

NCBI Pathogen Detection: Facilitating Traceback and Outbreak Investigation of Pathogen Genome Sequences
in Real-Time Using an Automated SNP Clustering Analysis Pipeline.

William Klimke

NIST

Standards for Pathogen Detection Workshop

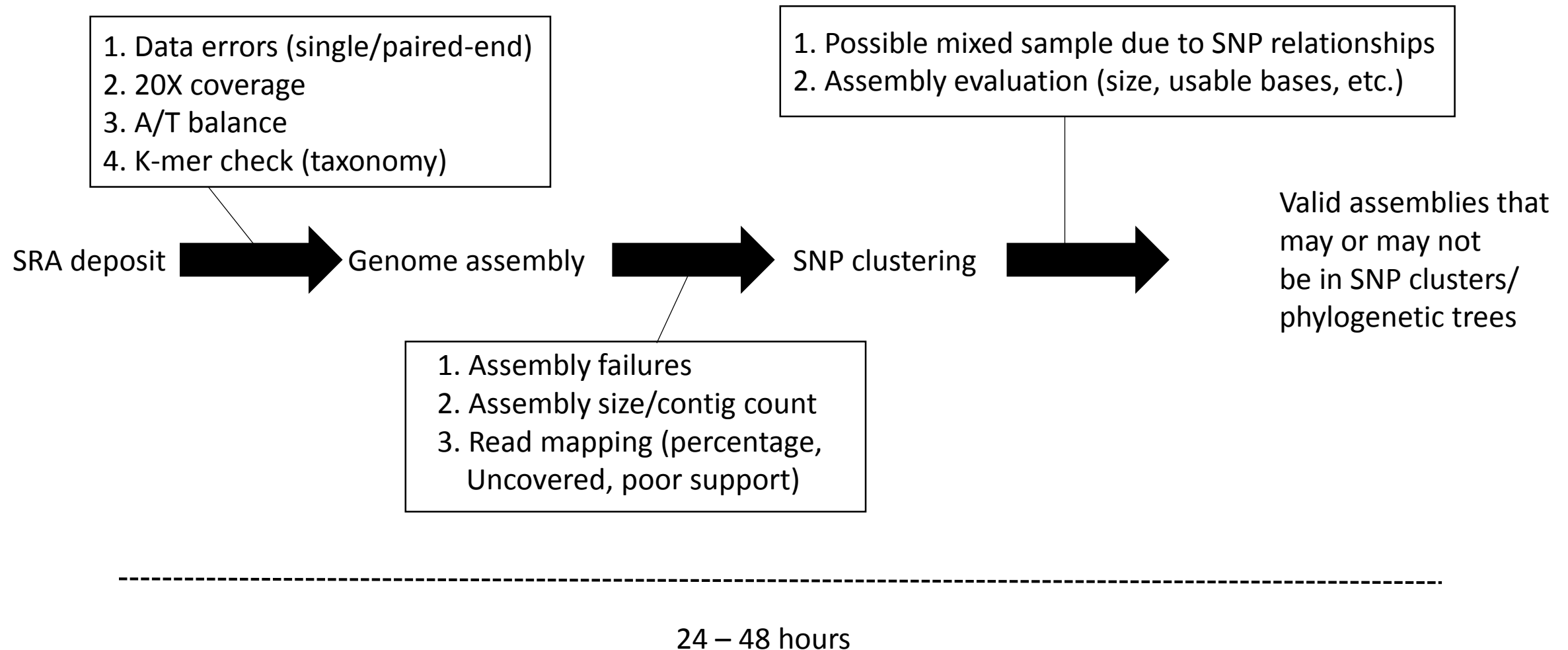
Aug 14-15

[Health](#) > [Pathogen Detection](#)

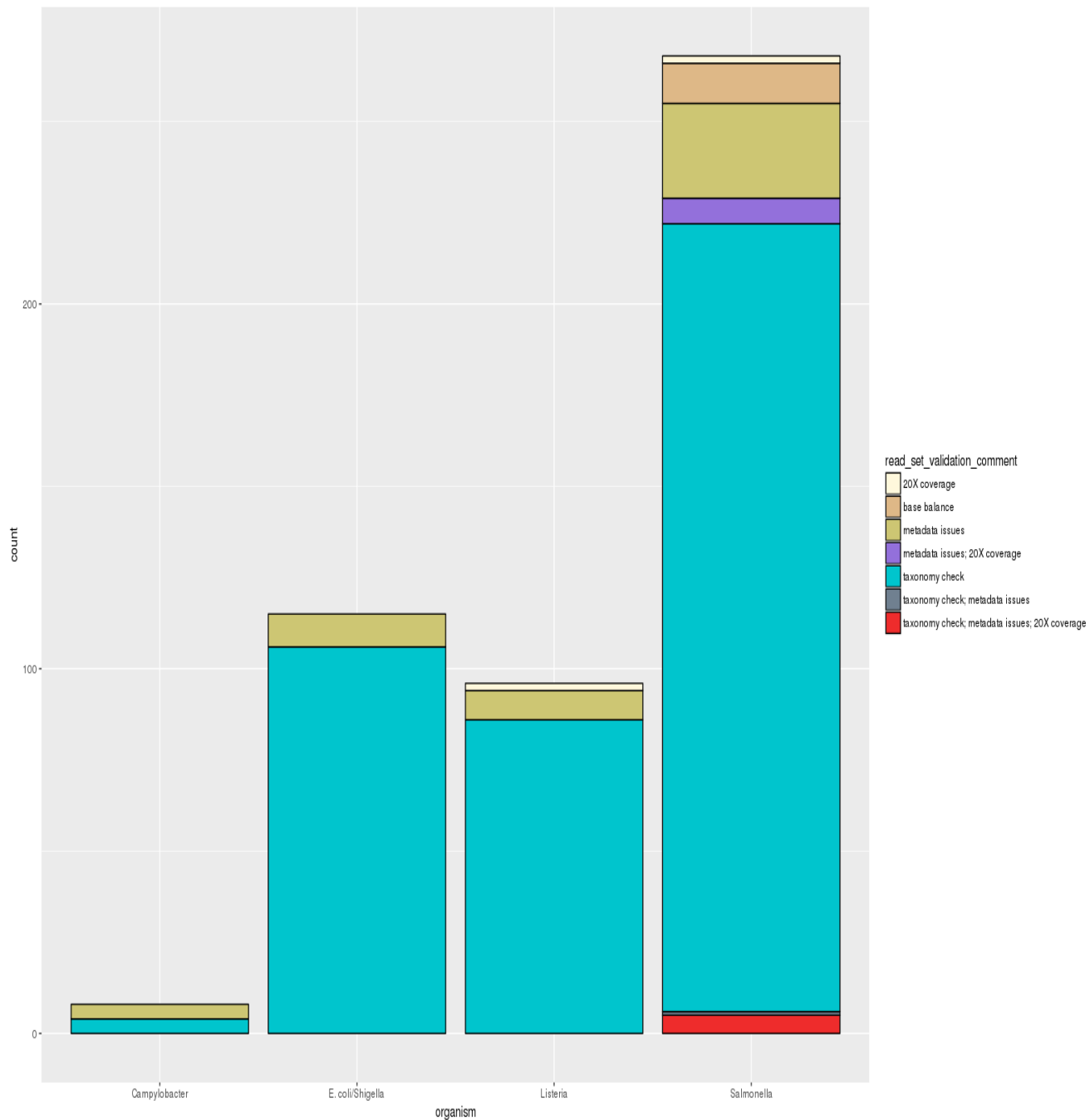
Pathogen Detection **BETA**

NCBI Pathogen Detection integrates bacterial pathogen genomic sequences originating in food, environmental sources, and patients. It quickly clusters and identifies related sequences to uncover potential food contamination sources, helping public health scientists investigate foodborne disease outbreaks.

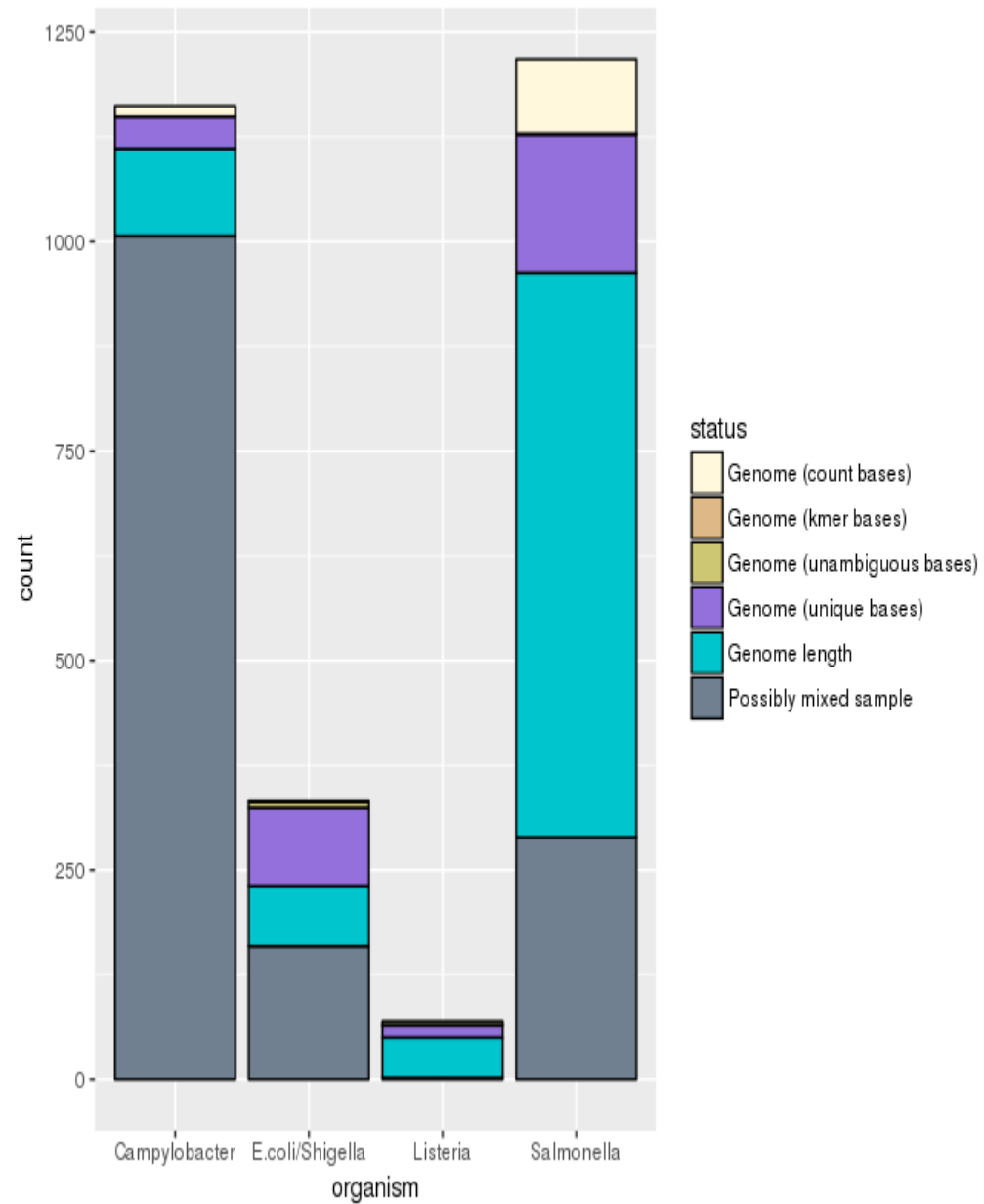
Basic Outline of NCBI Pathogen Detection Pipeline



Read validation issues Pathogen Detection
by organism



Number of isolates with exceptions in
SNP clustering step (post assembly)



Validaton Tests

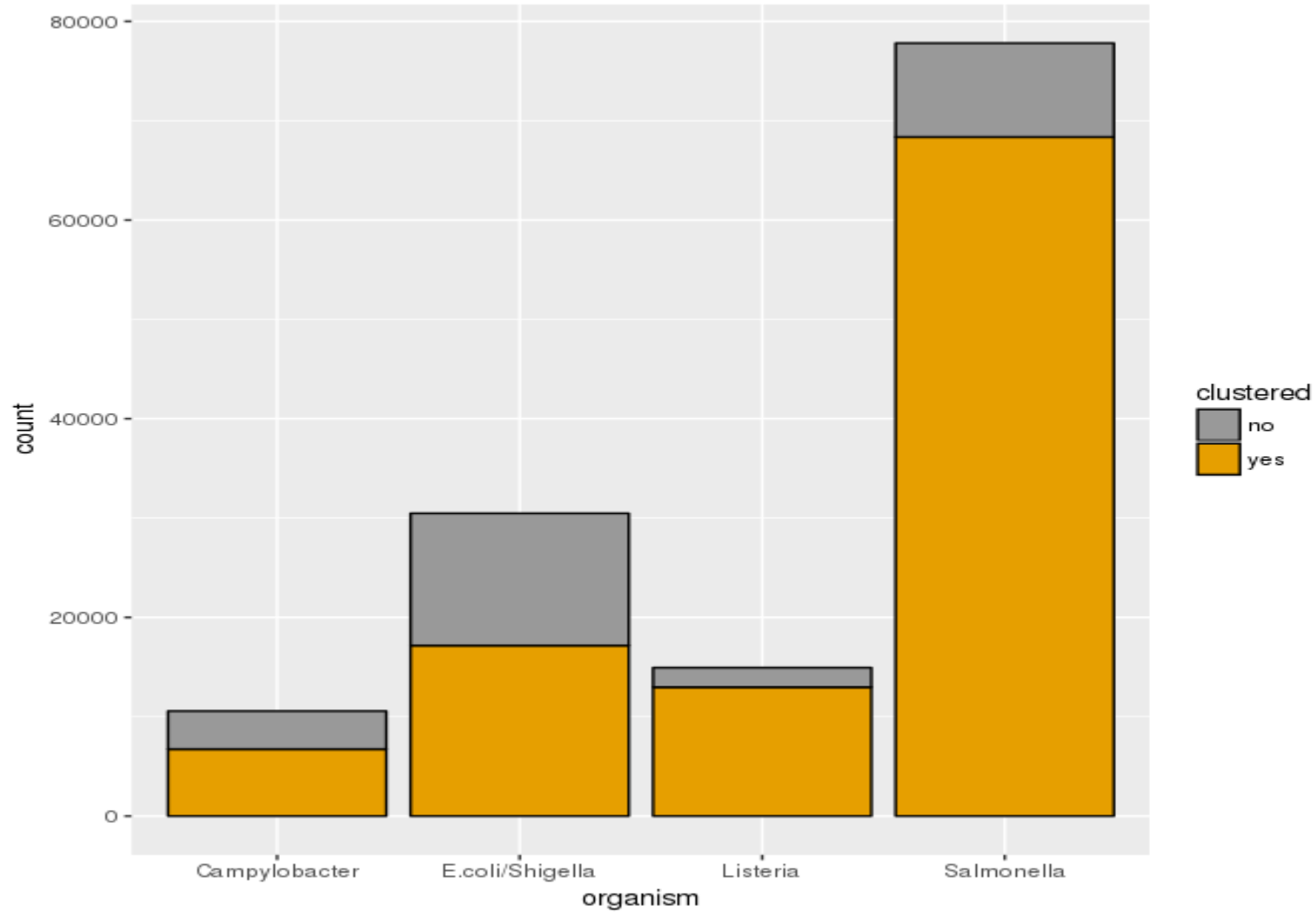
Pipeline stage

Status

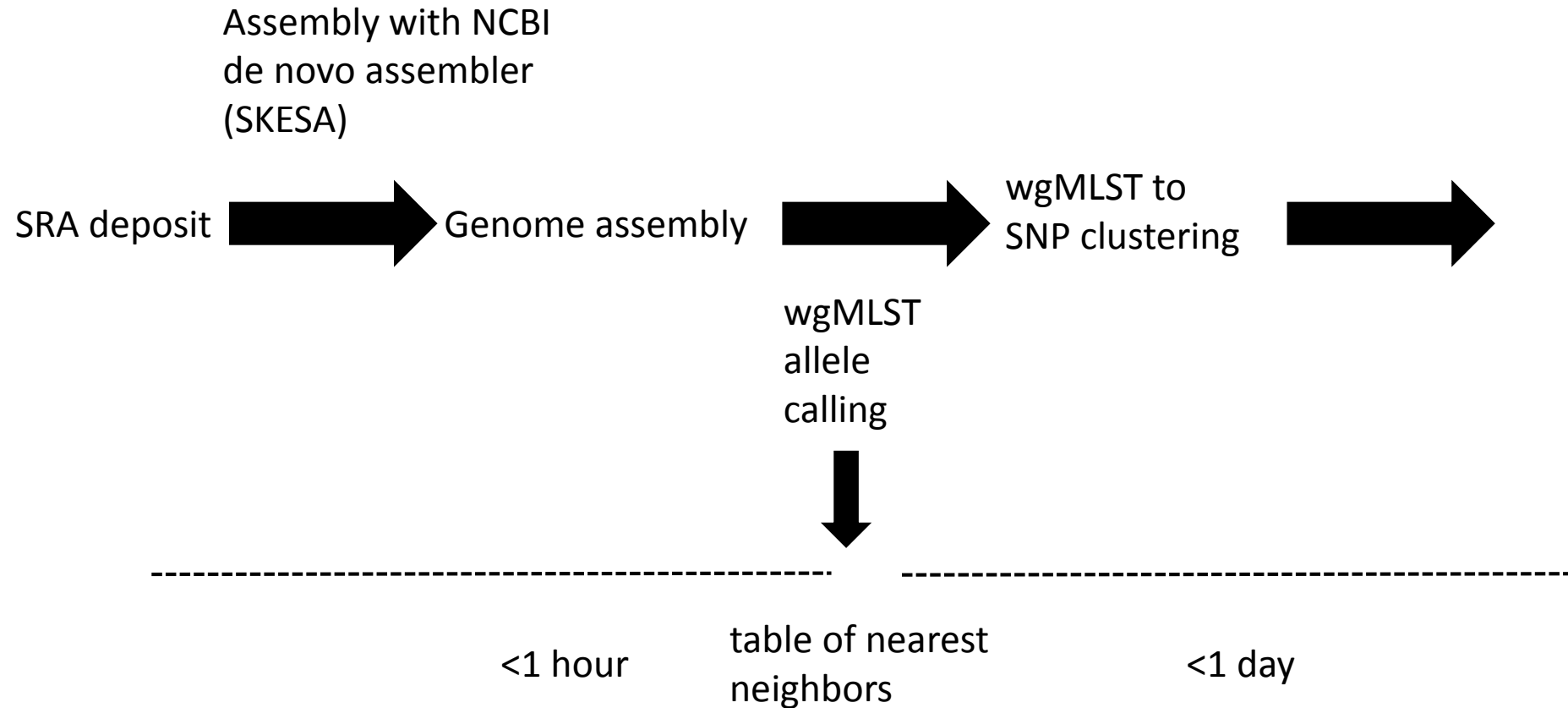
Test	Pipeline stage			Status		
	Read set validation/ submission	Assembly validation	SNP Processing Validation	Implemented	Validated	Discussed with collaborators
Duplicate data submission	done			done	done	done
20X coverage	done			done	done	done
A/T Balance	done			done	done	done
Taxonomy check	done			done	testing	done
Metadata errors	done			done	done	done
Multi-run check	done			testing	not done	not done
Assembly failures		done		done	done	done
Assembly length		done		done	testing	done
No. of contigs		done		done	not done	done
N50		done		done	not done	done
L50		done		done	not done	done
Filtered mapping rate		done		done	not done	done
Poor quality support bases		done		done	not done	done
Uncovered bases		done		done	not done	done
Contaminaton check		testing		not done	not done	not done
Assembly length			done	done	done	done
Unique bases			done	done	testing	not done
Unique k-mers			done	done	testing	not done
Unambiguous bases			done	done	testing	not done
No. of bases			done	done	testing	not done
Possible mixed samples			done	done	done	done

done
 testing
 not done

Total isolates clustered
vs. non clustered per organism



New Proposed Pipeline based on wgMLST



Organism	No. of loci	No. of alleles	No. of genomes	No. of loci for rapid reports
Campylobacter	6823	9813	7173	1175
Escherichia coli/Shigella spp.	12427	14901	21345	3105
Listeria	4992	5834	12771	2150
Salmonella	13362	16721	60840	3340

NCBI Pathogen Detection Isolates Browser

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Log in

[Health](#) > [Pathogen Detection](#)

Pathogen Detection BETA

NCBI Pathogen Detection integrates bacterial pathogen genomic sequences originating in food, environmental sources, and patients. It quickly clusters and identifies related sequences to uncover potential food contamination sources, helping public health scientists investigate foodborne disease outbreaks.

[Find isolates now!](#)

Explore the Data

Species	New Isolates	Total Isolates
Salmonella enterica	30	55,300
E.coli and Shigella	94	23,701
Listeria monocytogenes	23	12,604
Campylobacter jejuni	13	4,625
Acinetobacter baumannii	68	2,897

Learn More

- [About](#)
- [FAQ](#)
- [Antimicrobial Resistance](#)
- [Contributors](#)

Data Resources

- [Isolates Browser](#)
- [Antimicrobial resistance reference gene database](#)
- [Isolates with antibiotic resistant phenotypes](#)
- [Beta-lactamase resources](#)
- [Download analysis results \(FTP\)](#)

Submit

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

Pathogen Isolates Browser

Escherichia_coli_Shigella ✕ Q Search

Filters 1 ✕

Organism group E.coli and Shigella (3,580)

Location CA (234) CO (144) LONDON (121) MIDLANDS AND EAST OF ENGLAND (102) NONE (118) NORTH OF ENGLAND (138) PA (434) SOUTH OF ENGLAND (129) UNITED KINGDOM (609) USA (2,622)

Source bovine, cow-Feces (66) environmental isolate (75) Feces-Bovine (44) Feces-Chicken (51) food (43) Human (608) porcine Pleural Cavity (279) Spinach (44) Stool (144) Urine (46)

Collected by CDC (519) CDPH FDLB (18) FDA (515) FLUFL (62) GA (16) Penn State E. coli Reference Center (741) PHE (608) TX (47) USDA-FSIS (64) Xiaomin Zhao (15)

Host Bos taurus (1) bovine (1) cattle (2) chicken (1) Gallus gallus breed Broiler (1) Gallus gallus domesticus (1)

Target Creation
Thu Jul 07 2016 05:31:34 GMT-0400 (Eastern Daylight Time) -> Fri Sep 16 2016 17:13:25 GMT-0400 (Eastern Daylight Time) [reset](#)

Scientific Name filter by Scientific Name

Select columns

14 items selected Remove all Add all

<input checked="" type="checkbox"/>	scientific_name	-	asm_acc	+ +
<input checked="" type="checkbox"/>	Target	-	asm_display_name	+ +
<input checked="" type="checkbox"/>	Cluster	-	asm_level	+ +
<input checked="" type="checkbox"/>	New?	-	asm_stats_contig_n50	+ +
<input checked="" type="checkbox"/>	Min-same	-	asm_stats_length_bp	+ +
<input checked="" type="checkbox"/>	Min-opp	-	asm_stats_n_contig	+ +
<input checked="" type="checkbox"/>	attribute_package	-	assembly_method	+ +
<input checked="" type="checkbox"/>	biosample_acc	-	bioproject_acc	+ +
<input checked="" type="checkbox"/>	sample_name	-	clustered	+ +
<input checked="" type="checkbox"/>	Links	-	complete_fl	+ +
<input checked="" type="checkbox"/>	collected_by	-	Epitype	+ +
<input checked="" type="checkbox"/>	collection_date	-	erd_group_acc	+ +
<input checked="" type="checkbox"/>	isolation_source	-	fullasm_id	+ +
<input checked="" type="checkbox"/>	target_creation_date	-	geo_loc_name	+ +
<input type="checkbox"/>			HHS_region	+ +

Pathogen Isolates Browser

Escherichia_coli_Shigella ✕ Q Search

Filters 1 **Columns** **Download**

minimum SNP distance to isolates from all different epitypes in the SNP cluster

1 of 180

.	scientific_name	Target	Cluster	New?	Min-same	Min-opp	attribute_package	biosample_acc	collection_date	isolation_source	geo_loc_name	target_creation_date	SNP_Tree
1	Escherichia coli	PDT000137182.1	PDS000003441.41	false	0	0	Pathogen: environmental/food/other	SAMN05245391	FDA	2016-06-02	All-Purpose Flour	2016-07-07	SNP_Tree
2	Escherichia coli	PDT000137183.1	PDS000003441.41	false	0	0	Pathogen: environmental/food/other	SAMN05245392	FDA	2016-06-02	All-Purpose Flour	2016-07-07	SNP_Tree
3	Escherichia coli	PDT000137185.1	PDS000003441.41	false	0	0	Pathogen: environmental/food/other	SAMN05245394	FDA	2016-06-02	All-Purpose Flour	2016-07-07	SNP_Tree
4	Escherichia coli	PDT000137186.1	PDS000003441.41	false	0	0	Pathogen: environmental/food/other	SAMN05245395	FDA	2016-06-02	All-Purpose Flour	2016-07-07	SNP_Tree
5	Escherichia coli	PDT000137188.1	PDS000003441.41	false	0	0	Pathogen: environmental/food/other	SAMN05245618	FDA	2016-06-02	All-Purpose Flour	2016-07-07	SNP_Tree
6	Escherichia coli	PDT000137189.1	PDS000003441.41	false	0	0	Pathogen: environmental/food/other	SAMN05245624	FDA	2016-06-02	All-Purpose Flour	2016-07-07	SNP_Tree
7	Escherichia coli	PDT000137191.1	PDS000003441.41	false	0	0	Pathogen: environmental/food/other	SAMN05245621	FDA	2016-06-02	All-Purpose Flour	2016-07-07	SNP_Tree
8	Escherichia coli	PDT000137812.1	PDS000006928.2	false	n/a	0	Generic	SAMEA1905156				2016-07-13	SNP_Tree
9	Escherichia coli	PDT000137880.1	PDS000007044.1	false	n/a	0	Generic	SAMEA1905135				2016-07-13	SNP_Tree
10	Escherichia coli	PDT000139198.1	PDS000003441.41	false	0	0	Pathogen: clinical or host-associated	SAMN05389709	PNUSAE003616	CDC		2016-07-19	SNP_Tree

NCBI Isolates Browser (search/filter/sort – links to SNP trees)

E.coli and Shigella (PDG00000004.535)
 Updated: 2016-09-15T19:51:05Z Previous:2016-09-13T16:59:56Z Targets: 18747
 Tree (PDS000003441.41)

Target Metadata

Target	Cluster	New?	Min-same	Min-opp	attribute_package	biosample_acc	sample_name	serovar	asm_acc
PDT000025649.2	PDS000003441.41	false	3		39: Pathogen: clinical or host-associated	SAMN02353016	2011C-3108	O121:H19	GCA_000614215.2

E.coli and Shigella (PDG00000004.535)
 Updated: 2016-09-15T19:51:05Z Previous:2016-09-13T16:59:56Z Targets: 18747
 Tree (PDS000003441.41)

Target Metadata

New?	Min-same	Min-opp	attribute_package	biosample_acc	sample_name	asm_acc	collected_by	collection_date	geo_loc_name	isolation_source
false	0	0	Pathogen: environmental/food/other	SAMN05245391			FDA	2016-06-02	USA	All-Purpose Flour
false	0	0	Pathogen: environmental/food/other	SAMN05245392			FDA	2016-06-02	USA	All-Purpose Flour
false	1	2	Pathogen: environmental/food/other	SAMN05245393			FDA	2016-06-02	USA	All-Purpose Flour
false	0	0	Pathogen: environmental/food/other	SAMN05245394			FDA	2016-06-02	USA	All-Purpose Flour
false	0	0	Pathogen: environmental/food/other	SAMN05245395			FDA	2016-06-02	USA	All-Purpose Flour
false	1	1	Pathogen: environmental/food/other	SAMN05245396			FDA	2016-06-02	USA	All-Purpose Flour
false	0	0	Pathogen: environmental/food/other	SAMN05245618			FDA	2016-06-02	USA	All-Purpose Flour
false	0	0	Pathogen: environmental/food/other	SAMN05245624			FDA	2016-06-02	USA	All-Purpose Flour
false	1	1	Pathogen: environmental/food/other	SAMN05245744			FDA	2016-06-02	USA	All-Purpose Flour

NCBI SNP Tree Viewer (examine closely related isolates)
subtree in red is E. coli O121 flour outbreak

New version of SNP Tree Viewer to enhance navigation

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information

Home / Isolates Browser / Collapsed Tree View | Whole Tree View

Escherichia_coli_Shigella - 715 isolates
PDG000000004.757 / PDS0000000952.163

12 Isolates Selected ✖ Clear

Distance between selected isolates:
minimum=0 SNPs, maximum= 13 SNPs, average=5 SNPs

Target creation date range among selected isolates:
2015-12-31 to 2017-04-25

Filter selected isolates

- 2017 environmental/other | USA:NV 2017-04-25
- 2016 environmental/other | USA:ID 2016-04-26
- 2015 clinical | USA 2015-12-31
- clinical | USA 2015-12-31
- clinical | USA 2015-12-31
- clinical | USA 2015-12-31
- clinical | USA 2015-12-31
- clinical | USA 2015-12-31
- clinical | USA 2015-12-31
- clinical | USA 2015-12-31
- clinical | USA 2015-12-31
- clinical | USA 2015-12-31
- clinical | USA 2015-12-31
- clinical | USA 2015-12-31

#	Strain	Serovar	Isolate ID	Create Date	Location	Isolation Source	Isolation type	Host	Min-same	Min-diff	BioSample
1	PNUSAE0018	E. coli O157:H7	PDT000101092.1	2015-12-31	USA	Stool	clinical		1		3 SAMN04351450
2	PNUSAE0018	E. coli O157:H7	PDT000101093.1	2015-12-31	USA	Stool	clinical		0		4 SAMN04351451
3	PNUSAE0018	E. coli O157:H7	PDT000101094.1	2015-12-31	USA	Stool	clinical		1		4 SAMN04351452
4	PNUSAE0018	E. coli O157:H7	PDT000101095.1	2015-12-31	USA	Stool	clinical		2		5 SAMN04351453
5	PNUSAE0018	E. coli O157:H7	PDT000101096.1	2015-12-31	USA	Stool	clinical		1		4 SAMN04351454

Choose Columns Page 1 of 143

Spacing DIM UNSELECTED OFF TREE TIPS OFF PDT000101093

Acknowledgements

**Richa Agarwala
Azat Badretdin
Slava Brover
Joshua Cherry
Vyacheslav
Chetvernin
Robert Cohen
Michael DiCuccio
Boris Fedorov
Mike Feldgarden
Lewis Geer
Dan Haft
Lianyi Han
William Klimke
Alex Kotliarov
Arjun Prasad
Edward Rice
Kirill
Rotmistrovskyy**

**Stephen Sherry
Sergey Shiryev
Martin Shumway
Tatiana Tatusova
Igor Tolstoy
Chunlin Xiao
Leonid Zaslavsky
Alexander
Zasytkin
Alejandro A.
Schaffer
Lukas Wagner
Aleksandr Morgulis**

**David Lipman
James Ostell**

**CDC
FDA/CFSAN
USDA-FSIS
PHE/FERA
NIHGRI
NIAID
WRAIR
Broad
Wadsworth/MDH**

pd-help@ncbi.nlm.nih.gov

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

<http://www.ncbi.nlm.nih.gov>

National Center for Biotechnology Information – National Library of Medicine – Bethesda MD 20892 USA