Whole Genome Metagenomics Workshop: Centrifuge

Centrifuge: rapid and sensitive classification of metagenomic sequences[?]

Introduction

What is Centrifuge?

Centrifuge is a very rapid and memory-efficient system for the classification of DNA sequences from microbial samples, with better sensitivity than and comparable accuracy to other leading systems. The system uses a novel indexing scheme based on the Burrows-Wheeler transform (BWT) and the Ferragina-Manzini (FM) index, optimized specifically for the metagenomic classification problem.

Tutorial

For this tutorial we will use data from this publication "Metagenomic Characterization of the Human Intestinal Microbiota in Fecal Samples from STEC-Infected Patients" [?]: https://www.frontiersin.org/articles/10.3389/fcimb.20 All data used in this tutorial was retrieved from SRA with fastq-dump. SRAccession: PRJEB23207.

In this study, whole genome metagenomic sequencing was used to investigate possible changes in the composition of the intestinal microbiota in samples from patients with Shiga toxin-producing E. Coli (STEC) infection (N = 2) compared to Crohn's Patients (N =4), healthy (N = 2) and healed controls (N = 3). Faeces samples, collected during an outbreak, from STEC infected patients showed a lower intestinal abundance of beneficial microorganisms in comparison to controls where those microorganisms predominated. These differences were observed with bioinformatic approaches and seemed to be related with the STEC infection. So, using the metagenomics data from this study can you identify the STEC positive samples? Can you identify the specific strain of STEC?

Many of the steps have been done for you. However, I have provided instructions and code so that you can build the full database indexes that we will use for classifying reads with Centrifuge.

Activate the Appropriate Environment

- Remember to activate a screen.

```
# Activate fastqdump environment
source activate fastqdump
```

Download Metagenomic Data from Sequence Read Archive (SRA)

- Download data from SRA using fastq-dump

```
# Create a directory to store results

mkdir -p /home/<username>/centrifuge/fastq
mkdir -p /home/<username>/centrifuge/taxonomy

# Copy the reads into your fastq directory

cp /home/elavelle/centrifuge/fastq/*.fastq /home/<username>/centrifuge/fastq
```

```
# Detach screen while this takes place

# This is an example for downloading a single fastq file from SRA:

prefetch -O . -p ERR2271237

fastq-dump --split-files --outdir . --gzip --skip-technical \
--readids --read-filter pass --dumpbase --split-e --clip \
/home/<username>/ERR2271237

conda deactivate

# Do NOT do for this tutorial.

# Use a for loop to download many files sequentially. First, write all of the SRA accession numbers
# that you want to download into a file, e.g., PRJEB23207_accessions.txt

# Use a for loop to store the SRA accession numbers from the PRJEB23207_accessions.txt file
# into a variable named accessions

accessions=`for i in PRJEB23207_accessions.txt; do cat $i;done`

# Use a for loop to run prefetch fastq-dump for each SRA
# accession number in the accessions variable

for j in $accessions; do prefetch -O . -p $j; done

for j in $accessions; do fastq-dump --outdir . --gzip --skip-technical \
--readids --read-filter pass --dumpbase --split-e --clip \
/home/condasw/data/Metagenomics/centrifuge/$j; done
```

Retrieve RefSeq and EuPathDB Data

Centrifuge provides a few scripts that will allow you to download the NCBI taxonomy tree and RefSeq reference genomes.

- To map sequence identifiers to taxonomy IDs, and taxonomy IDs to names and their parents, three taxonomy tree files are necessary in addition to the reference fasta files:
  - taxonomy tree files:
    - nodes.dmp from the NCBI taxonomy dump. Links taxonomy IDs to their parents names.
    - names.dmp from the NCBI taxonomy dump. Links taxonomy IDs to their scientific name.
    - seqid2taxid.map is a tab-separated file that maps sequence IDs to taxonomy IDs.
- When using the provided scripts to download the genomes, these files are automatically downloaded or generated. When using custom taxonomy or sequence files, please refer to the manual section TODO to learn more about their format. Here is the link to the manual https://ccb.jhu.edu/software/centrifuge/manual.shtml
- command flags:

  -o

  specifies the folder to which the files are downloaded.

  taxonomy

  Argument indicates the database to download. The options are refseq, genbank, contaminants or taxonomy.

```
# It took 20 seconds to complete during testing

source activate centrifuge
```

```
centrifuge-download -o /home/<username>/centrifuge/taxonomy taxonomy
```

- Download complete fungal genomes from RefSeq with the centrifuge–download commands. Do this exercise for this tutorial on a reduced data set.
- command flags:

  -a Only download genomes with the specified assembly level. Default: 'Complete Genome'. Use 'Any' for any assembly level.

  -m Mask low-complexity regions using dustmasker. This is part of the blast program that was installed in the metagenomics environment.

  -P Number of threads or processes to use when downloading.

  -d What domain to download. One or more of bacteria, viral, archaea, fungi, protozoa, invertebrate, plant, vertebrate_mammalian, vertebrate_other (comma separated).

  refseq Tells the program to use the refseq database.

```
# This took about 1 minute to run during testing

centrifuge-download -o library -a "Complete Genome" -m -P 4 -d "fungi" refseq > seqid2taxid.map
```

- Do not do: This database was pre-built for you to use in the tutorial. Download complete archaea, bacteria, viral and fungal genomes from RefSeq with the centrifuge–download commands. This is what you should do to build a more comprehensive RefSeq database containing only complete genomes.
- At the time of download, there were 289 Archea, 13,287 bacterial, 8,583 viral, and 10 fungal genomes.

```
centrifuge-download -o library -a "Complete Genome" -m -P 30 -d "archaea,bacteria,viral,fungi" refseq \
> seqid2taxid.map
```

- Download the RefSeq chromosome level human genome reference with centrifuge-download commands.
- Command flags are the same as the previous step:

  -t Only download the specified taxonomy IDs, comma separated. Default: any. 9606 is the taxonomy id for Human.

  -c Only download genomes in the specified refseq category. Default: any.

  >> Appends the human seqid2taxid. map to the previously generated fungi seqid2taxid.map file.

```
# This step took 14 minutes during testing.

centrifuge-download -o library -a "Chromosome" -m -P 4 -d "vertebrate_mammalian" \
-t 9606 -c "reference genome" refseq >> seqid2taxid.map
```

- *Optional. Customize your database by adding EuPathDB release 28 reference genomes. This step was done for you. The EuPathDB data comes from this publication [?].
  - Here is the link to the publication https://doi.org/10.1371/journal.pcbi.1006277 . Here is the webpage for the data download https://ccb.jhu.edu/data/eupathDB/

```
# Make new directory called eupath in the library directory. Something like this

mkdir -p /home/<username>/centrifuge/library/eupath

# Move into the /home/<username>/centrifuge/library/eupath directory.

cd /home/<username>/centrifuge/library/eupath

# This wget command will download EuPathDB data from the following link.

wget https://ccb.jhu.edu/data/eupathDB/dl/eupathDB.tar.gz

# Uncompress the eupathDB.tar.gz file into the eupath directory. You should see
# a directory named library

tar -zxvf eupathDB.tar.gz

# Move the fasta files from the /home/<username>/centrifuge/library/eupath/library directory
#to /home/<username>/centrifuge/library/eupath/ directory

mv /home/<username>/centrifuge/library/eupath/library/*.fna \
/home/<username>/centrifuge/library/eupath

# Extract and append the contig/scaffold and taxonomy id information from the prelim_map.txt that was
# downloaded with eupathDB.tar.gz file to the seqid2taxid.map file

awk -F'\t' '{print $2,$3}' /home/<username>/centrifuge/library/eupath/library/\
prelim_map.txt | awk -F'\|' '{print $1,$3}' | awk -F' ' '{print $1"\t"$2}' \
>> /home/<username>/centrifuge/seqid2taxid.map
```

- Concatenate the reference genome fasta files that were downloaded with the centrifuge-download commands. We are using a for loop to do this; otherwise, we would have to type every file name on the command-line that we want to concatenate- e.g., cat fasta_1.fna fasta_2.fna .... >> all_fasta_files.fna . For practice, we will concatenate the fungi reference genome fasta files and the human fasta file.

```
for i in /home/<username>/centrifuge/library/*/*.fna; do cat $i \
>> fungi_Hum.fna;done

# Centrifuge doesn't like the header format of some sequences. Reformat the headers:

sed 's/|kraken[^|]*|/ /' fungi_Hum.fna > fungi_HumanFixed.fna
```

Build the Centrifuge Database Indexes

Build the Centrifuge indexes using the centrifuge-build command. This has been done for you already. It took 14 hrs and ~ 410 GBs of RAM using 30 threads to build this database index on logrus. And, this was just using the "Complete Genome" Refseq genomes for archea, bacteria, viruses, fungi, human, and EuPathDB. I attempted to build a Centrifuge index using "All" RefSeq genomes for archea, bacteria, viruses, fungi, human plus the EuPAthDB but ran out of memory on logrus. I tried re-bulding this index on a higher memory server and Centrifuge was using 890 GBs of RAM before I killed the program. So, unless you have high-capacity compute you may not be able to build the indexes on your system. But, the Centrifuge developers have provided various premade indexes that you can download here, ftp://ftp.ccb.jhu.edu/pub/infphilo/centrifuge/data/p+h+v.tar.gz, with wget.

- Do this exercise for this tutorial. Build the Centrifuge database indexes with the centrifuge-build command. This is to practice building an index.

- command flags:

  -p The number of processes/threads to use for creating the index.

  --conversion-table The seqid2taxid.map file that you created with the centrifuge-download commands.

  --taxonomy-tree The nodes.dmp file that was downloaded with the first centrifuge-download command.

  --name-table The nodes.dmp file that was downloaded with the first centrifuge-download command.

```
# Each run will use about 24 GB of RAM.
# It took 39 minutes to run during testing.

centrifuge-build -p 4 --conversion-table /home/<username>/centrifuge/seqid2taxid.map --taxonomy-tree \
/home/<username>/centrifuge/taxonomy/nodes.dmp --name-table \
/home/<username>/centrifuge/taxonomy/names.dmp \
/home/<username>/centrifuge/library/fungi_HumanFixed.fna \
/home/<username>/centrifuge/fungi_Human_indeces
```

Classify Reads with Centrifuge

In this portion of the tutorial we will use the pre-built indexes that contain complete RefSeq genomes for archea, bacteria, viruses and fungi. The human chromosome level reference genome was included and we also added the EuPathDB database as well. Each file will use about 36 GB of RAM. So obviously, not every one can run samples at the same time so coordinate with one another.

- Classify reads with Centrifuge.
- command flags:

  -x Path to the indexes

  - Index location /home/metag/centrifuge/complete_genomes/Arc_Bac_Vir_Hum_Eupath_v2

  -1 Read 1 fastq file

  -2 Read 2 fastq file

  -t Print wall-clock time taken by search phase. This is optional.

  -p Number of alignment threads to launch

  --met-file Send metrics to file at <path>

  --met-stderr Send metrics to stderr

  -S Output file name

```
# Make a directory to store your results:

mkdir -p /home/<username>/centrifuge/centrifuge_results

# Run Centrifuge. It will take 10 - 20 minutes per sample to run.

centrifuge -x /home/condasw/db/centrifuge/complete_genomes/Arc_Bac_Vir_Hum_Eupath_v2 \
-1 /home/condasw/data/Metagenomics/centrifuge/ERR2271042_1.fastq \
-2 /home/condasw/data/Metagenomics/centrifuge/ERR2271042_2.fastq -t \
-p 4 --met-file /home/<username>/centrifuge/centrifuge_results/ERR2271042_meta.txt --met-stderr \
-S /home/<username>/centrifuge/centrifuge_results/ERR2271042_cent.out
```

- If you have many samples to analyze, use a for loop to run centrifuge on each sample sequentially.

```
# Use a for loop to store the names of the fastq files in a variable (like a list)
# named fastq using the basename command

fastq=`for i in /home/condasw/data/Metagenomics/centrifuge/*_1.fastq; do basename -s _1.fastq $i;done`

# You can view the contents of the fastq variable with the echo command

echo $fastq

# Use a second for loop to read through the variable list and run centrifuge on each file in the list
# sequentially.

for j in $fastq; do centrifuge -x /home/condasw/db/centrifuge/complete_genomes/Arc_Bac_Vir_Hum_Eupath_v2 \
-1 /home/condasw/data/Metagenomics/centrifuge/${j}_1.fastq \
-2 /home/condasw/data/Metagenomics/centrifuge/${j}_2.fastq -t -p 4 \
--met-file /home/<username>/centrifuge/centrifuge_results/${j}_meta.txt --met-stderr \
-S /home/<username>/centrifuge/centrifuge_results/${j}_cent.out; done
```

Convert the Centrifuge Output to a Kraken Style Report

- Create a Kraken style report from the Centrifuge output using the centrifuge-kreport command. This command used about 17 GB of memory and took about 1 minute per file to complete during testing.
- command flags:

  -x Path to the indexes.

```
# Make a directory to store your output files:

mkdir -p /home/<username>/centrifuge/centrifuge_krn_results

centrifuge-kreport -x /home/condasw/db/centrifuge/complete_genomes/Arc_Bac_Vir_Hum_Eupath_v2 \
/home/<username>/centrifuge/centrifuge_results/ERR2271042_cent.out \
> /home/<username>/centrifuge/centrifuge_krn_results/ERR2271042_cent.out.krn
```

- Use a for loop if you have many samples to analyze.

```
# Use a for loop to store the names of the fastq files in a variable (like a list)
# named output using the basename command

output=`for i in /home/metag/centrifuge/centrifuge_results/*_cent.out; do basename -s _cent.out $i;done`

# You can view the contents of the fastq variable with the echo command

echo $output

# Use a second for loop to read through the variable list and run centrifuge on each file in the list
# sequentially.

for j in $output; do centrifuge-kreport \
-x /home/condasw/db/centrifuge/complete_genomes/Arc_Bac_Vir_Hum_Eupath_v2 \
/home/<username>/centrifuge/centrifuge_results/${j}_cent.out > \
/home/<username>/centrifuge/centrifuge_krn_results/${j}.krn; done
```

Parse the Metadata File to Find Out which Samples Contain STEC

- Use awk to extract columns 2, 10, and 18 from the metadata file located at /home/metag/centrifuge/data/PRJEB23207_metadata.txt . Then, pipe the output to grep and exclude samples that were sequenced with the “Ion Torrent PGM” platform

```
# Copy the /home/condasw/data/Metagenomics/centrifuge/PRJEB23207_metadata.txt to your working directory

cp /home/condasw/data/Metagenomics/centrifuge/PRJEB23207_metadata.txt \
/home/<username>/centrifuge

# Use awk and grep to parse the metadata file
# -F is the field separator flag. In this case the metadata file is tab-separated '\t'.
# -i case-insensitive, -v exclude the search string, i.e., exclude lines that contain "ion"

awk -F'\t' '{print $2,$10,$18}' PRJEB23207_metadata.txt | grep -i -v "ion"
```

Visualize the Kraken Reports with Pavian

- First, download all *.krn files to your computer using secure copy (scp) with unix, linux, or MobaXterm terminals.

```
# Download the kraken report to you Documents directory
scp -P 44111 <username>@gateway.training.ncgr.org:\
/home/condasw/data/Metagenomics/centrifuge/centrifuge_krn_results/*.krn Documents/
```

- Install Pavian on you computers by following the instruction below.
- This assumes that you have installed R on your computer.

```
# Run R by opening a terminal

R

# Or click on your R application from R-studio or R-console.
```

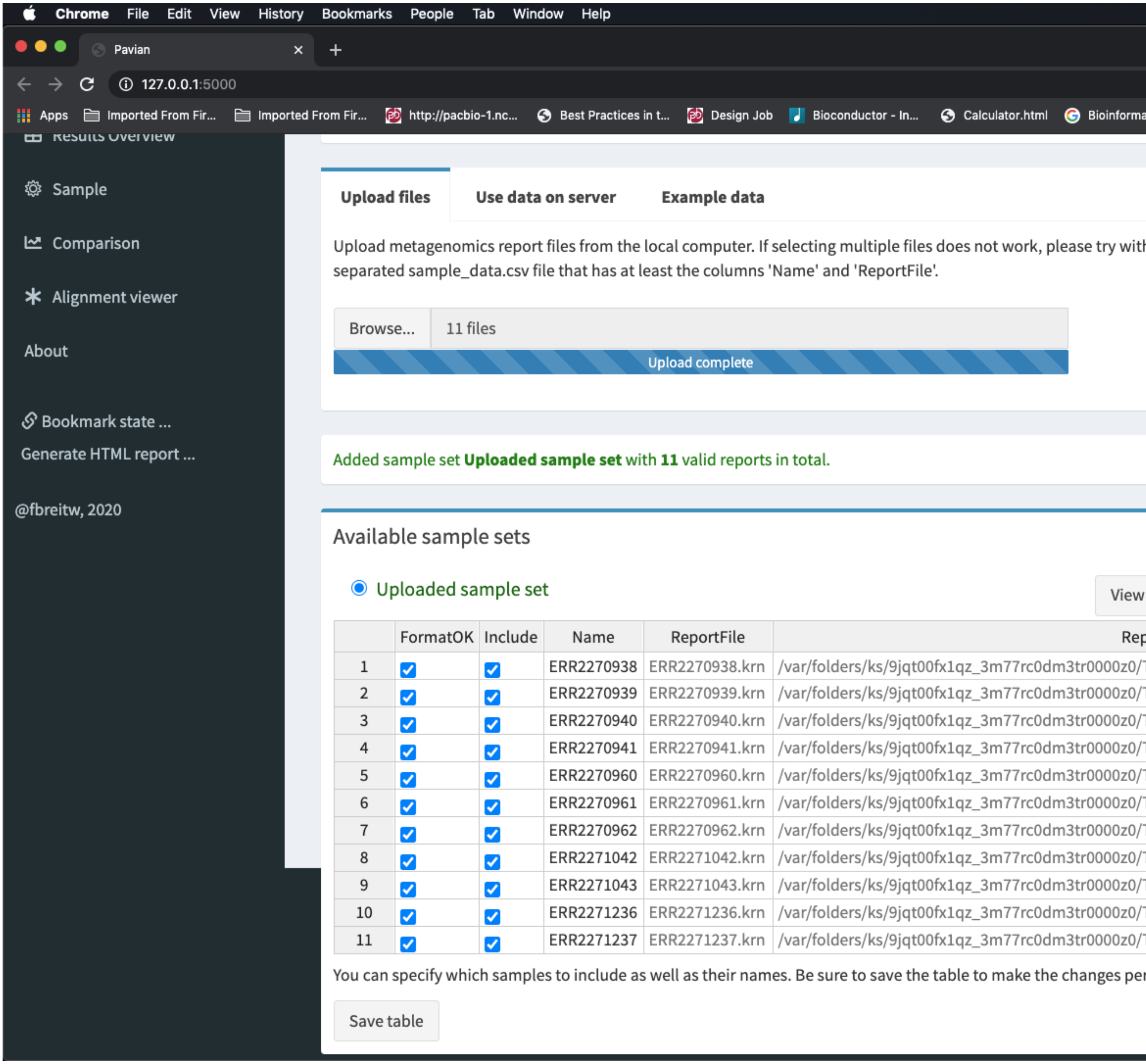
- Install Pavian if you haven’t done so.

```
if (!require(remotes)) { install.packages("remotes") }

## Loading required package:  remotes

remotes::install_github("fbreitwieser/pavian")

## Skipping install of 'pavian' from a github remote, the SHA1 (81d784d8) has not changed since last install.
##  Use `force = TRUE` to force installation
```

- Start Pavian in your browser from R

```
pavian::runApp(port=5000)
```

- A web browser should pop up. If not, Pavian can be accessed by entering this url into your browser. http://127.0.0.1:5000
- Drag the .krn files to the "Browse" bar under the "Upload files" tab.



Upload metagenomics report files from the local computer. If selecting multiple files does not work, please try with separated sample_data.csv file that has at least the columns 'Name' and 'ReportFile'.

Added sample set **Uploaded sample set** with **11** valid reports in total.

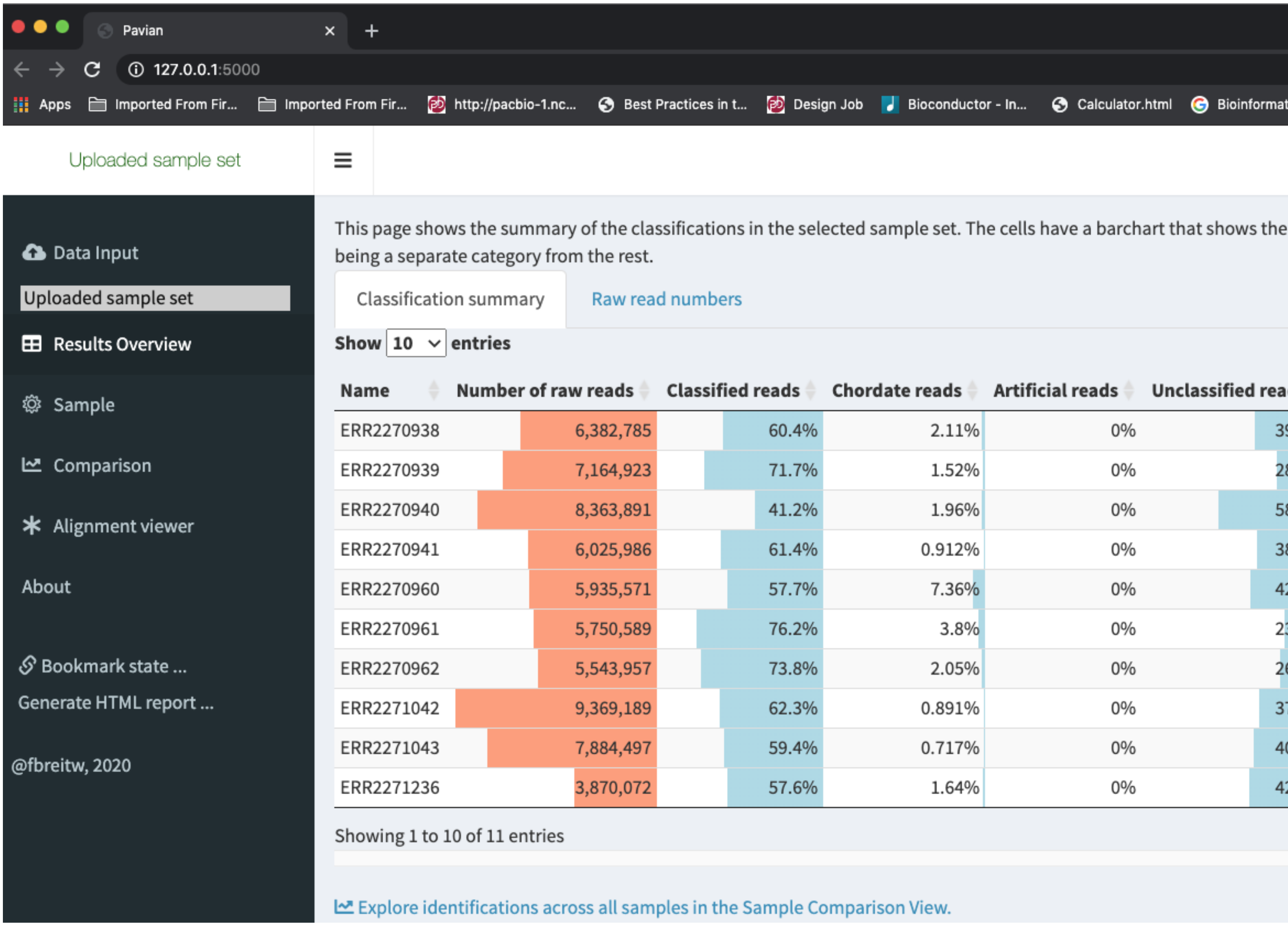
Available sample sets

● Uploaded sample set

	FormatOK	Include	Name	ReportFile	Rep
1	☑	☑	ERR2270938	ERR2270938.krn	/var/folders/ks/9jqt00fx1qz_3m77rc0dm3tr0000z0/
2	☑	☑	ERR2270939	ERR2270939.krn	/var/folders/ks/9jqt00fx1qz_3m77rc0dm3tr0000z0/
3	☑	☑	ERR2270940	ERR2270940.krn	/var/folders/ks/9jqt00fx1qz_3m77rc0dm3tr0000z0/
4	☑	☑	ERR2270941	ERR2270941.krn	/var/folders/ks/9jqt00fx1qz_3m77rc0dm3tr0000z0/
5	☑	☑	ERR2270960	ERR2270960.krn	/var/folders/ks/9jqt00fx1qz_3m77rc0dm3tr0000z0/
6	☑	☑	ERR2270961	ERR2270961.krn	/var/folders/ks/9jqt00fx1qz_3m77rc0dm3tr0000z0/
7	☑	☑	ERR2270962	ERR2270962.krn	/var/folders/ks/9jqt00fx1qz_3m77rc0dm3tr0000z0/
8	☑	☑	ERR2271042	ERR2271042.krn	/var/folders/ks/9jqt00fx1qz_3m77rc0dm3tr0000z0/
9	☑	☑	ERR2271043	ERR2271043.krn	/var/folders/ks/9jqt00fx1qz_3m77rc0dm3tr0000z0/
10	☑	☑	ERR2271236	ERR2271236.krn	/var/folders/ks/9jqt00fx1qz_3m77rc0dm3tr0000z0/
11	☑	☑	ERR2271237	ERR2271237.krn	/var/folders/ks/9jqt00fx1qz_3m77rc0dm3tr0000z0/

You can specify which samples to include as well as their names. Be sure to save the table to make the changes per

- Click on the “Results Overview” tab on the left side of the screen.



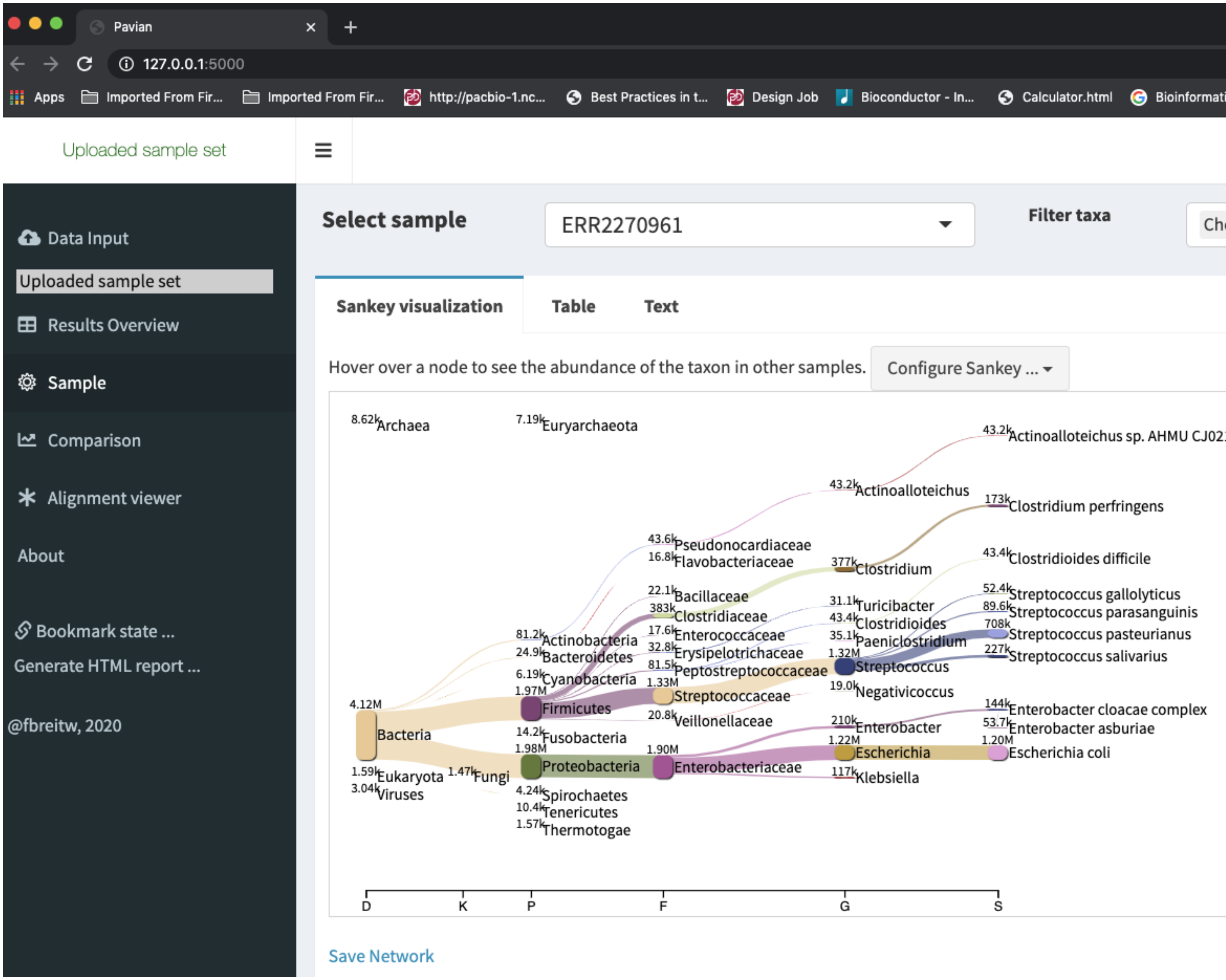
This page shows the summary of the classifications in the selected sample set. The cells have a barchart that shows the being a separate category from the rest.

Name	Number of raw reads	Classified reads	Chordate reads	Artificial reads	Unclassified rea
ERR2270938	6,382,785	60.4%	2.11%	0%	39
ERR2270939	7,164,923	71.7%	1.52%	0%	28
ERR2270940	8,363,891	41.2%	1.96%	0%	58
ERR2270941	6,025,986	61.4%	0.912%	0%	38
ERR2270960	5,935,571	57.7%	7.36%	0%	42
ERR2270961	5,750,589	76.2%	3.8%	0%	23
ERR2270962	5,543,957	73.8%	2.05%	0%	26
ERR2271042	9,369,189	62.3%	0.891%	0%	37
ERR2271043	7,884,497	59.4%	0.717%	0%	40
ERR2271236	3,870,072	57.6%	1.64%	0%	42

Showing 1 to 10 of 11 entries

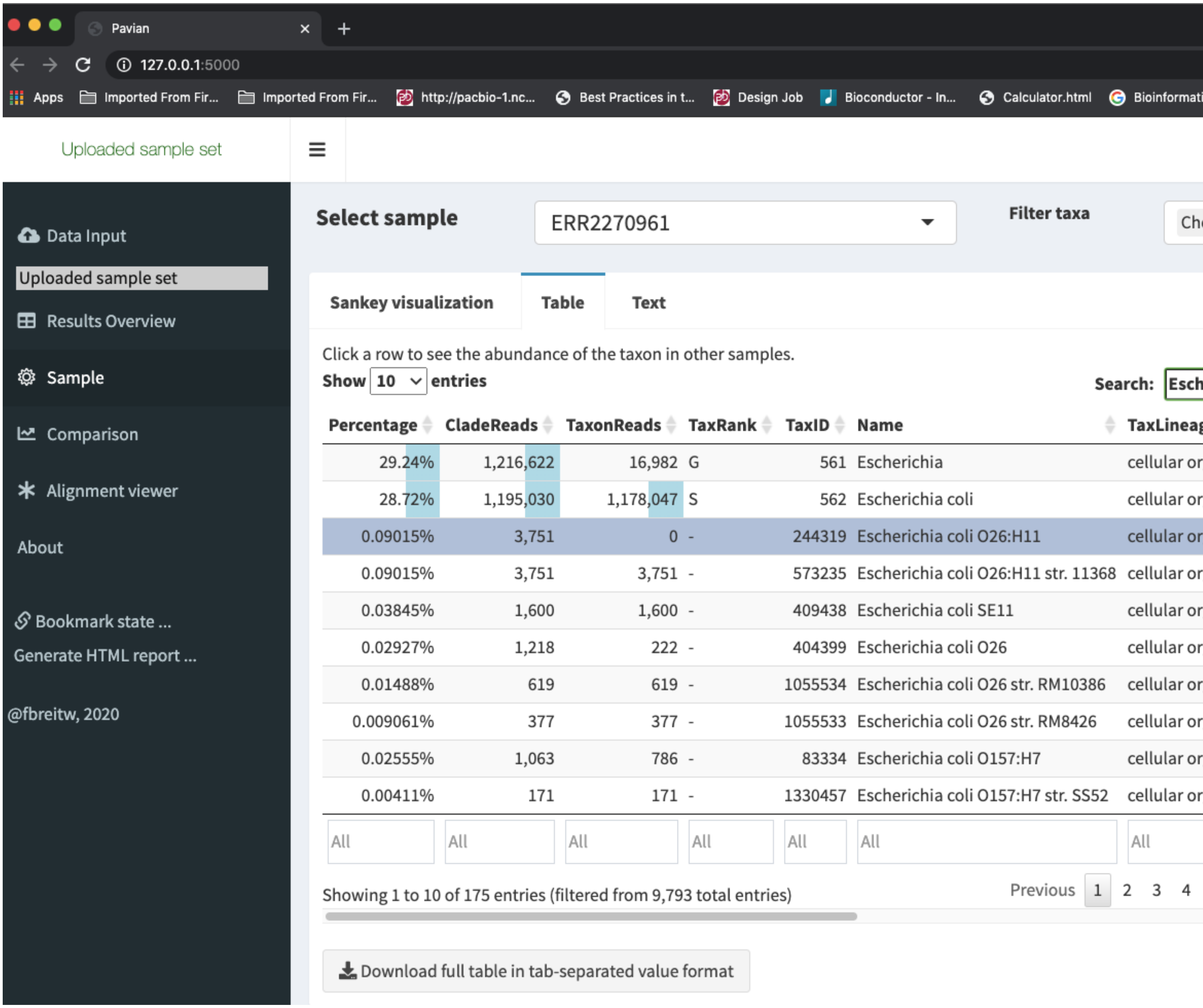
Explore identifications across all samples in the Sample Comparison View.

- Click on the "Sample" tab on the left side of the screen. You will notice an interactive Sankey chart that displays different classified taxa. You can click and drag nodes to reorder the nodes if you want. You can select different samples by clicking on the down arrow next to the "Select sample" header. Select sample ERR2270960 or ERR2270961 since we know that these two samples are likely infected with STEC.



Select sample: ERR2270961

- From the "Sample" page, click on the "Table" tab. You should see an interactive, filterable and searchable table that summarizes the results of Centrifuge. For example, type "Escherichia coli" into the "Search:" box. A histogram will appear that displays the numbers of reads across all samples that were classified as "Escherichia coli". Sort the table by the most abundant taxa by clicking on the up arrow next to "TaxonReads".



Select sample: ERR2270961

Filter taxa: Ch

Click a row to see the abundance of the taxon in other samples.

Show 10 entries

Search: Esch

Percentage	CladeReads	TaxonReads	TaxRank	TaxID	Name	TaxLineag
29.24%	1,216,622	16,982	G	561	Escherichia	cellular or
28.72%	1,195,030	1,178,047	S	562	Escherichia coli	cellular or
0.09015%	3,751	0	-	244319	Escherichia coli O26:H11	cellular or
0.09015%	3,751	3,751	-	573235	Escherichia coli O26:H11 str. 11368	cellular or
0.03845%	1,600	1,600	-	409438	Escherichia coli SE11	cellular or
0.02927%	1,218	222	-	404399	Escherichia coli O26	cellular or
0.01488%	619	619	-	1055534	Escherichia coli O26 str. RM10386	cellular or
0.009061%	377	377	-	1055533	Escherichia coli O26 str. RM8426	cellular or
0.02555%	1,063	786	-	83334	Escherichia coli O157:H7	cellular or
0.00411%	171	171	-	1330457	Escherichia coli O157:H7 str. SS52	cellular or
All	All	All	All	All	All	All

Showing 1 to 10 of 175 entries (filtered from 9,793 total entries)

- Can you can you identify the strain(s) of STEC that samples ERR2270960 and ERR2270961 likely contain?

References

Gigliucci F, von Meijenfeldt FAB, Knijn A, Michelacci V, Scavia G, Minelli F, Dutilh BE, Ahmad HM, Raangs GC, Friedrich AW, Rossen JWA, Morabito S. Metagenomic Characterization of the Human Intestinal Microbiota in Fecal Samples from STEC-Infected Patients. Front Cell Infect Microbiol. 2018 Feb 6;8:25. doi: 10.3389/fcimb.2018.00025. eCollection 2018. PubMed PMID: 29468143; PubMed Central PMCID: PMC5808120.

Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016 Dec;26(12):1721-1729. Epub 2016 Oct 17. PubMed PMID: 27852649; PubMed Central PMCID: PMC5131823.

Lu J, Salzberg SL. Removing contaminants from databases of draft genomes. PLoS Comput Biol. 2018 Jun 25;14(6):e1006277. doi: 10.1371/journal.pcbi.1006277. eCollection 2018 Jun. PubMed PMID: 29939994; PubMed Central PMCID: PMC6034898.