Benchmark datasets for phylogenomic pipeline validation

GenomeTrakr Meeting
Sept. 2018

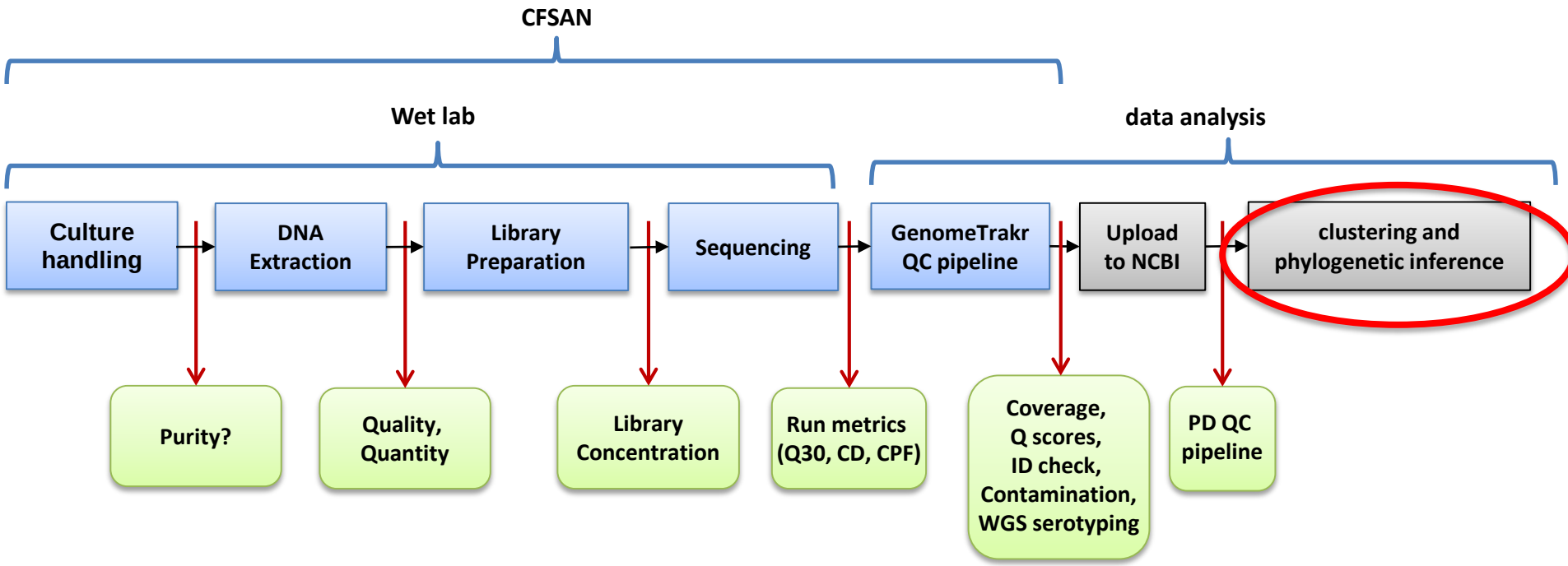
Ruth E. Timme, PhD
Research Microbiologist
GenomeTrakr data coordinator

Validation for phylogenomics

Phylogenomic pipeline = raw reads -> phylogeny

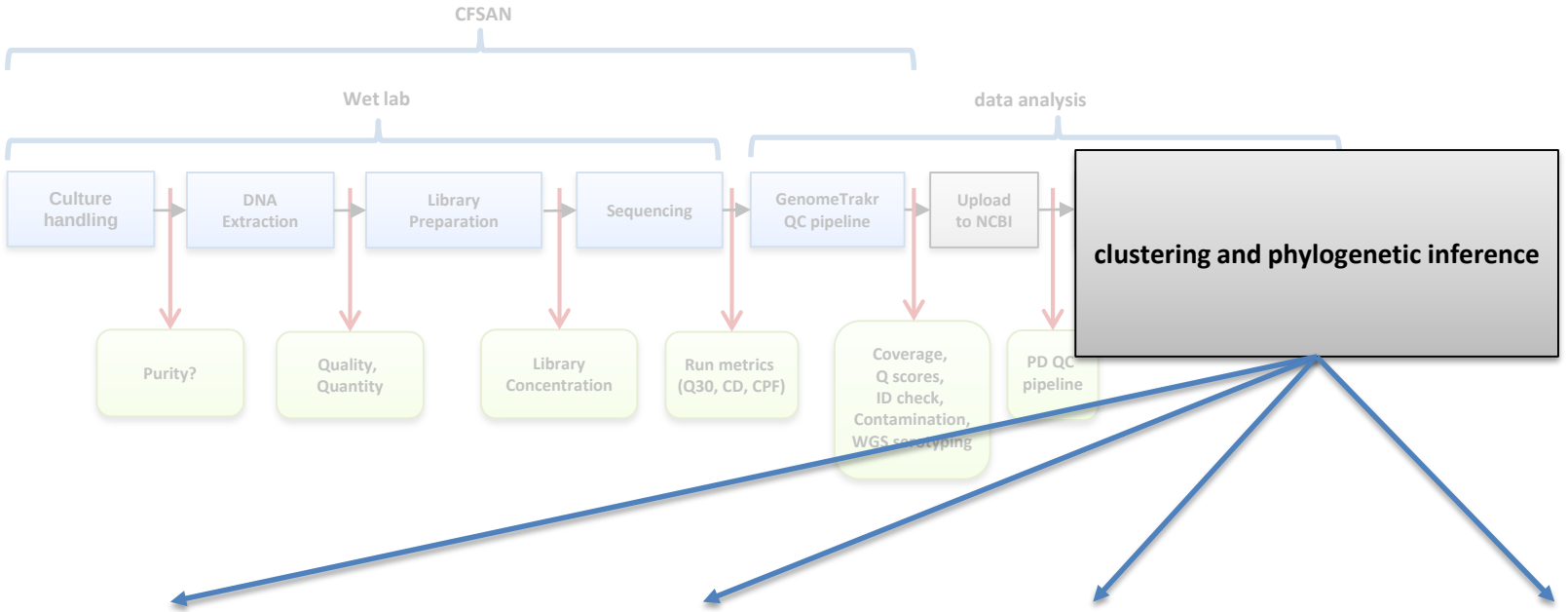
Goal: validate results to ensure accuracy and reproducibility. Do not enforce standards to overall approach.

CFSAN's WGS workflow modules

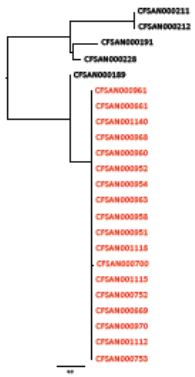


- 6 modules: each is independently validated
- Each module has routine QC checks

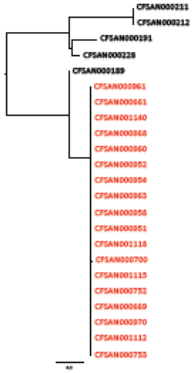
Phylogenomic analysis within the Gen-FS community



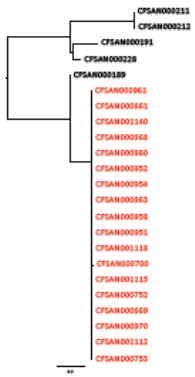
NCBI PD pipeline
Maximum compatibility phylogeny



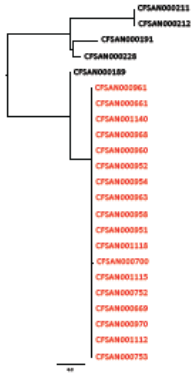
CFSAN SNP pipeline
Maximum likelihood phylogeny



BioNumerics wgMLST
Distance phylogeny



Lyve-SET pipeline
Maximum likelihood phylogeny



Compare pipeline results to ”truth”

1. SNPs and other variants
2. tree topology

What is “truth”?

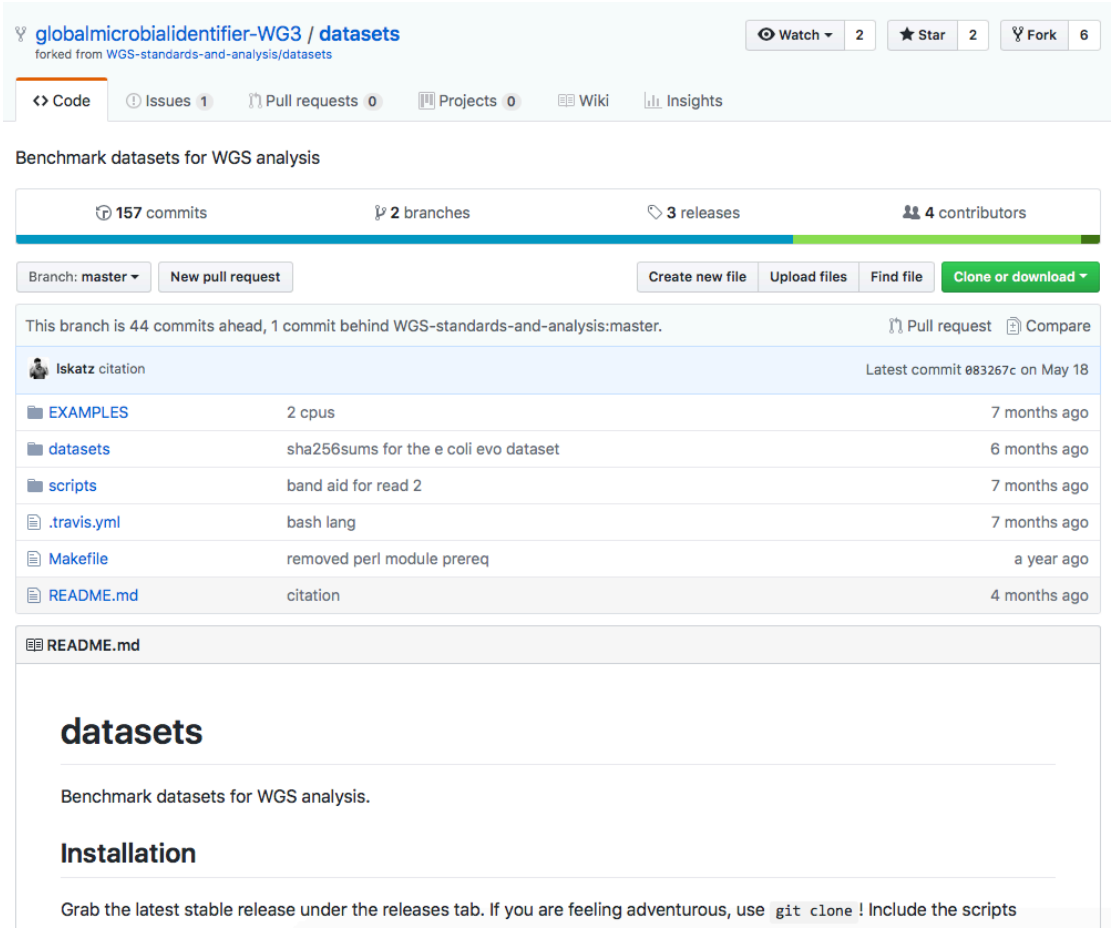
- Empirical dataset:
  - well-studied outbreak
  - concordant epi and WGS signal
  - **validated SNPs
- Synthetic dataset
  - known SNPs and tree topology

Benchmark datasets

Dataset	Organism	# isoaltes	Type of dataset
Stone Fruit recall	*L. monocytogenes*	31	Empirical
Spicy Tuna outbreak	*S. enterica*	23	Empirical
Raw Milk Outbreak	*C. jejuni*	22	Empirical
Sprouts outbreak	*E. coli*	10	Empirical
Simulated outbreak	*S. enterica*	23	Synthetic
Experimental evo.	*E. coli*	100	in vitro evolution

Manuscript:
R Timme et al. 2017. Bacterial pathogen genome datasets for bioinformatics pipelines. *PeerJ* 5, e3893.

grab downloading script at GitHub



https://github.com/globalmicrobialidentifier-WG3/datasets

Standard template for datasets

header

Organism	Salmonella enterica subspecies enterica Serovar Bareilly
Outbreak	1203NYJAP-1 - Tuna Scrape Outbreak
pmid	25995194
tree	http://api.opentreeoflife.org/v2/study/ot_301/tree/tree1.tre
source	Ruth Timme
dataType	empirical

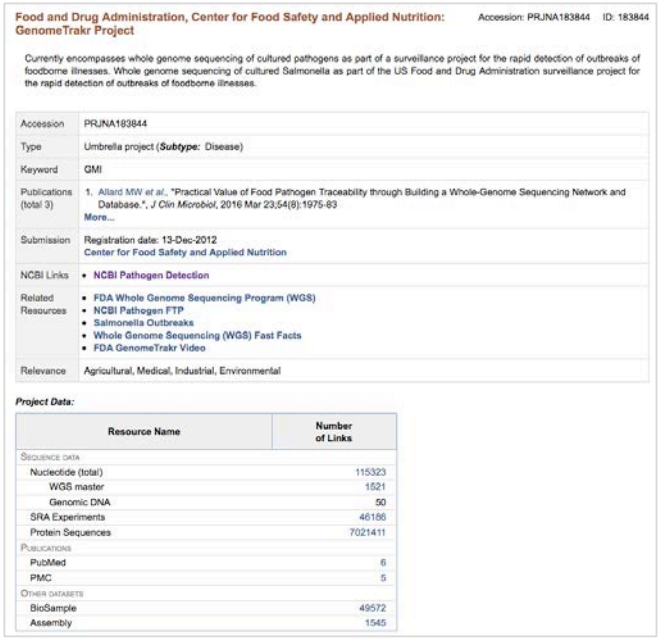
body

biosample_acc	strain	genBankAssembly	SRArun_acc	outbreak	dataSetName	suggested Reference	sha256sum Assembly	sha256su mRead1	sha256sum Read2	voucherContact
SAMN01823701	CFSAN000189	CP006053, CP006054	SRR498276	outgroup	1203NYJAP-1 - Tuna Scrape Outbreak	TRUE	checksums			Dwayne Roberson, FDA
SAMN00860590	CFSAN000191	JMMH00000000	SRR498369	outgroup	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN00989085	CFSAN000211	JMMM00000000	SRR498373	outgroup	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN00862341	CFSAN000212	JRDM00000000	SRR500494	outgroup	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN00862340	CFSAN000228	JRCY00000000	SRR500493	outgroup	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN00991042	CFSAN000661	JMMG00000000	SRR498397	1203NYJAP-1	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN00991044	CFSAN000669	JRCQ00000000	SRR498399	1203NYJAP-1	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN00991046	CFSAN000700	JRCO00000000	SRR498402	1203NYJAP-1	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN00991047	CFSAN000752	JRCN00000000	SRR498403	1203NYJAP-1	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN00991048	CFSAN000753	JRCM00000000	SRR498404	1203NYJAP-1	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN00991050	CFSAN000951	JRCJ00000000	SRR498422	1203NYJAP-1	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN00991051	CFSAN000952	JRCI00000000	SRR498423	1203NYJAP-1	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN00991053	CFSAN000954	JRCG00000000	SRR498425	1203NYJAP-1	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN00991081	CFSAN000958	JRCD00000000	SRR498431	1203NYJAP-1	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN00991087	CFSAN000960	JRCB00000000	SRR498433	1203NYJAP-1	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN00991100	CFSAN000961	JRCA00000000	SRR498434	1203NYJAP-1	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN01942246	CFSAN000963	JRBZ00000000	SRR498436	1203NYJAP-1	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN01942268	CFSAN000968	JMMF00000000	SRR498442	1203NYJAP-1	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN00996629	CFSAN000970	JRBT00000000	SRR498444	1203NYJAP-1	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN01816358	CFSAN001112	JRBL00000000	SRR1258439	1203NYJAP-1	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN01816359	CFSAN001115	JRBK00000000	SRR1258442	1203NYJAP-1	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN01816351	CFSAN001118	JRBJ00000000	SRR1258443	1203NYJAP-1	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA
SAMN01816352	CFSAN001140	JRBI00000000	SRR1258440	1203NYJAP-1	1203NYJAP-1 - Tuna Scrape Outbreak	FALSE				Dwayne Roberson, FDA

Data Archiving

Permanent repository for data:

- NCBI – sequence data, assemblies, biosamples
  - www.ncbi.nlm.nih.gov/pathogens
  - www.ncbi.nlm.nih.gov/biosample
  - www.ncbi.nlm.nih.gov/bioproject
  - www.ncbi.nlm.nih.gov/sra
  - www.ncbi.nlm.nih.gov/assembly



Food and Drug Administration, Center for Food Safety and Applied Nutrition: GenomeTrakr Project Accession: PRJNA183844 ID: 183844

Currently encompasses whole genome sequencing of cultured pathogens as part of a surveillance project for the rapid detection of outbreaks of foodborne illnesses. Whole genome sequencing of cultured Salmonella as part of the US Food and Drug Administration surveillance project for the rapid detection of outbreaks of foodborne illnesses.

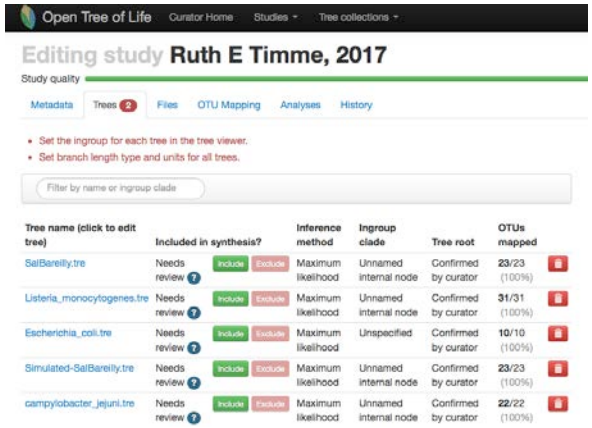
Accession	PRJNA183844
Type	Umbrella project (**Subtype:** Disease)
Keyword	GMI
Publications (total 3)	1. Allard MW *et al.*, "Practical Value of Food Pathogen Traceability through Building a Whole-Genome Sequencing Network and Database.", *J Clin Microbiol*, 2016 Mar 23;54(8):1975-83 More...
Submission	Registration date: 13-Dec-2012 Center for Food Safety and Applied Nutrition
NCBI Links	• NCBI Pathogen Detection
Related Resources	• FDA Whole Genome Sequencing Program (WGS) • NCBI Pathogen FTP • Salmonella Outbreaks • Whole Genome Sequencing (WGS) Fast Facts • FDA GenomeTrakr Video
Relevance	Agricultural, Medical, Industrial, Environmental

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (total)	115323
WGS master	1521
Genomic DNA	50
SRA Experiments	46186
Protein Sequences	7021411
PUBLICATIONS	
PubMed	6
PMC	5
OTHER DATASETS	
BioSample	49572
Assembly	1545

- openTreeOfLife – tree files

  tree.opentreeoflife.org/curator/study/edit/ot_301



Open Tree of Life Curator Home Studies Tree collections

Editing study Ruth E Timme, 2017

Study quality

Metadata Trees 2 Files OTU Mapping Analyses History

- Set the ingroup for each tree in the tree viewer.
- Set branch length type and units for all trees.

Filter by name or ingroup clade

Tree name (click to edit tree)	Included in synthesis?	Inference method	Ingroup clade	Tree root	OTUs mapped
SalBareilly.tre	Needs review Include Exclude	Maximum likelihood	Unnamed internal node	Confirmed by curator	23/23 (100%)
Listeria_monocytogenes.tre	Needs review Include Exclude	Maximum likelihood	Unnamed internal node	Confirmed by curator	31/31 (100%)
Escherichia_coli.tre	Needs review Include Exclude	Maximum likelihood	Unspecified	Confirmed by curator	10/10 (100%)
Simulated-SalBareilly.tre	Needs review Include Exclude	Maximum likelihood	Unnamed internal node	Confirmed by curator	23/23 (100%)
campylobacter_jejuni.tre	Needs review Include Exclude	Maximum likelihood	Unnamed internal node	Confirmed by curator	22/22 (100%)

Working with the Datasets

GenFSgopher.pl -o outdir spreadsheet.dataset.tsv

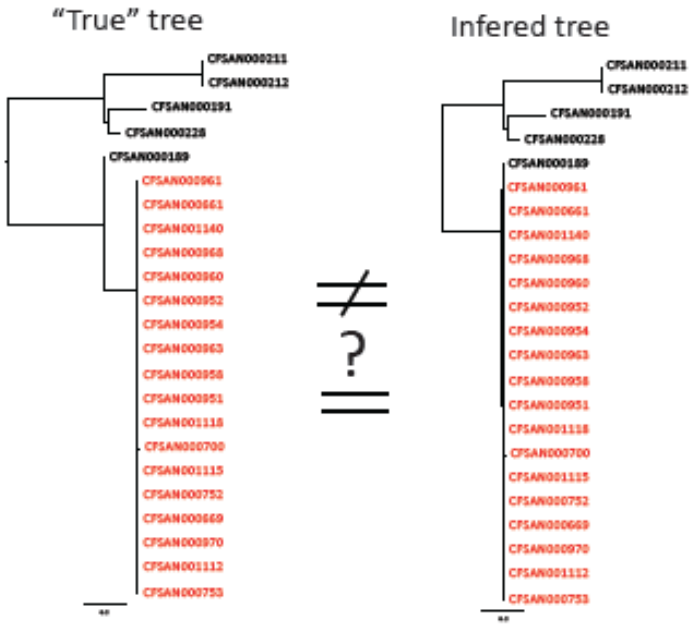
↓

Fastq files, fasta files, newick tree file

↓

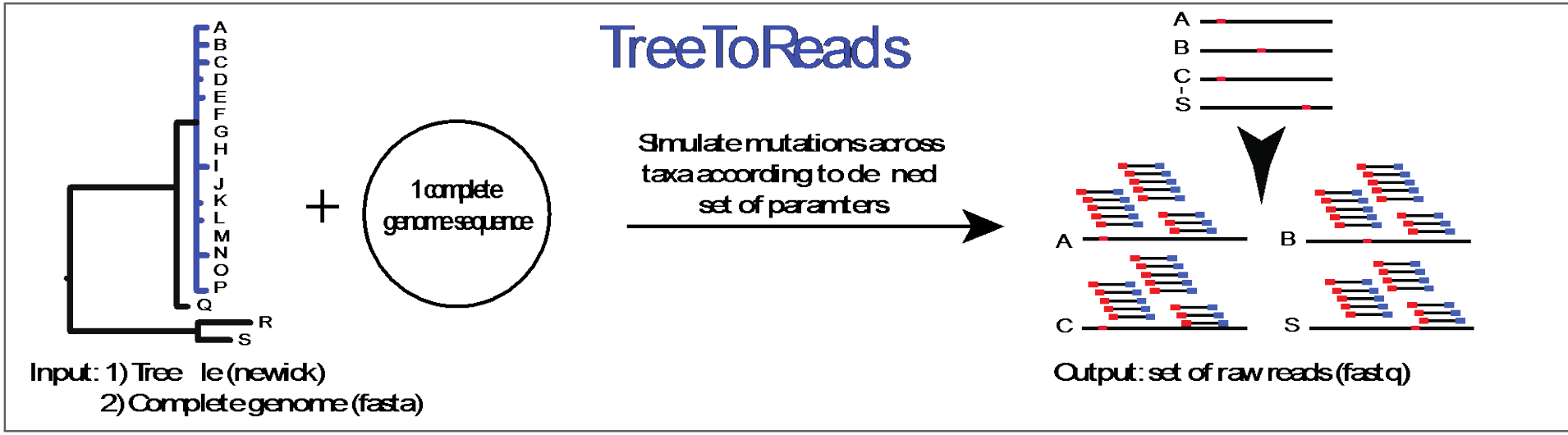
Perform your own clustering pipeline(s)
(FDA: SNP pipeline, CDC: LyveSet, NCBI: Pathogen Detection, …)

Compare results!



TreeToReads: Simulated data pipeline

Collaboration with Emily Jane McTavish, UC Merced



Download here:

https://github.com/snacktavish/TreeToReads

McTavish, E. J., Pettengill, J., Davis, S., Rand, H., Strain, E., Allard, M., & Timme, R. E. (2017). TreeToReads - a pipeline for simulating raw reads from phylogenies. *BMC Bioinformatics*, *18*(1), 178.

Acknowledgements

- Gen-FS WGS Standards and Analysis workgroup (CDC/FDA/NCBI/FSIS)
  - FDA: Ruth Timme, Hugh Rand, Steven Davis
  - CDC: Lee S. Katz, Eija Trees, Heather A. Carleton
  - NCBI: Richa Agarwala, Martin Shumway, Bill Klimke
  - FSIS: Mustafa Simmons, Glenn Tillman, Philip Bronstein, Stephanie Defibaugh-Chávez
- Global Microbial Identifier – WG3
- Other contributors:

  Errol Strain

  Darlene Wagner

  Cheryl Tarr

  Maria Hoffmann

  Maria Sanchez-Leon


FDA
U.S. FOOD & DRUG
ADMINISTRATION