Update on GenomeTrakr Submissions and Bioinformatics Tools

Errol Strain, Ph.D.
Director, Biostatistics and Bioinformatics Staff
FDA Center for Food Safety and Applied Nutrition
GenomeTrakr Meeting 2018

GenomeTrakr Report Card

What's Going Well?

- NCBI Submission volume
- Epi & Pathogen Detection Portal
- Improvements to Metadata
- Guidelines for Interpretation
- Local Bioinformatics

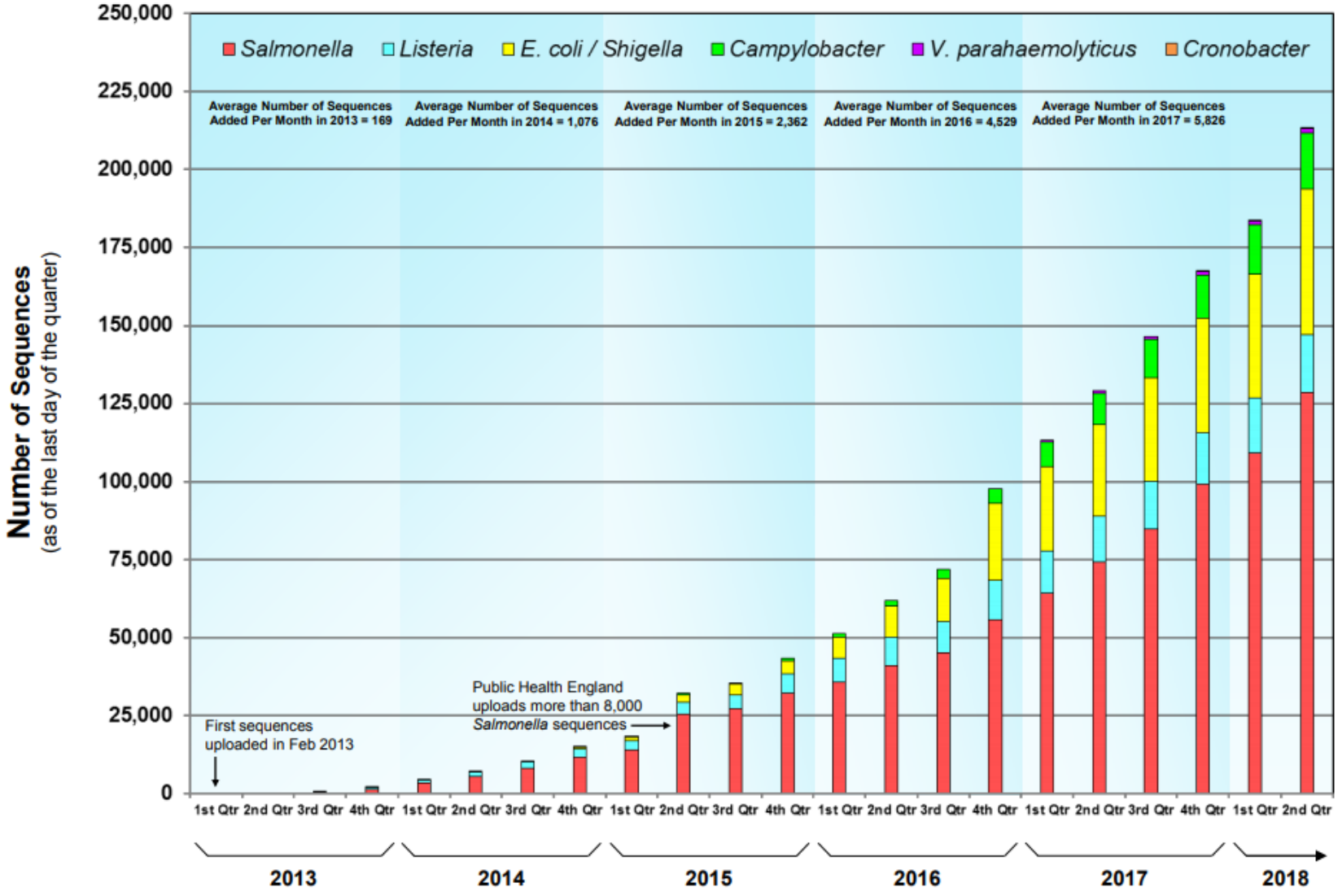
Needs Improvement

- Local QA/QC
- Submission speed

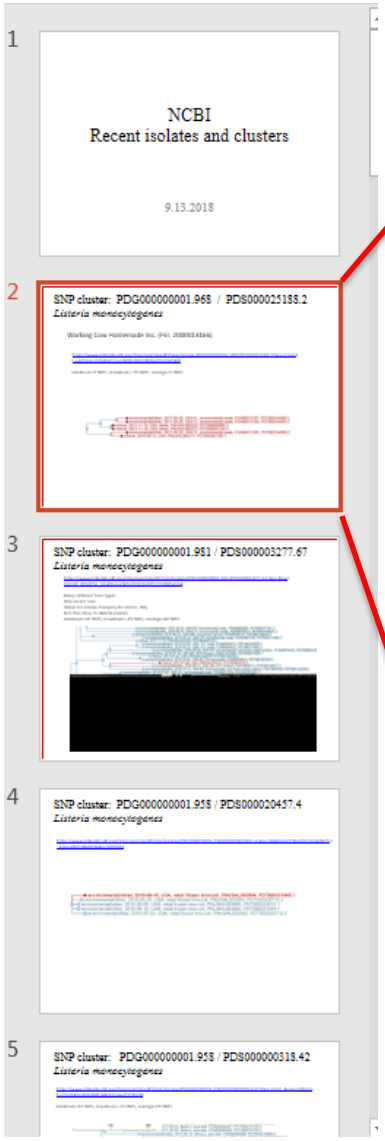
Concerns

- Local submission
- Rapid Typing/Matching

> 275K Total Pathogens



FDA CORE Signals



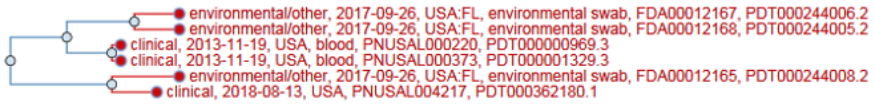
SNP cluster: PDG000000001.968 / PDS000025188.2

Listeria monocytogenes

Firm A (FEI: XXXX)

https://www.ncbi.nlm.nih.gov/Structure/tree/#!/tree/Listeria/PDG000000001.968/PDS000025188.2?key=dznvnFr-n4rlSqjmqSrNghgCcSxddB0NJXNVGjRXbx0tUmMSA1I

minimum=0 SNPs, maximum= 15 SNPs, average=9 SNPs



https://www.ncbi.nlm.nih.gov/pathogens

FDA Sample ID

Full

Pathogen: environmental/food/other sample from Listeria monocytogenes

Identifiers	BioSample: SAMN07702406; Sample name: CFSAN069610; SRA: SRS2542029
Organism	Listeria monocytogenes cellular organisms; Bacteria; Terrabacteria group; Firmicutes; Bacilli; Bacillales; Listeriaceae; Listeria
Package	Pathogen: environmental/food/other; version 1.0

Attributes

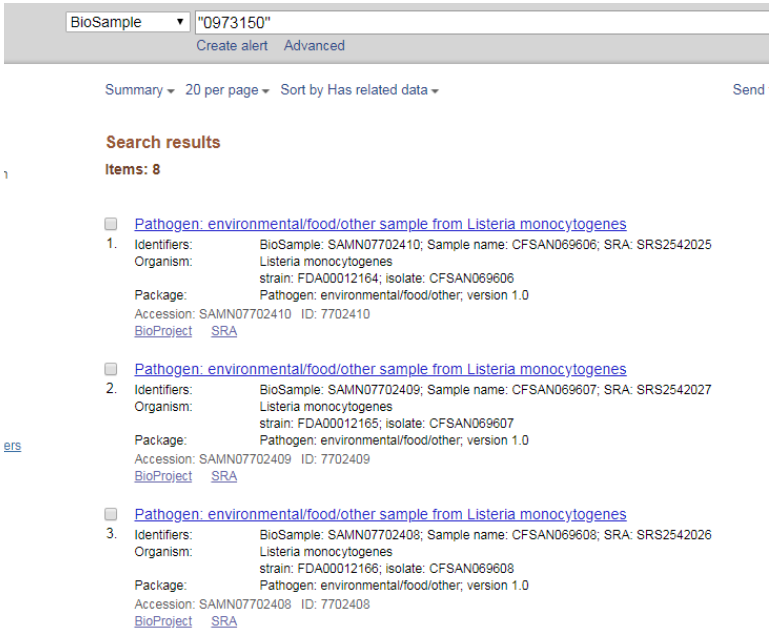
strain	FDA00012168
isolate	CFSAN069610
collected by	FDA
collection date	2017-08-30
geographic location	USA:FL
isolation source	environmental swab
latitude and longitude	missing
FDA_lab_id	0973150-032-001

BioProject	PRJNA215355 Listeria monocytogenes Retrieve all samples from this project
Submission	FDA, Maria Sanchez Leon; 2017-09-26

Accession: SAMN07702406 ID: 7702406
BioProject SRA

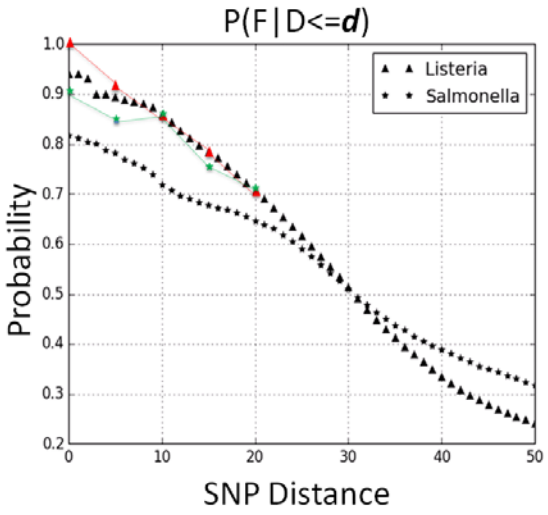
Increased Transparency

FDA Sample Key – “0973150”



Interpretation

Thresholds



SNP cutoff	*Sal.*	*Lis.*
0	0.82	0.94
5	0.78	0.89
10	0.72	0.86
15	0.68	0.79
20	0.65	0.70

12

Wang, Yu et al. JFP Galley Proof

Includes FEI (firm) numbers

Guidelines

WGS for Food Safety: Common Source

	Supports	Neutral	Does Not Support
SNP distance	< 20	20 – 100	> 100
Bootstrap support	> 0.90	0.80 – 0.90	< 0.80
Tree topology	Monophyletic	Paraphyletic	Polyphyletic

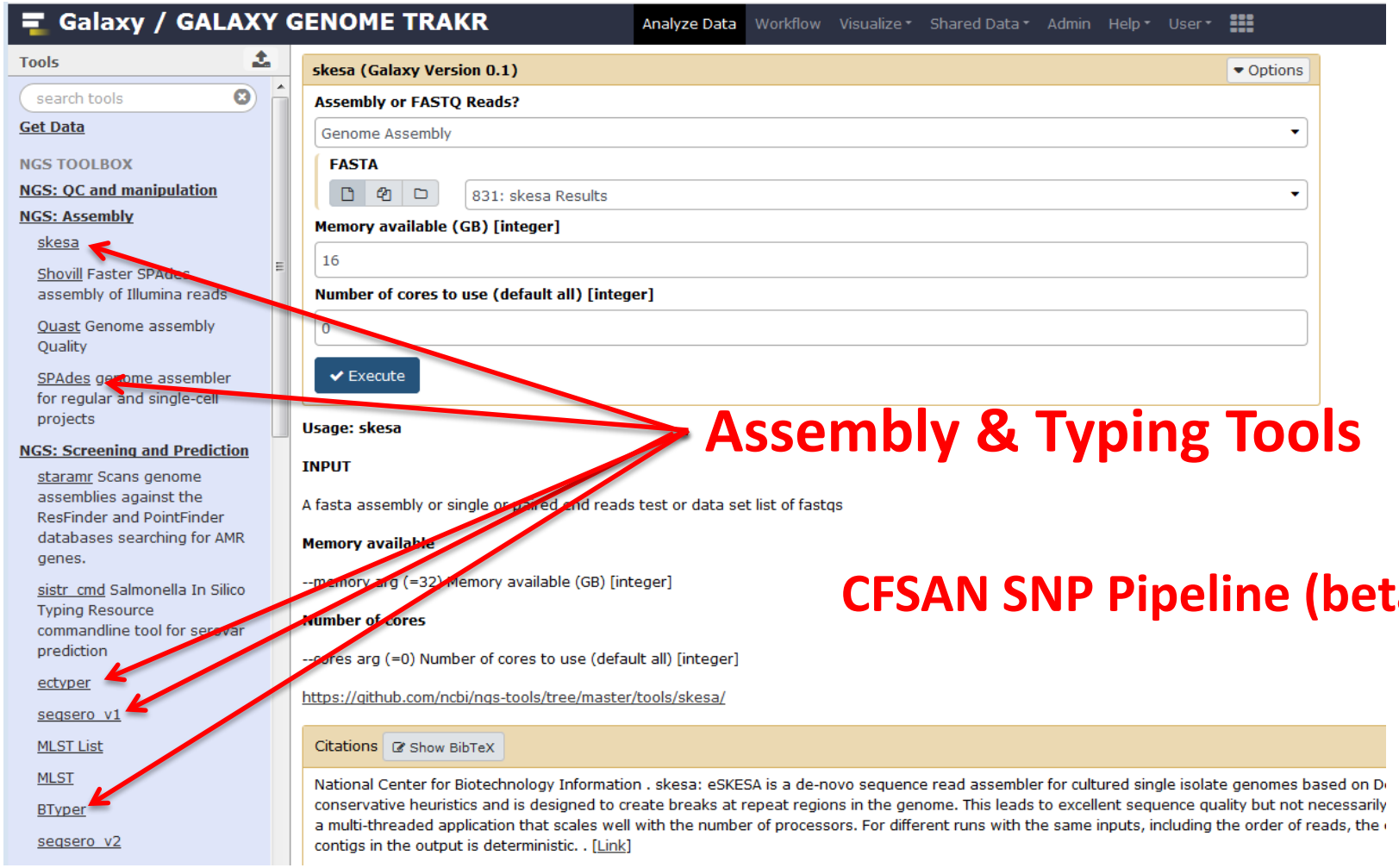
1. Supporting epidemiology or traceback information are required to justify decisions.

2. Isolates with any combination of evidence may ultimately be determined to match (but see point 1).

This approach reduces the chance that minor variations in a category of evidence will lead to significant changes in the interpretation of WGS analyses.

Pightling AW, Pettengill JB, Luo Y, Baugher JD, Rand H, Strain E. Interpreting Whole-Genome Sequence Analyses of Foodborne Bacteria for Regulatory Applications and Outbreak Investigations. Front Microbiol. 2018 Jul 10;9:1482. doi: 10.3389/fmicb.2018.01482. eCollection 2018. PubMed PMID: 30042741; PubMed Central PMCID: PMC6048267.

Galaxytrakr.org



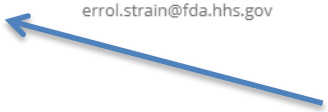
Assembly & Typing Tools

CFSAN SNP Pipeline (beta)

Galaxytrakr.org

Top 10 Users by Total Job Processing - Non Admin

Username	Email Address	Location	Total Process Time (Min)	Total CPU Slots Used
zli	zhen.li@doh.wa.gov	Washington State Department of Health, Public Health Laboratories	491135	173
jcarey	jaclyn.carey@health.ny.gov	Wadsworth Center New York State Department of Health	281150	14863
swirth	samantha.wirth@health.ny.gov	Wadsworth Center New York State Department of Health	158814	4182
mprarat	melanie.prarat@agri.ohio.gov	Ohio Department of Agriculture - Animal Disease Diagnostic Laboratory	61808	11247
mtay	moontayyuefeng@gmail.com	JIFSAN	60404	8426
hhoyt	hhoyt323@gmail.com	Wadsworth Center New York State Department of Health	51766	7423
whottel	wesley-hottel@uiowa.edu	University of Iowa State Hygienic Lab	21071	3508
anguyen	angela.nguyen@fda.hhs.gov	FDA CFSAN CPK1	17521	1493
astewart	alesha.stewart@dshs.texas.gov	Texas Department of State Health Services	15132	2394
estrain	errol.strain@fda.hhs.gov	Unknown	13283	560



#10 - estrain

State, Federal, Academic lab usage is up

Training – James Madison, JIFSAN, …

Needs Improvement

Tools for Local QA/QC

- 12 recent runs shared w/ GT, 4 out of 12 had isolates for other projects on the run
- Mix of local, FDA, and CDC isolates
- Labs need to be able to check an entire run, difficult to identify problems with an isolate

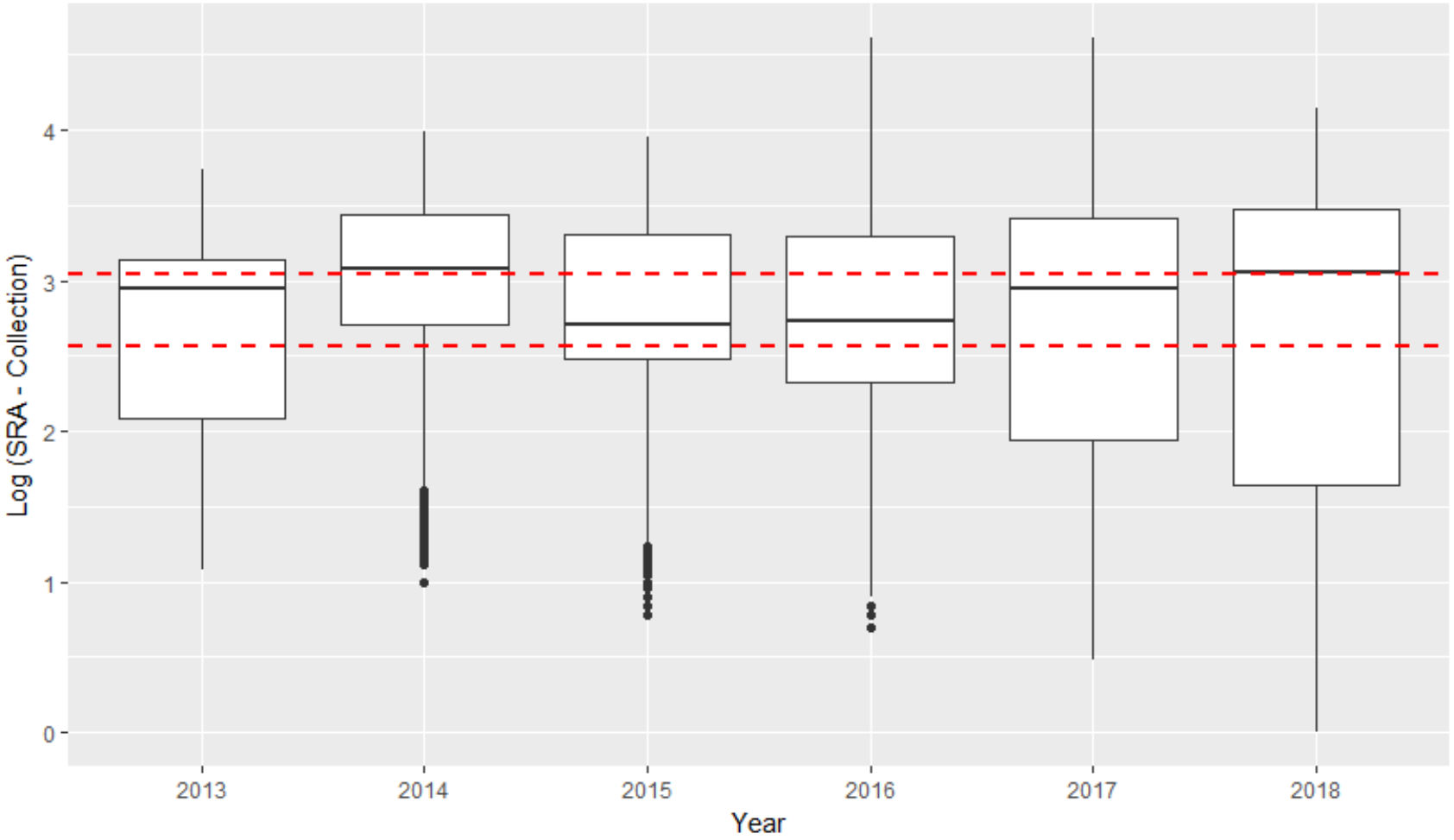
Not a New Problem

- I mentioned this last year, what's changed?
- Galaxy Workflow, ~10-15 minutes
  - FASTQ quality – SKESA assembly & coverage
  - Taxonomy – MLST and/or Serotype

Lab	GT	Other
FL Ag	16	0
MI Ag	13	0
NY Vet Lirn	14	8
Vet-LIRN-LSU	12	0
Penn State	18	2
Ny Ag	16	0
VA Health	0	16
MI Ag	16	0
VA	16	0
NM Unvi	16	0
CA Health	16	0
HI Health	0	20

Were any samples switched? Re-run or submit?

SRA Submission vs Collection DATE



32,317 out of215,873 with collection day

FDA Field Lab To NCBI

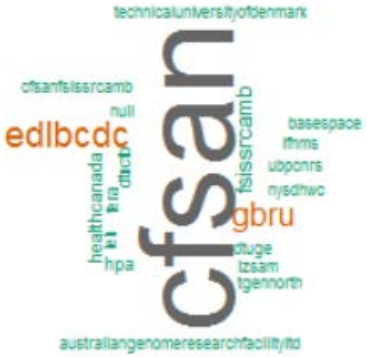
Lab	Run	Isolate	MiSeq	SRA	# Days
CF-NGS-FNE	180918_M05996_0007_000000000-BW88F	FDA00005334	9/20/2018 4:38	9/21/2018	**1**
CF-NGS-FDN	180905_M05930_0006_000000000-BYNG4	FDA00004835	9/7/2018 13:26	9/13/2018	**6**
CF-NGS-FNW	180914_M01599_0074_000000000-BYN9C	FDA00005342	9/16/2018 11:53	9/18/2018	**2**
CF-NGS-FSE	180905_M01404_0184_000000000-BLK94	FDA00013464	9/7/2018 3:57	9/10/2018	**3**
CF-NGS-FAR	180906_M01430_0008_000000000-BNBTC	FDA00013430	9/12/2018 11:52	9/13/2018	**1**
CF-NGS-FSF	180821_M01444_0085_000000000-BWLWM	FDA00013386	8/24/2018 10:40	8/27/2018	**3**
CF-NGS-FSW	180810_M01449_0078_000000000-BMPF6	FDA00013380	8/14/2018 13:33	8/16/2018	**2**

Latest uploads from Field Labs to shared genomics drive

Concerns – SRA Submission

2013

2014



2015

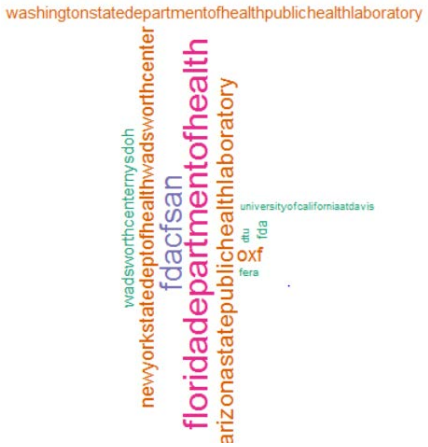
2016

2017

2018

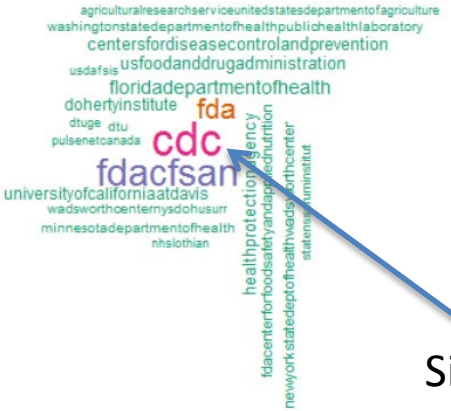
Sequencing Is Distributed (BioProjects)

2013



2014

2015



Single BP For PN

2016

2017

2018

Rapid Searching

Labs (and other customers) need rapid strain level reports for isolates on a run before submission

- Is the isolate in the outbreak cluster? Did I switch samples?
- NCBI – Roughly 2-3 days from SRA to Pathogen Detection
- BioNumerics – Faster, but labs must be part of PulseNet (and isolates must be in PN DB)

CFSAN Solution – cgMLST & MASH

- kmers of assemblies don't work
- Salmonella, Listeria, & E.coli/Shigella
- must keep assembly & cgMLST databases up to date – expensive
- Next year in GalaxyTrakr

U.S. FOOD & DRUG ADMINISTRATION

Thank You