



Advanced Molecular Detection Southeast Region Bioinformatics



SC2 Data Submissions, Part1: General Overview

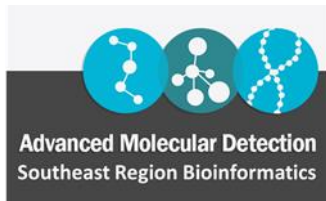
September 15, 2021

BPHL-SEbioinformatics@flhealth.gov

SARS-CoV-2 Data Submission Training Series

- **Part 1: General Overview**
- Part 2: Sample Review, Batch, and Multi-Fasta File Prep
- Part 3: Submissions to GISAID and NCBI
- Part 4: FASTQ de-host and SRA Submissions
- Part 5: Flagged Sample Review, Variant Confirmation, and Assembly Correction

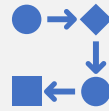
Advanced Molecular Detection
Southeast Region Bioinformatics



Outline



AMD Southeast Region and HiPerGator Updates



SARS-CoV-2 Data Submissions Workflow



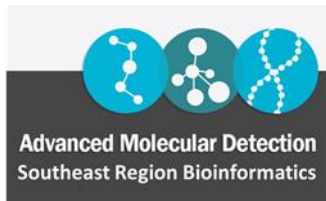
GISAID Overview



NCBI Overview

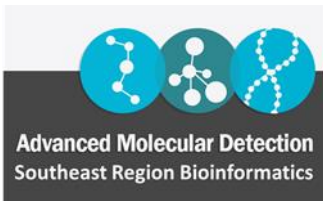
AMD Southeast Region Updates

- 2021-2022: FL and GA are the Training/WFD Lead
 - Annual week-long training – estimated Spring 2022 in Atlanta, GA
 - Online trainings/webinars – as needed or requested throughout the year
 - One-on-one or group trainings
 - Linux 101 Training
 - HiPerGator Training
 - BaseSpace Command Line Interface Training
 - Bacterial WGS analysis Training
 - Viral NGS analysis Training
- New Bioinformatics Staff in FL to support BRR
 - Jiaqi Li, PhD
- Increased compute resources on HiPerGator



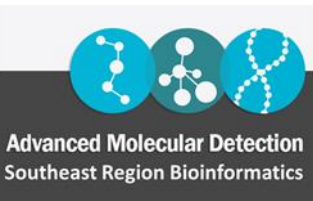
HiPerGator Updates

- **Effective today, 9/15/2021**, each jurisdiction has access to the following increased compute resources on HiPerGator
 - NCUs (1 NCU = 1 CPU, 8 GB RAM) = 150
 - Blue storage (for current analyses, high-performance) = 2 TB
 - Orange storage = 10 TB
- *Blue and orange storage increase are available upon request and based on usage
- **REMINDER** – HiPerGator is free to all Southeast Region jurisdictions
 - Unlimited analyses
 - Analysis runs will just be queued if more than 150 NCUs are in use at a single time.



HiPerGator Resources/Information

- Visit <https://github.com/StaPH-B/southeast-region> for the following information:
 - Request access for HiPerGator account
 - Help and Documentation for HiPerGator
 - Past trainings and webinars (recordings and slides)
- Requirements/required trainings for SC2 submissions using HiPerGator:
 - HiPerGator account
 - Linux 101 Training
 - HiPerGator Training (walk-through HiPerGator Analysis Reference Guide with BRR) and compute setup
 - BaseSpace CLI Training

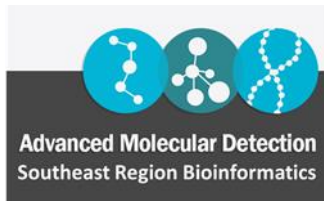




Current Analysis Options

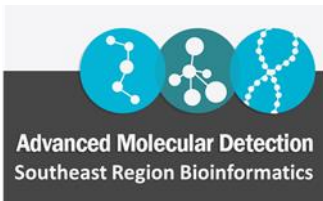
Scripts and pipelines available for state PHL use on HiPerGator

Southeast Region Bioinformatics



Pipelines and Scripts

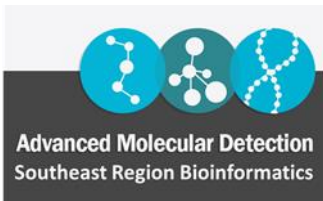
- FLAQ (FLorida Assembly Quality)
 - Generates annotated, de novo assemblies and quality metrics for WGS of bacterial species
- FLAQ-AMR (FLAQ - Antimicrobial Resistance)
 - Generates annotated, de novo assemblies and quality metrics for WGS of bacterial species
 - Determines species ID
 - Determines ST using MLST schemes from PubMLST
 - Identifies AMR genes, virulence genes, and plasmids
 - Performs serotyping of *Salmonella* and *E. coli*, if applicable (more species to come)
- FLAQ-Lp (FLAQ – *Legionella pneumophila*)
 - Includes CDC's Lp Species ID Tool and Lp Serotyping Tool
- FLAQ-USC (FLAQ – Unknown Species Classification)
 - Consensus identification from Mash, Metaphlan3, and Kraken2 (mini and std db)



Pipelines for Comparative Genomics

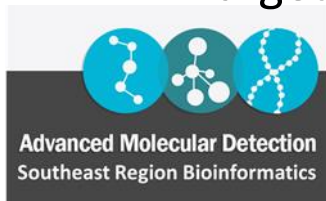
- FL-cgSNP (Core-Genome SNP analysis) – reference-free method
 - Pan-genome analysis (identifies core genes shared by all isolates) and generates a multiple sequence alignment
 - Identifies pair-wise SNPs between isolates and outputs a pairwise SNP matrix
 - Generates a maximum-likelihood phylogenetic tree
- hqSNP (High-Quality SNP analysis) – reference-based method
 - Uses CDC's Lyveset pipeline to identify hqSNPs
 - Generates a pairwise SNP matrix and maximum-likelihood phylogenetic tree
- Custom scripts to output annotated tree (based on metadata input)

Advanced Molecular Detection
Southeast Region Bioinformatics



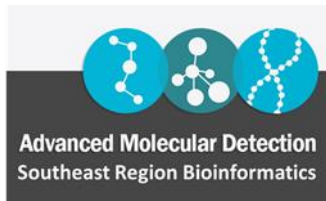
Viral Pipelines and scripts

- FLAQ - SC2 (FLAQ – SARS-CoV-2)
 - Generates SARS-CoV-2 consensus assemblies from ARTIC V1, V2, V3, V4 targeted amplicon sequencing using Illumina (e.g., Nextera XT or Flex) and non-Illumina (e.g., PrimalSeq or MN Tailed) library prep
 - Outputs variant file and final report with quality metrics (including a PASS/FAIL quality flag and a PASS/REVIEW annotation flag based on public repository submission criteria)
 - Now compatible with QiaSEQ and COVIDSeq (at least 150bp read lengths)
- FLAQ – SC2-ClearLabs
 - Assembly quality metrics, pangolin lineage call, nextclade report, and VADR review from ClearLabs generated data
- SC2 associated scripts to prepare/format assemblies for batch submissions to GISAID and NCBI's Genbank
- SC2-Correct-Assembly
 - Removes indels/SNPs that are likely PCR or sequencing artifacts/errors (requires prior manual review of mapped reads in IGV (or other program))
- Targeted Amplicon Variant Calling and Consensus Sequence Generation



Scripts and Individual Tools

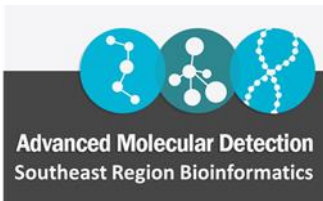
- Quick Species ID (screen against RefSeq database)
- Species ID, contamination check, and metagenomic classification
- Pull out gene sequence of interest from assembly (e.g., pull out AR genes of interest)
- Merge fastqs from multiple lanes on a NextSeq into one R1 and one R2 file (compatible with BioNumerics)
- Download fastqs from NCBI's Sequence Read Archive
- Run any tool individually
 - SeqSero2, mlst, abricate, etc.
- Run any tool or workflow in the Staph-B Toolkit (https://staph-b.github.io/staphb_toolkit/)
- Batch runs (i.e., run >2 analysis scripts, multiple samples at a time)
- **Custom scripts and pipelines as requested or needed**



Analysis Request and/or support

- Email BPHL-SEbioinformatics@flhealth.gov
- One-on-one video conference sessions to set up your HiPerGator environment and walk-through the use of each pipeline/script needed
- Custom pipeline/script development as requested
- Scripts (and data, if applicable) can be shared through your /blue/bphl-<state>/public-share/ directory on HiPerGator

Advanced Molecular Detection
Southeast Region Bioinformatics

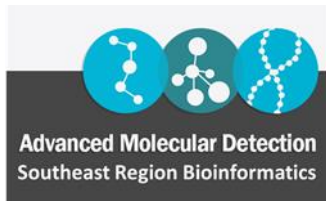




SARS-CoV-2 Data Submissions

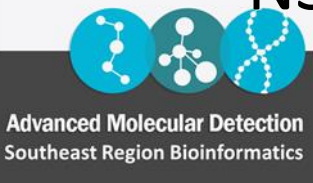
General Overview

Southeast Region Bioinformatics



SC2 Sequencing Resources - APHL

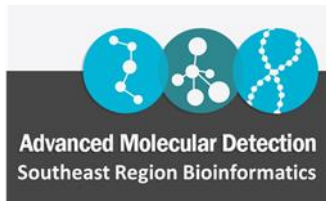
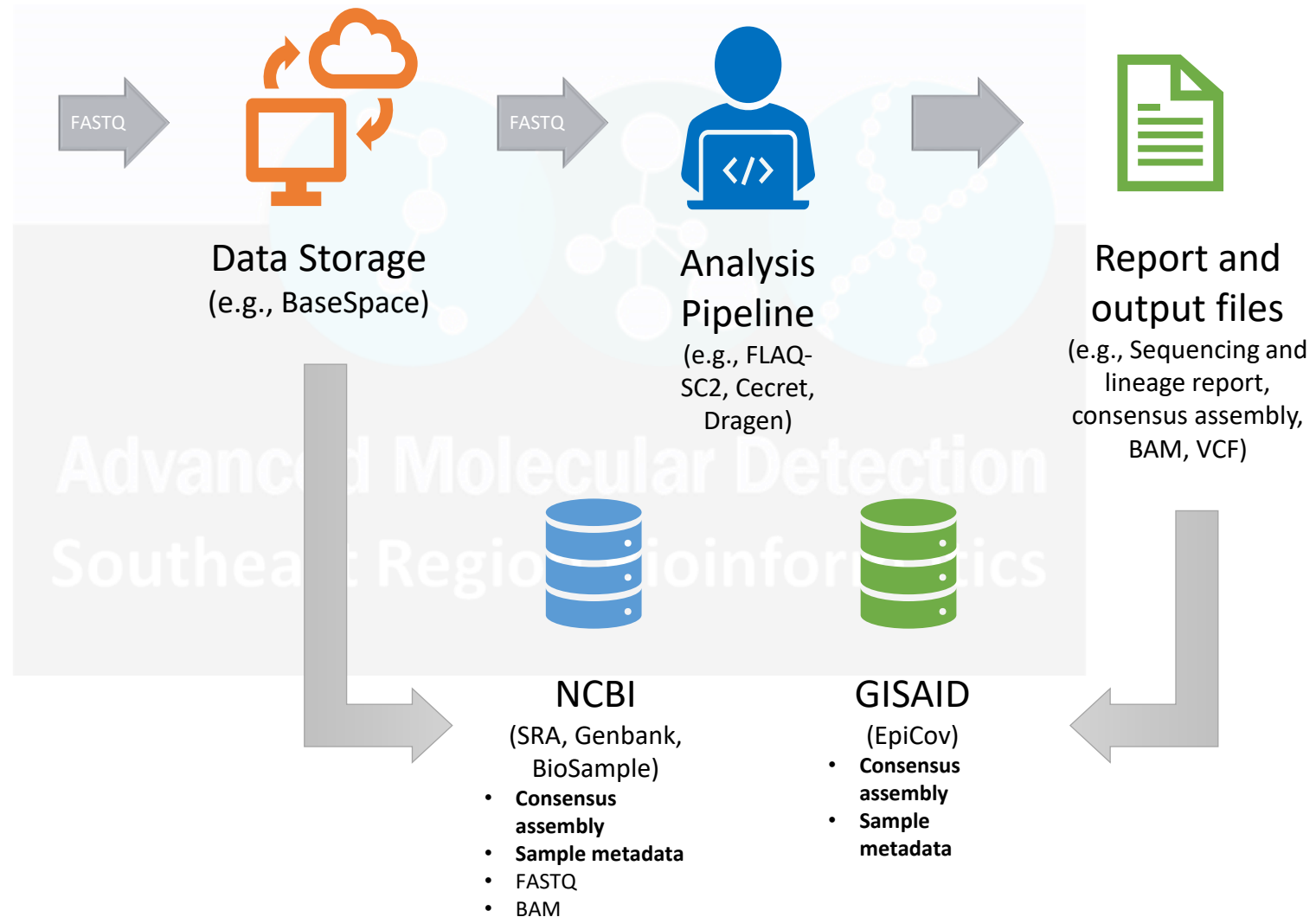
- <https://www.aphl.org/programs/preparedness/Crisis-Management/COVID-19-Response/Pages/Sequencing.aspx>
- Wet Lab Protocols
- Genome Submissions to Public Repositories
 - **APHL SARS-CoV-2 Sequencing Recommendations**
 - <https://www.aphl.org/programs/preparedness/Crisis-Management/Documents/APHL-SARS-CoV-2-Sequencing.pdf> (*currently being updated*)
- Bioinformatics
- Training Resources
- NS3



SARS-CoV-2 Sequencing Workflow



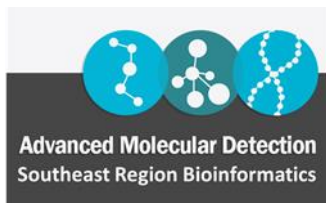
Tiled-Amplicon or
Enrichment-based
sequencing



SC2 Consensus Assembly Submissions to GISAID and NCBI

- Submission Process

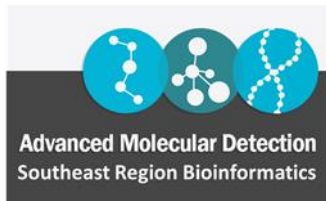
- ✓ Screen passing QC samples for submission (VADR – HiPerGator)
- ✓ Select samples for submission
- ✓ Collect relevant sample metadata needed for submission
- ✓ Assign public repository sample names
- ✓ Prepare formatted multi-fasta files for GISAID and Genbank (HiPerGator)
- ✓ Submit to GISAID – submit metadata template and multi-fasta file
- ✓ Retrieve GISAID accessions
- ✓ Submit to NCBI Biosample - submit metadata template (with linked GISAID accessions)
- ✓ Save NCBI Biosample accessions
- ✓ Submit to NCBI Genbank – submit metadata template (with linked GISAID and Biosample accessions) and multi-fasta file
- ✓ Save NCBI Genbank accessions
- ✓ **SUBMISSION COMPLETE and all data is now linked!!!**



GISAID

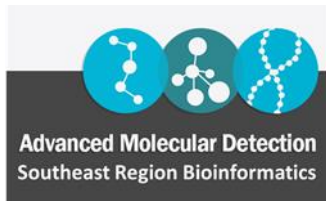
- GISAID (Global Initiative on Sharing Avian[All] Influenza Data)
 - <https://www.gisaid.org/>
- Public-Private Partnership – non-profit, Germany, Singapore, USA, private and corporate philanthropy
- EpiFlu and EpiCov databases
- Encourages sharing of data prior to publication, protection of intellectual rights to the data, and acknowledgement for data usage
- >3.5 million SARS-CoV-2 genomes

Southeast Region Bioinformatics



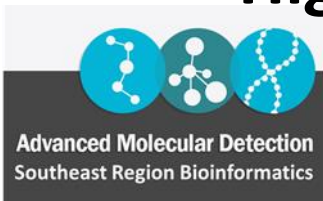
NCBI

- BioSample
 - Database for specimen metadata
- Genbank
 - Database for nucleotide sequences (e.g., consensus assemblies)
 - >1.3 million SC2 sequences
- Sequence Read Archive (SRA)
 - Database for raw sequencing data (e.g., FASTQs, BAMs)
 - >1.1 million SC2 datasets
- Other NCBI SC2 resources
 - <https://www.ncbi.nlm.nih.gov/sars-cov-2/>
 - NCBI Virus (<https://www.ncbi.nlm.nih.gov/labs/virus/>)
 - NCBI Datasets (<https://www.ncbi.nlm.nih.gov/datasets/>)



Why do we have to submit to both?

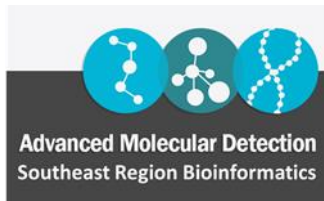
- GISAID
 - Public-private partnership
 - Data usage policies and terms of use apply
 - Cannot download and host data elsewhere without permission
- NCBI
 - Public, open data
- Differences in features and ways to query data
- Differences in acceptance criteria
- Data are pulled from GISAID and NCBI for different purposes
- **Highly recommended to submit to both!!!**



GISAID and NCBI Account Set-up

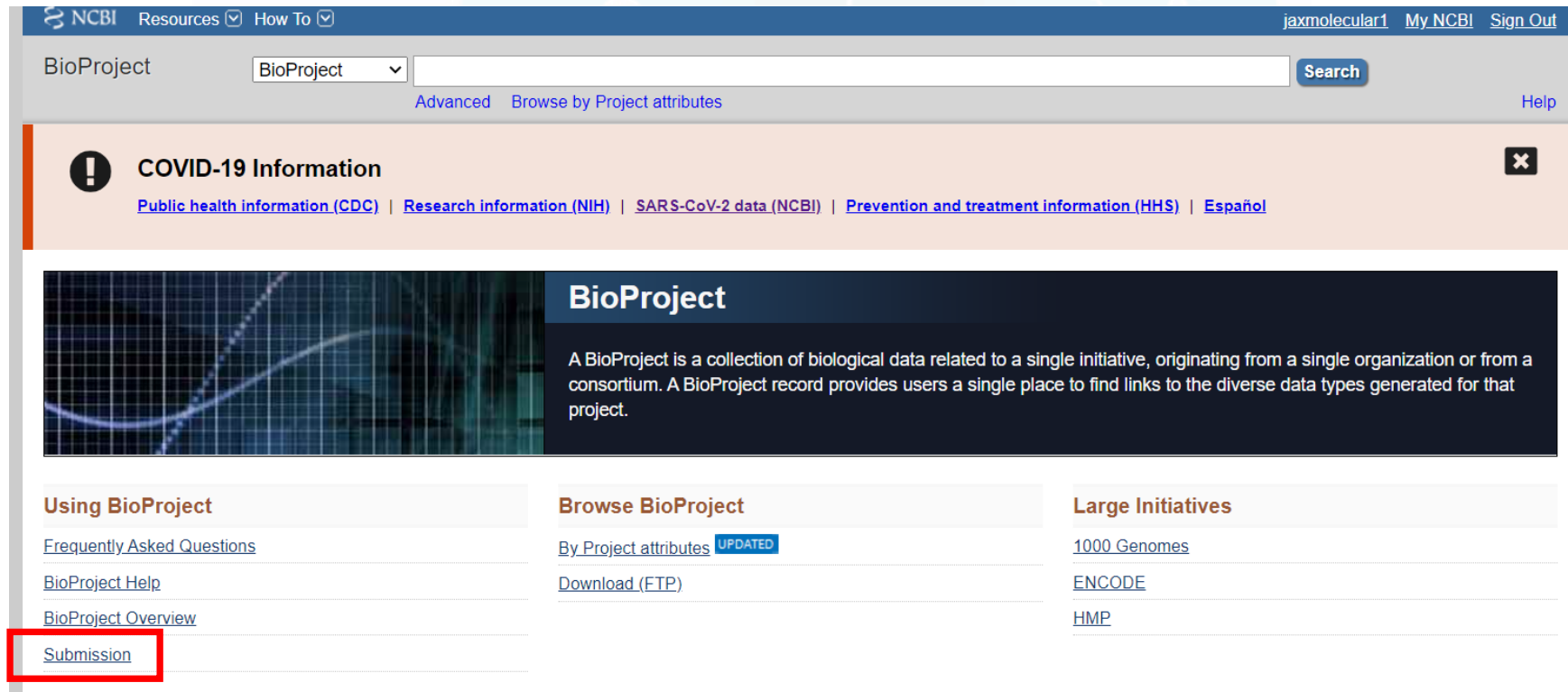
- GISAID
 - <https://www.gisaid.org/registration/register/>
- NCBI
 - Create a lab NCBI account if you do not already have one. Most PHLs should have one for other WGS data (e.g., PulseNet, ARLN, etc).
 - **Create a SARS-CoV-2 sequencing BioProject for your lab.**
 - This will link all SC2 sequence data from your PHL.
 - <https://www.ncbi.nlm.nih.gov/bioproject/>

Advanced Molecular Detection
Southeast Region Bioinformatics



NCBI BioProject

- Create a SARS-CoV-2 sequencing BioProject for your lab.
 - Follow the instructions in the Submission wizard.
 - Link to CDC Umbrella BioProject - **PRJNA615625**
 - Demo



NCBI Resources How To jaxmolecular1 My NCBI Sign Out

BioProject BioProject Search

Advanced Browse by Project attributes Help

COVID-19 Information

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

BioProject

A BioProject is a collection of biological data related to a single initiative, originating from a single organization or from a consortium. A BioProject record provides users a single place to find links to the diverse data types generated for that project.

Using BioProject

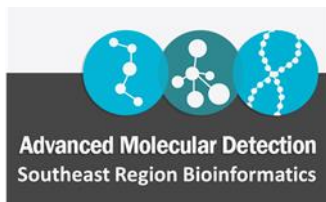
- [Frequently Asked Questions](#)
- [BioProject Help](#)
- [BioProject Overview](#)
- [Submission](#)

Browse BioProject

- [By Project attributes](#) **UPDATED**
- [Download \(ETP\)](#)

Large Initiatives

- [1000 Genomes](#)
- [ENCODE](#)
- [HMP](#)

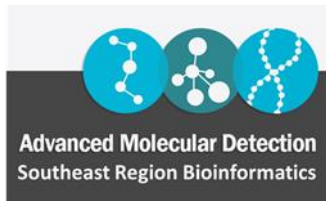


GISAID Features (Demo)

- Search for data by state
- Search by lineage/variants/mutations
- Download data for downstream analysis or reporting
- <https://www.gisaid.org/>



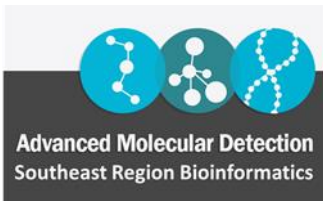
Advanced Molecular Detection
Southeast Region Bioinformatics



GISAID “Lineage Pull” (Demo)

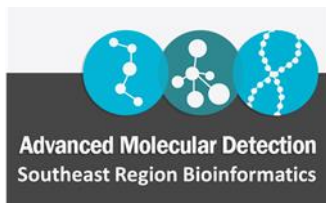
- GISAID (and NCBI) run the latest version of Pangolin on all data every day.
- Up-to-date state specific lineage data can be downloaded.
- GISAID only allows up to 10,000 genomes/associated metadata to be downloaded at once using their web portal.
- If your state has >10,000 genomes submitted, then a “metadata download package” can be parsed to retrieve your state’s data.

Advanced Molecular Detection
Southeast Region Bioinformatics



GISAID “Lineage Pull” – Florida Example

- Click on “Downloads” tab
- Scroll down to “Download packages”
- Click on “metadata”
- Check “I agree to terms and conditions”. Click “Download” button.
- Transfer downloaded metadata file to HiPerGator via WinSCP
 - /blue/bphl-<state>/<user>/gisaid_lineage_pull
 - First time only – make new directory via WinSCP in your user directory



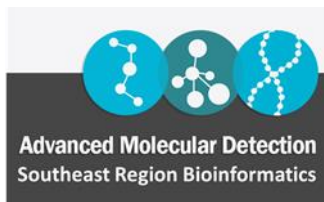
GISAID “Lineage Pull” – Florida Example

- Log in to HiPerGator via Putty

```
[usr@login]$ cd /blue/bphl-<state>/<usr>/gisaid_lineage_pull
[usr@login]$ ls                                     #You should see your file you just transferred
[usr@login]$ unxz metadata_tsv_<date>.tar.xz
[usr@login]$ tar -xf metadata_tsv_<date>.tar
[usr@login]$ head -n 1 metadata.tsv > 20210915_FL_lineages_all.tsv    #This creates a new file with your header
[usr@login]$ grep "North America / USA / Florida" >> 20210915_FL_lineages_all.tsv  #New file with FL only data
```

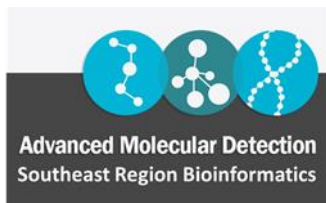
- Transfer lineage file back to your local computer via WinSCP
- The lineage file can be opened in Excel

Advanced Molecular Detection
Southeast Region Bioinformatics



Next Trainings

- **Thursday, 9/16/21** – SARS-CoV-2 Data Submissions, Part 2: Sample Review, Batch, and Multi-Fasta File Prep
- **Friday, 9/17/21** – SARS-CoV-2 Data Submissions, Part 3: Submissions to GISAID and NCBI
- **Follow-up calls with each jurisdiction for hands-on submission walk-throughs, if requested**
- **TBD** – SARS-CoV-2 Data Submissions, Part 4: FASTQ de-host and SRA Submissions
- **TBD** – SARS-CoV-2 Data Submissions, Part 5: Flagged Sample Review, Variant Confirmation, and Assembly Correction
- The recording from each training, slides, and associated training materials will be available at <https://github.com/StaPH-B/southeast-region>.





Advanced Molecular Detection Southeast Region Bioinformatics

Questions???

BPHL-SEbioinformatics@flhealth.gov

Sarah Schmedes, PhD

Lead Bioinformatician

BRR/WFD Lead, Southeast Region

Jiaqi Li, PhD

Bioinformatician

Jason Blanton, PhD

Molecular Administrator

State Sequencing Coordinator