Annaliese Meyer & Matthew Baldes
Project Proposal

Paper: A comprehensive metagenomics framework to characterize organisms relevant for planetary protection
Authors: Danko et al. 2021

This study analyzed 51 samples from clean rooms at the JPL Spacecraft Assembly Facility. Surface, air filter, and vacuum debris samples were collected from clean rooms ranging from ISO 5, most protected, to ISO 8.5, least protected. A subset of samples were replicated and analyzed for reproducibility within the study. The samples were tested for species richness and diversity, growth rate, and extremophile characteristics. The authors found no significant difference in species diversity between ISO 6 - ISO 8.5 level facilities. ISO 5 facilities were found to have lower levels of diversity; however, all facilities had measurable microbial populations, active growth, and microbes with traits that increased resistance to extreme conditions. The study places emphasis on reproducibility and the need to have standards of analysis to characterize clean room microbial communities, especially given growing planetary protection concerns in public and private space exploration initiatives.

Much of the data cleaning and initial taxonomic analysis conducted by Danko et al. (2021) utilized tools wrapped in the publicly available MetaSUB Core Analysis Pipeline (https://github.com/MetaSUB/MetaSUB_CAP). Metagenome-assembled genomes (MAGs) were assembled, binned and annotated with other open-sources tools such as MetaSPAdes and Prodigal. The raw sequencing data is provided at https://pngb.io/jpl-clean-rooms along with associated metadata.

We propose to reanalyze the samples listed below, which represent a selection of 12 samples from the ISO 6-8.5 facilities and 12 samples from ISO 5 facilities. Each sample consists of two sets of approximately 2.5 Gb of raw data. We estimate that the 24 samples represent 144 Gb of data to be processed. We intend to recreate all figures except those pertaining to novel genomes discovered in the dataset as a whole (e.g. Fig.4), as we will be unable to perform analyses on the entire complement of studied samples.

Table 1. Proposed samples for reanalysis.

Sample #	Sample Type	Sample Description	total kmer count
2_5	Wipe Solution	ISO 6 (103-102C) Facility	23994241
2_6	Wipe Solution	ISO 8.5 (103-110) Facility	23990109
2_7	Wipe Solution	ISO 8.5 (103-110) Facility	23990401
2_8	Wipe Solution	ISO 8.5 (103-110) Facility	23992452
2_9	Wipe Solution	ISO 8.5 (103-110) Facility	23995634
3_7	Filter Solution	ISO 8 (318-123) non-carbon filter	23987807
3_8	Filter Solution	ISO 7 (306) non-carbon filter	23990145
4_1	Vacuum Particle Solution	ISO 7 (179-121) Vacuum #1	23992641
4_2	Vacuum Particle Solution	ISO 7 (179-121) Vacuum #2	23993799
4_3	Vacuum Particle Solution	ISO 7 (179-121) Vacuum #3	23991246ls
4_4	Vacuum Particle Solution	ISO 7 (179-121) Vacuum #4	23992198
4_5	Vacuum Particle Solution	ISO 7 (179-121) Vacuum #5	23991599
2_2	Wipe Solution	ISO 5 (233-151) Facility	23983173
2_4	Wipe Solution	ISO 5 (233-151) Facility	23987611
2_10	Wipe Solution	ISO 5 (233-151) Facility	23992936
3_1	Filter Solution	ISO 5 (233-151) Carbon Filter	22033416
3_3	Filter Solution	ISO 5 (233-151) Carbon Filter	23990894
3_5	Filter Solution	ISO 5 (233-151) Carbon Filter	23989452
3_4	Filter Solution	ISO 5 (233-151) Carbon Filter	23986851
3_6	Filter Solution	ISO 5 (233-151) Carbon Filter	23990849
3_9	Filter Solution	ISO 5 (233-151) Carbon Filter	23988907
4_6	Vacuum Particle Solution	ISO 5 (233-141) Vacuum #1	23990680
4_7	Vacuum Particle Solution	ISO 5 (233-141) Vacuum #2	23990861
4_8	Vacuum Particle Solution	ISO 5 (233-141) Vacuum #3	23992706

The proposal was quite good. If the number of samples is sufficient to create mags. When designing an experiment the amount of data needs to be sufficient to create a sufficient number of mags. Might have to go with more samples. 2.5 Gb per sample isn't that much. Focus on a couple of sample types rather than different conditions. Start some of the analysis as soon as possible. Will go through some important things over the next week. Go through one sample and do some assessment. MAGS per sample. See how it goes. E prepared to analyze more. Maybe 15 samples. Won't be a huge computational burden. For some of the analysis may not be able to run on poseidon metahit and metaspade (very memory hungry) you will run out of memory in the normal nodes. Since we have very few nodes that can handle that memory we can't use them for the class. Will be part of our results, these are the restrictions in memory, they cannot produce the results that require high memory. Goal not to reproduce exactly but to go through the pipeline and see what we can reproduce and what we cannot. What are our restrictions? Not necessary to go for all the figures. Should not be memory intensive to recover MAGS. Highly selective environment. First and second figure based on classified reads, do not involve MAGS. Are the species in fig 2a from database or from MAGS? Small metagenomes because diversity is low. Fast processes. It would be nice to have replicate samples to see how they behave. Fig 3 microbial profiling. Logical next step, introduce replicates in order to create diagram / timeline of which tools we need to recreate each figure and we can have a discussion about the amount of work required. Should we constrain to a few of them or go for all 4. Go through methods to see which are MAG related and which are read related. Which do you want to replicate and which do you want to do a few samples. If the metagenomes are small we may be able to work with many of them. Make a plan start from a few of the samples and reassess time and limitations. Reconstructing entire MAGS won't be possible. But get some numbers and compare to theirs.

Get our pipeline first before we expand the samples. Start with two different sample types and their replicates. Check consistency of results. Their pipeline may be reproducible but it will be good to see if their replicates are reproducible. How far can we expand. May be using reference genomes for mapping, these go very fast. Most time consuming part may be to download and decompress the data. For comparative reasons to see how close we are to their figures we may need 10-12 samlpes. Replicates are just from the same environment. Go for the sample types that generate the greatest observable difference. Work with two sample types to start and go from there. Start drafting analysis and we can meet to discuss again.

The project develops as we go through.