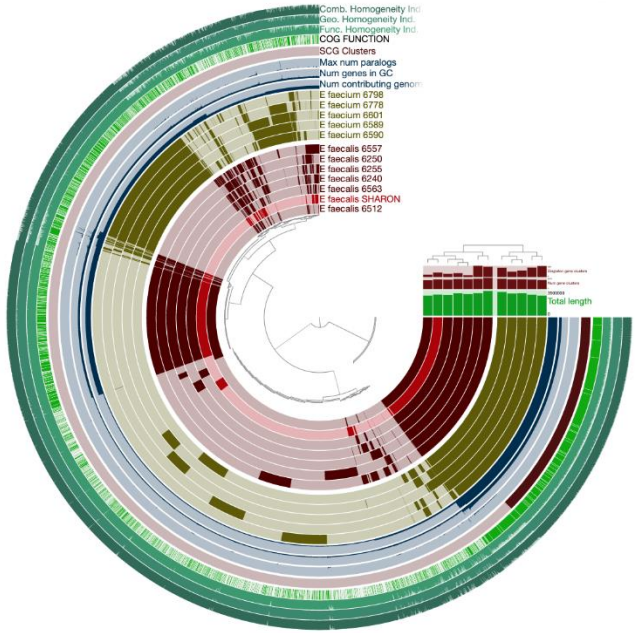
Introduction to (meta)genomic visualization

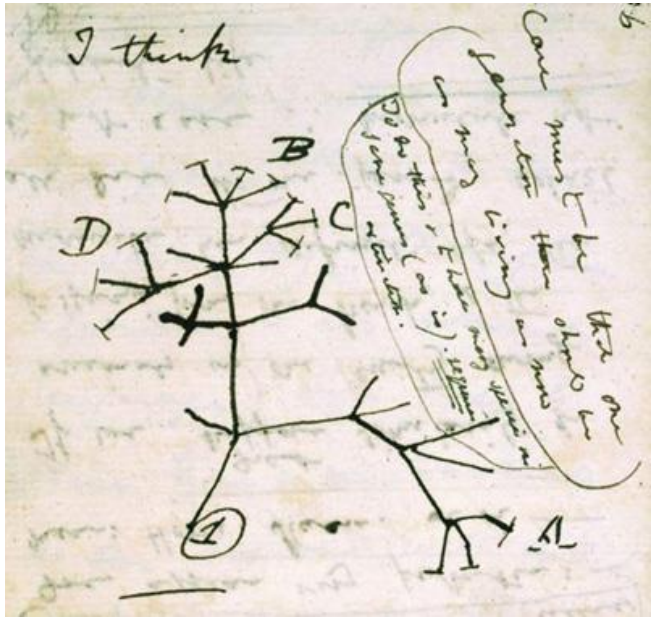
- (Meta)genomic data are complex and difficult to visualize.
- Craft a clear take away for each figure/panel. Refine towards this message.
- “Rules” for selecting plot, color, typeface, etc. are often simply trends. BUT Knowing the strengths/limits of design elements improves communication.
- Comprehensive software like ANVIO exists, and looks great (side panel). But, its constrained ecosystem and high memory footprint create accessibility issues or are inconvenient.
- Programmatic data visualization can require custom solutions/software.
- We will explore the genomic visualization I use in my research. For in-depth resources for scientific data visualization see: bit.ly/GenomeViz



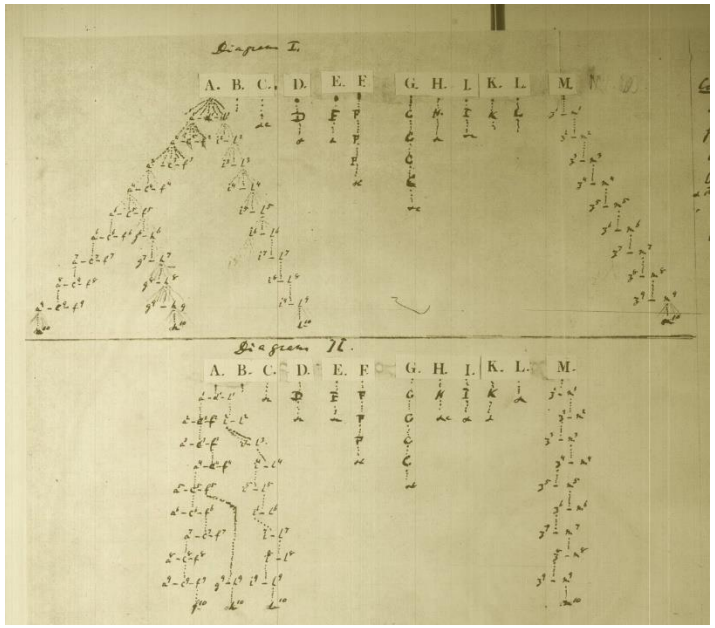
Genome-Resolved Metagenomics Tutorial [http://merenlab.org]

Phylogenetic Trees

- Originally described morphological divergence
- Terminal leaves represent species
- Nodes represent ancestors
- Branch lengths can represent degree of change (Phylogram)



Darwin, 1837. Notebook B.

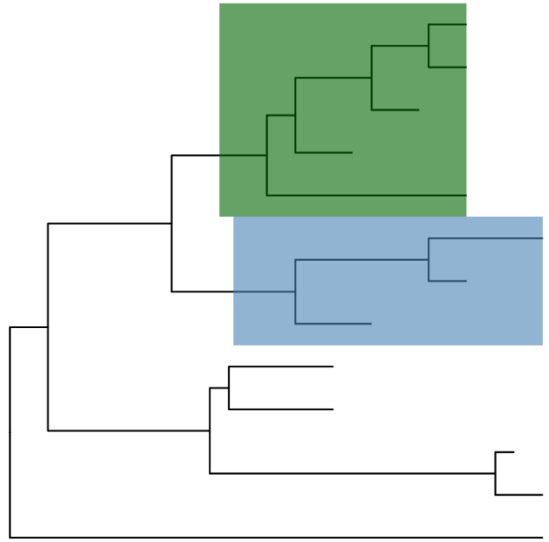


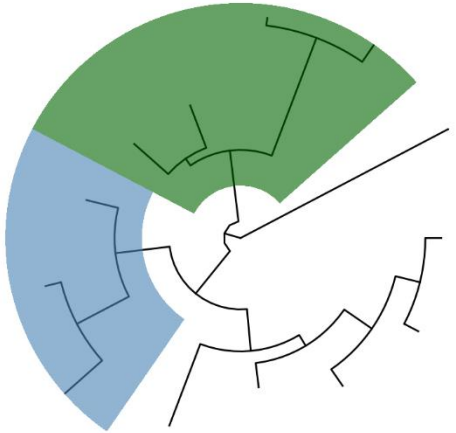
Darwin, 1856-1858. Unfinished sketch.

Phylogenetic Trees

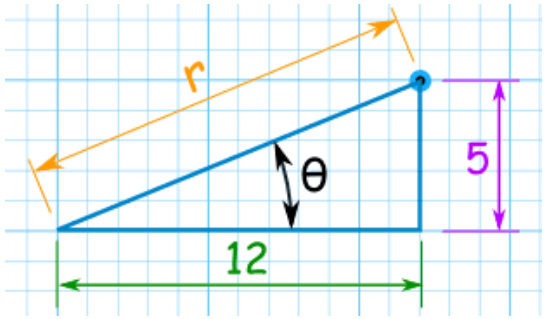
- Rectangular trees:
  - Easy to read with few (<50) leaves
  - Linear structure of categorical variable along axis

- Circular trees:
  - Compact visualization of many (>50) leaves
  - Data must be mapped to polar coordinates





Converting linear data to polar coordinates (e.g., in *circlize* R package)



$r = \sqrt{(12^2 + 5^2)} = 13$

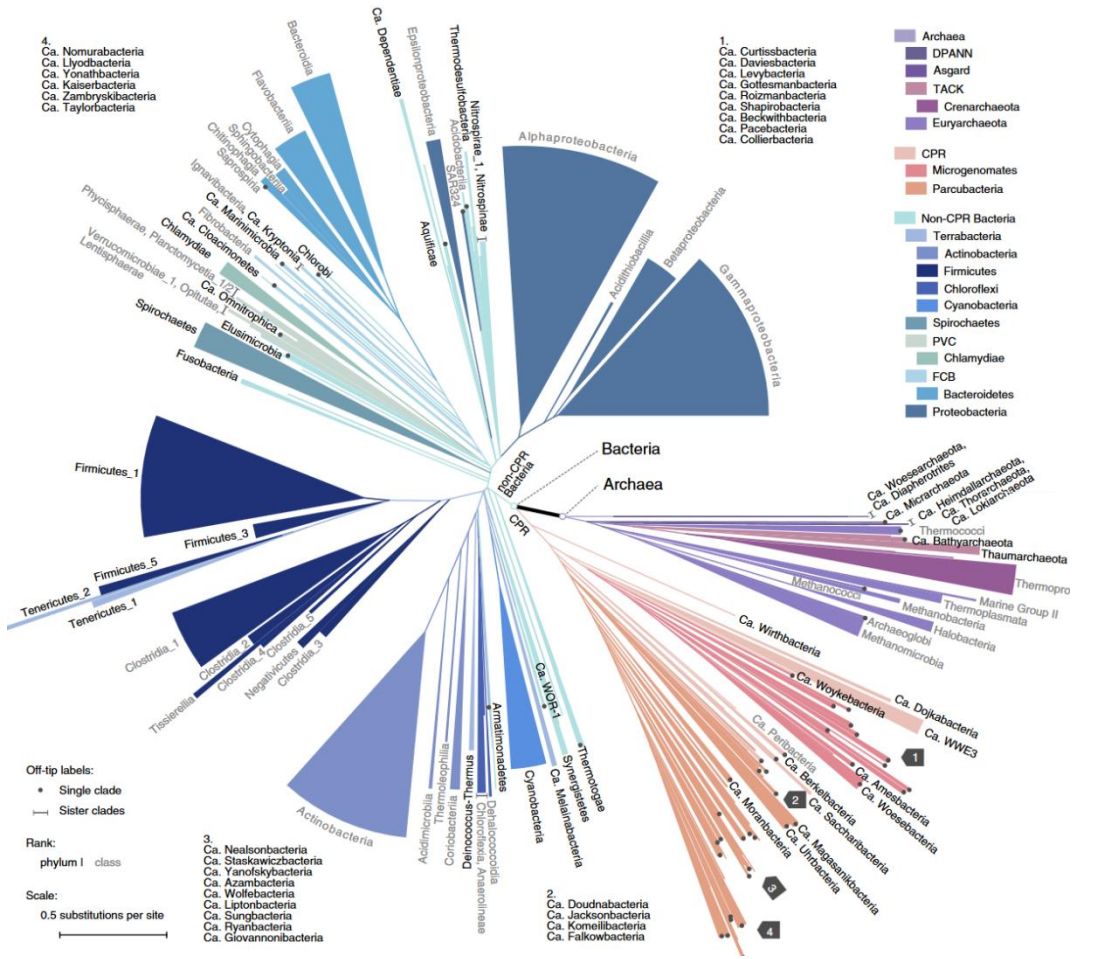
$\theta = \tan^{-1}(5 / 12) = 22.6°$

github.com/YuLab-SMU/ggtree

Phylogenomics

- Whole-genome phylogeny can be analyzed with conserved genes or ribosomal proteins.
  - Contrast with 16S rRNA or ANI
- Maximum Likelihood (ML) trees show divergence of nucleic or amino acid characters according to substitution models.
- Calculated by a variety of algorithms, e.g.,
  - RAxML (slower, more accurate)
  - FastTree (faster, less accurate)
  - IQTree (comprehensive model selection and branch length calculations)

Combined 381 protein alignments into 1 FastTree and 2 RAxML trees per alignment. Reconstructed to final tree with RAxML. Visualized with iTOL v4.



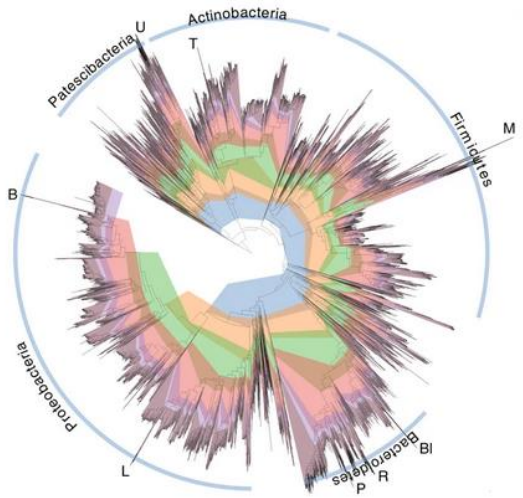
Zhu et al., 2019 Nat. Comm.10:5477

Genome Taxonomy Database (GTDB)

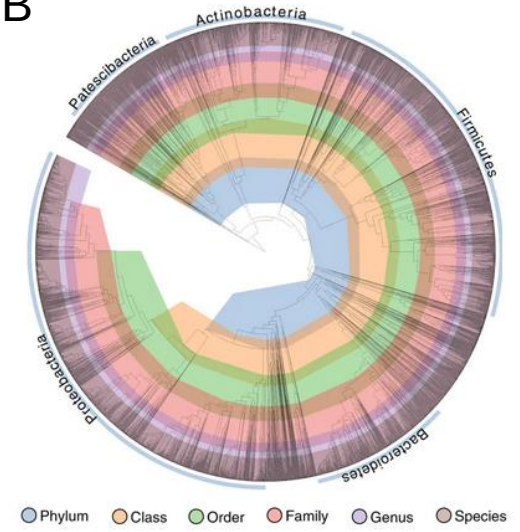
- Whole-genome phylogeny using 120 bacterial or 122 archaeal single-copy genes.
- Comprehensive genome-based taxonomy using relative evolutionary distance.
- Characterize taxonomy of novel genomes using tree placement and average nucleotide identity (ANI) with GTDB-Tk (https://github.com/Ecogenomics/GTDBTk)
- We will use GTDB taxonomy (https://gtdb.ecogenomic.org/) to organize data generated in metagenomic and metatranscriptomic datasets.

RED: Taxonomic rank normalization

NCBI



GTDB



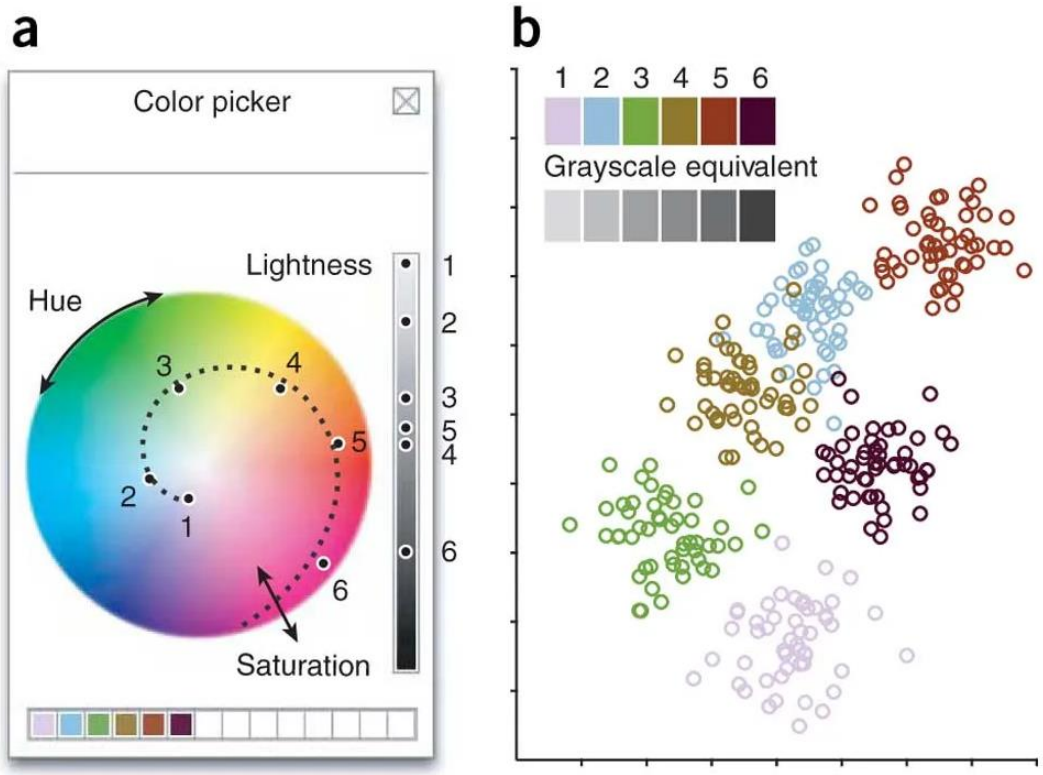
Park et. al., 2018 Nat. Biotech. 36, 996

A note on colour.

Be careful assigning multiple colours to quantitative data.

Try using either a single colour or two with a neutral midpoint.

If you do use multiple, altering hue, saturation and lightness can generate a colour set distinguishable even in grayscale.

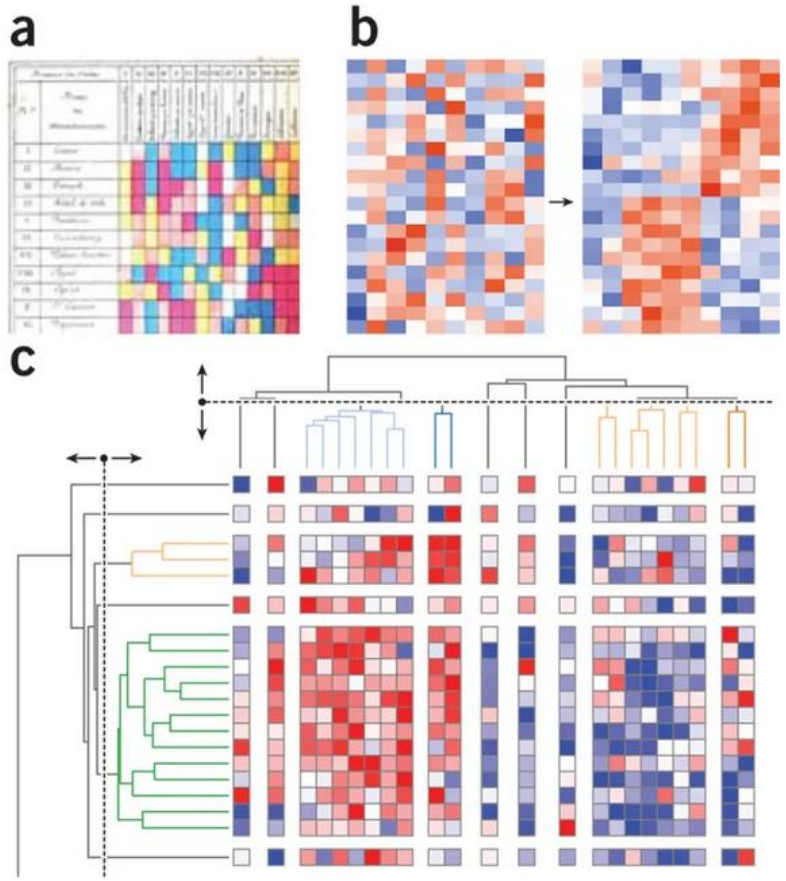


Single Colour:

Two Colour:

Three Colour:

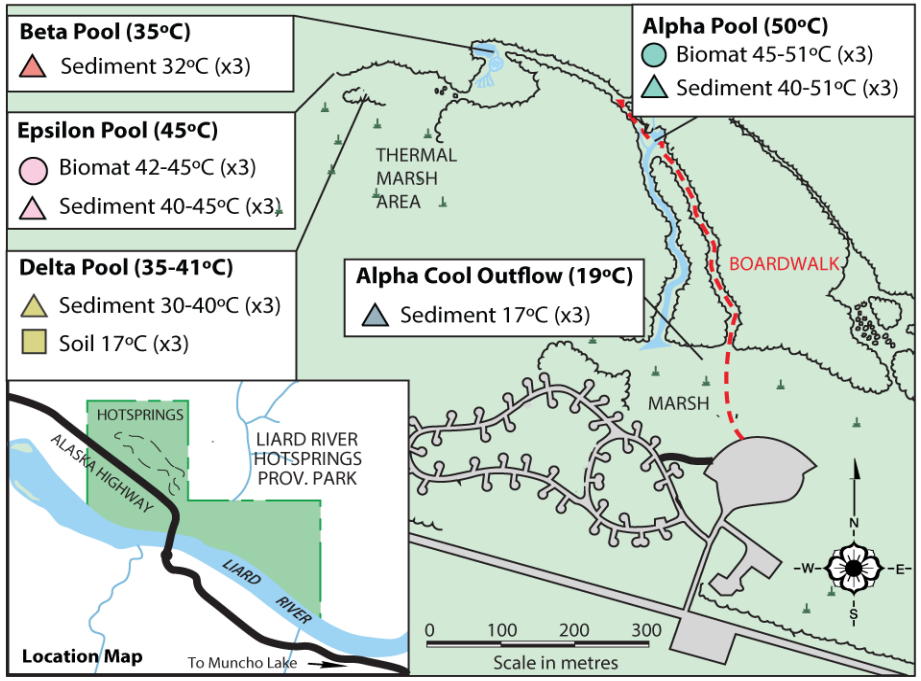
Multiple Colour:



We will use metagenomics and metatranscriptomic approaches to characterize the metabolic capacity of microbial consortia in a **thermal swamp.**

What is a “thermal swamp” anyway?



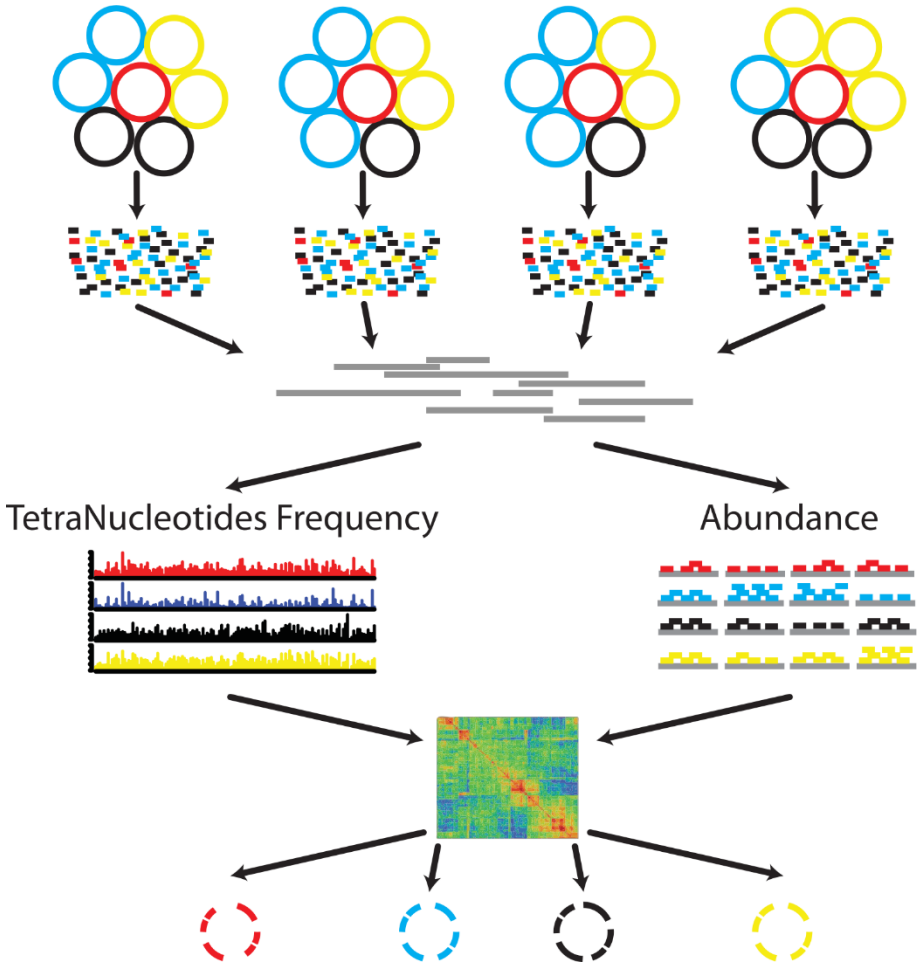



Liard River Hot Springs in northern British Columbia

- Temperature from 30 to 55ºC

- Hot spring complex and continually-warmed marshland harbouring unique flora, fauna and microorganisms

Metagenome Assembly and Binning



Preprocessing

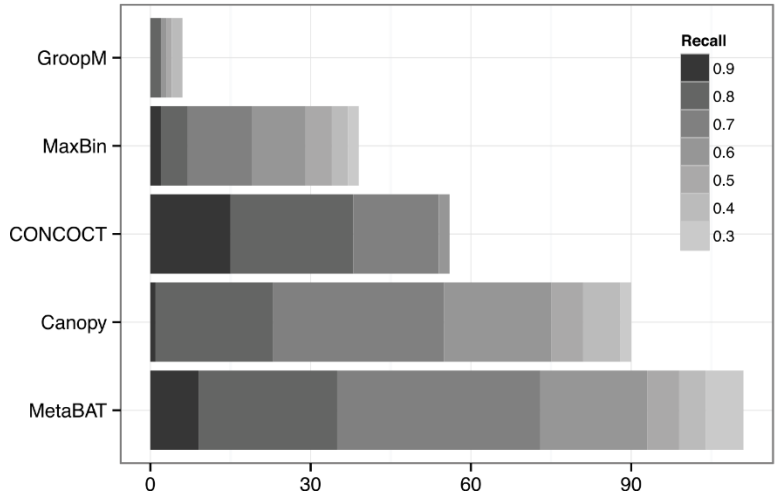
1. Samples from multiple sites or times
2. Metagenome libraries
3. Initial de-novo assembly using the combined library

MetaBAT

4. Calculate TNF for each contig
5. Calculate Abundance per library for each contig
6. Calculate the pairwise distance matrix using pre-trained probabilistic models
7. Forming genome bins iteratively

Assembly: MegaHit 1.1.3

Binning: MetaBAT2



Kang et al., 2015. PeerJ 26336640

For each genome assembled from the Liard River Hot Springs metagenome, our job will be to answer the following biological questions:

1. What are they? (Taxonomic classification)
2. Can we trust them? (Assembly statistics)
3. Where are they? (Abundance calculations)
4. What can they do? (Function/Pathway analysis)
5. What are they doing? (Metatranscriptomics)