

# ***Kraken on HPC***

## **What is Kraken?**

Kraken is a program that is used to classify DNA sequences from metagenomic data. It is a fast and accurate tool that is widely used in metagenomic studies to identify and quantify the microorganisms present in a sample. To use Kraken, you provide it with a set of DNA sequences and a database of known sequences, and it uses a set of rules to identify the taxonomic classification of each sequence based on its similarity to the known sequences. The output of Kraken is a list of the taxonomic classifications of the input sequences, along with the confidence of each classification. Kraken is widely used in metagenomic studies because it is fast and accurate, and because it can handle very large datasets.

Links:

[Official Website](#)

[Manual](#)

## **Versions Available:**

The following versions are available on the cluster:

- Kraken 0.10.5

## **How to load Kraken?**

To load Kraken, use the following commands:

```
#Load the KRAKEN module  
module load bio/kraken/0.10.5
```

To verify if the module and dependencies are loaded correctly, use the following command.

```
#Show all the modules loaded
module list
```

This should list all the KRAKEN dependencies that are loaded.

## How to use Kraken?

The command line arguments for kraken are,

```
Need to specify input filenames!
Usage: kraken [options] <filename(s)>

Options:
  --db NAME                Name for Kraken DB
                           (default: none)
  --threads NUM            Number of threads (default: 1)
  --fasta-input            Input is FASTA format
  --fastq-input            Input is FASTQ format
  --gzip-compressed        Input is gzip compressed
  --bzip2-compressed        Input is bzip2 compressed
  --quick                  Quick operation (use first hit or hits)
  --min-hits NUM           In quick op., number of hits req'd for
classification
                           NOTE: this is ignored if --quick is not
specified
  --unclassified-out FILENAME
                           Print unclassified sequences to filename
  --classified-out FILENAME
                           Print classified sequences to filename
  --output FILENAME        Print output to filename (default: stdout);
  "-" will                  suppress normal output
  --only-classified-output
                           Print no Kraken output for unclassified
sequences
  --preload                Loads DB into memory before classification
  --paired                  The two filenames provided are paired-end
reads
```

```
--check-names          Ensure each pair of reads have names that
agree
                        with each other; ignored if --paired is not
specified
--help                 Print this message
--version              Print version information
```

If none of the \*-input or \*-compressed flags are specified, and the file is a regular file, automatic format detection is attempted.

A sample slurm script to build a database using kraken is given below,

```
#!/bin/bash
#SBATCH -J kraken_database_build
#SBATCH -p main
#SBATCH --qos main
#SBATCH --ntasks=4
#SBATCH --cpus-per-task=4
#SBATCH --time=4:00:00
#SBATCH --mem=20GB

#Load module
module load bio/kraken/0.10.5

# set the number of threads to use
export OMP_NUM_THREADS=$SLURM_CPUS_PER_TASK

# SEE DOCS FOR MORE INFO ON DATABASE
#krakenDB="Choose a database"

# build the database
kraken-build --threads $OMP_NUM_THREADS --db $krakenDB
```

See manual for more information on Kraken and it's usage.

### ***Where to find help?***

If you are confused or need help at any point, please contact OIT at the following address.

<https://ua-app01.ua.edu/researchComputingPortal/public/oitHelp>