Jet Propulsion Laboratory
California Institute of Technology

Genetic Inventory of the Mars 2020 Spacecraft and Associated Surfaces

Lisa Guan

Biotechnology and Planetary Protection Group

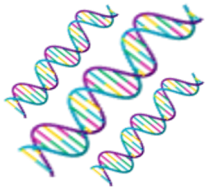
June 9, 2022

Mars 2020 Planetary Protection

- PLRA-L1-16: The project shall identify, quantify, document, and archive potential pre-launch terrestrial contamination sources, both organic compounds and organisms, and provide mechanisms **to support characterization of round-trip terrestrial contamination**.
- PS L2 – 44244: The PS shall identify, quantify, document, and archive potential pre-launch terrestrial contaminants, both organic compounds and organisms, as well as contamination sources.
- Not limited to just spores or any one type of organism in particular but rather we wanted to capture all forms of microbial biological contamination
- We implemented a DNA-based metagenomics approach to address the above requirements

Genetic Inventory Workflow

Sample Collection

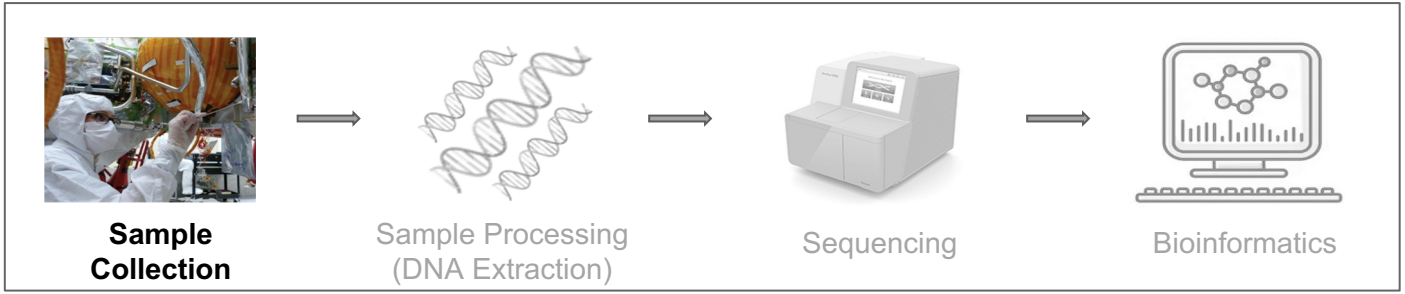


Sample Processing
(DNA Extraction)

Sequencing

Bioinformatics

Genetic Inventory Workflow



- The GI sampling approach targeted hardware locations with highest probability of access to cached samples and areas with a direct vector to the Martian environment.
- Hardware was broken down into 3 levels of priority, and only high priority samples (Priority Areas 1 and 2) were exhaustively sampled for GI.
- A subset of samples were sent for long-term storage at Johnson Space Center

Sampling Scope (2018-2020)

Priority 1

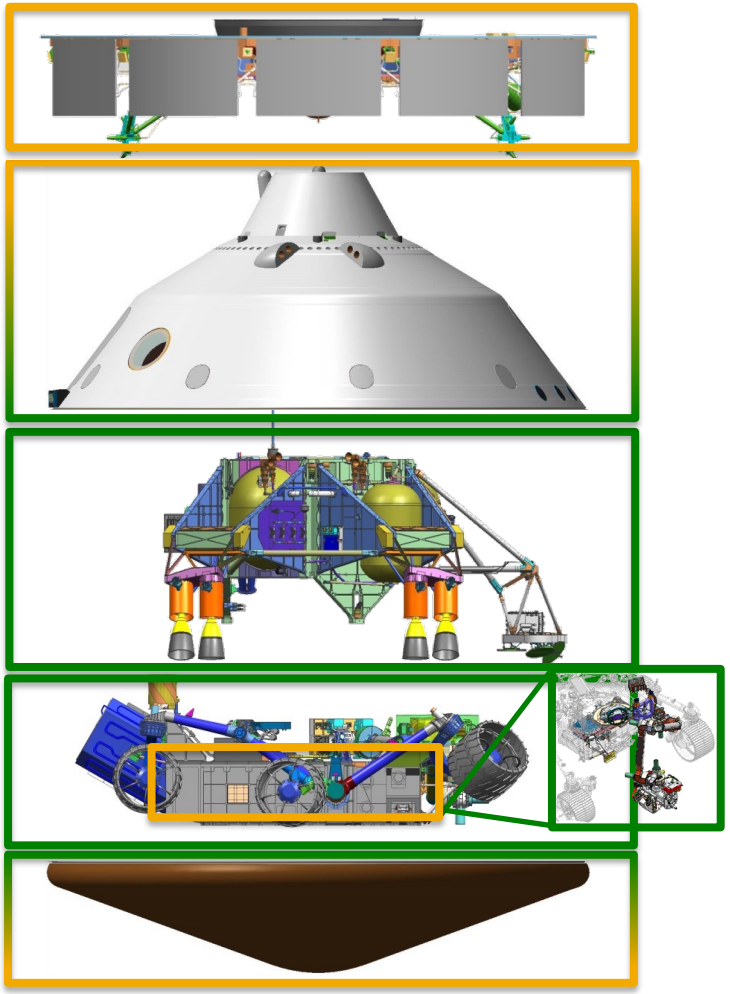
- All Sample Caching System Hardware (including Adaptive Caching Assembly and corer)
- Rover Exposed Exterior

Priority 2

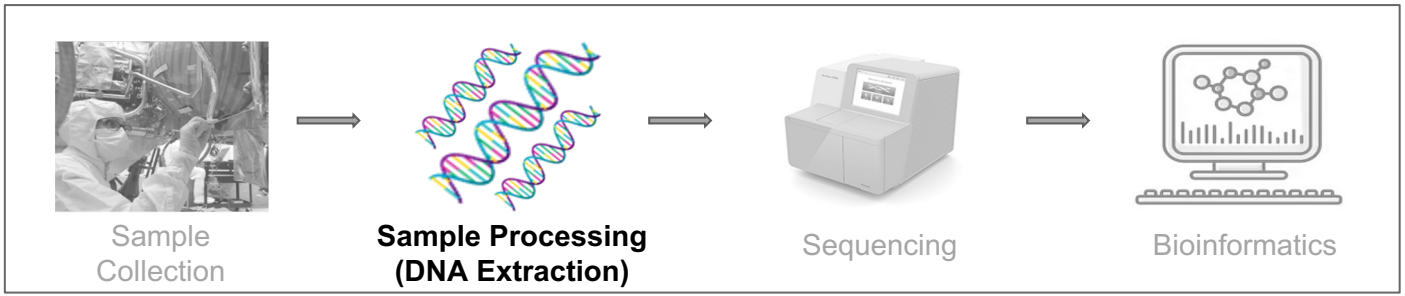
- Descent Stage Exposed Exterior
- Aeroshell/Heatshield Interior (with recontamination potential during launch)

Priority 3

- Cruise Stage Vehicle
- Exterior Aeroshell/Heatshield (including Parachute)
- Descent Stage Interior/Mated
- Rover Interior/Mated

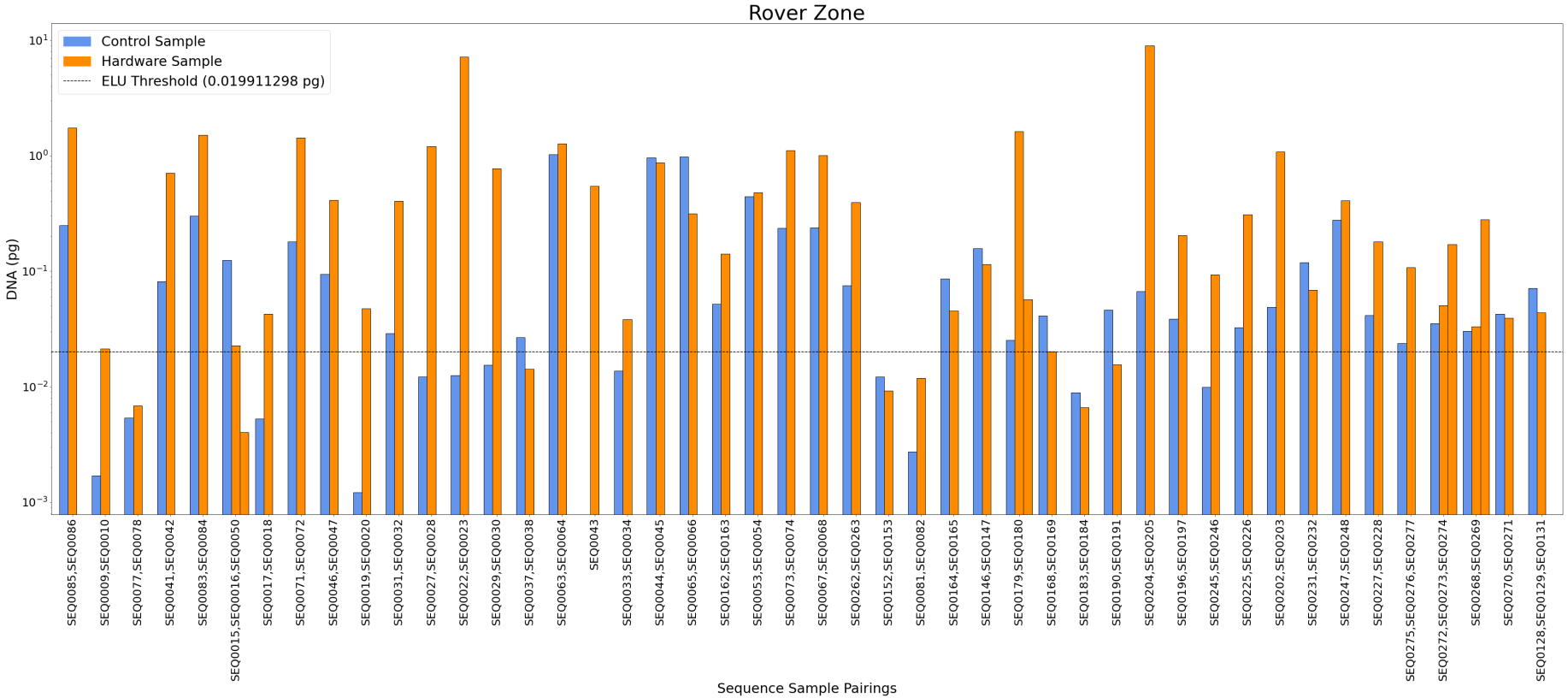


Genetic Inventory Workflow



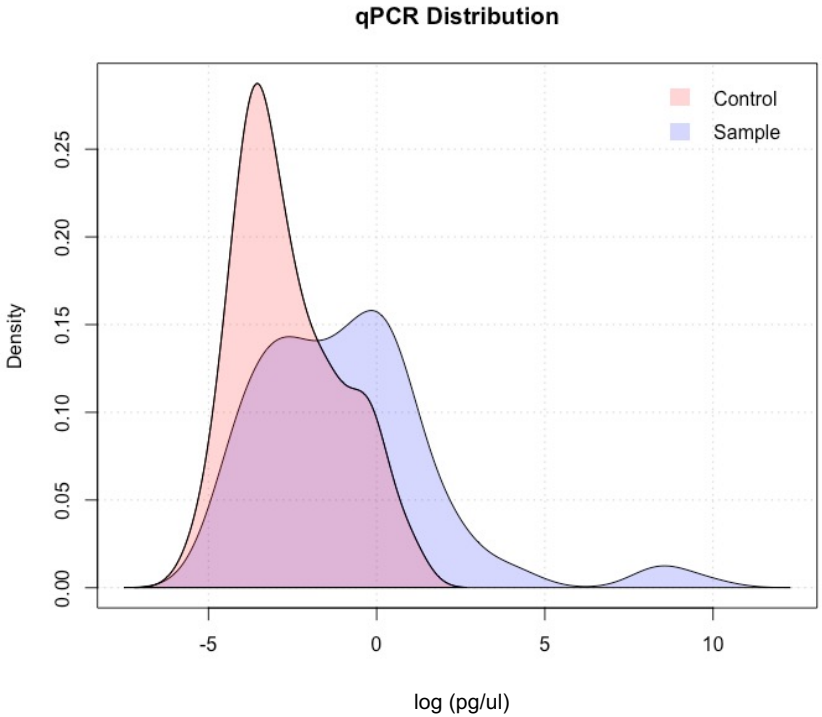
- Extremely low biomass, difficult to extract quantifiable amounts
- Dedicated facility for GI sample processing
- Extraction steps:
  - Double dissociation wash
  - Centrifugation basket to spin swabs dry
  - Chemical and physical lysis
  - **Pooling samples of same spacecraft grouping** (covering hundreds of cm^2 per DNA sample)
  - **Sample concentration with Amicon filters**
- Manual pre-process, automated extraction with Qiacube (column based)

qPCR Results

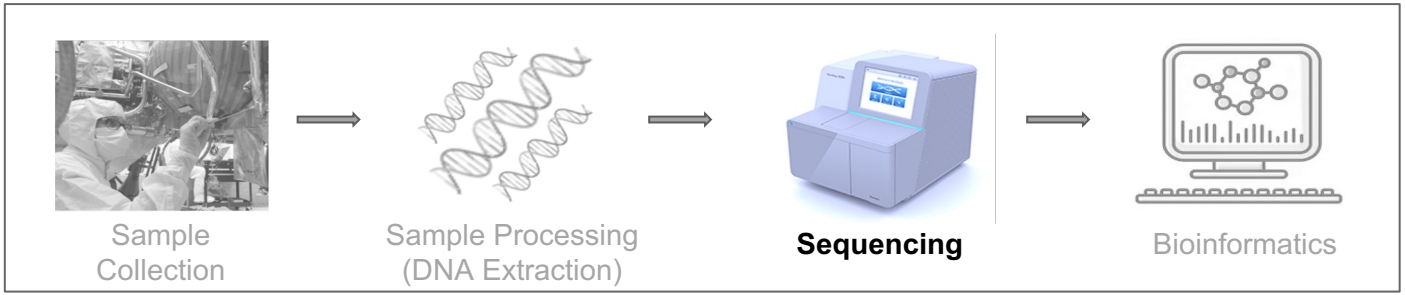


qPCR Results

- Due to large range of control concentrations, an absolute threshold for determining “cleanliness” for hardware may not be appropriate.
- Relative concentrations compared to handling controls may be more informative
- In a linear log-normal mixed model with observed data, it’s estimated that the controls have 0.19 pg/ul (95% CI: 0.13 – 0.29) *less* DNA than samples.
  - We controlled for “Facility/Spacecraft” location and control-sample sets

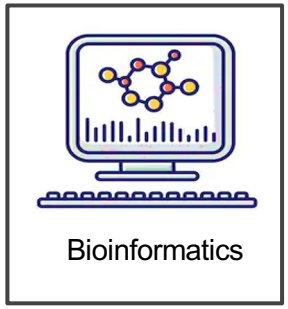


Genetic Inventory Workflow



- Samples were sent to Fulgent Genetics for library prep and sequencing
  - Samples must contain at least 1pg of DNA
  - Batches of at least 40
    - Minimum amount for us to have a whole sequencing run just for GI samples
    - Avoids potential carryover contamination from other higher-biomass samples
  - Library prep included linear amplification
  - Illumina HiSeq

Control Types Included in Study	What It Shows	Why It's Necessary
Negative Handling	Contamination from cleanroom environments	Handling controls encompass all possible sources of contamination.
Negative Extraction – Spanning different lots and various laboratory consumables (i.e Extraction kit, buffers, water, beads)	Contamination from various steps in the extraction process	Find steps in laboratory process that are significantly contributing to contamination.
Negative Sequencing	Contamination from library preparation	Shows if library preparation is contributing contamination. Can show sequencing batch effects
Positive Extraction	Bias in extraction	Extraction may miss certain taxa (tough to lyse, very low in abundance). A known, log-distributed mixed community can show if low abundance taxa are missing.
Positive Sequencing	Variation between runs Bias in PCR/sequencing Bias from input DNA concentration	Can show sequencing batch effects Shows differences in expected vs actual distribution and if input DNA concentration affects distribution



Sequencing Vendor

FTP transfer

Raw FASTQ Files

Data Processing

Database: Hg38

Remove Human Reads
(Bowtie2, Samtools, Bedtools)

NCBI SRA

Submission

Human Read Removed
FASTQ Files

Quality Trimming & Adapter Removal
(Trimmomatic)

Quality Trimmed
Human Read Removed FASTQ Files

Confidence score informed by simulated metagenomes (InSilicoSeq) – understand bioinformatic processing noise

Taxonomic Classification
(**Kraken2 + Bracken**)

Database: Custom*

Taxonomy Results
('Counts' tables, Species Level)

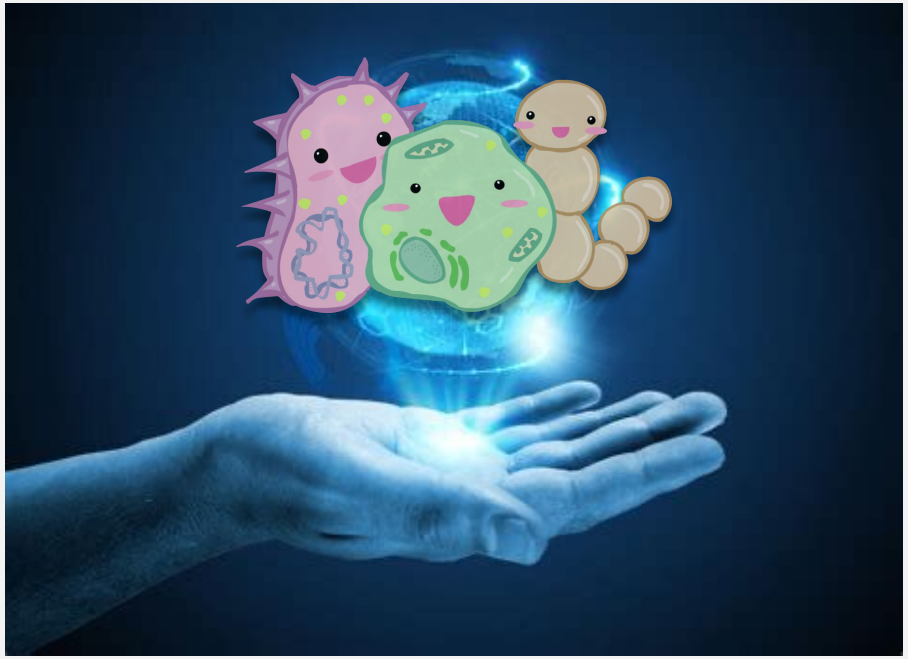
*Bacteria, Virus, Fungi, Archaea, Arthropoda, Plants

Data Analysis

Statistics, Data Visualization

SIMULATED METAGENOME

- Contains 25 bacterial species
- Created using InSilicoSeq (sequencing simulator)
- 5 M reads
- HiSeq error-model
- Uniform abundance



www.pngaaa.com

Slide provided by Shirin Nataneli

KRAKEN 2'S CONFIDENCE SCORING SCHEME

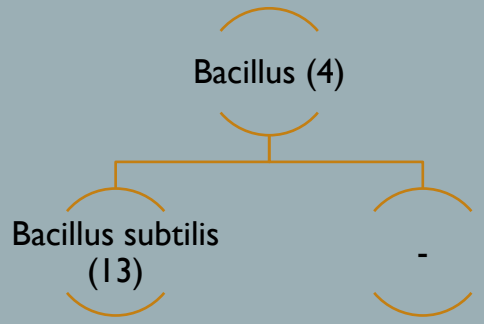
$$confidence\ score = \frac{C}{Q}$$

$C = Number\ of\ kmers\ mapped\ to\ lowest\ common\ ancestor\ in\ a\ taxonomic\ tree$
(e.g. family, genus, species etc.)
$Q = Number\ of\ kmers\ in\ the\ sequence\ that\ don't\ have\ an\ ambiguous\ nucleotide$

* kmers = subsequence of a certain length (k) within a sequencing read

Slide provided by Shirin Nataneli

CONFIDENCE SCORE EXAMPLE



A:31 0:1

the first 13 k-mers mapped to *B. subtilis*

the next 4 k-mers mapped to *Bacillus*

the next 31 k-mers contained an ambiguous nucleotide

One k-mer was not in the database

Slide provided by Shirin Nataneli

$$confidence\ score = \frac{C}{Q}$$

$C = Number\ of\ kmers\ mapped\ to\ lowest\ common\ ancestor\ in\ a\ taxonomic\ tree$
$Q = Number\ of\ kmers\ in\ the\ sequence\ that\ don't\ have\ an\ ambiguous\ nucleotide$

B. subtilis

$$\frac{13}{13+4+1} = \frac{13}{18}$$

Bacillus

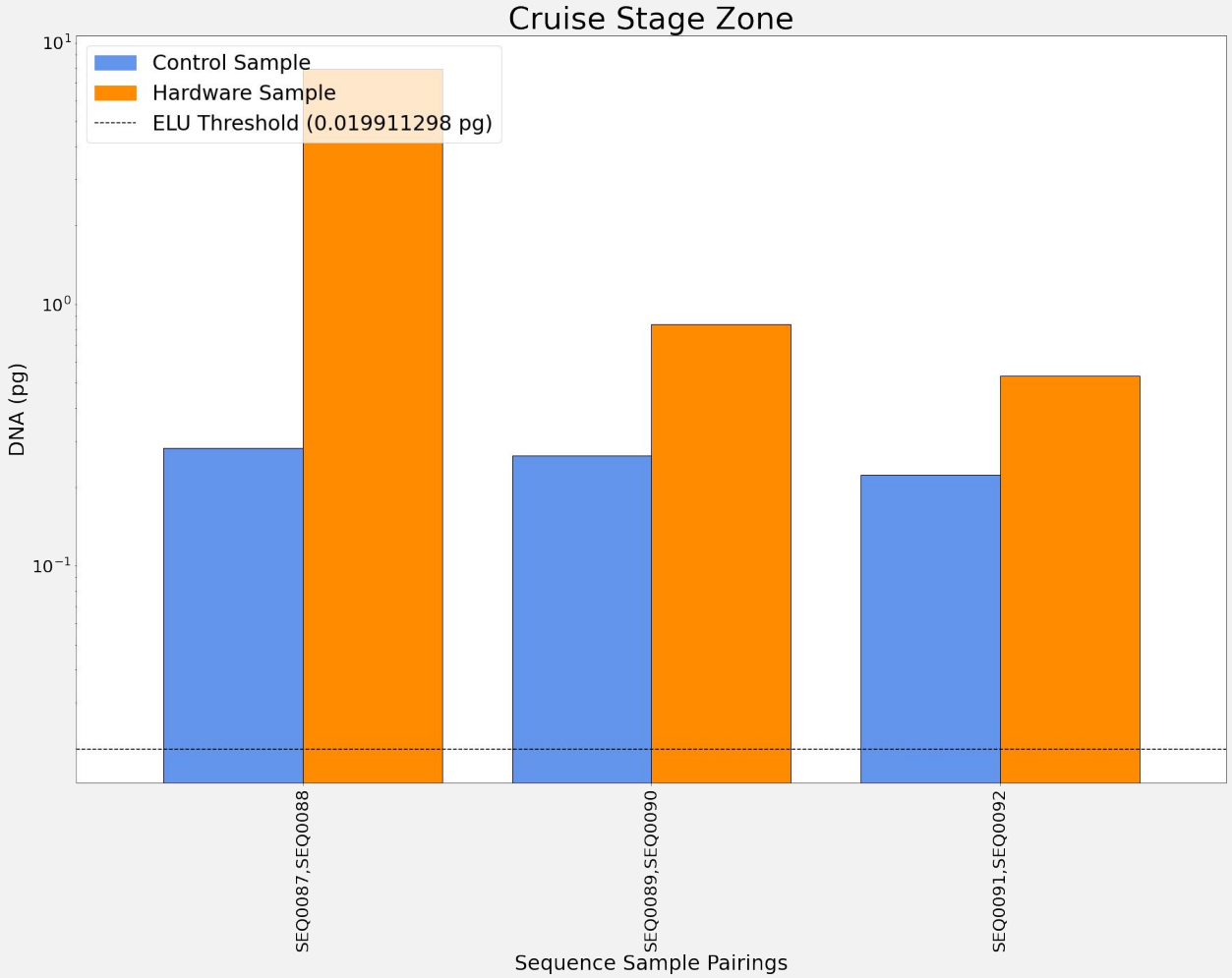
$$\frac{13+4}{13+4+1} = \frac{17}{18}$$

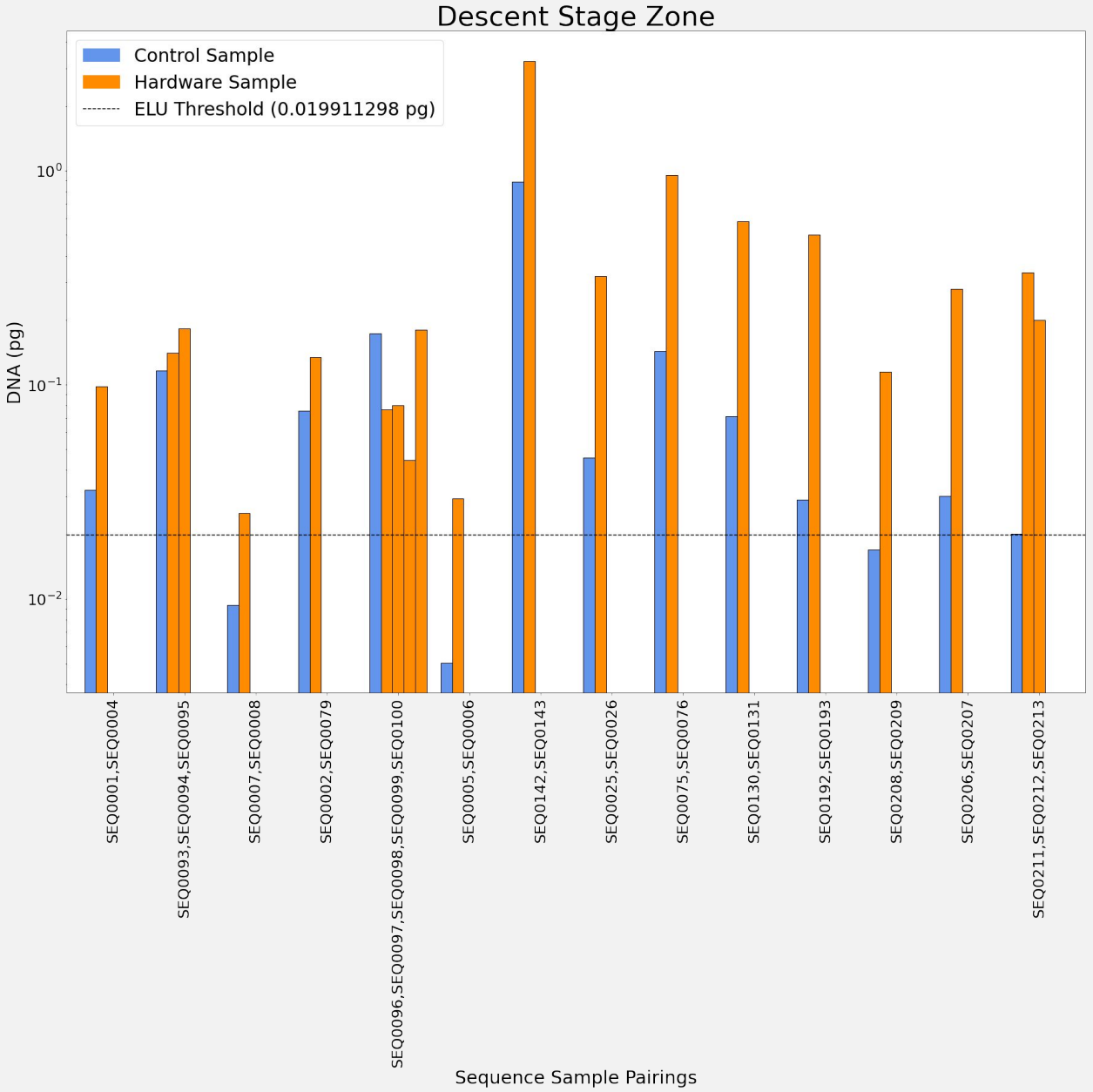
If user specifies threshold over $\frac{13}{18}$, *B. subtilis* is adjusted to *Bacillus*.

If user specifies a threshold above $\frac{17}{18}$ the sequence would become unclassified

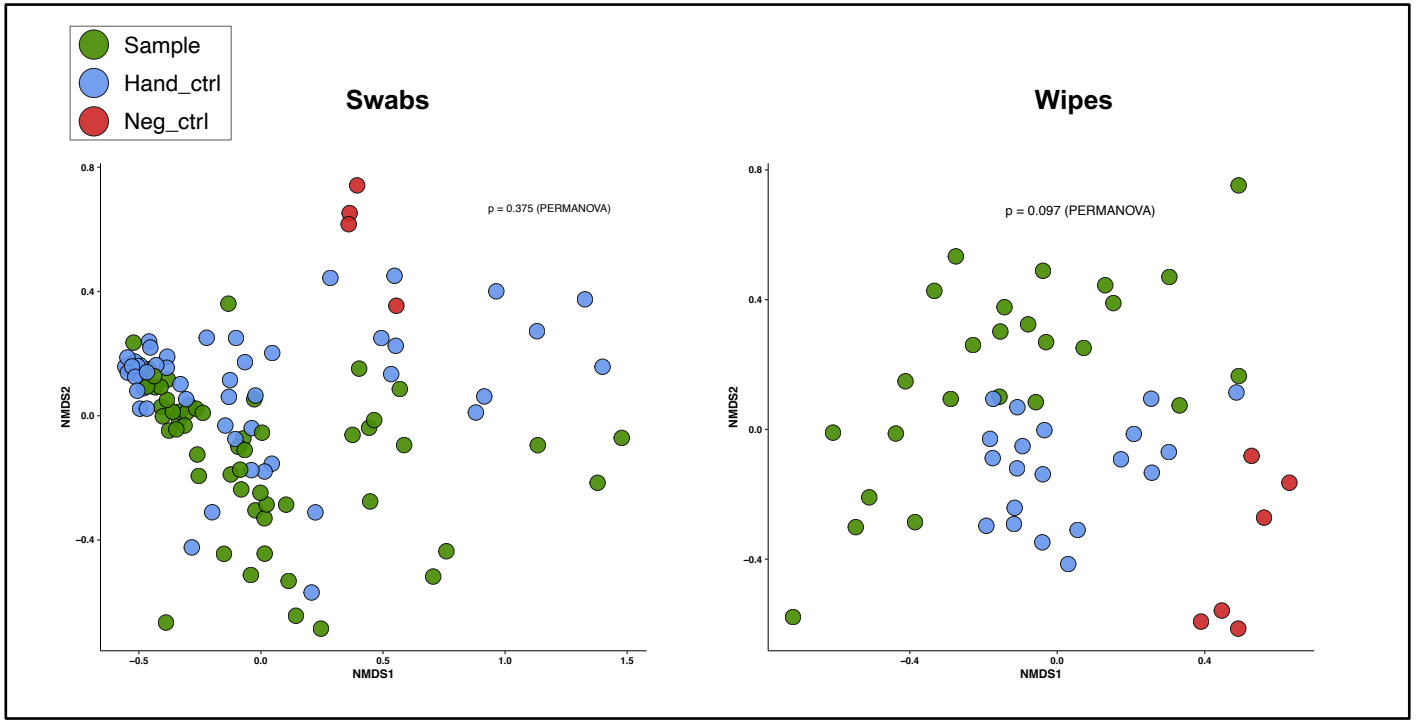
CONFIDENCE SCORE RESULTS

Slide provided by Shirin Nataneli





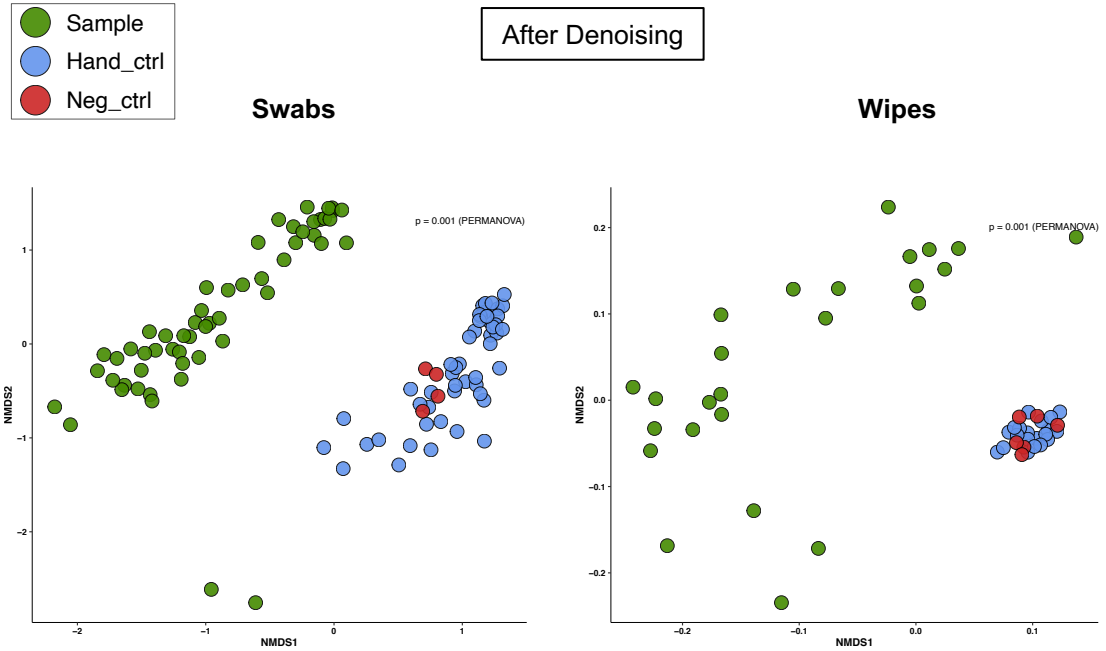
Slide provided by Shirin Nataneli



NMDS ordinations of hardware samples (green), handling controls (blue), and sequencing controls (red) based on species-level taxonomic profiles, from sequencing batches 1-7. Microbial profiles of hardware samples and handling controls were often indistinguishable, especially for the swab method, indicating presence of shared species.

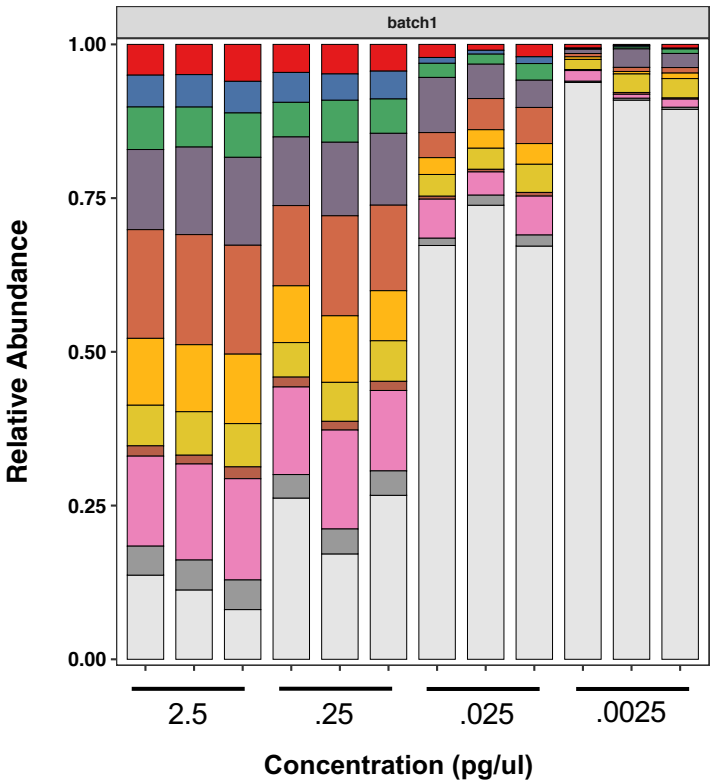
Decontam Results

- NMDS ordinations showed that profiles of hardware samples & controls were different from each other after denoising, for both the wipe and swab samples
- Handling controls (swabs/wipes) were virtually indistinguishable from NTCs, suggesting that these do not significantly contribute to the overall biodiversity of the samples

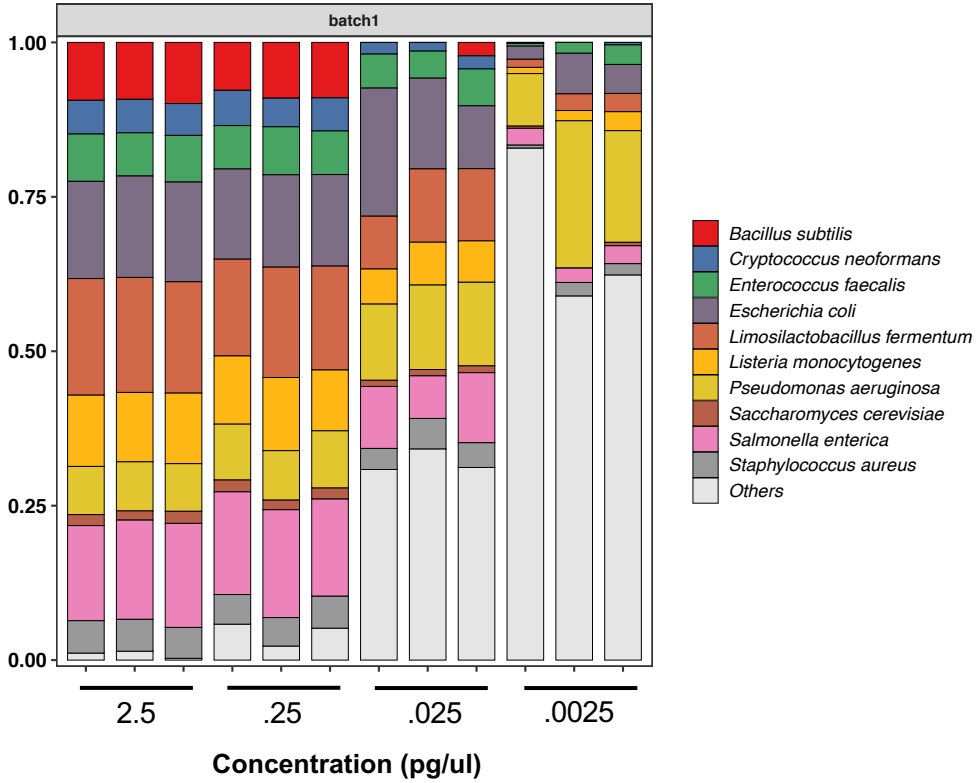


Low Biomass Challenges

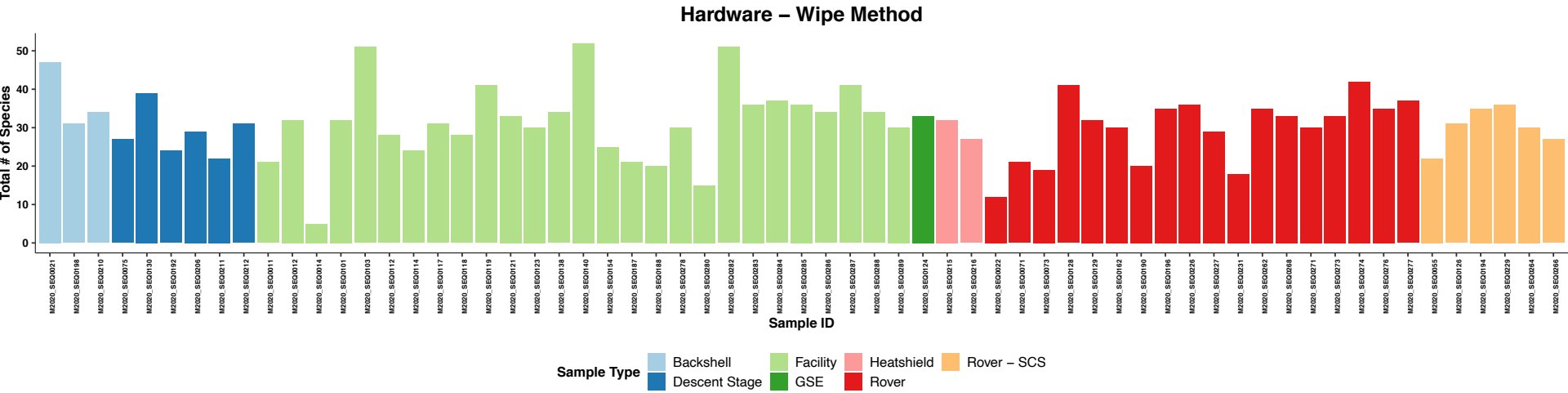
Positive Control Samples -
Raw Species Abundance



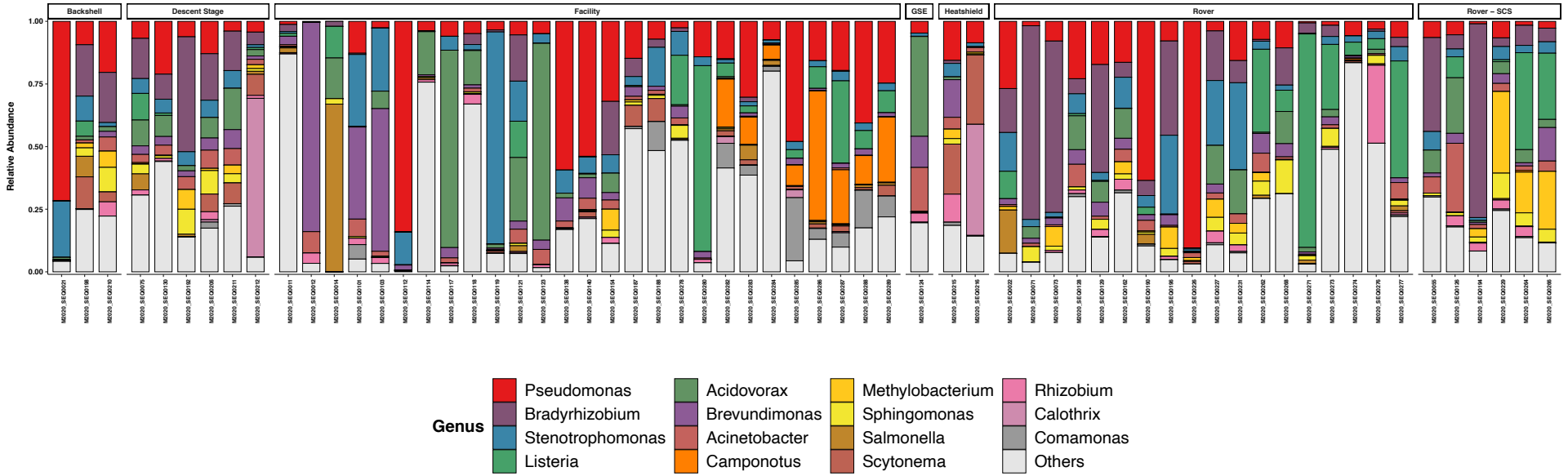
Positive Control Samples –
Post Confidence Score and Decontam



Hardware Sample Species Counts



Hardware Sample Taxonomic Results



Lessons Learned from GI

- Improved sampling methods can increase yield (swabs, peels, tapes, etc)
- Sample to results turnaround time currently too long due to batch size requirements
- Can complement metagenomics-based method with rapid, quantitative method such as qPCR/ddPCR
  - Use ddPCR for targeted quantification without amplification bias; ddPCR product can be used for sequencing
  - Use metagenomics for comprehensive identification and downstream functional analysis
- Simulated metagenomes can help determine data processing parameters for unique datasets
- Extensive controls are needed for quality filtering and interpreting process contamination/false positives

Acknowledgements

- GI Team
  - Nitin Singh
  - Parag Vaishampayan
  - Chiranjit Mukherjee
  - Shirin Nataneli
  - Kathryn van Artsdalen
- Mars 2020 PP Team
- Mars 2020
- JPL Mars Program Office

REFERENCES

- Gourlé, H., Karlsson-Lindsjö, O., Hayer, J., & Bongcam-Rudloff, E. (2019). Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics*, *35*(3), 521-522.
- Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome biology*, *20*(1), 1-13.

Slide provided by Shirin Nataneli

Jet Propulsion Laboratory
California Institute of Technology

jpl.nasa.gov