Data Modernization: Improving the usefulness of genomic data

Kelsey Florek, PhD, MPH
Senior Genomics and Data Scientist
Wisconsin State Laboratory of Hygiene

May 21, 2024

Slides live at:
www.k-florek.net/talks

Wisconsin State
Laboratory of Hygiene
UNIVERSITY OF WISCONSIN–MADISON

Supported By

AWS Diagnostic Development Initiative (DDI)

1. **Necessities of Next Generation Sequencing Capacity Building**
2. Blueprints for an NGS Data Solution
3. Simplifying Genomics for Public Health Partners

Expanding Genomic Sequencing Capacity

Pre SARS-CoV-2 Pandemic

- 4x Illumina MiSeq
- 1x ONT MinION

Expanding Genomic Sequencing Capacity

Pre SARS-CoV-2 Pandemic

- 4x Illumina MiSeq
- 1x ONT MinION

Post SARS-CoV-2 Pandemic

- 4x Illumina MiSeq
- 2x NextSeq 2000
- 1x ONT GridION
- 1x Eppendorf epMotion
- 1x Tecan Fluent 780 NGS Dream Prep

Expanding Genomic Sequencing Capacity

Pre SARS-CoV-2 Pandemic

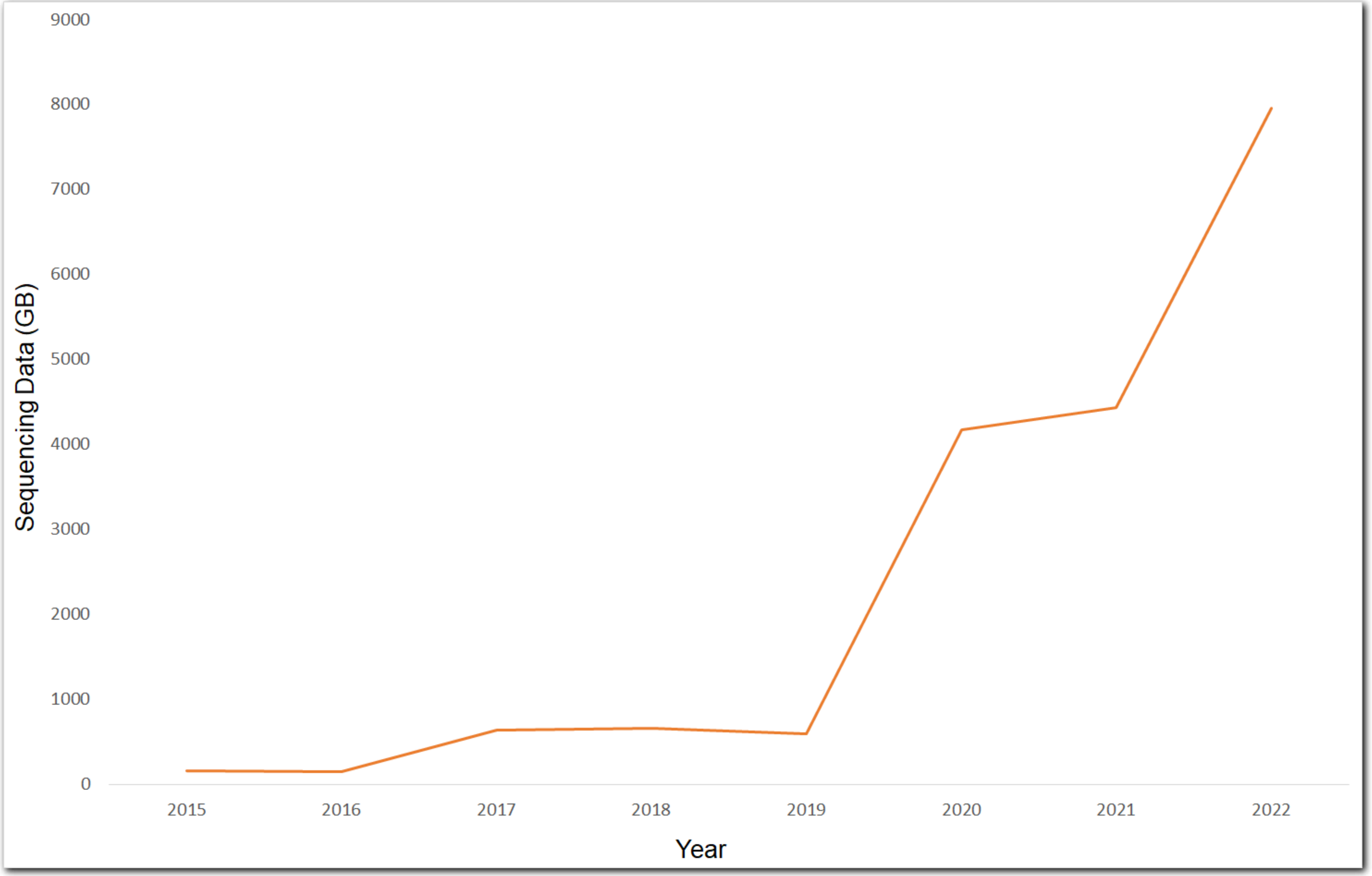
- 4x Illumina MiSeq
- 1x ONT MinION

Post SARS-CoV-2 Pandemic

- 4x Illumina MiSeq
- 2x NextSeq 2000
- 1x ONT GridION
- 1x Eppendorf epMotion
- 1x Tecan Fluent 780 NGS Dream Prep

Over 900% increase in sequencing data generation capacity

NGS Data Storage



Improvements in Analytical Approaches

Old Approach

- Entirely Python Based
- Limited logging and fault tolerance
- Required installing complicated and often conflicting dependencies

Improvements in Analytical Approaches

Old Approach

- Entirely Python Based
- Limited logging and fault tolerance
- Required installing complicated and often conflicting dependencies

New Approach

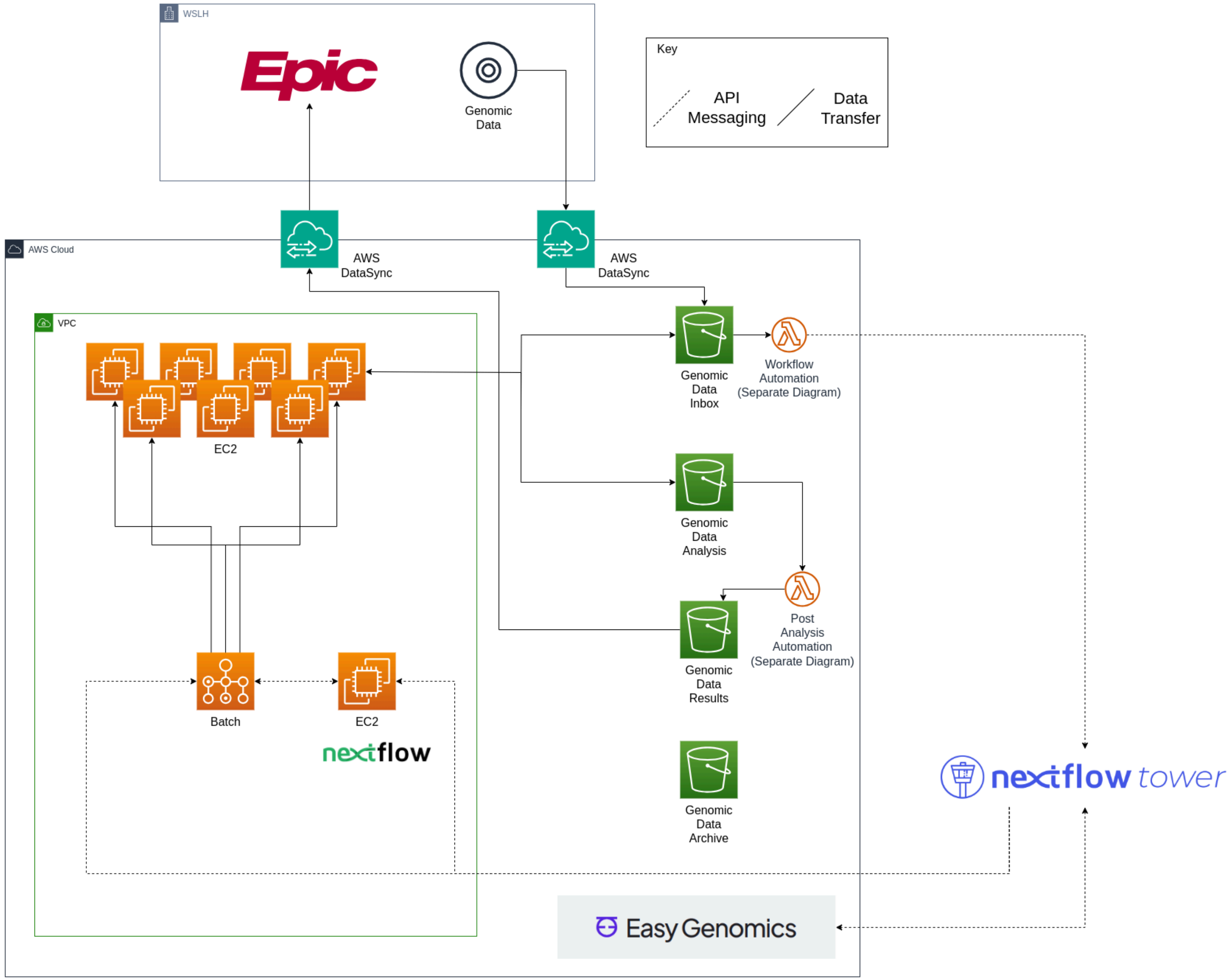
- Nextflow - Nf-Core Based
- Containerized Steps
- Detailed Logging
- Compatible with a variety of Cloud and HPC environments
- Supports a high degree of job parallelization and horizontal scalability

1. ~~Necessities of Next Generation Sequencing Capacity Building~~
2. **Blueprints for an NGS Data Solution**
3. Simplifying Genomics for Public Health Partners

Bioinformatics analytical infrastructure

- Highly scalable and capable of managing burst data
- Highly reliable and fault tolerant
- Cost effective
- Adaptable to changing needs
- Detailed logging and traceability

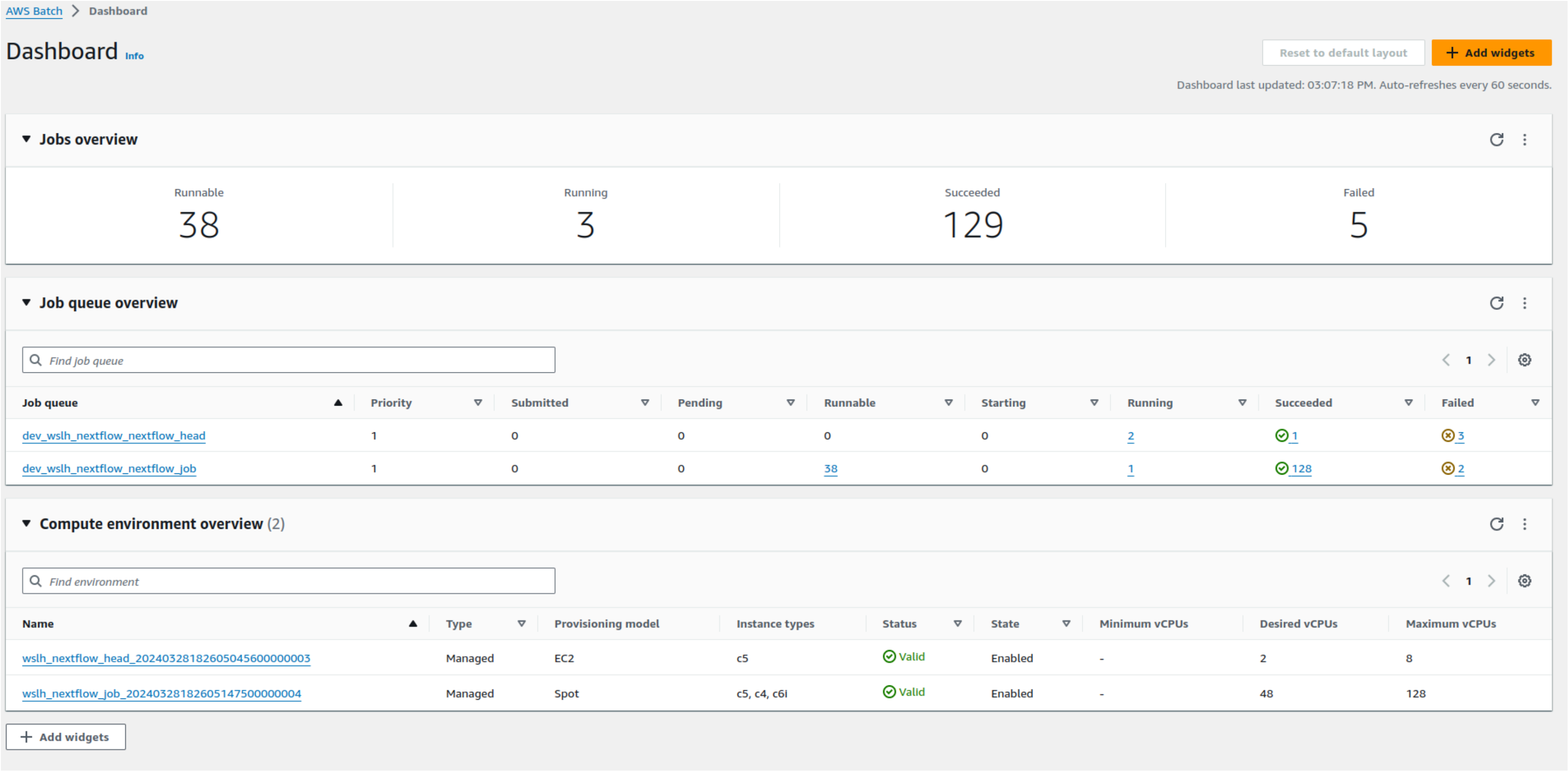
WSLH Bioinformatics Analytical Infrastructure



AWS Batch

AWS Batch automatically provisions compute resources and optimizes the workload distribution based on the quantity and scale of the workloads.

AWS Batch



AWS Batch > Dashboard

Dashboard Info

Dashboard last updated: 03:07:18 PM. Auto-refreshes every 60 seconds.

Jobs overview

Runnable	Running	Succeeded	Failed
38	3	129	5

Job queue overview

Find job queue

Job queue	Priority	Submitted	Pending	Runnable	Starting	Running	Succeeded	Failed
dev_wslh_nextflow_nextflow_head	1	0	0	0	0	2	✓ 1	⊗ 3
dev_wslh_nextflow_nextflow_job	1	0	0	38	0	1	✓ 128	⊗ 2

Compute environment overview (2)

Find environment

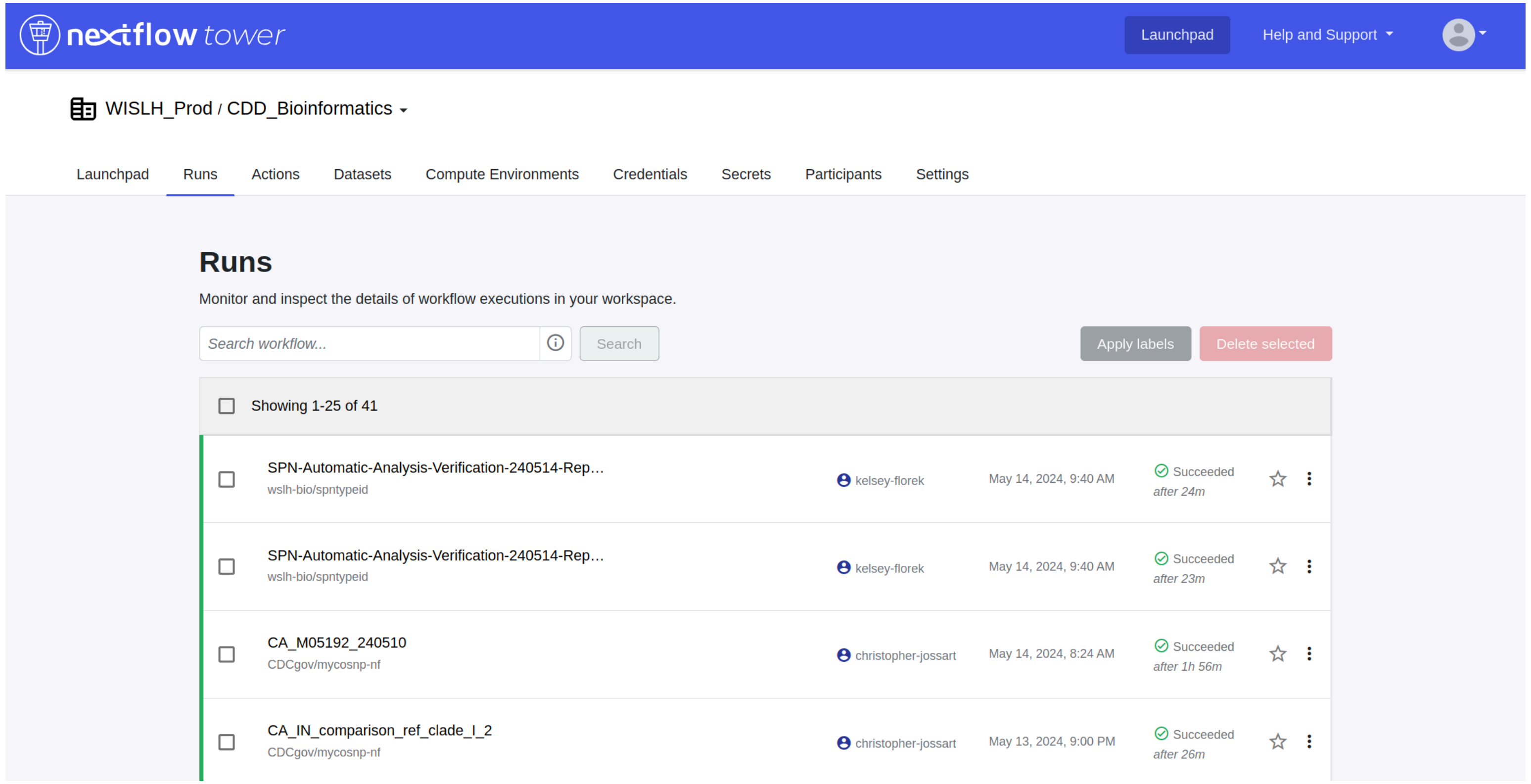
Name	Type	Provisioning model	Instance types	Status	State	Minimum vCPUs	Desired vCPUs	Maximum vCPUs
wslh_nextflow_head_20240328182605045600000003	Managed	EC2	c5	✓ Valid	Enabled	-	2	8
wslh_nextflow_job_20240328182605147500000004	Managed	Spot	c5, c4, c6i	✓ Valid	Enabled	-	48	128

nextflow tower

Nextflow Tower is an intuitive centralized command post that enables data analysis at scale. With Tower, users can easily launch, manage, and monitor scalable Nextflow data analysis pipelines and compute environments on-premises or across the cloud providers of their choice.

Seqera Labs - Nextflow Tower

Nextflow Tower - Dashboard



Runs

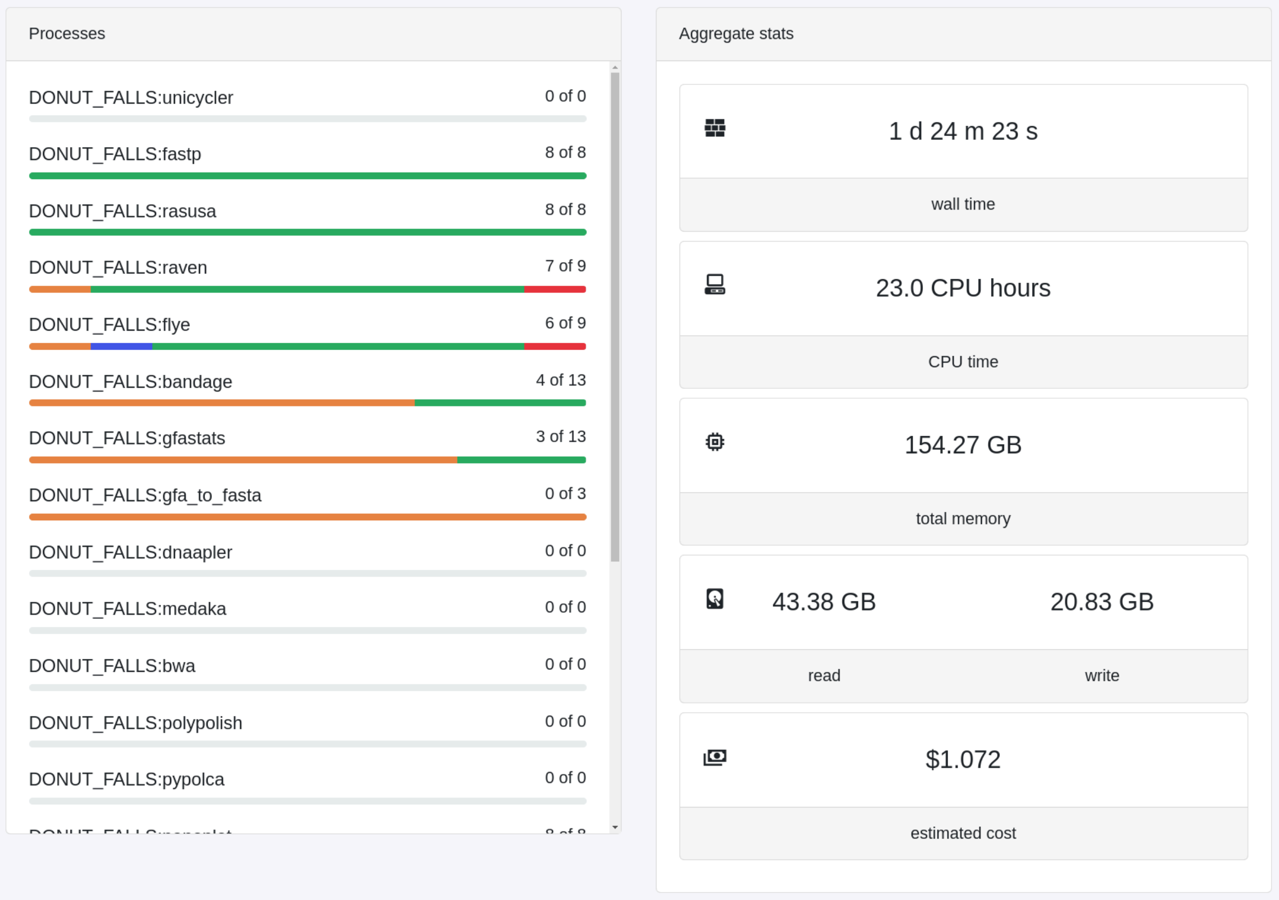
Monitor and inspect the details of workflow executions in your workspace.

□ Showing 1-25 of 41

□	SPN-Automatic-Analysis-Verification-240514-Rep… wslh-bio/spntypeid	kelsey-florek	May 14, 2024, 9:40 AM	Succeeded *after 24m*	☆
□	SPN-Automatic-Analysis-Verification-240514-Rep… wslh-bio/spntypeid	kelsey-florek	May 14, 2024, 9:40 AM	Succeeded *after 23m*	☆
□	CA_M05192_240510 CDCgov/mycosnp-nf	christopher-jossart	May 14, 2024, 8:24 AM	Succeeded *after 1h 56m*	☆
□	CA_IN_comparison_ref_clade_I_2 CDCgov/mycosnp-nf	christopher-jossart	May 13, 2024, 9:00 PM	Succeeded *after 26m*	☆

Seqera Labs - Nextflow Tower

Nextflow Tower - Monitor



Processes

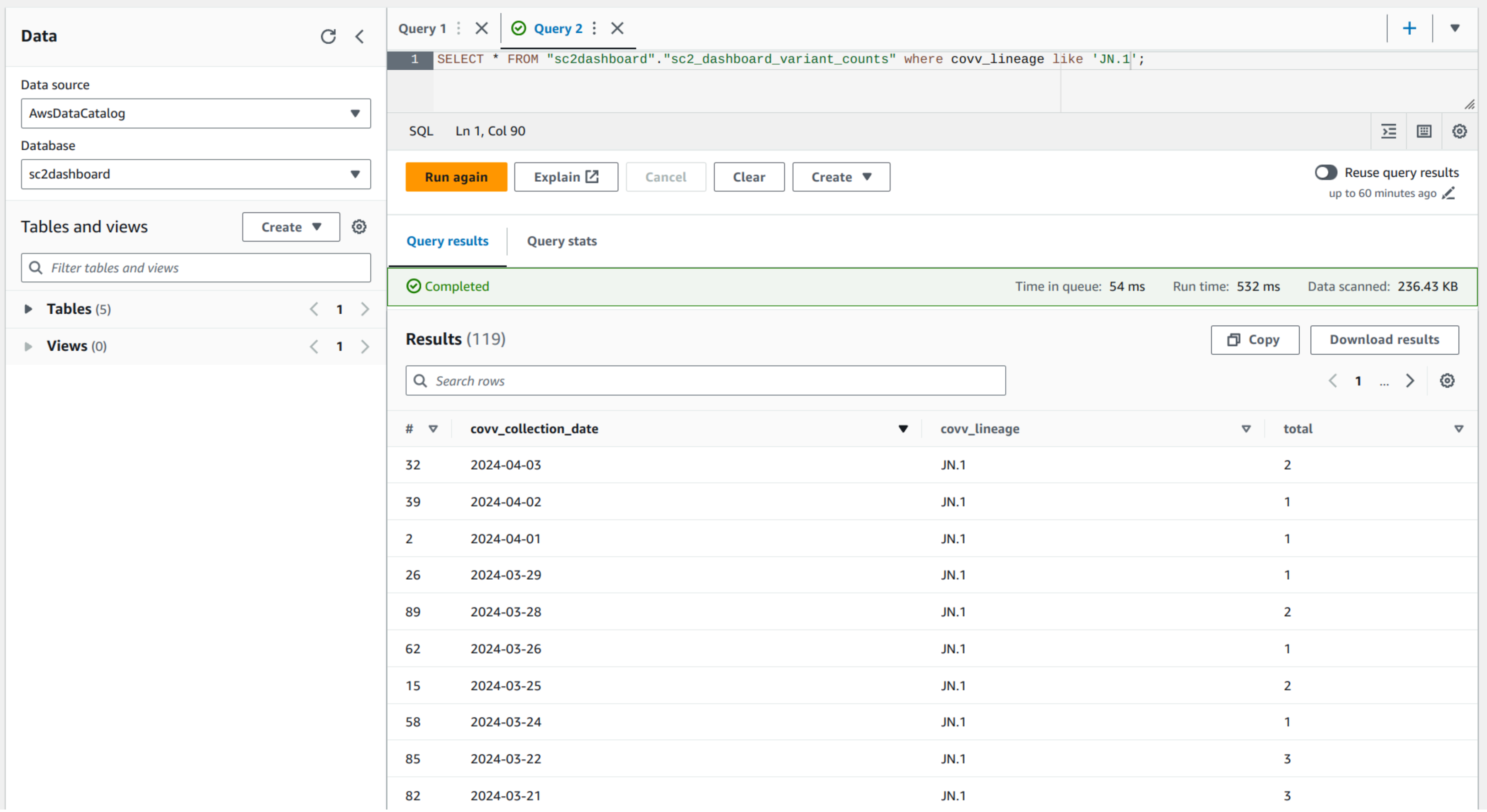
DONUT_FALLS:unicycler	0 of 0
DONUT_FALLS:fastp	8 of 8
DONUT_FALLS:rasusa	8 of 8
DONUT_FALLS:raven	7 of 9
DONUT_FALLS:flye	6 of 9
DONUT_FALLS:bandage	4 of 13
DONUT_FALLS:gfastats	3 of 13
DONUT_FALLS:gfa_to_fasta	0 of 3
DONUT_FALLS:dnaapler	0 of 0
DONUT_FALLS:medaka	0 of 0
DONUT_FALLS:bwa	0 of 0
DONUT_FALLS:polypolish	0 of 0
DONUT_FALLS:pypolca	0 of 0
DONUT_FALLS:nanoplot	8 of 8

Aggregate stats

1 d 24 m 23 s	wall time
23.0 CPU hours	CPU time
154.27 GB	total memory
43.38 GB	read
20.83 GB	write
$1.072	estimated cost

Seqera Labs - Nextflow Tower

AWS Athena



Data source: AwsDataCatalog

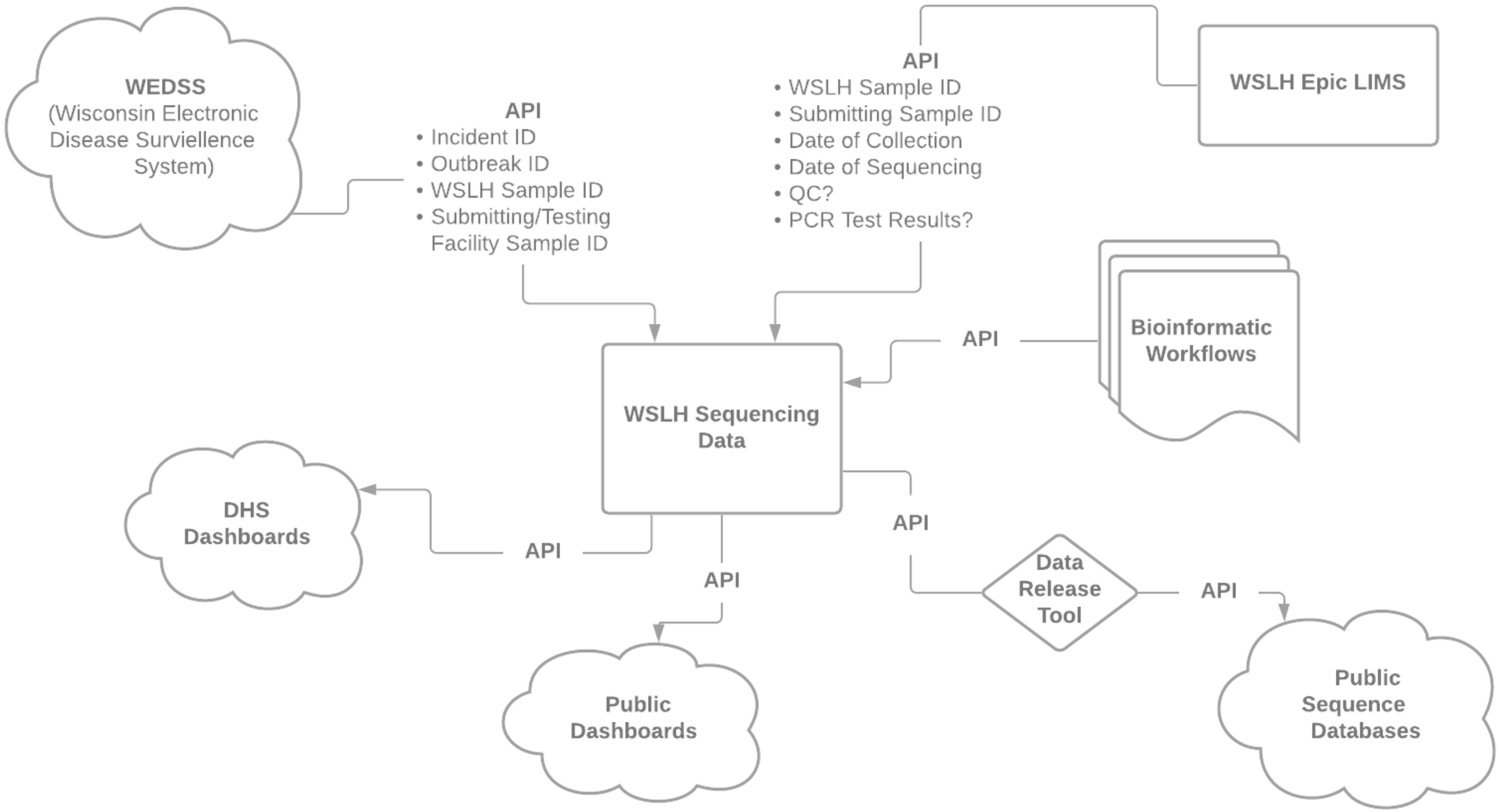
Database: sc2dashboard

```
SELECT * FROM "sc2dashboard"."sc2_dashboard_variant_counts" where covv_lineage like 'JN.1';
```

Completed   Time in queue: 54 ms   Run time: 532 ms   Data scanned: 236.43 KB

#	covv_collection_date	covv_lineage	total
32	2024-04-03	JN.1	2
39	2024-04-02	JN.1	1
2	2024-04-01	JN.1	1
26	2024-03-29	JN.1	1
89	2024-03-28	JN.1	2
62	2024-03-26	JN.1	1
15	2024-03-25	JN.1	2
58	2024-03-24	JN.1	1
85	2024-03-22	JN.1	3
82	2024-03-21	JN.1	3

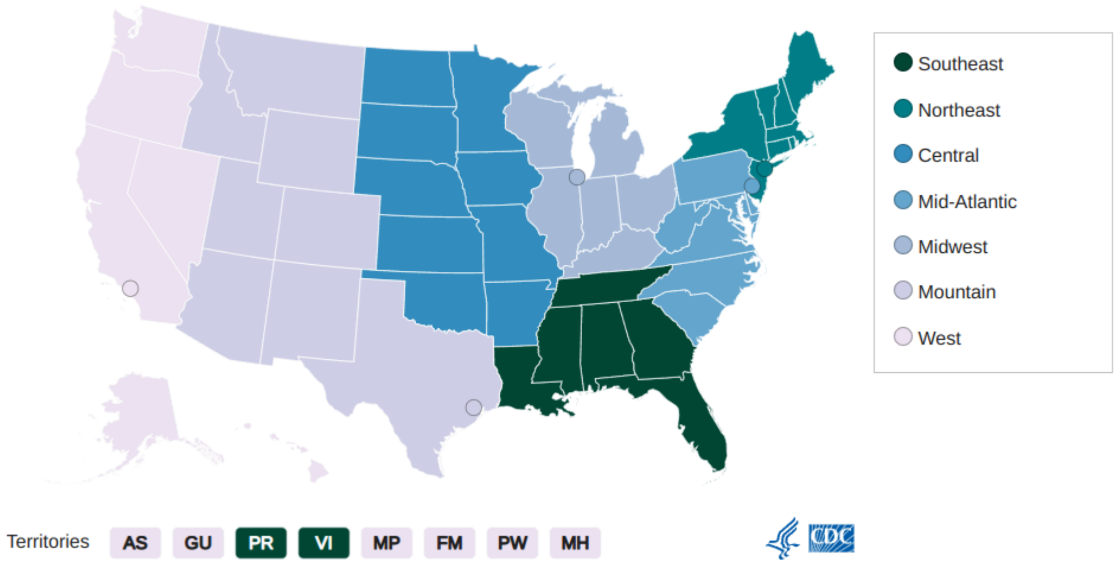
Connecting Data Across Siloed Systems



1. ~~Necessities of Next Generation Sequencing Capacity Building~~
2. ~~Blueprints for an NGS Data Solution~~
3. **Simplifying Genomics for Public Health Partners**

Need for a centralized resource

AMD Bioinformatics Regional Resource - Midwest Region

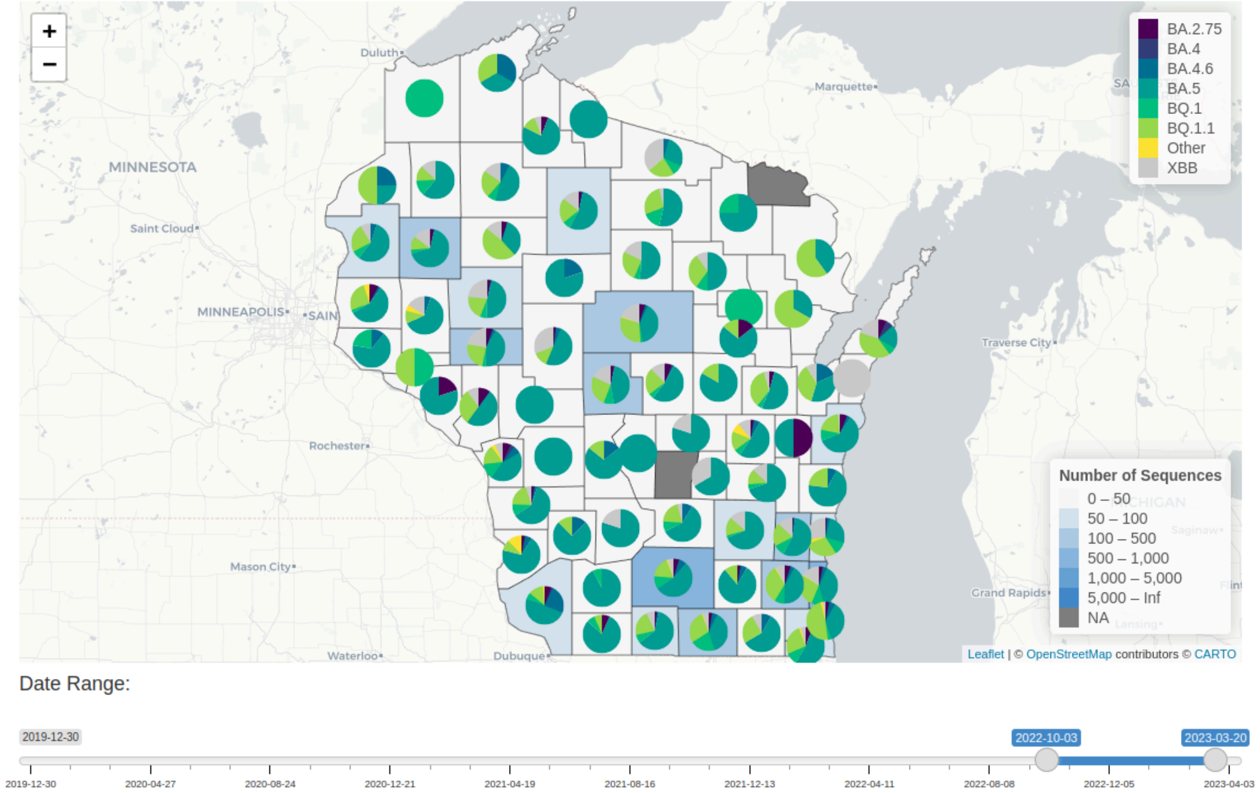


Ad-hoc Analytical Support

Provision of Computational Resources

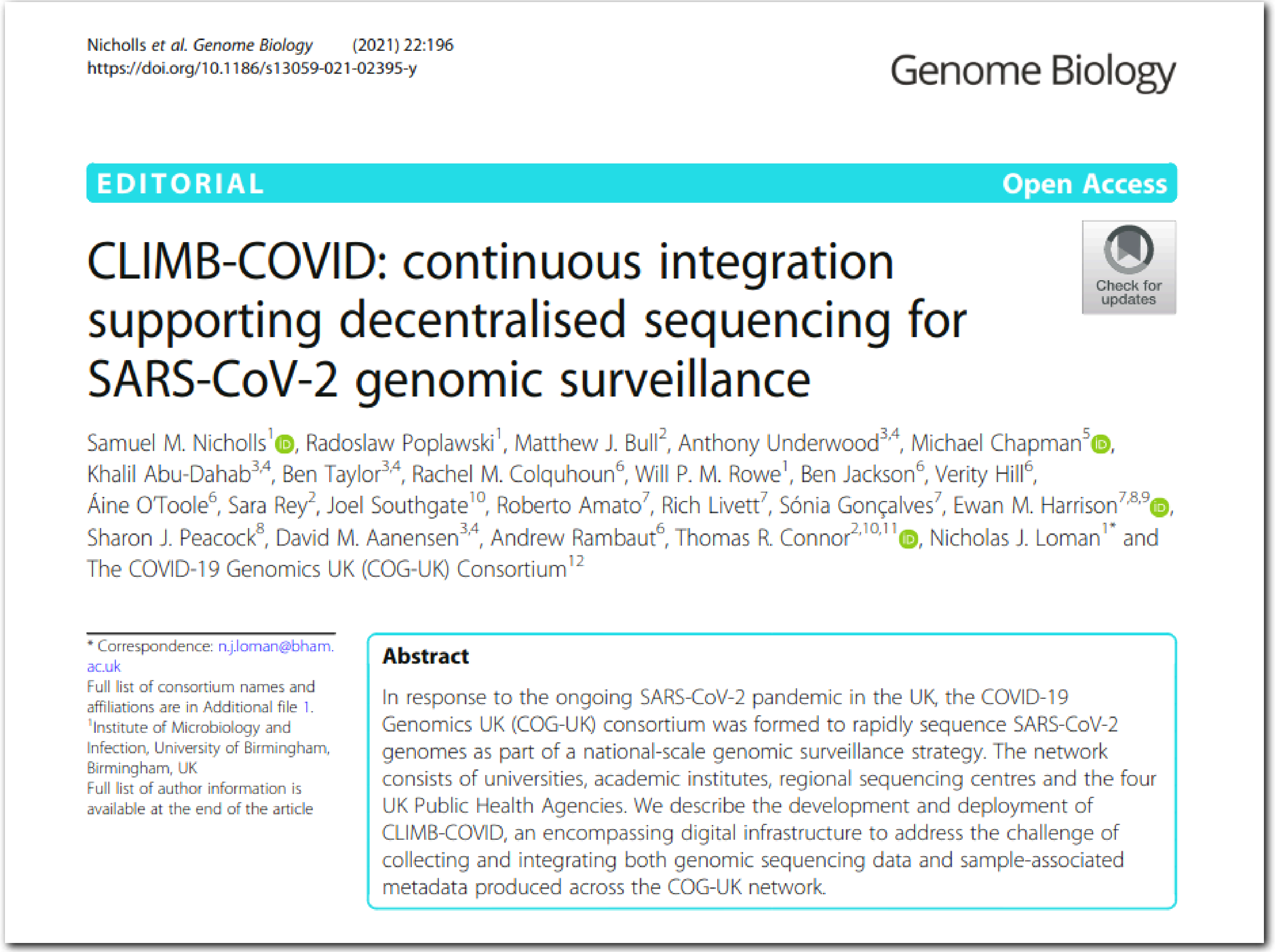
Need for a centralized resource

SARS-CoV-2 Genomic Surveillance



WSLH SARS-CoV-2 Dashboard

COVID-19 Genomics UK (COG-UK) CLIMB-COVID



Nicholls *et al. Genome Biology* (2021) 22:196
https://doi.org/10.1186/s13059-021-02395-y

Genome Biology

EDITORIAL Open Access

CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance

Samuel M. Nicholls[1], Radoslaw Poplawski[1], Matthew J. Bull[2], Anthony Underwood[3,4], Michael Chapman[5], Khalil Abu-Dahab[3,4], Ben Taylor[3,4], Rachel M. Colquhoun[6], Will P. M. Rowe[1], Ben Jackson[6], Verity Hill[6], Áine O'Toole[6], Sara Rey[2], Joel Southgate[10], Roberto Amato[7], Rich Livett[7], Sónia Gonçalves[7], Ewan M. Harrison[7,8,9], Sharon J. Peacock[8], David M. Aanensen[3,4], Andrew Rambaut[6], Thomas R. Connor[2,10,11], Nicholas J. Loman[1*] and The COVID-19 Genomics UK (COG-UK) Consortium[12]

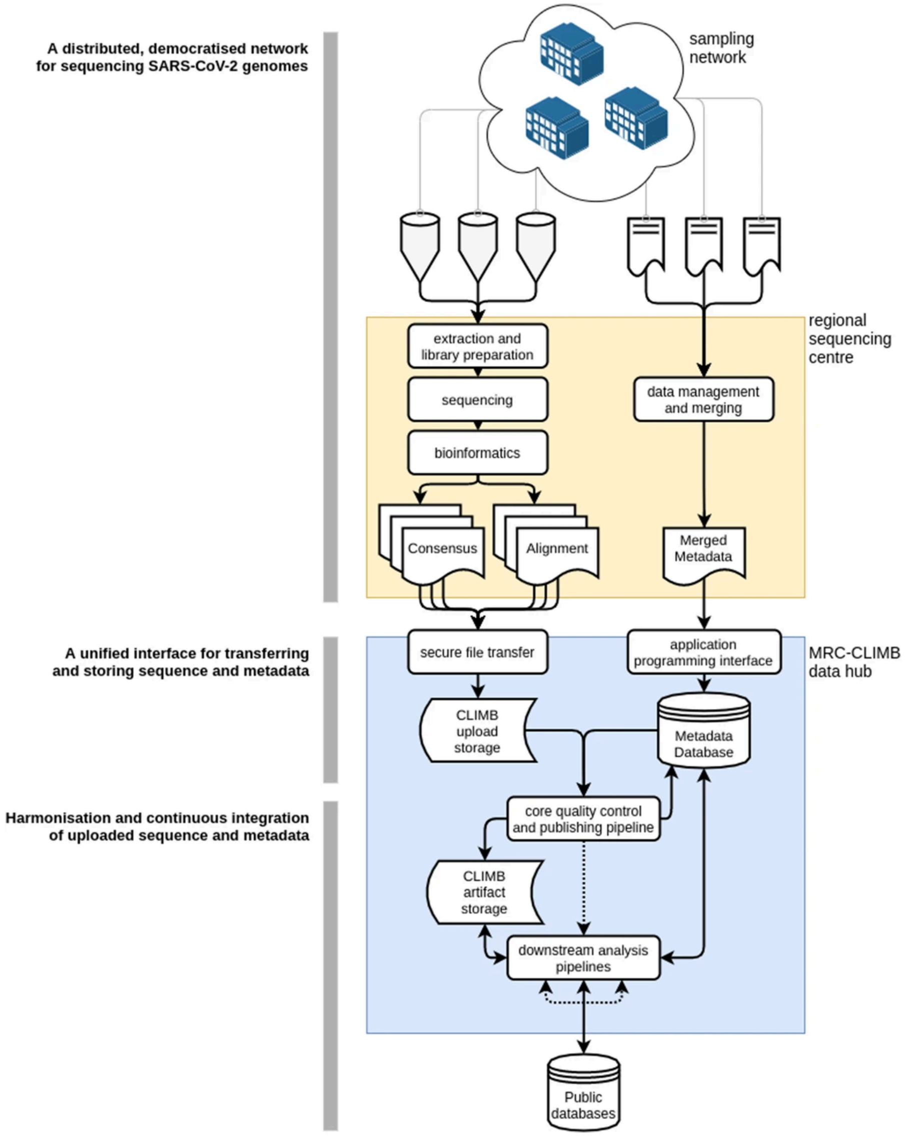
* Correspondence: n.j.loman@bham.ac.uk
Full list of consortium names and affiliations are in Additional file 1.
[1]Institute of Microbiology and Infection, University of Birmingham, Birmingham, UK
Full list of author information is available at the end of the article

Abstract

In response to the ongoing SARS-CoV-2 pandemic in the UK, the COVID-19 Genomics UK (COG-UK) consortium was formed to rapidly sequence SARS-CoV-2 genomes as part of a national-scale genomic surveillance strategy. The network consists of universities, academic institutes, regional sequencing centres and the four UK Public Health Agencies. We describe the development and deployment of CLIMB-COVID, an encompassing digital infrastructure to address the challenge of collecting and integrating both genomic sequencing data and sample-associated metadata produced across the COG-UK network.

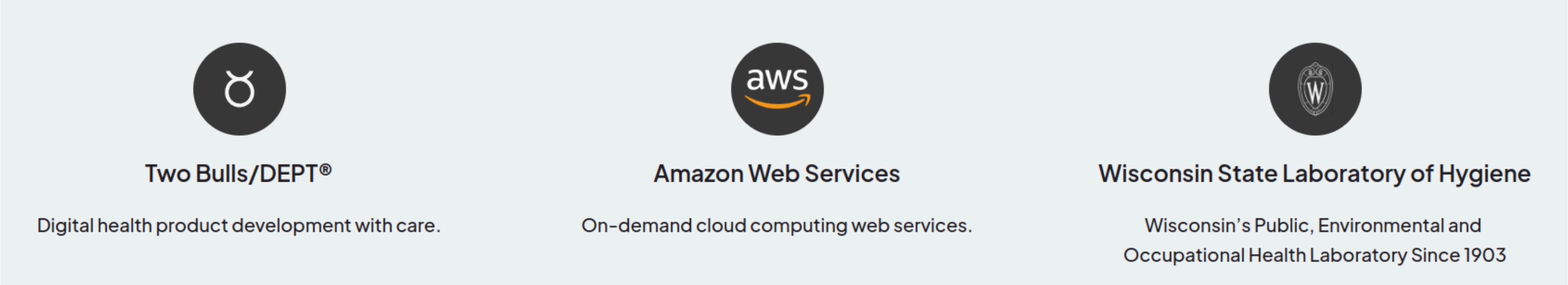
Nicholls et al. 2021 *Genome Biology*

COVID-19 Genomics UK (COG-UK) CLIMB-COVID


A distributed, democratised network for sequencing SARS-CoV-2 genomes
A unified interface for transferring and storing sequence and metadata
Harmonisation and continuous integration of uploaded sequence and metadata
sampling network
regional sequencing centre
extraction and library preparation
sequencing
bioinformatics
Consensus
Alignment
data management and merging
Merged Metadata
MRC-CLIMB data hub
secure file transfer
application programming interface
CLIMB upload storage
Metadata Database
core quality control and publishing pipeline
CLIMB artifact storage
downstream analysis pipelines
Public databases

Nicholls et al. 2021 *Genome Biology*
StaPH-B Monthly Webinar Oct 2021

Easy Genomics Partnership



Two Bulls/DEPT®

Digital health product development with care.

Amazon Web Services

On-demand cloud computing web services.

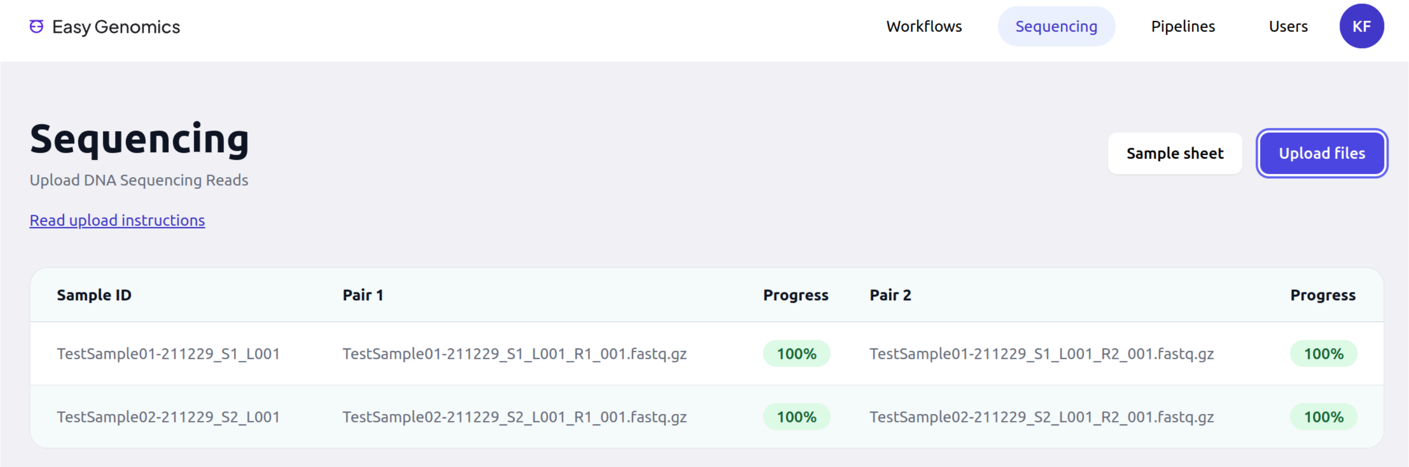
Wisconsin State Laboratory of Hygiene

Wisconsin's Public, Environmental and
Occupational Health Laboratory Since 1903

Easy Genomics - Minimal Viable Product

- Simplify the process of launching and monitoring workflows
- Provide the ability for users to upload sequence data through the web browser
- Allow users to download analysis results through the web browser
- User/Lab separation

Easy Genomics - Sequence Data Upload



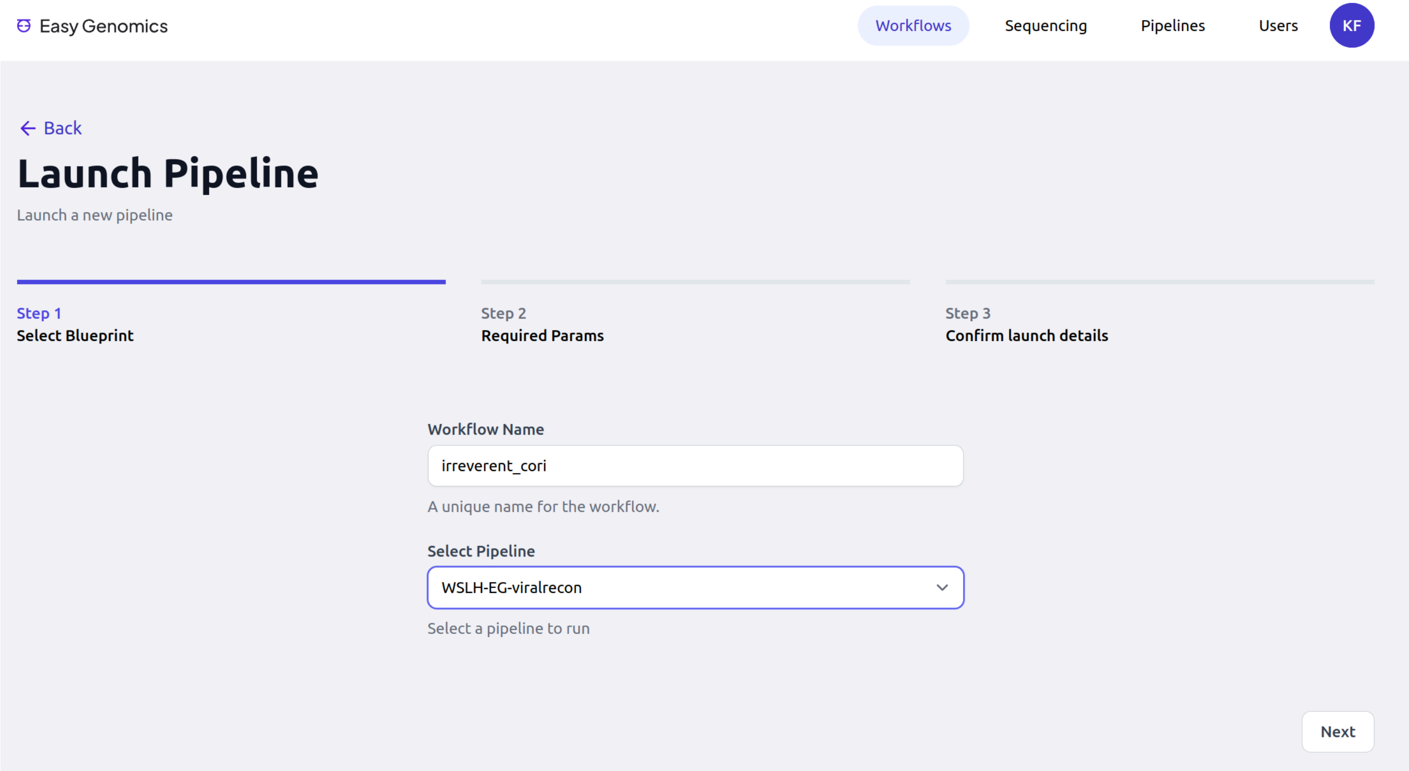
Sequencing

Upload DNA Sequencing Reads

Read upload instructions

Sample ID	Pair 1	Progress	Pair 2	Progress
TestSample01-211229_S1_L001	TestSample01-211229_S1_L001_R1_001.fastq.gz	100%	TestSample01-211229_S1_L001_R2_001.fastq.gz	100%
TestSample02-211229_S2_L001	TestSample02-211229_S2_L001_R1_001.fastq.gz	100%	TestSample02-211229_S2_L001_R2_001.fastq.gz	100%

Easy Genomics - Launch



Launch Pipeline

Launch a new pipeline

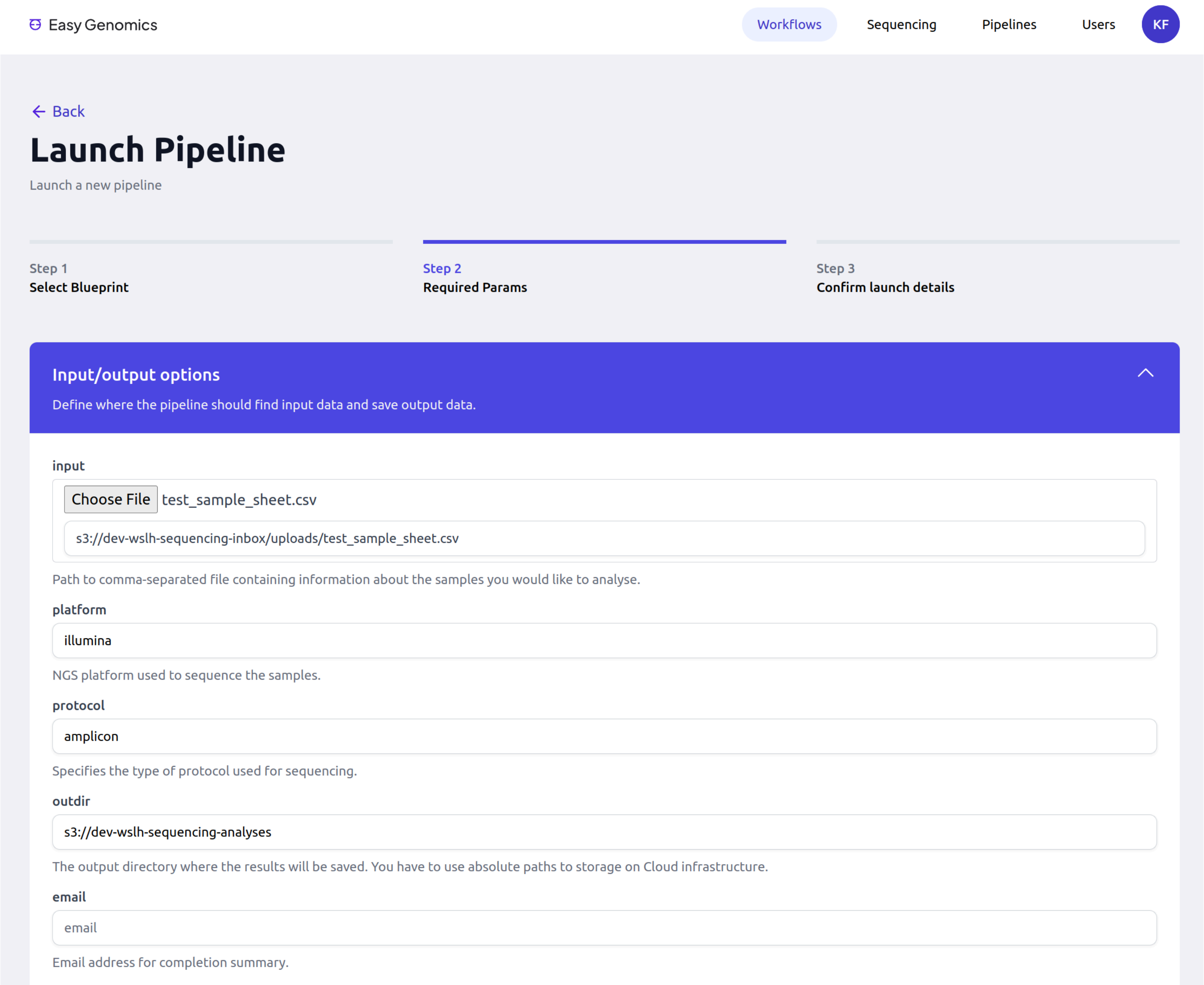
Workflow Name: irreverent_cori

A unique name for the workflow.

Select Pipeline: WSLH-EG-viralrecon

Select a pipeline to run

Easy Genomics - Launch



Launch Pipeline

Launch a new pipeline

Step 1
Select Blueprint

Step 2
Required Params

Step 3
Confirm launch details

Input/output options

Define where the pipeline should find input data and save output data.

input

Choose File test_sample_sheet.csv

s3://dev-wslh-sequencing-inbox/uploads/test_sample_sheet.csv

Path to comma-separated file containing information about the samples you would like to analyse.

platform

illumina

NGS platform used to sequence the samples.

protocol

amplicon

Specifies the type of protocol used for sequencing.

outdir

s3://dev-wslh-sequencing-analyses

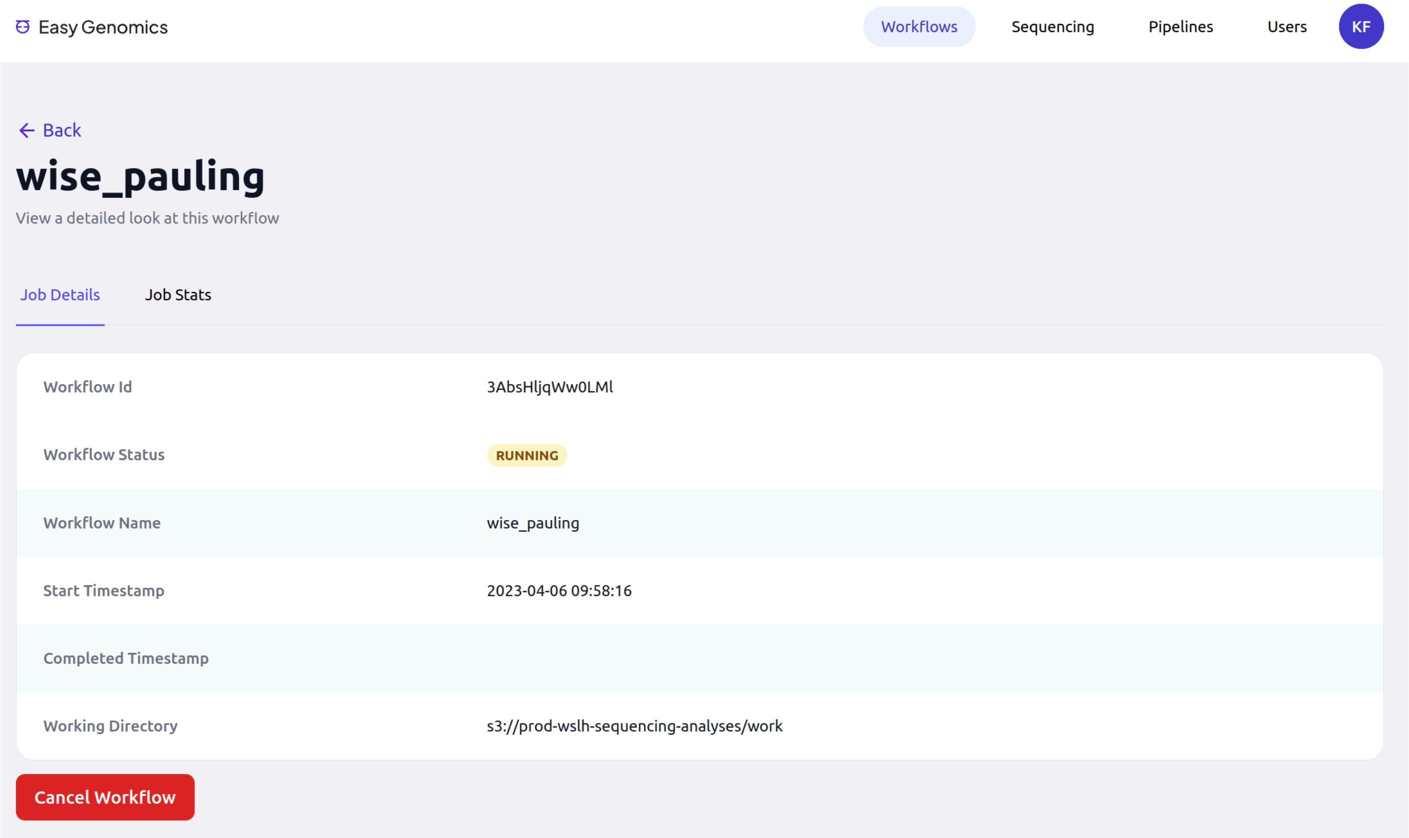
The output directory where the results will be saved. You have to use absolute paths to storage on Cloud infrastructure.

email

email

Email address for completion summary.

Easy Genomics - Monitor

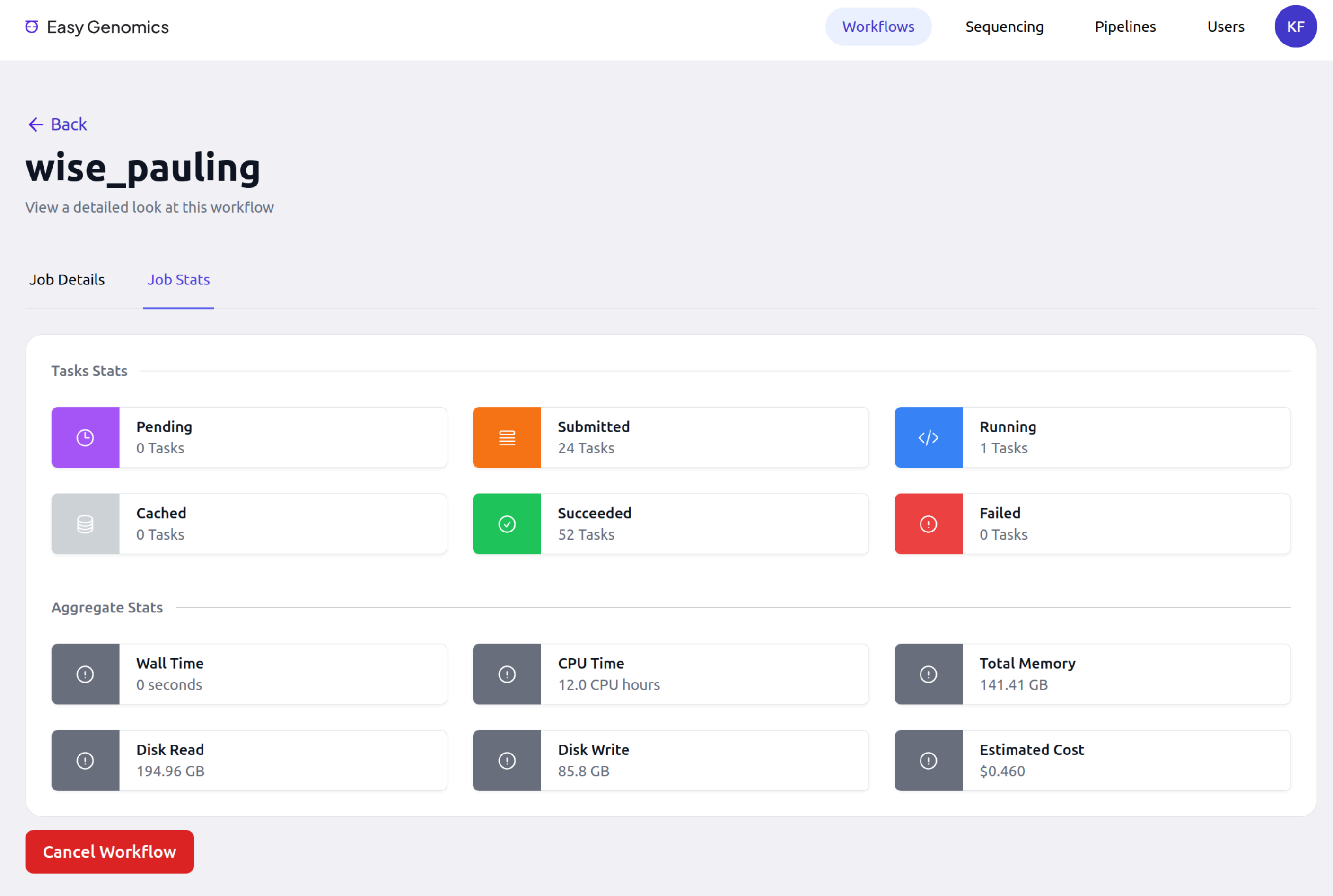


wise_pauling

View a detailed look at this workflow

Workflow Id	3AbsHljqWw0LMl
Workflow Status	RUNNING
Workflow Name	wise_pauling
Start Timestamp	2023-04-06 09:58:16
Completed Timestamp	
Working Directory	s3://prod-wslh-sequencing-analyses/work

Easy Genomics - Monitor



wise_pauling

View a detailed look at this workflow

Tasks Stats

Pending 0 Tasks	Submitted 24 Tasks	Running 1 Tasks
Cached 0 Tasks	Succeeded 52 Tasks	Failed 0 Tasks

Aggregate Stats

Wall Time 0 seconds	CPU Time 12.0 CPU hours	Total Memory 141.41 GB
Disk Read 194.96 GB	Disk Write 85.8 GB	Estimated Cost $0.460

Easy Genomics - Roadmap

Easy Genomics - Roadmap

- 2024 Spring - Deploy Easy Genomics for internal use

Easy Genomics - Roadmap

- 2024 Spring - Deploy Easy Genomics for internal use
- 2024 Early Summer - Open Access to SARS-CoV-2 Sequencing Laboratories

Easy Genomics - Roadmap

- 2024 Spring - Deploy Easy Genomics for internal use
- 2024 Early Summer - Open Access to SARS-CoV-2 Sequencing Laboratories
- 2024 Mid Summer - Easy Genomics MVP Update

Acknowledgments

Abigail Shockey, PhD

Thomas Blader

Christopher Jossart, MPH

Eva Gunawan, MS

Dustin Lyfoung, MS

Special Thanks

- UW-Madison Public Cloud Team
- UW-Madison Office of Cybersecurity

Reveal.JS