

Developing Threat Agnostic Sequencing Capabilities and the Utility in Public Health Laboratories

A framework for local analysis of NGS data in real-time

Brian Merritt M.S.
Johns Hopkins University Applied Physics Laboratory
Laurel, MD
brian.merritt@jhuapl.edu

Pre-event/early warning metagenomic sequencing

System level considerations

National Biodefense
Strategy integration

Scope, Scale,
Sustainability

Infrastructure
requirements

Translation into
practice

Regulatory
requirements

Pre-analytical

Surveillance

- Sample source (human, animal, environmental)
- Sampling scheme
- Metadata

Clinical diagnostics

- Specimen type (swab, stool, blood, serum)
- Sampling scheme
- Clinical Data

Methods development

Replicable wet & dry laboratory workflows

- Wet laboratory workflows
- Flexible bioinformatic pipeline (TaxTriage)
- Database curation/testing
- Technical support
- Interlaboratory Comparison
- Transition to additional laboratories
- Proficiency Testing

Post-analytical

Surveillance Reporting

- Integration into surveillance systems
- Data management requirements
- Platform evaluation

Clinical Reporting

- Electronic Laboratory Reporting (ELR)
- Data management requirements
- Patient care



Implementing Sustainable Genomic Epidemiology Research

The Need for Rapid and Simplistic Deployment for NGS Analysis

Intro and advanced workshops to enable researchers to:

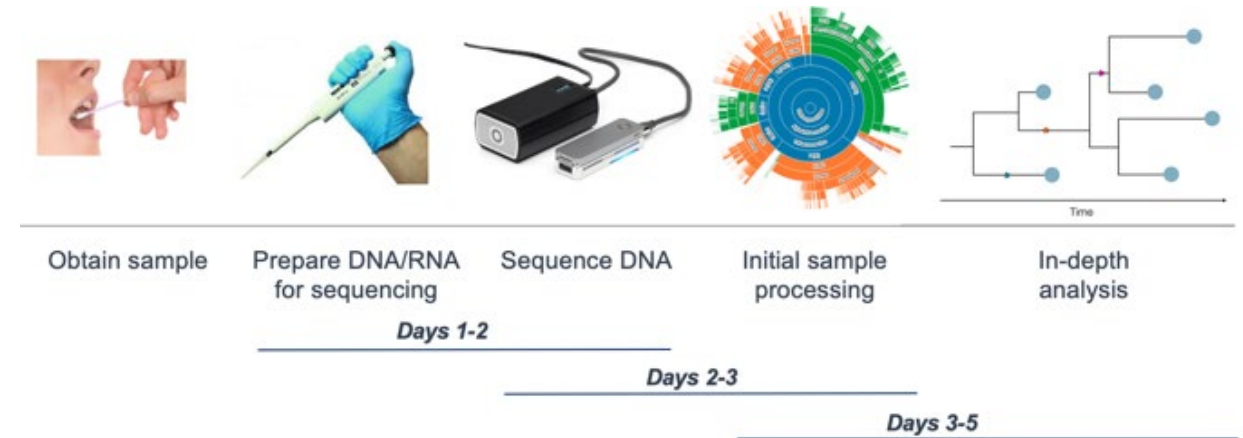
- Establish best practices for genomic surveillance
 - Influenza initially
 - SARS-CoV-2 in 2020 and beyond
- Prepare RNA virus samples for sequencing
- Analyze raw DNA sequence data
- Perform phylogenetic and phylodynamic analysis

Seven workshops across FY19-22:

- Asia (x3), North/South America (2), Africa (x2)

Limitations

- Limited Compute/Development Knowledge
- Cost to keep a bioinformatician
- Data Storage
 - Setup and egress time/costs
 - Limited Internet prevents rapid analysis
 - Some institutes have unreliable internet



High Throughput DNA Sequencing Platforms

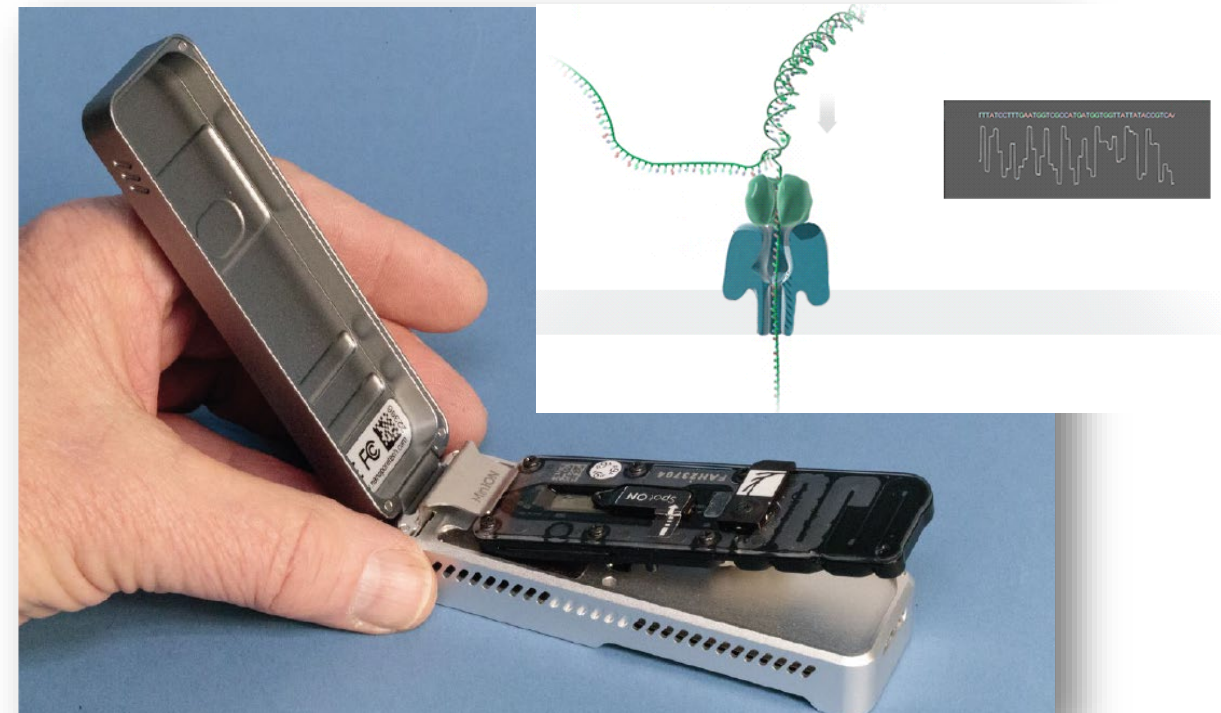
Short-read sequencing

- Indirect DNA analysis
- Ultra-accurate
- <300bp read lengths
- 1-2 days from sample to data
- Laboratory instruments (\$30-300k)



Long-read sequencing

- Direct DNA analysis
- Moderate accuracy
- Any length DNA
- <1 hour to 3 days from sample to data
- Laptop connected instruments (<\$1k)



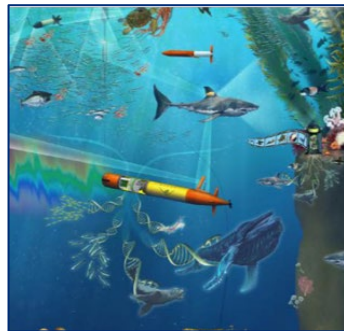
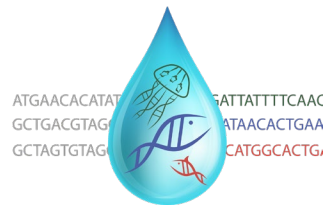
DNA Surveillance

How we can establish analysis from sample collection (metagenomics)

Initial Collection

eDNA collection or Swabs

Sample preparation for Illumina/Oxford
Nanopore



DNA Surveillance

How we can establish analysis from sample collection (metagenomics)

Initial Collection

eDNA collection or Swabs

Sample preparation for Illumina/Oxford
Nanopore

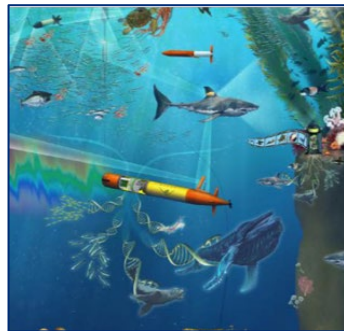
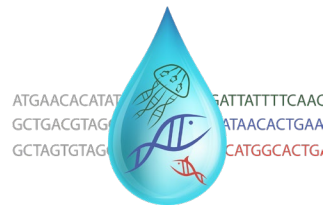
Problems

Technical limitations in in-field technicians

Hardware accessibility

Internet capabilities

Familiarity with report analysis format



DNA Surveillance

How we can establish analysis from sample collection (metagenomics)

Initial Collection

eDNA collection or Swabs

Sample preparation for Illumina/Oxford Nanopore

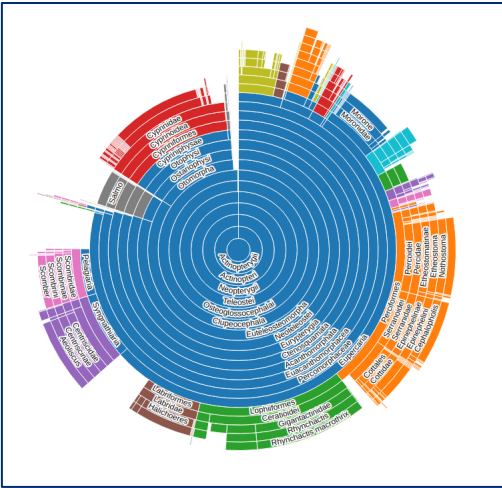
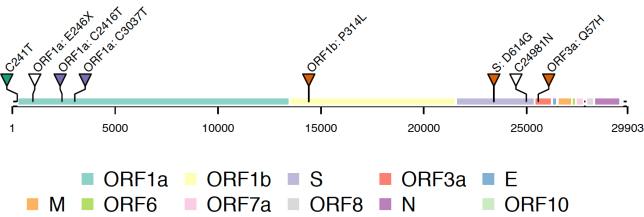
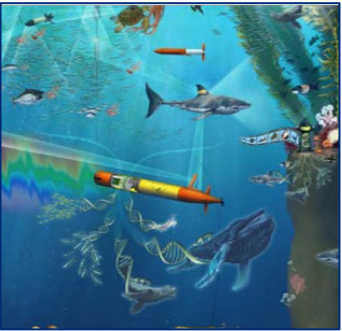
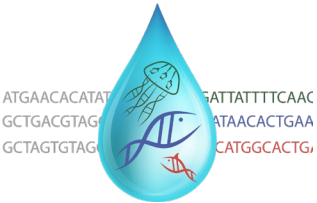
Problems

Technical limitations in in-field technicians

Hardware accessibility

Internet capabilities

Familiarity with report analysis format



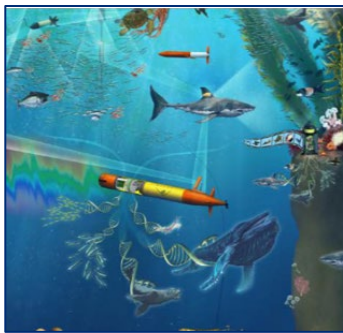
DNA Surveillance

How we can establish analysis from sample collection (metagenomics)

Initial Collection

eDNA collection or Swabs

Sample preparation for Illumina/Oxford Nanopore



Implementation of Software Frameworks

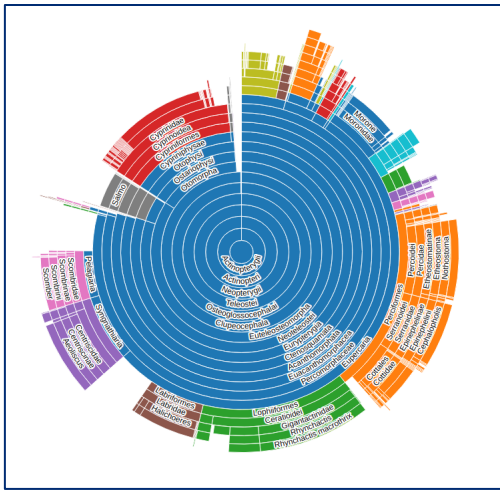
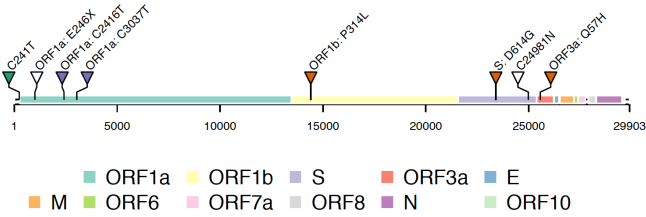
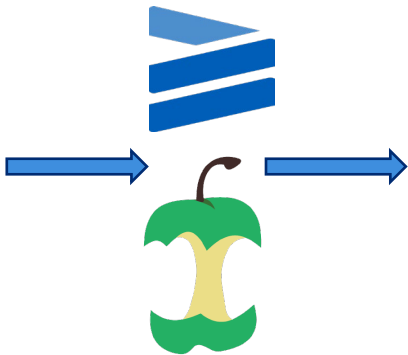
Automated Pipeline Build Process

Agnostic OS Requirements

Containerized Environments

Community Driven / Open Source

Local and Cloud Capable



TaxTriage

Classifications, filtering and qc in a multi level reports

TaxTriage

Modularized Nextflow Process for Detailed Processing and Report building of Genomics Data

What it is

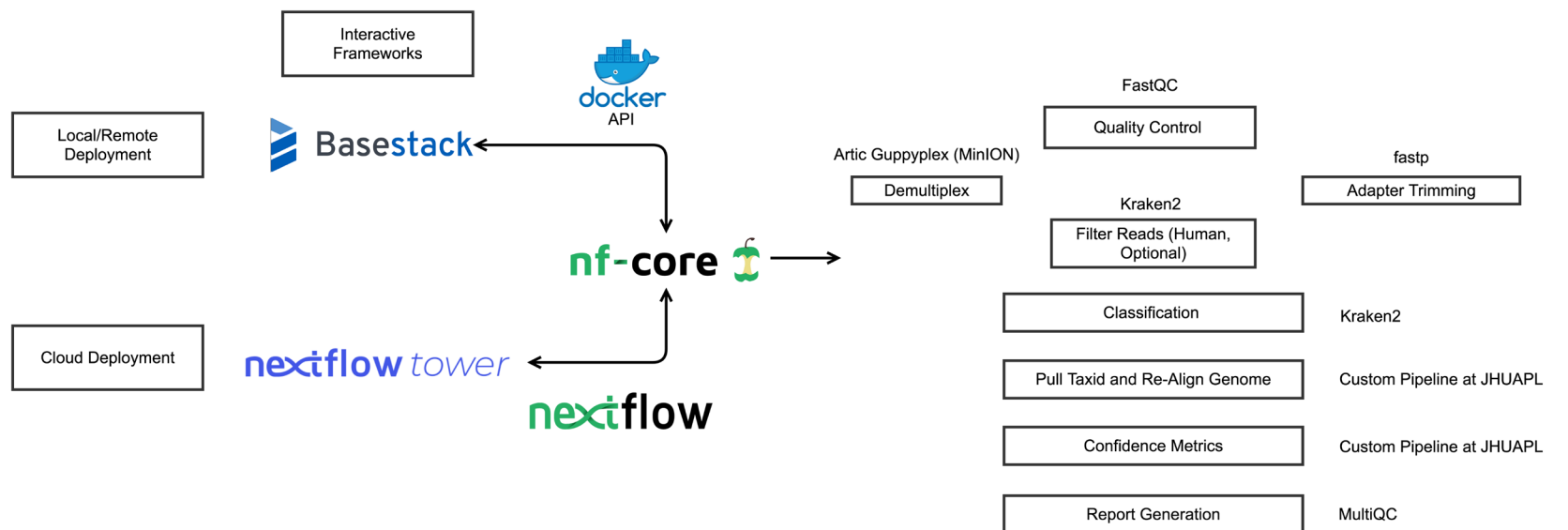
- Nextflow and nf-core pipeline designed to deliver Interactive CLIA style reports
- Rapid Taxonomic Filtering and Classification from Illumina or Oxford Nanopore Samples

How to use it

- Deployable Windows 10, Linux, Mac OSX
 - Includes Mac M1
- Integrated into Basestack
 - Integrated user interface
- Cloud Compute Capability
 - Nextflow Tower

Future Updates

- Improved Taxa Classification Predictions
- Integrated Remote Databases



Copy table Configure Columns Plot Showing ⁵⁴/₅₄ rows and ¹¹/₁₁ columns.										
Acc ▲	Name	Rank	Length	Reads Aligned	Abu Total Aligned	Abu Total Bases	Total Bases Adjusted	Breadth Genome Coverage	Depth Coverage Mean	Depth Coverage
1280_HEC_Background	Staphylococcus aureus	S	2 980 548.0	384 093.0	0.2	0.5	0.5	0.8	26.8	49.2
1280_Sample1	Staphylococcus aureus	S	2 980 548.0	170 361.0	0.1	0.2	0.5	0.8	11.9	6.6
1280_Sample2	Staphylococcus aureus	S	2 980 548.0	387 292.0	0.2	0.5	0.6	0.9	27.1	26.5
1280_Sample3	Staphylococcus aureus	S	2 980 548.0	298 095.0	0.2	0.4	0.5	0.8	21.1	32.0
1280_Sample4	Staphylococcus aureus	S	2 980 548.0	282 619.0	0.2	0.4	0.5	0.8	19.7	26.6
1280_Sample5	Staphylococcus aureus	S	2 980 548.0	407 102.0	0.2	0.4	0.4	0.8	28.8	44.7
1280_Sample6	Staphylococcus aureus	S	2 980 548.0	541 550.0	0.2	0.4	0.5	0.8	37.6	56.5
1280_Sample7	Staphylococcus aureus	S	2 980 548.0	627 277.0	0.2	0.5	0.5	0.9	43.7	69.1
12875_HEC_Background	Rothia	S	2 282 716.0	133 203.0	0.1	0.1	0.1	0.2	7.1	115.4

Base Report

Staph

Sample Name	% Staphylococcus aureus	% Top 5 Species	% Unclassified	% Dups	% GC	Read Length	% Failed	M Seqs
Staph_1	91.1%	91.5%	2.1%					
Staph_1_1				7.3%	32%	228 bp	9%	0.1
Staph_1_2				6.9%	32%	228 bp	9%	0.1
Staph_2	77.5%	81.0%	11.9%					
Staph_2_1				1.0%	33%	205 bp	9%	0.0
Staph_2_2				0.9%	33%	206 bp	9%	0.0
Staph_3	35.4%	52.0%	38.5%					
Staph_3_1				0.1%	36%	173 bp	18%	0.0
Staph_3_2				0.1%	36%	175 bp	9%	0.0
Staph_4	59.0%	68.6%	26.6%					
Staph_4_1				0.0%	36%	159 bp	27%	0.0
Staph_4_2				0.0%	37%	218 bp	36%	0.0
Staph_5	0.5%	24.6%	65.7%					
Staph_5_1				0.3%	41%	128 bp	27%	0.0
Staph_5_2				0.2%	41%	129 bp	27%	0.0
Staph_6	0.0%	19.5%	72.0%					
Staph_6_1				0.2%	41%	109 bp	27%	0.0
Staph_6_2				0.2%	42%	110 bp	27%	0.0

Base Report

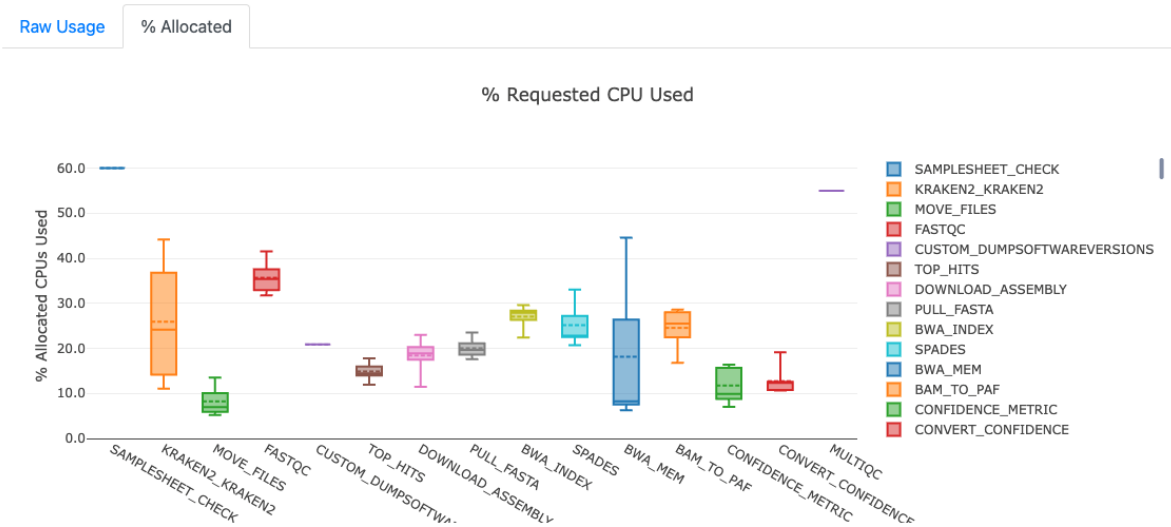
An Example Output of Sample Stats

Acc	Name	Rank	Length	Reads Aligned	Abu Total Aligned	Abu Total Bases	Total Bases Adjusted	Breadth Genome Coverage	Depth Coverage Mean	Depth Coverage Stdev	Depth Coef. Variation
619591_Sample_1	Monkeypox virus Zaire-96-l-16	S1	196858.0	6931.0	0.8	0.8	0.8	1.0	8.6	4.4	0.5
619591_Sample_2	Monkeypox virus Zaire-96-l-16	S1	196858.0	1003.0	0.8	0.8	0.8	0.7	1.2	1.2	0.9
10244_Sample_1	Monkeypox virus	S	197209.0	1546.0	0.2	0.2	0.2	0.5	1.9	2.8	1.5
12643_Sample_1	Ectromelia virus	S	209771.0	331.0	0.0	0.0	0.0	0.2	0.4	0.8	2.3
10244_Sample_2	Monkeypox virus	S	197209.0	218.0	0.2	0.2	0.2	0.2	0.3	0.6	2.4
619591_Sample_3	Monkeypox virus Zaire-96-l-16	S1	196858.0	90.0	0.9	0.9	0.9	0.1	0.1	0.3	3.0
None	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
10244_Sample_3	Monkeypox virus	S	197209.0	11.0	0.1	0.1	0.1	0.0	0.0	0.1	8.6
44283_Sample_4	Pasteurella multocida subsp. multocida	S1	2137782.0	1.0	0.1	0.0	0.1	0.0	0.0	0.0	319.1
619591_Sample_4	Monkeypox virus Zaire-96-l-16	S1	196858.0	16.0	0.9	1.0	0.8	0.0	0.0	0.1	7.5
1328324_Sample_4	Klebsiella pneumoniae subsp. pneumoniae KPN1H27	S2	5241638.0	1.0	0.1	0.0	0.1	0.0	0.0	0.0	511.9

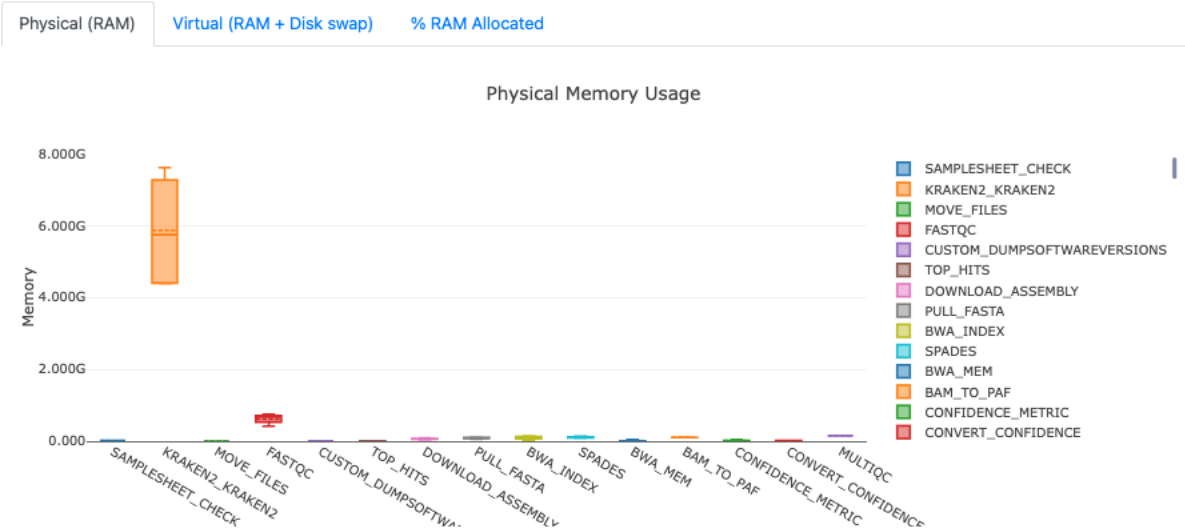
Example

Resource Monitoring

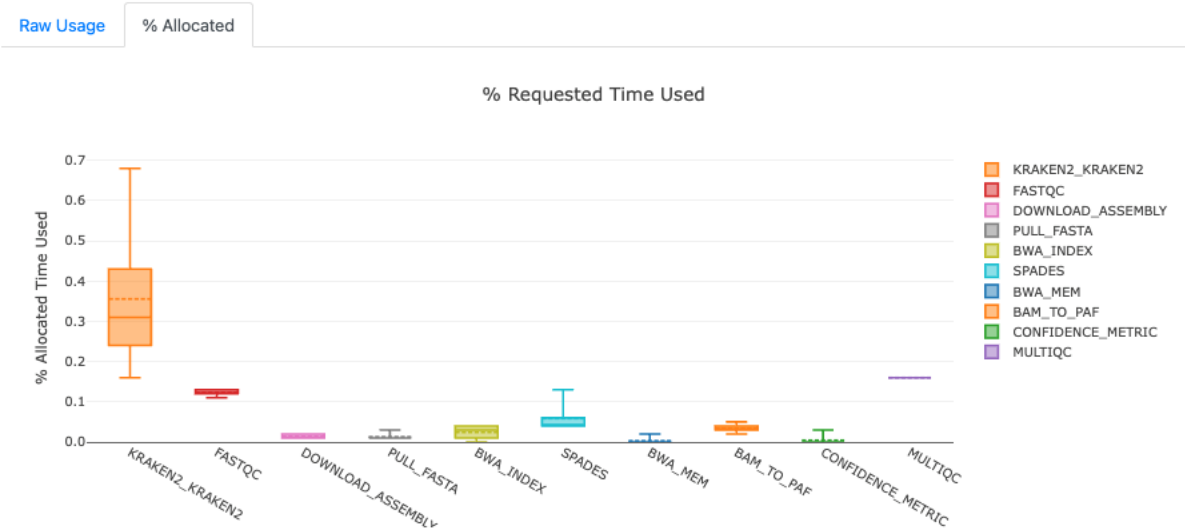
CPU



Memory



Job Duration



Real-Time Reporting

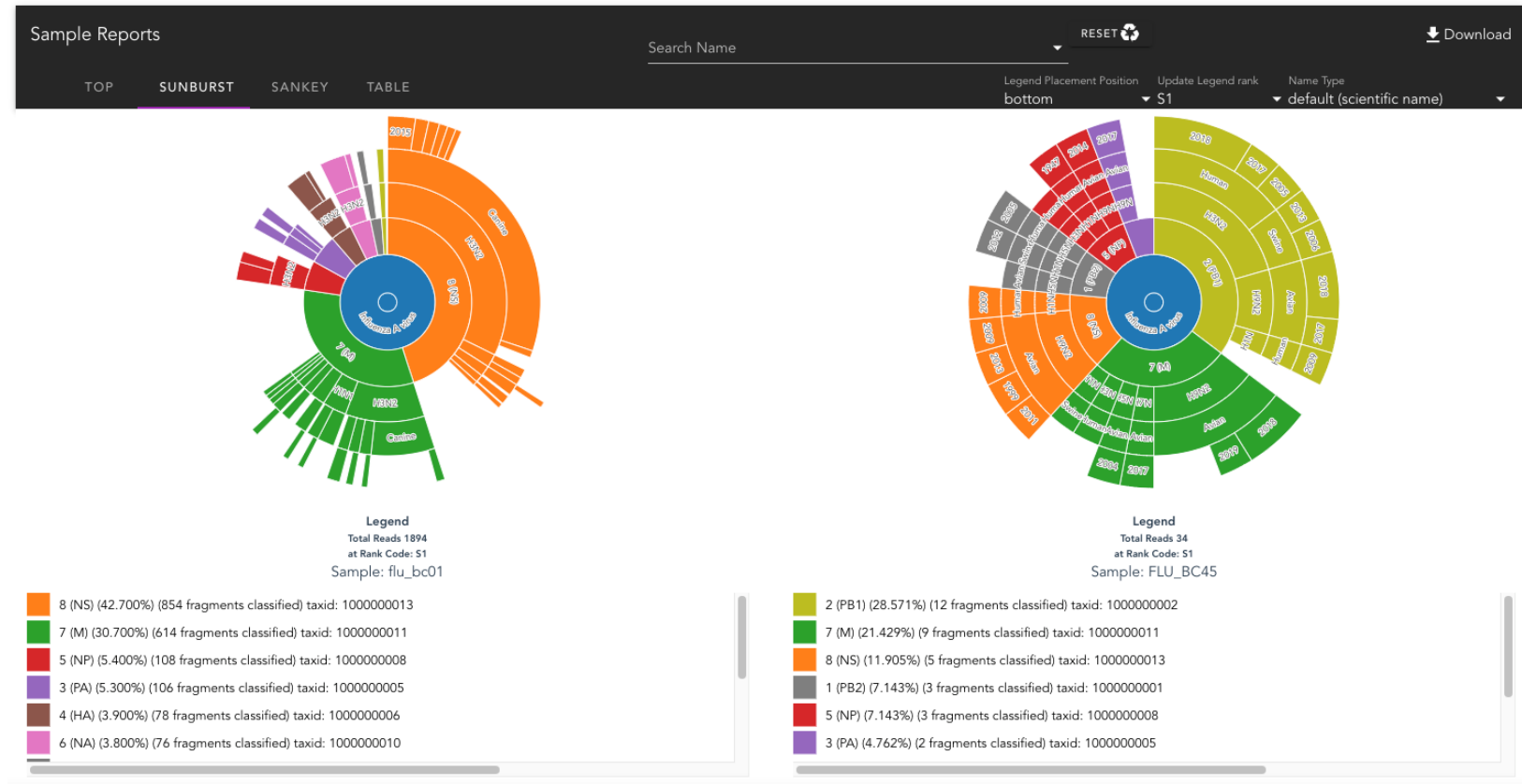
Generating Sequence Classifications in Real-time

- Use Case

- Real time reporting for sequencing reads (Oxford Nanopore, Illumina (paired or single))
- Interactive Design for detailed exploration
- Hierarchical representation of Lowest Common Ancestor (LCA) for reads
- Provides the Initial Triage step before running full analysis pipelines

- Features or Benefits

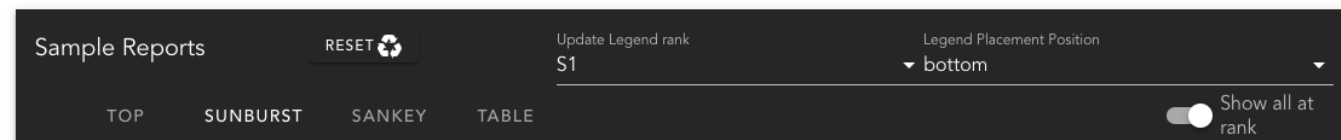
- Sequencing reads are fed, one-by-one through Kraken2 metagenomic classifier
- **OPTIONAL:** Demultiplexing Barcoding Run Directories automatically
- Watching for reads as generated, automated classification



Mytax2

Influenza samples classified in real-time

- Fully available in GitHub as source code
Deployable from standalone Docker containers & pushed to Docker Hub
- Full GPU support for demultiplexing Oxford Nanopore reads



Real Time Nanopore Report Analysis

Samplesheet

DOWNLOAD CSV

PAUSE UPDATES

RESTART REPORT RUN

NEW ITEM

ADVANCED

						(gz)	Available
NB11	/Users/merribb1/Documents/Projects/real-time-reporting/data/NB11	directory	oxford	/Users/merribb1/Desktop/mytax/flukraken2		✓	
NB03	/Users/merribb1/Documents/Projects/real-time-reporting/data/NB03	directory	oxford	/Users/merribb1/Desktop/mytax/flukraken2		✓	
ERR6913101	/Users/merribb1/Documents/Projects/real-time-reporting/data/ERR6913101_1.fastq.gz	file	illumina	/Users/merribb1/Desktop/mytax/flukraken2		✓	
ERR6913102	/Users/merribb1/Documents/Projects/real-time-reporting/data/ERR6913102_1.fastq.gz	file	illumina	/Users/merribb1/Desktop/mytax/flukraken2		✓	

/data/Samplesheet.csv

0 files (0 B in total)

skipping report making for /Users/merribb1/Documents/Projects/real-time-reporting/data/NB03/fastq_runid_77d58da2fa1dfd2379616f6e43b369b2c89c93f5_3.fastq

SHOW FULL

Rows per page:

10

1-7 of 7

Samples

sample ERR6913102

ERR6913101 flu_bc01

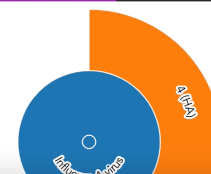
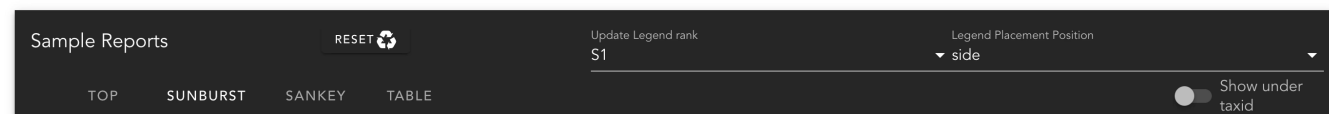
test NB03 NB11

20

Max Depth of Tax Tree

0

Min Depth of Tax Tree



Legend

Sample: sample

4 (HA) (0.160 %), taxid: 11



Legend

Sample: ERR6913102

4 (HA) (0.750 %), taxid: 11

6 (NA) (0.520 %), taxid: 11

1 (PB2) (0.290 %), taxid: 11

2 (PB1) (0.250 %), taxid: 11

Deployment Methods

The limitations of deployment and usage

Nf-core / Nextflow / TaxTriage

- Cloud (Nextflow Tower)
 - Requires regular maintenance or incur heavy costs for data storage or compute use
 - Requires data transfer time
 - Internet-requiring
- Local
 - Complex environments if running natively
 - No Native Desktop App

Mytax2

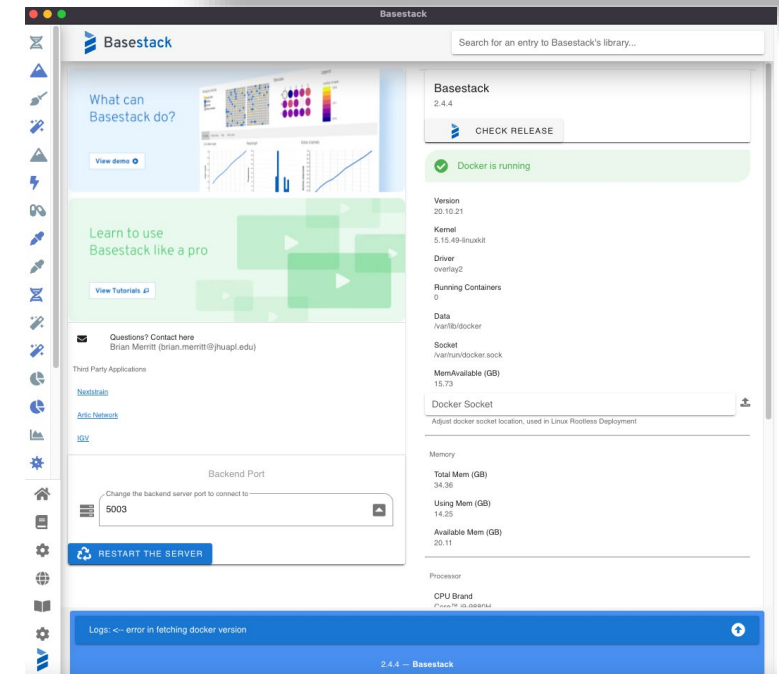
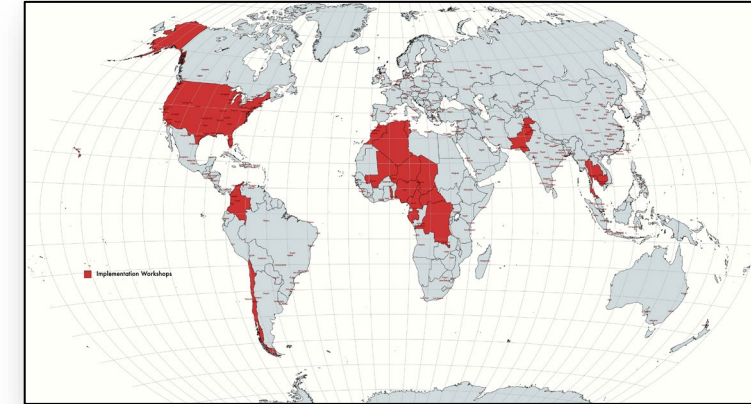
- Requires server build knowledge
- Initialization requires command-line use
- Extensive conda-build process

Basestack: An Open Software Platform for Operational Genomics

Purpose-built to deploy reproducible genomics data processing and analysis

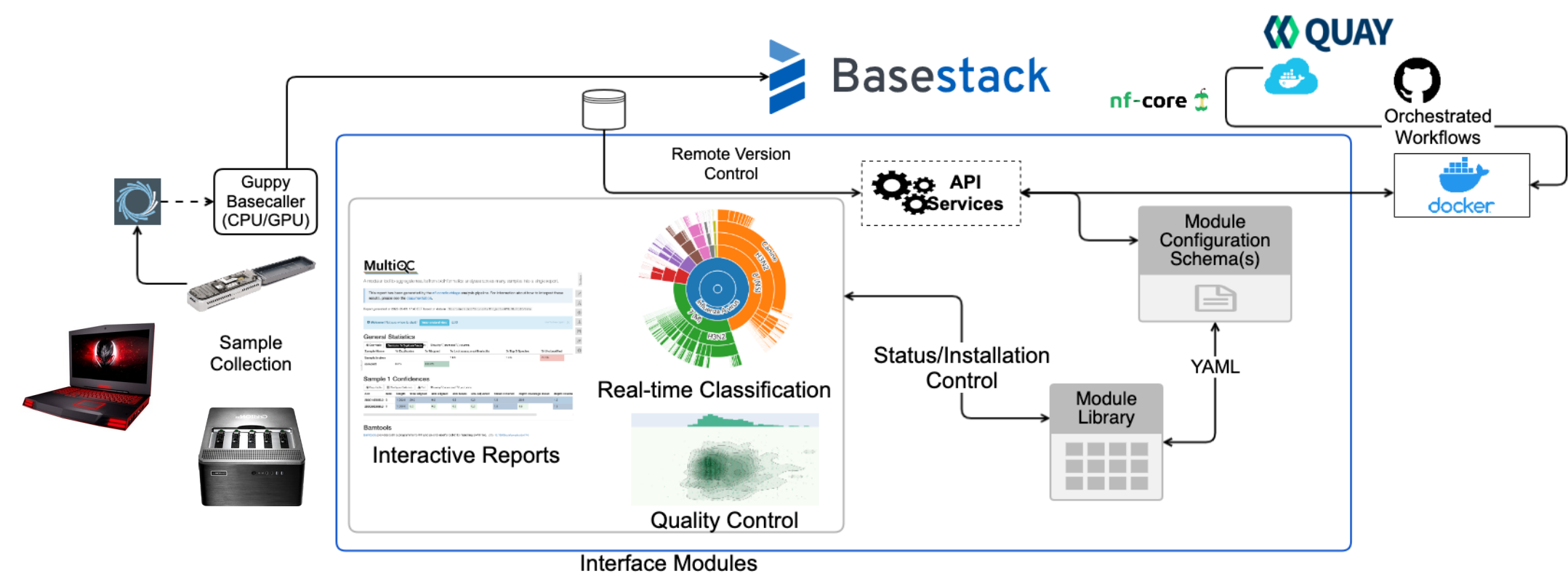
Purpose-built to enable minimally trained users to generate and analyze DNA sequencing data

- **Analysis pipelines for SARS-CoV-2, Influenza, and complex microbial community analysis**
 - ARTIC, myTax (custom meta'omics), Nextflow (ViralRecon, TaxTriage), Medaka Consensus, IVAR, etc.
- **Laptop-deployable**
 - Eliminates need to transfer large quantities of data for analysis
 - Enables fast-response deployment
- **Tested and deployed in a variety of use-cases**
 - NIH Fogarty International Center: 25+ countries
 - Defense Threat Reduction Agency (DTRA) Technical Services Group
 - Institut Pasteur du Cambodge (IPC)



Basestack

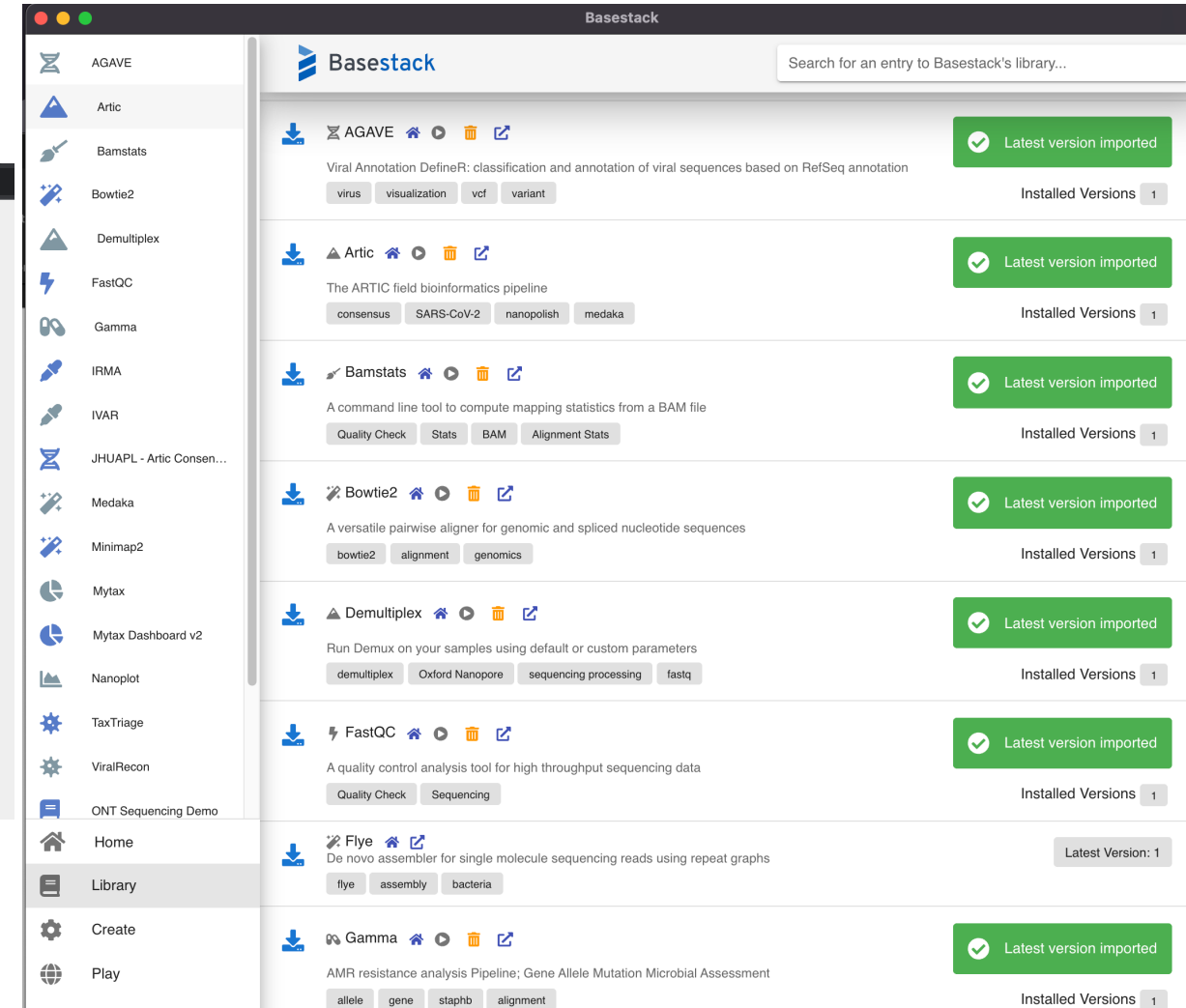
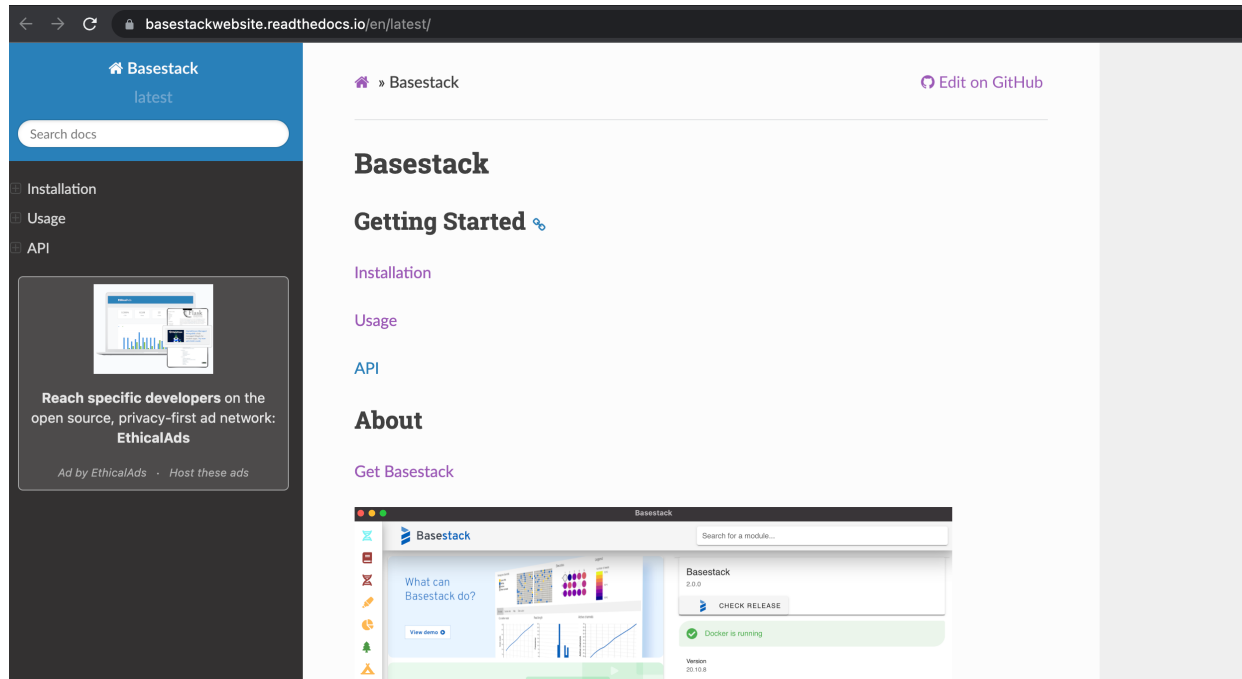
Diagram of Architecture and Usage Case



Basestack

Platform for Rapid Bioinformatic Deployment for both Remote and Local Capabilities

- Website : <https://basestackwebsite.readthedocs.io/en/latest/>
- Repository, Software: <https://github.com/jhuapl-bio/Basestack>



Basestack

Platform for Rapid Bioinformatic Deployment for both Remote and Local Capabilities

- Purpose:
 - Bioinformatic Toolkit ecosystem for running pipelines from a local desktop application
 - No terminal/command-line is required
- Fully available in GitHub as source code
 - Available on Windows, Mac OSX, Linux (Most distros)
 - Default installation requires root but non-root is available
- Dependencies
 - Docker
- Additional Features
 - Update Scheduler (Automated)
 - Real-time status reporting
 - Choose modules best fit for your needs, reduces installation requirements/size
 - Customizable modules for one-off runs
 - Cached input, reloading the app brings previous settings back up



A screenshot of the Basestack web application interface. The top bar is dark with the Basestack logo and a search bar. Below the search bar, there's a section for 'Mytax' with a 'CUSTOMIZE' button and a 'Version' dropdown set to 'Latest' (v1). The main content area is divided into two columns. The left column has a 'Module Status' section (green background) showing '3 / 3 up-to-date dependencies' and 'All dependencies installed', with 'CHECK' and 'BUILD' buttons. Below this is an 'Inputs' section with three rows: 'FASTA/Q file (Single file) -> \${file}' with a file selection button and a file path '/Users/mer...NB03.fastq (61 B)'; 'Memory Load Database -> \${memory_mapping}' with a 'Pre-load' dropdown; and 'Nodes.dmp file, gathered from ncbi taxonomy -> \${nodes}' with a 'default' dropdown. The right column has an 'Output Locations' section with a table showing two rows: 'Report From classification, in kraken-style' and 'JSON Visualization File'. Both rows show 'Access' and 'Completed' status as '1 / 1'. There's also a 'Kraken2 Report' section with a 'Create Report JSON' button.

Improvements

What we aim to achieve with Basestack

- Integrated Module Management
 - User-interface for defining parameters and inputs for all steps
 - Promotes increased community pipeline development and support
 - Interactive force-layout node-based network for linking services/commands
- Cloud Integration (Nextflow Tower and AWS)
- Non-Docker options
- Queuing System
 - Schedule and prioritize jobs based on user-preferences
- Integrated Terminal for recreating jobs from the command-line
- Improved Documentation and UI/UX
- More bioinformatics modules
- Personalized history mode

The screenshot displays the Basestack dashboard, which is organized into several sections. On the left is a vertical sidebar with navigation links: DASHBOARD, CONCENSUS, MODULE 2, MODULE 3, Show all, LIBRARY, TUTORIALS, LOGS, and SYSTEM. The main content area is titled 'Your Basestack' and includes a 'Recent Modules' section with three cards: 'Concensus' (last used 1 day ago), 'Module Title' (last used 1 day ago), and another 'Module Title' (last used 1 day ago). Below this is a section titled 'What can Basestack do?' featuring a 'View demo' button and a complex visualization of a network graph with various data points and lines. Further down is a 'Learn to use Basestack like a pro' section with a 'View Tutorials' button. On the right side of the dashboard is a 'SYSTEM SUMMARY' section with a link to 'All System Details'. This section provides details about the installed version (1.1.1), the processor (Core i7-7820HQ), memory (17.18 GB total, 5.43 GB used), and operating system (Linux darwin). It also includes a 'CONTACT US' section with developer information and a 'Third Party Apps' section listing Nextstrain, ARTIC Network/RAMPART, and IGV. The Johns Hopkins Applied Physics Laboratory logo is visible at the bottom right.

