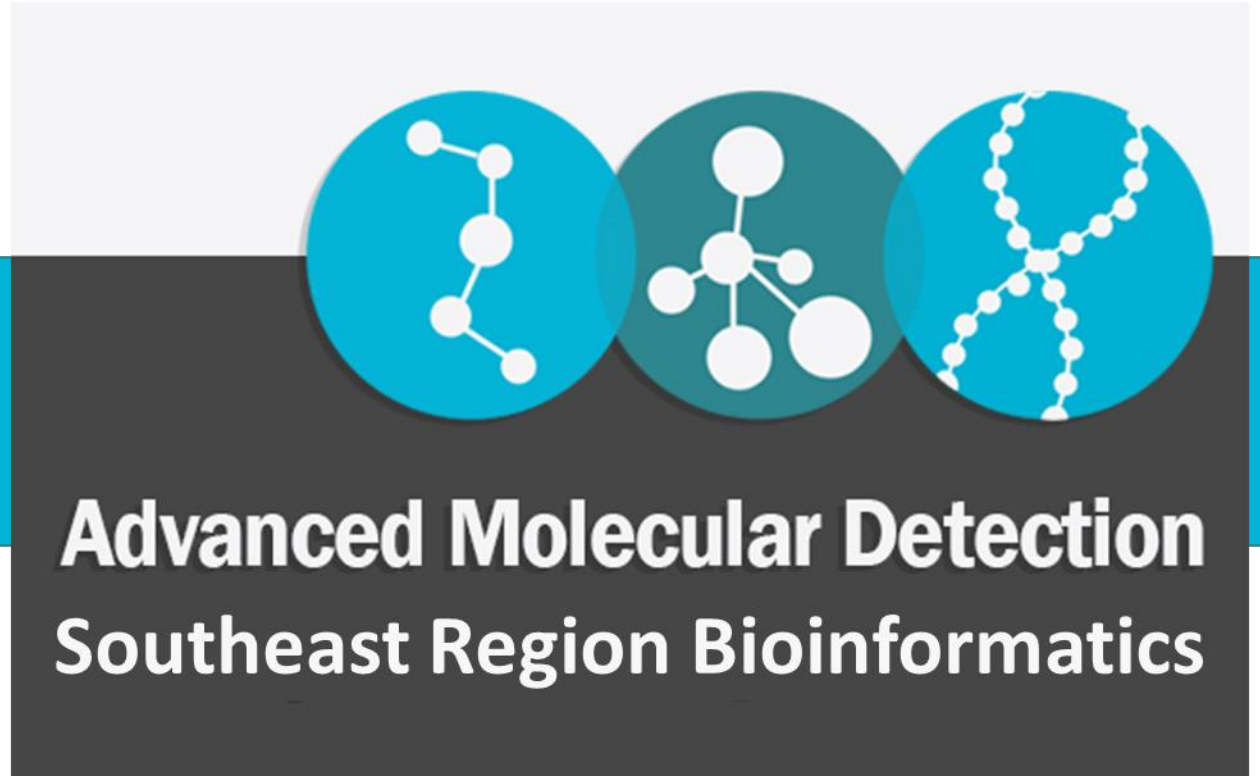


SC2 Data Submissions, Part2: Sample Review & Fasta Prep

September 16, 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

BPHL-SEbioinformatics@flhealth.gov

Outline

Sample review and batch samples

Assign public repository names and collect metadata

Generate formatted multi-fasta file

SARS-CoV-2 Sequencing Workflow

Tiled-Amplicon or Enrichment-based sequencing

FASTQ

Data Storage
(e.g., BaseSpace)

FASTQ

Analysis Pipeline
(e.g., FLAQ-SC2, Cecret, Dragen)

Report and output files
(e.g., Sequencing and lineage report, consensus assembly, BAM, VCF)

NCBI
(SRA, Genbank, BioSample)

- **Consensus assembly**
- **Sample metadata**
- FASTQ
- BAM

GISAID
(EpiCov)

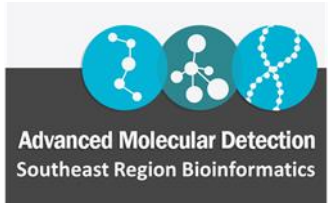
- **Consensus assembly**
- **Sample metadata**

Advanced Molecular Detection
Southeast Region Bioinformatics

BPHL-SEbioinformatics@flhealth.gov

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!!!**



BPHL-SEbioinformatics@flhealth.gov

Select Samples for Submission

- Select PASS QC samples
  - Here in Florida: ≥80% genome coverage and ≥100x mean read depth
- Sample review to flag samples with any annotation errors
  - Run NCBI's VADR on HiPerGator
  - NOTE: This step is only necessary if your analysis pipeline does not include VADR
- **FLAQ-SC2 users:** VADR is part of the pipeline. View the report.txt file for the QC and VADR flag.
- Recommended to prioritize PASS/PASS (QC/VADR) samples
  - Review and submit flagged (PASS/REVIEW) samples when available

NCBI's VADR

- Viral Annotation DefineR
- "VADR is a suite of tools for classifying and analyzing sequences homologous to a set of reference models of viral genomes or gene families. It has been mainly tested for analysis of Norovirus, Dengue, and SARS-CoV-2 virus sequences in preparation for submission to the GenBank database."
- https://github.com/ncbi/vadr/wiki/Coronavirus-annotation
- List of VADR alerts
  - https://www.ncbi.nlm.nih.gov/genbank/sequencecheck/virus/

Recent VADR Updates

- Version 1.3
- Many alerts/errors in ORF3a, ORF6, ORF7a, ORF7b, ORF8, and ORF10 do not cause a sequence to FAIL
  - “As of August 5, 2021, many common alerts (e.g. early stop codons, frameshift mutations, etc.) in ORF3a, ORF6, ORF7a, ORF7b, ORF8, and ORF10 will no longer cause a sequence to fail VADR as they did previously. Instead such problems will cause that feature to not be annotated as a CDS but rather as a misc_feature, and no protein translation product and corresponding entry in the GenBank Protein database will be created. (Since February 2021, ORF8 was misc_featur-izable in this way, but ORF3a, ORF6, ORF7a, ORF7b, and ORF10 were not.)” - https://github.com/ncbi/vadr/wiki/Coronavirus-annotation

Run VADR to screen samples prior to submission (Demo)

- If your analysis pipeline already runs VADR, then this step is complete
  - FLAQ-SC2 users: See the VADR_flag field in report.txt
- Run VADR on HiPerGator
  - Transfer consensus assemblies to HiPerGator (if not already there)
    - Use WinSCP to transfer files from local computer or you can use the BaseSpace CLI to transfer fasta files from BaseSpace if you used the Dragen COVID Lineage Pipeline

```
[usr@login]$    bs download project –n <project_name> --extension=fasta –o <output_folder>
```

  - Run **sc2_review_vadr.py** with **sbatch_vadr_review.sh**
  - Both scripts are located in /blue/bphl-<state>/public-share/scripts/

Batch samples for submission and assign public sample names (Demo)

- Log samples for submission and collect relevant metadata
- Assign public repository names
  - Example: USA/FL-BPHL-0001/2021

GISAID Prefix	NCBI Prefix	Public Base Sample Name
hCoV-19/	SARS-CoV-2/Human/	USA/FL-BPHL-0001/2021

Country of Sample Collection	State-Lab-SampleID	Sample Collection Year
USA	FL-BPHL-0001	2021

Advanced Molecular Detection
Southeast Region Bioinformatics

BPHL-SEbioinformatics@flhealth.gov

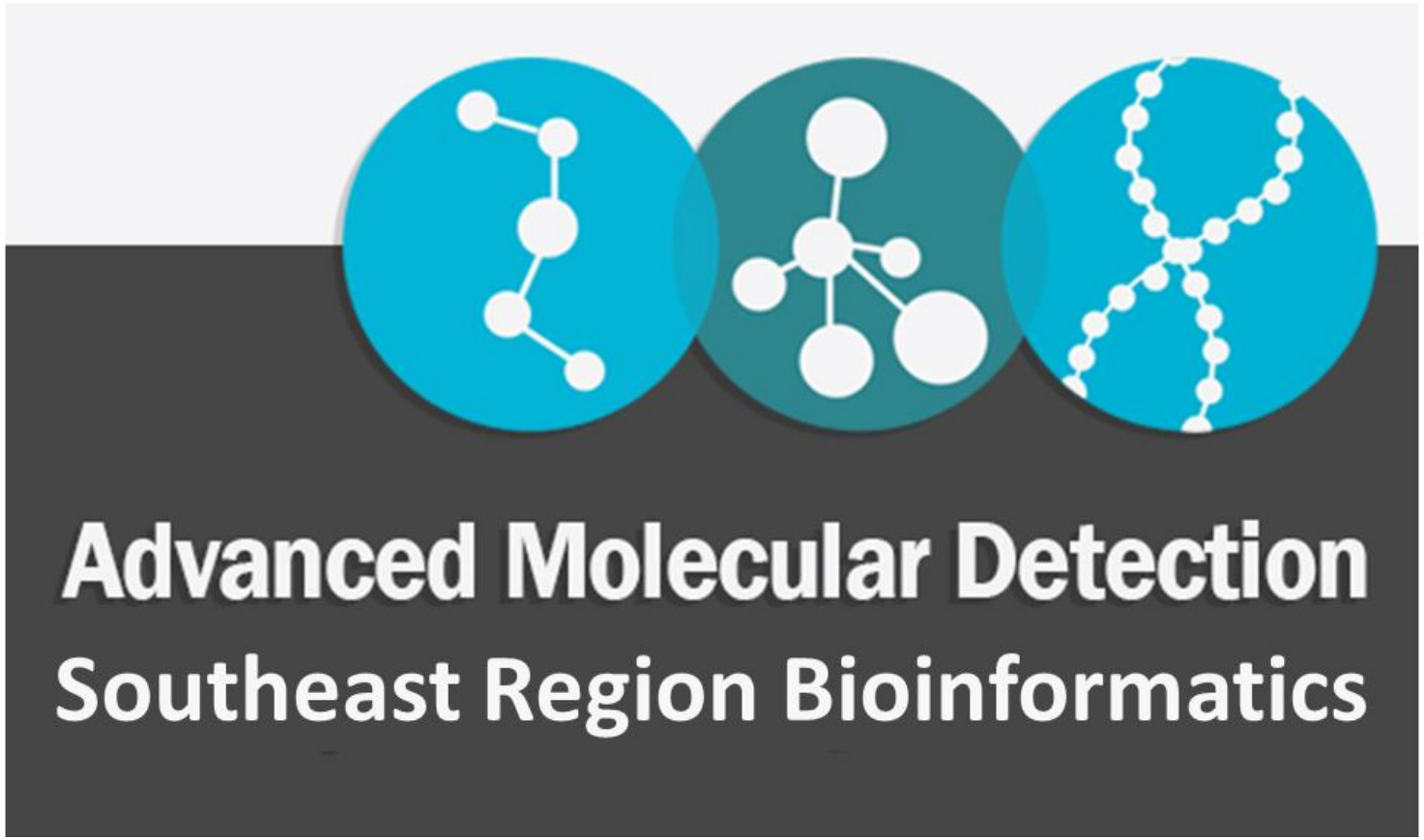
Multi-Fasta File Prep (Demo)

- GISAID and NCBI Genbank both require one multi-fasta file per batch submission
- All samples to be submitted must be renamed and properly formatted in a single file
- Run **sc2_fasta_for_sub.py** with **sbatch_sc2_fasta_sub.sh**
  - *Supports single or concatenated fastas as input. Also, supports single line and multi-line fastas as input.*
  - *Tested with assemblies generated by FLAQ-SC2 and Dragen COVID Lineage pipelines*
  - *Requires a tab-delimited input file with two columns (lab sample name, public base sample name)*
- Both scripts are located in /blue/bphl-<state>/public-share/scripts/

Next Trainings

- **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.



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