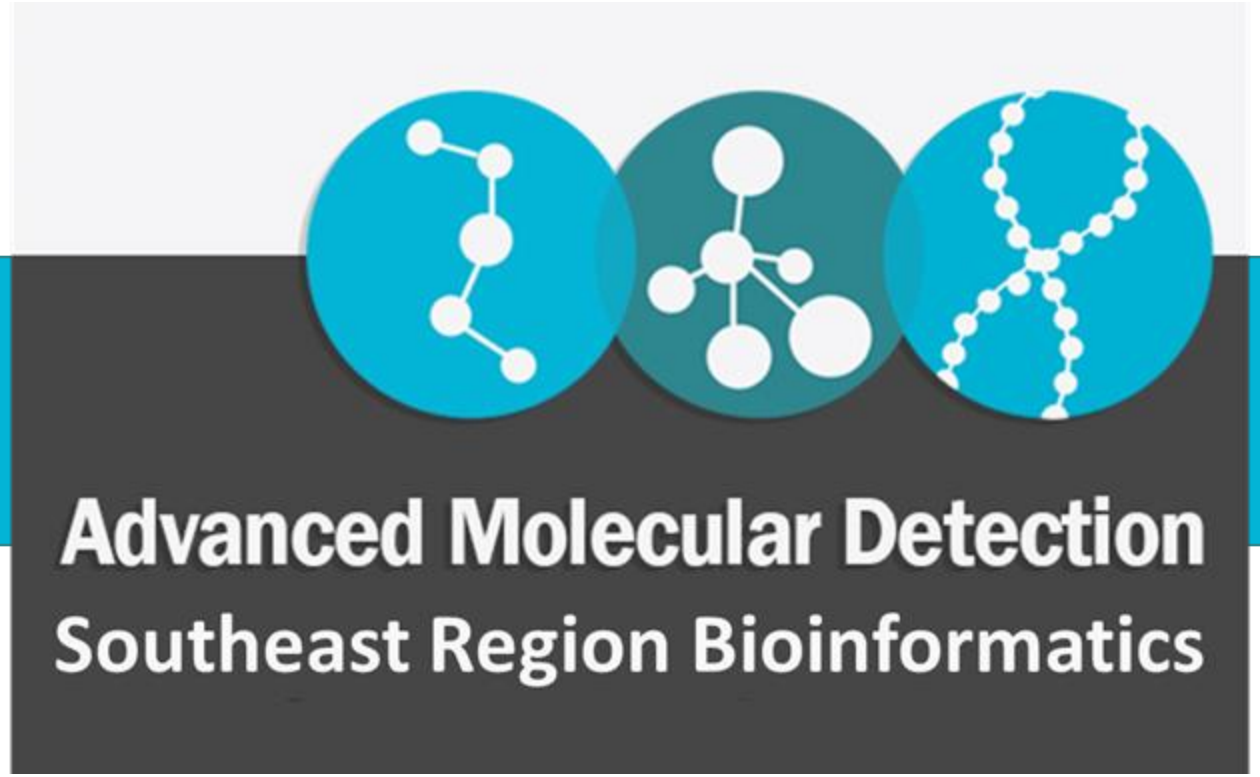


SRA Human Scrubber

Sam Marcellus, MPH

8/5/2024

- Florida's IT Issues have been resolved (-ish)
- Planned one-day HPG outage sometime between August 12th and 21st. Exact day TBA.
- HPG updates effective August 18th
  - TN now has 13 TB Blue and 10 TB Orange (Previously 10 and 5)
  - GA now was 8 TB Blue and 15 TB Orange (Previously 5 and 10)
- PhaME (Phylogenetic and Molecular Evolution Analysis Tool) is now available on our GitHub
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6997174/
  - Yibo debugged it and made sure it runs on HPG
    - https://github.com/BPHL-Molecular/PhaME_m

Advanced Molecular Detection
Southeast Region Bioinformatics

SRA Human Scrubber

What is HRRT/Human Scrubber?

Pulling Scrubber

Applying Scrubber

NCBI Data Sets

The Human Read Removal Tool
AKA Human Scrubber

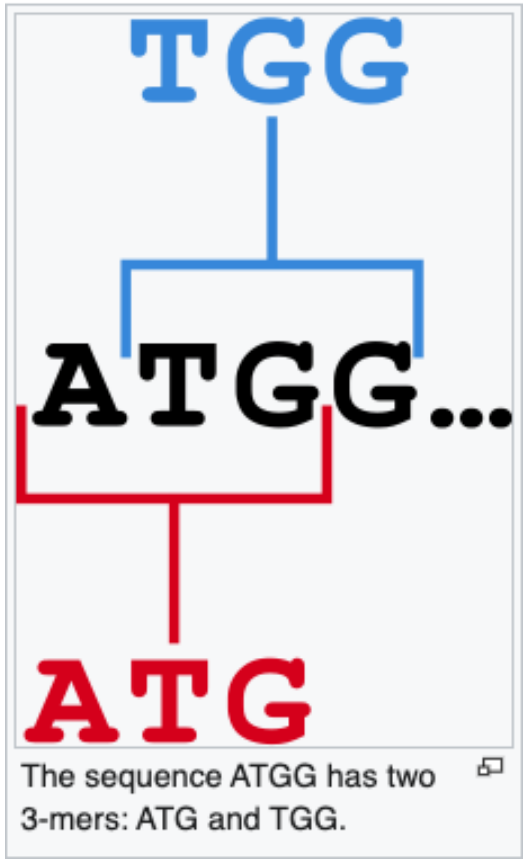
- Removes any probable human reads from pathogen genomes
- Input .fastq -> Output .fastq.clean
  - Reads identified as potentially human as labelled "N"
- Based on SRA Taxonomy Analysis Tool (STAT)
  - MinHash-based k-mer tool (k=32bp)
  - Developed for QA assessment of SRA submissions

Advanced Molecular Detection
Southeast Region Bioinformatics

https://ncbiinsights.ncbi.nlm.nih.gov/2023/02/02/scrubbing-human-sequences-sra-submissions/

K-Mers

- Sub-string of a longer DNA segment
  - K is the length of the sequence (4-mer, 2-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



The sequence ATGG has two 3-mers: ATG and TGG.

***k*-mers for GTAGAGCTGT**

k	*k*-mers
1	G, T, A, C
2	GT, TA, AG, GA, AG, GC, CT, TG
3	GTA, TAG, AGA, GAG, AGC, GCT, CTG, TGT
4	GTAG, TAGA, AGAG, GAGC, AGCT, GCTG, CTGT
5	GTAGA, TAGAG, AGAGC, GAGCT, AGCTG, GCTGT
6	GTAGAG, TAGAGC, AGAGCT, GAGCTG, AGCTGT
7	GTAGAGC, TAGAGCT, AGAGCTG, GAGCTGT
8	GTAGAGCT, TAGAGCTG, AGAGCTGT
9	GTAGAGCTG, TAGAGCTGT
10	GTAGAGCTGT

Advanced Molecular Detection
Southeast Region Bioinformatics

https://en.wikipedia.org/wiki/K-mer

How to Pull Human Scrubber-GitHub

- GitHub
  - https://github.com/ncbi/sra-human-scrubber
  - Quick Start Guide
    - Clone the repo.
    - `pushd` or `cd` to directory `sra-human-scrubber`.
      - Alternatively, download the zip file from the green 'Code' button, unzip it, then cd to directory `sra-human-scrubber-master`.
    - Execute `./init_db.sh` in directory `sra-human-scrubber` - this will retrieve the default (newest) pre-built db from ftp and place it in the directory `sra-human-scrubber/data` where it needs to be located.
    - Please note binary `aligns_to` in bin was compiled on x86_64 GNU/Linux.
    - Please refer to CHANGELOG for recent changes.

Advanced Molecular Detection
Southeast Region Bioinformatics

https://github.com/ncbi/sra-human-scrubber?tab=readme-ov-file#quick-start

GitHub Scrubber Test

Here the command is simply given the (file) argument `test` `./scripts/scrub.sh test`

```
./scripts/scrub.sh test
2022-08-31 14:35:08     aligns_to version 0.707
2022-08-31 14:35:08     hardware threads: 32, omp threads: 32
2022-08-31 14:35:09     loading time (sec) 1
2022-08-31 14:35:09     /tmp/tmp.EpHdBbPYzb/temp.fasta
2022-08-31 14:35:09     FastaReader
2022-08-31 14:35:09     100% processed
2022-08-31 14:35:09     total spot count: 2
2022-08-31 14:35:09     total read count: 2
2022-08-31 14:35:09     total time (sec) 1
1  spot(s) masked or removed.

test succeeded
```

Other useful options:

```
./scripts/scrub.sh -h
Usage: scrub.sh [OPTIONS] [file.fastq]
OPTIONS:
        -i <input_file>; Input Fastq File.
        -o <output_file>; Save cleaned sequence reads to file, or set to - for stdout.
                NOTE: When stdin is used, output is stdout by default.
        -p <number> Number of threads to use.
        -d <database_path>; Specify a database other than default to use.
        -x ; Remove spots instead of default 'N' replacement.
                NOTE: Now by default sequence length of identified spots replaced with 'N'.
        -r ; Save identified spots to <input_file>.spots_removed.
        -u <user_named_file>; Save identified spots to <user_named_file>.
                NOTE: Required with -r if output is stdout, otherwise optional.
        -t ; Run test.
        -s ; Input is (collated) interleaved paired-end(read) file AND you wish both reads mas
        -h ; Display this message.
```

Advanced Molecular Detection
Southeast Region Bioinformatics

https://github.com/ncbi/sra-human-scrubber?tab=readme-ov-file#quick-start

How to 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

How to 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

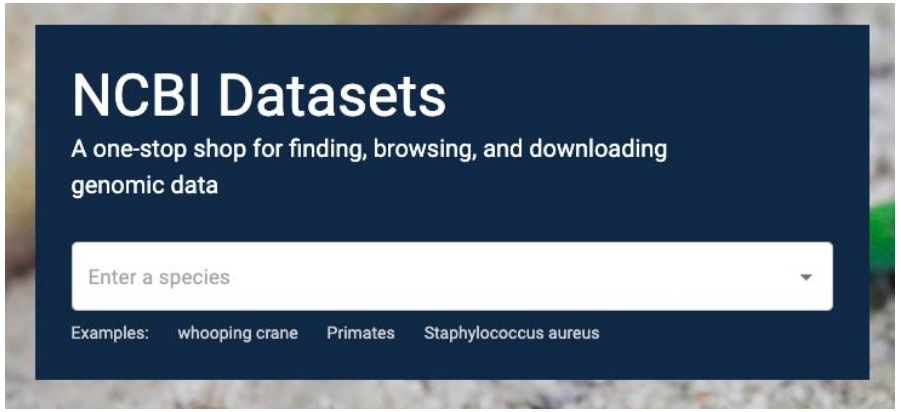
Applying Human Scrubber

- 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

https://ncbiinsights.ncbi.nlm.nih.gov/2023/02/02/scrubbing-human-sequences-sra-submissions/

NCBI Data Sets

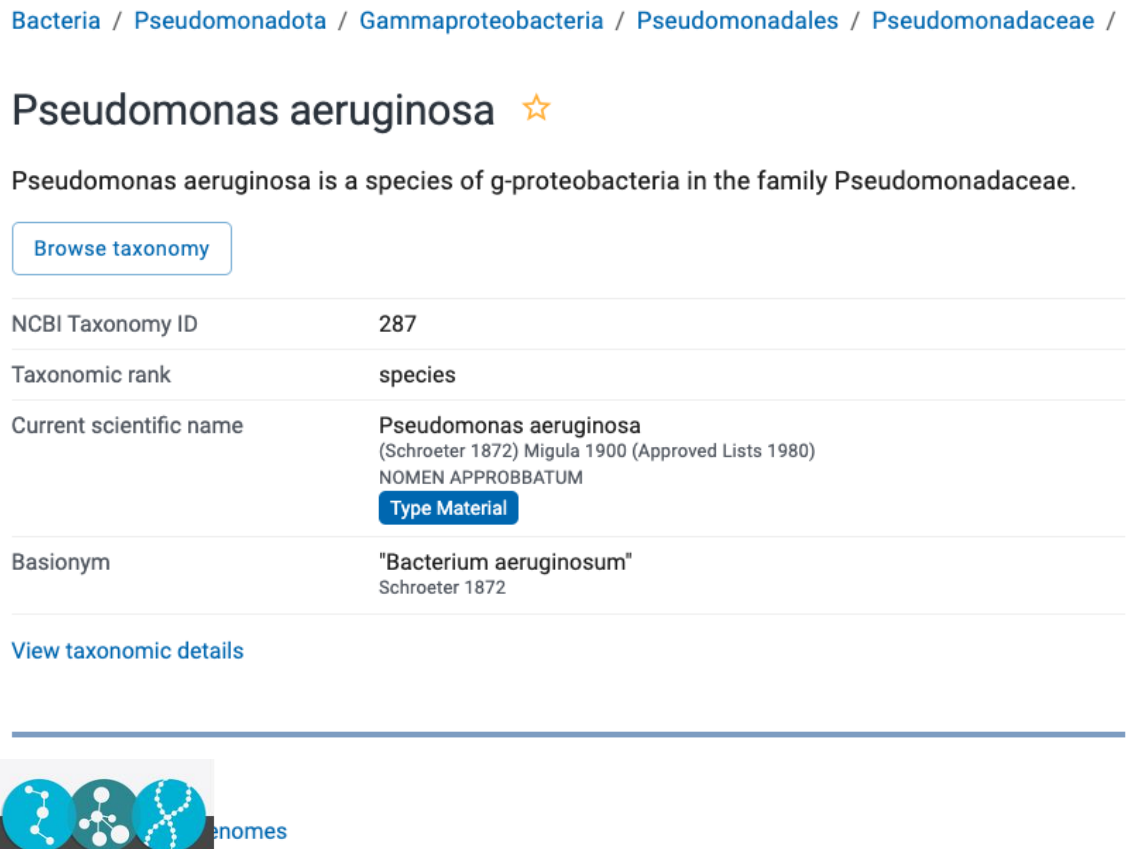
- NCBI has data sets for almost anything you could ever want
- Taxonomy, gene, and genome level
  - Special data set for viruses
- Can access in 3 ways
  - CLI
  - GitHub
  - NCBI Website
- Excellent How-To Guides on their website
  - https://www.ncbi.nlm.nih.gov/datasets/docs/v2/how-tos/

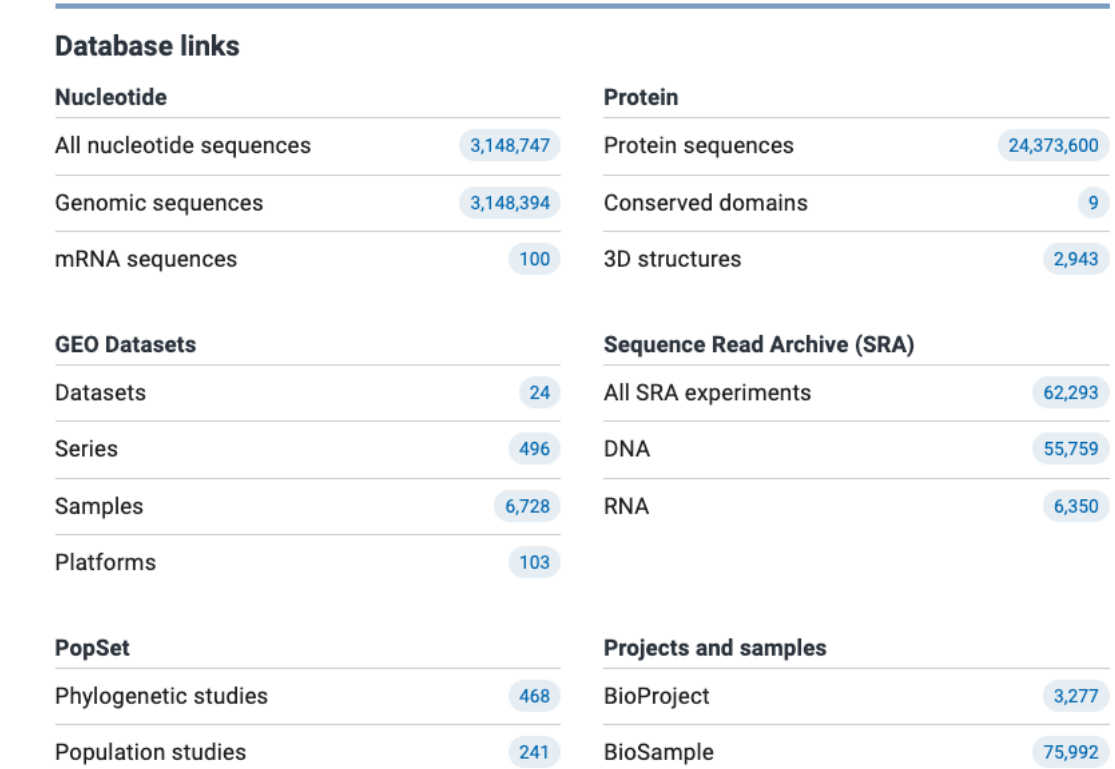


Advanced Molecular Detection
Southeast Region Bioinformatics

NCBI Data Sets GUI

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

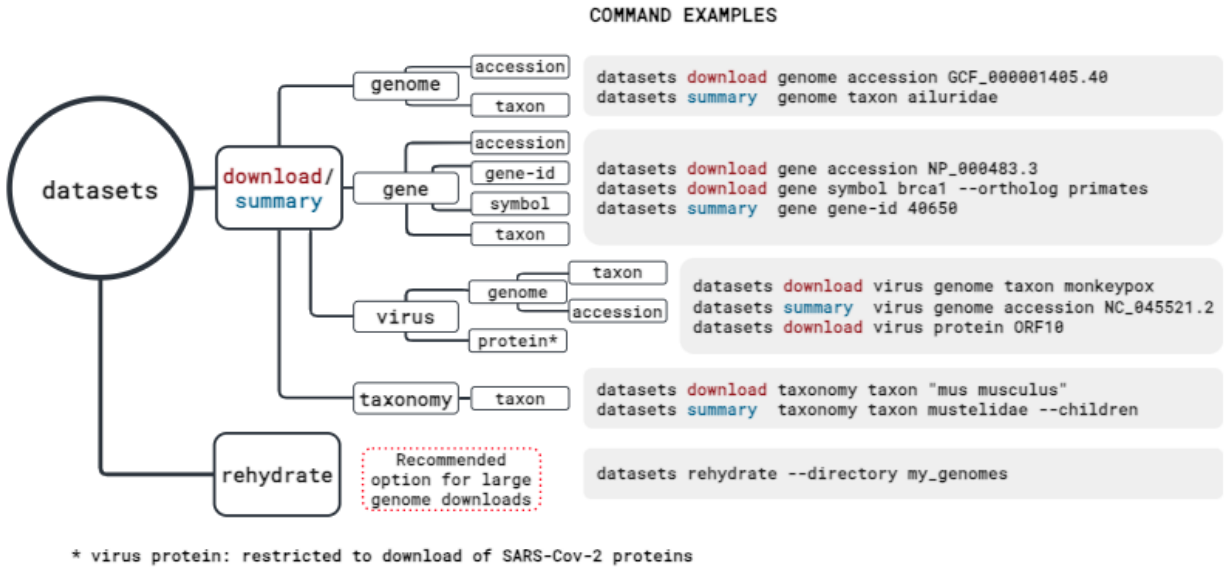




Advanced Molecular Detection
Southeast Region Bioinformatics

NCBI Data Sets 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



Advanced Molecular Detection
Southeast Region Bioinformatics

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

NCBI Data Sets CLI

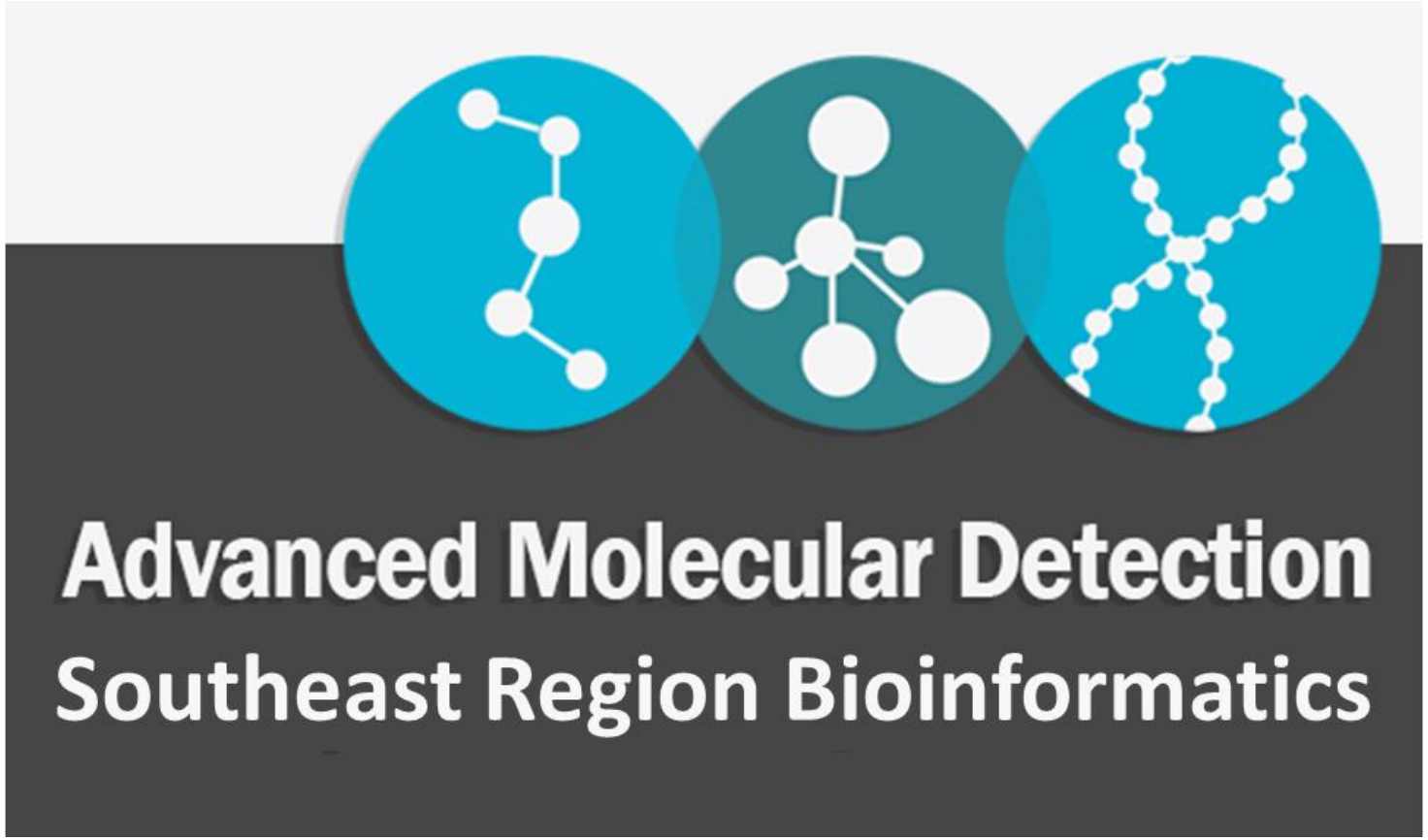
- 3-Step Conda install (includes both datasets and dataformats in the conda package)
  1. Create the conda environment: conda create –n ncbi_datasets
  2. Activate the environment: conda activate ncbi_datasets
  3. Install the datasets conda package: conda install -c conda-forge ncbi-datasets-cli
     - 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 Data Sets GitHub

- Request a new feature or submit a bug report
  - .github/ISSUE_TEMPLATE
    - bug_report.md
    - feature_request.md
- ⚠ "The NCBI Datasets command-line tools (CLI) v13.x and older, as well as the API v1, will be deprecated in June 2024 and then retired in December 2024. Please download and install the latest version."
  - Current version as of 8/1/24 is v16.x

Advanced Molecular Detection
Southeast Region Bioinformatics



Questions?

bphl-sebioinformatics@flhealth.gov

Sam Marcellus, MPH
Bioinformatician
Samantha.marcellus@flhealth.gov

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

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