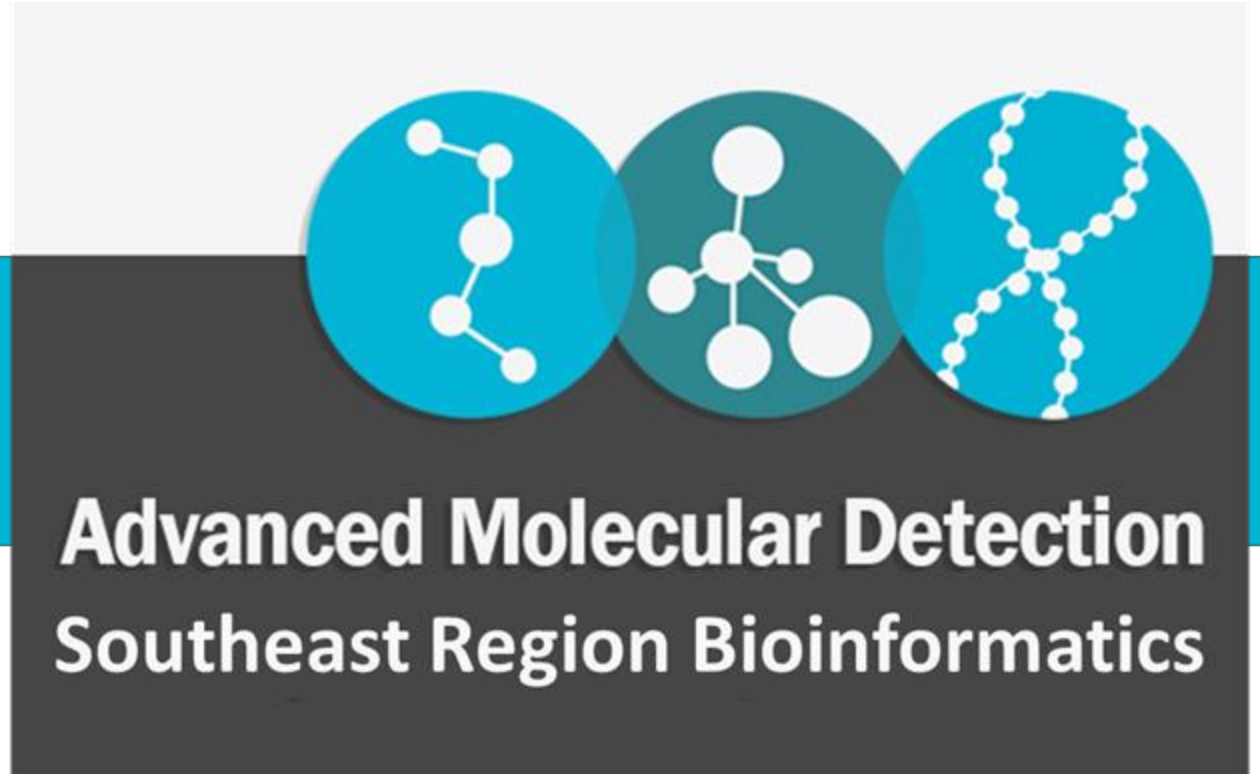


Where to Start Guide

Sam Marcellus, MPH

07/22/2024

Updates

- New Staff Introductions
  - Nikhil Reddy
  - Arnold Rodríguez-Hilario
- On-going Florida computer issues to be resolved soon

Advanced Molecular Detection
Southeast Region Bioinformatics

IT Department Communication

Communicating Needs

Understanding Capabilites

Defining Tasks

Bioinformatics Resources

Communicating with IT: The Basics

- Buy-in from leadership
  - Lab Leadership
  - Upper-Level Agency Leadership
- Identify allies in IT
  - Who is familiar with Linux?
  - Does anyone have experience with Cloud or computing clusters?
  - Who has worked well with you in the past?
- Are there any existing contracts to be aware of?
- Clearly (and repeatedly) state your needs (in writing)

Advanced Molecular Detection
Southeast Region Bioinformatics

Statement of Needs

- Could range from bullet points in an email to a formal memo
- Next-Generation Sequencing has become an essential tool
  - Define NGS
  - List the uses of NGS in your lab
    - Surveillance
    - Clinical
    - Outbreak Investigation
    - Name all the groups or pathogens using NGS
    - Note that uses will only expand in the future
  - Discuss the increase in staff on the lab side to handle sequencing

Advanced Molecular Detection
Southeast Region Bioinformatics

Statement of Needs

- NGS and Bioinformatics have unique computational needs
  - Clouds "attached" to sequencers (BaseSpace)
  - Large amounts of data
  - Need larger than usual computational power
- Sample Tracking
  - Is current system working for you?
  - What's your ideal set up?
  - Can you offer what another state is using as an example?

- Current computing set up and why it's insufficient
  - HiPerGator?
    - It's down two weeks per year
    - Share resources with the region

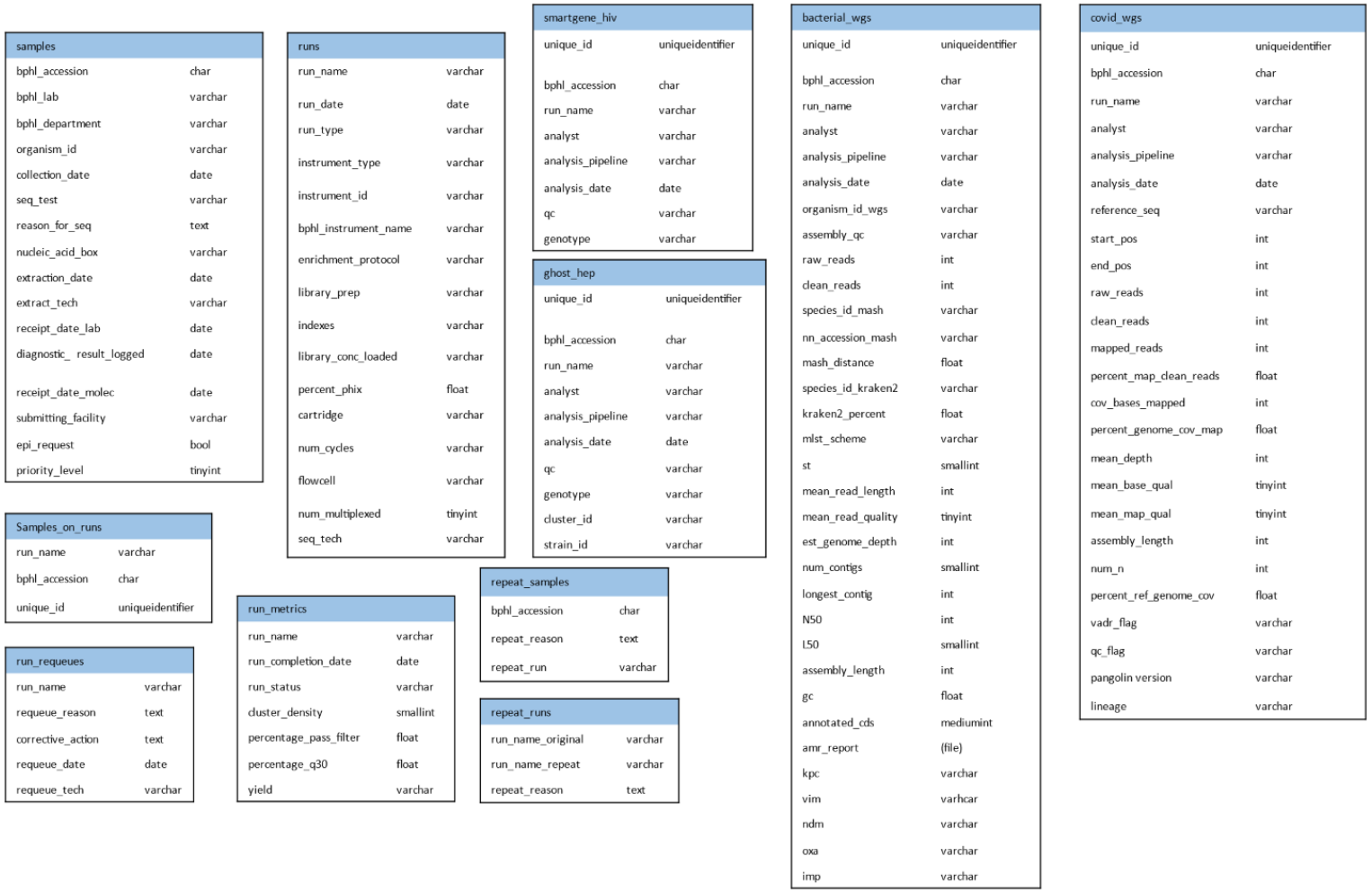
Advanced Molecular Detection
Southeast Region Bioinformatics

Statement of Needs

- List your sequencers and their capacities
  - GB of data per run
  - Samples per run
- List dry lab positions and their responsibilities
- Briefly describe epis and what they do with the sequencing data
- Any useful diagrams can be added as appendices
- Short-, medium-, and long-term goals for your team and its infrastructure

Advanced Molecular Detection
Southeast Region Bioinformatics

Statement of Needs



samples	
bphl_accession	char
bphl_lab	varchar
bphl_department	varchar
organism_id	varchar
collection_date	date
seq_test	varchar
reason_for_seq	text
nucleic_acid_box	varchar
extraction_date	date
extract_tech	varchar
receipt_date_lab	date
diagnostic_ result_logged	date
receipt_date_molec	date
submitting_facility	varchar
epi_request	bool
priority_level	tinyint

runs	
run_name	varchar
run_date	date
run_type	varchar
instrument_type	varchar
instrument_id	varchar
bphl_instrument_name	varchar
enrichment_protocol	varchar
library_prep	varchar
indexes	varchar
library_conc_loaded	varchar
percent_phix	float
cartridge	varchar
num_cycles	varchar
flowcell	varchar
num_multiplexed	tinyint
seq_tech	varchar

smartgene_hiv	
unique_id	uniqueidentifier
bphl_accession	char
run_name	varchar
analyst	varchar
analysis_pipeline	varchar
analysis_date	date
qc	varchar
genotype	varchar

ghost_hep	
unique_id	uniqueidentifier
bphl_accession	char
run_name	varchar
analyst	varchar
analysis_pipeline	varchar
analysis_date	date
qc	varchar
genotype	varchar
cluster_id	varchar
strain_id	varchar

bacterial_wgs	
unique_id	uniqueidentifier
bphl_accession	char
run_name	varchar
analyst	varchar
analysis_pipeline	varchar
analysis_date	date
organism_id_wgs	varchar
assembly_qc	varchar
raw_reads	int
clean_reads	int
species_id_mash	varchar
nn_accession_mash	varchar
mash_distance	float
species_id_kraken2	varchar
kraken2_percent	float
mlst_scheme	varchar
st	smallint
mean_read_length	int
mean_read_quality	tinyint
est_genome_depth	int
num_contigs	smallint
longest_contig	int
N50	int
L50	smallint
assembly_length	int
gc	float
annotated_cds	mediumint
amr_report	(file)
kpc	varchar
vim	varhcar
ndm	varchar
oxa	varchar
imp	varchar

covid_wgs	
unique_id	uniqueidentifier
bphl_accession	char
run_name	varchar
analyst	varchar
analysis_pipeline	varchar
analysis_date	date
reference_seq	varchar
start_pos	int
end_pos	int
raw_reads	int
clean_reads	int
mapped_reads	int
percent_map_clean_reads	float
cov_bases_mapped	int
percent_genome_cov_map	float
mean_depth	int
mean_base_qual	tinyint
mean_map_qual	tinyint
assembly_length	int
num_n	int
percent_ref_genome_cov	float
vadr_flag	varchar
qc_flag	varchar
pangolin version	varchar
lineage	varchar

Samples_on_runs	
run_name	varchar
bphl_accession	char
unique_id	uniqueidentifier

run_requeues	
run_name	varchar
requeue_reason	text
corrective_action	text
requeue_date	date
requeue_tech	varchar

run_metrics	
run_name	varchar
run_completion_date	date
run_status	varchar
cluster_density	smallint
percentage_pass_filter	float
percentage_q30	float
yield	varchar

repeat_samples	
bphl_accession	char
repeat_reason	text
repeat_run	varchar

repeat_runs	
run_name_original	varchar
run_name_repeat	varchar
repeat_reason	text

Advanced Molecular Detection
Southeast Region Bioinformatics

Understanding Capabilities

- Current policies in place
  - IT Policies
  - Data Retention Policies
  - Privacy Policies
  - *If it's not written down, it's not a real policy*
- List of available software and current partnerships
  - Any cloud agreements with other agencies?
  - IT staff familiar with Linux working in a different agency
- Overarching Data Modernization Plans

Advanced Molecular Detection
Southeast Region Bioinformatics

Defining Tasks

Laboratory Tasks

- Day-to-day management of resources
- Data analysis
- Implementation and use of bioinformatic pipelines

IT Tasks

- Install of new resources on state computers
- Data encryption
- Data protection
- Database construction
- Technical Support

Advanced Molecular Detection
Southeast Region Bioinformatics

Bioinformatics Resources

- Trainings helpful for new staff or IT staff interested in bioinformatics
- StaPH-B (State Public Health Bioinformaticians)
  - https://staphb.org/training.html
- CDC Genomic Epidemiology Toolkit
  - https://www.cdc.gov/advanced-molecular-detection/php/training/index.html
- Data Camp
  - https://app.datacamp.com/
  - Data Scientist in R Track
  - Data Analyst in SQL Track

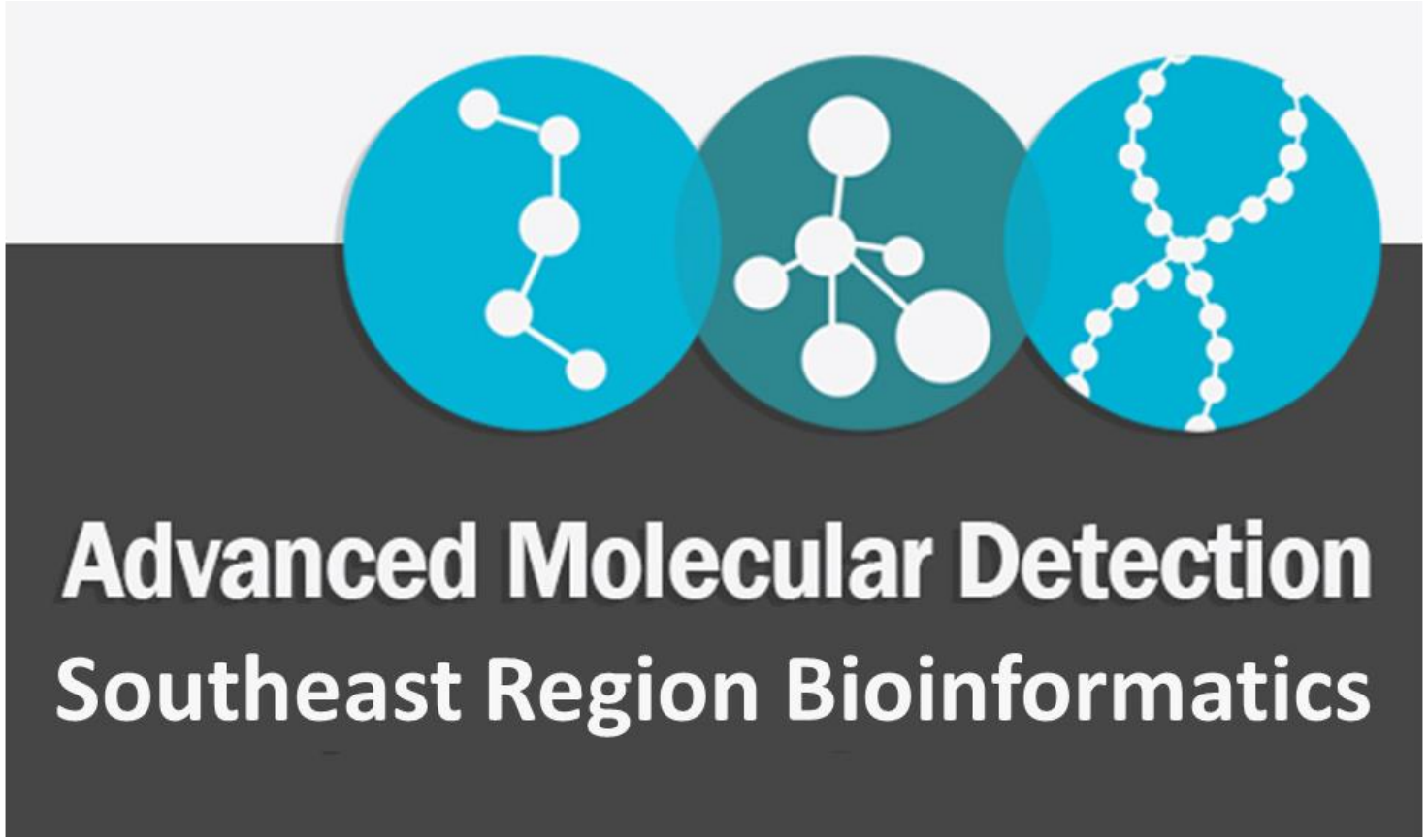
Advanced Molecular Detection
Southeast Region Bioinformatics

Bioinformatics Resources

- PulseNet
  - https://www.cdc.gov/pulsenet/hcp/about/next-gen-wgs.html
  - An Overview of PulseNet Databases
    - https://doi.org/10.1089/fpd.2019.2637

- Bioinformatics in Public Health
  - Libuit KG, Doughty EL, Otieno JR, Ambrosio F, Kapsak CJ, Smith EA, Wright SM, Scribner MR, Petit Iii RA, Mendes CI, Huergo M, Legacki G, Loreth C, Park DJ, Sevinsky JR. Accelerating bioinformatics implementation in public health. Microb Genom. 2023 Jul;9(7):mgen001051. doi: 10.1099/mgen.0.001051. PMID: 37428142; PMCID: PMC10438813.

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