Taxonomic Classification with Kraken

1. Introduction:

For this current lab which is based on virome capture metagenomics data captured in the lab, we performed **taxonomic classification** using **Kraken2**. Kraken2 is a tool used for the analysis and metagenomic sequence classification. This was introduced first in a paper in Genome Biology by Wood and Salzberg (2014). To find the lowest common ancestor taxa in each read, it compares read-originating k-mers against a database of reference genomes. This allows the precise composition of complicated samples to be presented.

The **data** we are using for the current lab were collected from **clinical and wastewater sources**. The samples were **sequenced** using Illumina's NextSeq 2000 instrument. The data is located on the HPC cluster at the location: /projects/class/binf6203_001/ViromeLab-2023/

2. Methodology:

2.1 Data:

The data was copied from the above-mentioned location to the working directory along with the Kraken2 database. Once these files were transferred, they were unzipped.

2.2 Kraken2 script

1)

```
#!/bin/bash
#SBATCH --partition=Centaurus
#SBATCH --job-name=kraken
#SBATCH --time=72:00:00
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=16
#SBATCH --mem=64GB

module load kraken2
module load blast/2.11.0+
for i in `ls /users/dmehta12/UNK1_S11_L001_R1_001.fastq
UNK1_S11_L001_R2_001.fastq | grep R1`
do
   name=`basename $i | cut -d_ -f1`
   kraken2 \
```

```
    --db /users/dmehta12/k2_viral_db \
    --paired \
    --use-mpa-style \
    --classified-out classified_${classifiedout}_R#.fastq \
    --unclassified-out unclassified_${unclassifiedout}_R#.fastq \
    --output ${Output1}_kraken11.txt \
    --report ${output2}_kraken11.report \
    $i ${i/R1/R2}
done
```

2)

```
#!/bin/bash
#SBATCH --partition=Centaurus
#SBATCH --job-name=kraken
#SBATCH --time=72:00:00
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=16
#SBATCH --mem=64GB
module load kraken2
module load blast/2.11.0+
for i in `ls /users/dmehta12/UNK2_S1_L001_R1_001.fastq UNK2_S1_L001_R2_001.fastq |
grep R1`
do
    name=`basename $i | cut -d_ -f1`
    kraken2 \
    --db /users/dmehta12/k2_viral_db \
    --paired \
    --use-mpa-style \
    --classified-out classified_${classifiedout}_R#.fastq \
    --unclassified-out unclassified_${unclassifiedout}_R#.fastq \
    --output ${Output1}_krakenS1.txt \
    --report ${output2}_krakenS1.report \
    $i ${i/R1/R2}
done
```

3)

```
#!/bin/bash
#SBATCH --partition=Centaurus
```

```
#SBATCH --job-name=kraken
#SBATCH --time=72:00:00
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=16
#SBATCH --mem=64GB
module load kraken2
module load blast/2.11.0+
for i in `ls /users/dmehta12/UNK3_S5_L001_R1_001.fastq UNK3_S5_L001_R2_001.fastq | grep R1`
do
   name=`basename $i | cut -d_ -f1`
   kraken2 \
   --db /users/dmehta12/k2_viral_db \
   --paired \
   --use-mpa-style \
   --classified-out classifiedS5_${classifiedout}_R#.fastq \
   --unclassified-out unclassifiedS5_${unclassifiedout}_R#.fastq \
   --output ${Output1}_krakenS5.txt \
   --report ${output2}_krakenS5.report \
   $i ${i/R1/R2}
done
```

2.3 Transforming Kraken2 reports into Krona-ready text using metaphlan2krona.py script.

```
module load anaconda3
python /users/dmehta12/useful_scripts/metaphlan2krona.py -p _krakenS5.report -k _kronaS5.txt
python /users/dmehta12/useful_scripts/metaphlan2krona.py -p _krakenS1.report -k _kronaS1.txt
python /users/dmehta12/useful_scripts/metaphlan2krona.py -p _kraken11.report -k _kronaS11.txt
```

2.4 Generating HTML files:

```
module load anaconda3
conda create -n KronaTools
conda activate KronaTools
conda install -c bioconda krona
```

ktImportText _kronaS1.txt -o kronaS1.html
ktImportText _kronaS5.txt -o kronaS5.html
ktImportText _kronaS11.txt -o kronas11.html

3. Results:

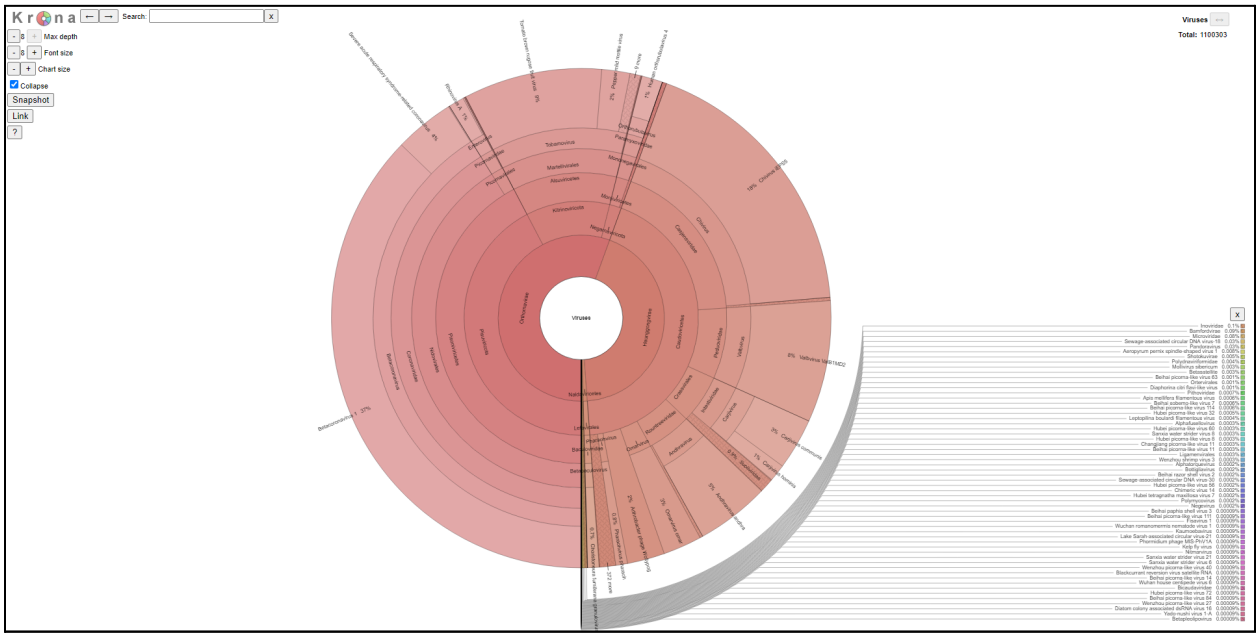

Fig: kronas11.html > Sample S11

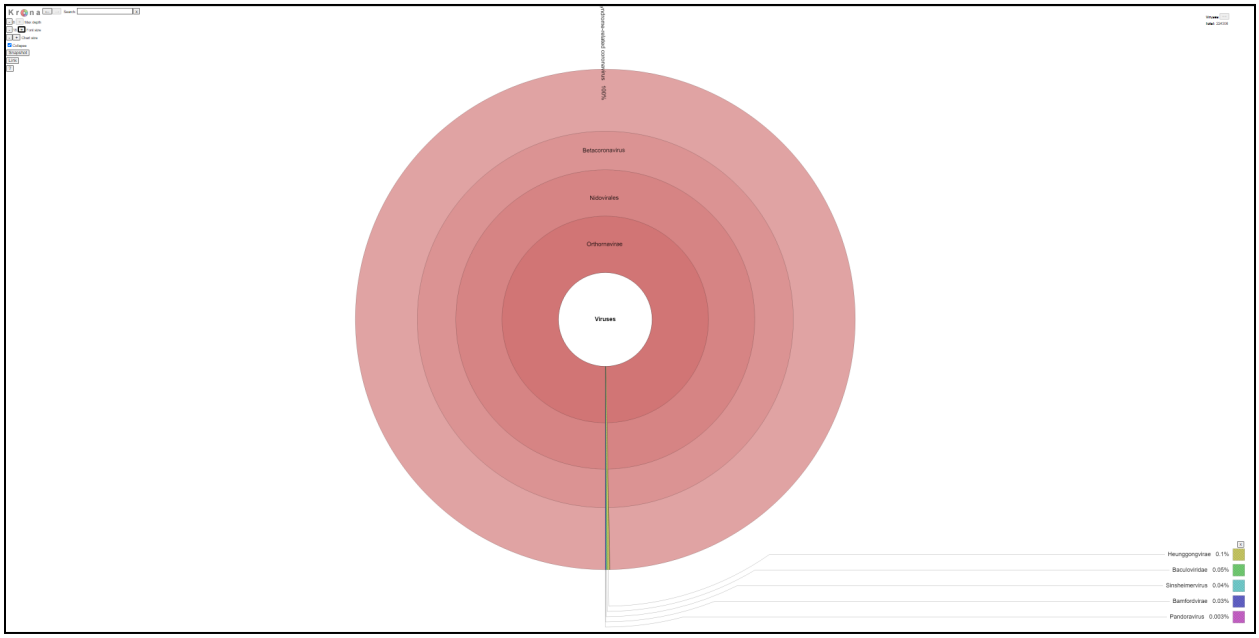

Fig: kronas5.html > Sample S5

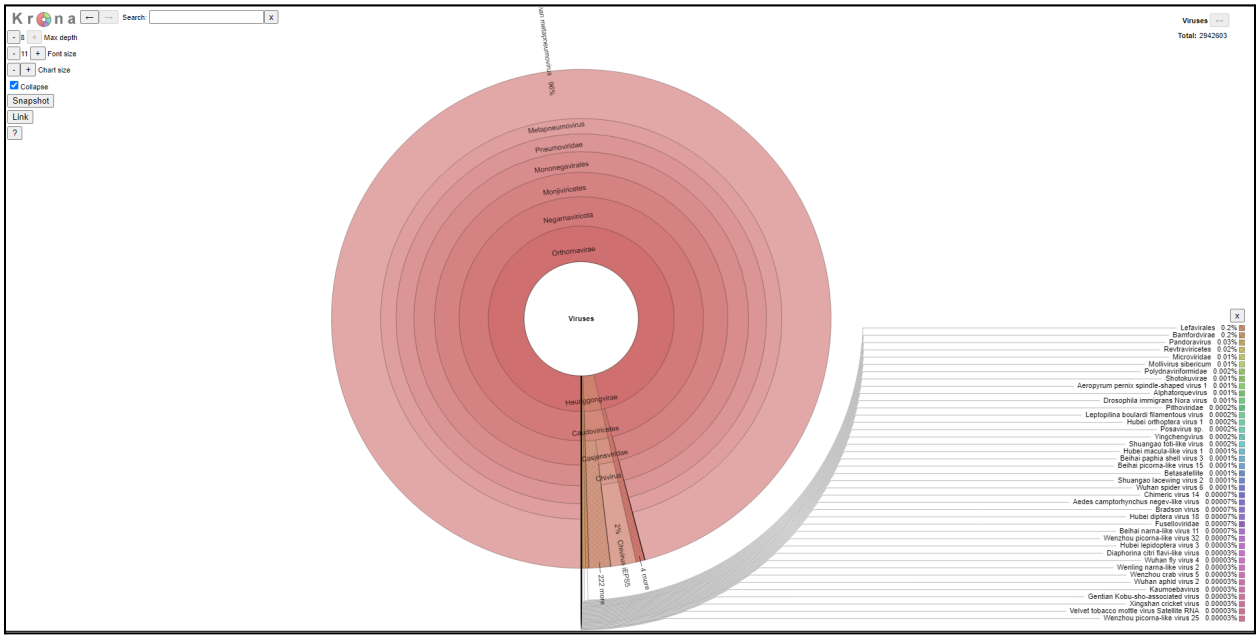

Fig: kronas1.html > Sample S1

4. Discussion:

When you see the above images, you can understand that the depth for the taxonomic classification is different for all the 3 HTML files. The kronas5 has the least depth and shows an overview of the classification of the viruses

Sample S11 was the **wastewater sample** because it contained a variety of viruses like Chivirus, Rhinovirus, Sewage-associated circular DNA virus, … etc. **Sample S1** was the **clinical "negative"** sample. We can say this because Orthornavirae is around 96% of the total and Heuggongvirae is around 3%. They state to be respiratory system-associated viruses. And **sample S5** was the **clinical "positive" sample**. This contained coronaviridae.