CDC’s Advanced Molecular Detection (AMD)
Sequence Data Analysis and Management

Scott Sammons
Technology Officer
Office of Advanced Molecular Detection

National Center for Emerging and Zoonotic Infectious Diseases
Office of Advanced Molecular Detection
**Advanced Molecular Detection (AMD)**

2011: “Bioinformatics Blue Ribbon Panel”

2014: Congress establishes AMD initiative

- 5-year, $30m-per-year modernization program
- Core goals:
  - Improving *pathogen detection and characterization*
  - Developing *new diagnostics* to meet public health needs
  - Supporting genomic and *bioinformatics* needs in the US public health system
  - Implementing enhanced, sustainable, *integrated information systems*
  - Developing tools for prediction, modeling and *early recognition of emerging infectious threats*

2016: AMD External Review
AMD’s Focus

- NGS & bioinformatics: transforming pathogen detection and characterization
  - At CDC
  - In state and local health laboratories
- Other technologies: MALDI-TOF, optical mapping
- Infrastructure: storage, networking, high-performance computing, cloud
- Workforce development
- Collaboration
AMD: role with State Health Labs

- AMD Portal – A centralized portal for the State Health Lab partners to gain access to CDC’s infectious disease programs.

- In FY2016, AMD awarded more than $3 million to State Health Labs through the ELC for capacity building and workforce development in next generation sequencing, genomics, and bioinformatics.
APHL Implementation Guide and QA/QC Documentation

- CDC Next Generation Sequencing Quality Team is working closely with APHL to provide QA/QC guideline documents to State Health Labs through APHL.org.
Sequencing and Analysis Workflows

- Sample Preparation
- Library Preparation
- Sequence Generation
- Sequence Analysis
- Result Reporting

**Sufficient DNA quality and quantity**
- Sufficient library quantity and proper size range
- Instrument metrics, average read quality scores and metrics
- Bioinformatics QC
- Minimum coverage of targeted regions
- Updated mutation databases and interpretative guides

Adapted from Gargis, Quality Assurance and Validation of Next-Generation Sequencing
Next Generation Sequence Analysis

• AMD sponsored/funded commercial products
  o CLCbio Workbench
  o Geneious
  o LaserGene
  o BioNumerics

• Others
  o SeqScape, Sequencher,
  o Freeware packages (windows): Mauve, MEGA, MEGAN, MrBayes, PHYLIP
  o Open Source software: over 250 packages offered though our LINUX machines and high performance computing cluster
  o In-house developed analysis pipelines
Analysis Workflow: Simple Metagenomics Example

- **Trimmomatic**: quality filtering and trimming
- **RepeatMasker**: removal of repetitive and low complexity sequences
- **BLAST**: remove hits to the human genome
- **BLAST**: keep hits to GenBank nr
- **MEGAN**: taxonomic classification based on blast hits
Analysis Workflow: wgMLST Example

Quality filtering and trimming

Spades: *de novo* assembly

Reference Mapping: map raw reads against an allele database

BLAST: contigs against an allele database

Combine Results and create trees (minimum spanning)
Analysis Workflow: Newborn Screening Example

Cutadapt: primer and adapter trimming

BWA: Alignment of reads to a reference genome

Variant Calling: GATK, FreeBayes, lofreq

VCFintersect: report only variants in area of interest
Sequencing Instrument Data Management

- Instrument directory access
  - Read access managed for Linux & Windows users
  - Write access from instrument
- Most instruments write directly to the shared storage
- Analysis pipelines read from shared storage – minimize data duplication

Average MiSeq run generates 15G – 20G of data. Most Miseqs ship with 500G hard drive. Storing runs on an external server is best option for long term and disaster recovery.
Group Data Management

- **Group Working Area**
  - **Read/write** access to all group members for base directory

- **Group Analysis Area**
  - **Read/write** access for Application-specific service accounts
  - **Read/write** access for all group members

- **Group Data Area**
  - **Read/write** access for data generators, e.g. Core Facility, OID Bioinformatics RFBS Results
  - **Read** access to “golden” set of data

- **Group Share Area**
  - **Bi-directional** share access for group collaboration

SciComp Storage

/scicomp/groups/

Group Directories

CDC Organizational Hierarchy

Analysis/

Shiny/

CLC/

NCBI-Submission/

...

Data/

BCFB/

OIDBIO/

...

Share/

in/

out/
Laboratory Workflow Data Management

- Clarity LIMS
  - End to end sample and workflow tracking
  - Stores metadata and complete sample history
  - Instrument integration
  - Reduces human error
  - Built-In QC steps
  - Allows for extensive customization (workflows/protocols + EPP)
    - Creating sample sheets for MiSeq
    - Kick off demultiplexing automatically
    - Starting a post-processing script
Clarity LIMS Workflows/Protocols

Nextera DNA for MiSeq 5.0

Workflow Status: Pending

1. DNA Initial QC 5.0
2. Nextera DNA Library Prep 5.0
3. Library Validation QC 5.0
4. Illumina SBS [MiSeq] 5.0

Illumina SBS [MiSeq] 5.0

Step 1 » Sort MiSeq Samples [MiSeq] 5.0
Step 2 » Library Normalization [MiSeq] 5.0
Step 3 » Library Pooling [MiSeq] 5.0
Step 4 » Denature, Dilute and Load Sample [MiSeq] 5.0
Step 5 » MiSeq Run [MiSeq] 5.0
Data Management: Food for Thought

- Evaluations/Pilots
  - Epidemiologic and sequence data integration and analysis
    - Cloudera Hadoop – influenza
    - Collaborative Advanced Analytics and Data Sharing (CAADS) – HIV, tuberculosis
  - Metadata and Sequence integration
    - CKAN, DSpace, Socrata
Bioinformatics Support

- Computing infrastructure: HPC, storage, virtual machines
- Central support team: bioinformaticians, sys admins, software developers
- Online forums: helpdesk, technical support, bioinformatics, programming, user groups
- GITLAB: code repository
- User groups: bioinformatics, R, sequencing, etc.
- Pipeline Development Tools: AGAVE, SciLuigi (eval)
- Workforce development: targeted to lab scientists, bioinformaticians, epidemiologists
AMD: Scientific Infrastructure

Computing/Virtualization
- Support for 250+ Linux applications
- Accommodates virtual LINUX boxes for researchers for development, testing, and deployment of scientist applications and pipelines

High Performance Computing
- ~1800 core HPC cluster
- Support of multiple queues to accommodate different types of analysis
- Support for workflows directly coupled with scientific instrumentation
- Over 1 million jobs/month

Storage
- 4 petabytes (1,000 terabytes) of networked attached storage (NAS)
- Supports all scientific instrument data, application data, and cluster data
The findings and conclusions in this presentation are those of the author and do not necessarily represent the official position of the Centers for Disease Control and Prevention. Use of trade names is for identification only and does not imply endorsement by the Centers for Disease Control and Prevention or by the U.S. Department of Health and Human Services.

www.cdc.gov/amd
twitter: @cdc_amd