Understanding the SRA

Amelia Harrison

Presented: April 20, 2022

The Sequence Read Archive

- Publicly available repository of high-throughput sequencing data
  - Specifically, raw sequencing data (“straight out of the sequencer”)
  - Data is available from NCBI servers or through some cloud services
- Vocabulary
  - Run – same combination of sample + library + strategy + layout + instrument model
    - For paired-end data, forward and reverse reads belong in the same run
  - Library – A pool of DNA fragments ready for sequencing
  - Read – The DNA sequence from one strand of DNA
  - Spot – Essentially, the information from a location on the flow cell
- The equivalent in Europe is the ENA (European Nucleotide Archive)

Other NCBI Submission Types

- BioProject
  - “A BioProject is a collection of biological data related to a single initiative, originating from a single organization or from a consortium. A BioProject record provides users a single place to find links to the diverse data types generated for that project.”

Ocean Microbiome

Accession: PRJEB7988 ID: 277448

Tara Oceans Ocean Microbiome project

The Tara Oceans expedition (Karsenti et al. 2011) has collected seawater samples from all over the globe and the metagenomic analysis of a subset of these samples has been published (Sunagawa et al. **More...**

Accession	PRJEB7988
Data Type	Genome sequencing and assembly
Scope	Monoisolate
Publications	1. Guidi L *et al.*, "Plankton networks driving carbon export in the oligotrophic ocean.", *Nature*, 2016 Apr 28;532(7600):465-470 2. Sunagawa S *et al.*, "Ocean plankton. Structure and function of the global ocean microbiome.", *Science*, 2015 May 22;348(6237):1261359
Submission	Registration date: 6-Mar-2015 **EMBL HEIDELBERG**
Locus Tag Prefix	BN1885

NAVIGATE UP

This project is a component of the Tara-oceans samples barcoding and shotgun sequencing

NAVIGATE ACROSS

28 additional projects are components of the Tara-oceans samples barcoding and shotgun sequencing.

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (WGS master)	232
PUBLICATIONS	
PubMed	2
PMC	1
OTHER DATASETS	
BioSample	233
Assembly	231

Other NCBI Submission Types

- BioProject
  - “A BioProject is a collection of biological data related to a single initiative, originating from a single organization or from a consortium. A BioProject record provides users a single place to find links to the diverse data types generated for that project.”
- BioSample
  - “A BioSample contains descriptive information about the physical biological specimen from which your experimental data are derived. Typical examples of a BioSample include a cell line, a tissue biopsy or an environmental isolate.”

TARA_Y100001970; TARA_20110728T1718Z_122_Combined-EVENTS_CAST_MB_D_(115 m)_GIRUS_NUC-dry_W0.22-0.45_TARA_Y100001970

Identifiers: BioSample: SAMN07286193; Sample name: Alphaproteobacteria bacterium MarineAlpha2_Bin1

Organism: Alphaproteobacteria bacterium MarineAlpha2_Bin1
cellular organisms; Bacteria; Proteobacteria; Alphaproteobacteria; unclassified Alphaproteobacteria

Package: Microbe; version 1.0

Attributes

isolation source	saline water including plankton
collection date	2011-07-28
geographic location	Pacific Ocean: South Pacific Ocean
latitude and longitude	9.0063 S 139.1394 W
isolate	MarineAlpha2_Bin1
depth	115 m
sample type	metagenomic assembly
temperature	24.69625
broad-scale environmental context	marine biome (ENVO:00000447)
collected by	Tara Oceans
note	This BioSample is a metagenomic assembly obtained from the marine metagenome BioSample: SAMEA2622695.
metagenome_source	marine metagenome
environmental_sample	TRUE
metagenomic	TRUE

Description: This sample (TARA_Y100001970) was collected during the Tara Oceans expedition (2009-2013) at station TARA_122 (latitudeN=-9.0063, longitudeE=-139.1394) on date/time=2011-07-28T17:15:56, using a ROSETTE sampler with CTD (sbe9C) and 10 Niskin bottles. The sample material (saline water (ENVO:00002010), including plankton (ENVO:xxxxxxxx)) was collected at a depth of 115-115 m, targeting a deep chlorophyll maximum layer (ENVO:xxxxxxxx) in the marine biome (ENVO:00000447). The sample was size-fractionated (0.22-0.45 micrometres), and stored at -20 degC for later detection of large DNA virus nucleic acid sequences by pyrosequencing methods, and for later metagenomics analysis.

BioProjects: PRJNA390581 marine metagenome
Retrieve all samples from this project

PRJEB7988 Ocean Microbiome
Retrieve all samples from this project

SRA tools

- SRA Run Selector
  - Used to browse **runs** in from Studies, Samples, or Experiments
- SRA Run Browser
  - Used to browse **metadata** related to a particular **run**
- SRA Tool Kit
  - Command line tool used to interact with the SRA

Activity

Agricultural pond microbes and viruses

- Sampled an agricultural pond once a month for 3 months
- Water was filtered through 2 filter sizes (1 µm and 0.2 µm)
- Filters were cut into 4 pieces and prepped for 16S sequencing
- Water that made it through the filter was prepped for shotgun metagenomic sequencing

Questions?

- How many samples are there?
- How many BioProjects? BioSamples? Runs?
- What information can you find in each?

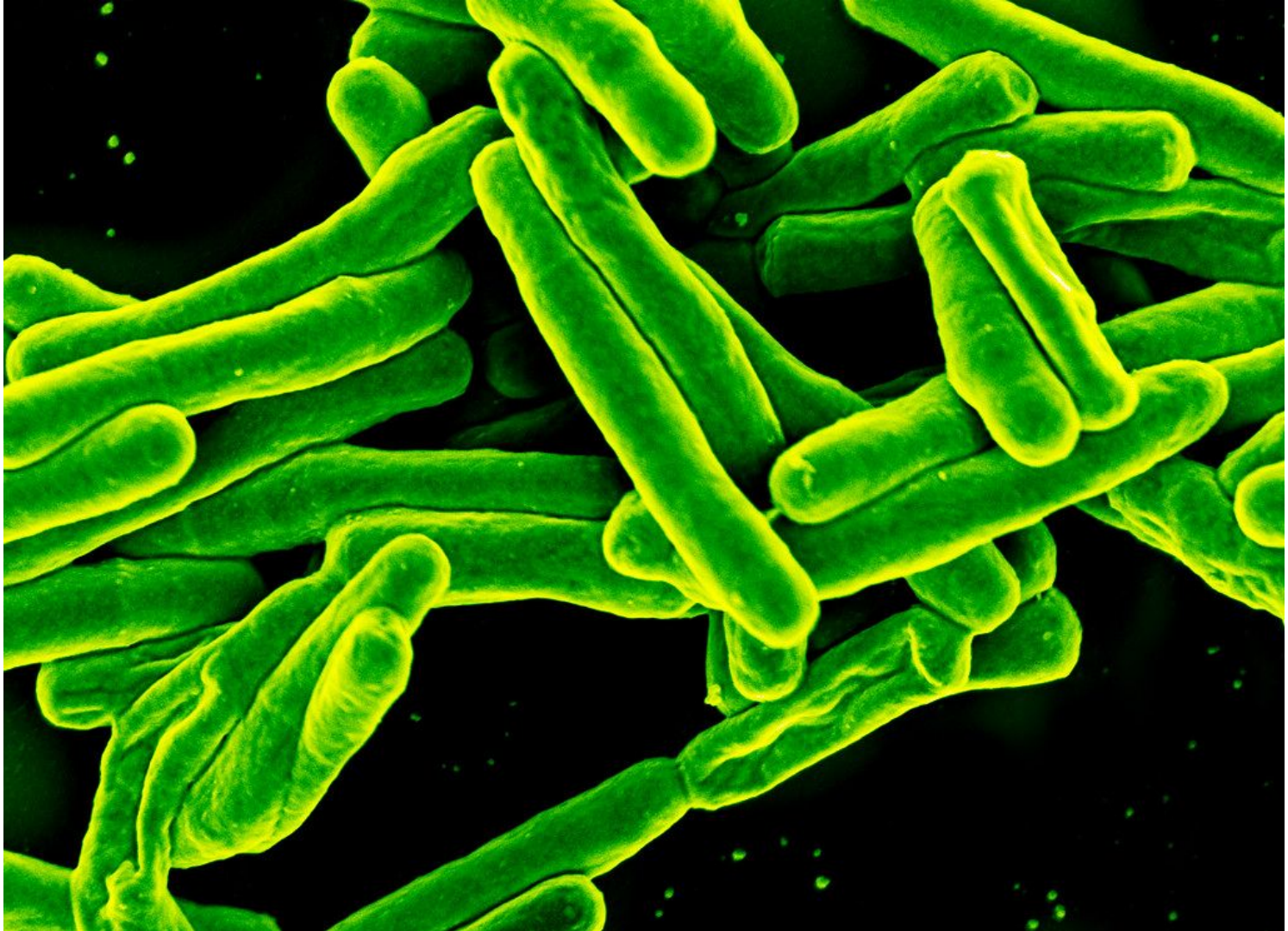
Agricultural pond microbes and viruses

Go to this paper: https://doi.org/10.3389/fmicb.2018.00792

Find:

- BioProject
- BioSamples
- Runs

Can use the run explorer and/or NCBI search, but start with the accession(s) reported in the paper.



M. tuberculosis studies

Explore these studies in the SRA Run Selector:

- Blouin et al. 2012: ERP001885
- Lee et al. 2015: SRP039605

What do you notice about the metadata for these studies?