**Methanothermococcus SAGs: outline of findings**

**Story:**

What type of strain-level variation do we see among Methanothermococcus SAGs isolated from a single hydrothermal fluid sample?

Insights into:

Extent of pangenomic variation in the vent environment

What type of variation we see (what types of variable genes/islands etc)

Whether specific strains are more successful in specific environments and why

Whether these show signs of natural selection or if we evidence of drift.

Overall goal: learn about natural genomic variation and selection in vents. (niche partitioning)

**General results:**

Isolated from Ginger Castle, Von Damm, pretty sure 2012

Used CISA to create a meta-assembly based on a5, SPAdes, CLC, idba-ud

Varying completeness (see spreadsheets), up to 85% complete

Phylogenetic tree (Phylosift): E23 and N22 are very similar; M21 and C09 are very similar; K20 in a clade with C09 and M21

All-v-all similarity matches clonal phylogeny

**SAG abundance in metagenomes and metatranscriptomes: “success” and “activity”**

Mapped MCR metagenomes to SAGs. Some of them were REALLY successful in specific bins. We see variation. E23 and N22 in particular seem to be quite successful, C09 does pretty well in some. M21 middling, K20 terrible (but also very incomplete)

So, question: Why are E23 and N22 so succesful?

All of the SAGs were most abundant in FS848, FS866, FS881. All are Von Damm. E23 and N22 were disproportionately abundant in FS866 and FS881, which are two of the hottest samples we took (though FS874 was hot too). FS848 was actually not that hot. Note that they WERE highly expressed in both FS848 and FS881 (no data for FS866). The place where these were the most “turned on” was the vent they were sampled from.

**Variable genome:**

***Pangenome:***

We added 665 new genes to the Methanothermococcus pangenome (adding 5 strains to the two that were prev. sequenced)

*Differences between SAGs:*

We did ITEP (and JGI PA) on the five genomes and refs. Definitely some genes present in SAGs that are missing from refs. Include some restriction-modification genes in C09 and M21. N22 has cell motility genes that the others lack.

Some mcrA and nif genes were different in E23 and N22 that were different from others.

Made gene trees (mcr, nif) that show that the gene trees kinda sorta follow the clonal phylogeny. N22 has no nitrogenase genes, E23 only has a few weird ones. Many nitrogen-related proteins in C09+M21+K20 that are missing in E23+N22.

E23 and N22 also have glycosyltransferases, membrane proteins, restriction/modification systems, random genes that differentiate them from the others.

Some genomic islands apparent but hard to locate; mapping shows variation from SAG to SAG in which some genes are definitely present and others are not; unclear what those genes are though…

Lots of CRISPR genes in C09 and M21, which matches observation of CRISPR loci (see below).

**Variation due to viral infection:**

***CRISPRs***

Looked at CRISPRs to determine whether there was cross-infection, whether there is a lot of variation in the CRISPRs🡪 can give insight into how much viruses might be contributing to genomic variation, and whether different viruses infect different strains in the same vent (how much spacer overlap?)

Used CRT to look at CRISPRs. Used e-value 0.001 as match (Held and Whitaker standards). Found CRISPRs in C09, M21. K20 has a single little tiny one, none in E23 or N22.

Need to see if there are cas genes in N22 and E23: are they resistant? Use other methods?

Only spacer matches are between C09 and M21: 237 spacers in C09, 109 in M21, 18 matches between them, so 16.5% of M21 and 7.6% of C09 matched each other—rest of spacers were unique. Matching spacers were not in a row, showed that they occasionally got hit by the same viruses, but not always.

I am noticing that many of the ones that are repeats in the same locus are also found on other genomes. So the common viruses hit a lot.

M. okinawensis has many CRISPR matches with the SAGs, and was isolated in 2000.

(note that I removed one CRISPR locus from M21 (see line 14870 in notes) because I think it’s a repeat assembly)

This means either there is lots of active, frequent viral infection, with some degree of host specificity but also evidence of cross- infection. May also show that there are lots of different Methanothermococcus viruses. Or both.

CRISPR genes: C09 and M21 both have lots of CRISPR genes, no other genomes do. The csm genes all fall on the same contig in C09 and one contig in M21, but there are CRISPR loci on other contigs too. Doesn’t completely rule out the possibility that there are CRISPRs in the others, though I haven’t found any CRISPR genes in the others!

*Prophage*

Found prophage on C09 using VirSorter. Two contigs (Ga0127414\_1095 and Ga0127414\_1066 in JGI terms) have phage proteins on them. Probably a single prophage.

Contain genes related to phage tail and head; some genes match those previously seen on Methanothermococcus okinawensis.

Found the prophage on M. okinawensis using VirSorter; they do seem very similar (not identical) to each other.

There are several matches of CRISPR spacers from C09 and M21 that match prophage from M. okinawensis, with a few matches to the prophage from C09, but not that many! Interesting.

TO DO NEXT:

-write to Kelly wrighton about CRISPR pipeline

-do the SNV thing in other metagenomes, not just FS848. If there’s enough coverage.

-look at SAAVs. Are other variants more abundant? (can rely on consensus here!)

-do dN/dS if possible to compare between strains (not mapping, direct comparison)

-test of selection vs drift?

-explore the virus story more. Are there prophage? Do they match the CRISPRs? Who else matches the CRISPRs? (can we figure out who other hosts are, do they share viruses?)