**SAMPLE geospatial information:** SAMPLE\_geospatial\_metadata.xlsx

* **sample\_ID:** official names of samples.
* **well\_ID:** the alias name of the fracking well that each sample was collected from.
* **sample\_type:** input or produced fluid. Inputs are various ‘topside’ treatments of the water before being used to frack the well. Produced fluids are what we call the water that comes out of the well and represents the subsurface microbial communities.
* **sample\_subtype:** another level of sample description: either ‘day#’ for the produced fluids or a 2-3 letter abbreviation for the input types (for example: DM=drill muds).
* **days\_since\_frack:** how long has the well been in production (this is essentially a proxy of how old the well is when the sample was taken).
* **basin:** which basin the sample was recovered from.
* **formation:** which formation (within a basin) the sample was recovered from.
* **timeseries:** is the sample part of a well with temporal sampling? (yes/no/NA)

Other columns I would like to add but I need to troubleshoot some data issues first:

* Date sampled: month/year
* Salinity (conductivity):
* Salinity (Cl mg/L)
* % Fermenters
* % Methanogens
* % Sulfate reducers
* % Sulfide producers

A couple notes:

* I need to chat with Mike about how to best handle lat/long or generally location of wells. For now planning to just jitter points within a basin.
* It would be great to be able to visualize timeseries samples both in the geospatial map but also in another viewer on the page.
* Though sample\_subtype and days\_since\_frack are somewhat repetitive, I thought it would be helpful on the back end to have both columns. Would love your thoughts on this as well!

**GENOME geospatial information:** GENOME\_geospatial\_metadata.xlsx

* **MAG ID:** official name of the genome recovered.
* **basin:** the basin that the sample from which the MAG was recovered belongs to.
* **sample type:** input or produced fluid (see SAMPLE geospatial info).
* **completeness (%):** estimated genomic completeness of the MAG
* **contamination (%):** estimated genomic contamination in the MAG
* **columns F-L:** taxonomic classification, broken down by the classification level (domain, phylum class, etc. etc.).

Other columns I would like to add but I need to troubleshoot some data issues first:

* Average relative abundance by basin (assuming here each basin [n=11] would need to be an individual column)
* Produced sulfide? (Yes/No)
* Methanogen? (Yes/No)
* Sulfate Reducer (Yes/No)
* Fermenter? (Yes/No)
* Core to basin? (would list all basins that genome is considered ‘core’.. if possible?)
* NCBI taxonomy (either full string, or lowest classification)

A couple notes:

* I broke down the taxonomy by column because again I assume it will be easier for the user to be able to parse information this way, and I assume having levels of classification separated on the back end of the data would help for filtering.
* I think there is a lot of potential for cool visualizations here and would love to brainstorm more on this. Users will care most about the functional potential of a MAG(genome) as well as where it was recovered, what basins it is detected in, what topside samples it is detected in, if it is considered a ‘core’ genome for a basin... possibly others as I brainstorm.