SRA HUMAN SCRUBBER

This resource was made possible through funding provided under the Epidemiology and Laboratory Capacity for Prevention and Control of Emerging Infectious Diseases (ELC) Cooperative Agreement (CK24-0002), Project D: Advanced Molecular Detection to the Florida Department of Health. The conclusions, findings, and opinions expressed by authors do not necessarily reflect the official position of the U.S. Department of Health and Human Services, the Public Health Service, or the Centers for Disease Control and Prevention.

OVERVIEW

Purpose

- Identify and remove or mask human-origin reads from FASTQ files

Usage

- Used by public health labs and researchers to clean sequencing data by removing human-derived reads, helping ensure privacy compliance before sharing or analyzing pathogen genomes to study outbreaks and viral evolution.

Dependencies

- Unix Tools, Curl and Pre-Built k-mer reference database (init_db.sh)

Advanced Molecular Detection
Southeast Region Bioinformatics

REVIEW OF K-MERS

- Short, fixed-length subsequences of DNA sequences
  - “K” refers to the length of the sequence (Ex. 3-mer, 4-mer, 32-mer
- Human Scrubber uses a 32-mer
  - Human scrubber uses a reference k-mer database to map query reads to known pathogen genomes

Advanced Molecular Detection
Southeast Region Bioinformatics

WORKFLOW

Single-End fastq file

↓

Run the scrub.sh on HPG

↓

Scrubber_test_fastq.clean

Advanced Molecular Detection
Southeast Region Bioinformatics

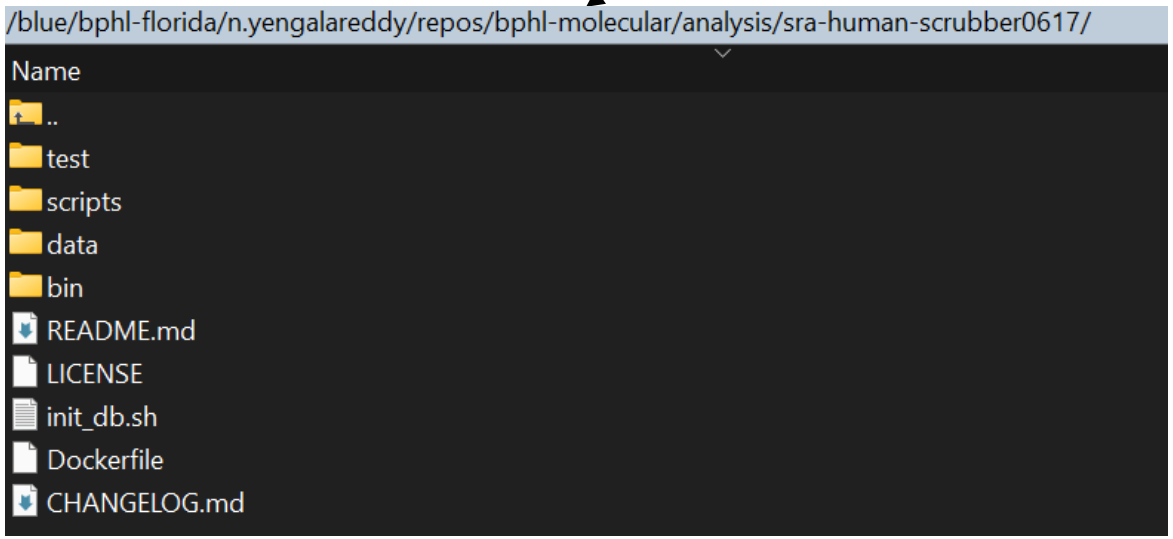
APPLICATION

Objective:

Use a single end fastq file and make a clean dataset using Human Scrubber

APPLICATION CONT.

```
cd /blue/bphl-<state>/<user>/repos/bphl-molecular/
git clone https://github.com/ncbi/sra-human-scrubber
mkdir analysis/
cd analysis/
cp /blue/bphl-<state>/<user>/repos/bphl-molecular/human-scrubber/*
```



Advanced Molecular Detection
Southeast Region Bioinformatics

APPLICATION CONT.

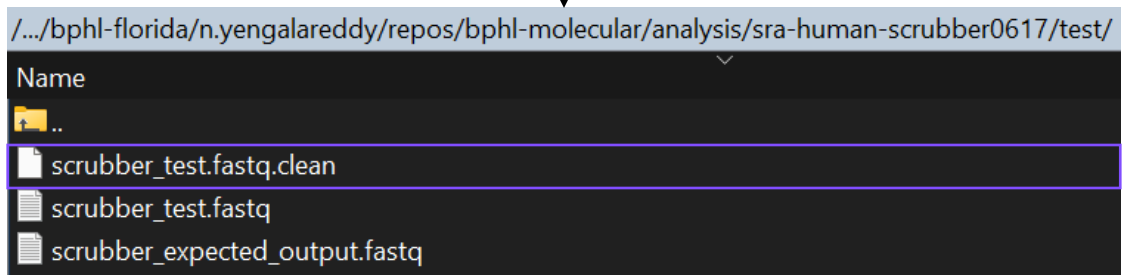
```
[n.yengalareddy@login8 sra-human-scrubber0617]$ ./scripts/scrub.sh test
2025-06-19 22:37:11     aligns_to version 0.801
2025-06-19 22:37:11     hardware threads: 64, omp threads: 64
2025-06-19 22:37:17     loading time (sec) 6
2025-06-19 22:37:17     /tmp/tmp.bjRmMbNeAI/temp.fasta
2025-06-19 22:37:17     FastaReader
2025-06-19 22:37:17     100% processed
2025-06-19 22:37:17     total spot count: 2
2025-06-19 22:37:17     total read count: 2
2025-06-19 22:37:17     total time (sec) 6
1  spot(s) masked or removed.
DB version is 20250325v2

test succeeded
```

Advanced Molecular Detection
Southeast Region Bioinformatics

APPLICATION CONT.

```
./scripts/scrub.sh -p 8 /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/sra-human-scrubber0617/test/scrubber_test.fastq
```



Advanced Molecular Detection
Southeast Region Bioinformatics

APPLICATION CONT.

/.../bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/sra-human-scrubber0617/test/

Name

..

scrubber_test.fastq.clean

scrubber_test.fastq

scrubber_expected_output.fastq

```
@test.2 2 length=286
GGCAATGCATTTACAGTACAAAAGACATACTGTTCTAATGTTGAATTCACTTTGAATTTATCAAAACACTCTACACGAGCACGTGCAGGTATAATTCT
+test.2 2 length=286
A?AA1C1@1BFFDGB1GGGGGGFHCGBGHHH3DB3F33FHFGHGFHGHBFHHHHGHHHHHHHHEGBEFGHHHHHHE?/EEAE/B/CEFGFGHDFFGHH
@test.216524 216524 length=194
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+test.216524 216524 length=194
BBBBBFBFFFFFGGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHA
```

Advanced Molecular Detection
Southeast Region Bioinformatics

PULL HUMAN SCRUBBER-DOCKER

- DockerHub
  - https://hub.docker.com/r/ncbi/sra-human-scrubber
  - docker pull ncbi/sra-human-scrubber

Here the command is given the path to your local fastq file as argument `docker run -it -v $PWD:$PWD:rw -w $PWD ncbi/sra-human-scrubber:latest /opt/scrubber/scripts/scrub.sh path-to-fastq-file/filename.fastq`

Example: `docker run -it -v $PWD:$PWD:rw -w $PWD ncbi/sra-human-scrubber:latest /opt/scrubber/scripts/scrub.sh MyFastqFile.fastq`

```
2022-09-06 21:35:04    aligns_to version 0.707
2022-09-06 21:35:04    hardware threads: 8, omp threads: 8
2022-09-06 21:35:04    loading time (sec) 0
2022-09-06 21:35:04    /tmp/tmp.Ccqruccyoq/temp.fasta
2022-09-06 21:35:04    FastaReader
2022-09-06 21:35:04    0% processed
2022-09-06 21:35:06    100% processed
2022-09-06 21:35:06    total spot count: 216859
2022-09-06 21:35:06    total read count: 216859
2022-09-06 21:35:06    total time (sec) 2
129  spot(s) masked or removed.
```

Advanced Molecular Detection
Southeast Region Bioinformatics

PULL HUMAN SCRUBBER-BIOCONDA

- Bioconda
  - https://anaconda.org/bioconda/sra-human-scrubber
  - conda install bioconda::sra-human-scrubber
- On HPG:
  - module load sra_human_scrubber
  - Run *module spider sra-human-scrubber* to see what environment modules are available for sra human scrubber
- Note: SRA Human Scrubber is included in Bactopia

Advanced Molecular Detection
Southeast Region Bioinformatics

APPLYING HUMAN SCRUBBER

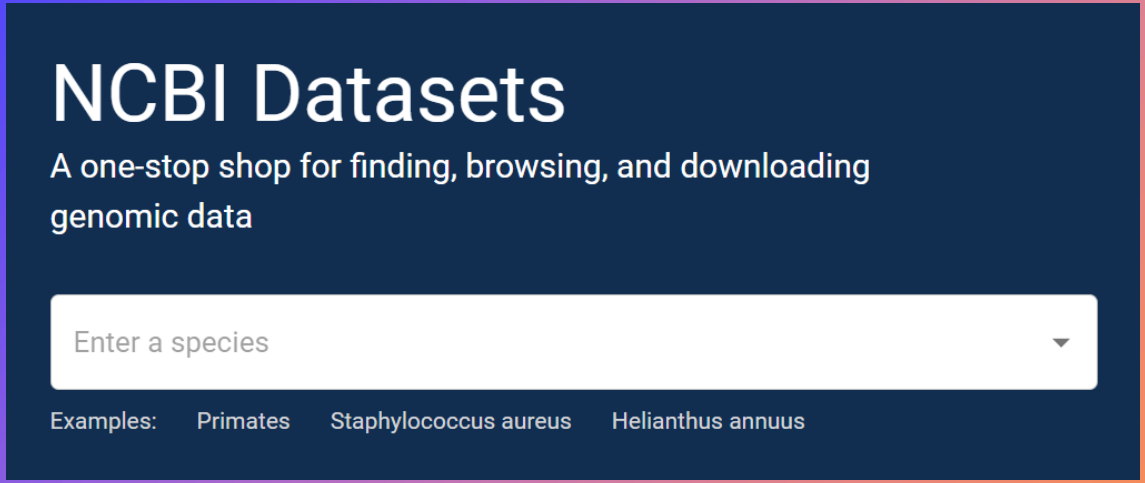
- To retroactively apply SRA Human Scrubber to your SRA submissions, email the SRA Help Desk
  - sra@ncbi.nlm.nih.gov
  - Request HRRT be activated for your BioProject
    - Include your BioProject Number
    - Depending on the number of samples it'll take about a week for Human Scrubber to be applies.
- Will also be applied to future submissions
  - Better to do it in-house before SRA submission to protect possible PHI breach

Advanced Molecular Detection
Southeast Region Bioinformatics

NCBI DATASETS

NCBI DATASETS OVERVIEW

- NCBI has data sets for almost anything you could ever want
- Taxonomy, gene, and genome level
  - Special data set for viruses
- Can access in 3 ways
  - CLI
  - GitHub
  - NCBI Website
- Excellent How-To Guides on their website
  - https://www.ncbi.nlm.nih.gov/datasets/docs/v2/how-tos/



Advanced Molecular Detection
Southeast Region Bioinformatics

NCBI DATASETS GUI

Bacteria / Pseudomonadota / Gammaproteobacteria / Pseudomonadales / Pseudomonadaceae /

Pseudomonas aeruginosa ☆

Pseudomonas aeruginosa is a species of g-proteobacteria in the family Pseudomonadaceae.

Browse taxonomy

NCBI Taxonomy ID	287
Taxonomic rank	species
Current scientific name	Pseudomonas aeruginosa (Schroeter 1872) Migula 1900 (Approved Lists 1980) NOMEN APPROBBATUM Type Material
Basionym	"Bacterium aeruginosum" Schroeter 1872

View taxonomic details

Genome

Browse all 31,553 genomes

Database links

Nucleotide	
All nucleotide sequences	3,148,747
Genomic sequences	3,148,394
mRNA sequences	100

Protein	
Protein sequences	24,373,600
Conserved domains	9
3D structures	2,943

GEO Datasets	
Datasets	24
Series	496
Samples	6,728
Platforms	103

Sequence Read Archive (SRA)	
All SRA experiments	62,293
DNA	55,759
RNA	6,350

PopSet	
Phylogenetic studies	468
Population studies	241

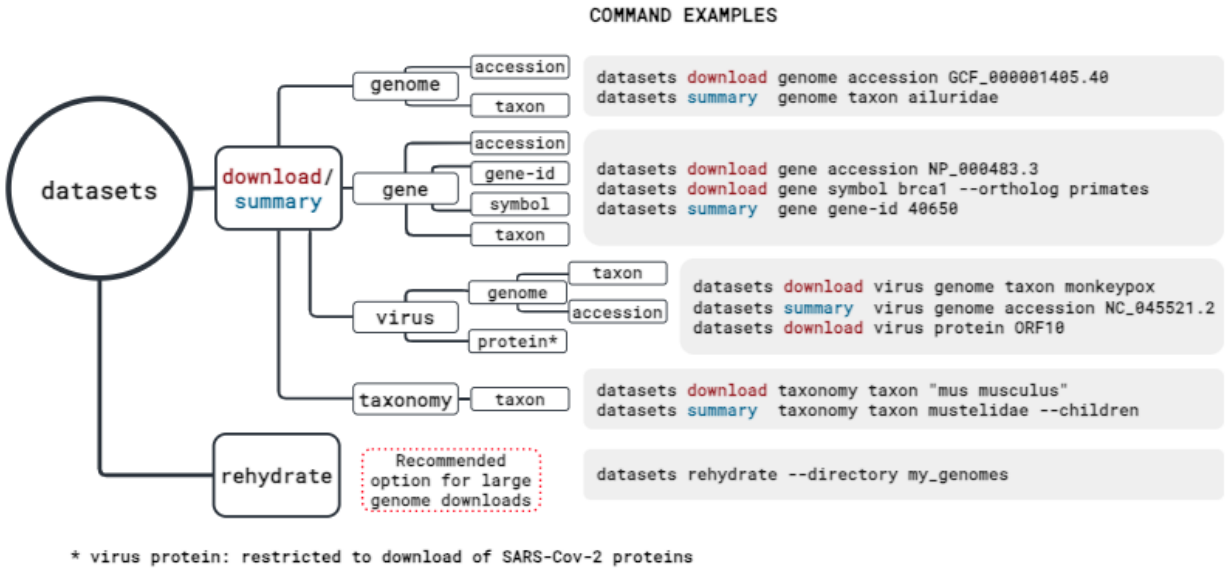
Projects and samples	
BioProject	3,277
BioSample	75,992

https://www.ncbi.nlm.nih.gov/datasets/

Advanced Molecular Detection
Southeast Region Bioinformatics

NCBI DATASETS CLI

- Two CLI tools
  - Datasets: download sequence data across all domains of life
  - Dataformat: convert metadata from JSON to other formats
- Commands follow a standard syntax



https://www.ncbi.nlm.nih.gov/datasets/docs/v2/download-and-install/

Advanced Molecular Detection
Southeast Region Bioinformatics

NCBI DATASETS CLI CONT.

- 3-Step Conda install (includes both datasets and dataformats in the conda package)
  1. Create the conda environment: conda create –n ncbi_datasets
  2. Activate the environment: conda activate ncbi_datasets
  3. Install the datasets conda package: conda install -c conda-forge ncbi-datasets-cli
     - Note the switch from _ to – in ncbi-datasets
  - Example code
    - datasets download genome accession GCA_020809405.1
    - datasets download genome accession GCA_020809405.1 GCA_020748185.1
      - An example of multiple genomes

Advanced Molecular Detection
Southeast Region Bioinformatics

NCBI DATASETS GITHUB

- Request a new feature or submit a bug report
  - .github/ISSUE_TEMPLATE
    - bug_report.md
    - feature_request.md
  - Current version as of 6/23/25 is v18.x

Advanced Molecular Detection
Southeast Region Bioinformatics

CONCLUSION

- Fundamentals of Human Scrubber and NCBI Datasets
- Installation and setup of Human Scrubber in HPG
- Successfully executed job query for Human Scrubber
- Generated output file for Human Scrubber

Advanced Molecular Detection
Southeast Region Bioinformatics

Advanced Molecular Detection
Southeast Region Bioinformatics

Questions?

bphl-sebioinformatics@flhealth.gov

Molly Mitchell, PhD
Bioinformatics Supervisor
Molly.Mitchell@flhealth.gov

Nikhil Reddy, MS
Bioinformatician
Nikhil.Yengala@flhealth.gov

Sam Bernhoft, MPH
Bioinformatician
Samantha.bernhoft@flhealth.gov