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

**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

Partnership with CDC, Public Health Laboratories, Association of Public Health Laboratories, and Johns Hopkins University

Not for distribution
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)

<table>
<thead>
<tr>
<th>Initial Collection</th>
<th>Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>eDNA collection or Swabs</td>
<td>Technical limitations in in-field technicians</td>
</tr>
<tr>
<td>Sample preparation for Illumina/Oxford Nanopore</td>
<td>Hardware accessibility</td>
</tr>
</tbody>
</table>

- 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

---

[Map of global DNA surveillance locations]

[Sequence alignment diagram]

[DNA strand image]
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
## Base Report

### Staph

<table>
<thead>
<tr>
<th>Sample Name</th>
<th>% Staphylococcus aureus</th>
<th>% Top 5 Species</th>
<th>% Unclassified</th>
<th>% Dups</th>
<th>% GC</th>
<th>Read Length</th>
<th>% Failed</th>
<th>M Seqs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staph_1</td>
<td>91.1%</td>
<td>91.5%</td>
<td>2.1%</td>
<td>7.3%</td>
<td>32%</td>
<td>228 bp</td>
<td>9%</td>
<td>0.1</td>
</tr>
<tr>
<td>Staph_1_1</td>
<td></td>
<td></td>
<td></td>
<td>4.9%</td>
<td>32%</td>
<td>228 bp</td>
<td>9%</td>
<td>0.1</td>
</tr>
<tr>
<td>Staph_1_2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Staph_2</td>
<td>77.5%</td>
<td>81.0%</td>
<td>11.9%</td>
<td>1.0%</td>
<td>33%</td>
<td>205 bp</td>
<td>9%</td>
<td>0.0</td>
</tr>
<tr>
<td>Staph_2_1</td>
<td></td>
<td></td>
<td></td>
<td>4.0%</td>
<td>33%</td>
<td>205 bp</td>
<td>9%</td>
<td>0.0</td>
</tr>
<tr>
<td>Staph_2_2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Staph_3</td>
<td>35.4%</td>
<td>52.0%</td>
<td>38.5%</td>
<td>0.1%</td>
<td>36%</td>
<td>172 bp</td>
<td>18%</td>
<td>0.0</td>
</tr>
<tr>
<td>Staph_3_1</td>
<td></td>
<td></td>
<td></td>
<td>0.0%</td>
<td>36%</td>
<td>172 bp</td>
<td>18%</td>
<td>0.0</td>
</tr>
<tr>
<td>Staph_3_2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Staph_4</td>
<td>58.0%</td>
<td>58.6%</td>
<td>20.4%</td>
<td>0.0%</td>
<td>36%</td>
<td>159 bp</td>
<td>27%</td>
<td>0.0</td>
</tr>
<tr>
<td>Staph_4_1</td>
<td></td>
<td></td>
<td></td>
<td>0.0%</td>
<td>37%</td>
<td>159 bp</td>
<td>27%</td>
<td>0.0</td>
</tr>
<tr>
<td>Staph_4_2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Staph_5</td>
<td>0.0%</td>
<td>26.6%</td>
<td>66.3%</td>
<td>0.3%</td>
<td>41%</td>
<td>128 bp</td>
<td>27%</td>
<td>0.0</td>
</tr>
<tr>
<td>Staph_5_1</td>
<td></td>
<td></td>
<td></td>
<td>0.2%</td>
<td>41%</td>
<td>128 bp</td>
<td>27%</td>
<td>0.0</td>
</tr>
<tr>
<td>Staph_5_2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Staph_6</td>
<td>0.0%</td>
<td>19.5%</td>
<td>72.0%</td>
<td>0.0%</td>
<td>41%</td>
<td>100 bp</td>
<td>27%</td>
<td>0.0</td>
</tr>
<tr>
<td>Staph_6_1</td>
<td></td>
<td></td>
<td></td>
<td>0.2%</td>
<td>42%</td>
<td>100 bp</td>
<td>27%</td>
<td>0.0</td>
</tr>
<tr>
<td>Staph_6_2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

18 June 2024 | 11
<table>
<thead>
<tr>
<th>Acc</th>
<th>Name</th>
<th>Rank</th>
<th>Length</th>
<th>Reads Aligned</th>
<th>Abu Total Aligned</th>
<th>Abu Total Bases</th>
<th>Total Bases Adjusted</th>
<th>Breadth Genome Coverage</th>
<th>Depth Coverage Mean</th>
<th>Depth Coverage Stdev</th>
<th>Depth Coef. Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>619591_Sample_1</td>
<td>Monkeypox virus Zaire-96-l-16</td>
<td>S1</td>
<td>196858.0</td>
<td>6931.0</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>1.0</td>
<td>8.6</td>
<td>4.4</td>
<td>0.5</td>
</tr>
<tr>
<td>619591_Sample_2</td>
<td>Monkeypox virus Zaire-96-l-16</td>
<td>S1</td>
<td>196858.0</td>
<td>1003.0</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.7</td>
<td>1.2</td>
<td>1.2</td>
<td>0.9</td>
</tr>
<tr>
<td>10244_Sample_1</td>
<td>Monkeypox virus</td>
<td>S</td>
<td>197209.0</td>
<td>1546.0</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.5</td>
<td>1.9</td>
<td>2.6</td>
<td>1.5</td>
</tr>
<tr>
<td>12643_Sample_1</td>
<td>Ectromelia virus</td>
<td>S</td>
<td>209771.0</td>
<td>331.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.8</td>
<td>2.3</td>
</tr>
<tr>
<td>10244_Sample_2</td>
<td>Monkeypox virus</td>
<td>S</td>
<td>197209.0</td>
<td>218.0</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
<td>0.6</td>
<td>2.4</td>
</tr>
<tr>
<td>619591_Sample_3</td>
<td>Monkeypox virus Zaire-96-l-16</td>
<td>S1</td>
<td>196858.0</td>
<td>90.0</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.1</td>
<td>0.1</td>
<td>0.3</td>
<td>3.0</td>
</tr>
<tr>
<td>None</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>10244_Sample_3</td>
<td>Monkeypox virus</td>
<td>S</td>
<td>197209.0</td>
<td>11.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>8.6</td>
</tr>
<tr>
<td>44283_Sample_4</td>
<td>Pasteurella multocida subsp. multocida</td>
<td>S1</td>
<td>2137782.0</td>
<td>1.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>319.1</td>
</tr>
<tr>
<td>619591_Sample_4</td>
<td>Monkeypox virus Zaire-96-l-16</td>
<td>S1</td>
<td>196858.0</td>
<td>16.0</td>
<td>0.9</td>
<td>1.0</td>
<td>0.8</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>7.5</td>
</tr>
<tr>
<td>1328324_Sample_4</td>
<td>Klebsiella pneumoniae subsp. pneumoniae KPNH27</td>
<td>S2</td>
<td>5241638.0</td>
<td>1.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>511.9</td>
</tr>
</tbody>
</table>
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
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)

https://github.com/jhuapl-bio/Basestack
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

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
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