Now that we have covered our basic workflow, it’s time to start thinking about projects. There are two general approaches: looking for differences in an organism by environment and looking for differences in environments by the types of organisms present.

1. Examining a single organism across many metagenomes
2. Examining a few datasets for all organisms.

Two example projects would be:

1. Looking at the variation in E. coli, across 100 different human gut metagenomes
2. Comparing two sets of human fecal metagenomes (before and after fecal transplants) across all organisms to see how the entire community changes.

Not all experiments fall completely into one of these two categories. There are many approaches to metagenomics but these two types are a good place to start. For this homework, you will plan two sample projects of each of type. The purpose of this mini project is to give you experience analyzing data – do not worry about the impactfullness of your research. Choose something simple that you will get results for.

**Comparing an organism across multiple environments**

For this experiment, find an interesting organism and a few data sets that will almost certainly contain the organism and answer the questions below.

1. What organism are you investigating? Why is this organism relevant?
2. What environments are you investigating? What are the Run accessions for these datasets? (less than 100)
3. What differences do you expect to find in organism between samples?

**Comparing environments by organism(s)**

For this experiment, find a few datasets that are similar – ideally the runs should be part of the same experiment or project. An example of a good dataset would be 5 runs from healthy individuals and 5 runs from sick individuals – or perhaps the runs are before and after a treatment. Then answer the following questions:

1. What SRA Runs are you investigating? How are these runs related and organized?
2. What organism(s) are you looking for in these samples?
3. What differences do you expect to find between these environments?

**Helpful Hints and Reminders**

* If you run out of space on your instance, delete any left over .fastq files, these are almost certainly the largest files. Also, delete the ~/.ncbi/ folder, this contains the unextracted versions of the fastQ files.
* How To Search for Datasets: Go to the SRA website -> search for your search terms -> send to -> run viewer
* Once in the run viewer, be sure to narrow down your search by selecting “Assay Type” on the right side of the viewer and checking “wgs”. Remember, we only want WGS datasets (not amplicon/16s).
* Find related SRA datasets by searching for Projects (SRP, ERP, DRP).
* Use this repo to automate larger scans. If you want a new feature, email me.
  + git clone https://github.com/kylelevi/BAM\_Scripts