Setting the Course: Virginia's experience navigating information technology and bioinformatics needs for whole genome sequencing

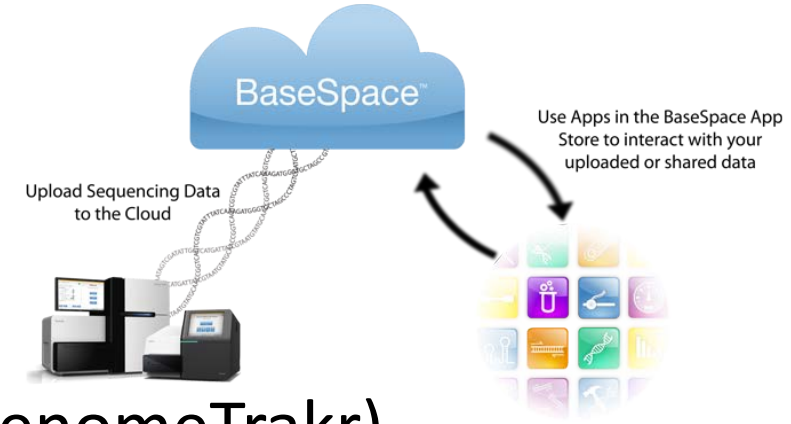
Lauren Turner, Ph.D.

Virginia Division of Consolidated Laboratory Services

DCLS' WGS Odyssey

- **Prior to 2013:** Capillary array electrophoresis and pyrosequencing
- **2013: DCLS joined GenomeTrakr Project**
  - Resources: MiSeq, reagents, ORISE Fellow
  - Expectation: Sequencing ~400 food, animal and environmental isolates/year
- **Immediate Considerations**:
  - Data Sharing/Transfer
  - Data Storage

Data Transfer Options

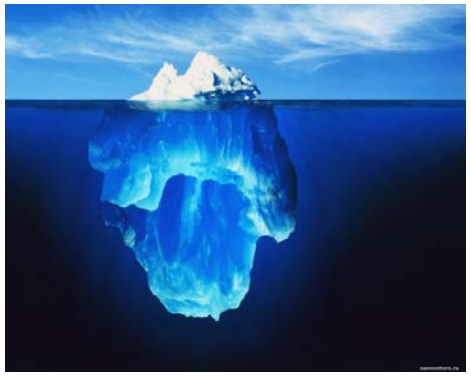


- **Several options**
  - FTP site (PulseNet only)
  - BaseSpace (PulseNet and GenomeTrakr)
- **Data is streamed from the MiSeq instrument in "real-time" to the BaseSpace Cloud**
- **Path through Firewall (in and out) for data and instrument troubleshooting**

Instating Data Transfer

- **Permission to use cloud-based resource**
  - VA Lab ID and submitting lab anonymized through CDC, no patient identifying information
- **Collaborations with DCLS IT, VA IT Group, Illumina, and NY State DoH/Wadsworth Center (Bill Wolfgang's group).**

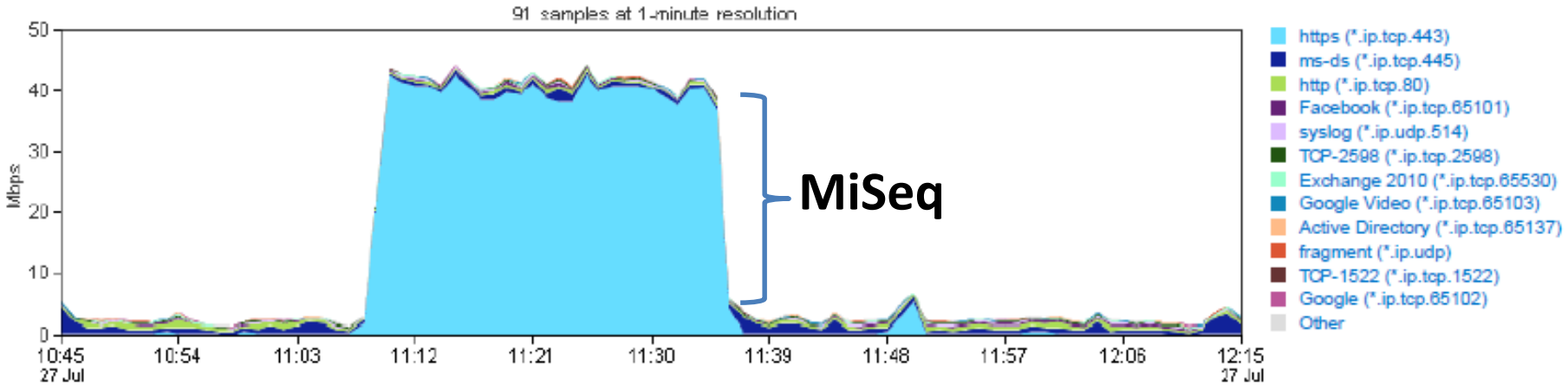
Unforeseen Obstacles



- **Network control**
  - DCLS network is not building-controlled and, with other applications running or being updated in the background, data streaming my be interrupted
  - During 39 hour runs, interruptions are very common
    - De-prioritization of instrument network use?
    - Applications running nightly?
- **Exempting the MiSeq during network patching**

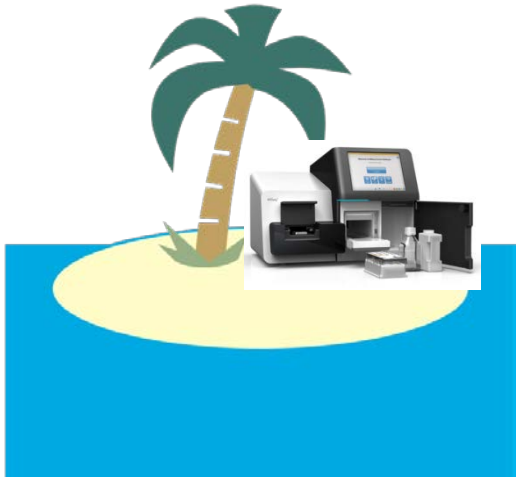
Unforeseen Obstacles

- **Interruptions in data streaming require restarting the MiSeq to complete data upload**
- **Significant network issues:**
  - Building network bandwidth is 45 Mbps, WGS raw data files are ~10 Gb
  - Streaming can cause major strain on the overall building network

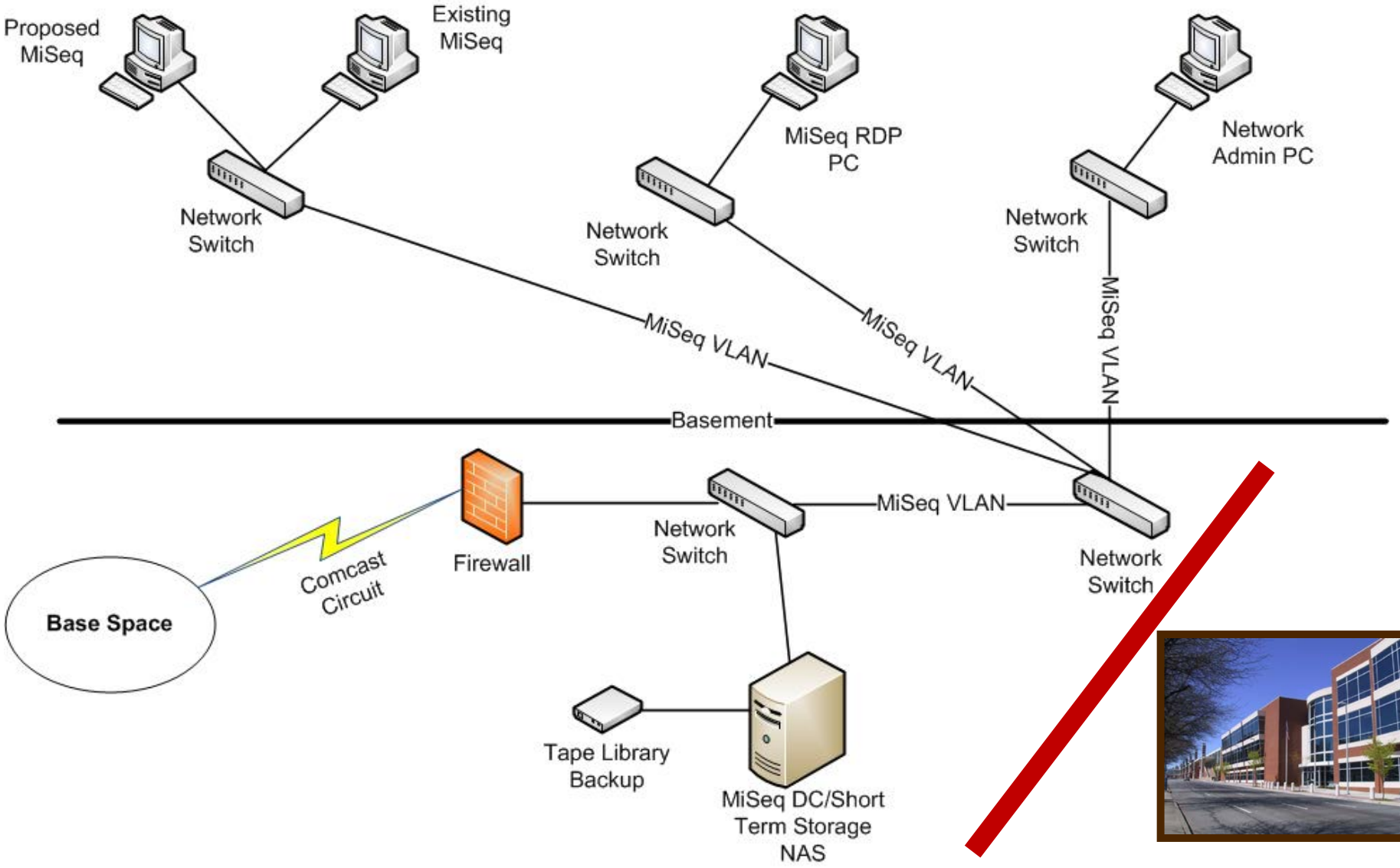


Data Bottleneck Resolution

- **IT monitoring**
- **Working with Illumina to diagnose issues**
- **Recognition of true and projected future needs**
  - Simultaneous data streaming with multiple MiSeqs?
  - Advantages of cloud for data storage, access, informatics and sharing?
  - Can MiSeqs be exiled?



Coming in March 2016…



Data Storage and Backup

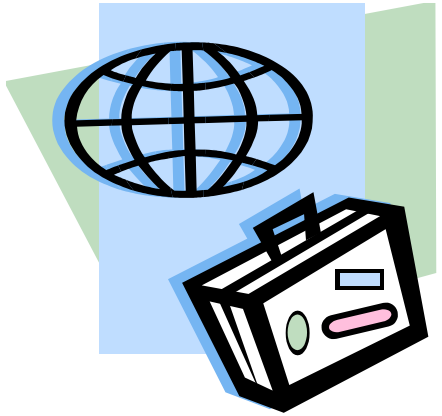
- **To date:**

Program	Circa	Isolates Sequenced	Data Shared
GenomeTrakr	2013	645	~0.5 TB
PulseNet	2014	396	~0.3 TB

- **MiSeq hard drive: 500 MB (~10 runs)**
  - Not a long term solution
- **NCBI stores the final, QC'd raw reads**
  - Backed up to DDBJ and EMBL
- **Data retention policies…**
  - CLIA (3 yrs), DCLS (10 yrs)

Data Storage Solutions

- **Current short-term:**
  - Back-up all files to external hard drives (1 – 3 TB)
  - Remove image thumbnails after 6 months
  - Save all fastq.gz files longer term
    - Data backup
    - Future validations using “raw data”



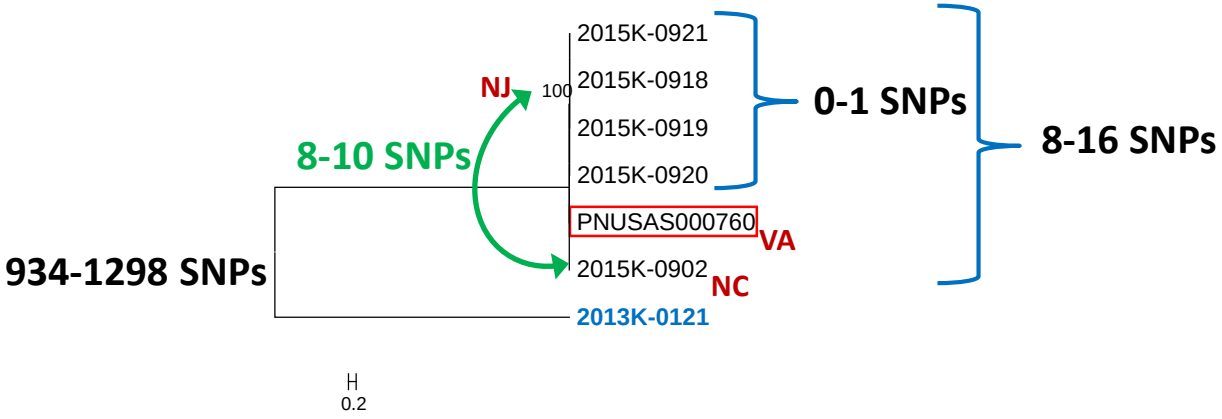
- **Future planning:**
  - Tape storage (off-site, secure, redundant)
  - Cloud-based cold storage (Amazon Cloud)

amazon web services™

Additional IT considerations

- **WGS workflow coordination through LIMS**
- **Building bioinformatic capacity**
  - Apply WGS to VA clusters
  - Apply to non-PulseNet organisms

Laboratory Information Management Systems



Current Workflow Coordination

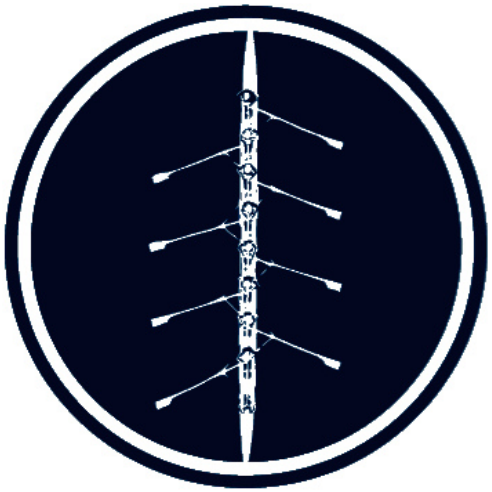
- **WGS supported by several DCLS Groups**
  - Reference Laboratory: Receipt and ID confirmation
  - PFGE Lab: Isolate storage, WGS reagent inventory
  - Molecular Group: DNA extraction, library prep and run
- **Cross-group communication and tracking**
  - Multi-group distribution list
  - Workflow tracking, reagent inventories on network drive

VA Key	Organism	Serotype for Salmonella	WGS_ID	request receive date	DNA extraction date	DNA quantification value	Index 1	Index 2	MiSeq run date	MiSeq Run #	NCBI accession # upload date by PFGE	Accession	CDC data availability notification date	total sequences	seq. length	coverage	pass/fail	BioProject id
R150302492	non-O157 STEC		PNUSAE000621	N/A	5/15/2015	N/A	N701	S502	5/20/2015	DCLS_052015	5/22/2015	SAMN03732436	5/22/2015	494356	250	49	pass	218110
R150303327	non-O157 STEC		PNUSAE000622	N/A	5/15/2015	N/A	N701	S517	5/20/2015	DCLS_052015	5/22/2015	SAMN03732437	5/22/2015	468095	250	47	pass	218110
R150303382	non-O157 STEC		PNUSAE000623	N/A	5/15/2015	N/A	N701	S503	5/20/2015	DCLS_052015	5/22/2015	SAMN03732438	5/22/2015	535795	250	54	pass	218110
R150303710	non-O157 STEC		PNUSAE000624	N/A	5/15/2015	N/A	N701	S504	5/20/2015	DCLS_052015	5/22/2015	SAMN03732439	5/22/2015	587794	250	59	pass	218110
R150400759	non-O157 STEC		PNUSAE000625	N/A	5/15/2015	N/A	N702	S502	5/20/2015	DCLS_052015	5/22/2015	SAMN03732440	5/22/2015	546417	250	55	pass	218110
R150400923	non-O157 STEC		PNUSAE000626	N/A	5/15/2015	N/A	N702	S517	5/20/2015	DCLS_052015	5/22/2015	SAMN03732441	5/22/2015	701987	250	70	pass	218110
R150500118	non-O157 STEC		PNUSAE000627	N/A	5/15/2015	N/A	N702	S503	5/20/2015	DCLS_052015	5/22/2015	SAMN03732442	5/22/2015	778657	250	78	pass	218110
R150500277	non-O157 STEC		PNUSAE000628	N/A	5/15/2015	N/A	N702	S504	5/20/2015	DCLS_052015	5/22/2015	SAMN03732443	5/22/2015	413143	250	41	pass	218110
R150500765	non-O157 STEC		PNUSAE000725	N/A	5/21/2015	21.3	N705	S504	6/3/2015	M02323_150603	6/5/2015	SAMN03763794	6/5/2015	1260973	250	126	pass	218110
R150500753	non-O157 STEC		PNUSAE000724	N/A	5/21/2015	20.1	N705	S503	6/3/2015	M02323_150603	6/5/2015	SAMN03763793	6/5/2015	1074987	250	107	pass	218110
R150500565	non-O157 STEC		PNUSAE000723	N/A	5/21/2015	22.2	N705	S517	6/3/2015	M02323_150603	6/5/2015	SAMN03763792	6/5/2015	443016	250	44	pass	218110
R150500443	non-O157 STEC		PNUSAE000722	N/A	5/21/2015	29.1	N705	S502	6/3/2015	M02323_150603	6/5/2015	SAMN03763791	6/5/2015	715108	250	72	pass	218110
R150301890	non-O157 STEC		PNUSAE000721	N/A	5/21/2015	19.9	N704	S504	6/3/2015	M02323_150603	6/5/2015	SAMN03763790	6/5/2015	889458	250	89	pass	218110
R150201678	non-O157 STEC		PNUSAE000720	N/A	5/21/2015	24.3	N704	S503	6/3/2015	M02323_150603	6/5/2015	SAMN03763789	6/5/2015	807072	250	81	pass	218110
R150201033	non-O157 STEC		PNUSAE000719	N/A	5/21/2015	20.7	N704	S517	6/3/2015	M02323_150603	6/5/2015	SAMN03763788	6/5/2015	958026	250	96	pass	218110
R150200932	non-O157 STEC		PNUSAE000718	N/A	5/21/2015	21.8	N704	S502	6/3/2015	M02323_150603	6/5/2015	SAMN03763787	6/5/2015	1357233	250	136	pass	218110
R140900134	non-O157 STEC		PNUSAE000717	N/A	5/21/2015	21.1	N703	S504	6/3/2015	M02323_150603	6/5/2015	SAMN03763786	6/5/2015	1429963	250	143	pass	218110

Run Log | NARMS- GT | Vtech_Salm-GT | Pending Historical non-O157 | Pending WGS | library prep training | need WGS ID

Improved Workflow Coordination

- **Use LIMS to replace interim measures for:**
  - Isolate tracking
  - Testing coordination and prioritization
  - QC metric documentation
- **Expand to include:**
  - Pathotype
  - Virulence factors, resistance markers
  - MLST types?
  - Cluster tracking information
  - Reporting/electronic messaging to Department of Heath
  - Other DCLS areas in LIMS (NBS, TB, Virology?)



Projecting Future Needs

- **Next Steps:**
  - Enhance VA bioinformatics
  - **Options:**
  - BioNumerics 7.5 wgMLST
  - Commercial software (CLC Genomics Workbench)
  - Bioinformatic partnerships- VA government, VA universities, other state partners
    - Bioinformatics expertise and robust hardware for processing
    - Trained staff

DCLS' Odyssey Continues

- **Charting a new course for the future**
  - Assessment of IT capacity and needs
  - Interagency collaborations to address IT and security demands
  - Identifying new resources and strategies for data storage and analysis