



# **Advanced Molecular Detection**

## **Southeast Region Bioinformatics**

**AMD Southeast Region Check-in Call**  
02/02/2023

# Outline



Workforce Development & BRR Staff Introduction



HPG Updates & Trainings



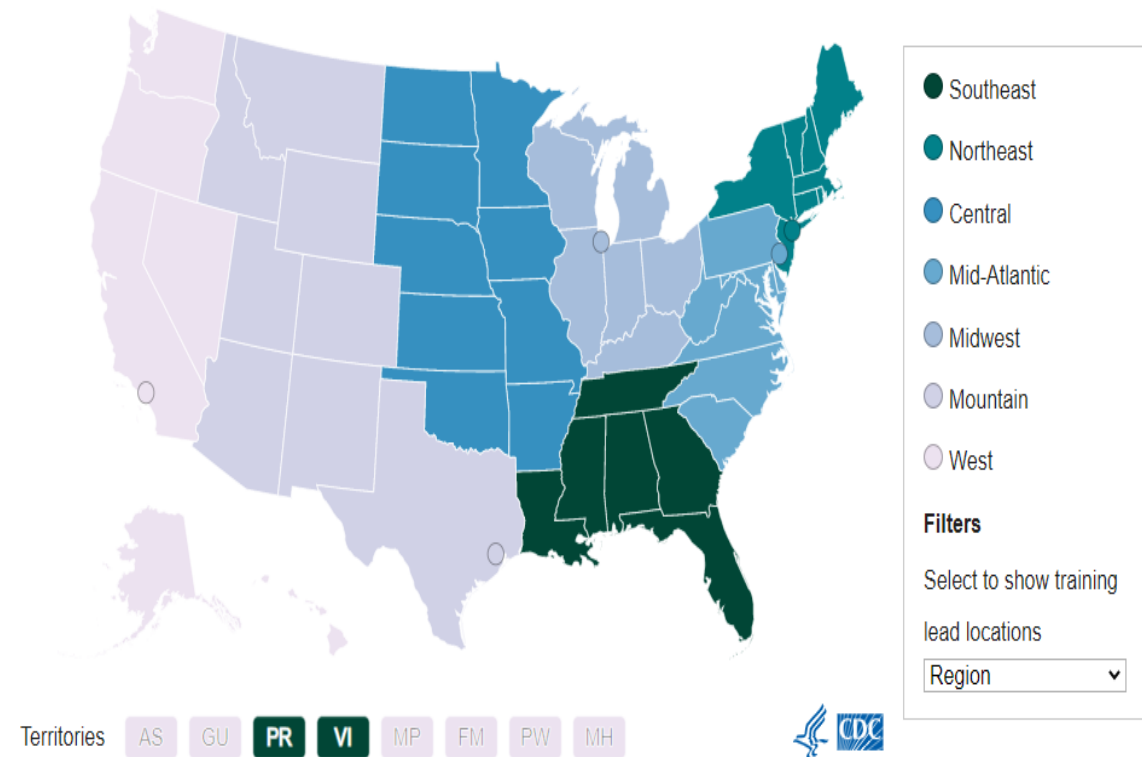
Pipeline Updates (FL & CDC)



ELC Updates

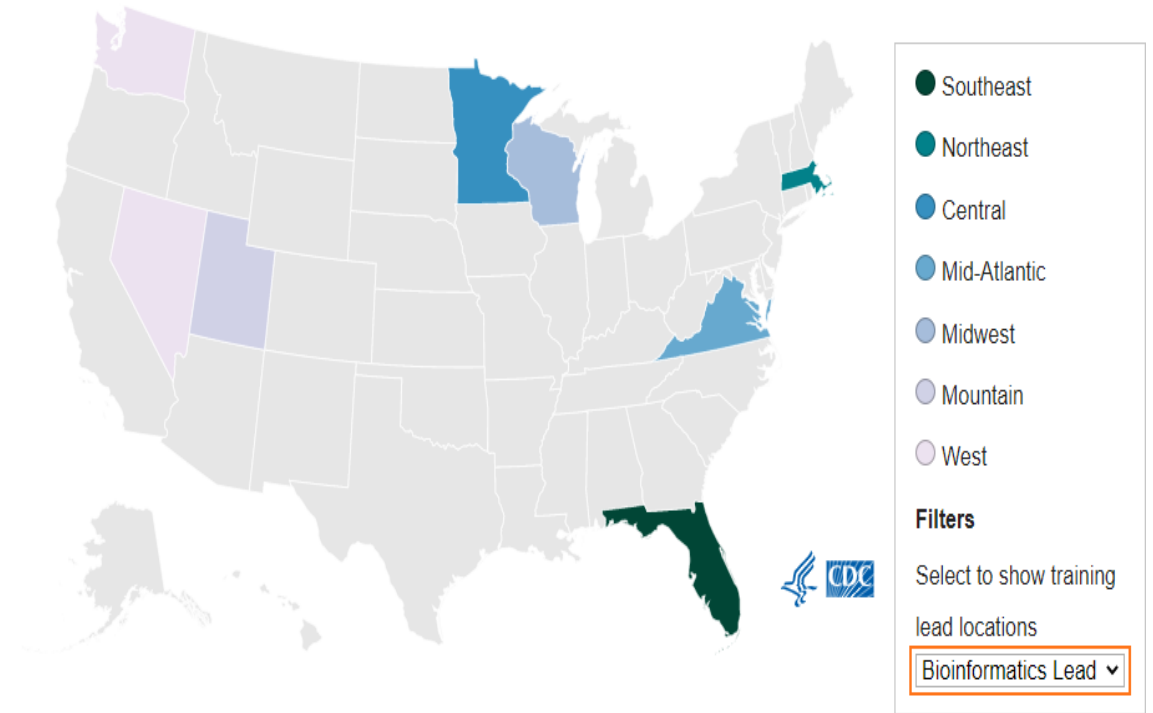
# CDC AMD Regions

- CDC's Advanced Molecular Detection (AMD) program builds and integrates laboratory, bioinformatics, and epidemiology technologies across CDC and nationwide.
- AMD has received funding to implement these technologies in public health programs.
- In 2018, the AMD program established seven workforce development regions across the country. Each region has an AMD training lead and a bioinformatics lead.
- This provides a network of customized AMD support which helps develop skills and provides training assistance to public health labs across the country.
- Florida and Georgia serve as leads for the Southeast Region.



# Bioinformatics Regional Resource Lead

- AMD Regional Bioinformatics Regional Resource Lead acts as a regional consultant.
- Provides support to labs within the region on data analysis & how to interface with IT departments within the region & across regions to help develop national bioinformatics.
- Also helps in implementing or expanding the use of AMD technologies.
- Florida serves as BRR Lead for the Southeast Region.



# BRR Team



Sarah Schmedes, PhD  
Lead Bioinformatician  
Sarah.Schmedes@flhealth.gov



Molly Mitchell  
Bioinformatician  
Molly.Mitchell@flhealth.gov



Lakshmi Thsaliki, MS  
Bioinformatician  
Lakshmi.Thsaliki@flhealth.gov

For any BRR requests, please send an email to [bphl16bioinformatics@flhealth.gov](mailto:bphl16bioinformatics@flhealth.gov)  
Based on the requests one of us from the team will respond as soon as possible.

# Our Expertise

- Sequencers
  - ISeq, MiSeq, NextSeq, PacBio, ClearLabs, and MinION
- NGS Workflows and Pipelines (from wet-lab to bioinformatic analyses)
  - Bacterial WGS and viral targeted amplicon (clinical and wastewater)
- Expertise in pipeline development, high-performance computing, tool evaluation, validation, benchmarking, testing,  $\beta$ -testing, phylogenetic evaluations, data visualizations, genomic epidemiology, and Ad-hoc analysis

# Our Services

## Consultations

- Public Health Bioinformatics
- Bioinformatics Concepts
- Data management
- Data analysis strategy and implementation

## Data Analysis & Pipeline Development

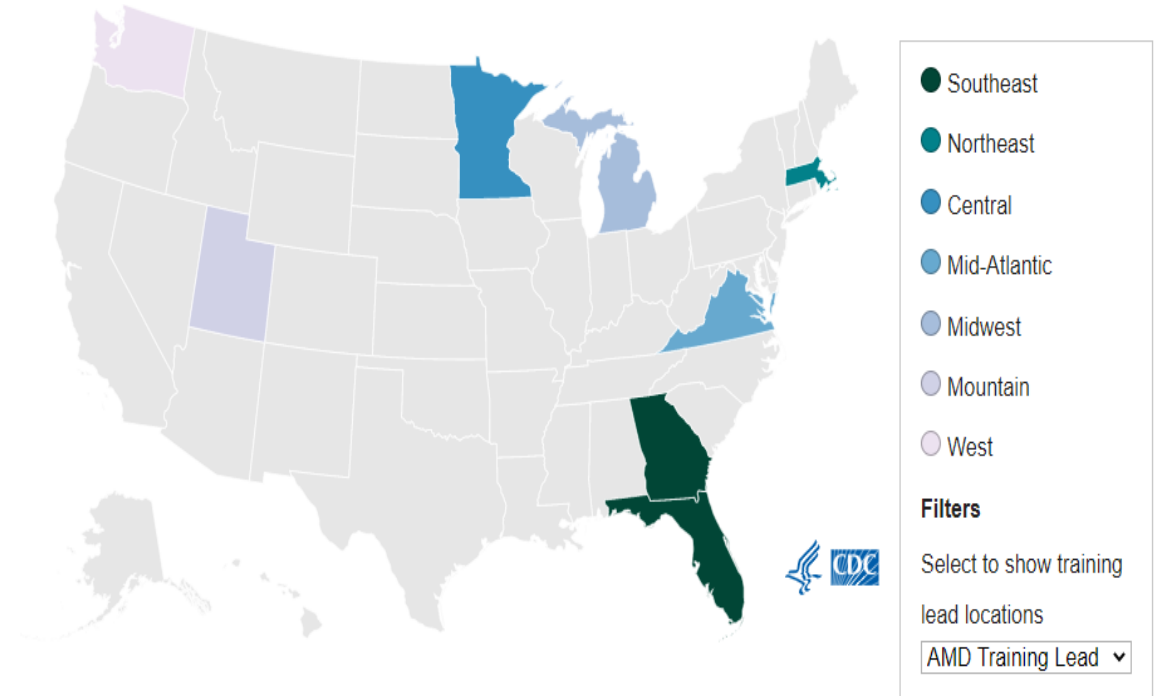
- WGS, targeted amplicon, metagenomics
- Linux & command-line
- Containerization
- Coding, bash scripting, R, Python, Perl, Nextflow

## Training and Support

- One-on-one calls
- Webinars/Training calls (Teams/Zoom)
- Site visits

# AMD Training Lead

- AMD Regional Workforce Development Training Leads provide support to labs within the region and across regions on cross cutting AMD training to help staff develop the critical skills necessary to extract , analyze, and interpret sequencing data.
- Regional training may incorporate local or regional resources or collaboration with academic institutions.
- Georgia & Florida are the training leads for the Southeast Region.

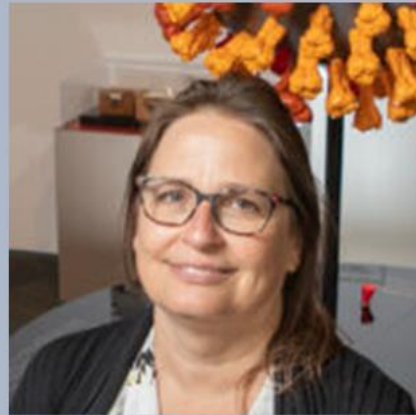




# Georgia AMD BTL Team



**Tatyana Kiryutina, MS**  
Bioinformatics AMD Manager  
<tatyana.kiryutina@dph.ga.gov>

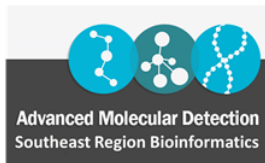


**Tonia Parrott, PhD, HCLD [ABB]**  
Deputy Director of GA-PHL  
<Tonia.Parrott@dph.ga.gov>



**Vijay Reddy, PhD**  
Sr. Bioinformatician (CDC Consultant)  
<vijay.reddy@dph.ga.gov>

The [Applied Bioinformatics Laboratory \(ABiL\)](#) is a collaboration between [IHRC, Inc.](#), [ASRT, Inc.](#), and the [GaTech bioinformatics program](#).



**Georgia Institute of Technology**

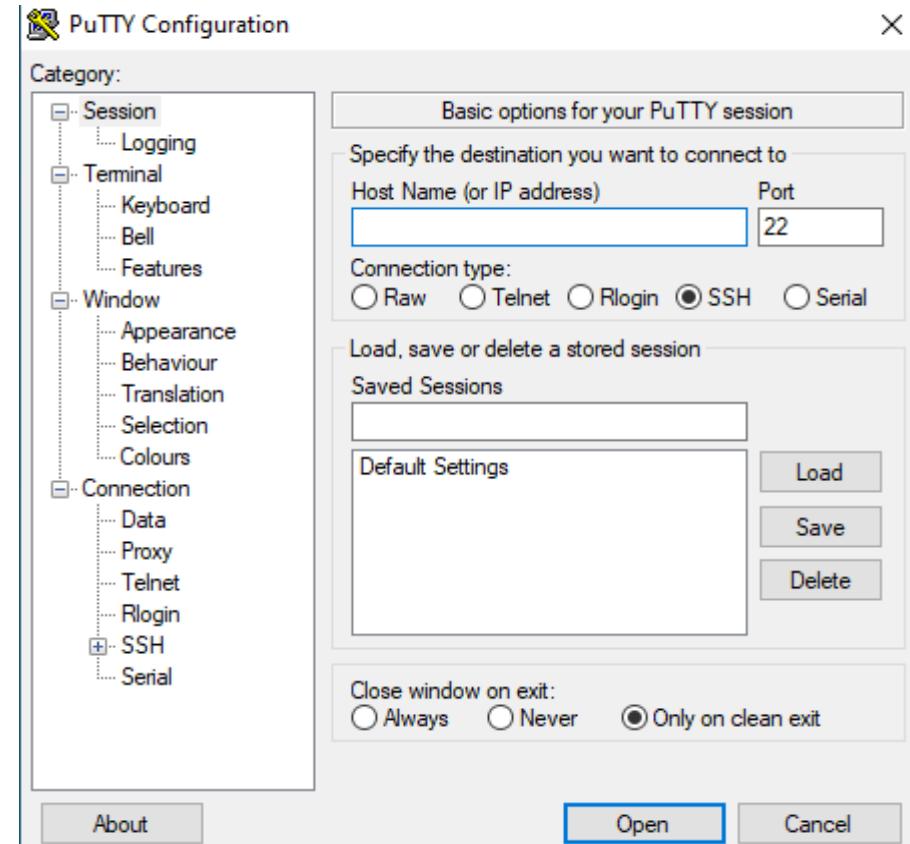
A new Bioinformatics Data Specialist will be joining our GA BTL team in the future!

# What is HiPerGator?

- HiPerGator is a powerful supercomputer unveiled by University of Florida which includes the latest generation of processors and offers nodes for memory-intensive computation.
- HiPerGator can be accessed through Linux command line via ssh, or through galaxy.
- Work on HiPerGator is submitted to SLURM scheduler to run in the batch system when resources are available.
- Compute capacity is sold in NCUs which is 1 CPU core & 7.8GB RAM
- **Storage:** Blue for high performance file system and orange for low performance file system; can be purchased in terabyte quantities.
- **Southeast Region jurisdictions receive computational resources for FREE (supported by BRR funding from CDC ELC)!**

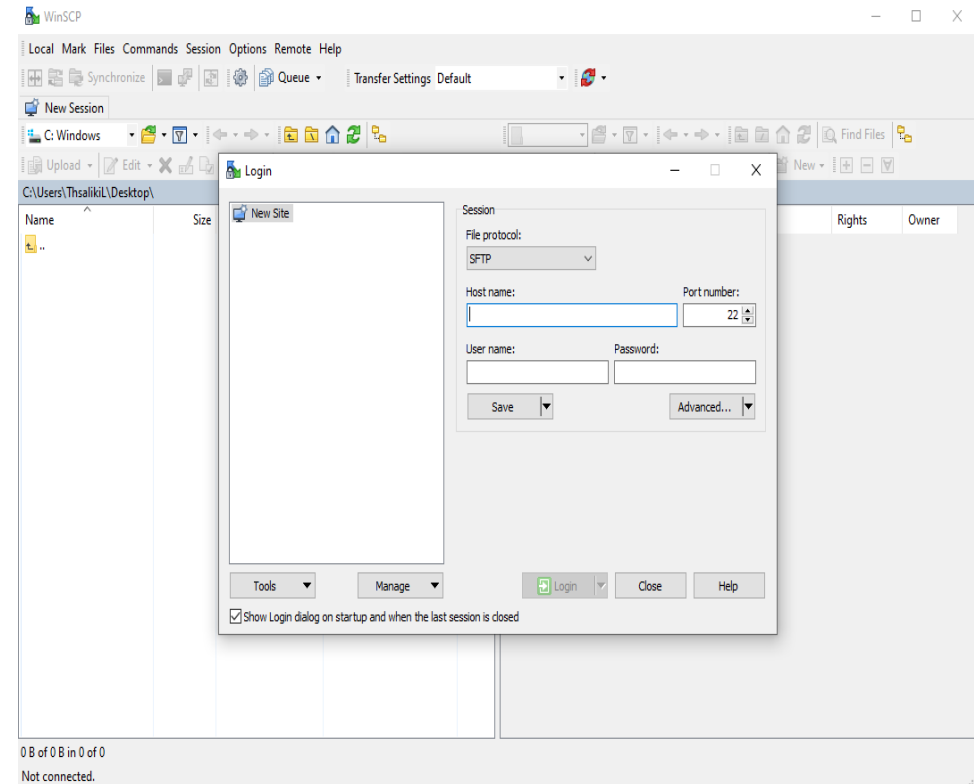
# Access to HPG

- Every person should have their own HPG account to perform data analysis.
- Detailed instructions on how to request & access your HiPerGator account is available in the attached link ([HiPerGator Account Access PDF](#))
- There are various software programs to access HPG through your Windows computer which may need to be installed by your IT department to access HPG.



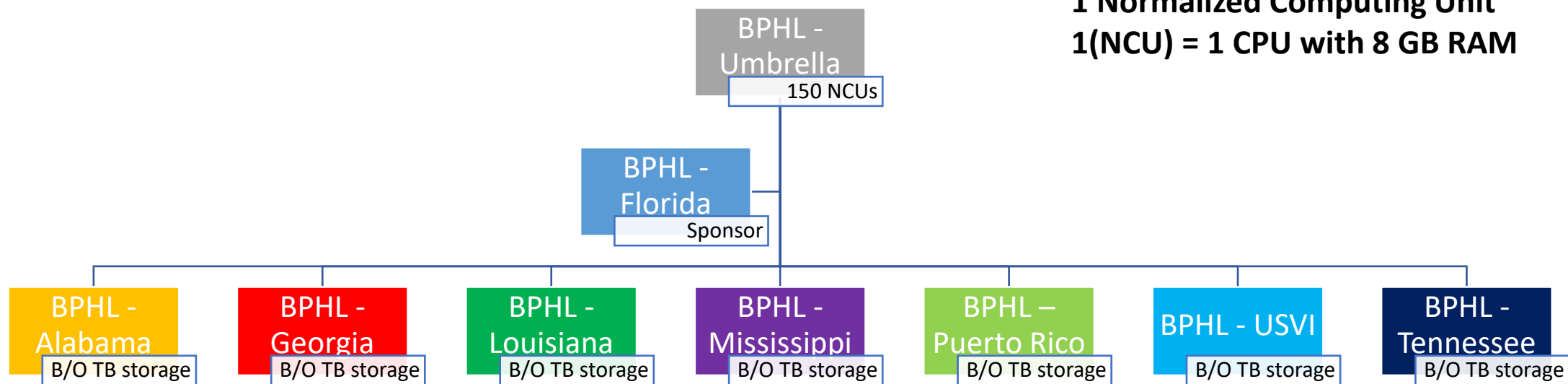
# Access to HPG

- Putty allows you to login to HPG and work in the command-line via a terminal.
- WinSCP allows you to transfer files from your local Windows computer to HPG.
- Putty & WinSCP are free software that are usually approved for use by most IT departments but there are other programs besides these that work the same.
  - You can always discuss those options with your IT department.

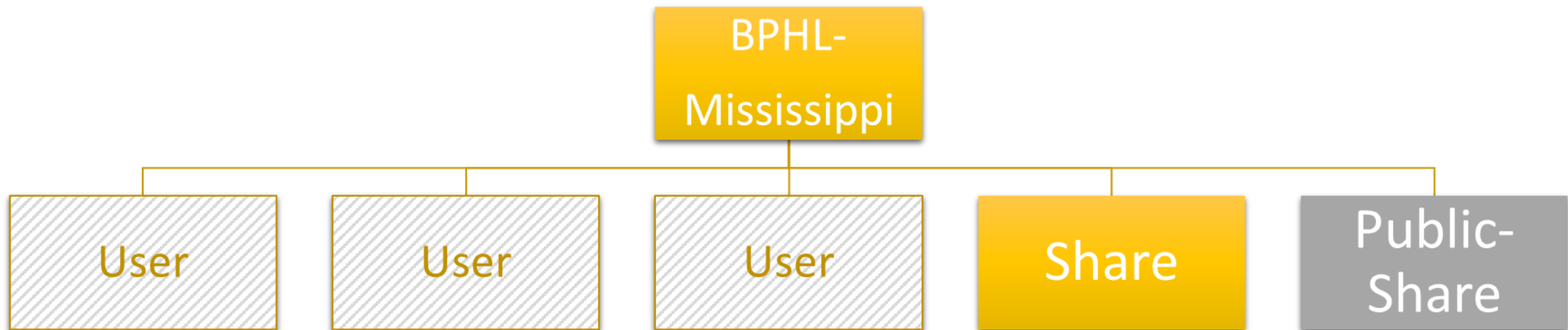





# HiPerGator State Groups

**1 Normalized Computing Unit**  
**1(NCU) = 1 CPU with 8 GB RAM**



# HiPerGator State Groups



-  Data/files private to group
-  Data/files private to group & BPHL-Florida
-  Data/files private to user

# Storage on HPG

- **Home storage:** /home/<user>
  - 40 GB limit
  - scripts, code, compiled applications
- **Blue storage:** /blue/<group>/<user>; short-term, high-performance storage
- **Orange storage:** /orange/<group>/; long term storage
- To check your storage, you can navigate to the command line and enter  
blue\_quota  
orange\_quota
- Orange and Blue storage quotas are at the group level and based on investment.
- Also, if you have data that you are not actively using on HiPerGator, make sure you move the data to orange storage or to separate storage off HiPerGator.
- As every lab is different, make sure your laboratory has a long-term data storage plans like **cloud, server or hard drive backups** that follow your lab's data retention policies.

# Interactive Sessions

- For testing out tools in the command line that require more resources like space and time, you can use the below command

```
srun --qos=bphl-umbrella --account=bphl-umbrella --cpus-per-task=<number of cpus> --  
mem=<memAmount>gb --time=<time limit> --pty bash -i
```

- Interactive work (other than managing jobs) should be done on a login node when the following are true:
  - No more than 16 cores
  - No longer than 10 min (wall time)
  - No more than 64 GB of RAM
- Otherwise, you need to start an interactive session, as described above, which will allow you to work interactively on a compute node.



# HiPerGator Trainings

- Introduction to HPC & HiPerGator
- Introduction to Linux – Part 1
- BaseSpace Command Line Interface
- Introduction to Linux – Part 2
- Compute environment setup & software install
- Individual pipeline trainings

Please let us know how we could improve these trainings.

Updated trainings will be shared on the [StaPH-B/Southeast-region GitHub](#)

# HiPerGator Updates

- Please note:
  - Singularity is now apptainer (<https://apptainer.org/>)
  - Now when you want to load the module that was singularity, use:

**module load apptainer**

- Now you can continue as normal

# Pipelines

## **FLAQ-AMR (FLAQ – Antimicrobial Resistance)**

- Generates annotated, *de novo* assemblies and quality metrics for WGS of bacterial species.
- Determines species ID, serotyping using MLST schemes from PubMLST.
- Identifies AMR genes, virulence genes, and plasmids.

## **FL-cgSNP (Core-genome SNP analysis) - reference free method**

- Pan-genome analysis & generates a multiple sequence alignment
- Identifies pair-wise SNPs between isolates and outputs a pairwise SNP matrix and phylogenetic tree.

## **HqSNP (High-Quality SNP analysis) - reference free method**

- Uses CDC's Lyveset pipeline to identify hqSNPs
- Generates a pairwise SNP matrix and maximum-likelihood phylogenetic tree.

# Pipelines

- Targeted Amplicon Variant Calling and Consensus Sequence Generation – identify variants and generate a consensus sequence for HIV genotyping (will work for any target/amplicon – just need to provide a reference sequence).

## **FLAQ-SC2 (FLAQ-SARS-CoV-2)**

- Generates SARS-CoV-2 consensus assemblies from ARTIC targeted amplicon sequencing using Illumina (e.g., Nextera XT or DNA Prep) and non-Illumina (e.g., Tailed) library prep.
- Outputs variant file and final report with quality metrics (including a PASS/FAIL quality flag based on public repository submission criteria). Automatically generates flags if indels or internal stop codons are present, indicating the need for manual review, based on NCBI's VADR annotation tool.
- Individual scripts also are available to prepare and format assemblies for batch submissions to GISAID and NCBI's Genbank, and to remove indels/SNPs that are likely PCR or sequencing artifacts/errors prior to submission).

# Pipelines

## **FLAQ-SC2-meta**

- SARS-CoV-2 metagenomics pipeline developed primarily for wastewater samples
- Automatically runs the Freyja tool to estimate lineage populations

## **FLAQ-SC2-clearlabs**

- FL BPHL's SARS-CoV-2 Clearlabs analysis pipeline
- Secondary analysis pipeline using Clearlabs data outputs to generate a report with assembly metrics, pangolin lineage, and VADR flags

## **FLAQ-MPX**

- FL BPHL's Monkeypox analysis pipeline for clinical specimens
- Generates consensus assembly from tiled-amplicon data and variant calling

**Custom scripts and pipelines as requested or needed.**

# Pipeline Updates - Nextflow Versions (Under Active Development)

## **Daytona**

- A Nextflow version of FLAQ-SC2 and FLAQ-SC2 Clearlabs

## **Sanibel**

- A Nextflow version of the FLAQ-AMR pipeline including additional subtyping & serotyping tools specific for the species

# Pipeline Updates - Nextflow Versions (Under Active Development)

## Talbot

- A Nextflow version of our FL-cgSNP pipeline

## Enterovirus\_assembler

- A Nextflow pipeline for *de novo* assembly of Enterovirus, Rhinovirus, Parechoviruses using *fastq* data
- *fasta* pipeline outputs can be used as input to CDC's PiType tool ([Picornavirus Typing Tool \(cdc.gov\)](https://www.cdc.gov/pitype/))

# Analysis Request and/or support

- Email [bphl16bioinformatics@flhealth.gov](mailto:bphl16bioinformatics@flhealth.gov)
- One-on-one Teams sessions to go through each New HPG User Training, including a walk-through to set up your HPG environment and hands-on experience with each pipeline/script needed
- Custom pipeline/script development as requested
- All pipelines & scripts will be shared via GitHub ([BPHL-Molecular · GitHub](#)) or the public- share directory in HiPerGator via /blue/bphl-<state>/public-share/directory
- All pipelines are currently being converted to Nextflow workflows which are compatible to work on and off HiPerGator



# Pipeline Updates from CDC

## **PHoeNix pipeline (Portable Healthcare Nextgen Informatics pipeline)**

- This is the replacement of Quaisar-H pipeline.
- A short-read pipeline for healthcare-associated and antimicrobial resistant pathogens.
- PHoeNix is a bioinformatics analysis pipeline built using [Nextflow](#), a workflow tool to run tasks across multiple compute infrastructures in a portable manner.
- It uses Docker/Singularity containers making installation trivial and results highly reproducible.
- The use of PHoeNix provides a standardized approach for identifying and characterizing healthcare-associated bacterial pathogens, specifically for public health partners.
- The [Nextflow DSL 2](#) implementation of this pipeline uses one container per process which makes it easier to maintain and update software dependencies.

# PHoeNIx pipeline

PHoeNIx takes in **Illumina paired-end reads** and was designed for use with pathogens causing healthcare-associated bacterial infections. This comprehensive pipeline performs:

- Quality control
- Checks for contamination
- Confirms taxa ID
- Performs sequence typing
- Assembles reads into scaffolds
- Detects antibiotic resistance and hypervirulence genes
- Searches for plasmid markers

Pipeline and documentation can be found at ([GitHub - CDCgov/phoenix](https://github.com/CDCgov/phoenix))

- This pipeline is available to run on Terra, Nextflow tower, CLI, & HiPerGator and is incorporated into the [StaPH-B toolkit](#).

# 2023 AMD Southeast Region Trainings

- 2023 Bioinformatics Trainings in collaboration with [ABiL](#)
  - Training needs assessment (TNA) survey:
    - ~ 60 responses received and reviewed (thanks!)
  - ABiL is working on quotes to include in the current contract. We will begin discussing concrete curriculums, taking your survey responses into account.
- Target training deliverables (with tentative dates)
  1. Access to modules from the existing [ABiL Training services](#).
  2. Intro to NGS and Bioinformatics: develop new online course (~May 2023).
  3. In-person, four-day, advanced bioinformatics workshop in Atlanta, GA (~Fall 2023).

# CDC AMD Platform Development Activities

- The AMD platform "...*integrates the latest next-generation sequencing technologies with bioinformatics and epidemiology expertise across CDC and the nation to help us find, track, and stop disease-causing pathogens faster than ever before.*"
- AMD platform will serve CDC programs & STLT partners by providing a common infrastructure to perform genomic epidemiology and contribute high-quality data to publicly available data repositories.
- The office of AMD has established five communities of practice (CoP) to build processes and tools for relevant interests, concerns, and priorities regarding the AMD platform.

# CDC AMD Platform Development Activities

- The five CoPs:
  - Argile Architecture, Pipeline Development, and Automation
  - Data Modernization
  - IT Security and Privacy
  - Genomic Epidemiology
  - Quality and Standards
- Florida's Domain Leader facilitates collaboration between OAMD and the public health community for the Quality and Standards CoP.

# ELC Funding & Updates

(Program Strategy for BP4)

Program A: Cross-Cutting Epidemiology and Laboratory Capacity

Strategy 1: Enhance Workforce Capacity

- 1f) AMD Regional Workforce Development Training Participant
- Please apply to receive funding to send participants to the in-person workshop in August/September 2023.

Strategy 6: AMD Platform Support

- 6a) Core Activities – Collaborate with OAMD to assess the landscape of genomic epidemiology in public health and provide subject matter expertise for AMD Platform planning (64 awards)

# BRR Trainings and Updates

- There will be an in-depth training on **PHoeNix** - watch for this email
- Quarterly meeting with each lab will be reinstated
- Implementing a bimonthly office hours call with Molly and Lakshmi
  - Hour long call for questions, quick demos, and other issues
  - Look for this email this month

# Links

- StaPH-B GitHub link ([StaPH-B Github](#))
- StaPH-B Southeast Region GitHub link ([StaPH-B Southeast Region](#))
- StaPH-B YouTube ([StaPH-B – YouTube](#))
- BPHL-Molecular GitHub link ([BPHL-Molecular · GitHub](#))





# Advanced Molecular Detection Southeast Region Bioinformatics

## Questions?

[bphl16bioinformatics@flhealth.gov](mailto:bphl16bioinformatics@flhealth.gov)

**Sarah Schmedes, PhD**

Lead Bioinformatician

[Sarah.Schmedes@flhealth.gov](mailto:Sarah.Schmedes@flhealth.gov)

**Molly Mitchell**

Bioinformatician

[Molly.Mitchell@flhealth.gov](mailto:Molly.Mitchell@flhealth.gov)

**Lakshmi Thsaliki, MS**

Bioinformatician

[Lakshmi.Thsaliki@flhealth.gov](mailto:Lakshmi.Thsaliki@flhealth.gov)