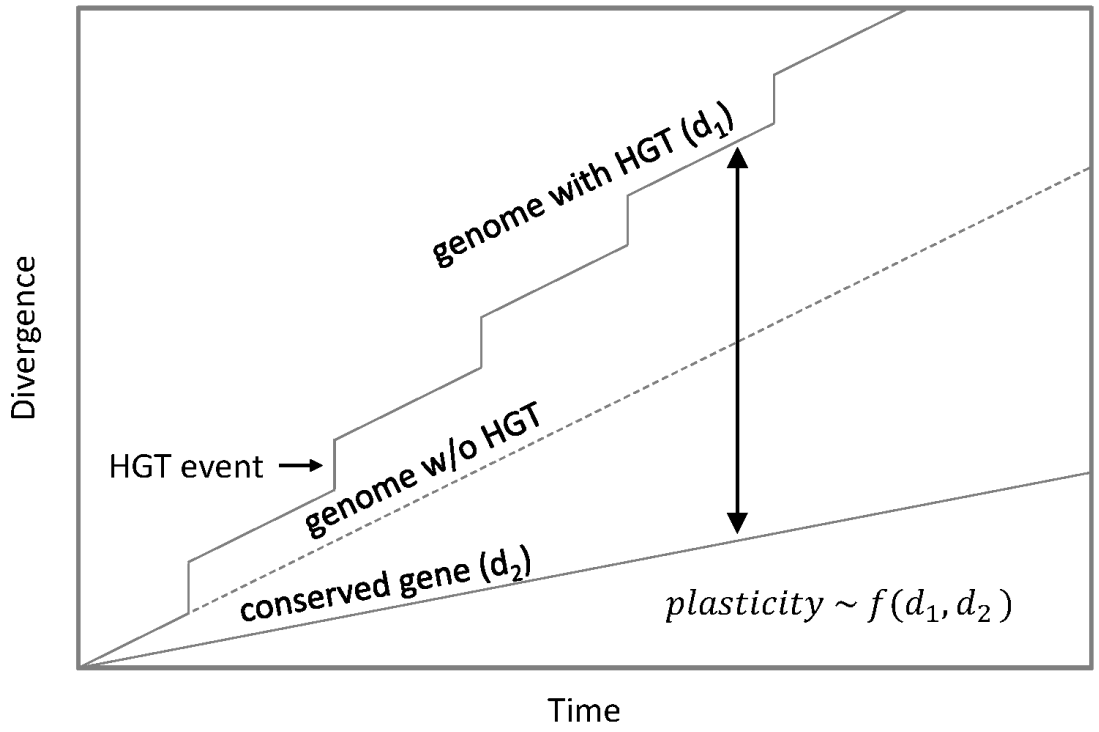


The confidence score also takes into account the predicted genomic plasticity of each node.

$$c = \frac{S_{core}}{mean(S_{clade})} * (1 - mean(\phi))$$

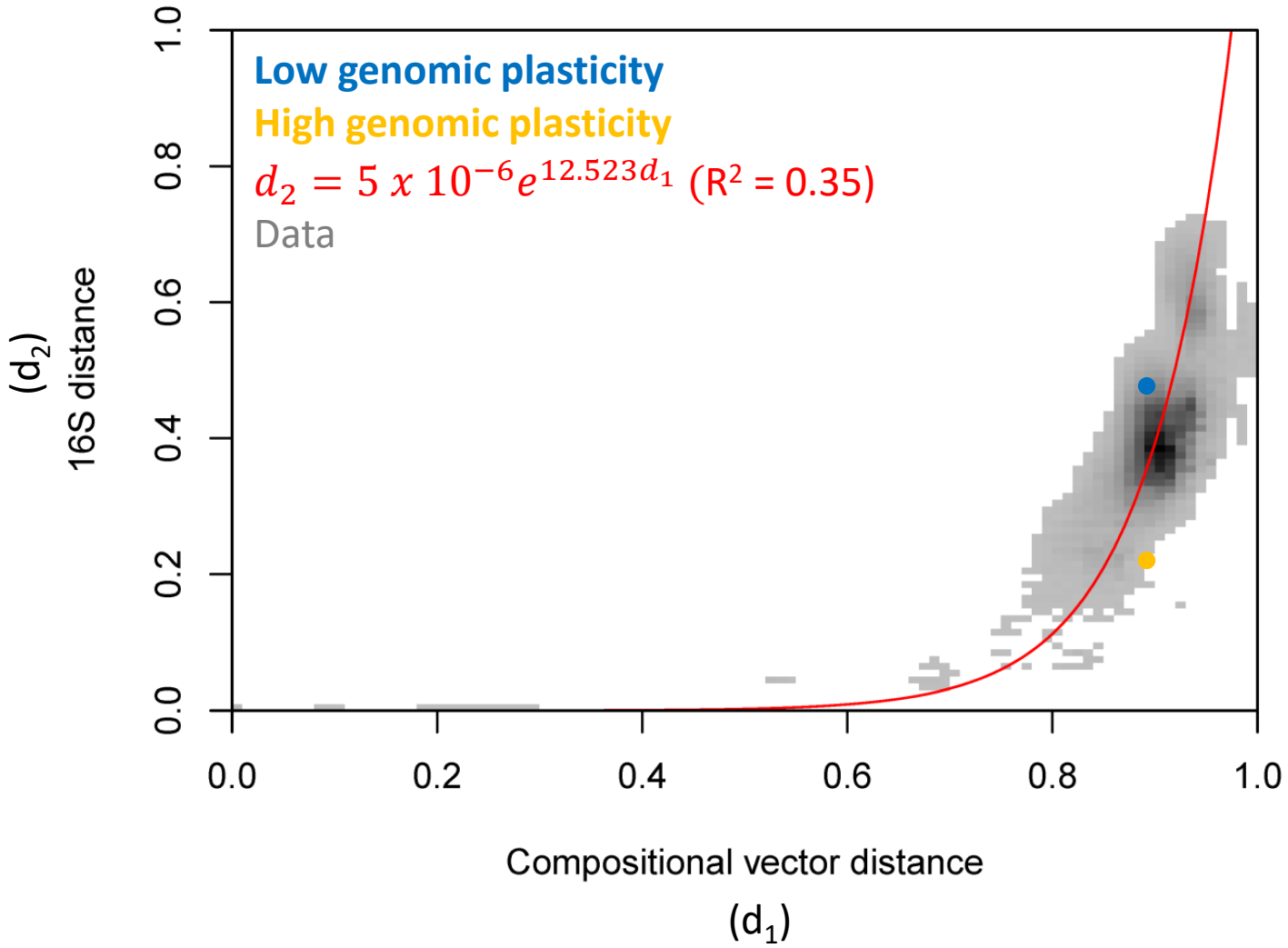
S_{core} = size core genome
S_{clade} = mean clade genome size
ϕ = genomic plasticity of all clade members

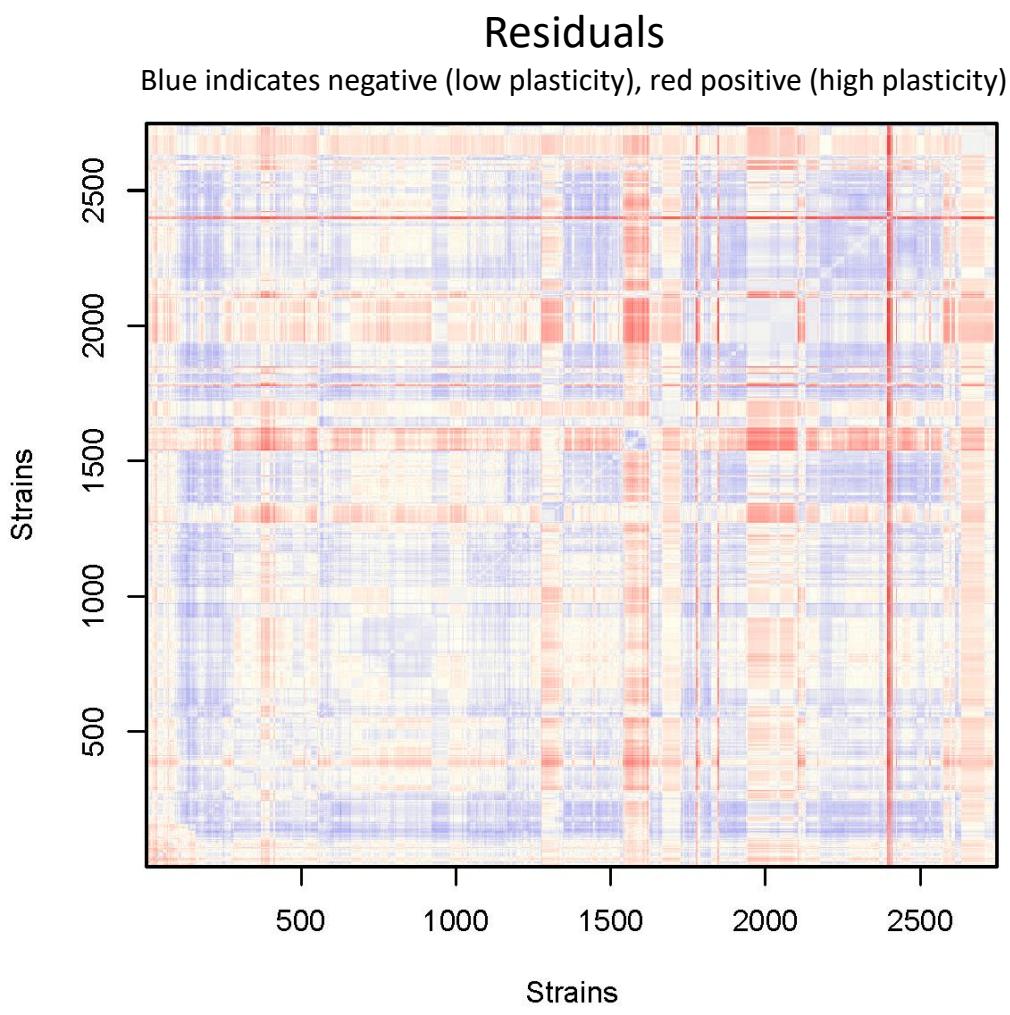
Our measure of genomic plasticity (ϕ) is based on the degree of divergence between two genomes relative to the divergence between their 16S rRNA genes.

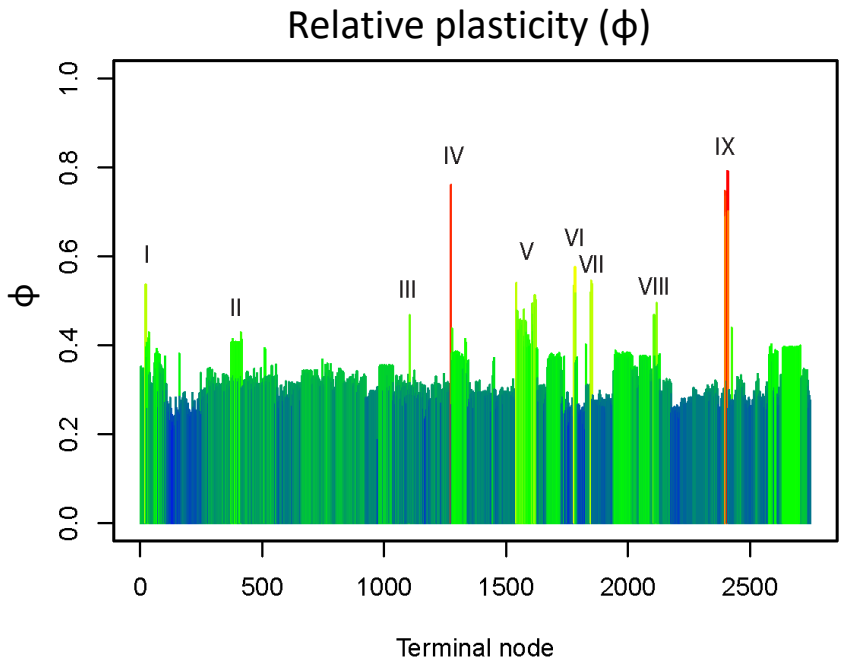


Genome distance is determined as the distance between predicted proteome *compositional vectors*

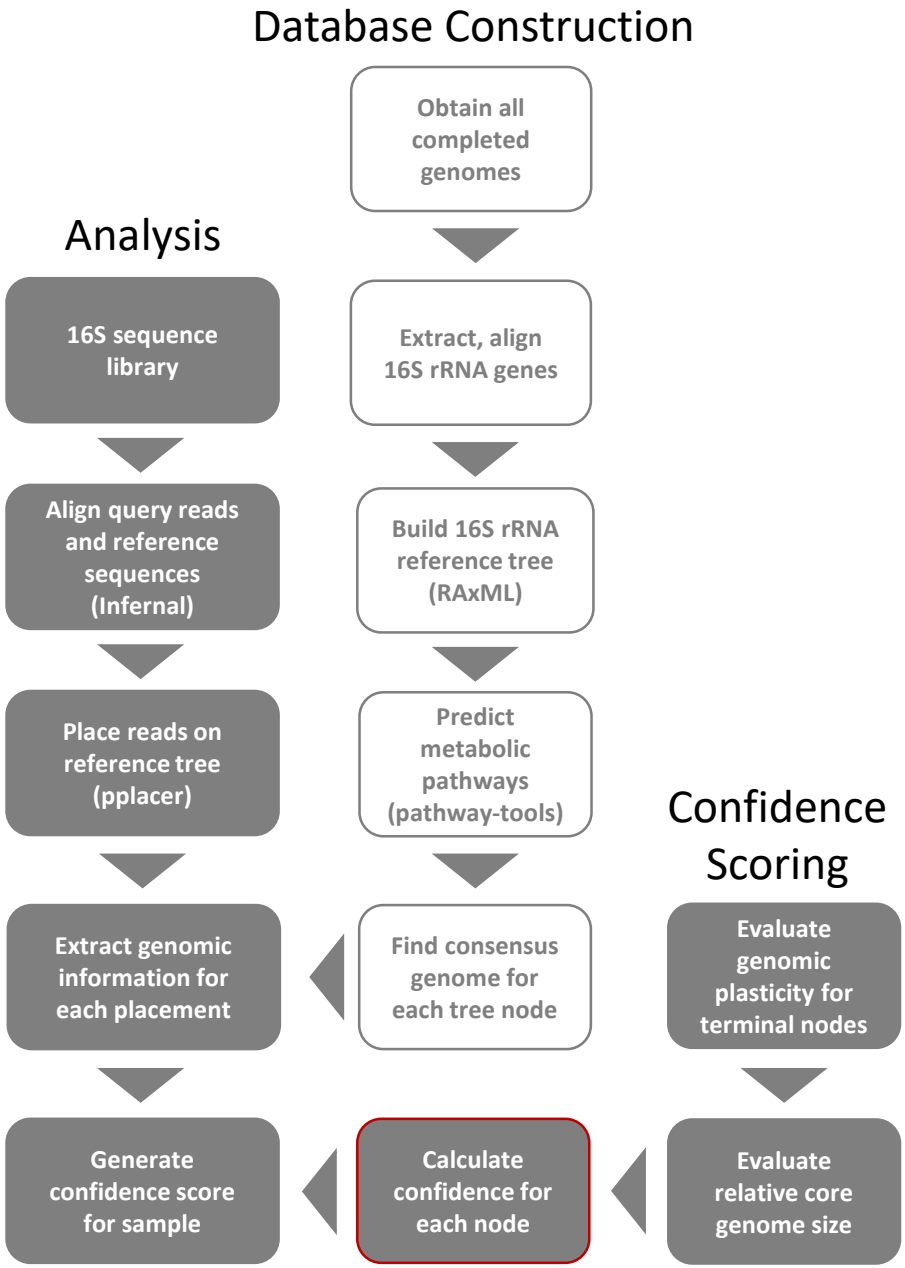
- Red line = 16S distance predicted from CV distance for all possible pairwise comparisons
- Deviations from prediction indicate more or less plasticity than expected for one or both of the genomes being compared





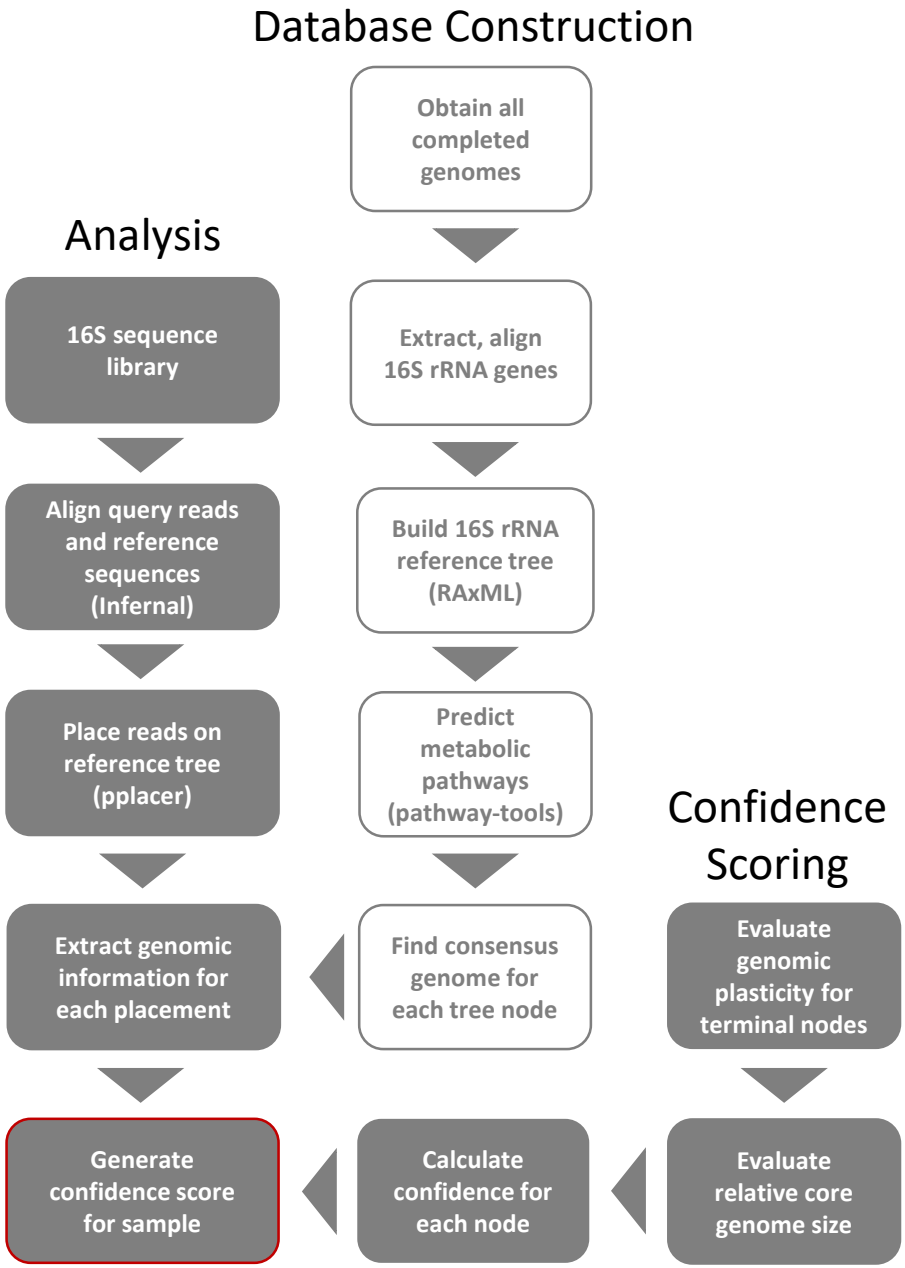


I) *Nanoarchaeum equitans*
II) the *Mycobacteria*
III) a butyrate producing bacterium within *Clostridium*
IV) *Candidatus Hodgkinia circadicola*
V) the *Mycoplasma*
VI) *Sulcia muelleri*
VII) *Portiera aleyrodidanum*
VIII) *Buchnera aphidicola*,
IX) symbiotic *Oxalobacteraceae*



$$c = \frac{S_{core}}{mean(S_{clade})} * (1 - mean(\phi))$$

S_{core} = size core genome
S_{clade} = mean clade genome size
ϕ = genomic plasticity of all clade members



Sample confidence score is the weighted mean of scores for all nodes with placements.

$$c = \frac{S_{core}}{mean(S_{clade})} * (1 - mean(\phi))$$

S_{core} = size core genome
S_{clade} = mean clade genome size
ϕ = genomic plasticity of all clade members

3 ways to run paprica:

- Locally (or on a workstation/server) with OSX, Linux, or Cygwin (Windows)
  - This method is preferred
- Using the provided Amazon Machine Instance
  - This is the second choice
- Using the provided VirtualBox
  - Recommended for testing only

Two shell scripts that execute Python scripts

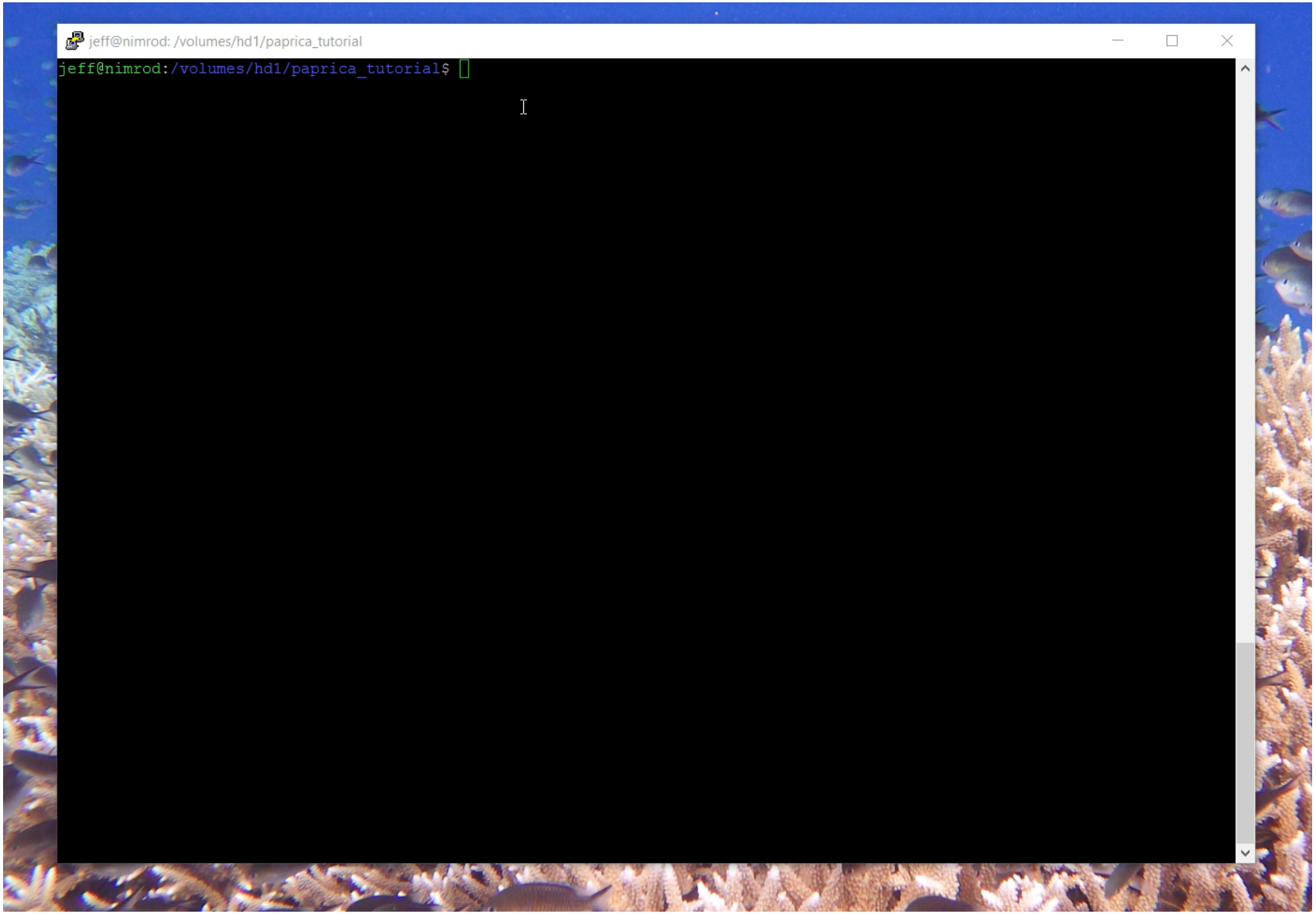
- paprica-build.sh [domain] builds a database for the specified domain
  - Not necessary for most users, pre-built databases are provided
- paprica-run.sh [input fasta] [domain]
  - Actually does the analysis, given a QC'd fasta file

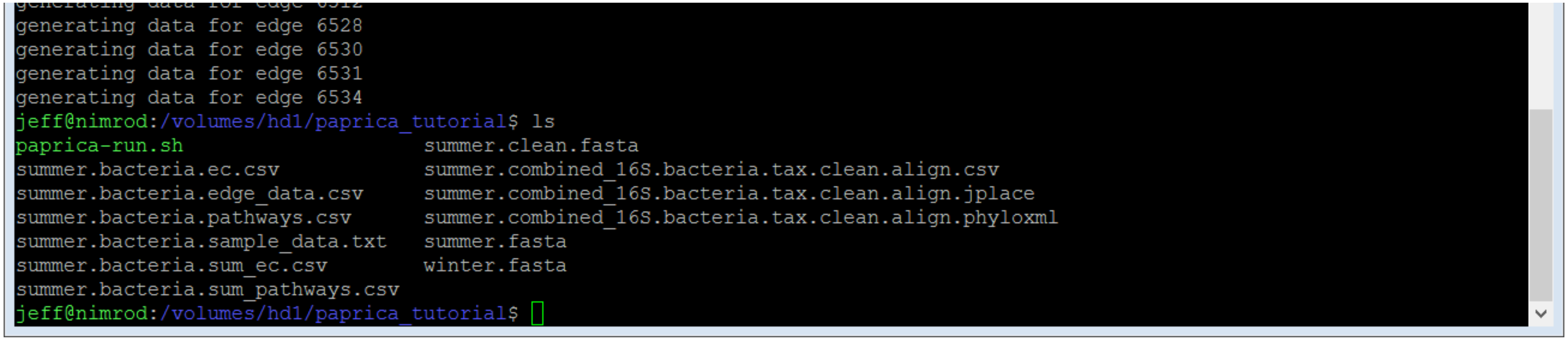
https://github.com/bowmanjeffs/paprica

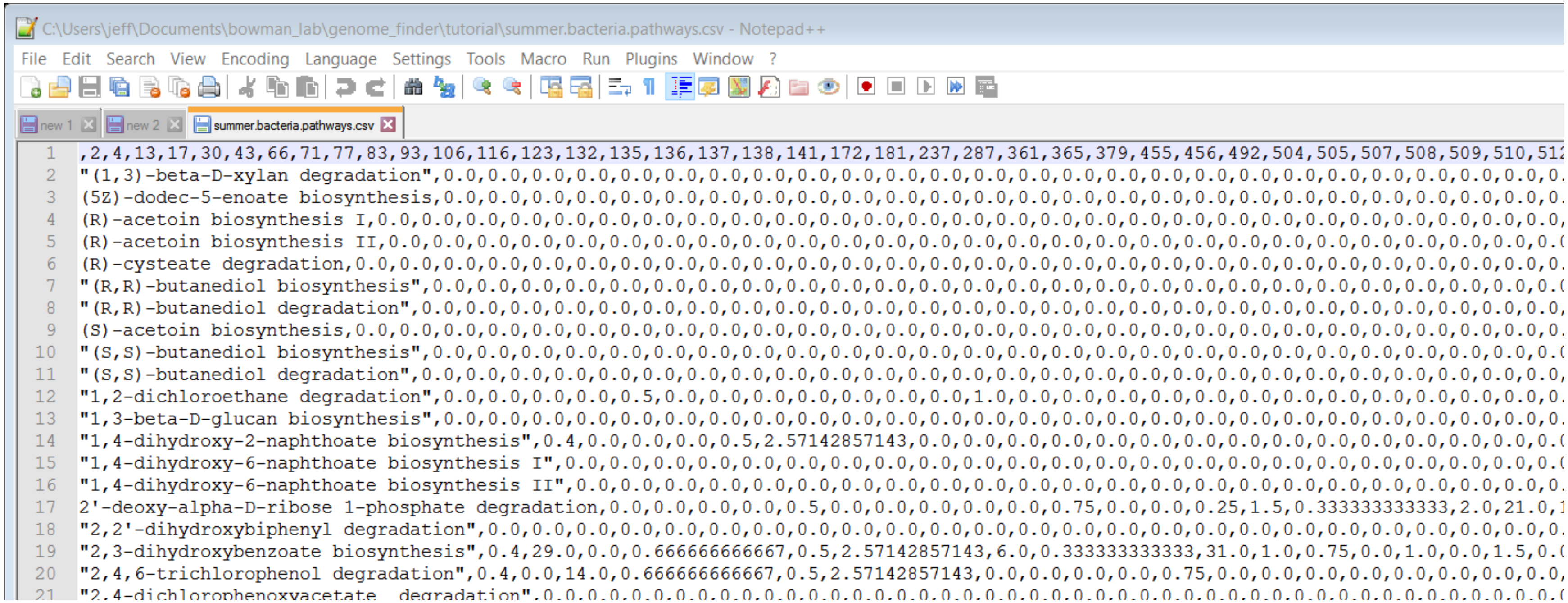
Inside paprica-run.sh…

```
#!/bin/bash

#### These are the critical steps for using paprica if you are use the provided database (a.k.a. ref_genome_database) or
#### have already built it using paprica_build.sh.  Used in this way paprica is nice and lightweight, but you won't have
#### access to the PGDBs if you want to do something more sophisticated than just tally up the number of metabolic pathways
#### that have been inferred.

#### If you have a large number run this script in a loop (see the manual for an example).  Because the bottleneck is alignment,
#### and infernal is parallelized, it is best not to run samples in parallel.

#### Execute this script as ./paprica-run.sh [query] [domain].

query=$1
domain=$2

## Select gene based on domain.

if [ $domain = "eukarya" ];then
                gene=18S
else
                gene=16S
fi

## 1. phylogenetic placement of query reads

paprica-place_it.py -ref_dir ref_genome_database -query $query -ref combined_$gene.$domain.tax -splits 1 -domain $domain

## 2. find pathways and other information associated with edges.  if you subsampled in the previous step (i.e. with -n) your input
##    file is $query.sub.combined_$gene.tax.clean.align.csv

paprica-tally_pathways.py -ref_dir ref_genome_database -i $query.combined_$gene.$domain.tax.clean.align.csv -o $query.$domain -cutoff 0.5 -domain $domain
```

Part 2 – A practical tutorial




generating data for edge 6528
generating data for edge 6530
generating data for edge 6531
generating data for edge 6534
jeff@nimrod:/volumes/hd1/paprica_tutorial$ ls
paprica-run.sh
summer.clean.fasta
summer.bacteria.ec.csv
summer.combined_16S.bacteria.tax.clean.align.csv
summer.bacteria.edge_data.csv
summer.combined_16S.bacteria.tax.clean.align.jplace
summer.bacteria.pathways.csv
summer.combined_16S.bacteria.tax.clean.align.phyloxml
summer.bacteria.sample_data.txt
summer.fasta
summer.bacteria.sum_ec.csv
winter.fasta
summer.bacteria.sum_pathways.csv
jeff@nimrod:/volumes/hd1/paprica_tutorial$


C:\Users\jeff\Documents\bowman_lab\genome_finder\tutorial\summer.bacteria.pathways.csv - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
new 1
new 2
summer.bacteria.pathways.csv
,2,4,13,17,30,43,66,71,77,83,93,106,116,123,132,135,136,137,138,141,172,181,237,287,361,365,379,455,456,492,504,505,507,508,509,510,51
"(1,3)-beta-D-xylan degradation",0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
(5Z)-dodec-5-enoate biosynthesis,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
(R)-acetoin biosynthesis I,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
(R)-acetoin biosynthesis II,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
(R)-cysteate degradation,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
"(R,R)-butanediol biosynthesis",0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
"(R,R)-butanediol degradation",0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
(S)-acetoin biosynthesis,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
"(S,S)-butanediol biosynthesis",0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
"(S,S)-butanediol degradation",0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
"1,2-dichloroethane degradation",0.0,0.0,0.0,0.0,0.5,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
"1,3-beta-D-glucan biosynthesis",0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
"1,4-dihydroxy-2-naphthoate biosynthesis",0.4,0.0,0.0,0.0,0.5,2.57142857143,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
"1,4-dihydroxy-6-naphthoate biosynthesis I",0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
"1,4-dihydroxy-6-naphthoate biosynthesis II",0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
2'-deoxy-alpha-D-ribose 1-phosphate degradation,0.0,0.0,0.0,0.0,0.5,0.0,0.0,0.0,0.0,0.0,0.0,0.75,0.0,0.0,0.25,1.5,0.333333333333,2.0,21.0,1
"2,2'-dihydroxybiphenyl degradation",0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
"2,3-dihydroxybenzoate biosynthesis",0.4,29.0,0.0,0.666666666667,0.5,2.57142857143,6.0,0.333333333333,31.0,1.0,0.75,0.0,1.0,0.0,1.5,0.0
"2,4,6-trichlorophenol degradation",0.4,0.0,14.0,0.666666666667,0.5,2.57142857143,0.0,0.0,0.0,0.0,0.75,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
"2,4-dichlorophenoxyacetate degradation",0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.

Part 2 – A practical tutorial

```
generating data for edge 6528
generating data for edge 6530
generating data for edge 6531
generating data for edge 6534
jeff@nimrod:/volumes/hd1/paprica_tutorial$ ls
paprica-run.sh                    summer.clean.fasta
summer.bacteria.ec.csv            summer.combined_16S.bacteria.tax.clean.align.csv
summer.bacteria.edge_data.csv     summer.combined_16S.bacteria.tax.clean.align.jplace
summer.bacteria.pathways.csv      summer.combined_16S.bacteria.tax.clean.align.phyloxml
summer.bacteria.sample_data.txt   summer.fasta
summer.bacteria.sum_ec.csv        winter.fasta
summer.bacteria.sum_pathways.csv
jeff@nimrod:/volumes/hd1/paprica_tutorial$
```

C:\Users\jeff\Documents\bowman_lab\genome_finder\tutorial\summer.bacteria.sum_pathways.csv - Notepad++

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?

new 1 | new 2 | summer.bacteria.pathways.csv | summer.bacteria.sum_pathways.csv

```
"(1,3)-beta-D-xylan degradation",
(5Z)-dodec-5-enoate biosynthesis,
(R)-acetoin biosynthesis I,
(R)-acetoin biosynthesis II,110.604007104
(R)-cysteate degradation,
"(R,R)-butanediol biosynthesis",
"(R,R)-butanediol degradation",
(S)-acetoin biosynthesis,18.0
"(S,S)-butanediol biosynthesis",
"(S,S)-butanediol degradation",
"1,2-dichloroethane degradation",47.2346072187
"1,3-beta-D-glucan biosynthesis",
"1,4-dihydroxy-2-naphthoate biosynthesis",987.547433547
"1,4-dihydroxy-6-naphthoate biosynthesis I",
"1,4-dihydroxy-6-naphthoate biosynthesis II",21.8333333333
2'-deoxy-alpha-D-ribose 1-phosphate degradation,1390.96452305
"2,2'-dihydroxybiphenyl degradation",183.0
"2,3-dihydroxybenzoate biosynthesis",1219.36858097
"2,4,6-trichlorophenol degradation",20.8880952381
"2,4-dichlorophenoxyacetate  degradation",1.0
"2,4-dichlorotoluene degradation",34.7187793427
```

Part 2 – A practical tutorial

```
generating data for edge 6528
generating data for edge 6530
generating data for edge 6531
generating data for edge 6534
jeff@nimrod:/volumes/hd1/paprica_tutorial$ ls
paprica-run.sh                    summer.clean.fasta
summer.bacteria.ec.csv            summer.combined_16S.bacteria.tax.clean.align.csv
summer.bacteria.edge_data.csv     summer.combined_16S.bacteria.tax.clean.align.jplace
summer.bacteria.pathways.csv      summer.combined_16S.bacteria.tax.clean.align.phyloxml
summer.bacteria.sample_data.txt   summer.fasta
summer.bacteria.sum_ec.csv        winter.fasta
summer.bacteria.sum_pathways.csv
jeff@nimrod:/volumes/hd1/paprica_tutorial$
```

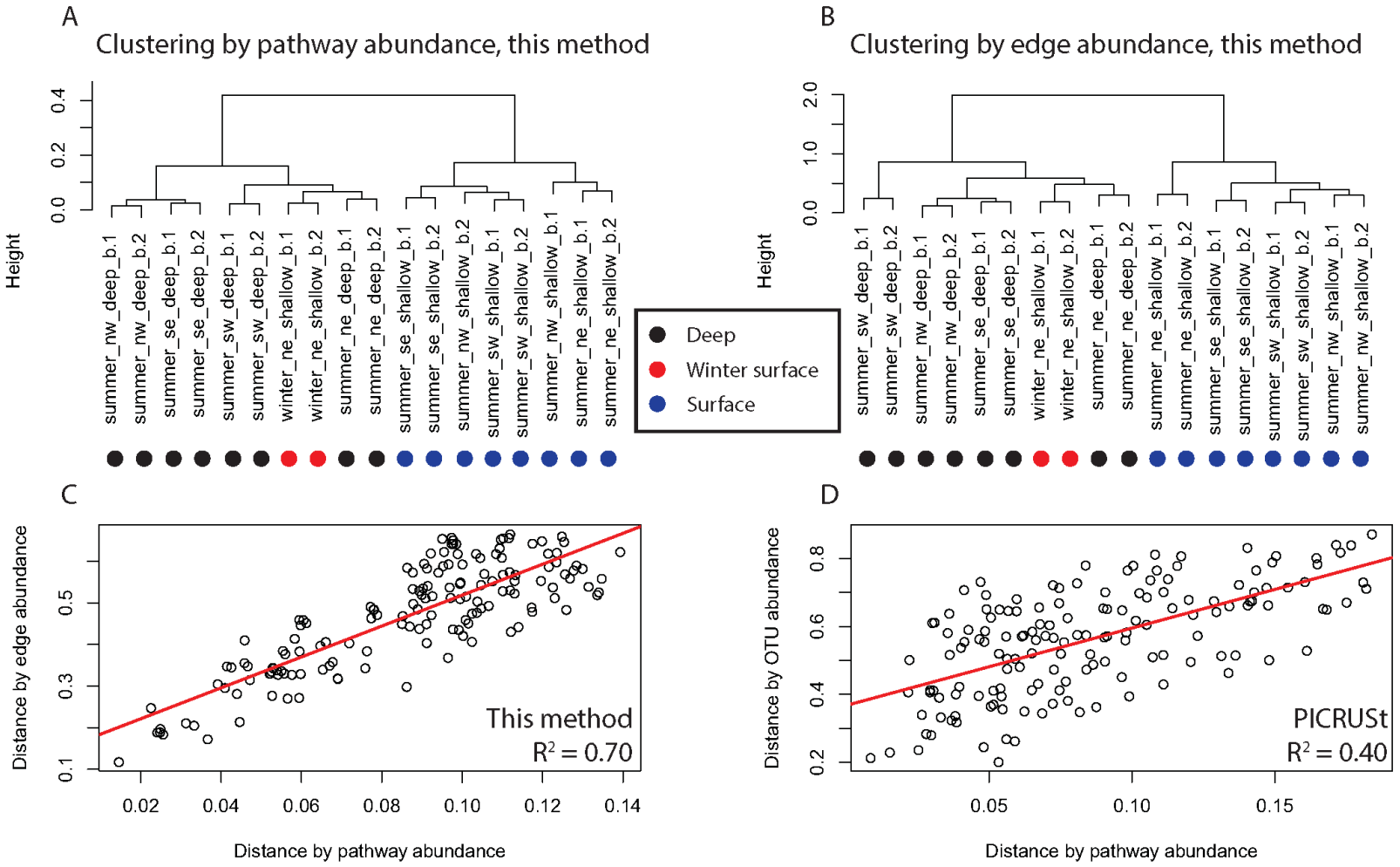
C:\Users\jeff\Documents\bowman_lab\genome_finder\tutorial\summer.bacteria.edge_data.csv - Notepad++

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?

new 1 | new 2 | summer.bacteria.pathways.csv | summer.bacteria.sum_pathways.csv | summer.bacteria.edge_data.csv

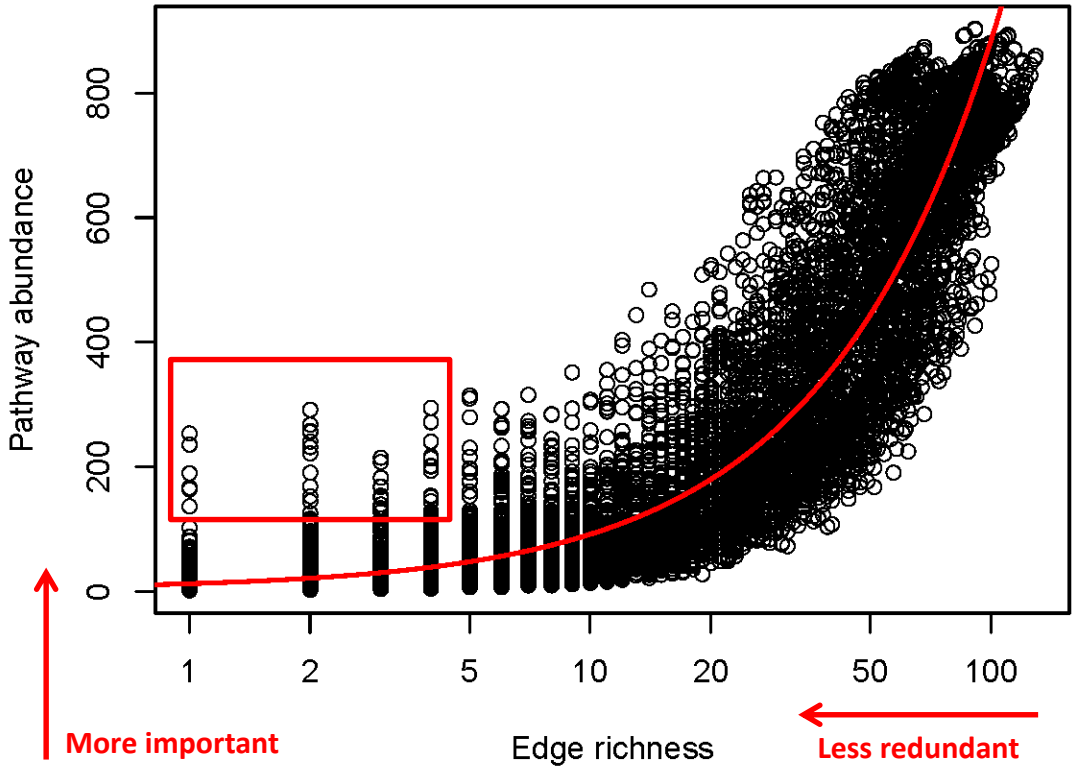
```
edge_num,nedge,post_prob,map_ratio,map_overlap,taxon,n16S,nedge_corrected,nge,ncds,genome_size,phi,clade_size,npaths_terminal,nec_termi
2,1,5.7894e-05,0.752137,117.0,Xanthobacteraceae,2.5,0.4,1.5,4774.5,5497435.0,0.583917535385,2.0,196.5,232.5,0.0781762,67.3320710715,239
4,58,0.332230982759,0.656298913793,116.431034483,Rhizobiales,2.0,29.0,1.33333333333,4633.66666667,5253297.66667,0.588265927053,3.0,200.
13,28,0.0243517321429,0.701797428571,119.892857143,GCF_000013745.1_Rhodopseudomonas palustris BisB18_strain=BisB18,2.0,14.0,1.0,4827.0,
17,1,0.0320233,0.710744,121.0,Rhodopseudomonas palustris,1.5,0.666666666667,1.0,4496.0,5112186.5,0.58557909086,2.0,220.5,570.0,0.020969
30,1,0.504984,0.626016,123.0,Bradyrhizobium,2.0,0.5,1.33333333333,6957.33333333,8071421.66667,0.584830640773,3.0,265.333333333,1041.0,0
43,4,0.0037217525,0.684211,114.0,Bradyrhizobiaceae,1.55555555556,2.57142857143,1.22222222222,5903.05555556,6657237.44444,0.576422832052
66,18,0.168918611111,0.637327277778,111.055555556,GCF_000019845.1_Beijerinckia indica subsp. indica ATCC 9039_strain=ATCC 9039,3.0,6.0,
71,1,0.106355,0.646018,113.0,GCF_001402875.1_Blastochloris viridis_strain=ATCC 19567,3.0,0.333333333333,1.0,3095.0,3724841.0,0.60223949
77,31,0.40700216129,0.669976903226,114.677419355,GCF_000503895.1_Hyphomicrobium nitrativorans NL23_strain=NL23,1.0,31.0,1.0,3244.0,3653
83,1,0.130227,0.657895,114.0,Hyphomicrobium,1.0,1.0,1.0,3813.33333333,4068394.66667,0.58265791538,3.0,183.0,428.666666667,0.120205,59.8
93,2,0.2090395,0.5963305,109.0,Rhodobacteraceae,2.66666666667,0.75,2.0,4892.33333333,5479988.33333,0.579124061687,3.0,210.0,36.6666666
106,1,0.275013,0.713115,122.0,GCF_000144605.1_Brevundimonas subvibrioides ATCC 15264_strain=ATCC 15264,2.0,0.5,1.0,3238.0,3445263.0,0.6
116,1,0.893128,0.8,120.0,GCF_000013025.1_Hyphomonas neptunium ATCC 15444_strain=ATCC 15444,1.0,1.0,1.0,3430.0,3705021.0,0.61602330171,1
123,1,0.281426,0.60177,113.0,GCF_000444995.1_Paracoccus aminophilus JCM 7686_strain=JCM 7686,4.0,0.25,6.0,4391.0,4873118.0,0.562415143
132,3,0.288385016667,0.763779666667,127.0,GCF_000018145.1_Dinoroseobacter shibae DFL 12 = DSM 16493_strain=DFL 12,2.0,1.5,6.0,4009.0,44
135,1,0.999999,0.704,125.0,GCF_000812665.2_Halocynthiibacter arcticus_strain=PAMC 20958,3.0,0.333333333333,2.0,3888.0,4376120.0,0.5577
136,2,0.171879,0.589744,117.0,GCF_000154785.2_Roseobacter litoralis Och 149_strain=Och 149,1.0,2.0,4.0,4392.0,4745450.0,0.525175414733,
137,21,0.0279111571429,0.691457714286,124.095238095,GCF_000014045.1_Roseobacter denitrificans OCh 114_strain=OCh 114,1.0,21.0,5.0,3977.
138,1,0.0716669,0.616667,120.0,Rhodobacteraceae,1.0,1.0,4.5,4184.5,4538342.0,0.538203807184,2.0,235.0,828.5,0.0262632,58.0555744627,264
141,24,0.4833585,0.685811625,121.625,GCF_000738435.1_Planktomarina temperata RCA23_strain=RCA23,2.0,12.0,1.0,3081.0,3288122.0,0.546286
```

Part 3 – Some fun applications: Metabolic redundancy



Bowman and Ducklow, 2015

Can we identify *functional redundancy* in microbial communities?



- The abundance of a pathway across samples can be accurately predicted from the number of genomes it appears in (red line)
- We consider abundance to be a proxy of ecological importance
- Abundant pathways with low redundancy *could* be ecologically important but not widely distributed taxonomically

Bowman and Ducklow, 2015

Table 5. Pathways of special biogeochemical interest identified through metabolic inference[a]

Function	Pathway[b]	Sanger studies	Hatam et al. (2014)	Bowman et al. (2012)
CO_2 fixation	CO_2 fixation into oxaloacetate (anapleurotic)	*Pseudoalteromonas haloplanktis* TAC125	*Polaribacter* MED152, *Acidimicrobiales* YM16-304	*Psychrobacter cryohalolentis* K5, *Polaribacter* MED 152
Antibiotic resistance	Triclosan resistance	*Pelagibacter ubique* HTCC1062, *Polaribacter* MED152	*Polaribacter* MED152, *Leadbetterella byssophila* DSM17132, *Thiomicrospira* spp., *Gloeocapsa* PCC7428, *Acidimicrobiales* YM16-304, *Janthinobacterium* spp.	*P. cryohalolentis* K5, *Polaribacter* MED152, GSOS
C1 metabolism	Formaldehyde oxidation II (glutathione-dependent)	*Colwellia psychrerythraea* 34H	*Gloeocapsa* PCC7428, *Marinobacter* BSs20148, *Glaciecola nitratireducens* FR1064	*Octadecabacter antarcticus* 307
Choline degradation	Choline degradation 1	*C. psychrerythraea* 34H	*Acidimicrobiales* YM304	*P. cryohalolentis* K5, *O. antarcticus* 307
Glycine betaine production	Glycine betaine biosynthesis I (Gram-negative bacteria)	*C. psychrerythraea* 34H	*Acidimicrobiales* YM304	*P. cryohalolentis* K5, *O. antarcticus* 307
Halocarbon degradation	2-chlorobenzoate degradation	*P. cryohalolentis* K5	*Polaromonas naphthalenivorans* CJ2	*P. cryohalolentis* K5
Mercury conversion	Phenylmercury acetate degradation	*Marinobacter* BSs20148, *P. haloplanktis* TAC125, *Octadecabacter arcticus* 238	*Belliella baltica* DSM15883, *Bordetella petrii*	*O. antarcticus* 307
Nitrogen fixation	Nitrogen fixation	*Coraliomargarita akajimensis* DSM45221	*C. akajimensis* DSM45221, *Methylomonas methanica* MC09, *Aeromonas* spp.	*C. akajimensis* DSM45221
Sulfite oxidation	Sulfite oxidation II/III	*Pelagibacter ubique* HTCC1062	*Cellvibrio japonicus* UEDA107	GSOS
Sulfate reduction	Sulfate reduction IV/V	*Halomonas elongata* DSM2581, *Psychrobacter arcticum* 273	*Vibrio vulnificus* YJ016	GSOS
Denitrification	Nitrate reduction I/VII	*C. psychrerythraea* 34H	*C. japonicus* UEDA107	-

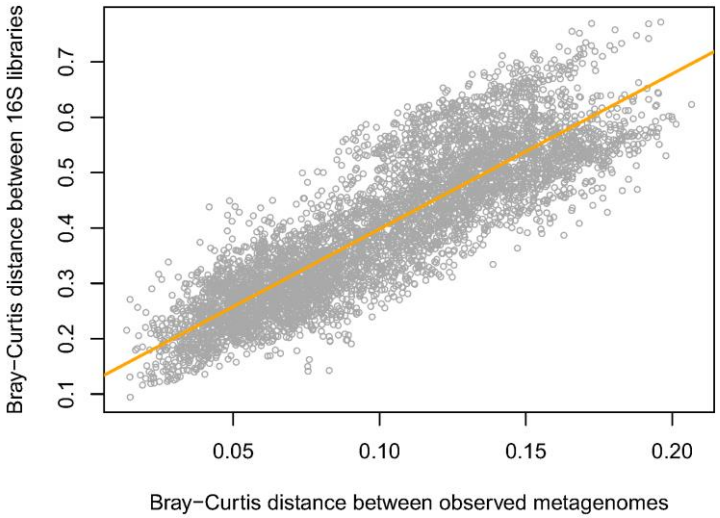
[a] Taxonomy of the nodes contributing the pathways in each dataset are given in the respective columns.
[b] Complete pathway names are according to the MetaCyc nomenclature.

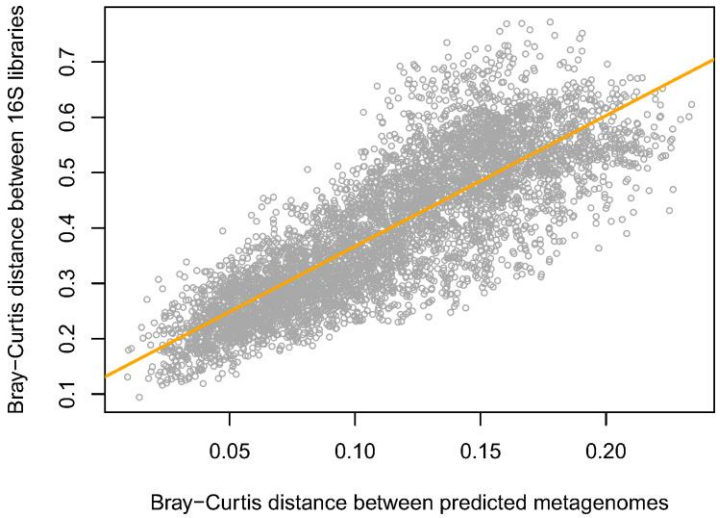
doi: 10.12952/journal.elementa.000072.t005

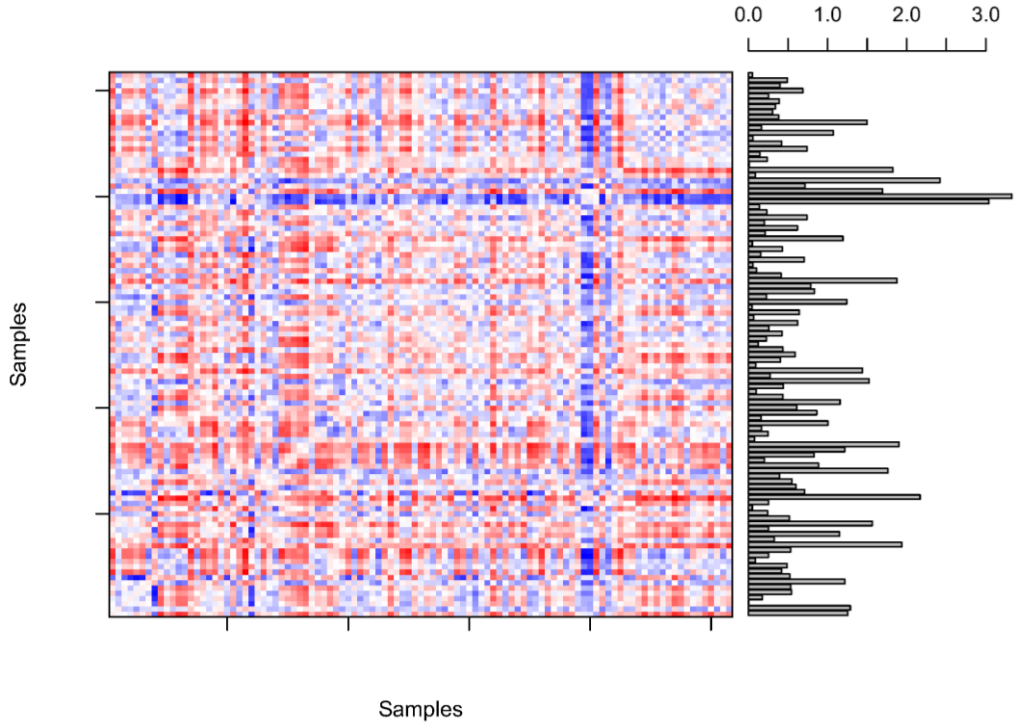
Bowman, 2015

Reminder – paprica and similar techniques just formalize what we already do…

Part 3 – Some fun applications: Microbial dark matter







Bowman and Ducklow, in prep

The paprica database consists of enzymes organized by genome, why not exploit this for metagenomics/transcriptomic analysis?

Building the paprica-mgt database

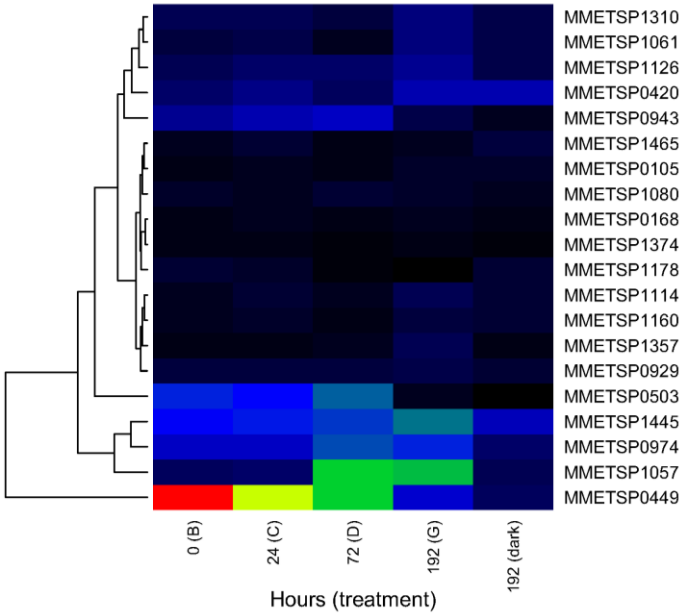
1. Collapse database to nonredundant CDS (nt space)
2. Create csv database of nonredundant CDS, enzyme number, product name, genome, etc.

Using the paprica-mgt database

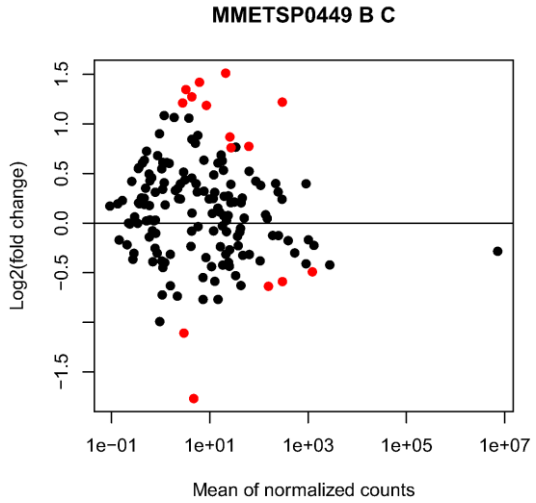
1. Use BWA to search MTs against database
2. Parse SAM format alignments to produce an abundance table of EC numbers, by genome

- This allows you to observe differential expression even as the relative abundance of the target organism(s) change over time

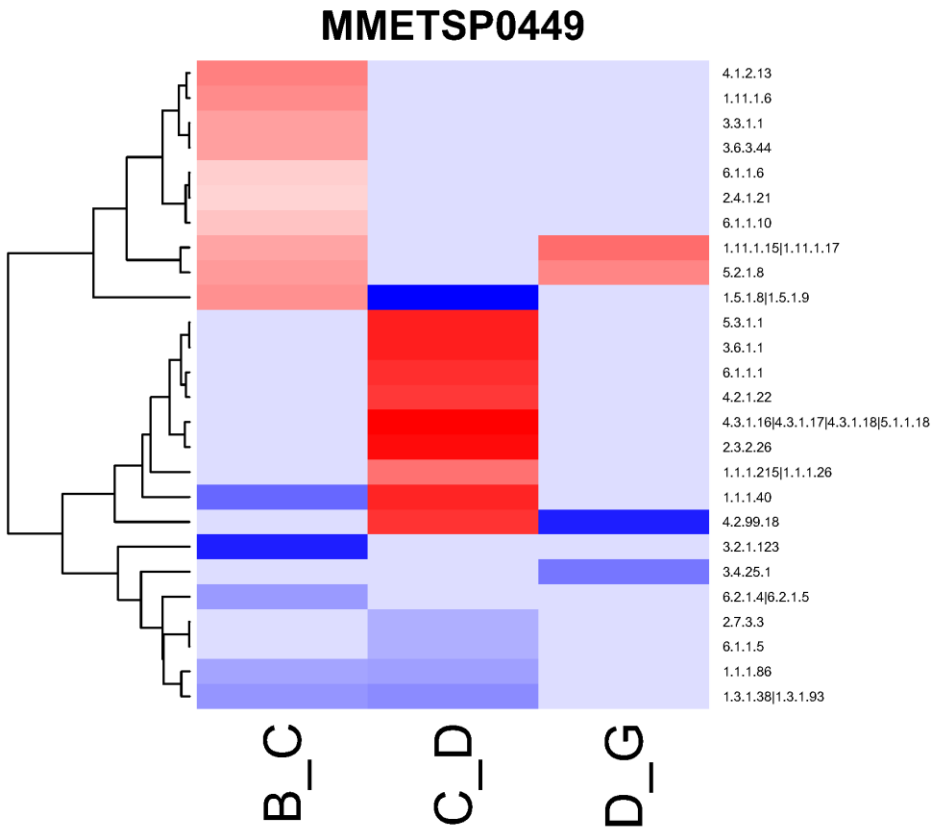
"Genome" abundance based on total reads placed



Differentially expressed genes for MMETSP0449



Differentially expressed genes for MMETSP0449, a closer look



Thanks!

Tutorials: www.polarmicrobes.org
Code: https://github.com/bowmanjeffs/paprica
Vbox appliance: http://www.polarmicrobes.org/extras/paprica-demo.ova

If there is interest would be happy to organize a training session. Let me know.