

Building cloud-based bioinformatic capacities – a public health laboratory perspective

Jie Lu, PhD

Texas Department of State Health Services Public Health Laboratory

Team Lead, Bioinformatics



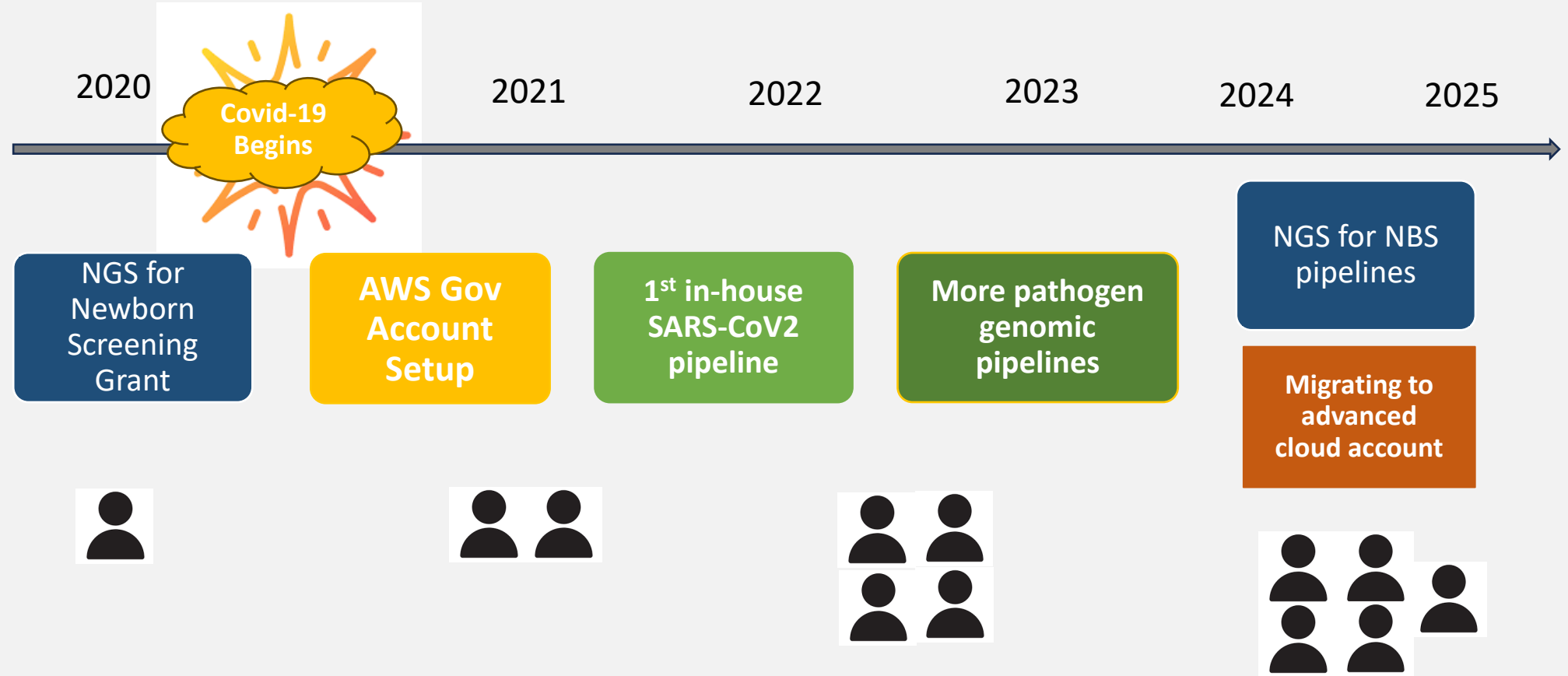
TEXAS
Health and Human
Services

Texas Department of State
Health Services

Conflicts of Interest

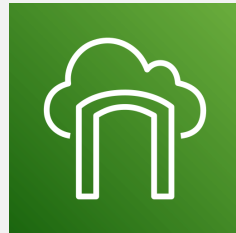
None.

Timeline of the bioinformatic capacity expansion at DSHS



Amazon Web Services (AWS) Government Cloud Overview

High speed date transfer



Direct Connect/Storage
Gateway

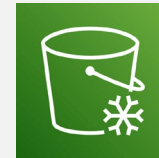
SSL encryption

Unlimited Storage



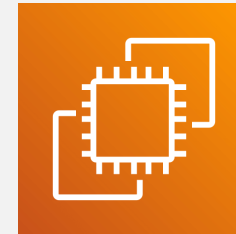
Simple Storage Service
(Amazon S3)

Server-Side Encryption (SSE)



Glacier Deep
Archive

Robust computing power



Amazon EC2 Linux
Instances (Static)

Bioinformatics Pipelines



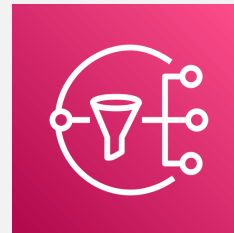
EBS Volume

Real-time Pipeline kickoff



Lambda Function

Seamless communication



Simple Notification
Service (SNS)

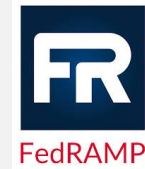
Security and Compliance



AWS Identity and Access
Management (IAM)



Firewall
Security Group



FedRAMP



DoD SRG



Texas Department of State
Health Services

Bioinformatic Pipelines running in house – Reference Based mapping/assembly



- **Cecret** (<https://github.com/UPHL-BioNGS/Cecret>)
Developed and maintained by Dr. Erin Young, Utah Public Health Laboratory.
Sars-CoV-2 sequencing with the [artic](#)/Illumina library prep workflow for MiSeq data (19,000+ since pandemic)
Now switched to Clear Labs long reads sequencing.
Used for Sars-CoV2 wastewater sequencing, Mpox and Measles sequencing as well.
Processed 507 wastewater samples and 201 Mpox samples in 2024.
- **Mycosnp** (<https://github.com/CDCgov/mycosnp-nf>)
Developed and maintained by Centers for Disease Control and Prevention (CDC).
WGS analysis of fungal organisms, including *Candida auris*.
Processed 797 samples in 2024.
- **Varpipeline** (<https://github.com/CDCgov/NCHHSTP-DTBE-Varpipeline-WGS>)
Developed and maintained by Division of TB Elimination, CDC.
WGS analysis of *Mycobacterium Tuberculosis* data.
Influx of samples from VA, TN, AR, AZ, LA and Houston due to recall of specific reagents, preventing PST testing. APHL has rerouted samples to Texas. Processed 1161 samples in 2024.



Texas Department of State
Health Services

Bioinformatic Pipelines running in house – *De novo* Assembly – Species Agnostic

- **PHeoNix** (<https://github.com/CDCgov/phoenix>)

Developed and maintained by CDC

Taxa determination, antimicrobial resistance profiling and MLST typing in Carbapenem-Resistant Organisms (CRO) and now other bacteria as well. **Processed 823 samples in 2024.**

- **Grandeur** (<https://github.com/UPHL-BioNGS/Grandeur>)

Developed and maintained by Dr. Erin Young, UPHL

Core genome alignment and phylogenetic tree construction for sequence relatedness investigation. **Analyzed 22 outbreak cases in 2024.**

- **TaxTriage** (<https://github.com/jhuapl-bio/taxtrriage/>)

Developed and maintained by The Johns Hopkins University Applied Physics Laboratory LLC (funded by CDC)

Metagenomics sequencing within complex samples/specimens (e.g., respiratory swabs, lesion swabs, whole blood) using untargeted DNA or RNA sequencing data



Texas Department of State
Health Services

Critical building blocks – container and virtual environment



Container enables fast deployment of bioinformatics tools. Workflow description language provides improved reproducibility and scalability.

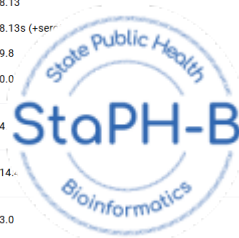


Virtual environment allows for isolated project environments, ensuring each project has its own set of dependencies and package versions, preventing conflicts when multiple pipelines on the same server need different software versions

State Public Health Bioinformatics Community

Available Docker images

Software	Version	Link
ABRicate docker pulls: 27k	0.8.7 0.8.13 0.8.13s (+ser) 0.9.8 1.0.0	https://github.com/tseemann/abricate
any2fasta docker pulls: 738	0.4	https://github.com/tseemann/any2fasta
ARIBA docker pulls: 1k	2.14.0	https://github.com/sanger-pathogens/ariba
artic-ncov2019 docker pulls: 71k	1.3.0	https://github.com/artic-network/fieldbioinformatics
artic-ncov2019-epi2me docker pulls: 44k	0.3.10	https://github.com/epi2me-labs/wf-artic



Github

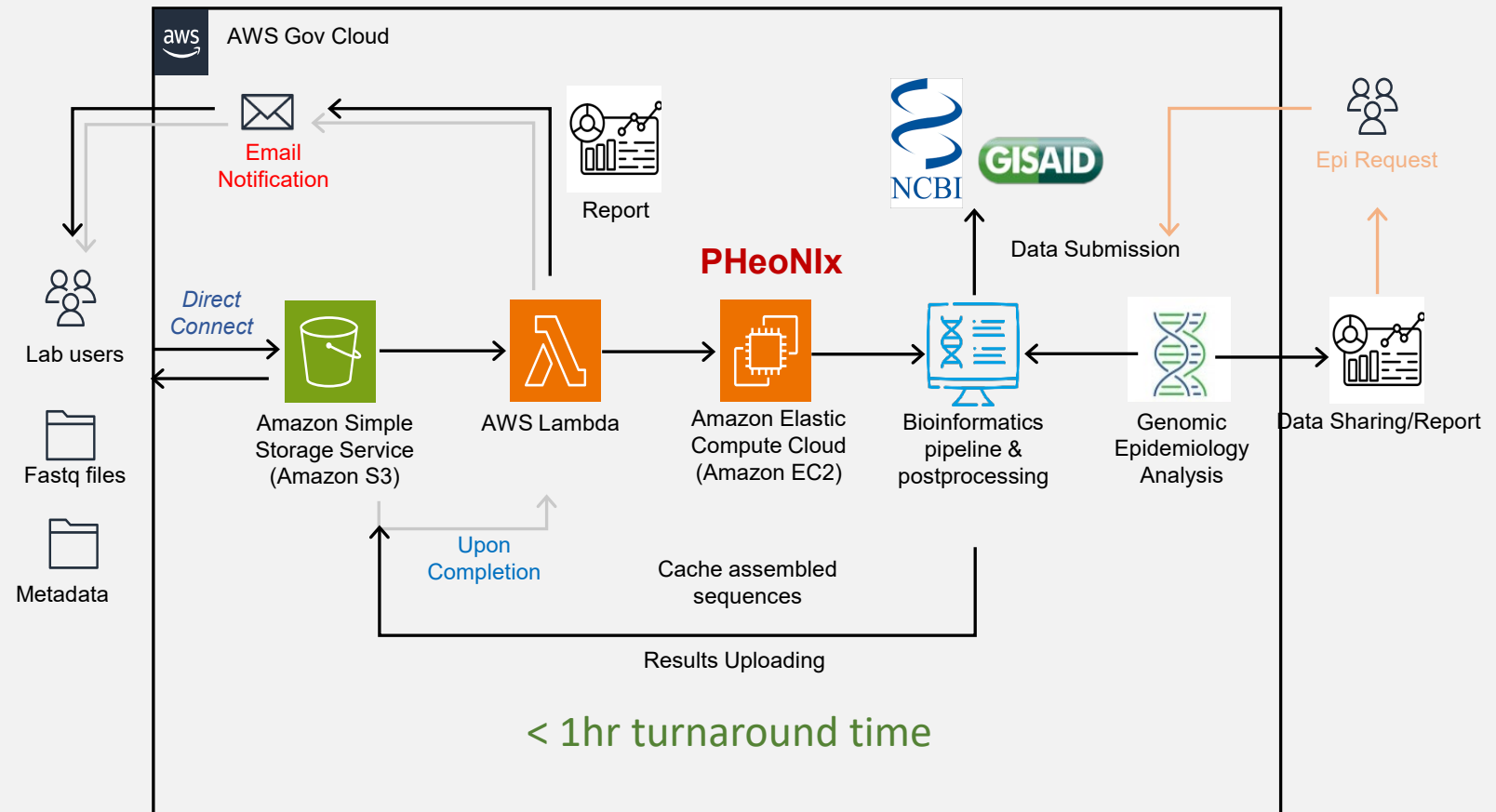


version control and collaborative programming



Texas Department of State Health Services

Automated Bioinformatic Workflow for CRO WGS



Automated Bioinformatic Workflow for CRO WGS



Direct Connect/Storage Gateway



Simple Storage Service (Amazon S3)

AWS email notifications upon data upload and completion

AWS Notification Message



DSHS Bacterial WGS Notification <no-reply@sns.amazonaws.com>
To: Lu, Jie (DSHS)

Fri 12/6/2024 10:23 AM

WARNING: This email is from outside the HHS system. Do not click on links or attachments unless you expect them from the sender and know the content is safe.

ARLN run AR_241205B_VH00729 was uploaded to 804609861260-bioinformatics-infectious-disease successfully. Pipeline is starting.

Batch QC and AMR profiling report

WGS_id	Species	blaKPC	blaNDM	blaOXA-48	Auto_QC_Outcome	Auto_QC_Failure_Reason	Estimated_Coverage	Genome_Length	Assembly_Ratio(%)	#_of_Scaffolds>50k	GC_%
2025CB-00117	Acinetobacter baumannii	NOT_DETECTED	NOT_DETECTED	NOT_DETECTED	PASS	missing	90.18	3817789	0.9575 (NA)	62	38.92
2025CB-00118	Citrobacter arsenatis	NOT_DETECTED	DETECTED	NOT_DETECTED	PASS	missing	79.56	5372980	1.0005 (NA)	33	51.91
2025CB-00119	Staphylococcus aureus	NOT_DETECTED	NOT_DETECTED	NOT_DETECTED	PASS	missing	134.3	2723787	0.9629 (NA)	21	32.66
2025CB-00120	Staphylococcus aureus	NOT_DETECTED	NOT_DETECTED	NOT_DETECTED	PASS	missing	162.8	2714061	0.9595 (NA)	27	32.63
2025CB-00121	Neisseria meningitidis	NOT_DETECTED	NOT_DETECTED	NOT_DETECTED	PASS	missing	149.1	2036566	0.9501 (NA)	102	51.97
CONPA0729250213	Pseudomonas aeruginosa	NOT_DETECTED	NOT_DETECTED	NOT_DETECTED	PASS	missing	88.35	6805977	1.0237 (NA)	88	66.08
CONVP0729250213	Vibrio parahaemolyticus	NOT_DETECTED	NOT_DETECTED	NOT_DETECTED	PASS	missing	96.77	5235049	1.0098 (NA)	61	45.2

metadata/attribute tables for NCBI submission

A	B	C		H	I	J	K	L	M
*sample_id	sample_title	bioproject_accession	*organism	platform	instrument	design_description	filetype	filename	file_name
2025CB-00117		PRJNA288601	Acinetobacter	ILLUMINA	NextSeq 2	ILLUMINA D	fastq	2025CB-00117	2025CB-00117
2025CB-00118		PRJNA288601	Citrobacter	ILLUMINA	NextSeq 2	ILLUMINA D	fastq	2025CB-00118	2025CB-00118
2025CB-00119		PRJNA533550	Staphylococcus	ILLUMINA	NextSeq 2	ILLUMINA D	fastq	2025CB-00119	2025CB-00119
2025CB-00120		PRJNA533550	Staphylococcus	ILLUMINA	NextSeq 2	ILLUMINA D	fastq	2025CB-00120	2025CB-00120
2025CB-00121		PRJNA1170207	Neisseria	ILLUMINA	NextSeq 2	ILLUMINA D	fastq	2025CB-00121	2025CB-00121

Next-Generation Sequencing (NGS) Analysis Report



TEXAS
Health and Human
Services

**Texas Department of State
Health Services**

Report date: 2025-02-14

WGS ID: 2025CB-00117

TX DSHS ID: AMD2500193

Species ID: *Acinetobacter baumannii*

MLST: ST195.ST1816-PARALOG

Table: Antimicrobial Resistance genes identified

Gene_symbol	product_name
Mbl	Description not found
Zn-dependent	Description not found
abaF	fosfomycin efflux MFS transporter AbaF
armvA	multidrug efflux MFS transporter ArmvA
ant(3'')-IIa	aminoglycoside nucleotidyltransferase ANT(3'')-IIa
aph(6)-Id	aminoglycoside O-phosphotransferase APH(6)-Id
blaADC-73	extended-spectrum class C beta-lactamase ADC-73
OXA-23	OXA-23 family carbapenem-hydrolyzing class D beta-lactamase OXA-23
blaOXA-66	OXA-51 family carbapenem-hydrolyzing class D beta-lactamase OXA-66
blaTEM-1D	Description not found
strA	Description not found
tet(B)	tetracycline efflux MFS transporter Tet(B)



TEXAS
Health and Human
Services

Texas Department of State Health Services

Limitations of government cloud account

- Government agencies must adhere to complex regulations and compliance standards, which can limit the features and flexibility of cloud services.
- Permission is limited and gradually taken away after Sars-CoV2 for AWS services such as Lambda function, SNS, etc. Automation becomes impossible to implement.



Advanced Cloud Account (New and Improved)

- Advanced cloud account (ACA) allows permission to most AWS features to build a more advanced architecture to facilitate automation and enable scalability.
- The migration of pipelines is complicated. Some of the measures we used to accelerate this process: reduce dependencies and hard coding, create virtual environments, and etc.
- Implement infrastructure-as-code using Terraform to deploy the pipelines automatically in the new account.



Texas Department of State
Health Services

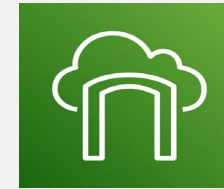
Shared responsibility model for cloud-based system

- Lab is responsible for architecture design, function implementation, code development, operating system management, account administration, application-level support.
- The Public Cloud Manager contracted via Texas DIR is responsible for provisioning, installing, and patching the operating system; installing and maintaining antivirus; and performing SIEM logging and critical watch reporting.
- Cloud solution architect of the central IT department assists to design the architecture and implement infrastructure-as-code.

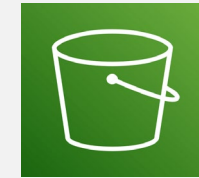


Work in progress -- NGS workflow for Newborn Screening

- The robust cloud-based infrastructure resulted from building pathogen genomic bioinformatic capacities has laid a solid foundation for NGS for Newborn Screening Program in Texas.
- We have explored the feasibility of using whole genome sequencing as well as targeted sequencing.
- We are building a hybrid solution for NGS workflow combining our EC2 instances, Illumina Basespace, and AWS HealthOmics in the ACA with more advanced architecture and automation settings.



Direct
Connect/Storage
Gateway



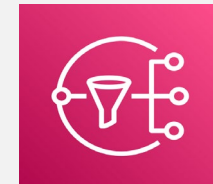
Simple Storage
Service (Amazon
S3)



Lambda



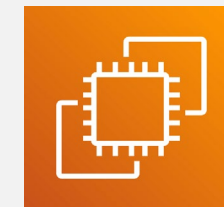
Step Functions



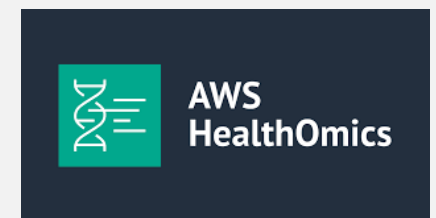
Simple
Notification
Service (SNS)



Illumina



EC2 instances



Acknowledgements



DSHS Bioinformatics Team

Jie Lu
Richard Bovio
Megan Partridge
Jessica Respress

Microbiology Unit

Yan Sun
Bonnie Oh
Maliha Rahman

Laboratory Leadership

Grace Kubin
[Deputy Commissioner,](#)
[Laboratory Director](#)
Susan Tanksley
[Deputy Laboratory Director](#)
Charles Hooks
[Laboratory Informatics Director](#)
Casey Schroeder
[Microbiology Unit Director](#)

HHSC IT Group

Scott Miller
Preeti Baranga

State Public Health Bioinformatics Community

Mountain Region Bioinformatics Resource



Thank you!

Building cloud-based bioinformatic capacities – a public health laboratory perspective

Jie Lu, PhD

Jie.lu@dshs.texas.gov

The many hats Bioinformaticians wear – continuing education

- Cloud architect – design and manage the cloud-based systems. Configure cloud infrastructure and implement automation functions.
- Programmer – develop customized data processing pipelines and continuous integration and continuous delivery (CI/CD) of existing pipelines to adapt to version updates.
- Linux administrator – Installing and maintaining applications, monitoring performance, managing user accounts and writing automation scripts.
- Data scientist – analyze, visualize and interpret DNA sequencing data to provide insights into genomic epidemiology.

