Data Sets

We have gathered several data sets for you to use in the course, and in this manual we use different data sets as examples.

The main thing you need to know is whether a data set is a 16S amplicon library or a random community metagenome.

coral_algae

This is coral, algae, CCA, and water (control) samples from Kevin Walsh in Liz Dinsdale's lab.

This is a random community metagenome data set

There are 50,000 reads in these data sets.

Abstract: Coral reefs are undergoing global microbialization as carbon and energy from higher trophic levels shift into the microbial food web. Increase in labile carbon resources has altered microbial community composition from phototrophs to copiotrophs and super-heterotrophs. Even in oligotrophic systems, super-heterotrophs persist in the rare biosphere. These rare organisms are likely to become abundant with microbialization, but their scarcity inhibits deep sequencing and metabolic analysis with metagenomics. Therefore, we utilized enrichment after pre-exposure of the water column microbiome to benthic organisms, 1) water control, 2) alga, Stypopodium zonale, 3) crustose coralline algae, and 4) the coral, Mussismilia hartti to identify super-heterotrophs in the coral reef environment. We compared enriched communities to metagenomes from coral reef water and the coral Mussismilia braziliensis to compare the presence of dominant genera as population genomes in these native microbial communities. Enrichment selected for super-heterotrophs in the genus, Vibrio, Pseudoalteromonas and Arcobacter with greater sequence coverage of Arcobacter in coral exposures. We assembled two Vibrio, a Pseudoatleromonas and an Arcobacter, which identified previously unannotated sequences. To determine genes that define the ecotypes of the environmental microbes, we compared the population genomes to genomes of related organisms sequenced from cultured isolates. We found the coral reef associated Vibrio population had a higher proportion of genes in the metabolic pathways: Type IV secretion and conjugative transfer, maltose utilization and urea decomposition. Pseudoalteromonas population had more genes involved in sugar utilization pathways while Arcobacter population was distinguished by genes contributed to type VI secretion systems and utilization of alkylphosphates and aromatic compounds. By assembling population genomes, we identified novel genes defining the ecotype relative to the pangenome of culture isolates. Novel gene identification informs the ecology of super-heterotrophs on corals reef independent of the limitations of reference genomes.

drinking_water

A drinking water study from the University of Adelaide, Australia

This is 16S amplicon dataset with SRP ID SRP059994

Publication: Shaw JLA, Monis P, Weyrich LS, Sawade E, Drikas M, Cooper AJ. 2015. Using Amplicon Sequencing To Characterize and Monitor Bacterial Diversity in Drinking Water Distribution Systems. Appl Environ Microbiol 81:6463–6473

Abstract: Drinking water assessments use a variety of microbial, physical, and chemical indicators to evaluate water treatment efficiency and product water quality. However, these indicators do not allow the complex biological communities, which can adversely impact the performance of drinking water distribution systems (DWDSs), to be characterized. Entire bacterial communities can be studied quickly and inexpensively using targeted metagenomic amplicon sequencing. Here, amplicon sequencing of the 16S rRNA gene region was performed alongside traditional water quality measures to assess the health, quality, and efficiency of two distinct, full-scale DWDSs: (i) a linear DWDS supplied with unfiltered water subjected to basic disinfection before distribution and (ii) a complex, branching DWDS treated by a four-stage water treatment plant (WTP) prior to disinfection and distribution. In both DWDSs bacterial communities differed significantly after disinfection, demonstrating the effectiveness of both treatment regimes. However, bacterial repopulation occurred further along in the DWDSs, and some end-user samples were more similar to the source water than to the postdisinfection water. Three sample locations appeared to be nitrified, displaying elevated nitrate levels and decreased ammonia levels, and nitrifying bacterial species, such as Nitrospira, were detected. Burkholderiales were abundant in samples containing large amounts of monochloramine, indicating resistance to disinfection. Genera known to contain pathogenic and fecal-associated species were also identified in several locations. From this study, we conclude that metagenomic amplicon sequencing is an informative method to support current compliance-based methods and can be used to reveal bacterial community interactions with the chemical and physical properties of DWDSs

This dataset has 5 sequence runs:

Read ID	Name
SRR2080423	SA 3 tank
SRR2080425	1.5kmPostDis
SRR2080427	SASourceWater2
SRR2080434	SA 2 CT
SRR2080436	WA1 outlet

You can download these datasets with this command:

```
for SRR_ID in SRR2080423 SRR2080425 SRR2080427 SRR2080434 SRR2080436; do
        fastq-dump --outdir fastq --gzip --skip-technical  --readids --read-filter pass --du
done
```

For some reason this data seems mixed up, as there are sequences with each tag in each file.

`split.py` will split based on the tag sequence, and you can edit it and provide a maximum number of sequences per tag

The original number of counts per tag are:

- 303,446 TCGCAGG
- 276,080 GCTCGAA
- 312,009 CTCGATG
- 293,741 ACCAACT
- 269,702 GGATCAA

ground_water

This data comes from SRA project SRP075429

Hernsdorf AW, Amano Y, Miyakawa K, Ise K, Suzuki Y, Anantharaman K, Probst A, Burstein D, Thomas BC, Banfield JF. 2017. Potential for microbial H2 and metal transformations associated with novel bacteria and archaea in deep terrestrial subsurface sediments. ISME J 11:1915–1929

Title: Potential for microbial H2 and metal transformations associated with novel bacteria and archaea in deep terrestrial subsurface sediments

Abstract: Geological sequestration in deep underground repositories is the prevailing proposed route for radioactive waste disposal. After the disposal of radioactive waste in the subsurface, H2 may be produced by corrosion of steel and, ultimately, radionuclides will be exposed to the surrounding environment. To evaluate the potential for microbial activities to impact disposal systems, we explored the microbial community structure and metabolic functions of a sediment-hosted ecosystem at the Horonobe Underground Research Laboratory, Hokkaido, Japan. Overall, we found that the ecosystem hosted organisms from diverse lineages, including many from the phyla that lack isolated representatives. The majority of organisms can metabolize H2, often via oxidative [NiFe] hydrogenases or electron-bifurcating [FeFe] hydrogenases that enable ferredoxin-based pathways, including the ion motive Rnf complex. Many organisms implicated in H2 metabolism are also predicted to catalyze carbon, nitrogen, iron and sulfur transformations. Notably, iron-based metabolism is predicted in a novel lineage of Actinobacteria and in a putative methane-oxidizing ANME-2d archaeon. We infer an ecological model that links microorganisms to sediment-derived resources and predict potential impacts of microbial activity on H2 consumption and retardation of radionuclide migration

This is a random community data set

Design: DNA was extracted from the biomass retained on 0.2 um filters using the Extrap Soil DNA Kit Plus ver. 2 (Nippon Steel & Sumikin Eco-Tech Corporation) and sent for 150 bp paired-end sequencing with a 550 bp insert size by Hokkaido System Science Co., Ltd. using an Illumina HiSeq2000

Runs:

- SRR3546457
- SRR3546455
- SRR3546454
- SRR3546453
- SRR3546452
- SRR3546451
- SRR3546450
- SRR3546449

The runs can be downloaded from SRA like this:

```
for SRR_ID in SRR3546457 SRR3546455 SRR3546454 SRR3546453 SRR3546452 SRR3546451 SRR3546450 S
        fastq-dump --outdir fastq --gzip --skip-technical  --readids --read-filter pass --du
done
```

gut

This data comes from SRA project SRP074153

This is a random community data set

Brooks B, Olm MR, Firek BA, Baker R, Thomas BC, Morowitz MJ, Banfield JF. 2017. Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. Nat Commun 8:1814

Title: Metagenomes from 11 human infant fecal samples hospitalized in the same intensive care unit

Abstract: Bacteria that persist in hospitals can contribute to the establishment of the microbiome in newborns and the spread of hospital-acquired diseases. Yet we know little about microbial communities in hospitals, or about the extent to which persistent vs. recently immigrated bacterial strains establish in the gastrointestinal tracts of hospitalized individuals.In combination with BioProject PRJNA273761 (10 infants / 55 samples) we analyzed strain-resolved genomes obtained from a total of 202 samples collected over a three-year period from 21 infants hospitalized in the same intensive care unit.Strains were rarely shared, consistent with prior analysis of a subset of these data. Enterococcus faecalis and Staphylococcus epidermidis, common gut colonists, exhibit diversity comparable to that of NCBI reference strains, suggesting no recent common ancestor

for all populations in this hospital setting. Thus, we infer multiple introduction events for these species. Despite the rarity of shared strains, strains of five species exhibiting a degree of sequence variation consistent with in situ diversification were identified in different infants hospitalized three years apart. Three were also detected in multiple infants in the same year, suggesting that these strains are unusually widely dispersed and persistent in the hospital environment. Persistent strains were not significantly different from non-persistent strains with regards to pathogenicity potential including antibiotic resistance. Notably, non-identical siblings had multiple abundant strains in common, even 30 days after birth and antibiotic administration, suggesting overlapping strain sources and/or genetic selection. Our approach can be used in order to study microbial dynamics in hospitals and provides an important step towards directing health-promoting colonization in hospitalized individuals.

Design: DNA was extracted using the MO BIO PowerSoil DNA Isolation kit; libraries were made Illuminas Nextera kit with average insert sizes of 500/900 bp

Runs:

- SRR3466404
- SRR3506419
- SRR3506420
- SRR3546776
- SRR3546778
- SRR3546779
- SRR3546780
- SRR3546781
- SRR3546782

The runs can be downloaded from SRA like this:

```
for SRR_ID in SRR3466404 SRR3506419 SRR3506420 SRR3546776 SRR3546778 SRR3546779 SRR3546780 S
        fastq-dump --outdir fastq --gzip --skip-technical  --readids --read-filter pass --du
done
```