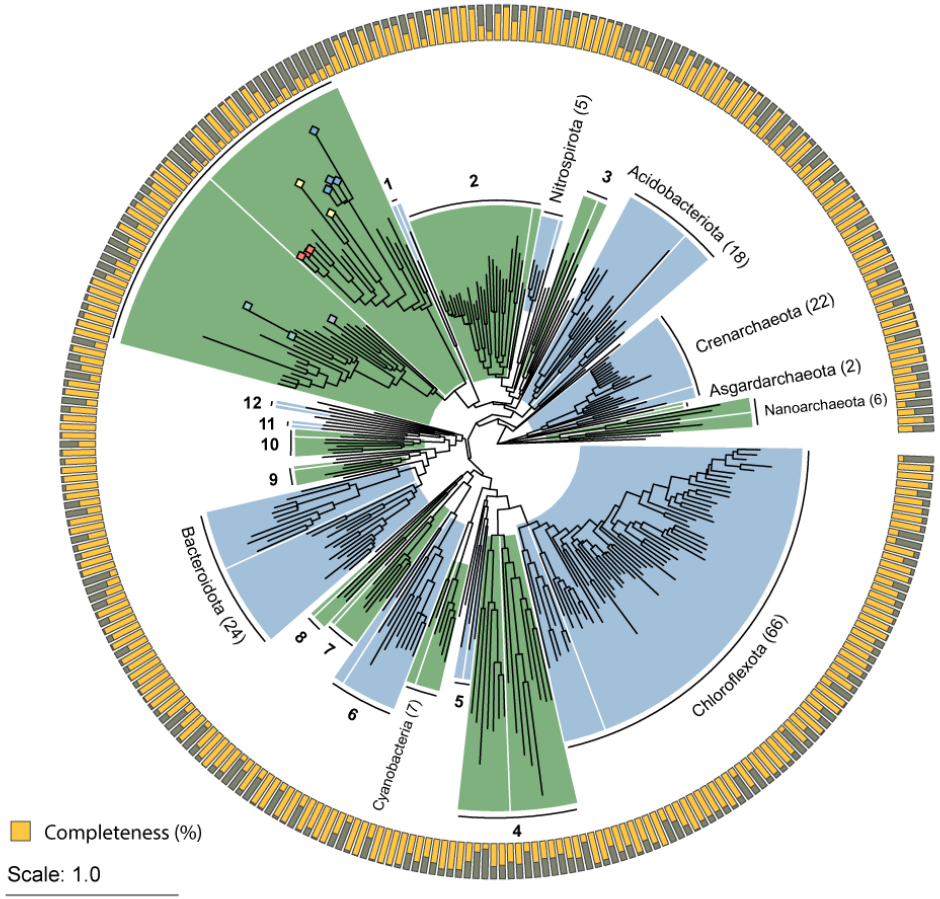
Visualization of Metagenomic and Metatranscriptomic Data in R

March 17 and 19, 2020



David Levy-Booth

dlevyboo@mail.ubc.ca

@DavidLevyBooth

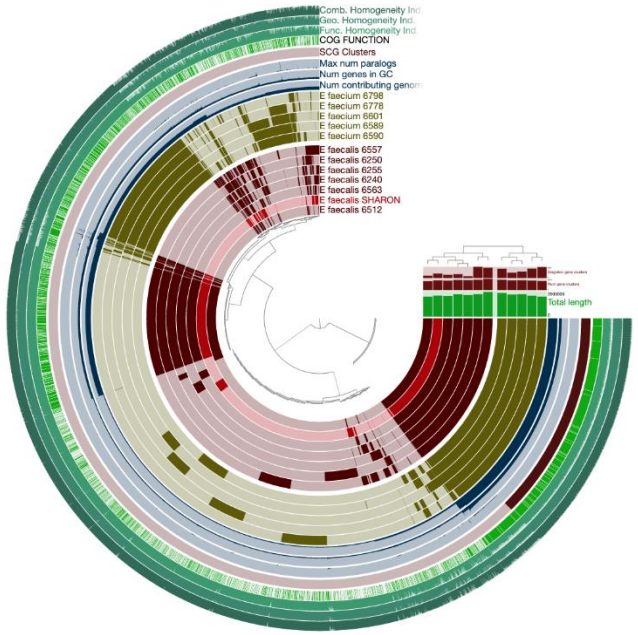
EDUCE

ECOSCOPE

routing the microcosmos

Introduction to (meta)genomic visualization

- (Meta)genomic data are complexity 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.
- Out-of-the-box software like ANVIO exists, and looks great (side panel). But, constrained ecosystem and high memory footprint creates accessibility issues or inconvenient. New approaches require custom solutions/software.
- We will explore the genomic visualization I use in my research. For an in-depth discussion of scientific data visualization see: bit.ly/GenomeViz



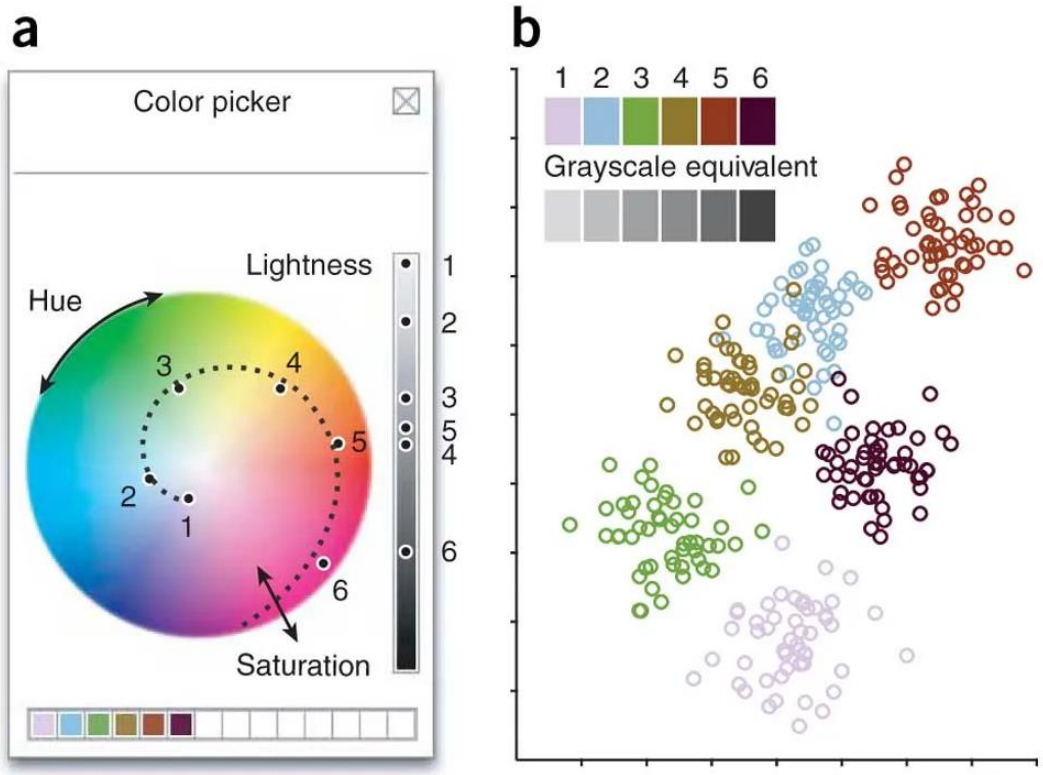
Genome-Resolved Metagenomics Tutorial [http://merenlab.org]

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.

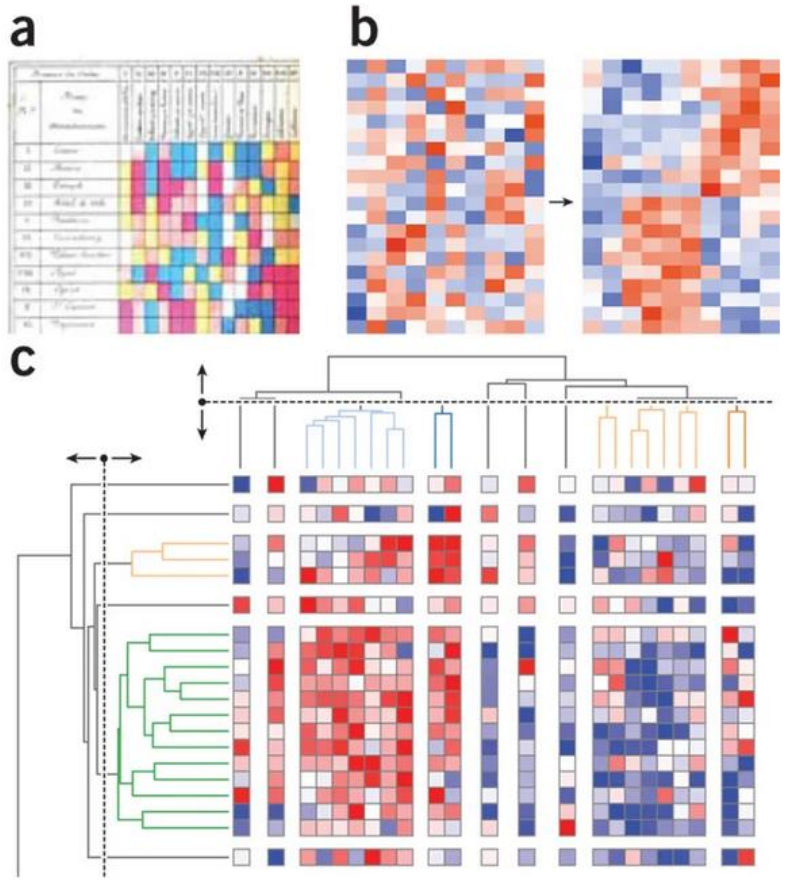


Single Colour:

Two Colour:

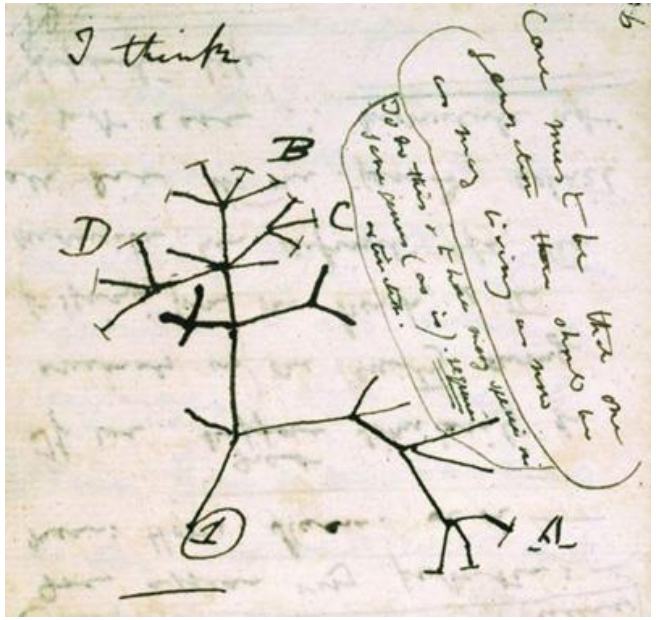
Three Colour:

Multiple Colour:

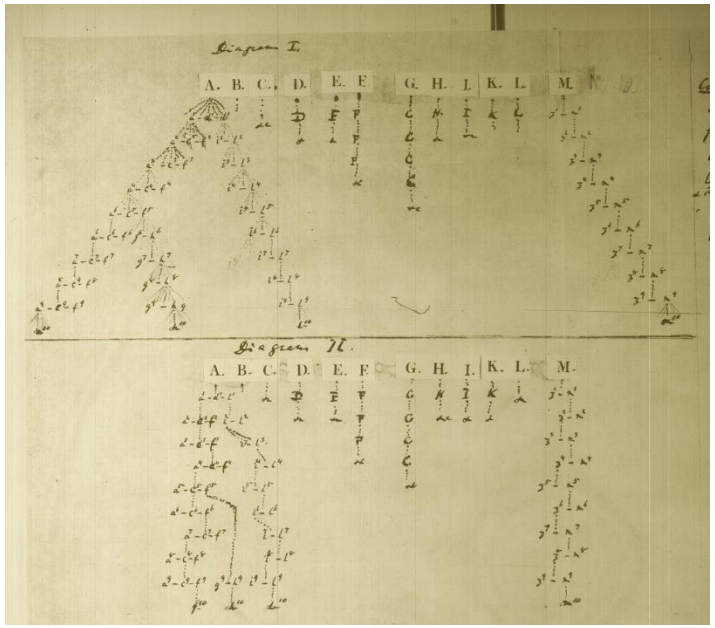


Phylogenetic Trees

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



Darwin, 1837. Notebook B.

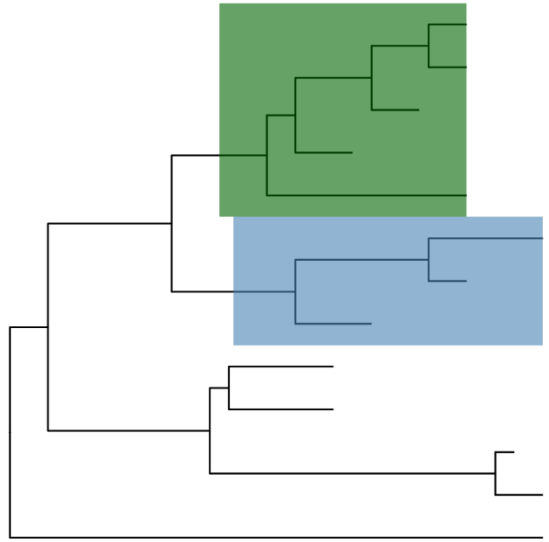


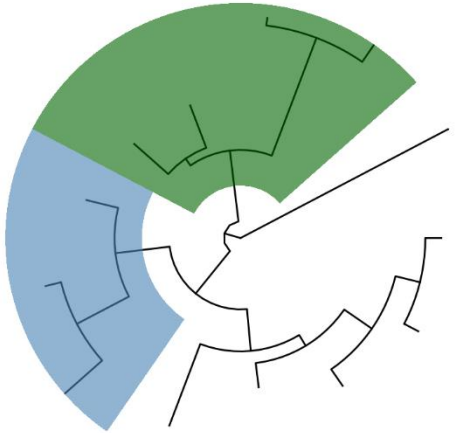
Darwin, 1856-1858. Unfinished sketch.

Phylogenetic Trees

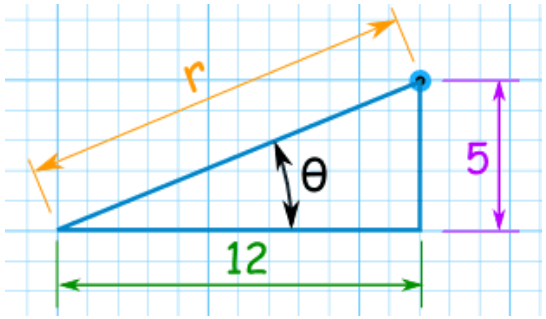
- Rectangular trees:
  - Easy to read with few (<50) leaves
  - Linear structure of categorical variable along axis

- 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)



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

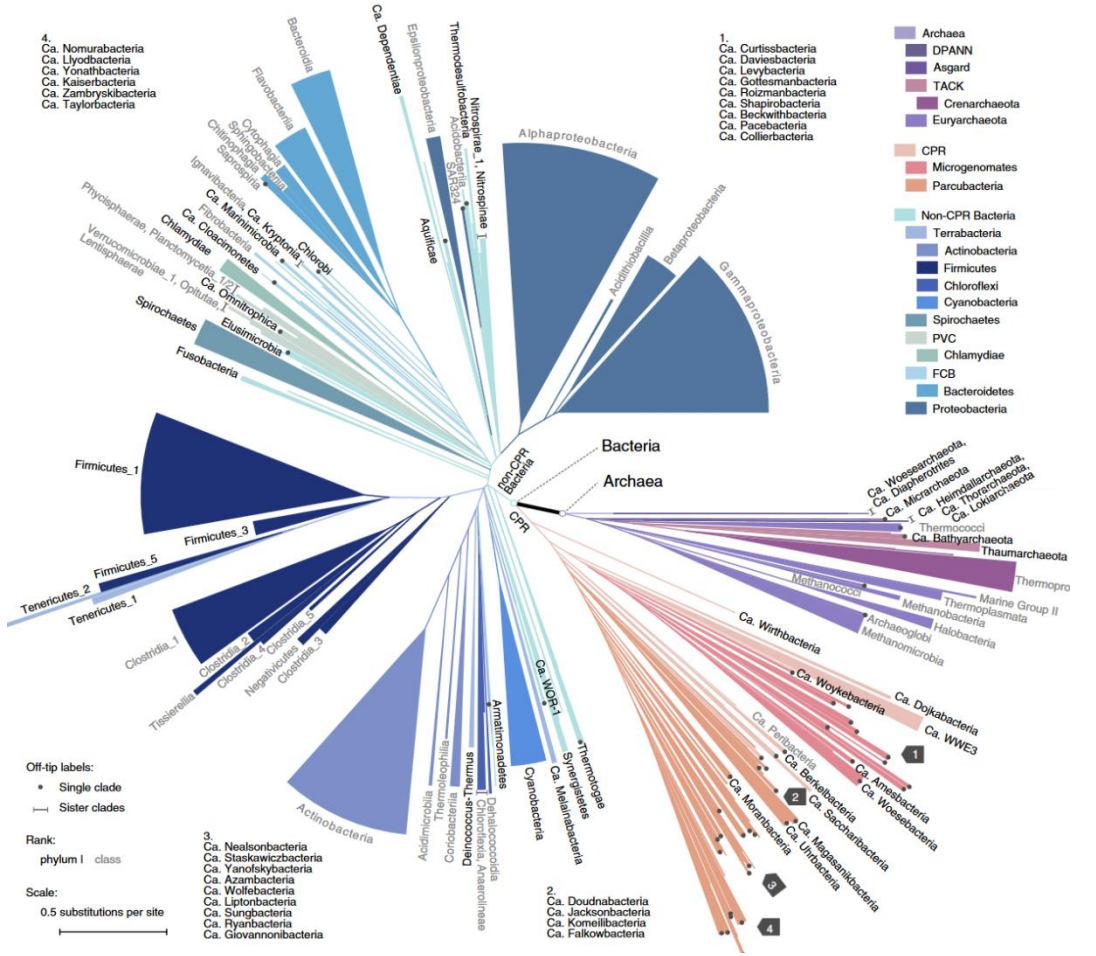
$\theta = \tan^{-1}(5 / 12) = 22.6°$

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)
- Can be visualized interactively (e.g., iTOL) or programmatically (e.g., ggtree)

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

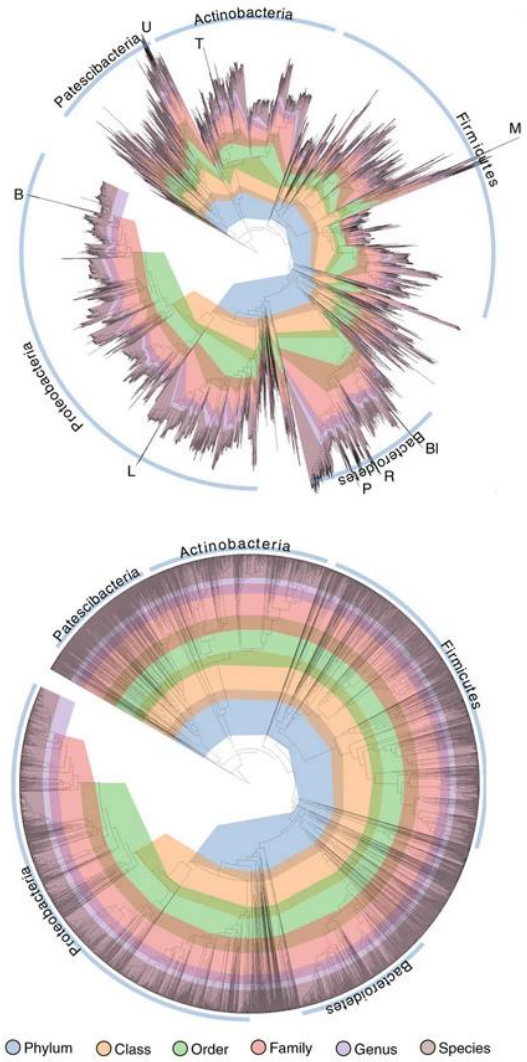


Zhu et al., 2019 Nat. Comm.10:5477

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.
- We will use GTDB taxonomy to organize data generated in metagenomic and metatranscriptomic datasets.

Rank normalization through RED

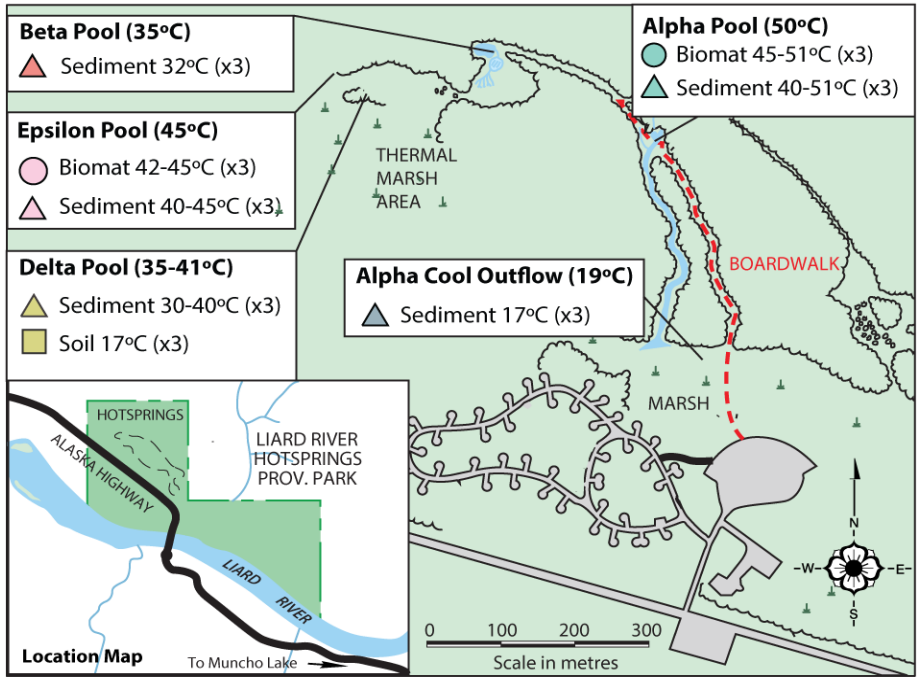


Park et. al., 2018 Nat. Biotech. 36, 996

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?



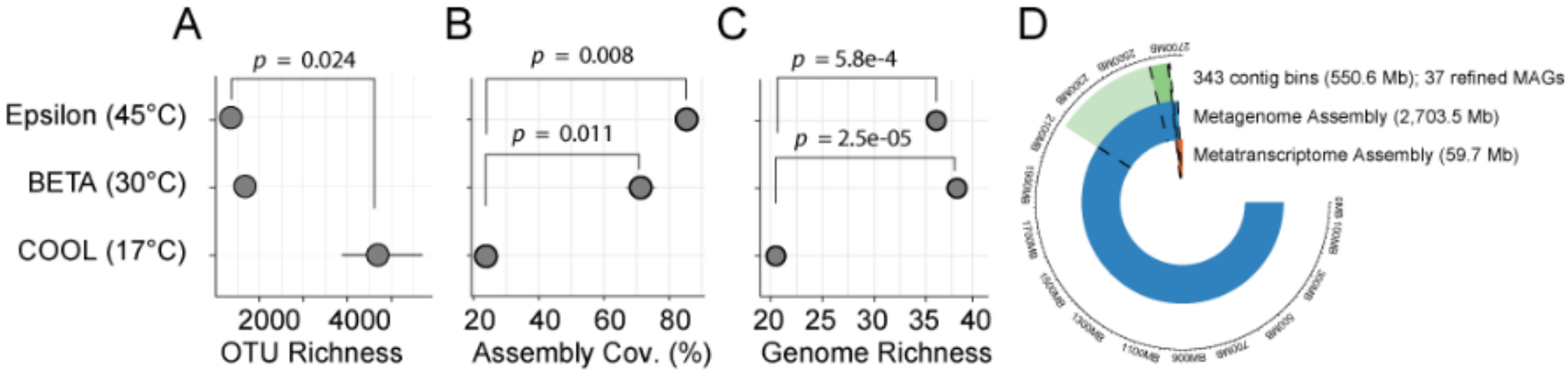



Liard River Hot Springs in northern British Columbia

- Temperature from 30 to 55ºC

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

Computationally tractable thermophilic communities



Targeting thermophilic metagenomes increased overall assembly by ~67% relative to mesophilic communities

For each genome assembled from the Liard River Hotsprings 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)