Next Generation Sequencing Implementation Guide

OCTOBER 2016

APHL ASSOCIATION OF PUBLIC HEALTH LABORATORIES®

Executive Summary

Next-generation sequencing (NGS) is a technology that allows for the sequencing of millions of DNA strands simultaneously. At first this technology was cost prohibitive for all but a few academic and industry institutions, however, as the instrumentation and technology continues to evolve, the accessibility of this technology has expanded for most laboratories and for public health laboratories (PHLs) in particular.

PHLs are now using NGS applications for foodborne pathogens, infectious disease and, in some cases, newborn screening. This technology allows for greater information about pathogens leading to more informed outbreak and surveillance studies. The initial instruments for NGS were error prone and cost prohibitive for most public health laboratories, however, this changed in 2010 when Life Technologies (now Thermo Fisher) introduced a benchtop sequencer, the Ion Torrent PGM, followed shortly thereafter by the Illumina MiSeq and the Roche 454 Junior. With these instruments, a fully equipped NGS laboratory can be set up for less than $200,000 and, while they are not suitable for sequencing large genomes, their speed and relative ease of operation make them attractive options for applications that are of interest to PHLs.

This guide is designed to give PHLs an overview of issues that will need to be considered and addressed during the implementation and use of NGS. This guide is broken into five chapters that provide a broad overview of general needs and considerations for sequencing—regardless of the program—without delving into specific applications for any particular pathogen. Because the scope of public health utility of NGS technologies is still being established, this guide is designed to help laboratories keep a broad perspective in their initial set-up of the methodology. However, knowing that many laboratories will begin with the sequencing of PulseNet pathogens, there is a specific appendix (Appendix 2) that goes into more detail about the considerations and requirements for PulseNet NGS implementation.

Given the fact that this technology is quickly evolving and that applications for PHLs are rapidly developing and changing, this document will continually be amended. Additional appendices with specific pathogens will be added as standardized protocols are developed.

This document was developed in partnership with subject matter experts from our membership and CDC partners. Specific attribution can be found at the end of each chapter.

Table of Contents

Chapter 1: Preparing to Purchase a Sequencer - Instrument Selection and Laboratory Preparations....4

Introduction....4

Instrument Placement....4

Additional equipment needs for NGS....5

Personnel considerations....7

Time considerations with maintaining legacy tests in addition to NGS....8

Supplemental Table 1: List of Instruments....10

Table 2: General Lab Supplies....11

Chapter 2: Information Technology Considerations....13

Introduction....13

Data Storage Considerations....13

Data Streaming....14

Chapter 3: Workforce....15

Introduction....15

Basic Steps....15

Chapter 4: Validation....17

Introduction....17

General considerations for designing validation protocols for NGS....17

Challenges in validation....17

One laboratory's validation experience....17

Chapter 5: What to Do With the Data?....20

Introduction....20

Quality Assurance (QA) of data....20

General pipeline of the data for Illumina MiSeq NGS....20

How to begin making sense of the data?....21

Bioinformatics tools....22

Appendix 1: Information Technology Checklist Tool for NGS Implementation....23

Appendix 2: What to Expect When Implementing Whole Genome Sequencing for PulseNet....28

Background....28

Chapter 1. Before implementation of WGS for PulseNet, what do I need to do?....28

Chapter 2: Is the Laboratory Ready for Data Streaming?....31

Chapter 3: Is the Laboratory Ready for Data Storage: Where to put data?....32

Chapter 4: How to get started with NCBI....33

Chapter 5: Setting Up Validation Studies: Best Practices....34

Chapter 6: Implementation of WGS in a PulseNet Laboratory....34

Glossary of Terms....38

Table 1: Checklist for Implementing WGS for PulseNet....38

Table 2: PulseNet Area Laboratories and Contact List for WGS....39

Resources....39

Chapter 1: Preparing to Purchase a Sequencer - Instrument Selection and Laboratory Preparations

Introduction

Currently, there are a wide variety of instruments on the market that differ in price, capacity, chemistry and read length, among other things. The Illumina MiSeq is the most popular sequencer among PHLs due to its relatively low cost, low error rate and ability to handle the moderate throughput required by most PHLs. Another popular instrument is the Ion Torrent by Thermo Fischer Scientific. Some CDC programs can provide support on one instrument over another (i.e. PulseNet protocols have been developed and validated on the use of the MiSeq instrument). Initial set-up of an instrument is dependent on the manufacturer's standard protocols, this can take as little as a day or a week. Working with the manufacturer to ensure that the laboratory space has been optimized to meet the requirements and working with IT (see chapter 2) to make sure the infrastructure is in place to handle the instrument will expedite the set-up process. It is a good practice to have a checklist for each aspect of work.

Considerations for instrument purchase should include:

- Cost
- Footprint
- Technical specifications that affect expected applications of the instrument
- Throughput requirements

Table 1: List of Instruments

Company	Ilumina	Illumina	Life Technologies	Oxford Nanopore
Instrument	MiniSeq	MiSeq	Ion Torrent PGM	MinION
Chemistry	Polymerase	Polymerase	Polymerase	Nanopore
Detection	Fluorescence	Flourescence	Proton/Semiconductor	Electrostatic
Reads	22-25 / 44- 50 M	10-25 M / 24-50 M	0.2- 6 M	~500 - 2,000+
Read Length	1x75 - 2 x150 bp	36-2x300 bp	200-400 bp	10,000+
Bases/run	1.65-7.5 Gb	0.5-15 Gb	20 Mb-2 Gb	
DNA Library prep time	1 day	0.75 - 1.5 days	1 day	2 hour
Run time	7-24 hours	4-40 hours	2-7 hours	real time
Sequencing Cost/run	~$1,000+	~$560-$1,500	~ $500- $1,000	~$1,000
Instrument cost	~ $50,000	~ $100,000	~ $80,000	

See Supplemental Table 1 at the end of this chapter for a complete comparison of the most common sequencers' instrument features.

Instrument Placement

The following factors should be considered when determining instrument placement:

- **Clean/dirty areas.** Any specimen preparation required prior to library preparation and sequencing will need to take place in a molecular area that is considered “clean.” The sequencer itself should be placed in a “dirty” area.

- **Ease of access for multiple groups**. Consider who needs access to the instrument initially, as well as who may need access in the future as sequencing activities expand to include additional pathogens and library prep methods.
- **Protection from vibration, temperature and humidity**:
    - Vibration – Sequencers can be especially sensitive to vibrations. Sources of vibration commonly include other instrumentation on the same bench, opening and closing of room doors and unstable table legs. Laboratories have reported issues when construction was taking place nearby and could only run the instrument in off-hours. Due to building constraints, some laboratories purchased vibration pads for HVAC and other equipment that produced high vibrations which affected sequencing instruments.
    - Temperature – Excessive heat can affect the performance of the instrument. The room the instrument is housed in should have adequate cooling capabilities, as the instrument can become overheated and will produce some heat on its own. Direct sunlight can also lead to an increase in heat that may impact the instrument, therefore it is recommended to use a screen or shading during operation if the instrument is in direct sunlight.
    - Humidity – can also be an issue both while operating the instrument and during the library preparation process. Manufacturers recommend that sequencers be placed in an area with 20-60% humidity for optimal performance.

      Note: Laboratories located in high altitude and low humidity areas should be aware that several steps during library preparation that require drying, may need to be optimized in a laboratory to account for humidity differences.
    - High traffic areas – Instrumentation should be kept away from high traffic areas that can produce vibrations and put the instrument at risk of other issues like accidental bumping by lab personnel.

Additional equipment needs for NGS

There are several pieces of ancillary equipment that are required or are extremely useful when performing NGS.

Equipment required for DNA isolation or library preparation

- **Nucleic Acid Quantitators** – It is crucial to accurately determine the amount of starting DNA. There are several options that give highly accurate quantitation of low quantities of DNA. Depending on the applications, consider placing one in each of the designated clean areas where DNA isolation and library preparation occur.
    - Qubit fluorometer (~$2500) by Thermo Fisher is a benchtop fluorometer that using the binding of a fluorescent dye to DNA, allows for highly sensitive quantitation of DNA even at low concentrations.
    - Nanodrop fluorometer (~$12K – special until end of 2016 for $7700) by Thermo Fisher.
    - Some laboratories use a real time PCR instrument like the ABI7500 with the Femto bacterial DNA kit for DNA quantitation.

- QIAxpert is a high-speed microfluidic UV/VIS Spectrophotometer that can analyze up to 16 samples in 2 minutes. (~$10K)
- Victor X3 plate reader (~$25K) - The above options are useful for small numbers of samples. For higher throughput a plate reader, which can accommodate 96 well plates is recommended for DNA quantitation. This plate reader has both fluorescence and UV absorbance detection methods.

- **Nucleic Acid Quality Analyzers** – These instruments are necessary in order to check DNA quality and size. Successful sequencing is dependent on starting with a high quality and sufficient DNA. These instruments are highly sensitive and necessary in order to determine whether or not the starting DNA quantity is sufficient to continue on in the sequencing protocol.
    - Agilent Bioanalyzer (~$17K) or TapeStation (~$26K) - used for the size measurements of DNA fragments, library or insert sizes. It is not that accurate for the quantitative measurements of DNA. The Bioanalyzer is less expensive but requires use of a chip that runs up to 12 samples. The TapeStation uses strip tubes, which allow for greater flexibility in the number of samples analyzed at a time. If only a few samples will be run at a time, the TapeStation may be more practical and cost efficient in the long run when the cost of consumables is considered. The TapeStation also has a high throughput option (costing ~$42K), which allows for analyzing DNA samples in a 96-well plate format.
    - QIAxcel (~$34K plus $2500 for installation and training) - similar to the Agilent Bioanalyzer and TapeStation, as it determines DNA quality and quantity. The downside of this in comparison to the Agilent options the initial instrument and software operation and optimization can require some optimization, however, once this initial optimization is completed, laboratories are impressed with the speed and efficiency of the QIAxcel.
- **Thermocyclers** – DNA amplification and labeling are necessary steps during the library prep step, therefore requiring the use of a thermocycler. While there are not recommendations for a specific thermocycler, some laboratories have reported having difficulties with needing to optimize protocols for their particular thermocycler. In some cases labs resorted to purchasing a new thermocycler that was recommended for a specific CDC protocol.

Optional equipment based on the application or pathogen

- **Ultrasonicator** - Ultrasonicators are used for DNA fragmentation, which is an important step in the TruSeq library preparation method. It is not required for Nextera XT library preparation
    - Covaris M220 (~$28K) is an ultrasonicator that is capable of shearing DNA into fragments in the range of 150 bp – 5 kb.
- **Rotating shaker with a 96-well plate format**.
- **Hybridization oven** – Certain applications such as TruSeq require a hybridization oven as opposed to a thermocycler for the binding of DNA to beads during the library prep.
- **Pipettors or pipetting robots** – Several dedicated multi-channel and single-channel pipets are a necessity for NGS set-up. Depending upon the workload, a pipetting robot could be programmed to automate many manual pipetting requirements.

See Table 2 at the end of this chapter for a suggested list of general supplies and equipment. The catalog numbers and companies are merely suggestions, as alternatives for these supplies may be found from other sources. While many of these may already be in laboratories, it may be useful to have dedicated pipets and supplies for sequencing in clean rooms.

Personnel considerations

In addition to the costs of instrumentation, there are considerations around time and personnel costs associated with set-up and implementation. Personnel time required for training is something that varies from laboratory to laboratory. Depending on a staff's molecular biology background, this process may take weeks to months to get the staff proficient. If a laboratory does not have staff experienced in performing NGS, it was recommended by one laboratory to dedicate two laboratory staff during the training period to focus entirely on mastering the sequencing protocols. If a laboratory already has a staff member proficient in the molecular techniques and protocols for sequencing, training new staff can be approached by having a new staff member observe the experienced staff member until they are comfortable with the protocol. Once they are comfortable with the procedure, the trainee will perform the procedure while being observed from start to finish.

When the trainer determines that the trainee is competent, the trainee then performs the procedure independently with immediate supervisory confirmation that the procedure was done correctly. Therefore, the minimal times for a procedure to be performed under observation and then performed independently for training is three times, however, this will be very dependent on the experience level of the trainee, and often additional training will be necessary. Some training programs may also include an additional "test" of the trainee where they are given "blinded" samples that include CDC certification strains or other sequenced isolates in order to verify the accuracy of the trainee's results. One laboratory's example of time required to achieve proficiency is listed in Table 3. It should be noted that the times listed are in an ideal setting and comes from a laboratory with high molecular biology technical abilities.

Table 3: Minimum time required for staff training

Procedure	Time required for training (h=hours or d= work days)	Additional time required to become proficient
DNA extraction		
Qiagen Blood Tissue Kit	6h	4h
Qiacube	6h	4h
MiSeq and PGM Ion Torrent		
DNA to Starting a run	4.5d	3d
Simple Data handling and analysis		
FastQC	<1h	<1h
Download from BaseSpace	<1h	<1h
NCBI Upload	<1h	<1h
Understanding Basic MiSeq metrics	<1h	<1h

These times assume previous experience with general molecular biological techniques (i.e. pipetting, PCR, centrifugation).

Time considerations with maintaining legacy tests in addition to NGS

In addition to the personnel training time that will be required, it is important to remember that in the beginning, concurrent laboratory procedures and analysis may be performed on the same isolates. For example, pulsed field gel electrophoresis (PFGE) will be performed in addition to whole genome sequencing (WGS) for all PulseNet pathogens. With this concurrent analysis, there will be both additional financial costs and staff time that need to be considered.

For example: A cost and labor analysis performed in 2014 at The Wadsworth Center compared PFGE to WGS. Assuming 18 bacterial samples were multiplexed on a MiSeq run. This level of multiplexing generally gave adequate read depth and coverage to perform SNP based cluster analysis. The PFGE cost analysis is comprehensive and includes all costs associated with running PFGE including staff time per sample. The WGS-MiSeq is not comprehensive, it includes the costs for reagents, setup/culture, isolation and DNA extraction and technician time for sample preparation through MiSeq run. It does not include costs associated with data analysis and sharing, including the personnel time to complete these tasks. For both analysis, technician cost was estimated at $50,000 salary plus fringe and indirect costs.

Table 4: Cost comparisons between WGS and PFGE*

	Cost per sample	Hours of labor per 18 samples
PFGE	$65.12	17.3
WGS-MiSeq	$278.21	21

*All numbers were calculated at Wadsworth; numbers may vary from lab to lab.

In this example, they have separate staff for PFGE and WGS testing with very limited cross training. This situation is unlikely to change during the transitional period as their PFGE staff are occupied fulltime with PulseNet PFGE duties. In cases of smaller labs with limited personnel where the same staff may be expected to perform both WGS and PFGE, longer training periods may be required and longer turn-around-times (TAT) may be seen.

As personnel changes are considered for the adoption of NGS, also think about the differences in TAT between current technologies and NGS. There are significant differences in TAT between the currently used methods and WGS. The summary of TAT is shown in the table below. TAT assumes no repeating of samples and does not take into account time to acquire multiple samples if batching is required (i.e., 18 samples batched to run on the MiSeq). These values will be laboratory dependent when taking into consideration testing algorithms and different methodologies.

Table 5: Approximate Turnaround times for current technologies and NGS

WGS-MiSeq	4 to 6 d
PFGE (wet bench time)	2 d
RT-PCR	<1 d
Pyro-sequencing	1 d

TAT for WGS is from DNA sample to reporting. TAT for PFGE it is from plug to reporting. TAT for RT-PCR and Pyro-sequencing is from sample to reporting.

Contribution:

This chapter was developed with input from the following staff members at The Wadsworth Center:

Patrick Van Roey, Ph.D., Director of the Scientific Cores
William J. Wolfgang, Ph.D., Faculty Bacteriology Laboratory
Lisa Mingle, Ph.D., Faculty Bacteriology Laboratory
Matthew Shudt, Sequencing Core
Kara Mitchell, Research Scientist
Samantha Wirth, M.S. Research Microbiologist
Michelle Dickinson, Molecular biologist

Supplemental Table 1: List of Instruments

Company	Ilumina	Illumina	Illumina	Illumina	Life Technologies	Life Technologies	Life Technologies	Pacific Biosciences	Pacific Biosciences	Oxford Nanop
Instrument	MiniSeq	MiSeq	NextSeq 500	HiSeq 2500	Ion Torrent PGM	Ion Torrent S5	Ion Proton	Sequel	RS II	MinIOI
Chemistry	Polymerase	Polymerase	Polymerase	Polymerase	Polymerase	Polymerase	Polymerase	Polymerase	Polymerase	Nanop
Detection	Fluorescence	Flourescence	Fluorescence	Fluorescence	Proton/Semi-conductor	Proton/semi-conductor	Proton/semi-conductor	SMRT	SMRT	Electro
Reads	22-25 / 44-50 M	10-25 M / 24-50 M	130-400 M/260-800 M	up to 4 B /86 B	0.2- 6 M	3 - 80 M	60-80 M	1 M	55 K	~500 2,000
Read Length	1x75 - 2 x150 bp	36-2x300 bp	75 - 2x150	36-2x125 bp	200-400 bp	200 -400 bp	200 bp	20,000+	20,000+	10,00
Bases/run	1.65-7.5 Gb	0.5-15 Gb	16-120 Gb	64-900 Gb	20 Mb-2 Gb	0.6-15 Gb	10 Gb		0.5-1Gb	
DNA Library prep time	1 day	0.75 - 1.5 days	0.75 - 1.5 days	1-3 days	1 day	1 day	1 day	6+ hours	6+ hours	2 hour
Run time	7-24 hours	4-40 hours	11-29 hours	1-11 days	2-7 hours	2-4 hours	2-4 hours	30 min - 6 hours	30 min - 6 hours	real ti
Sequencing Cost/run	~$1,000+	~$560-$1,500	~$1,000-$4240	~ $1,000-$9,800	~ $500-$1,000	~ $1,000+	~ $1200			~$1,0
Instrument cost	~ $50,000	~ $100,000	~$250,000	~ $650,000	~ $80,000	~ $60,000	~ $149,000	~300,000	~750,000	

Table 2: General Lab Supplies

COMPANY	ITEM NAME	CATALOG #
Fisher Scientific	Latex or Nitrile Gloves	19-177-520-24
	Vortex	
	Microcentrifuge w/ and without strip tube adaptor	
	Table top centrifuge	
	96-well format PCR Thermocycler System	
	96-well cold blocks	
Rainin	Single and Multi-channel Pipettes	
Rainin	Aerosol resistant pipette tips	
Brand Tech Scientific Inc.	PCR 8-tube strips, caps clear	781332
Life Technologies	PCR Plate, 96-well, semi-skirted, flat deck	AB-1400
	Sterile, nuclease free 1.5 ml micro-centrifuge tubes	
Bioexpress	GeneMate PCR sealing mats	T-3161-1
USA Scientific	TempPlate pierceable sealing foil, sterile	2923-0110
Axygen	25 ml disposable reagent reservoir, sterile	RES-V-25-S
BIO-RAD	Sealing Roller	MSR-0001
Ambion	Ultra-Pure Water (Molecular Biology Grade) Free of Detectable DNase, RNase and Protease	
Life Technologies	Quant-iT dsDNA High-Sensitivity Assay Kit	Q33120
Costar	50 ml Reservoir	4871
Costar	Assay Plate, 96-well, Black Polystyrene, non-treated	9315
Falcon	50 ml Falcon Tubes	35-2098
	25 ml Serological Pipette	
	Conical tubes	
Life Technologies	DynaMag-96 Side magnet magnetic plate (buy 2)	12331D
Sigma-Aldrich	Ethyl alcohol, Pure (200 proof, for molecular biology)	E7023
Beckman Coulter	Agencourt® AMPure® XP beads (60 ml, 450 ml)	A63881 or A63882

COMPANY	ITEM NAME	CATALOG #
Qiagen	Qiagen Buffer EB (10 mM Tris-Cl, pH 8.5)	19086
Thermo Fisher Scientific	Qubit dsDNA High Sensitivity (HS) Assay kit	Q32854
	Qubit® Assay Tubes	Q32856
Sigma-Aldrich	Tween-20	P7949
	Isopropanol	
Fisher	LENS PAPER 4X6IN BOOK; 600/PK	11-996
	Sodium Hypochlorite (household strength –typically 6-9% bleach)	
Fisher	TRAY PP BLUE 10 QUART	11-394-454
Benchmark Scientific, Inc	MyBlock Mini Digital Dry Bath/ Block, 15 x 1.5ml/2.0ml/Block, 24 x 0.5ml	BSH200*
Labsource	EZ-Read 8"Thermom -10 to 150C full imrsn	T70-053

The companies and catalogue numbers and supply list are merely suggestions. There are many comparable products that will work just as well.

Chapter 2: Information Technology Considerations

Introduction

IT set-up can be one of the most challenging parts of the NGS implementation process. Laboratories are encouraged to talk to their IT departments prior to purchasing and installing instrumentation. IT input is required for each step of the NGS process from instrument set-up to data transfer and analysis and everything in between. One essential rule of thumb for IT engagement is that they should be consulted as soon as possible. See resources for links to Illumina's guide for IT support as well as Thermo Fisher Scientific's guide.

Points where IT engagement is necessary include:

- Instrument set-up
- Investigation of data storage options
- Data streaming capabilities
- Software installation

Encouraging and arranging a meeting between the instrument manufacturers and IT staff early on could help mitigate downstream problems. In addition, having IT staff present with laboratory staff when the company comes to set-up and install the instrument can be incredibly useful. Appendix 1 includes a suggested template developed by the Missouri PHL for IT staff to use to help facilitate the conversation between an IT group and the instrument manufacturer.

Data Storage Considerations

Data storage is a consideration that laboratories should address prior to procurement of the instrument. Sequencing runs generate large amounts of data that must be stored and backed-up, potentially for years. Different PHLs have used the options below based on cost and ease of use.

- Local data storage on external hard drives – The advantage of this option is that it requires relatively little input from IT. It does not require any extra bandwidth or data streaming capabilities. Placement and storage of these local hard drives may eventually pose difficulties, as well as the potential for physical damage/destruction resulting in data loss. Another issue when using USB drives is the potential for viruses to be transferred between computers; proper precautions should be taken if this is the route chosen to store and backup data. Also, USB drives are not considered a secure long-term solution, as the drive can be compromised and the ability to retrieve data lost.
- Data can also be stored on cloud services. Illumina instruments come with an account for BaseSpace, a cloud based storage and analysis platform that is proprietary to Illumina instruments. Data is streamed from the instrument to BaseSpace and can be analyzed, stored and shared. BaseSpace comes with 1 TB of free storage, with additional storage available for purchase in one or 10 TB increments. Other cloud services such as Amazon Web Services may also be considered.
- In cases where internet security concerns do not allow for outside data streaming or when file storage infrastructure is well developed, some laboratories may use local servers for storage or scalable cluster and storage systems custom-designed for NGS analysis

Data Streaming

Data streaming is an important part of the sequencing protocol that allows laboratories to upload sequences for storage or to various databases and analysis pipelines.

Common concerns and potential solutions for data streaming

Data security in streaming

Concern: One common IT concern includes the security of streaming data with patient identifiers.

Potential solution: If a laboratory will have all samples de-identified for sequencing, this should be explained to IT. Typically, the samples are de-identified prior to sequencing.

Concern: Another common concern with data streaming is whether this will open a laboratory's network to potential security breaches.

Potential solution: IT can to build a separate network only for data streaming. This would provide added security but will take time and resources.

Potential solution: Only allow streaming in one direction into the various cloud storage spaces and databases and not allow any data to stream back into the instrument.

Potential Solution: Secure File Transfer Protocols (sFTP) sites are a potential solution if it is not possible for IT to allow incremental data streaming. These would allow a laboratory to share large amounts of data with partners, such as CDC, without the issues of data streaming. However, single transfers of large amounts of data require a lot of bandwidth, require manual actions to initiate transfer, and may be quite time consuming.

Data streaming bandwidth capabilities

Sequencing data files are very large. Streaming the data to any of the above data storage options will require networks to be able to handle this amount of data. There have been instances where networks have become overwhelmed. When using Illumina's BaseSpace and some cloud based storage, the data can be streamed in real time which is more manageable for most networks. If a sFTP is chosen for a data storage option, the data will be streamed in one bulk amount, which leads to a higher likelihood of overwhelming the network. This can be somewhat mitigated by doing these large data drops in non-peak network times.

This chapter was developed from contributions from Minnesota Public Health Laboratory Division and Missouri State Public Health Laboratory.

Chapter 3: Workforce

Introduction

NGS is a multi-step and multi-day process that has a number of opportunities for errors to be introduced. It is more complex than many commonly used methods in PHLs and management needs to consider laboratory set-up and staff training to incorporate NGS into the existing laboratory workflow. Depending on the pathogen you are sequencing, protocols will vary. Many of the various pathogen protocols are still in the pilot phase and comprehensive protocols do not yet exist. PulseNet pathogens do have specific protocols, which can be found in Appendix 2. As other CDC programs move from the pilot stages to PHL implementation, this document will be updated with other appendices. Below are the basic steps for workflow from isolation to analysis for NGS. While each pathogen may have variations, all will follow this basic workflow.

Basic Steps

Step 1: Nucleic acid isolation. Laboratories may be starting from an isolate or a clinical specimen, which will affect the protocol for DNA/RNA isolation. Each pathogen will likely require some troubleshooting to establish an effective protocol of nucleic acid isolation. Specifically, as testing shifts between bacteria, mycobacterium and viruses, etc., adjustments on kits and standard protocols will need to be taken. While protocols may vary for optimal nucleic acid isolation, the goal is to obtain high quality, pure DNA. While quantity is important, quality should not be sacrificed. Contamination of starting material can lead to poor sequencing data quality. The ancillary DNA quantitation equipment listed in Chapter 1 are crucial to ensure the starting material will be sufficient to continue through the sequencing process. Laboratories should also consider whether they use any robotic or high-throughput protocols for DNA isolation and ensure that these protocols are optimized and have proper quality assurance.

Step 2: Library preparation. Library preparation prepares the DNA for sequencing. NGS requires DNA to be random fragments of similar size. These random fragments are then fluorescently labeled for the sequencing process. Library preparation is typically a straightforward process, however, depending on laboratory instrumentation (e.g. thermocyclers) and laboratory conditions (such as the humidity issue discussed in Chapter 1), optimization will be required. Ancillary equipment listed in Chapter 1 such as the Qiagen QIAxel and Agilent Tapestation allow laboratories to preview the quantity and quality of the nucleic acid and success of the library preparation prior to moving on to the next step.

The library preparation kit chosen can affect the quality of sequencing data. There are a wide variety of library preparation kits. Choosing the appropriate kit for a sequencing run will depend on the sequencing instrumentation and the pathogen being sequenced.

Step 3: Sequencing. The actual sequencing is fairly straightforward and will be dependent on the sequencer in the laboratory. Once a sequencing run has been initiated, it does not require any hands-on time. Time for sequencing runs will vary based on the sequencer, the number of samples, the number of reads, etc. Several hours to days should be allotted for sequencing time.

One point of possible sequencing optimization that may need to occur is organism multiplexing (i.e. loading more than one organism-type) of the sequencing runs. It is recommended that in the initial implementation, restrict sequencing runs to all of the same pathogen. However, considering one of the biggest hurdles of NGS for smaller laboratories is the ability fill up a run, combining different pathogens on the same run would make the runs the most cost effective. Laboratories have had differing experiences in the multiplexing of pathogens. The most successful cases of multiplexing include similar bacteria in terms of genome size. More common solutions include filling runs by concurrently sequencing archival samples to provide valuable historical data. Illumina recently announced that it will offer lower capacity cartridges for sequencing which will make smaller (<24 samples) runs more cost effective.

Step 4: Data Analysis. From the sequencer, the data is stored and must eventually be analyzed. See chapter 5 about data analysis for more information. Each pathogen will have its unique data analysis workflow though there will be some similarity.

As laboratories begin to consider implementing workflows for various pathogens, please consider the following:

- Allow time for troubleshooting during the various steps any time a new sequencing protocol begins. At each step there will be a need to adjust the existing protocol and this time should be built in.
- Allow for additional staff training time
- As conventional methods are kept in parallel with NGS, ensure the laboratory information management system (LIMS) can track each sample through the parallel workflows that will be conducted.

Chapter 4: Validation

Introduction

As with any laboratory test, quality assurance and control is an important aspect of establishing NGS. Since NGS is a more complex method generating large amounts of data, quality assurance and control approaches are more complicated than standard laboratory tests. Validation for each application is an important component of quality assurance and control. To date, validation protocols developed and utilized by PulseNet have been the most widely applied.

General considerations for designing validation protocols for NGS

- Consider all steps in the validation process including nucleic acid extraction, enrichment, library preparation and data analysis.
- Establish thresholds for the alignment processes to flag reads that are not considered high enough quality.
- Establish accuracy by considering several NGS parameters including depth of coverage, number of reads, and Q score.
- Robustness and reproducibility: Consistency in performance independent of operator and reagent lots.

Challenges in validation

- **Limited bioinformatics capacity.** In the beginning, laboratories may have to rely heavily on outside assistance for bioinformatics analysis. The pathogens that could be validated may be limited by the capabilities and time of the collaborators for bioinformatics data.
- **Cost.** In addition to time needed, the cost of implementing a new workflow and validation of a new pathogen can be prohibitive. Consider the long-term funding needs to sustain and implement a new program.
- **Reagents.** In some cases reagent inconsistency can provide unnecessary challenges in the validation process. The inclusion of a control organism can help internal quality assurance.
- **Criteria for evaluation of sequencing results.** It is important to determine in advance the criteria that will be used to evaluate the sequencing results and what thresholds for acceptable/unacceptable results may be applied to aid in the interpretation of external analysis summaries. External partners providing analysis may vary in their approaches to data analysis and their acceptability thresholds.

One laboratory's validation experience

Background:
One PHL's validation included 50 well-defined isolates comprised of 10 of each of the following organism-types: *Salmonella spp.*, O157 Shiga toxin-producing *Escherichia coli* (STEC), non-O157 STEC, *Campylobacter jejuni* and *Listeria monocytogenes*. The set of isolates tested included eight previously analyzed isolates from the FDA-CFSAN GenomeTrakr PT panel from 2014, for which FDA had developed a bioinformatics pipeline, and 42 isolates sequenced in parallel with the CDC PulseNet Laboratory to assess the genomic accuracy of sequencing data generated. Isolates previously sent to the CDC PulseNet laboratory for sequencing were preferentially selected for the

validation panel to reduce the amount of sequencing support needed. Isolates were divided across four runs with 10-17 isolates per run, with the organism composition determining the number of isolates included. The sequencing runs were performed by two analysts and included multiple reagent lots. A quality control organism was not included in the runs, as this is not common practice for this lab. PhiX was spiked into runs at 5% and the percent aligned reviewed in Sequence Analysis Viewer to identify potential issues with the run performance, in addition to other metrics discussed below.

Validation testing

Validation runs were performed by two scientists, and data shared with FDA GenomeTrakr and CDC PulseNet laboratories for review. The four validation runs were performed over one month. Data were shared for external partner review following cursory quality assurance checks that included the predicted coverage for each organism sequenced by comparing the number of raw reads to the genome size for the pathogen, the cluster density on the flow cell, clusters passing filter, ≥Q30, and estimated yield. Each metric was compared to the guidelines provided by CDC PulseNet to determine whether results were acceptable for sharing with external partners.

Data review and summation, SOP preparation

Results were received from the FDA GenomeTrakr and CDC PulseNet partners within one to seven months of data sharing. Challenges to timely data sharing included the diversity of the panel selected and the need for contiguous reference genome data for each organism to determine genome coverage and to assess sequencing results through single nucleotide polymorphism comparison. The bulk of CDC-supported data analysis also coincided with the busiest season for enteric bacterial outbreaks, over the summer of 2015.

Data summaries received included the following metrics:

> CDC PulseNet: The number of mapped reads and average read length; median insert length; number of contigs; N50; longest contig and total genome length. CDC sequencing results were compared to the data from this lab to determine the number of SNPs differences, as an assessment of accuracy.
>
> FDA CFSAN GenomeTrakr: Results are assessed by mapping the sequencing data to a reference sequence; metrics that are returned include the number of reads, mean read depth, percentage of mapped reads and the number of SNPs vs. the reference sequence. Results were presented in a comparison with other sequencing sites to demonstrate acceptability relative to other laboratories.

Reports and data evaluations received from external partners were reviewed and summarized in a final report, which took approximately 5 months due to competing priorities requiring staffing to be dedicated to other critical tasks, and the time required for the LAB SOP and validation package preparation.

This laboratory is currently at the phase of final review and approval of standard operating procedure for routine testing, with the complete validation package moving through the final approval process by the QA Department and senior-level administrators. This laboratory has exclusively used the Nextera XT procedure and modifications described in the CDC PulseNet methods and have developed an internal SOP for this method.

References

Gargis A. (2013) Assuring the quality of next-generation sequencing in clinical laboratory practice. Nature Biotechnology 30(11): 10.1038/nbt.2403

Gargis, A. (2016) Quality Assurance and Validation of Next-Generation Sequencing. APHL Annual Meeting Preconference workshop https://www.aphl.org/conferences/proceedings/Documents/2016_AM_PreCon/01-Gargis.pdf

This chapter was developed with contributions from Lauren Turner, PhD, Foodborne and Advanced Pathogen Characterization Lead Scientist from the Virginia Division of Consolidated Laboratory Services.

Chapter 5: What to Do With the Data?

Introduction

An additional challenge with NGS is handling the massive amount of data. From determining the quality of the data that comes off of the sequencer to the analysis of it, the best practices and tools are still being developed. This chapter will attempt to give users a broad overview of what considerations they should take in how to look at the quality of the data and also an idea of the tools currently available to analyze the data. Bioinformatics support is one of the biggest needs PHLs have noted. With NGS being a relatively young technology, there are no real guidelines or best practice to guide the bioinformatician on which tools to use and thresholds to apply to filter the data. It is often up to the bioinformatics scientist to test and validate different tools, either by doing simulation or by testing samples with known features or mutations. Good bioinformatics support would require a person who has been specially trained in this field beyond some of the basic short term courses.

Quality Assurance (QA) of data

After all of the various steps, there are multiple places where things can go wrong and problems can occur. Contamination of the starting material is a common problem and sometimes isn't discovered until after the run when data analysis is performed. Confirmation of the size distribution of DNA fragments after shearing is important, as well as quantification and fragment size analysis of the DNA after library prep. Loading the right concentration onto the MiSeq is very important for optimal cluster density and therefore data output. A few points to consider in determining if data is good quality:

1. A first assessment of the quality of the data can be gathered by looking at the run metrics summary from the sequencer itself. More specifically, a good quality run would generate an amount of reads within the expected range of the machine, with a majority of Q Scores above Q20 or Q30 and a low error rate based on the sequencer's positive control (PhiX for Illumina instruments). Having a large amount of reads classified as "undetermined" after the de-multiplexing step can indicate that the sequencing run was of lower quality, issues exist related to the indexing (barcodes) of the reads, the sample sheet is incorrect or problems occurred with the flow cell during sequencing.

2. More advanced QA analysis can be performed by third party software such, as FastQC, which is one of the most popular (available on BaseSpace, Ion Reporter or Stand-alone version). A rapid or unexpected drop in per-base sequence quality is indicative of potential sequencing issues. Deviation from the expected GC content of the organisms sequenced, K-mer over-representation, or high level of duplicated sequences is usually the result of contaminating sequences, PCR artifacts or the presence of untrimmed adapter sequences.

3. Lastly, other quality control steps can be implemented later in the downstream analysis, such as percent of mapped and unmapped reads, mapping quality, overall depth of coverage, coverage uniformity, percent of genome covered, insert size distribution and level of homozygote vs. heterozygote SNP calls.

General pipeline of the data for Illumina MiSeq NGS

During the MiSeq run, the MCS software (installed on the instrument) can be used to view an overview of the quality statistics for the current run. The metrics can be viewed after cycle 25. For more detailed metrics during and after the run, the Sequence Analysis Viewer (SAV) software can be downloaded to an off-instrument computer and provides a greater detail of the sequencing metrics such as Qscore, intensity data by cycle, raw cluster numbers by lane. Illumina's BaseSpace can also be used to view the same metrics, as they are streamed in real-time during the run if the option to stream to BaseSpace was selected. Finally, after the run is complete, the Illumina Reporter software will execute according to pre-run selected options. Samples will be demultiplexed and the software will generate a summary table that includes per sample numbers such as number of reads, megabases sequenced, and mean quality score. There are many online tutorials on using the software mentioned above as well as how to interpret the run metrics at http://support.illumina.com.

After the run is completed and the samples are demultiplexed, one (or two if a run was a paired-end run) fastq files per sample will be left. It will probably be in a compressed format with .gz suffix to save space. The fastq format contains both the base calls and the quality scores for each base in every read for each sample. Next, assess the quality of the reads. There are several quality assessment software packages available, but the most widely used is FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). FastQC assesses the quality of the reads using ten different metrics including per base sequence quality, per base GC content, sequence duplication and overrepresented sequences. For most downstream data analysis applications, trimming and filtering reads for quality is recommended before proceeding further. Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic) and FASTX (http://hannonlab.cshl.edu/fastx_toolkit) are two commonly used tools for this purpose. These tools will identify and remove contaminants as well as trim the sequence based on quality score. Full execution of these tools requires coding expertise.

How to begin making sense of the data?

1. There are built-in apps available on BaseSpace (Illumina) and Ion Reporter (Ion Torrent) that can do many basic bioinformatics analysis such as metagenomic profiling, SNP calling and de novo genome assembly. There are also some commercial softwares available (CLC bio, DNA star, Geneious) that can be used, with minimal bioinformatics training to perform more intricate analysis or build workflows (a.k.a. pipelines). These are generally expensive to use. A free but less user-friendly alternative is also available in the form of a web server (usegalaxy.org). Finally, most of the day-to-day tools used for bioinformatics are open source (free), but require knowledge of Unix system and command lines to be installed and operated.

2. There are three main types of analysis that can be performed depending of the type of analyses needed. They are described below.

    1. **Resequencing Analysis** - the sequencing of samples for which a close relative has been fully sequence before, and is mostly used to identify high quality single nucleotide polymorphisms (SNPs) and short insertion/deletion sequences (INDELs). This is done by simply mapping sequencing reads over a reference genome and looking for positions where base calls on the reads disagree with the reference nucleotide at that position. The SNPs can be used to build a SNP alignment or make whole genome MLSTs assignments for phylogenetic purpose (outbreak tracking), or to identify changes causing phenotypic alteration (diseases or antibiotic resistances). Variations of genome re-sequencing are gene expression analysis (RNAseq), transcription (ChipSeq) and others (TNseq, RadSeq, etc).

    2. **De novo Genome Assembly** - This is required when no close relative genomes are available or when one is interested in structural genomic differences, gene presence/absences analysis or copy number variation (CNV).

    3. **Metagenomics** - sequencing directly from environmental or clinical samples. The resulting reads will be a mixture, at different proportions, of all the genetic material

of the organisms present in the starting material. This can either be amplicon based metagenomic (16s for instance) or whole genome. The major step in metagenomic analysis is the taxonomic assignment (a.k.a. binning) of the sequencing reads or assembled contigs.

Bioinformatics tools

Conventional tools

- Applied Maths BioNumerics - See the PulseNet Appendix 2 for information about BioNumerics
- There are several comprehensive commercial products available for NGS sequence analysis. BioNumerics, CLCbio, Geneious, and LaserGene are commonly used software packages. A variety of open source tools are available but most are Linux command-line and are more suitable for experienced bioinformaticians.
- Program specific tools are being developed at CDC and will be added to this document once they have been piloted.

Note: There seems to be a misconception that bioinformatics can be easily learned and with minimal involvement. Bioinformatics has now become a profession on its own. Every project has different requirements so custom pipelines have to be written and validated according to the specifications and the goals of the project. There are many bioinformatics workshops, online classes and tutorials that someone can attend or consult to get acquainted to bioinformatics, but these are usually focused on very specific types of analysis and data. Until more user-friendly pipelines and platforms are made available to researchers and clinical laboratories, it is advisable for laboratories who do not have a bioinformatician on staff, to collaborate with other states or CDC laboratories with bioinformatics experience.

For those that have greater bioinformatics support, know that all bioinformatics analysis can be performed using available open source software. In fact, many of these software programs are often the same programs that have been integrated and packaged into commercially sold bioinformatics suites. Advantages of using open source software, in addition of being free, include full control on the version that is installed on the laboratory system, timing of when to apply updates, and availability to new software that have not yet been integrated to commercial package. Some drawbacks of free software programs are they tend to be difficult to install, prone to software bugs, provide limited documentation for IT approval and support and are mostly command line-based only (no graphical interface available).

This chapter was developed with contributions from Pascal Lapierre, PhD, Bioinformatician from Wadsworth Center of the New York Department of Health and Scott Sammons from the Office of Advanced Molecular Detection at CDC.

Appendix 1: : Information Technology Checklist Tool for NGS Implementation

Question	Is this covered?	Document location where this covered	Notes
Baseline Requirements			
Data Classification - Restricted, Private, Public			
a. Can we get data through a sunshine request?			
b. Where is the single source of truth of the data?			
What are regulatory compliance requirements from department? Is there a data retention policy?			
Are interfaces required to connect to our Data Center? If so, what are they? What is the level of integration required with other State applications or databases?			
Does the department require compliance with any affected state/ federal regulations? If so, what are they? a. EXAMPLE: The state's data must be logically segregated from other customers. b. Will the State's data be "co-located" with other customers? If so, does the service provider have a documented set of controls that it uses to ensure the separation of data and security information? Please provide a copy.			
Does department have recovery time objectives and recovery point objective requirements?			
Is the solution compliant with Accessibility standards?			
Ask for 3 references from service provider where they are providing 'like' services to similar sized entities (preferably government)			
What is the risk to the Agency for failure?			

Question	Is this covered?	Document location where this covered	Notes
Compute			
Request system architecture diagram from vendor.			
Storage			
Network			
Request network diagram from the vendor.			
Security			
What security protocols or procedures does the vendor follow?			
How does the vendor ensure the data is safe? (EXAMPLES) a. All data stored on the service platform shall remain the property of the state. At the termination of this agreement for any reason, all data shall be transferred back to the state b. The contractor's hosting facilities and hardware must be protected from unauthorized access. c. The contractor must implement commercially reasonable security safeguards to prevent unauthorized access to and protect the integrity of the system interfaces, application software, systems files and state data within the contractor's Data Center Services environment, disaster recovery locations, and any off-site archiving location(s). d. The contractor must provide advanced security technologies such as managed firewalls, encryption, authentication, intrusion detection, site scanning, and the performance of security audits/penetration tests. e. The contractor must provide server-side anti-virus and anti-spam protection that is updated as necessary to be current.			

Question	Is this covered?	Document location where this covered	Notes
f. The contractor must monitor, review, respond to and report access violations and log security alerts and notification activity.			
g. Will the contractor provide a list of personnel with system-level and/or group-level authorities to the system application components and/or data. An annual review must be conducted to revise the list. Personnel no longer needing this access level should have access removed within 2 business days. *Gartner says this is not typical. An audit may be requested that shows that only authorized personnel have access to their systems*.			
h. The contractor personnel with access to sensitive data and/or operations must successfully complete a criminal background check prior to being given access.			
Does solution encrypt data at rest?			
Does solution encrypt data in motion?			
Does the vendor scan the application using a tool to check for security vulnerabilities like cross site scripting, etc. How often is this scanning done? Dynamic and Static?			
Does the department require a 3rd party assessment of the site? If so, how does the vendor typically handle the assessment? Is there a cost?			
Does the department require a 3rd party assessment of the application? If so, how does the vendor typically handle the assessment? Is there a cost? What can we audit with the vendor			
Support and Service Levels/Service Offerings			
What is vendor Breach Notification Policy?			

Question	Is this covered?	Document location where this covered	Notes
Vendor is required to notify through NOC. Any issues? a. Require to notify within 24 hours b. What is vendor Outage Notification Policy? Gartner says less than 1 hour			
Retention logs - Example: “The X system must have adequate monitoring, logging and assurance mechanisms including the monitoring and notification and reporting of application usage, security events, faults, and system health that are required to verify security safeguards and controls are in place and in effect.”			
How does State terminate the contract? a. How much time does provider have to return the data? b. What “transition out” support will the service provider deliver?			
Does Vendor require background checks of employees and subs?			
Does solution physically reside in US Data Center?			
Does vendor perform disaster recovery and business continuity planning exercises? If yes, briefly describe.			
Does provider give notice of upgrades or system changes that would impact performance? What is their policy?			
E-discovery and legal requests - what is their process?			
What type of logs are available?			
What level of access to logs and reports will ITSD/users have?			
What is the roadmap for the Cloud solution? (when and what enhancements will be provided in the future)			
What kind of training will be provided to the “End Users”?			

Question	Is this covered?	Document location where this covered	Notes
Management and DevOps			
What type of authentication does app require?			
Who maintains the application?			
What is state responsible for?			
Any browsers requirements/ dependencies?			
What is the experience level of the service providers' staff? What technical certifications do they hold?			
Price/Billing			
What is the contract vehicle?			
What is financial viability of the service provider?			
How is the solution priced? (per unit? If so, what are the units?)			

Appendix 2: What to Expect When Implementing Whole Genome Sequencing for PulseNet

Background

In 2013, PulseNet, the National Molecular Subtyping Network for Foodborne Disease Surveillance, began a pilot project to conduct real-time surveillance of all Listeriosis cases in the United States using whole genome sequencing (WGS) technology. This study and others have demonstrated the utility and feasibility of next generation sequencing technology for timely surveillance and outbreak detection of *Listeria monocytogenes* and other foodborne pathogens. (References)

The PulseNet USA network has implemented a gene-by-gene based approach, whole genome multi-locus sequence typing (wgMLST), as the primary method used to analyze whole genome sequence data for cluster detection. This method has several advantages over SNP and kmer based analytical methods in that it allows for: 1) standardization enabling comparison of data generated in different laboratories; 2) provides definitive naming scheme of sequence types for tracking subtypes over time; 3) method is computationally more amendable to program standardization and analytic automation; 4) simple and faster analysis requiring no substantial bioinformatics skills; 5) similar or better discriminatory power and epidemiological concordance compared to PFGE for cluster detection and outbreak investigations.

In addition to pathogen subtyping, WGS may replace most traditional reference testing by consolidating various workflow practices used by public health laboratories (PHLs). Antimicrobial susceptibility, serotype and other virulence markers can be predicted from the sequence data.

This document is intended to assist PHLs with the implementation of WGS for PulseNet. It is projected that whole genome sequencing will be implemented in PulseNet laboratories using a tiered approach through 2018. This guide will be a useful resource for bench level laboratorians, managers and laboratory directors by providing information to assist PHLs with the set up and implementation of WGS within the PulseNet network.

References

1. Use of Whole Genome Sequencing and Patient Interviews To Link a Case of Sporadic Listeriosis to Consumption of Prepackaged Lettuce. JFP, V 79, No5, 2016, p806-809

2. Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation. CID, 2016, Aug

Chapter 1. Before implementation of WGS for PulseNet, what do I need to do?

1.1 Requirements for Laboratory Instrumentation/Equipment/Reagents

The PulseNet Standard Operating Procedure, PNL32, provides specific requirements for the laboratory instrumentation, equipment, reagents and supplies for performing WGS using the standardized PulseNet protocol. All references to PulseNet Standard Operating Procedures (SOPs) in this document can be found on the PulseNet SharePoint Site in the Library of PulseNet Documents under QA/QC Manual.

- Laboratory instrumentation considerations at local/state levels (For more information see Chapter 1, p. 4-5 of the main document for information including cost and uses of ancillary equipment)

Several pieces of ancillary equipment are needed to perform the PulseNet WGS protocol from DNA extraction through library preparation and sequencing. Additional ancillary equipment specified in PNL32 includes: a Qubit® 3.0 Fluorometer, NanoDrop 2000 UV-Vis or equivalent spectrophotometer and a 96-well thermal cycler with heated lid. Some PHLs have connected the NanoDrop and the Qubit to one computer to make moving files easier (data from these two pieces of equipment are needed for PulseNet WGS worksheet). There is also a standalone NanoDrop instrument available from Thermo Fisher that does not necessarily require computer hook-up. (http://www.nanodrop.com/ProductOneOverview.aspx).

A few more factors to bear in mind when considering the MiSeq for a laboratory:

- For all PCR-based applications, including WGS sequencing, clean and dirty rooms are advised to minimize cross contamination. The MiSeq should be placed in the “dirty area” of the laboratory.
- Ensure the laboratory has sufficient bench space for equipment and processing of specimens. The MiSeq pushes a lot of air out of the back, so do not set any instrumentation behind it, if working on an open bench. Also, no centrifuges or other vibration-producing equipment can be placed on the same bench as the MiSeq due to the instrument’s sensitive optics. Some laboratories have placed their MiSeq instrument in a separate area to minimize any accidental bumping to the instrument while in use.
- It is advised to purchase an uninterrupted power supply (UPS) battery back-up pack for the sequencer in case power is disrupted during a run. Illumina can provide specifications on the appropriate UPS for the MiSeq.
- Ensure that a dedicated circuit line is installed for the instrument (Illumina site preparation guide, page 10).
- Some laboratories have requested a site visit with Illumina field application specialists prior to the installation of the instrument to provide recommendations for instrument placement and other workflow processes.

1.2 Staffing Requirements for PulseNet activities

It is expected that once WGS is fully implemented, the technology will combine several workflows from the enteric reference and/or PulseNet laboratories into a single workflow thereby making it cost neutral for staffing requirements. However, availability of pure cultures is still required for the current PulseNet WGS protocol. Existing laboratory staff can be trained to generate and analyze WGS data for PulseNet. PulseNet Central will provide push button analysis tools in BioNumerics 7.6 that do not require a bioinformatician for data analysis. Most state laboratorians trained at CDC to perform WGS have had experience in PFGE or MLVA. However, in the near-term while full implementation is being adopted at the state and local levels, it is expected that PFGE and WGS will be done concurrently. PFGE is expected to be phased out of the network in 2018-2019.

1.3 Workflow Considerations

Currently, DNA can be extracted for WGS from the same fresh culture plate that is used for PFGE. A fresh culture with minimal laboratory passage is required for DNA extraction prior to WGS. The laboratorian is expected to perform WGS and share sequence data with the CDC within seven working days after receiving the *Listeria* isolates in the PFGE laboratory. The turn-around-time is calculated from the receipt of the isolate at the WGS laboratory to pushing data to CDC via BioNumerics or, if BioNumerics 7.6 is not available in the laboratory, then sharing sequence data with CDC via Illumina BaseSpace or PulseNet's FTP site. Until WGS has been fully implemented, STEC, *Campylobacter* and *Salmonella* sequencing do not need to meet the 1 week turn-around-time (TAT) unless CDC requests expedited sequencing due to outbreak investigations.

1.4 Requirement Considerations for Information Technology (IT)

(For more in depth information about overall IT concerns, see chapter 2, pp. 10-11 in the main document.)

One of the greatest burdens of WGS implementation will be the development of an IT infrastructure for the MiSeq sequencer and subsequent data storage and access of analysis pipelines. Having good internet connection speeds to push WGS data out of the PHL to CDC in a reasonable amount of time and support for installation of BioNumerics 7.6 and setting-up connected databases is critical. Depending on each PHL's needs there may be further demands on IT support, particularly if PHLs decide to store and analyze data locally. Dependent on individual state requirements and needs, IT support for SQL, SQLite, MySQL and Oracle databases may be necessary for WGS analysis in BioNumerics 7.6.

- **What connection speeds are recommended when we receive a MiSeq and start doing sequencing analysis?**
  Network connection speeds (for streaming data to BaseSpace or uploading to CDC) will need to be a minimum of 10Mbps; however, 100Mbps-1Gbps is recommended for it to work consistently. To test internet upload and download speeds, try freely available websites such as speedtest.net or talk to laboratory IT support to determine speeds.

- **Some labs have mentioned an issue with people in their building not being able to connect to the internet during data uploads to BaseSpace. How can this be resolved?**
  CDC PulseNet has heard of this happening for the occupants in the building where the MiSeq is housed. The connection becomes saturated with just MiSeq data when participants are streaming data to BaseSpace and/or uploading to CDC particularly if streaming is interrupted during the run and a reconnection to BaseSpace is needed to complete data transfer. If this becomes an issue, talk to IT about segmenting the network or providing a dedicated connection for the MiSeq.

- **What is the MiSeq configuration?**
  The MiSeq runs with Windows 7 installed; the version on the MiSeq is specific to the instrument and should not be updated or modified in any way. Often IT groups have Group Policy Objects (GPOs) and will push down changes on computers on their network—an exception will be needed for the MiSeq if IT does this. Never turn on Windows Firewall or Windows Defender on the MiSeq. If needed, Illumina does have recommended anti-virus software to use and special configurations for that software since automatic scans or automatic updates cannot be done. Suggestions for anti-virus and details about MiSeq configuration are available in the MiSeq User Guide.

- **What are overall computer requirements?**
  The sequencers do not require additional computers to properly run. For BioNumerics 7.6, a computer (laptop or desktop) is required with at least 8 GB of RAM, a 64-bit processor and Windows 7 or later. Additionally, it is not recommended to install BioNumerics 7.6 on the same computer where BioNumerics 6.6 or earlier version databases are saved, since

opening the old databases in BioNumerics 7.6 leads to automatic data conversion and potential data loss. If possible, CDC recommends purchasing a new computer or having a dedicated computer for BioNumerics 7.6. However, BioNumerics 7.6 can be configured to access a different BioNumerics home directory if it must be installed on the same computer as other versions of BioNumerics.

1.5 Procurement Process

- **What funding mechanisms are available to support WGS for foodborne pathogens?** PHLs, who are eligible, should continue to use to use the "Epidemiology and Laboratory Capacity for Infectious Diseases (ELC)" grant funding as a major source of funding for implementation of WGS activities for PulseNet and other CDC infectious disease programs. Within the ELC grant, some funding for WGS implementation and support is provided through CDC AMD funds via the PulseNet transformational AMD project or directly from the Emerging Infectious Diseases Program (EIP) project. For 2016, a new source of funding through the President's Combating Antimicrobial Resistance Bacteria (CARB) will also provide funding for infrastructure and capacity building for WGS activities through the NARMS project. Finally, laboratories participating in the FDA GenomeTrakr project will receive some support from FDA for sequencing food and environmental isolates. Some support for training may also be available through APHL.

- **Sole source justifications** PulseNet Central at CDC has developed letters to support states with sole source procurement. A regional Illumina Sales Representative can also provide a manufacturer's Sole Source letter. An example of a sole source letter can be found in the Resources at the end of this Appendix. To request a letter of support, email the pulsenet@cdc.gov inbox. A reagent contract with Illumina may also be needed for the routine procurement of reagents over $5,000. A contract may be negotiated with the PulseNet preferred pricing through a Regional Sale Representative, which will also require a Sole Service Letter from Illumina.

- **Procurement considerations** Be prepared to dedicate time every day or weekly to follow the paperwork through the state fiscal process. It is not uncommon for a government's procurement process to span several months. Some state PHLs have had to undergo lengthy contract negotiations with Illumina and their state procurement office. Additionally, for some states, specific letters signed by the vendor and CDC may be required for the purchase of a sequencer.

Chapter 2: Is the Laboratory Ready for Data Streaming?

(See chapter 2, pp. 10-11 of the main document on considerations of data storage.)

2.1 What is cloud storage?

Cloud storage refers to a type of data storage where data are stored on off-site servers maintained by a hosting company (Amazon Web Services, Microsoft Azure Cloud Services). These hosting companies are responsible for keeping data in a secure and protected environment which can be accessed at any time.

2.2 What is Illumina BaseSpace?

Sequencing data generated by Illumina's sequencers (MiSeq, HiSeq, NextSeq, MiniSeq) can be directly transferred through an internet connection (i.e. "data streaming") to Illumina's cloud service provider, BaseSpace. This cloud platform can house sequence data for storage and processing of data for further upstream analysis. Currently, one terabyte (TB) or less of data (per account) is free to users. For additional information and pricing, see Illumina's website. Often, states have required special permissions to stream WGS data over their network for temporary storage on Illumina BaseSpace. It is strongly recommended that IT departments be engaged early on to discuss network requirements and established security measures for data streaming from the MiSeq to Illumina BaseSpace.

- Security measures to be aware of when using BaseSpace:

  BaseSpace uses Amazon Web Services (AWS). Transfer of data from the instrument to BaseSpace is encrypted and goes over Secured SSL Transport. The data in BaseSpace meets AWS security standards. For more details visit http://www.illumina.com/documents/products/technotes/technote_basespace_security.pdf.

  BaseSpace accounts may be set up free of charge for individual users. There are other account options available on BaseSpace, at an additional cost: http://www.illumina.com/informatics/research/sequencing-data-analysis-management/basespace.html

Some states are not allowed to use BaseSpace due to their institution's IT security policies. For these states, most have set up external hard drives to transfer sequencing files from the MiSeq instrument to another computer in order to send to sequence data to CDC via the FTP site.

Some states have set up separate port for connecting BaseSpace to their IT network.

2.3 What other options are available if my laboratory cannot access cloud services?

Raw read files can be transferred to CDC via the PulseNet FTP site. Please see SOP PND19 for instructions. Please contact PulseNet@cdc.gov to request the password to access the FTP site. Direct data transfer to the CDC will be available through BioNumerics 7.6 once the organism specific database is available.

Chapter 3: Is the Laboratory Ready for Data Storage: Where to put data?

Sequence raw read files will be stored at NCBI indefinitely, unless there is a requirement for storing a local back-up copy. In that case, an external storage device like an external hard drive, external server, computer with at least 2 TB storage space, or cold-storage cloud services (Illumina BaseSpace, Amazon) should be considered (see the IT infrastructure section below). The wgMLST allele designations and de novo assemblies will be stored in the local BioNumerics client database and also uploaded to the central PulseNet SQL databases for indefinite storage.

- Should an external storage computer/storage device be considered? The hard drive on the MiSeq is only 500 GB so external storage options should be considered since that hard drive will fill up within approximately 10 runs. An average yield (fastq files only) for 16 isolate *Salmonella/E.coli* run with 500 cycle chemistry is 6- 12 GB. However, the actual run folder is much larger due to files that are not needed unless troubleshooting a run. More runs can be stored on the MiSeq hard drive if contents are routinely deleted from the following folders on the Data (D) drive: Illumina>MiSeq Output>"your run folder">Thumbnail images, Illumina>MiSeq Output>your run folder>Images, Illumina> MiSeq analysis>"your run folder," and Illumina>MiSeq Temp. Since all of the data that passes PulseNet QC will be uploaded to NCBI, a copy of data is stored there. If the laboratory does not have a local data retention policy, then delete data from the MiSeq as needed and NCBI will serve as the data repository.

However, never delete local data before double-checking that it has been successfully transferred to its permanent storage location. If uploading data to BaseSpace, 1TB (terabyte) of free storage is available and supplementary storage space can be purchased for an annual fee. Additionally, if local storage is a requirement, an external hard drive can be used or a local storage system that has 2x3 TB drives that are mirrored (in a RAID array) on a computer that has at least 32 GB RAM. PulseNet Central at CDC can provide additional specifications. BaseSpace onsite hub is another option. This is a local computer that provides local data storage (http://www.illumina.com/informatics/research/sequencing-data-analysis-management/basespace/basespace-onsite.html). Storage space can also be purchased from cloud service provider, such as Amazon.

Chapter 4: How to get started with NCBI

The workflow for uploading the allele designations will be similar to what is currently done for uploading PFGE patterns and demographic information to PulseNet Central databases. There will be a new push button workflow in BioNumerics 7.6 to upload sequence raw reads and limited metadata to NCBI after the data has passed initial quality checks.

4.1 Setting up a NCBI Account

Until BioNumerics 7.6 can be used for all organisms (currently BioNumerics 7.5 can only be used for *Listeria* data submission) to upload sequence data to NCBI, the laboratories will have to submit the metadata that accompanies the sequences to NCBI separately. This process is called “biosample submission”. In order to be able to submit biosamples, a user account needs to be created. Please refer to the PulseNet SOP PND18, step 5.1 for instructions.

4.2 Navigating NCBI to upload my sequence data

The fastq-files are uploaded through BioNumerics 7.5/7.6 (see BioNumerics 7.5/7.6 training manual for instructions) but can only be done if a validated allele database is available to the PHLs. Currently such a database is only available for Listeria. For other organisms, CDC performs the fastq-file upload after the PHL shares the files via BaseSpace or the PulseNet ftp-site and uploads the metadata to NCBI.

4.3 Getting data back from NCBI

NCBI assigns multiple accession numbers to the submitted data. The most critical ones are the SAMN and SRR numbers. SAMN number is assigned to the biosample (metadata) of each strain. SRR number is assigned to the raw sequence data (fastq). If upload was done through BioNumerics 7.5/7.6 these accession numbers can be retrieved with a push of a button and will automatically populate the appropriate fields in the database. If biosample submission was done separately through the NCBI submission portal, an automated email from NCBI biosample helpdesk will return the SAMN numbers in a .txt-file. The SAMN numbers need to be uploaded to the NCBI accession field in the national PFGE databases.

To find sequence data on NCBI for any particular strain, navigate to http://www.ncbi.nlm.nih.gov/, copy and paste the SAMN number or the WGS_id assigned to the strain by CDC to the “Search” field and click on “Search” to access the “Results found” page. To view the metadata for the strain, click on the “Biosample” link. To view the information about raw data such as file size and upload date, click on the SRA link. To download sequences from NCBI to a local computer, first download the

SRA Toolkit at http://www.ncbi.nlm.nih.gov/Traces/sra/?view=software and follow the instructions for download. Alternately, fastq files may be directly downloaded through the European Molecular Biology Laboratory Nucleotide Archive (http://www.ebi.ac.uk/ena) by searching for associated reads by the SRA ID from NCBI.

4.4 How to view analyzed data on NCBI and looking up various trees (hpSNPs/Kmer)

In order to view the NCBI kmer trees and hqSNP results, navigate to http://www.ncbi.nlm.nih.gov/pathogens/, click on “Download results” and follow the instructions in the CDC training document titled “NCBI Genome Workbench and Mega tools.”

Chapter 5: Setting Up Validation Studies: Best Practices

(See chapter 4 of the main document, pages 15-16.)

Performing validation for the PulseNet WGS procedure is NOT required by CDC but may be performed if required by the laboratory. The individuals may either perform validation prior to or in conjunction with becoming certified. PulseNet Central will accept validation study data for the organisms for which certification is currently available (Refer to SOP PNQ08 for a list of available certification sets).

In order to get started, email PulseNet@cdc.gov with the subject line “WGS validation” to request assistance. PulseNet QA/QC personnel will respond by including an excel file template that will include a list of strains from the laboratory that have already been sequenced by CDC. Complete the template with a list of isolates the laboratory will sequence. If additional isolates from the laboratory need to be sequenced by CDC to have adequate representation of all organisms, please include those in the list as well.

When performing the sequencing for the validation study isolates, please adhere to the following isolate labeling format on the MiSeq sample sheet (SOP PNL32): **strainID-V-labID-MXXXX-YYMMDD_** where the strainID is the CDC assigned WGS ID, V is to indicate that the sequence belongs to validation, LabID is the laboratory ID assigned by PulseNet Central, MXXXX is the Machine ID and YYMMDD_is the date with an underscore at the end. Submissions that are not named in the proper format will not be accepted. Results must be submitted to the appropriate Bioproject on BaseSpace or shared on the PulseNet FTP site.

Sequences run by the PulseNet participating laboratory will be compared to the same isolates sequenced at CDC. Evaluation criteria will primarily be based on the average quality of R1 and R2, average coverage, average insert size, number of hqSNP differences and number of allele differences compared to the CDC-sequenced reference. A summary report will be emailed to the individuals included on the submission email. Results can be expected in 2-3 months after submission of the complete sequence set.

Chapter 6: Implementation of WGS in a PulseNet Laboratory

States that currently have sequencers are already submitting sequence data to PulseNet Central and have been supported with limited amount of AMD funding for reagents. Moving forward in 2016 more substantial amounts of funding will be available for instrumentation, staffing and reagents through the CARB initiative. The level of reagent support to be provided for sequencing clinical isolates will depend on the “expected” number of isolates received by the PHL annually. The target for 2017 is to sequence 100% Salmonella received in addition to *Listeria* and STEC. For data analysis (BioNumerics v7.5), we started piloting the *Listeria* wgMLST database in 10 states in November 2015. We expect that the pilot for the STEC and *Campylobacter* databases will start September 2016. We expect these databases to be available to all states towards the end of 2016. The *Salmonella* database is scheduled for release by early 2017 and databases for the remaining PulseNet organisms (*Vibrio*, *Cronobacter*, *Yersinia*) to become available later in 2017 and 2018.

6.1 Training

PulseNet Central will offer one week long, free-of charge training workshops 2-4 times per year depending on the demand and availability of funding. These workshops are announced by APHL and posted on the PulseNet SharePoint site and will cover the entire laboratory procedure and the BioNumerics 7.6 analysis workflow. The class size is limited to ten students. If the CDC workshops are full, contact an area laboratory for training. As part of the sequencing instrument installation, companies offer limited training mainly demonstrating the instrument interface, but not including library preparation. Illumina offers also more advanced training workshops at their San Diego headquarters for a charge. If a laboratory is interested in joining GenomeTrakr, FDA/JIFSAN also offers training workshops at a cost that are announced on the JIFSAN website.

6.2 Proficiency Testing and Certification

Certification and proficiency testing (PT) of isolates for WGS works similarly to PFGE and MLVA certification and PT. There are separate certifications for the lab procedure (generation of fastq files) and BioNumerics analysis. For instructions, please refer to the PulseNet SOP PNQ08. For the laboratory certification, one strain per organism (*Listeria*, STEC, *Salmonella* and *Campylobacter*) is included in the certification set. For cost efficiency, the certification set can be run with routine isolates, given that typically 16-20 isolates can be included in each MiSeq run. The raw reads for the certification strains are submitted to PulseNet Central via BaseSpace or the PulseNet ftp-site for quality assessment and analysis. The routine isolates included in the certification run may be submitted to PulseNet provided the certification files for that particular organism pass the CDC quality assessment. The analysis certification consists of two steps: 1) QC of fastq files provided by PulseNet Central and determination which files and accompanying metadata should be uploaded to NCBI, and 2) assessment of quality metrics for analyzed data provided by PulseNet Central in a BioNumerics bundle file

Starting in January 2017, PulseNet and GenomeTrakr will conduct annual PTs together so that those labs that participate in both networks only need to perform one PT per year. The PT strain set will consist of six isolates of two species, four of which will be the same serotype. The PT files will be uploaded to a PT database and evaluated using criteria agreed upon by CDC and FDA. For instructions, please refer to PulseNet SOP PNQ09.

6.3 PulseNet Naming and Nomenclature

The whole genome MLST nomenclature (pattern names) is currently being developed and will be based on a subset of the allele information. These pattern names will be assigned in BioNumerics 7.6. As of now, we expect to maintain the current PulseNet cluster code naming system.

- **How will a PFGE pattern library cross-reference with the new WGS pattern?** PulseNet laboratories are sequencing isolates representing the top PFGE patterns for all organisms in the respective national databases. While WGS data is not directly backwards compatible, (i.e., cannot determine the PFGE pattern from WGS data), representative sequence data from all top PFGE patterns will be available for comparison purposes.

6.4 Best practices for integrating subtyping- state perspectives

Some PHLs have integrated WGS activities into the PFGE laboratory while other PHLs have designated core facilities to centralize the library preparation and sequencing. PulseNet area laboratories are useful resources and should be able to provide guidance.

- **PFGE to WGS: useful tips to make workflows efficient** (See chapter 1 and 3 for more information)
    - Cross train as many people as possible in PFGE and WGS. There is a tremendous amount of organization that is required by the WGS protocol; multiple laboratory numbers (PFGE, WGS, NCBI), data that has to be captured in the WGS worksheet (NanoDrop and Qubit) and managing the extracted DNA (in the freezer, maximizing flow cell space). Each PulseNet laboratory will have to design systems to manage the organizational tasks and this is best done by a person familiar with both PFGE and WGS processes.
    - Create a log to track progress of isolates throughout the WGS process. This is especially useful if multiple people are involved in the various steps from DNA extraction to sequence submission to PulseNet, as well as the multiple identifiers assigned to a single isolate (state ID, WGS_ID, SAMN ID and SRA ID). Information tracked can include fields such as DNA extraction date, NCBI upload date, NCBI BioSample number export for PFGE database, date isolates are submitted to the CDC calculation engine and retrieved, and date the analyzed data (de novo assembly and allele calls) are uploaded to the PulseNet national database. The New York State template can be found under "Resources" at end of this Appendix.
    - When making PFGE plugs, consider doing DNA extractions at the same time since the same plate can be used. Plugs can be washed while the DNA extraction reaches the centrifugation and short incubation steps. These two activities can occur in the same work area. Some state PHLs try to perform the WGS PCR steps on Monday, the bead cleaning step on Tuesday and follow with the WGS run starting on Wednesday. This will allow the PFGE to stay on schedule. Alternatively, some laboratories have batched WGS extractions to 1-2 times per week.
    - Use export tools in BioNumerics to add metadata from BioNumerics 6.6 into BioNumerics 7.5/7.6 after completion of PFGE and PFGE patterns are assigned.
- Maintaining two databases-useful tips
    - Tracking log noted above is used to indicate which isolates are being WGS sequenced. NYS tracks the date when metadata is imported into BioNumerics 7.5 and dates of all the steps in the process since overall workflow spans several days and multiple people are involved.
- LIMS and BioNumerics 7.5/7.6-useful tips
    - Metadata from a LIMS system can be exported and linked to fields in BioNumerics using export tools in BioNumerics. Information related to exporting and linking Excel files into BioNumerics can be found on the PulseNet SharePoint Site under Library of PulseNet Documents in the BioNumerics training folder.

6.5 Reference laboratory testing: What a WGS Reference laboratory will look like for enteric pathogens

CDC Perspective:

Currently, tools are being developed within BioNumerics v7.6 to identify genus/species/subspecies and virulence genes as well as predict serotype and antimicrobial resistance from the WGS data. These tools will replace current phenotypic and PCR methods for characterization.

State Perspective:

- Remember that WGS will replace most reference and subtyping activities in the PHL's. For that reason all personnel working with foodborne pathogens should be trained in WGS.
- In New York State, WGS is being used to report the identification and drug resistance prediction of *Mycobacterium tuberculosis* isolates. Studies were done on 96 retrospective samples and 50 prospective samples to validate intra-assay reproducibility and specificity. A link to the New York State guidelines for validation of next generation sequencing based methods can be found under the "Resources" section of this document.

6.6 Guidance with FOIA's: what to expect

CDC Perspective:

Currently PulseNet Central still considers genomic data "experimental" for all other organisms except for *Listeria*. Consequently, only *Listeria* wgMLST trees are released for FOIA requests. For other organisms, the phylogenetic trees are only released if the genomic data is referred to on a web posting made by CDC. These postings also provide some interpretive language for phylogenetic trees posted onto SharePoint. PulseNet linelists are routinely released for FOIA requests and do contain the NCBI accession numbers for the sequenced isolates. CDC expects the state health departments to adhere to these guidelines, i.e. not to release phylogenetic hqSNP trees unless WGS data has been mentioned in a public CDC web posting.

State perspective:

For FOIA requests, New York State PulseNet laboratory has some experience with FOIA requests. The laboratory submits WGS interpretive reports along with accompanying dengrograms. State accession numbers are redacted by Records Access Office but anonymized IDs are not redacted.

6.7 PFGE

6.7.1 When to drop PFGE?

The PulseNet USA steering committee has formed a working group to come up with guidance how best to phase out PFGE in the network. No laboratory should drop PFGE for any of the PulseNet-tracked organisms until given go ahead by the PulseNet USA steering committee.

6.7.2 What happens to national PFGE databases

CDC's goal is to combine the national PFGE and wgMLST databases so that all information is in one place. However, this process has not yet begun, neither is there a firm time when it will happen or whether it is even feasible. For the time being continue performing all PFGE-related work in BioNumerics vs. 6.6 (or 5 if lab is still using that version) and perform WGS-related work in BioNumerics 7.5/7.6.

6.8 When to drop conventional/molecular serotyping? (CDC)

Drop conventional/molecular serotyping once alternate methods (serotype finder in BioNumerics) are validated and implemented in the laboratory.

Glossary of Terms

Allele	A variant of a gene which arise by mutation, insertion, deletion or recombination and found at the same position on an organism's chromosome
BaseSpace	Illumina cloud-based storage solution for sequence data with a direct connection to Illuminia's sequencing platforms.
BioNumerics	A bioinformatics software application developed by Applied Maths, Inc
Calculation Engine	Software that works with the BioNumerics client software and runs on a high performance computing cluster at the CDC that cleans the data and does analyses like assemblies and allele predictions on the WGS data
GenomeTraker	A FDA developed network of regulatory and public health laboratories performing whole genome sequencing on food and environmental isolates
hqSNP analysis	High quality single nucleotide polymorphism analysis. An analytical, sequence based method for subtyping bacterial strains that utilizes base position comparisons of the study sequences against a closely related reference sequence. The primary analysis method for the GenomeTrakr network.
MiSeq	A benchtop sequencer instrument manufactured by Illumina, Inc
NGS	Next Generation Sequencing, which is the newest sequencing technology using "sequencing by synthesis" as opposed to the Sanger method. It allows for faster sequencing and greater coverage of genomes.
WGS	whole genome sequencing. Refers to data output when sequencing all the DNA from a single isolate
wgMLST	whole genome multi-locus sequence typing. An analytical, sequence based method for subtyping bacterial strains using all detected open reading frames (genes) of the bacterial genome for analysis. The primary analysis method for the PulseNet network.

Table 1: Checklist for Implementing WGS for PulseNet

□	Download and read SOPs located on the PulseNet SharePoint Site
□	Discuss with IT Department set up, storage and data streaming requirements
□	Acquire the ancillary equipment and supplies and sequencer
□	Dedicate time every day to follow the paperwork through the state fiscal process for procurement of instrumentation
□	Ensure the laboratory has adequate storage space to handle all new equipment and supplies coming into the laboratory
□	After sequencer installed, contact CDC and/or PulseNet Area Laboratory for training
□	Request certification panel from CDC by emailing pulseNet@cdc.gov
□	Purchase and/or update version of BioNumerics to v7.6 (Applied Maths)

Table 2: PulseNet Area Laboratories and Contact List for WGS

Western Area Laboratory	Roxy Meek	Roxanne.Meek@doh.wa.gov
Mountain Area Laboratory	Justin Nucci	justin.nucci@state.co.us
Central Area Laboratory	Central Area PulseNet inbox	Health.PFGE@state.mn.us
Midwest Area Laboratory	Steve Dietrich	Dietrichs@michigan.gov
Southeast Area Laboratory	Christina Moore	Christina.Moore@tn.gov
Mid-Atlantic Area Laboratory	Lauren Turner	Lauren.Turner@dgs.virginia.gov
Northeast Area Laboratory	Northeast Area PulseNet inbox	NYAreaLab@health.ny.gov
CDC PulseNet	Laboratory protocol and BaseSpace troubleshooting	Pulsenetngslab@cdc.gov
CDC PulseNet	Other WGS related questions (certification, internal validation IT, bioinformatics support requests)	PulseNet@cdc.gov
Illumina Public Health Liaison	Dan Schoeffner, Lead Scientist	dschoeffner@illumina.com
Illumina Technical Support		techsupport@illumina.com 1.800.809.4566

Resources

1. New York State Guidelines for Validation on Next-generation Sequencing Based Methods
2. WGS Tracking log (New York State)
3. BioNumerics manual: http://www.applied-maths.com/knowledge-base/manual-bionumerics-version-7
4. Illumina manual
5. CDC Letter of support for sole source procurement

Association of Public Health Laboratories

The Association of Public Health Laboratories (APHL) works to strengthen laboratory systems serving the public's health in the US and globally. APHL's member laboratories protect the public's health by monitoring and detecting infectious and foodborne diseases, environmental contaminants, terrorist agents, genetic disorders in newborns and other diverse health threats.

This project was 100% funded with federal funds from a federal program of $ 228,529. This publication was supported by Cooperative Agreement # NU60OE000103 funded by the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of CDC or the Department of Health and Human Services.

APHL®

8515 Georgia Avenue, Suite 700
Silver Spring, MD 20910
Phone: 240.485.2745
Fax: 240.485.2700
Web: www.aphl.org