Construction, Communication and Collaboration: A fellow's Tale of ~~Two~~Twin Cities

Xiong (Sean) Wang DVM, Ph.D | APHL-CDC Bioinformatics Fellow (Alumni)

APHL Annual Meeting

June 2nd, 2018 Pasadena, CA

A little bit about me

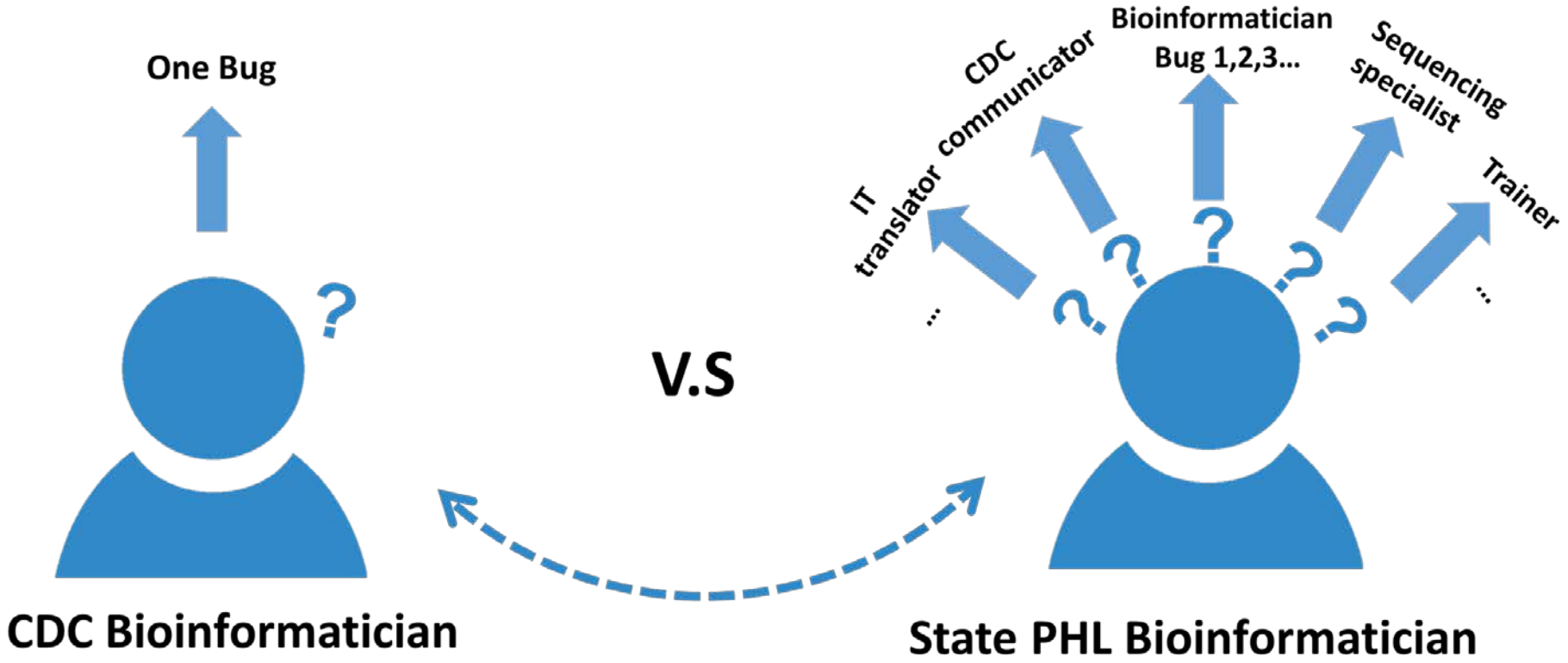
- I never read “ A Tale of Two Cities” before
- Trained as a veterinarian in China
- Have long-lasting interest in Public Health
- Ph.D in studying veterinary infectious disease viral genomics and host-pathogen interaction utilizing Next-generation sequencing (NGS)
- 2016 APHL-CDC Bioinformatics Fellow
- First fellow who was placed in a state public health laboratory
- As of yesterday, I took the role as sequencing and bioinformatics unit supervisor of Minnesota Public Health Laboratory

Dream VS. Reality

- What I thought I am stepping into:

- What I actually stepped into:

Bioinformatician at CDC vs. Bioinformatician at State PHL



Projects I was, am and will involved

- Minnesota Department of Health bioinformatics infrastructure building
  - High-performance computing cluster
  - Cloud-based computing environment
  - MDH-Minnesota Supercomputing Institute (MSI) liaison, contract maintenance and renew
- Salmonella serotyping by Whole Genome Sequencing (WGS)
  - Pipeline design and building
  - Validation dataset analyses
  - SOP and its related document
- Minnesota ELC-AMD WGS and bioinformatics training
- *Clostridium difficile* WGS
- Foodborne outbreak isolates whole genome SNP phylogenetic analyses
- Streptococcus pneumonia serotyping by WGS

…

Projects I was, am and will involved

- Minnesota Department of Health bioinformatics infrastructure building
  - High-performance computing cluster
  - Cloud-based computing environment
  - MDH-MSI liaison, contract maintenance and renew
- Salmonella serotyping by Whole Genome Sequencing (WGS)
  - Pipeline design and building
  - Validation dataset analyses
  - SOP and its related document
- Minnesota ELC-AMD WGS and bioinformatics training
- *Clostridium difficile* WGS
- Foodborne outbreak isolates whole genome SNP phylogenetic analyses
- Streptococcus pneumonia serotyping by WGS

Three current models to build the bioinformatics infrastructure

NY-Wadsworth Center
CDC

Pro:
- IT security;
- Ownership

Con:
- Cost;
- Stagnant compute resource;
- Maintenance

amazon web services™

Google Cloud Platform

IBM Cloud

Microsoft Azure

CO-PHL

Pro:
- Configurable computing resource;
- Portability
- Development environments

Con:
- IT security;
- Data transfer;
- Data ownership

MN-PHL
MA-PHL
IA-PHL
VA-PHL

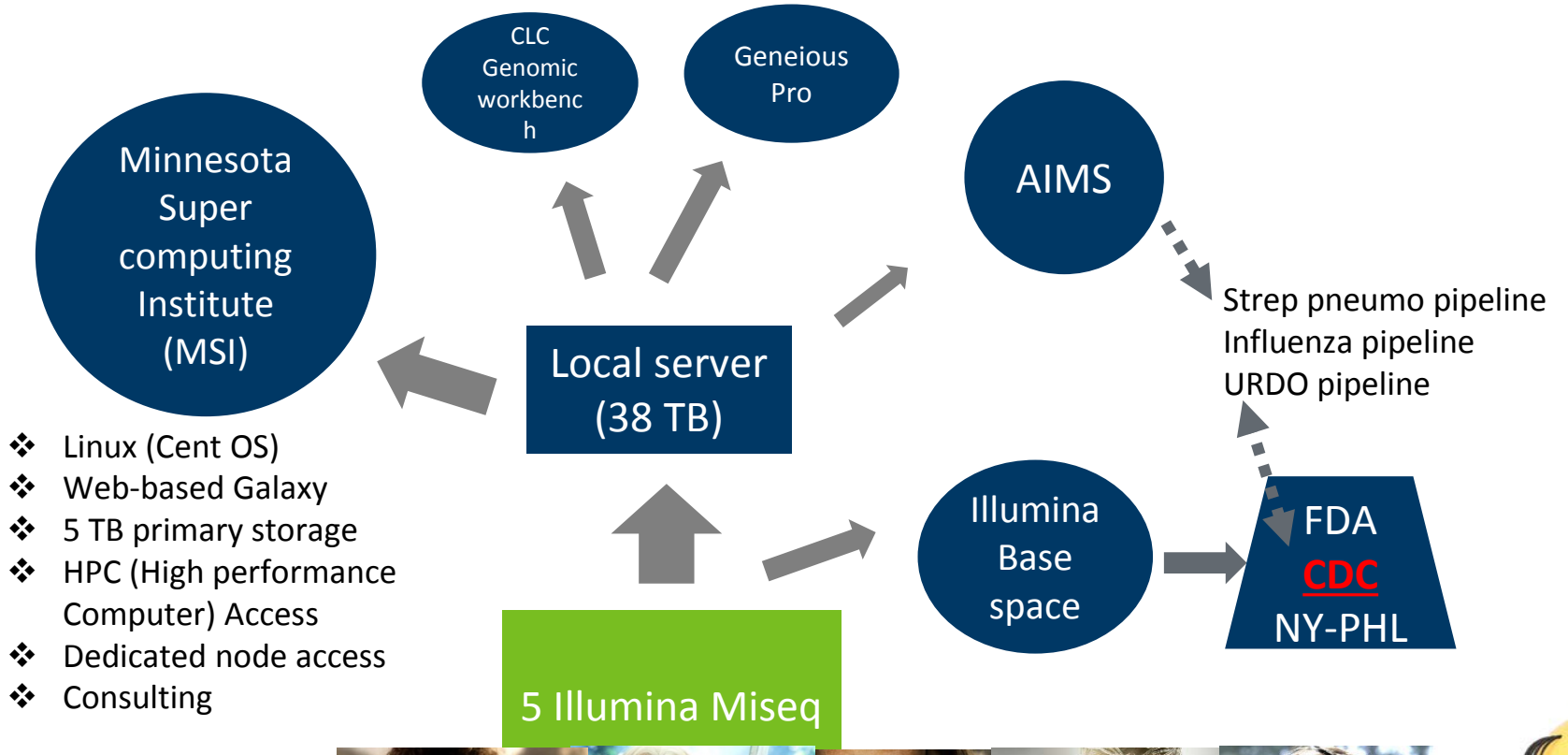
Pro:
- Configurable computing resource;
- Ready-to-use system
- Pre-installed software and dependencies
- Dedicated consulting

Con:
- No root access
- IT security;
- Data transfer;
- University-government contract

The bioinformatics infrastructure and NGS data flow of MDH

Stratus

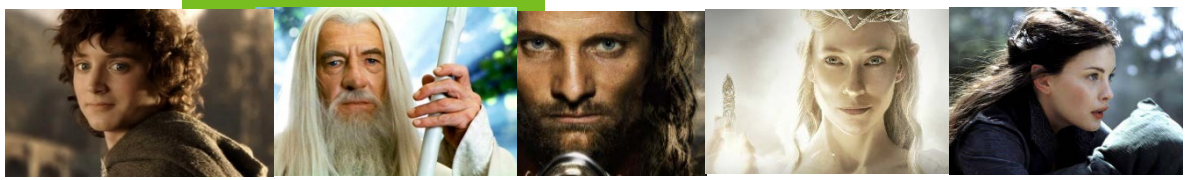
- Stratus cloud computing
  - 16vCPU
  - 2TB of ephemeral storage
  - Amazon S3 access



Data Sharing

Data flow

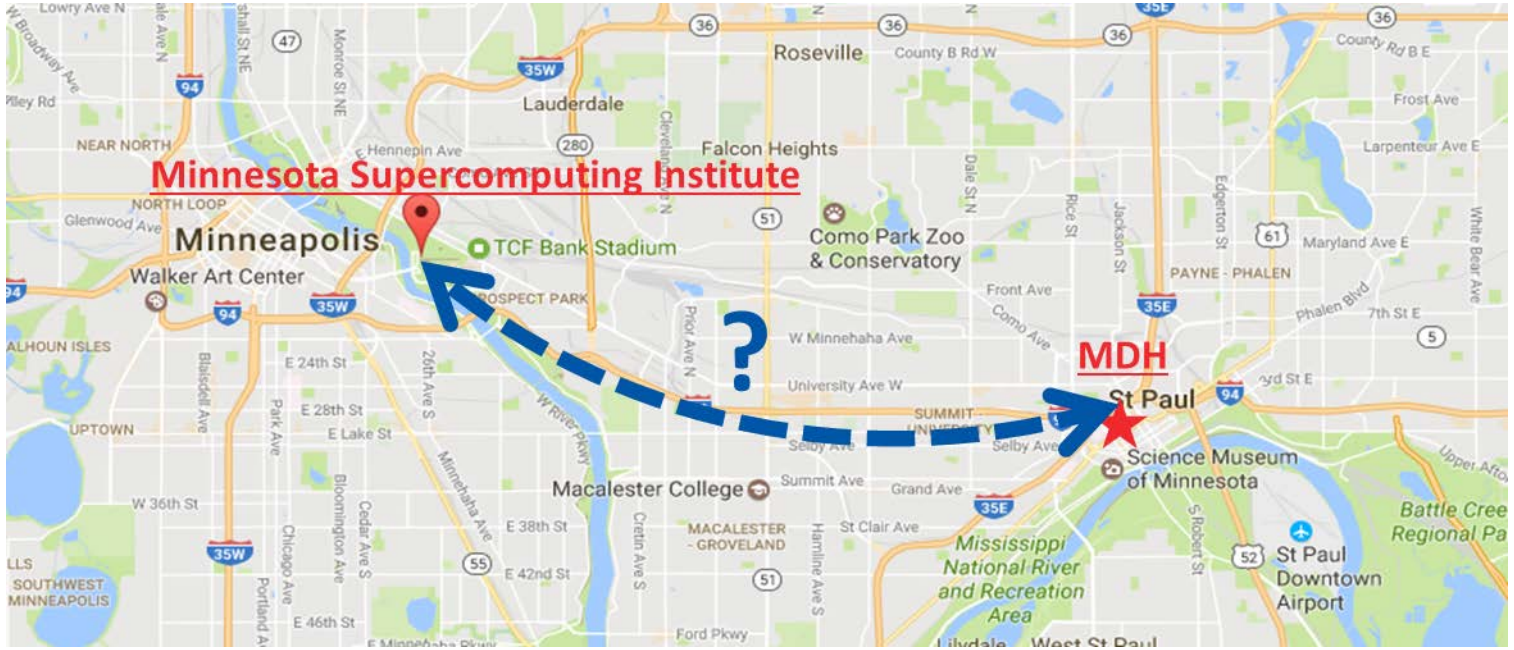
1 TB = 1000 GB
1 GB = 1000 MB
1 MB = 1000 B

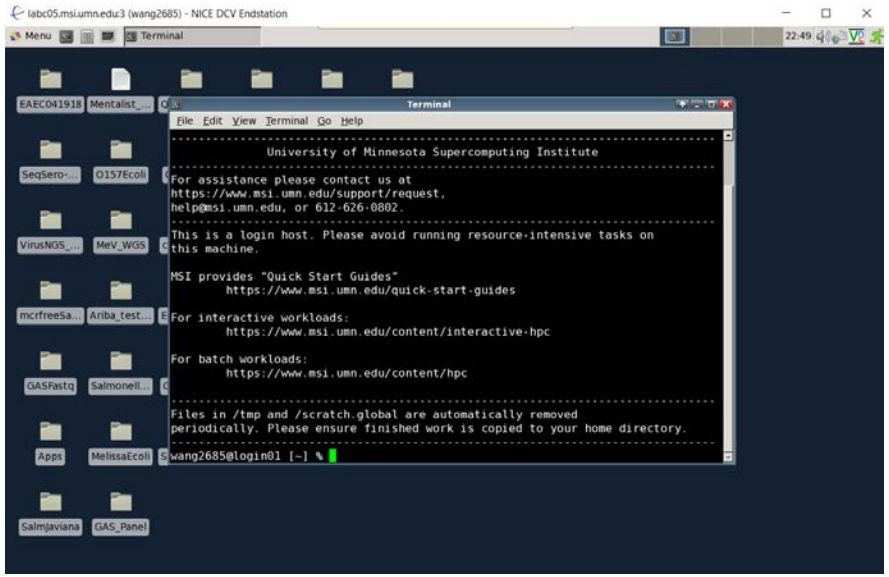


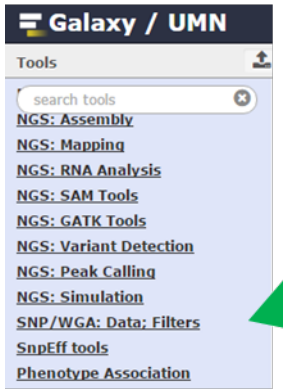
Lord of ~~Rings~~ Sequencers

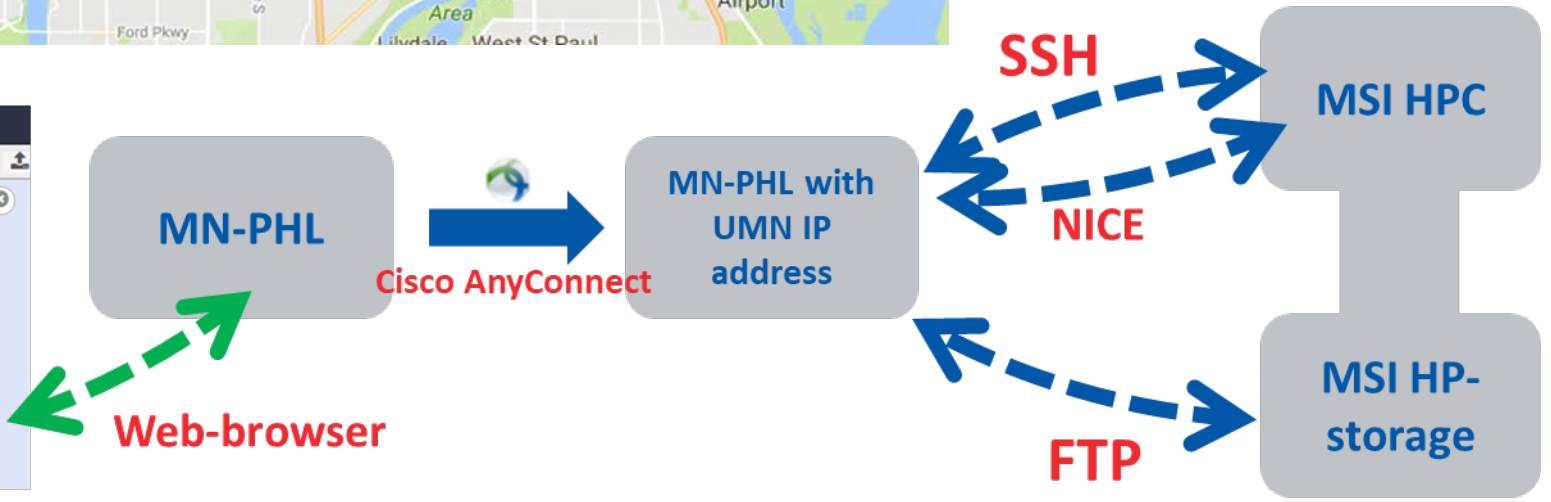
MinION

Current working mode for MSI access







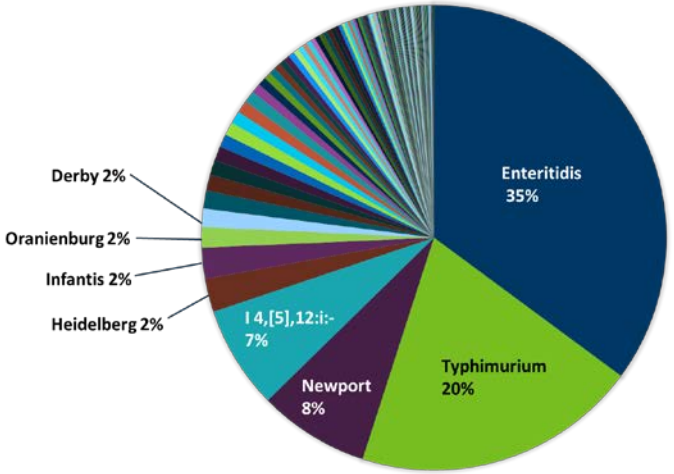


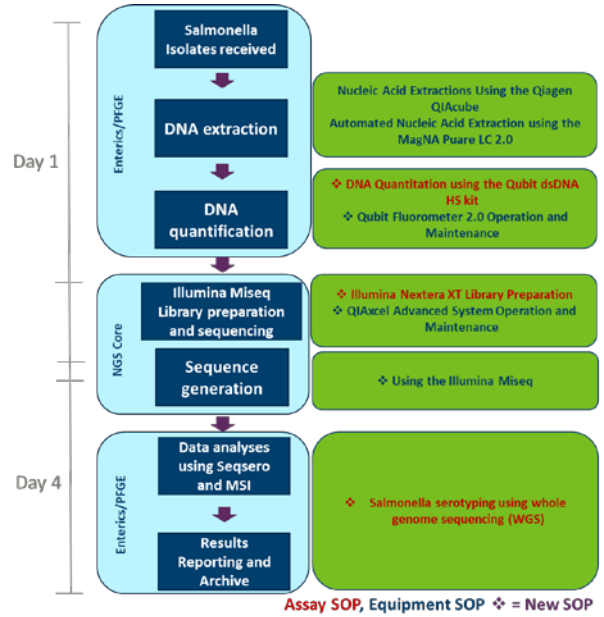
Projects I was, am and will involved

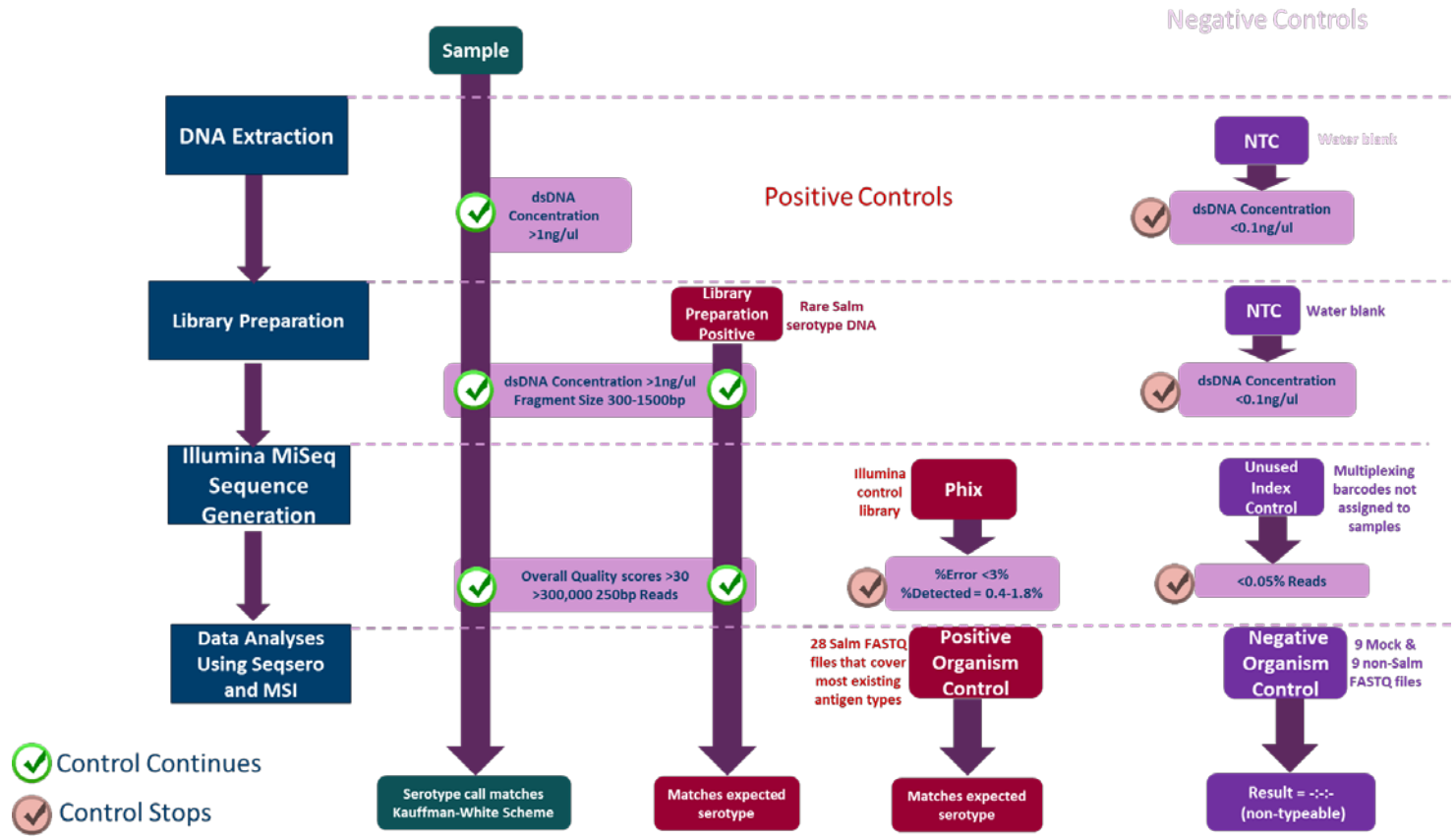
- Minnesota Department of Health bioinformatics infrastructure building
  - High-performance computing cluster
  - Cloud-based computing environment
  - MDH-MSI liaison, contract maintenance and renew
- Salmonella serotyping by Whole Genome Sequencing (WGS)
  - Pipeline design and building
  - Validation dataset analyses
  - SOP and its related document
- Minnesota ELC-AMD WGS and bioinformatics training
- *Clostridium difficile* WGS
- Foodborne outbreak isolates whole genome SNP phylogenetic analyses
- Streptococcus pneumonia serotyping by WGS

Salmonella serotyping by Whole Genome Sequencing (WGS)

Serotypes Analyzed with SeqSero (n=2414)

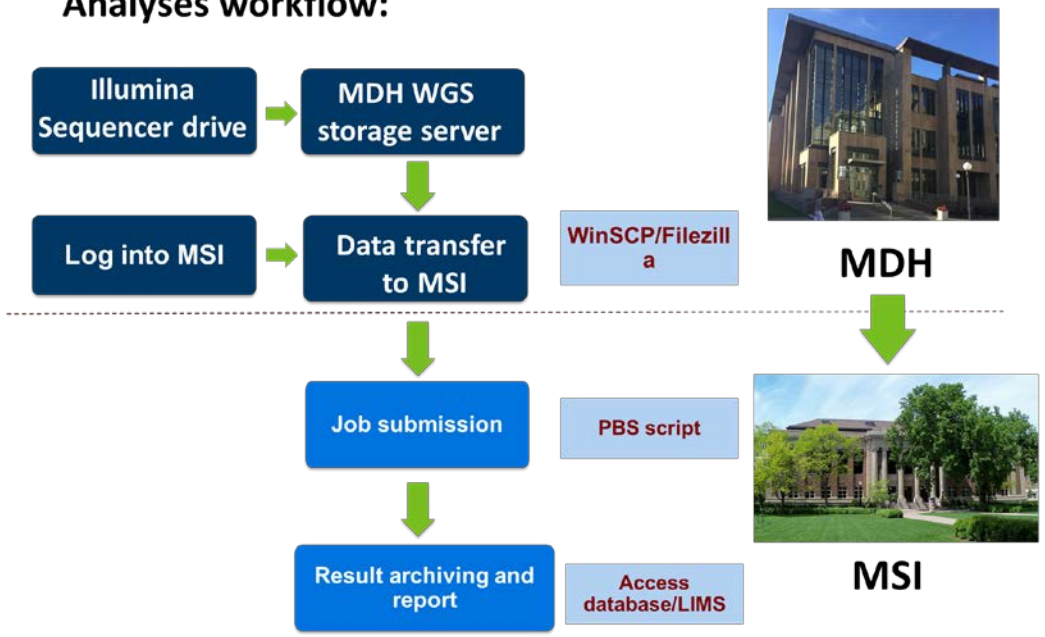




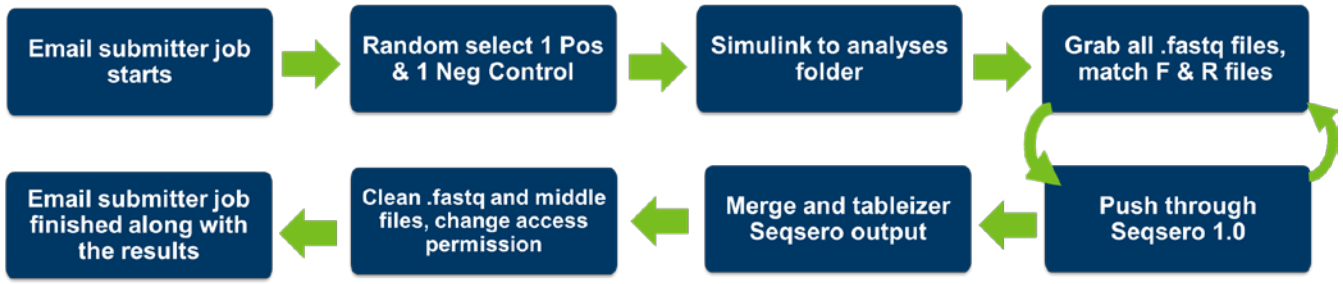


Salmonella serotyping by Whole Genome Sequencing (WGS)

Analyses workflow:



Bioinformatic pipeline workflow:



Speed: ~12 mins per sample under nodes=2:ppn=4,mem=48gb

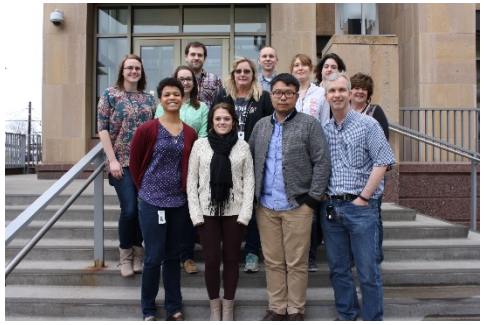
Projects I was, am and will involved

- Minnesota Department of Health bioinformatics infrastructure building
  - High-performance computing cluster
  - Cloud-based computing environment
  - MDH-MSI liaison, contract maintenance and renew
- Salmonella serotyping by Whole Genome Sequencing (WGS)
  - Pipeline design and building
  - Validation dataset analyses
  - SOP and its related document
- Minnesota ELC-AMD WGS and bioinformatics training
- *Clostridium difficile* WGS
- Foodborne outbreak isolates whole genome SNP phylogenetic analyses
- Streptococcus pneumonia serotyping by WGS

Minnesota ELC-AMD WGS and bioinformatics training

The Power of ATGC

ELC bioinformatic training



- April 2017

- Non-O157 STEC *stx* and *eae* subtyping
- Klebsiella *pneumoniae* vs. Klebsiella *variicola*

- March 2018

- WG SNP Clustering analyses using mock Salmonella outbreak
- K-mer based pathogen identification

Stratus

Cloud-computing based

- March/April 2019

Projects I was, am and will involved

- Minnesota Department of Health bioinformatics infrastructure building
  - High-performance computing cluster
  - Cloud-based computing environment
  - MDH-MSI liaison, contract maintenance and renew
- Salmonella serotyping by Whole Genome Sequencing (WGS)
  - Pipeline design and building
  - Validation dataset analyses
  - SOP and its related document
- Minnesota ELC-AMD WGS and bioinformatics training
- *Clostridium difficile* WGS
- Foodborne outbreak isolates whole genome SNP phylogenetic analyses
- Streptococcus pneumonia serotyping by WGS

Clostridium difficile WGS
- Crosswalk between Ribotype vs. MLST; Whole genome hq-SNP analyses using small scale data set

- 301 isolates from the 30 most common ribotypes (RT) in EIP collection
- Multilocus sequence typing (MLST) via WGS
- 7 house keeping gene scheme: adk, atpA, dxr, glyA, recA, soda, and tpi
- C. difficile MLST scheme was obtained from PubMLST.org
- StringMLST program used to determine sequence type (ST) and housekeeping gene alleles

- 22 (73%) of 30 RTs had a single ST
- 2 STs each had 2 RTs
- 8 (27%) RTs had multiple STs

Clostridium difficile WGS
- Crosswalk between Ribotype vs. MLST; Whole genome hq-SNP analyses using small scale data set

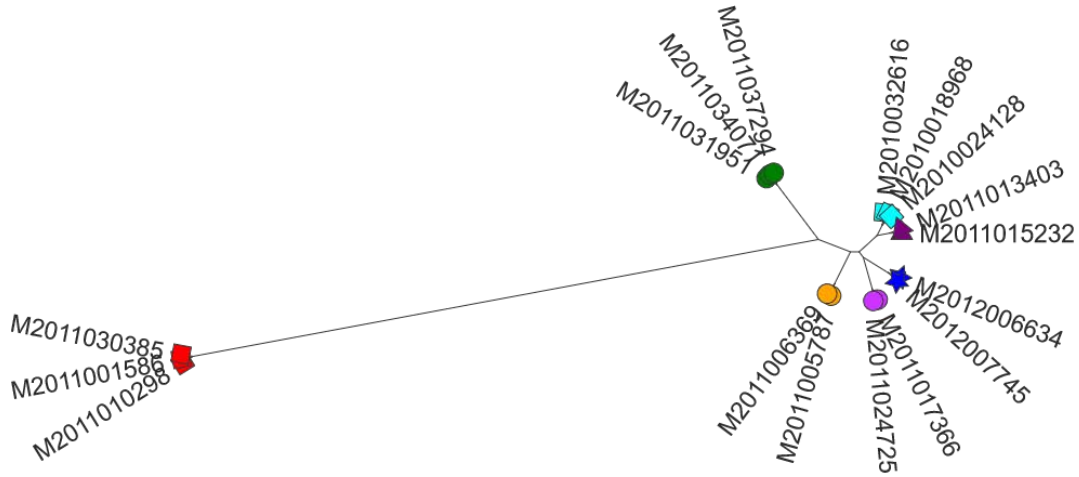
- 17 isolates representing 7 epidemiologically linked clusters

Accession number	Sequence Date	PFGE type	NAP type	MLST	SNP	Comment	Collection Date
M2011005787	1/23/2018	CDSMA.00387	unnamed	15	0-4	Mom	2/17/2011
M2011006369	1/23/2018	CDSMA.00387	unnamed	15		Infant daughter	2/20/2011
M2012006634	1/23/2018	CDSMA.00362	NAP6	8	0-1	Roomate	3/6/2012
M2012007745	1/23/2018	CDSMA.00362	NAP6	8		Roomate	3/19/2012
M2011013403	1/23/2018	CDSMA.00156	unnamed	58	0-3	Husband	4/11/2011
M2011015232	1/23/2018	CDSMA.00273	unnamed	58		Wife	4/29/2011
M2011001586	12/15/2017	CDF160	NAP7	11	0-1	Incident Isolate	1/7/2011
M2011010298	12/15/2017	CDF160	NAP7	11		1st Recurrent Isolate	3/22/2011
M2011030385	12/15/2017	CDF160	NAP7	11		2nd Recurrent Isolate	9/16/2011
M2010018968	6/5/2017	CDF138	unnamed	34	0-1	Incident Isolate	6/22/2010
M2010024128	1/8/2018	CDF138	unnamed	34		1st Recurrent Isolate	7/29/2010
M2010032616	1/8/2018	CDF138	unnamed	34		2nd Recurrent Isolate	10/8/2010
M2011031951	8/17/2017	CDF191	NAP1	67	0-1	Incident Isolate	10/5/2011
M2011034071	1/8/2018	CDF191	NAP1	67		1st Recurrent Isolate	10/24/2011
M2011037294	1/8/2018	CDF191	NAP1	67		2nd Recurrent Isolate	11/30/2011
M2011017366	1/8/2018	CDF1	NAP4	2	0	Incident Isolate	5/18/2011
M2011024725	12/15/2017	CDF1	NAP4	2		1st Recurrent Isolate	7/28/2011

Clostridium difficile WGS
- Crosswalk between Ribotype vs. MLST; Whole genome hq-SNP analyses using small scale data set

Maximum Likelihood Phylogeny

utilizing Micro**react** <microreact.org>



Pairwise SNP matrix

	M2011010298	M2011001586	M2011030385	M2011031951	M2011034071	M2011037294	M2011005787	M2011006369	M2012006634	M2012007745	M2011017366	M2011024725	M2010032616	M2010018968	M2010024128	M2011015232	M2011013403
M2011010298	0	0	1	69890	67636	65599	68630	68581	68398	68653	68219	67633	62640	70457	65335	69350	69496
M2011001586	0	0	1	80194	74901	71651	77075	77019	76480	76899	75704	74614	67366	80738	71170	77914	78016
M2011030385	1	1	0	68767	66529	64590	67430	67364	67247	67446	66982	66400	61488	69198	64164	68221	68290
M2011031951	69890	80194	68767	0	1	0	20673	20667	20191	20319	20402	19935	16844	22053	18048	20526	20671
M2011034071	67636	74901	66529	1	0	0	18944	18913	18610	18700	19131	18779	16408	19239	17340	18843	18880
M2011037294	65599	71651	64590	0	0	0	17950	17964	17660	17700	18284	17986	16021	18092	16767	17858	17910
M2011005787	68630	77075	67430	20673	18944	17950	0	4	10926	10893	11027	10799	9419	11560	10048	11177	11201
M2011006369	68581	77019	67364	20667	18913	17964	4	0	10915	10897	11003	10800	9455	11586	10055	11186	11257
M2012006634	68398	76480	67247	20191	18610	17660	10926	10915	0	1	9588	9425	8305	10080	8847	9679	9766
M2012007745	68653	76899	67446	20319	18700	17700	10893	10897	1	0	9569	9398	8287	10105	8821	9705	9781
M2011017366	68219	75704	66982	20402	19131	18284	11027	11003	9588	9569	0	0	8779	10361	9283	10366	10425
M2011024725	67633	74614	66400	19935	18779	17986	10799	10800	9425	9398	0	0	8690	10111	9130	10135	10211
M2010032616	62640	67366	61488	16844	16408	16021	9419	9455	8305	8287	8779	8690	0	1	1	3248	3307
M2010018968	70457	80738	69198	22053	19239	18092	11560	11586	10080	10105	10361	10111	1	0	0	3865	3946
M2010024128	65335	71170	64164	18048	17340	16767	10048	10055	8847	8821	9283	9130	1	0	0	3467	3515
M2011015232	69350	77914	68221	20526	18843	17858	11177	11186	9679	9705	10366	10135	3248	3865	3467	0	3
M2011013403	69496	78016	68290	20671	18880	17910	11201	11257	9766	9781	10425	10211	3307	3946	3515	3	0

In a APHL-CDC Bioinformatics Fellow perspective:

- Choose your host laboratory carefully based on your interests and future career plan (CDC vs. State Public Health Laboratory)
- Equipped yourself with great communication skills
- Ask for help whenever you need
- Collaboration Collaboration and collaboration
- Multi-tasking
- Eager to learn additional knowledge and skills outside of your expertise.

In a fellowship application reviewer perspective:

- The APHL-CDC Bioinformatics Fellowship should be a stepping stone for bioinformaticians to step into public health area.
- The fellowship should NOT be just another funding source.
- The fellowship should embrace diversified background yet hold ground rules:
  - Bioinformatics
  - Public Health
  - Infectious disease / human genetics

In a host laboratory/Advisor perspective (Dave Boxrud):

- Have a reasonable expectation for your fellow
- Have defined projects
- Mentorship is vital
- Need to give opportunities for advancement (training, conferences, present data at conferences…)
- Should have infrastructure for the fellow to be successful

Acknowledgement

Personnel:

Dave Boxrud (MDH)
Sara Vetter (MDH)
Joanne Bartkus (MDH)
Christin Hanigan (APHL)
Kelly Wroblewski (APHL)
Greg Armstrong (CDC)
Duncan MacCannell (CDC)
Scott Sammons (CDC)

Institutions:

Association of Public Health Laboratories (APHL)
Center for Disease Control and Prevention (CDC)
Minnesota Supercomputing Institute (MSI)
Minnesota Public Health Laboratory, Infectious disease lab
Minnesota State IT department

Picture credit: Google Inc. and Sean Wang

DEPARTMENT OF HEALTH

Thank you ! & Questions?

Xiong (Sean) Wang

Sean.Wang@state.mn.us

651-201-5050

@SeanWmpls