Analyze a uBiome sample yourself

The uBiome JSON taxonomy information will be more than enough for most people new to biology. For if you’re up for expanding your bioinformatics skills, the raw data files are the place to look.

This book uses the uBiome service for all examples because at the time of writing, that’s the most convenient and low-cost way to get your data. If you don’t have access to uBiome, you can still run all the examples using the techniques described here.

Please note that although this will give you a file that lets you run all the examples, your taxonomy data may be significantly different from the results you get from uBiome. Each lab uses its own database of 16S lookup data, but uBiome’s is proprietary and not available to other labs.

Scientists from different labs will argue fiercely for the correctness of their own particular way of analyzing data, but the science is too new for much agreement. Fortunately, this won’t generally affect the conclusions you make from your own experiments as long as you are consistent. It’s meaningless to compare data generated from uBiome’s lab with one from another lab, so just stick to one format.

FASTQ files are the text-based results of a run at your sample using one of Illumina’s high-end gene sequencers.

For example, the first line in my data is:

@NS500457:29:H2HLGAFXX:1:11101:16760:1033 1:N:0:ACTTACA+NTGCGAA

This contains information about the gene sequencing hardware uBiome used for this sample. The other lines in the raw text file contain the actual base pairs detected by the sequencer, along with quality information to indicate the certainty of a particular read. In my case, the entire file is over 170,000 lines long, and that’s just one of the eight files in the FASTQ archive. That’s a lot of data!

# Some background

Like most labs, uBiome starts with the V4 region of the 16s rRNA, which is on average 292bp (base pairs) long . They read with an Illumina NextSeq500 at 145-147bp from each end. If each forward and reverse were read from the same lane independently, then you have sequences of only 145-147bp to map to known sequences, which may lead to several alternative genuses to which annotate a sequence to. Instead, if you use both reads from a lane as one single biological entity, the number of 16S sequences to which it maps it will be substantially reduced and thus more accurate. In some experiments we have performed, we have seen that annotating the same sample using single reads vs pair-end reads can lead to dramatically different phylogenetic annotations.

To read the fastq files correctly, then,

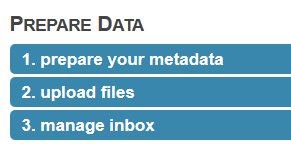
1. Pair up all forward and reverse reads from your raw data download,
2. Put all pair-end reads in just one big file

Although it’s possible to analyze this data on my own in a tool like BioPython, it’s much easier to submit the raw data to a public gene processing server, such as the Metagenomics Analysis Server (MG RAST) hosted by Argonne National Labs at this site:

<http://metagenomics.anl.gov/>

Note: You must use the Firefox browser (not Chrome, IE, or Safari), and before submitting any jobs you must apply for a (free) login ID, which you receive by email in a day or two.

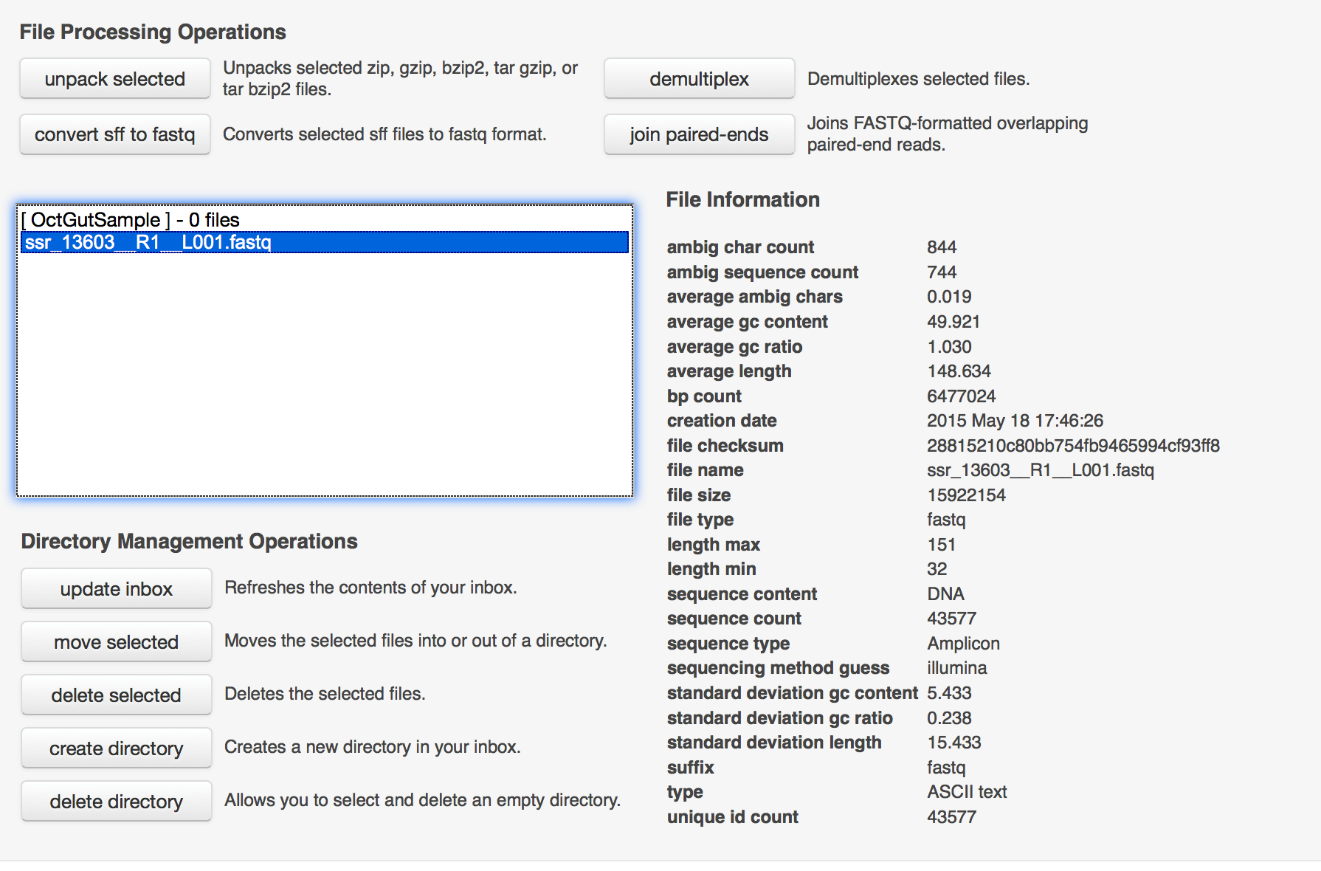
After logging in with your new account ID, click the “upload” icon on the home page and follow the process to prepare and submit your data:



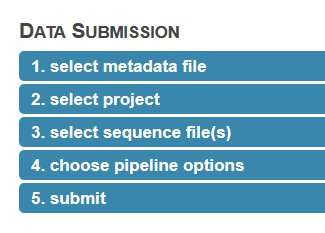
1. Ignore the part about preparing the meta data; you can add it later.
2. Upload files. The site offers to let you upload from the web browser, but I’ve had better luck using their api and the following curl command:

curl -H "auth: webkey" -X POST -F "upload=@/path\_to\_file/metagenome.fasta" "<http://api.metagenomics.anl.gov/1/inbox/>" > curl\_output.txt

I just type this from Terminal on my Mac. Note that you’ll need to get your own webkey (hit the ‘view webkey’ button on the submission page) which you will substitute into the curl command line above. Also change the filename part of the “upload=@” section to point to the correct file. Note that uBiome fastq files are compressed, but it’s okay to upload them as-is. You’ll uncompress them on the server after the upload.

1. Manage your inbox. Once the file has successfully uploaded, you’ll see it in your inbox. At this point, you should “unpack selected” if you uploaded a compressed file. Note depending on how busy the server is, it may be minutes or even hours of waiting, and pressing “update inbox” before you see the File Information shown below.

Next you’re ready to submit the data



Here are my suggestions for this section

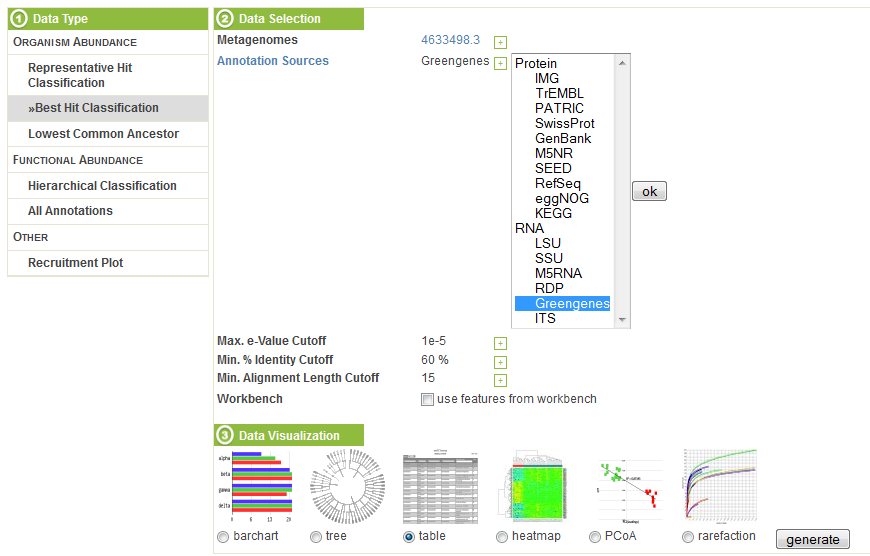
1. Check the box to say you won’t supply metadata.
2. Create a new project and select it.
3. Select the files you uploaded
4. Use the default values to choose pipeline options
5. Submit the job, being sure to make the data publicly accessible immediately, to increase the priority.

There you go! Now a job has been submitted and you wait for it to complete. It usually takes a day or two to complete, after which you’ll receive an email notification. After that, click the bar chart icon on the upper right of the page:

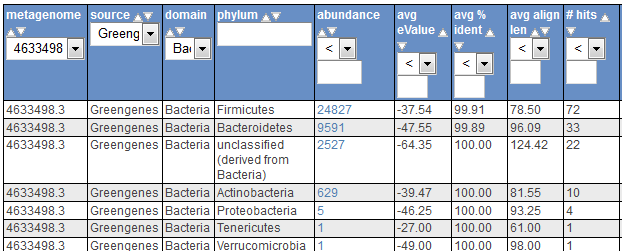
C:\Users\Richard Sprague\AppData\Local\Skitch\Screenshot_052115_112557_AM.jpg

This brings you to an analysis page where you can study your data in much more detail.

When I select my data under “Metagenomes” on the analysis page, I’m offered several options for Annotation Sources. Since the uBiome fastq data comes from 16S ribosomal RNA, I choose to compare my data against one of the large databases of known RNA. The popular Greengenes database is one of them, so I select it, and then ask to generate a table of the results.



The table, sorted by abundance and filtered to phylum, looks like this:



The results roughly conform to what we see from uBiome: the most common organisms are the same in each case, and although the scales are different, the abundance computed corresponds roughly to what we see from the uBiome count field.

Other MG RAST tools let you compare your data against other databases of bacteria and proteins, graph the results, and perform sophisticated statistical analysis to determine other unique features of your sample. You can also compare your sample to microbiomes that have been uploaded by others, enabling you to study your results in much more detail.