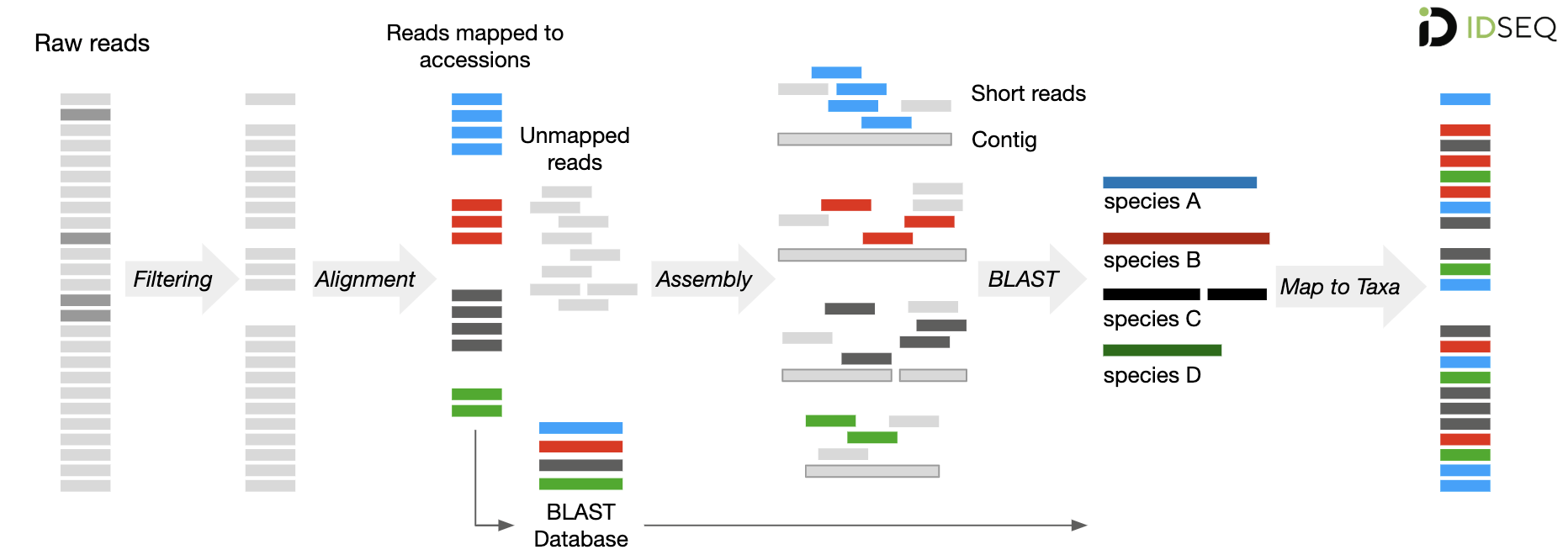
**Exercise 1: Clinical metagenomic analysis using IDseq**

EMD 531, Spring 2020

*Exercise template provided by Rebecca Egger, Product Lead, Chan Zuckerberg Initiative.*

The purpose of this exercise is to explore clinical metagenomic (mNGS) data using IDseq, a cloud-based infectious disease platform that automates many computationally-intensive steps. A visual representation of the pipeline steps is shown below.



IDseq then provides a user interface that simplifies data exploration and hypothesis generation.

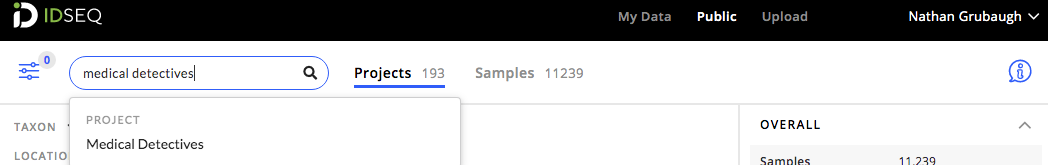
*Code Availability:* The code for the IDseq web-portal is open-source and available at the following url: [https://github.com/chanzuckerberg/idseq-dag/wiki.](https://github.com/chanzuckerberg/idseq-dag/wiki) The pipeline code is available at the following url: <https://github.com/chanzuckerberg/idseq-dag>.

*Documentation:* Pipeline documentation containing parameters used for each step of the pipeline can be found on the wiki: <https://github.com/chanzuckerberg/idseq-dag/wiki>

**Overview:**

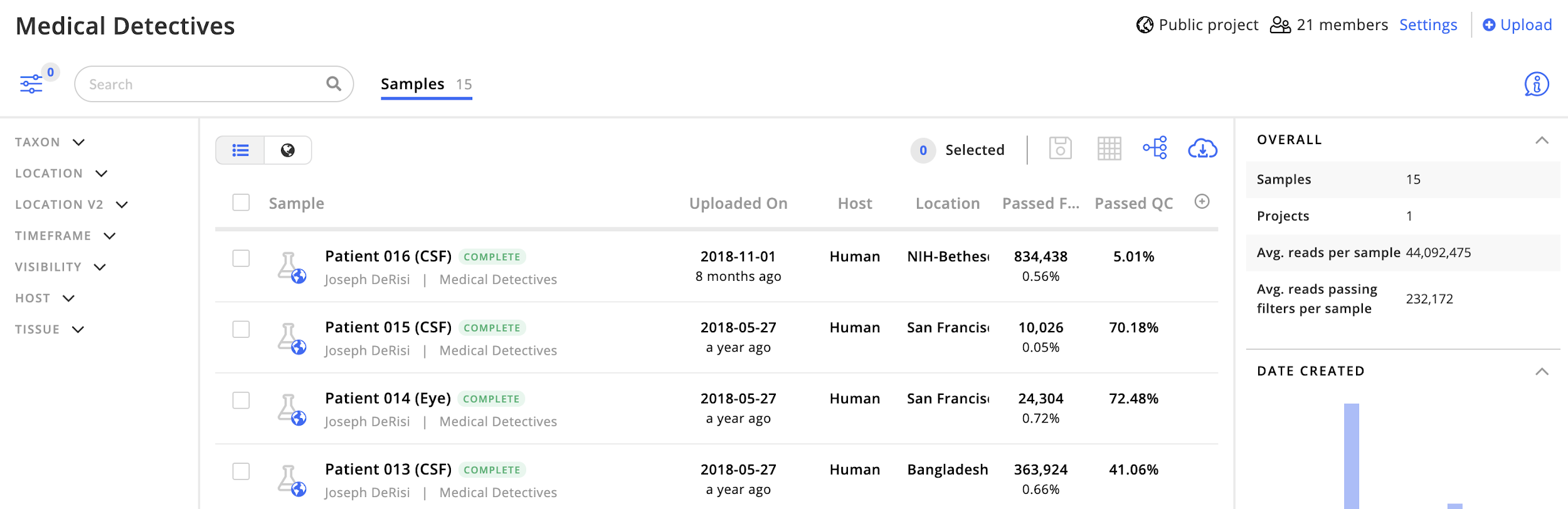
**Step 1:** Sign into IDseq and find the “Medical Detectives” data.

* You should have received an email from IDseq with your registration and log-in information.
* Register for an account and sign in.
* Select **Public** on the top tool bar and search for “medical detectives”

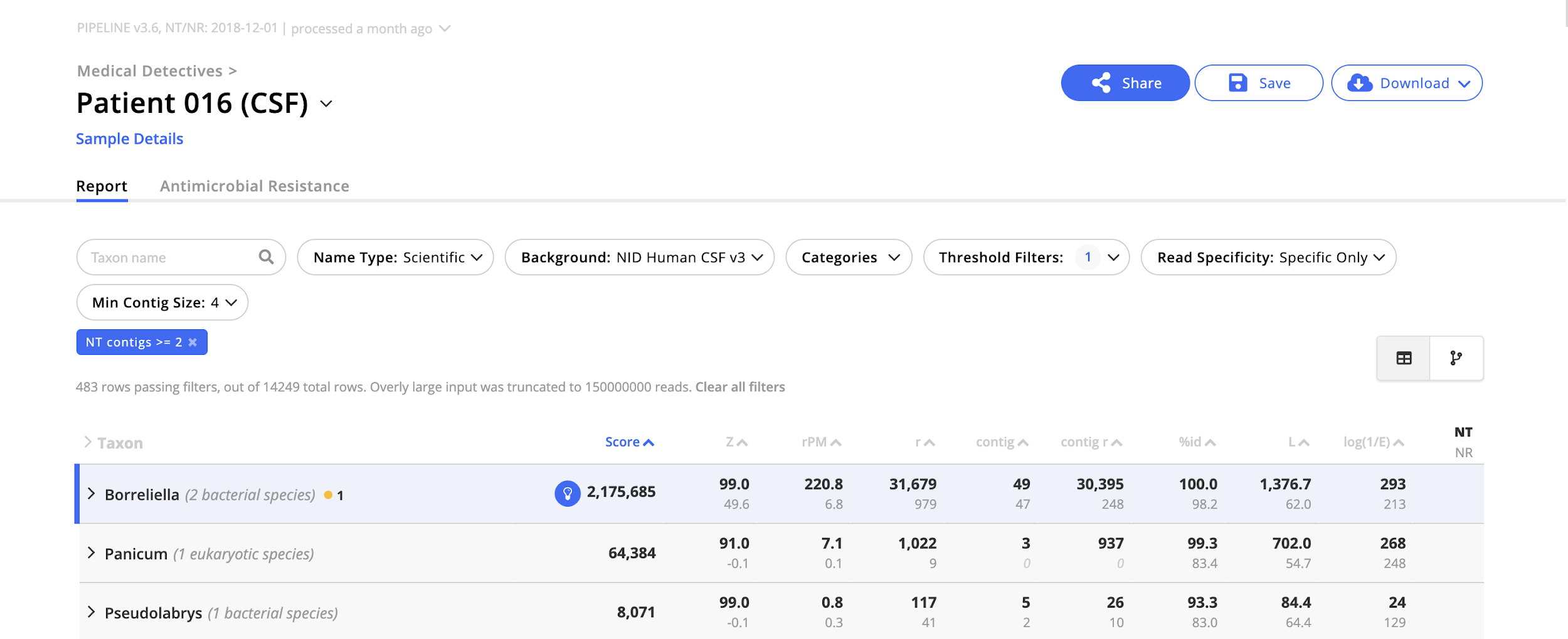
****

**Step 2**: Data reports

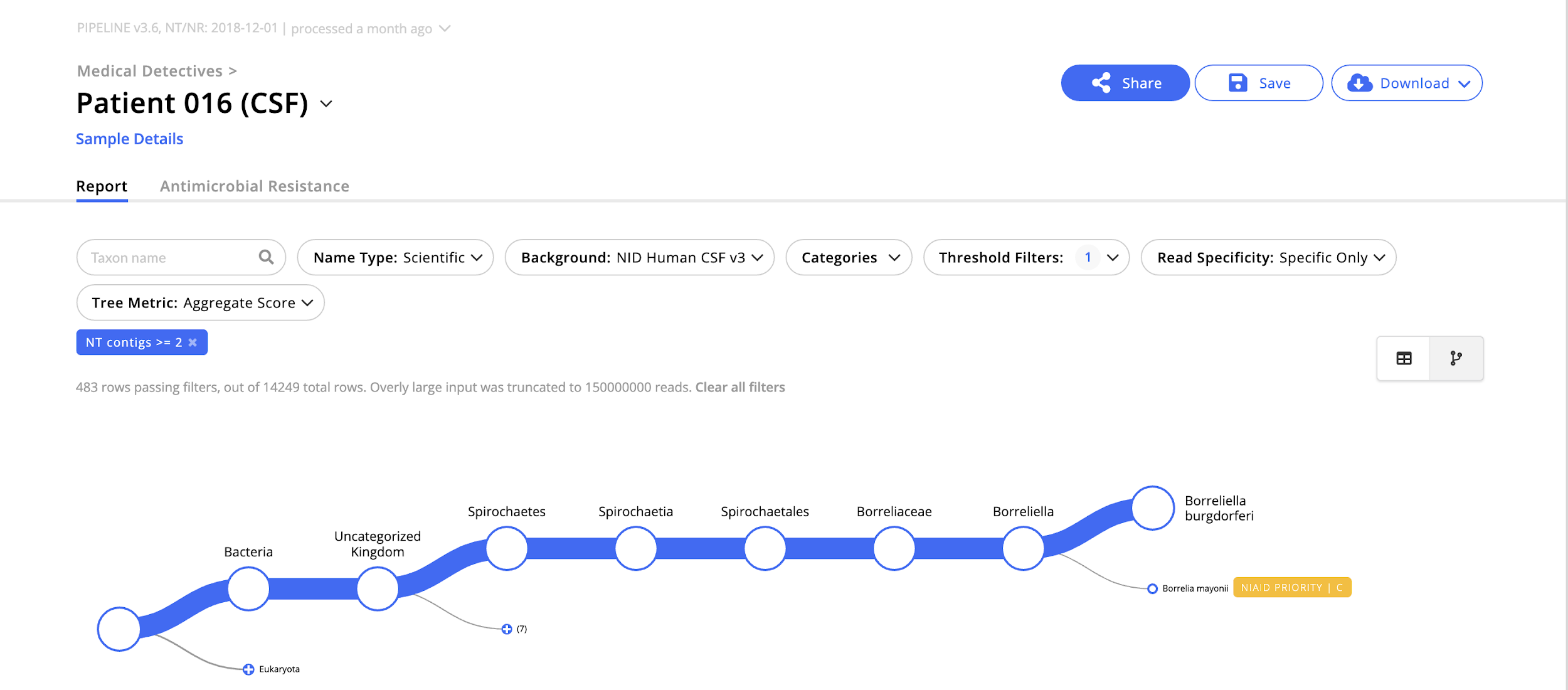
* There are 15 samples in this IDseq project. From the project page, you can see the dashboard of samples along with several metrics from the pipeline run.



* For the in-class demonstration, select sample **Patient 016 (CSF)** to follow along. For the questions at the end, select a different sample.
* IDseq provides two single-sample report views.
  + First, is the **Table View**, which provides a report containing several fields indicating the relative abundance of genera and species identified in the sample. An example is shown below:



* + Second, the **Taxonomic Tree View** provides a phylogram indicating the taxonomic relatedness of all microbes identified in the sample. The weight of the lines connecting tree nodes is indicative of the relative abundance. An example is shown below (we’ll come back to this later):

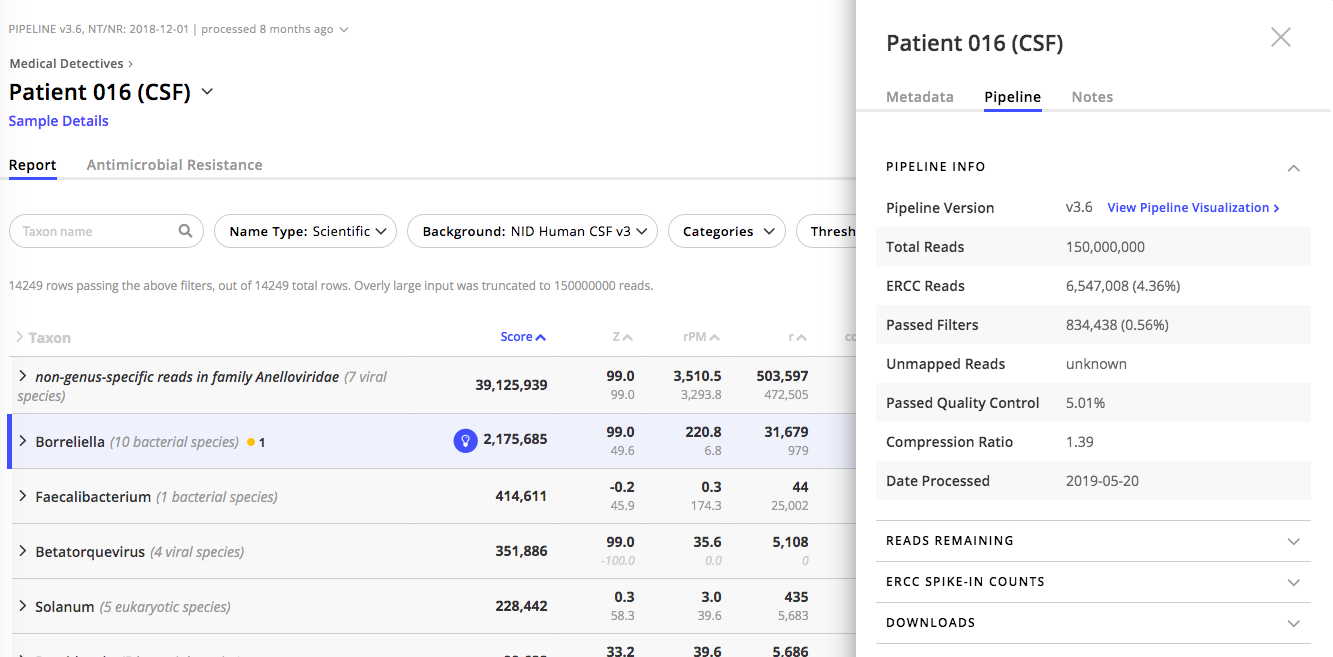


**Step 3**: Data QC and filtering

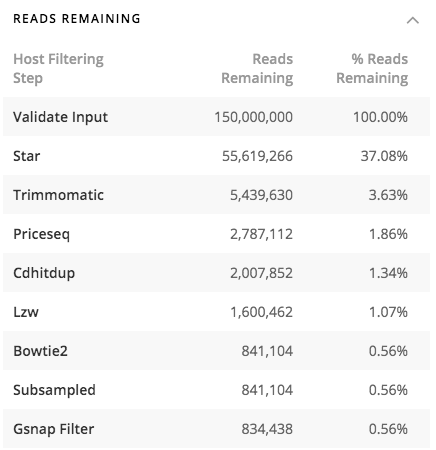
* The IDseq pipeline performs a series of filtering steps to remove *host* sequences and low-quality sequences. More details on the IDseq pipeline steps can be found here: <https://github.com/chanzuckerberg/idseq-dag/wiki>
* In particular, the following QC and filtering steps are performed:

|  |  |  |
| --- | --- | --- |
| **Step** | **Order** | **Purpose** |
| 1 | STAR Host Removal | Filter out most of the host sequences  (often > 90% of the sequencing reads) |
| 2 | Trimmomatic | Trim Illumina adapters |
| 3 | PriceSeq | Remove low-quality sequences |
| 4 | CD-Hit-Dup | Remove duplicate sequences generated through PCR amplification |
| 5 | LZW Complexity Filter | Remove low-complexity sequences  (ie long runs of a single nucleotide AAAAA) to improve alignment speed |
| 6 | Bowtie2 Host Removal | Filter out any remaining host sequences |
| 7 | GSNAP Human Removal | Filter out human sequences  (regardless of the host organism) |
| 8 | Subsampling | Subsample to 1 million fragments to increase pipeline speed |

* From the Table View, click on the **Sample Details** link (below sample ID, “Patient 016 (CSF)” in this case). Then in the Sample Details side-panel, select **Pipeline**:



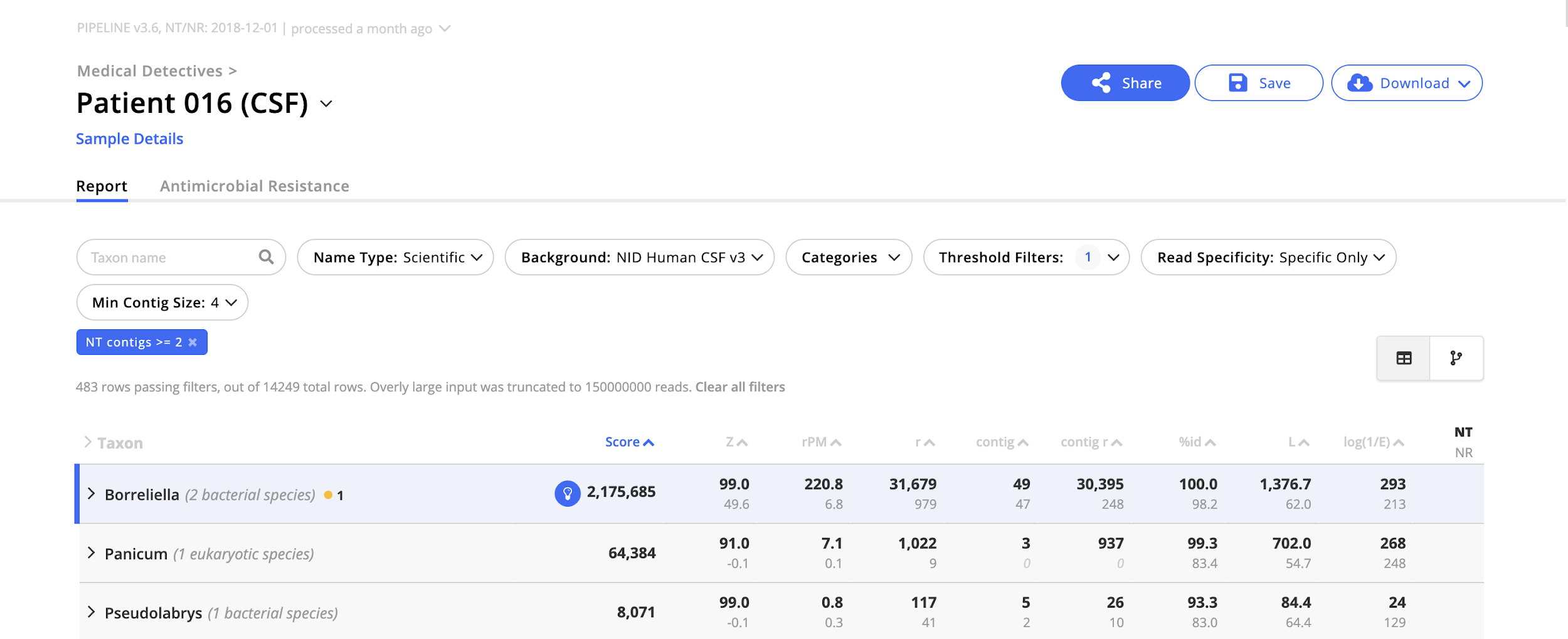
* The Total Reads is the amount of data coming from the sequencer.
* ERCC Reads are the number of positive control reads in the sample.
* Passed Quality Control is the percentage of reads that came out of PriceSeq, step (3) of the host filtration and QC steps, compared to what went in to Trimmomatic, step (2). Passed QC < 10% is generally considered low.
* Passed Filters is the percentage of reads that came out of step (8) of the host filtration and QC steps as compared to what went in at step (1). These are the number of reads used in the microbial analysis.
  + To see a breakdown of how many reads were removed at each QC and filtering step listed above, select the **Reads Remaining** dropdown menu:



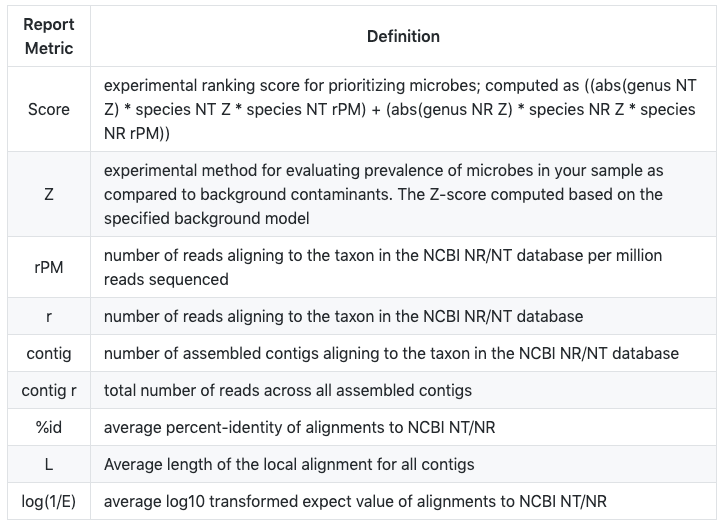
* The sequences remaining after QC and host filtering, 834,438 in this case, are potentially microbial. Some sample types typically have > 99% of reads removed during the host filtering steps (like this sample). Other sample types may have only 70% of reads removed during host filtering steps. This has to do with the ratio of host cells to microbial cells present in the sample.

**Step 4**: Report Table View and Filters

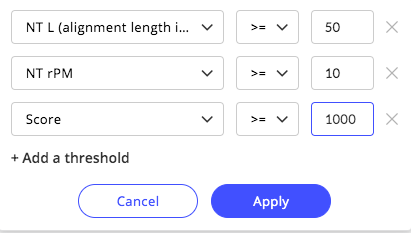
* To identify their species of origin of the “Passed Filters” reads, are aligned to the NCBI NT and NR databases. The NCBI NT (nucleotide) database consists of all nucleotide sequences from all organisms. Meanwhile, the NCBI NR (non-redundant) database consists of all protein sequences.
* To view how the reads aligned to different species, close the Sample Details side panel and look at the Report Table.



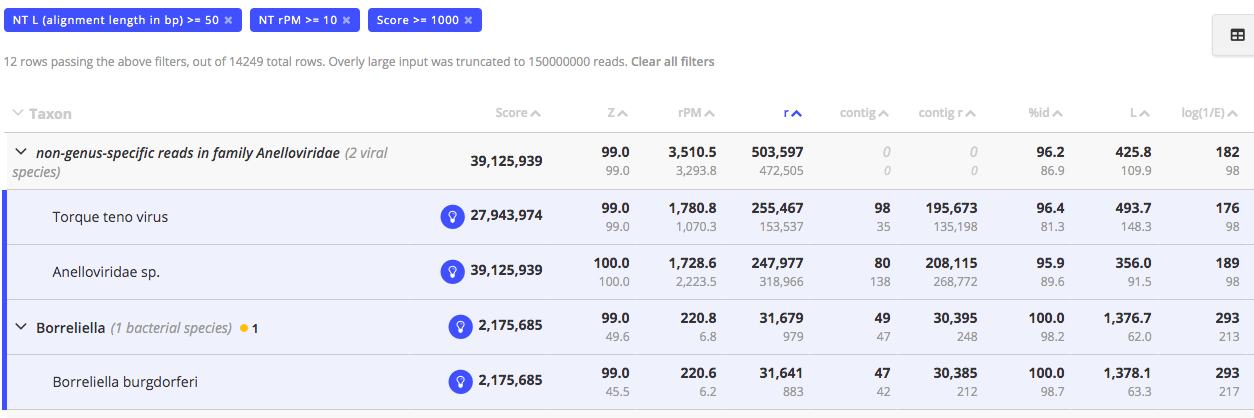
* There are a series of values in the Report Table. The definitions of these values can be identified by looking at the documentation here: <https://github.com/chanzuckerberg/idseq-dag/wiki/IDseq-Pipeline-Stage-%233:-Reporting-and-Visualization>



* When comparing between samples, rPM is used as a scaled metric of abundance. Each time we run a sequencing experiment, we may obtain different numbers of total reads. To scale the values across experiments we use rPM instead of raw read count.
* Sort the table by the **rPM** column.
* To reduce noise, filters are applied to the analysis. mNGS is a highly sensitive and it is possible to pick up low levels of sequences associated with environmental contaminants. These are often not relevant for making a decision regarding the infecting microbe. The filters allow us to focus on highly abundant species within particular categories (i.e. Viruses, Bacteria, etc.).
* At the top of the report page, there are a series of filters available. Select **Threshold Filters**
  + Select **NT L (alignment length in bp)** and set to **>= 50**
  + Select **NT rPM** and set to **>=10**
  + Select **Score** and set to **>= 1000**
  + Select **Apply**



* Click on the "**>**" next to Taxon to expand the table and view the species-level metrics.
* Now you should see that out of the 14,249 species represented in the total microbial data, we have now filtered it down to 12 to focus on.

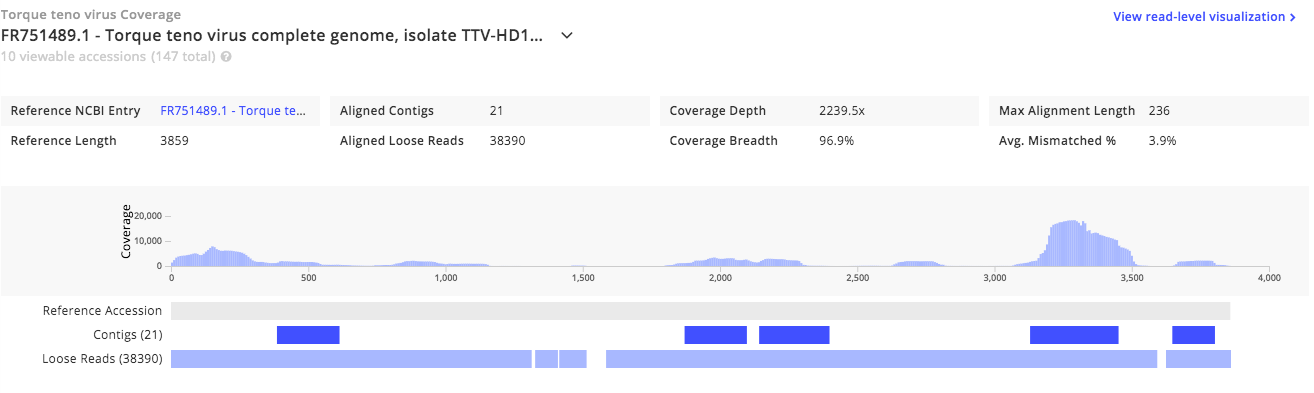


**Step 5**: Coverage Visualization

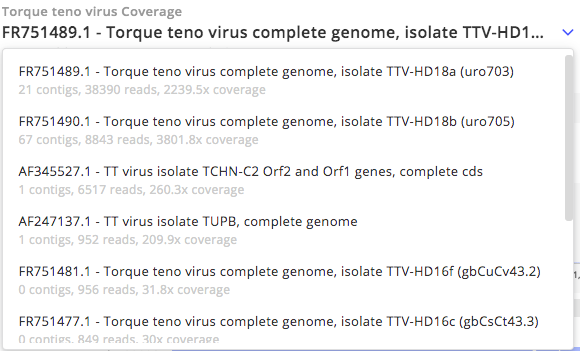
* For the microbe species with the greatest rPM value, select the Coverage Visualization icon.



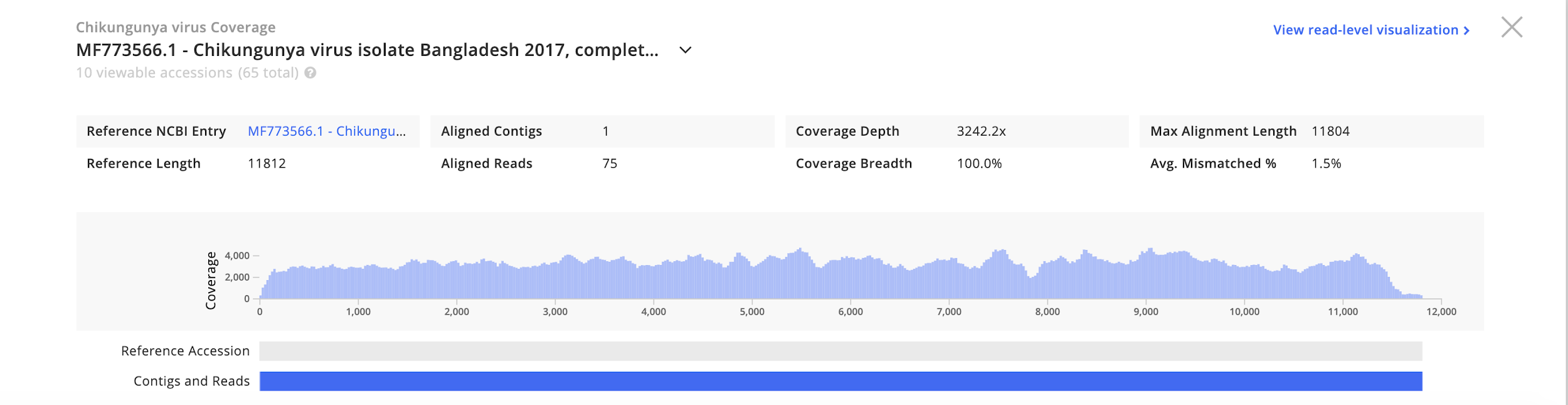
* This should open a pop-up window showing the coverage using microbial sequences from NCBI as references.



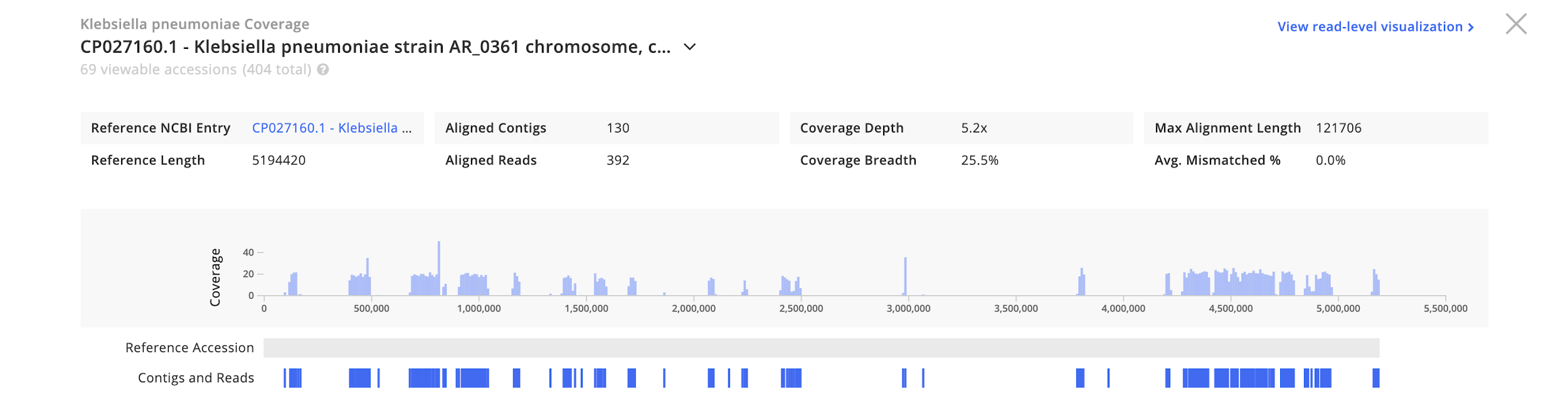
* The top coverage plot shows where individual reads align to the reference sequence. In this case, 38,390 “loose” reads aligned to the 3859 nt reference sequence.
* IDseq also takes the short reads and assembles them into contigs (longer, contiguous fragments of the same species). In this case, 21 contigs aligned to the reference sequence.
* Note that the alignments can be to several different sequences. The total number of reads (**r**) shown on the Table View (153,537 in this example) are the sum of all alignments with the reference sequences shown in the dropdown menu.



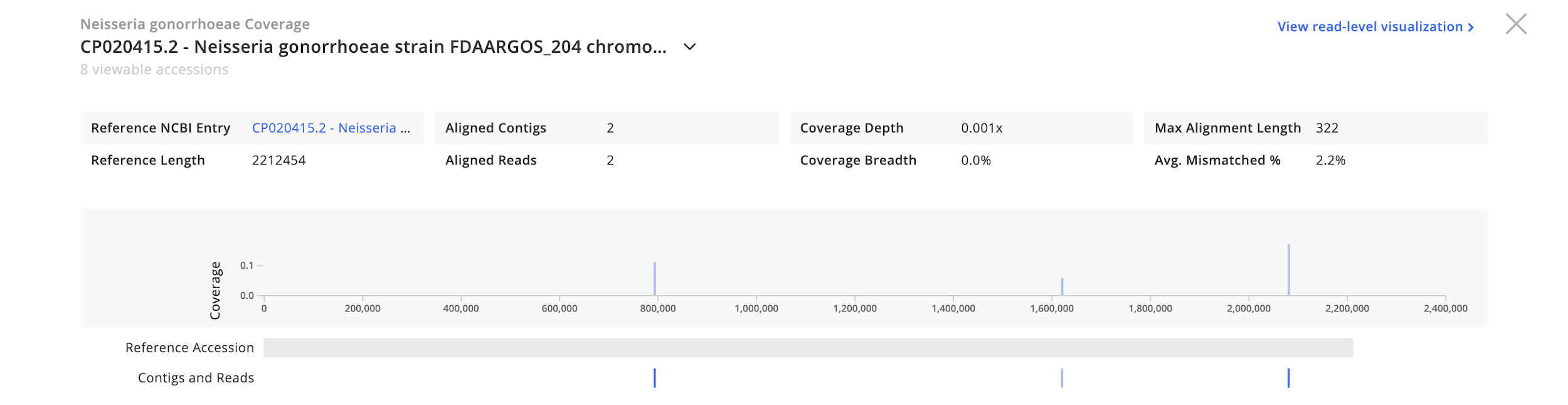
* The coverage visualization can be used to improve our confidence in the presence of a particular microbe. Specifically, a microbe that has high coverage or coverage at multiple loci across the reference sequence is likely to be a real hit. Whereas a microbe with low coverage of only a small portion of the genome may be a result of genomic similarity at conserved regions (e.g. bacterial 16S rRNA). Notably, since viral genomes are smaller than bacterial and eukaryotic genomes, viruses often have higher coverage of their genomes. As an example (not from Patient 016 (CSF)), here is high coverage of a complete viral genome.



* As another example (not from Patient 016 (CSF)), this shows high coverage of a bacterial microbe (obtained from a cultured isolate).

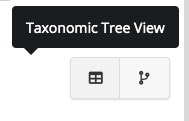


* Finally, here is a low-confidence bacterial hit with relatively few reads, isolated to a small region of the genome. This is likely not real.

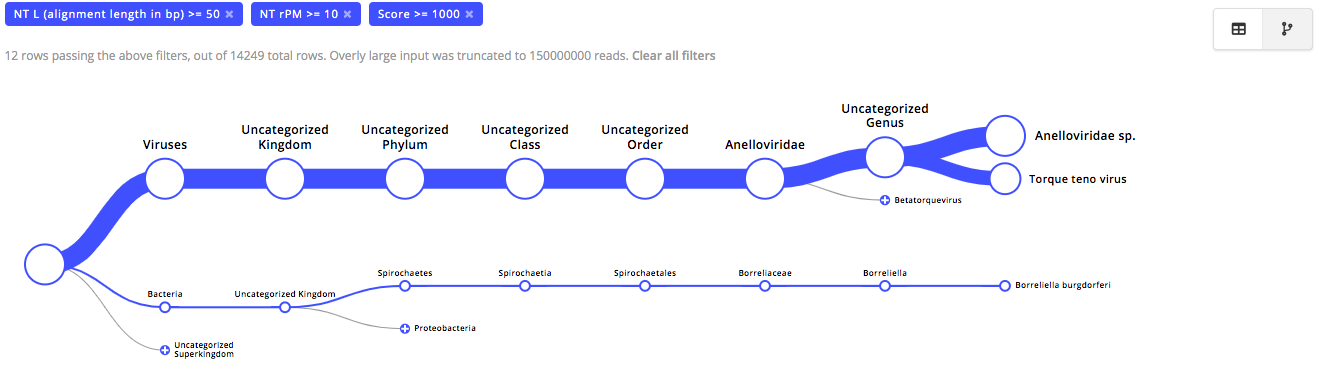


**Step 6**: Report Taxonomic Tree View

* Visualizing the report can be challenging. One way to view the data is to select the **Taxonomic Tree View** icon in the upper right corner of the Sample Report.



* You should see a tree showing the relative abundance (line widths) and taxonomic relationships of all the microbes identified in the samples. The following image shows the phylogenetic relationships for a single sample, including annotations for the taxonomic levels.



**Questions:**

*Combine answers with all exercises,* ***due before class 16 on March 5th****. Be prepared to discuss your answers during class 16.*

**Select a different sample than what was used for the demonstration (Patient 016 (CSF)) and perform the filter steps outlined in Step 4. Use the tutorial above to answer the following questions.**

**E1-1**: Identify your sample by sample name and determine the following metrics:

|  |  |  |  |
| --- | --- | --- | --- |
| **Sample** | **Total Reads** | **Passed Filters** | **Passed Quality Control** |
|  |  |  |  |

Why does only a small percent of reads pass filter? If you had a tumor biopsy and serum sample from the same patient, from which would you expect a high percent to pass filter (and why)? What library preparation approaches could you use to increase the percent that pass filter? (Max 300 words; 5 points)

**E1-2**: How many species (rows) were identified before and after you applied the threshold filters? What are the potential sources of many of the filtered species? What steps in the laboratory could you take to reduce these reads? (Max 300 words; 5 points)

**E1-3**: After applying the threshold filters, what is the species with the greatest number of reads (and how many reads)? What is the species with the fewest number of reads? Paste screenshots of the coverage plots for both. Based on the read numbers and distribution across the reference sequences, do you believe that both of these microbes are actually in the sample? (Max 300 words; 5 points)

**E1-4**: Provide a screenshot of the taxonomic tree report post threshold filters. Do you believe that you detected a human pathogen that may have been infecting the patient? What is your evidence for this based on the IDseq data and your knowledge of the microbe? (Max 300 words; 5 points)