Bacterial benchmark datasets for comparison and validation of phylogenomic pipelines

Standards for Pathogen Detection Workshop

NIST, 14 August 2017

Ruth E. Timme, PhD
Research Microbiologist
GenomeTrakr data manager

Genomics and Food Safety (Gen-FS)

Working Group: WGS Standards and Analysis

- Participants: FDA, CDC, NCBI and FSIS, APHL

Benchmarks allow:

- Evaluation of method performance
- Guidance on improvement of methods
- Standards for assessing reproducibility of results
- Bolster use of results for regulatory action

Benchmark datasets

Dataset	Organism	Number of Isolates [a]	Epidemiologically linked Isolates [b]	reference genome[c]	Type of dataset	Reference/Comment
Stone Fruit recall	*L. monocytogenes*	31	28	CFSAN023463	Empirical	PMID: 27694232
Spicy Tuna outbreak	*S. enterica*	23	18	CFSAN000189	Empirical	PMID: 25995194
Raw Milk Outbreak	*C. jejuni*	22	14	D7331	Empirical	http://www.outbreakdatabase.com/details/endricks-farm-and-dairy-raw-milk-2008/
Sprouts Outbreak	*E. coli*	10	3	2011C-3609	Empirical	http://www.cdc.gov/ecoli/2014/o121-05-14/index.html
Simulated outbreak	*S. enterica*	23	18	CFSAN000189	Synthetic	Simulated dataset based off the *S. enterica* spicy tuna outbreak tree and reference genome.

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	suggestedReference	sha256sumAssembly	sha256sumRead1	sha256sumRead2	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

GitHub:

https://github.com/WGS-standards-and-analysis/datasets

Gen-FS Gopher download script

dataset table files

NCBI:

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

Metadata: BioSample Database

Raw sequence data: SRA database

Assembled genomes: Assembly database

OpenTreeOfLife:

https://tree.opentreeoflife.org

newick formated tree files

Working with the Datasets

visit WGS standards GitHub site

↓

"git clone" to download tables and script (or zip file)

↓

downloadDataset.pl -o outdir spreadsheet.dataset.tsv

↓

Downloaded files, verify integrity with checksums:
- fastq files
- fasta assemblies
- tree file (newick format)

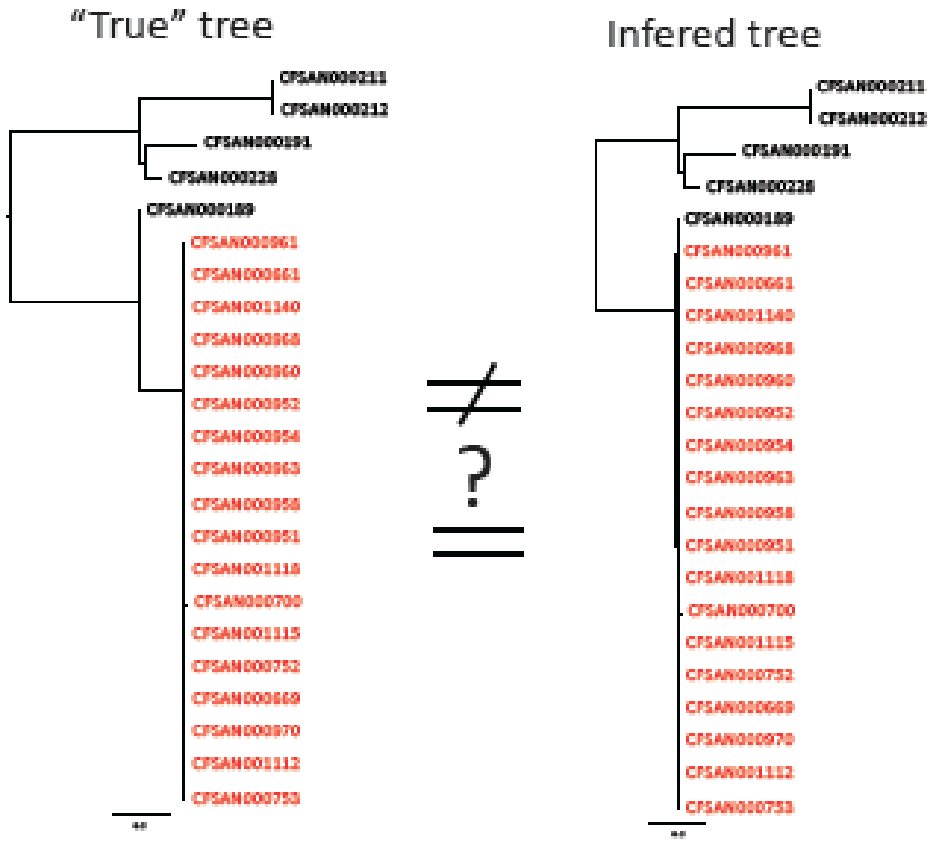
↓

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

↓

Compare results!

Compare results!



Manuscript accepted at PeerJ.