Module 1 tutorial

Rob Beiko

Since we have not yet covered much in the way of technical content, this tutorial will serve as an introduction to various data repositories and the corresponding files they contain. Repositories have different modes of access, contain various amounts of metadata, and offer different types of online analysis.

This is meant to be mostly a self-guided tour, with discussions about interesting questions that may come up. We will probably not have time to complete the entire tutorial, but I encourage you to choose those resources that are of greatest interest to you and continue to explore these after the tutorial session is done.

I have interspersed a few questions to focus your attention on certain aspects of these services.

Note on data access**:** in many cases repositories cannot take the strain of many simultaneous connections, and can even refuse connections if too much load is originating from one source. For example, NCBI's E-utilities (http://www.ncbi.nlm.nih.gov/books/NBK25497/) offer the following warning message:

> In order not to overload the E-utility servers, NCBI recommends that users post no more than three URL requests per second and limit large jobs to either weekends or between 9:00 PM and 5:00 AM Eastern time during weekdays. Failure to comply with this policy may result in an IP address being blocked from accessing NCBI.

This mode of access to NCBI can be very useful, but we will not be exploring it today. In general, you should not worry about querying databases unless you are writing software to access these resources automatically. The queries below are short in order to reduce our collective load on the servers.

Databases

(1) *MG-RAST*

Reference: Meyer et al. (2008): http://www.biomedcentral.com/1471-2105/9/386

Navigate to the MG-RAST home page at http://metagenomics.anl.gov/metagenomics.cgi?page=Home. Click on "Browse Metagenomes".

Warning! The browse function may take a while.

The Browse page gives summary statistics for the entire database, and a list of all 27,000+ metagenomes that are available. Click to take a look at the different terms you can use to refine your basic query: for example, under "Biome" you could choose "aquatic biome".

Q1. What is one immediate problem you can see with the refinement criteria?

Refine the Sequencing type to "Amplicon", and under "Biome" choose "black smoker". This should give you a list of 127 samples which are all associated with the same project. By clicking on the project you can get a conceptual overview of the project as well as associated publications.

Clicking on the name of a particular sample will bring you to the Overview page for the dataset. Let's choose "gua031" to investigate further. Feel free to try another example if you like; everything will still work but your results will be slightly different.

There is a lot of information here, including "Source Hits Distribution" which summarizes matches to Greengenes, RDP and SILVA; rarefaction curves; and taxonomic hit information. Let's take a look at the rank-abundance plot, which largely consists of different flavours of Archaea and, um, Eukaryotes (specifically fungi) as well.

Q2. Is it plausible that we might find fungi in these environments? Why or why not?

Let's go over and take a look at the download page. Here you can get the files corresponding to different steps in the analysis. Each type of file has a unique numeric identifier, for example rRNA gene clustering (which Will is going to talk about) which has an ID of 440. Right at the bottom of the page you can find a retrieval system for annotations. Let's download organism annotations based on RDP as reference database.

Q3. Open the tab-separated file that is produced. Which taxonomic assignments are least reliable at the genus level?

Finally, under "Analyze this metagenome" you can carry out visual analysis on the corresponding page. Click on "analysis page". Note the preselections; one criterion that must be changed is the annotation source. M5NR is the SEED curated set of functions, whereas we would like an RNA gene-based summary. Choose "SSU" from the dropdown menu and hit "ok". Then select "table" below.

Q4. From the resulting table, click on the link in the "Basidiomycota" row and download the supposedly fungal sequences. If you're feeling ambitious you can pop them into the NCBI nucleotide BLAST server and see what happens. What do you think has happened?

Be sure to check out the cool Krona graph as well.

The MG-RAST API
Reference: Wilke et al. (2015) http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004008

Browsing long lists of metagenomes can be an enormous pain, and MG-RAST has an application programming interface (API) documented at http://api.metagenomics.anl.gov/api.html; this allows you to generate custom queries directly via an http request. There is a tremendous amount of information that can be retrieved through this API. Note too as you hover over different links on the analysis page that everything is retrieved through the API as well.

The default return type is a JSON structure which contains relevant metadata as well as the answer to your query. There are libraries for importing JSON structs into (for example) Python; you might also want to install an add-on such as JSONView for Firefox (https://addons.mozilla.org/en-us/firefox/addon/jsonview/) to get slightly more readable results in your browser. Tabular results are returned as a BIOM-formatted file.

A few examples of calls to the API, related to our sample of interest:

http://api.metagenomics.anl.gov/1/download/4483535.3 - show the list of files that are available (results include URLs for each file)

http://api.metagenomics.anl.gov/1/compute/alphadiversity/4483535.3?level=phylum&source=RDP – show alpha diversity at the phylum level, using RDP as the reference source for annotation. The diversity score is exp(Shannon diversity).

[here is another tutorial from last year - http://adina-howe.readthedocs.org/en/latest/mgrast/]

Q5. Using the API, retrieve alpha diversity at the genus level for RDP, Greengenes and other sources if you like. Are they different? Why might they be different?

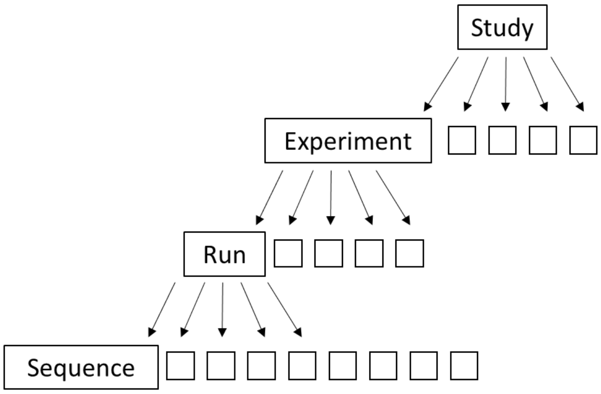
Note that our application GenGIS has a plugin that accesses MG-RAST through its API – you can look at the corresponding source code for inspiration if you are interested.

FTP access
ftp://ftp.metagenomics.anl.gov/projects (look at the README.ftp for different access modes)

This is a good way to get the reference databases that are used by MG-RAST. It is no longer browseable for metagenomic datasets; there are accessible through the API and the Python script outlined in the README file.

NCBI Sequence Read Archive (SRA) and FASTQ
URL: http://www.ncbi.nlm.nih.gov/Traces/sra/

The NCBI SRA has an overwhelming amount of raw data, all the way back to sequence traces. Included in this are marker-gene and metagenomic survey data from a remarkable range of projects. The organization of SRA is as follows:



Go to “Browse” to explore the different datasets that are available; instead of trolling through 55,000 studies, let’s search for “16S”. In my case this returned 1202 records. Add “robustness” to the search field to focus in on one set of studies in particular. Let’s choose DRP000903, the first entry in the list.

Each study in this set of 18 (here is the paper: http://dnaresearch.oxfordjournals.org/content/20/3/241.long) comes from a different individual in a probiotic intervention study. The study page contains several cross-referencing links to various other databases (such as BioProject) at NCBI, but let's stay within the SRA by clicking on "Runs".

On the resulting page you will see a table that summarizes the nine runs for this individual. The first was a pre-probiotic administration, the next four are identified as during the administration (eight weeks, with samples collected every two weeks), and the final four are unlabeled but likely correspond to the post-probiotic phase as identified in the paper. With the SRA we can drill as deeply as we want to into these data.

Click on the Run "DRR003411" to bring up a summary of this specific sequencing run, which corresponds to one time point from this single individual. We can into the Reads tab to look at each individual read, including the quality scores and intensity graph. You can flip through the reads on this page. Examining each read in this fashion is likely to be an unrewarding experience.

Want to download the data? This is where things get interesting. You can explore the FTP site with ftp://ftp.ncbi.nlm.nih.gov/sra/sra-instant/reads/ as a starting point, but NCBI recommends you not obtain the data in this manner. A more direct way of obtaining the sequence is to get the SRA Toolkit (https://github.com/ncbi/sra-tools/wiki/Downloads) and run utilities such as "fastq-dump".

There are many utilities and many options (see http://www.ncbi.nlm.nih.gov/Traces/sra/?view=toolkit_doc), but here is the short, short summary:
If you run "`fastq-dump.exe DRR003411`" from the Windows command prompt (or the corresponding program in Mac or Linux) you will get a FASTQ file from one run, corresponding to the reads, quality scores, etc. you saw on the web page. "`more DRR003411.fastq`" should give you the file contents.

Interpreting FASTQ

More will be said about FASTQ files in the afternoon, but the basic idea is that a FASTQ file stores the basecalls for a sequence along with associated Phred quality scores. The quality scores are represented as ASCII characters, with the choice of character = Q + an offset. There are three different encodings in FASTQ files: the original Sanger encoding (also used by the Roche 454 instrument), the Solexa encoding, and the early-middle Illumina encoding. I won't get into the details of this, but these are 454 reads so the encoding is of Sanger type.

Here is the Q score:

$$Q = -10log_{10}p$$

where p is the probability that a given base call is incorrect according to Phred. Sanger encoding adds 33 to the Q score to get the corresponding ASCII character. So if *p* = 0.001, Q = -10 * -3 = 30. And the entry in the FASTQ file will be the ASCII character whose code is 30+33 = 63. This is the question mark "?".

Q6. Look at the first few records in the FASTQ file you just downloaded. What is the maximum Q score and the corresponding minimum predicted error rate? How about the minimum Q score?

Q7. Do you see any tendencies for errors to occur more often in some parts of the sequence?

HMP DACC

URL: http://hmpdacc.org/

The Data Analysis and Coordination Center for the HMP contains information about reference genomes, 16S surveys (which you will be seeing soon enough!), metagenomes, with connections to taxonomy, function, pathway reconstructions, assemblies and protocols.

I'm not aware of an API for this system, so the primary mode of access is via the Web. For example, you can select "Illumina WGS Reads and Assemblies" and view the data table. If you open one of the sets (for example, "Anterior Nares" at the top) you can download the raw reads or assemblies via your browser. Files from the HMP are typically either in .gz or .bz2 format. Grab a file and take a look if you like; there will be no surprises in the contents.

Although there is no API, there are a couple of things you can do to semi-automate the process. These are unsupported but appear to work at the moment.

At the bottom of the table is a "Save as CSV" button. Click on this and you will get a comma-separated table of metadata. If you look at the HMASM table for example, you'll see that each sample has two associated URLs: one for assemblies and one for reads. We can make use of this.

There are three problems with this: (i) some of the csv tables are unavailable (for example, the processed 16S data at http://hmpdacc.org/HM16STR/), (ii) the csv files can be incomplete, and (iii) the URLs are *slightly* wrong. But we can deal with this in cases where the csv file is available.

Let's take SRS019379 as an example. The assembly file location is given as "/data/HMASM/IDBA/SRS019379.fna.bz2", which is obviously somewhere in some UNIX-like filesystem. One way to access it is through ftp:

```
ftp public-ftp.hmpdacc.org
```

It will ask you for a username and password; it makes no difference what you type here. A quick note that the HMP timeout is aggressive and callous, so if you stop typing for a minute or two you'll get booted.

Now you may notice that there is regrettably no /data/ directory. Instead you need to change directory (with the `cd` command) to HMASM, then to IDBA.

```
cd HMASM

cd IDBA

ls -1
```

will give you the directory contents. There are over 1000 files so it may take a moment or two. If you want a particular file you can use the 'get' command:

```
get SRS148408.fna.bz2
```

Want more than one file at a time? First we use "prompt" to switch off interactive mode, then use the "mget" command to grab a set of files that match a wildcard.

```
prompt

mget SRS1499*
```

will retrieve all files that start with "SRS1499". This is a relatively small number. Please DO NOT "mget *" during the workshop.

Want a simpler way to retrieve files, with more options, that can more easily be scripted? Of course you do. The solution is the GNU 'wget' command. The documentation is here: http://www.gnu.org/software/wget/

Here is the command (from the UNIX prompt) to retrieve a single file:

```
wget ftp://public-ftp.hmpdacc.org/HMASM/IDBA/SRS149961.fna.bz2
```

As before, we can use wildcards:

```
wget ftp://public-ftp.hmpdacc.org/HMASM/IDBA/SRS1499*
```

We can customize things further:

```
wget -A *1499* ftp://public-ftp.hmpdacc.org/HMASM/IDBA/*
```

which accepts any file from the directory containing "1499" in its name.

For the grand finale, we can operate directly on the csv file (this is all one line):

```
    grep bz2 HMASM.csv | grep 149 | cut -d"," -f3 | sed "s_/data_wget
    ftp://public-ftp.hmpdacc.org_" | sh
```

What the..?

The pipe command "|" allows us to chain commands together. If you want to understand this mess, I encourage you to build the command piece by piece. In order:

The first `grep` gets all lines from the file that contain "bz2".
The second `grep` gets all lines that contain "149"
The `cut` command splits lines on the comma character, and takes the third field (i.e. the one that contains the directory field)
The `sed s` command modifies the directory string into the wget command. The underscores split up the parts of the substitution: "s_poodle_chihuahua_" will replace "poodle" with "chihuahua". The most common delimiter for sed is "/", but this makes a mess when there are a number of "/"s inside the substitution string as well.
And `sh` makes it all run.

There are other options, including `lftp` (which has some basic scripting functionality) and `curl`. For text parsing, `sed` and `awk` are extremely powerful: you may never need Python again (not really).

QIITA
URL: http://qiita.microbio.me/

QIITA (Github: https://github.com/biocore/qiita/) is a repository for metagenomic data, built on QIIME, that hosts a large number (about 45,000) of studies from the Earth Microbiome Project. Although we used the EMP QIIME database in the past, QIITA is fairly new and requires registration. The basic workflow involves selecting studies, and submitting them for analysis (e.g., beta diversity).

Other resources include:
EBI Metagenomics: https://www.ebi.ac.uk/metagenomics/, reference http://nar.oxfordjournals.org/content/42/D1/D600.long

JGI IMG/M: https://img.jgi.doe.gov/cgi-bin/m/main.cgi (account required)

RDP: https://img.jgi.doe.gov/cgi-bin/m/main.cgi

The MetaHIT consortium (of "enterotypes" fame):
https://www.sanger.ac.uk/resources/downloads/bacteria/metahit/