Culture Club: Exploring Genome Annotations

We'll be working with a metagenome-assembled genome (MAG) from a Trunk River sample taken in 2018 (near_lem_2018). This MAG had 13x coverage in the metagenome (compared to 2000x for a green sulfur bacteria in the same sample!).

Bin ID: `near_lem_2018_bin.22`

Bin Id	near_lem_2018_bin.22
Coverage	13.0
Completeness	96.28
Contamination	3.21
Genome size (bp)	2,558,784
# scaffolds	244
N50 (scaffolds)	15,575
Mean scaffold length (bp)	10,486
Longest scaffold (bp)	62,331
GC	64.1
# predicted genes	2,655

Open reading frames (ORFs): 2,614 (called with prodigal v2.6.3)

- *Why are there fewer ORFs than predicted genes in the table above?*

Method 1: KEGG

genes Annotations for this bin were done using blastKOALA (https://www.kegg.jp/blastkoala/). KEGG provides a web server where you can upload an amino acid .fasta file of your ORFs and receieve a table of K numbers (KEGG identifiers) assigned to each. This can take a little while, so you can download pre-computed results here: https://drive.google.com/file/d/1M1rQzisXkrHvo8h74WL3eRanD_0kxRwq/view?usp=sharing.

KEGG also offers online tools for viewing and interpreting these annotations. Upload the BlastKOALA results to the Reconstruct tool here: https://www.genome.jp/kegg/mapper/reconstruct.html. Click "Browse" to upload your file, then "Exec" to execute.

You'll see a results page with four tabs. Generally, click on the number to the left of an entry to open page with more information. Clicking the number in parentheses to the right will show associated contigs and their K-number assignments.

- Pathway: This tab places your annotations in the context of metabolic pathways. Clicking the number to the left will open a pathway map with genes your organisms has highlighted in green.
- Brite: BRITE is a hierarchical classification system, rather than pathway-focused. This view is useful if you want to look at all the enzymes of a certain type (i.e. glycosyltransferases)
- Brite table: Shows some classes of conserved enzymes across domains of life. The ones your organism has are in red.
- Module: This another pathway-based view which lets you easily see which functional subsets of metabolism your organism has. By default, only complete modules (where your organism has all the genes it needs) are shown; this can be adjusted at the top of the page.

Method 2: METABOLIC

METABOLIC (https://github.com/AnantharamanLab/METABOLIC) is a tool that profiles metabolic and biogeochemical traits. It works on both genomic and metagenomic data.

Selected outputs for bin22 can be dowloaded here: https://drive.google.com/drive/folders/1ksZNMByXQwWw1VZu9nszujBGE1DRcEPd?usp=sharing. You should find a spreadsheet with hit information and a series of PDFs showing which parts of major cycles are present in the data.

Culture Club Questions

Compare KEGG vs METABOLIC output. Find two differences that seem MAJOR to you. What are they?

Does it have...

- A complete TCA cycle?
- Complete set of genes to do glycolysis? If so, which version (e.g. EntnerDoudoroff vs EmbdenMeyerhofParnas)?
- The ability to respire? If so, which terminal electron acceptors?
- The ability to degrade polysaccharides? Any clues on which ones?
- Any genes involved in the many sulfur-compound transformations we have talked about?
- The ability to fix nitrogen?
- The ability to fix carbon?