

Group Project: Alien Experiments

Description of challenge:

In summer 2024, a number of cancer patients mysteriously disappeared from the University Hospital Basel. They inexplicably turned up a single day later, with few memories of what happened. The only common thread in their story: Everyone believes they had been abducted by aliens. These friendly aliens supposedly administered a special, gut microbiota-modifying treatment to these patients in efforts to cure their disease. Lo and behold, some of these patients do appear to miraculously recover.

As a member of the clinical diagnostics team at the University of Zürich Institute of Medical Microbiology, you are tasked with investigating the potential ramifications of these “treatments”. To this end, you are given a dataset containing fecal microbiome sequences (“Sequences”) and a metadata table with additional information on the samples (“Sample metadata”) from these patients. To identify the effects of alien abduction, you should examine the relationship between the fecal microbiome, patient covariates, and recovery times.

The patients and their families await your (positive and negative) results with bated breath. The scientific community is already very interested in learning about your approach.

Some specific questions*:

1. Do you find any associations within your metadata (e.g. between stool consistency and sample day)?
 - Hint: This is a good moment to practice some pandas summary statistics on your metadata!
2. Explore the microbial communities of the alleged alien abductees. What characteristics do you observe? Are you able to identify predictive biomarkers for the speed of recovery?
 - Bonus: Some patients gave more than one stool sample. How do microbial communities change between these timepoints?

Information on provided datasets:

1. **Sequences:** The sequences originate from the V3-V4 region of the 16S rRNA gene and were sequenced on an Illumina MiSeq machine. The sequence files contain demultiplexed paired-end sequences and are provided as a QIIME 2 artifact (semantic type: `SampleData[PairedEndSequencesWithQuality]`).
Forward primer: ACTCCTRCGGGAGGCAGCAG
Reverse primer: GGACTACHVGGGTATCTAATCCTGTT
2. **Sample metadata:** The metadata table contains additional information on each collected sample. Each row represents a unique sample that can be identified with its unique Sample_ID. The remaining columns depict health data of the patient. All date-related values are given relative to the day of abduction e.g. Sample_Day of 13 means 13 days post-abduction.

Where to get this data:

(Use wget to download these links in your analysis notebooks for fully executable analyses!)

1. Sequences: <https://polybox.ethz.ch/index.php/s/lCmJYEDg5Xz1FEA>
2. Sample metadata: <https://polybox.ethz.ch/index.php/s/JrwER5E9kpgGuWd>

Questions? Your course instructors are here to help. Meet with instructors during in-class group work sessions or ask questions via email.

Contact TA: Lina Kim, lina.kim@hest.ethz.ch

** **Note on questions:** This is not a list of the only questions to answer about this dataset. All group semester projects are expected to incorporate most analysis topics covered in the course when possible (see Group Semester Projects assignment document). This is merely a list of specific questions that should be answered/ attempted in addition to the “standard” analyses covered. Not everything will work or give a clear result! But that does not mean that you should not try. Your goal as scientists is to figure out why.*