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

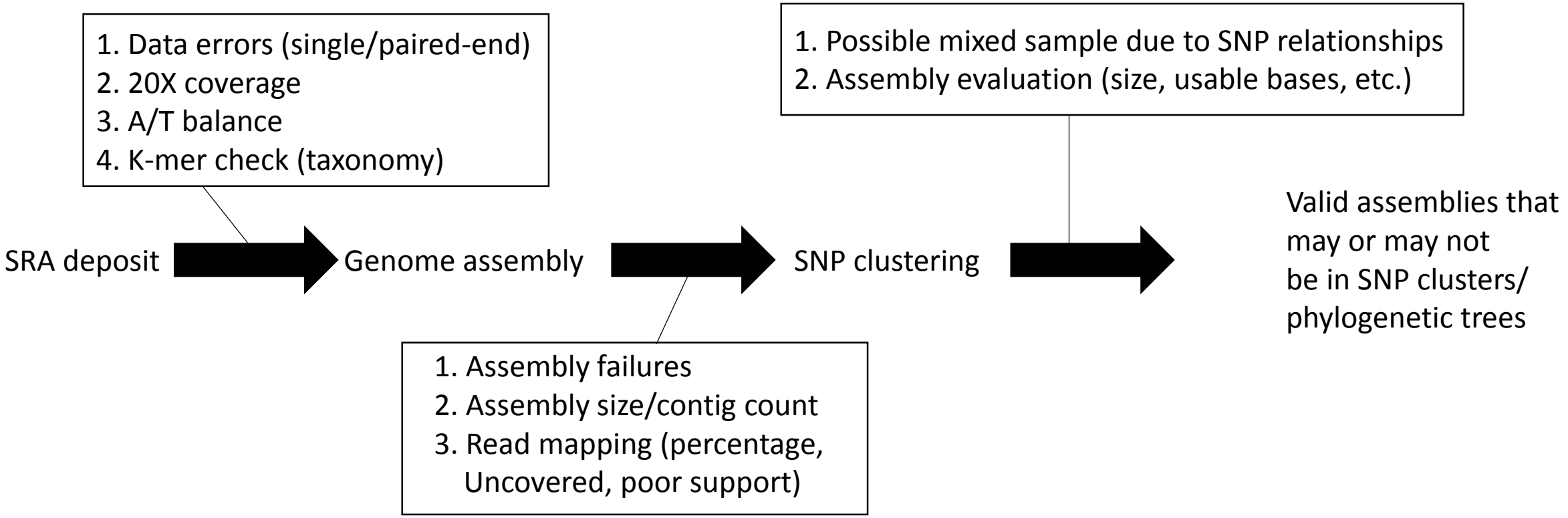
William Klimke
NIST
Standards for Pathogen Detection Workshop
Aug 14-15

Health > Pathogen Detection

Pathogen Detection BETA

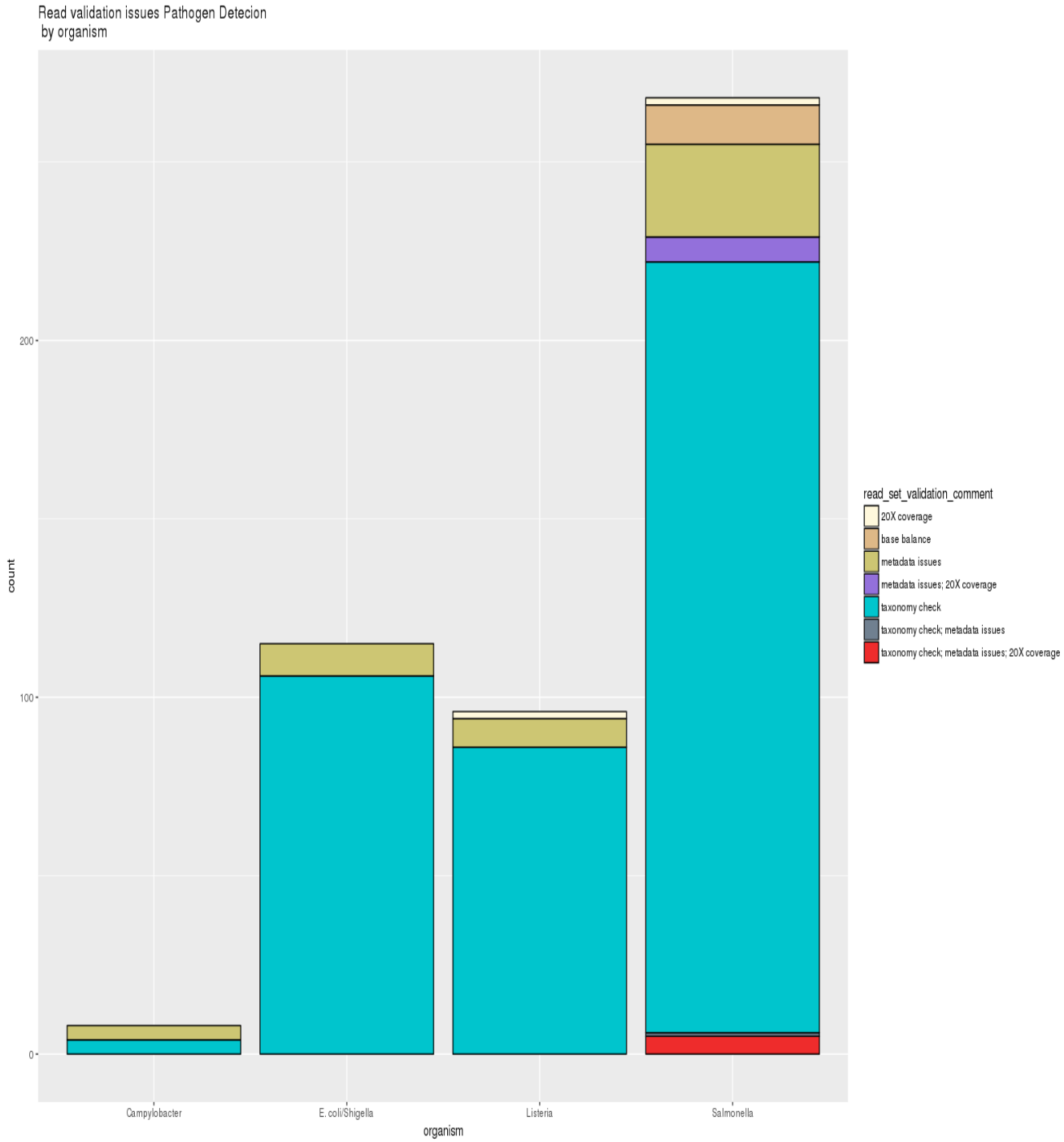
NCBI Pathogen Detection integrates bacterial pathogen genomic sequences originating in food, environmental sources, and patients. It quickly clusters and identifies related sequences to uncover potential food contamination sources, helping public health scientists investigate foodborne disease outbreaks.

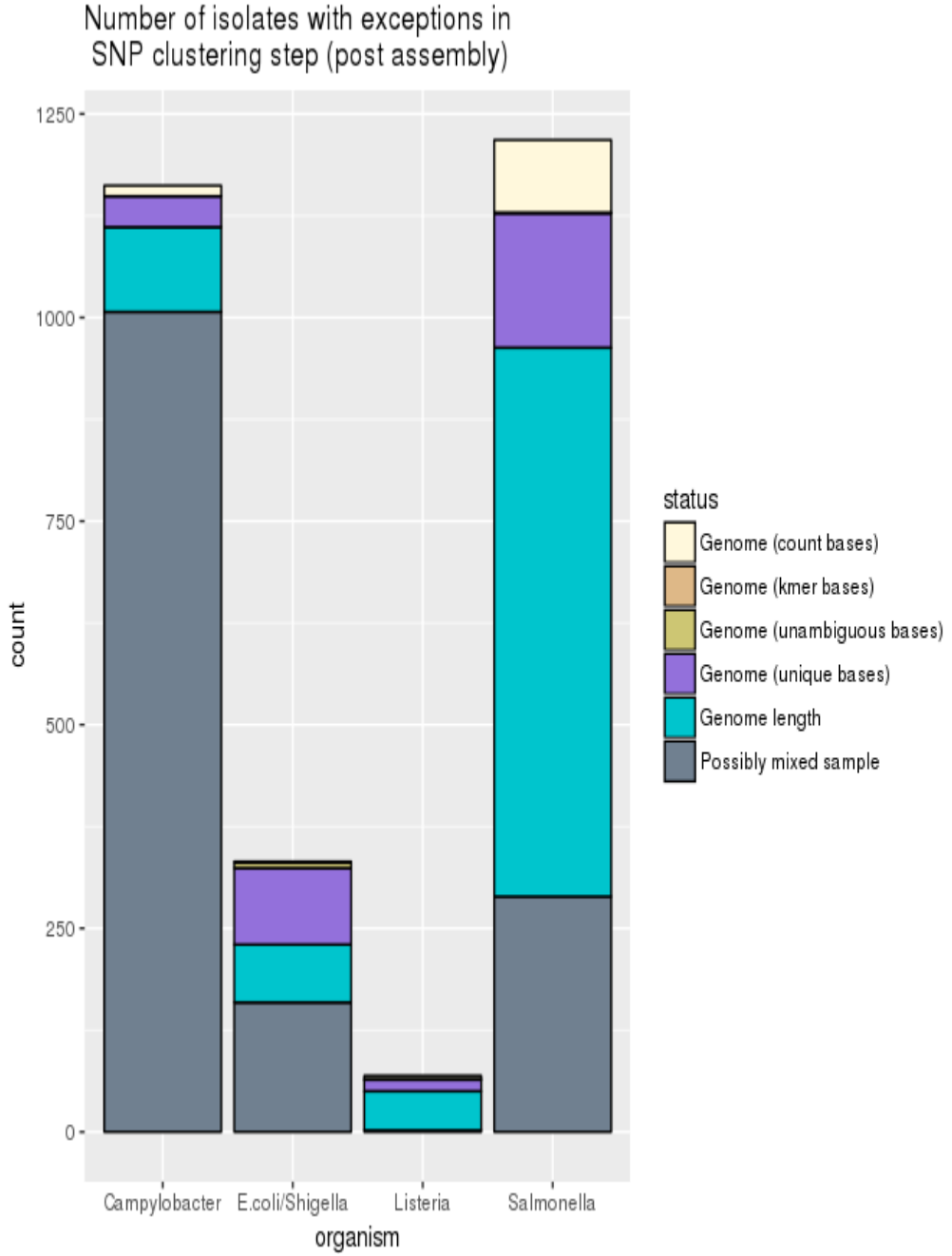
Basic Outline of NCBI Pathogen Detection Pipeline



---

24 – 48 hours

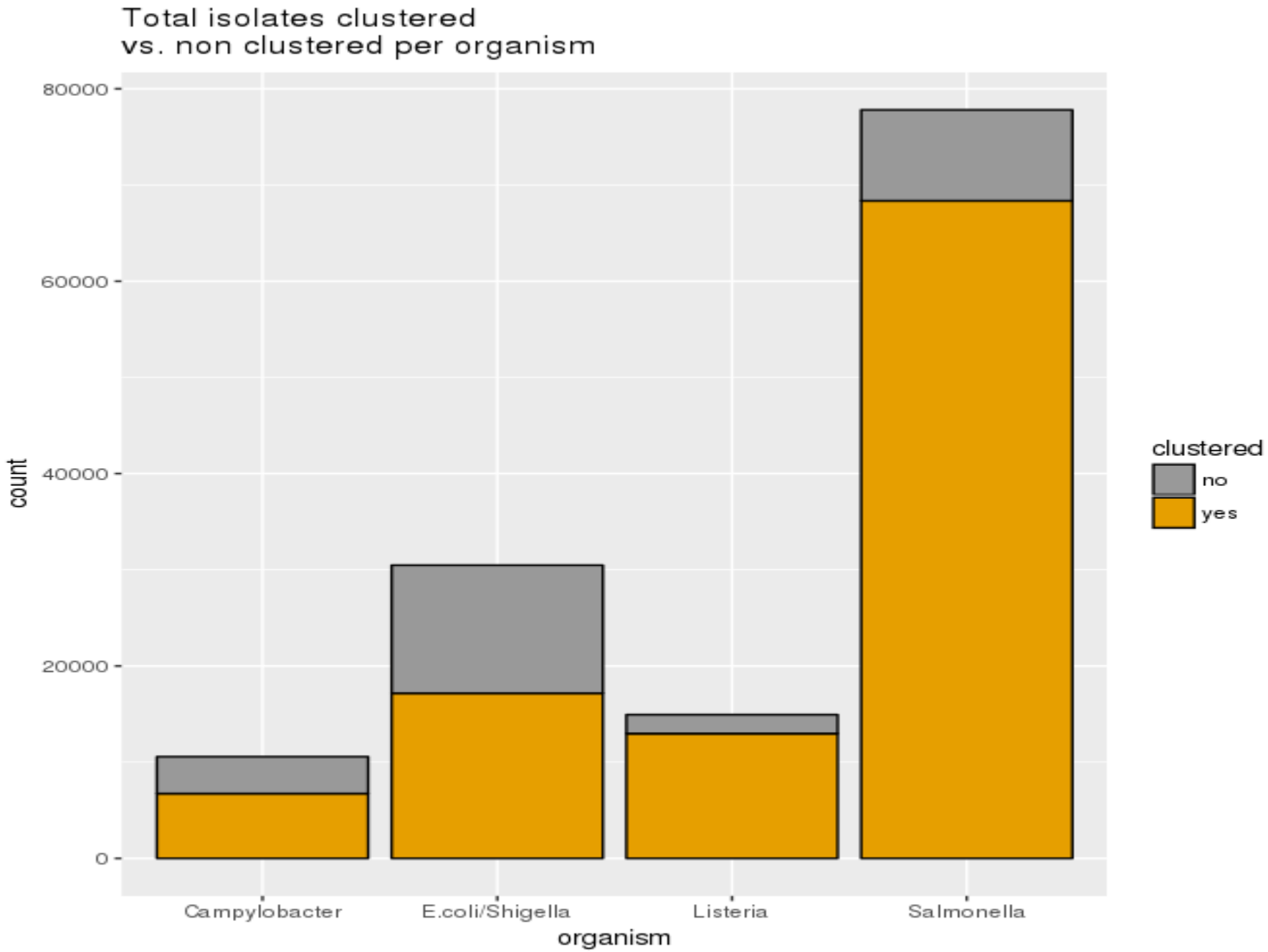

Read validation issues Pathogen Detecion
by organism
count
200
100
0
Campylobacter
E.coli/Shigella
Listeria
Salmonella
organism
read_set_validation_comment
20X coverage
base balance
metadata issues
metadata issues; 20X coverage
taxonomy check
taxonomy check; metadata issues
taxonomy check; metadata issues; 20X coverage


Number of isolates with exceptions in
SNP clustering step (post assembly)
count
1250
1000
750
500
250
0
Campylobacter
E.coli/Shigella
Listeria
Salmonella
organism
status
Genome (count bases)
Genome (kmer bases)
Genome (unambiguous bases)
Genome (unique bases)
Genome length
Possibly mixed sample

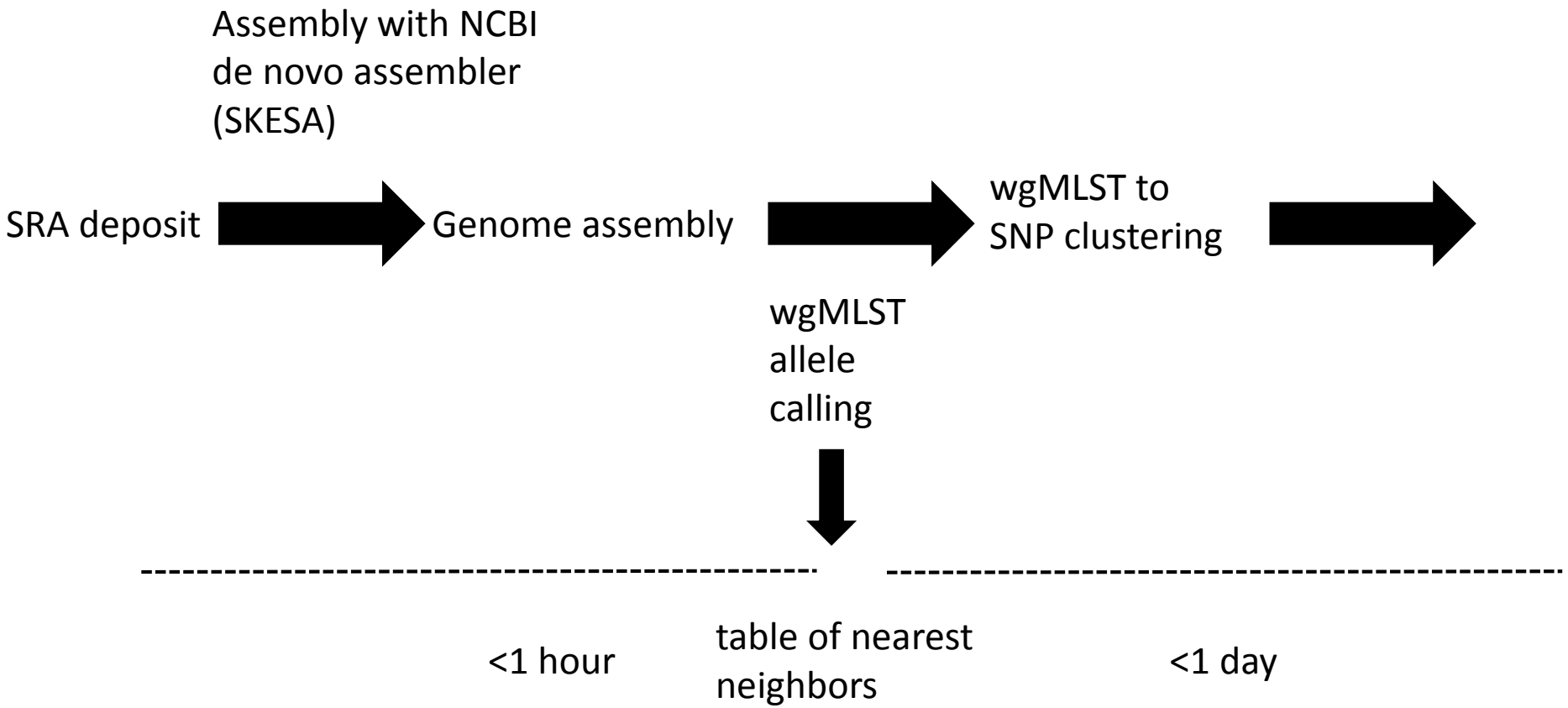
Validaton Tests | **Pipeline stage** | **Status**

Test	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	done	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	testing	not 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	testing	not done

done
testing
not done

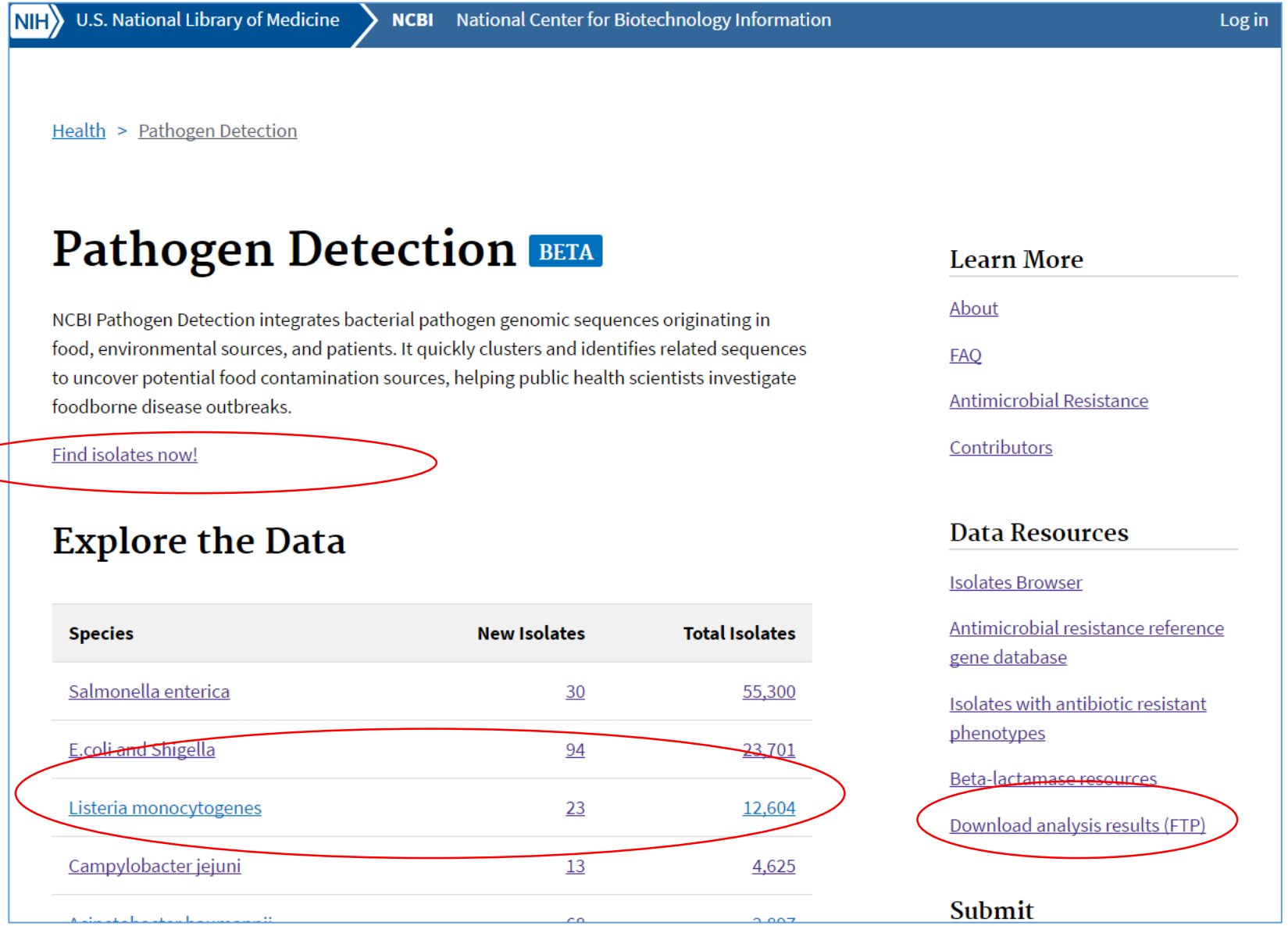

Total isolates clustered
vs. non clustered per organism
80000
60000
40000
20000
0
count
Campylobacter
E.coli/Shigella
Listeria
Salmonella
organism
clustered
no
yes

New Proposed Pipeline based on wgMLST

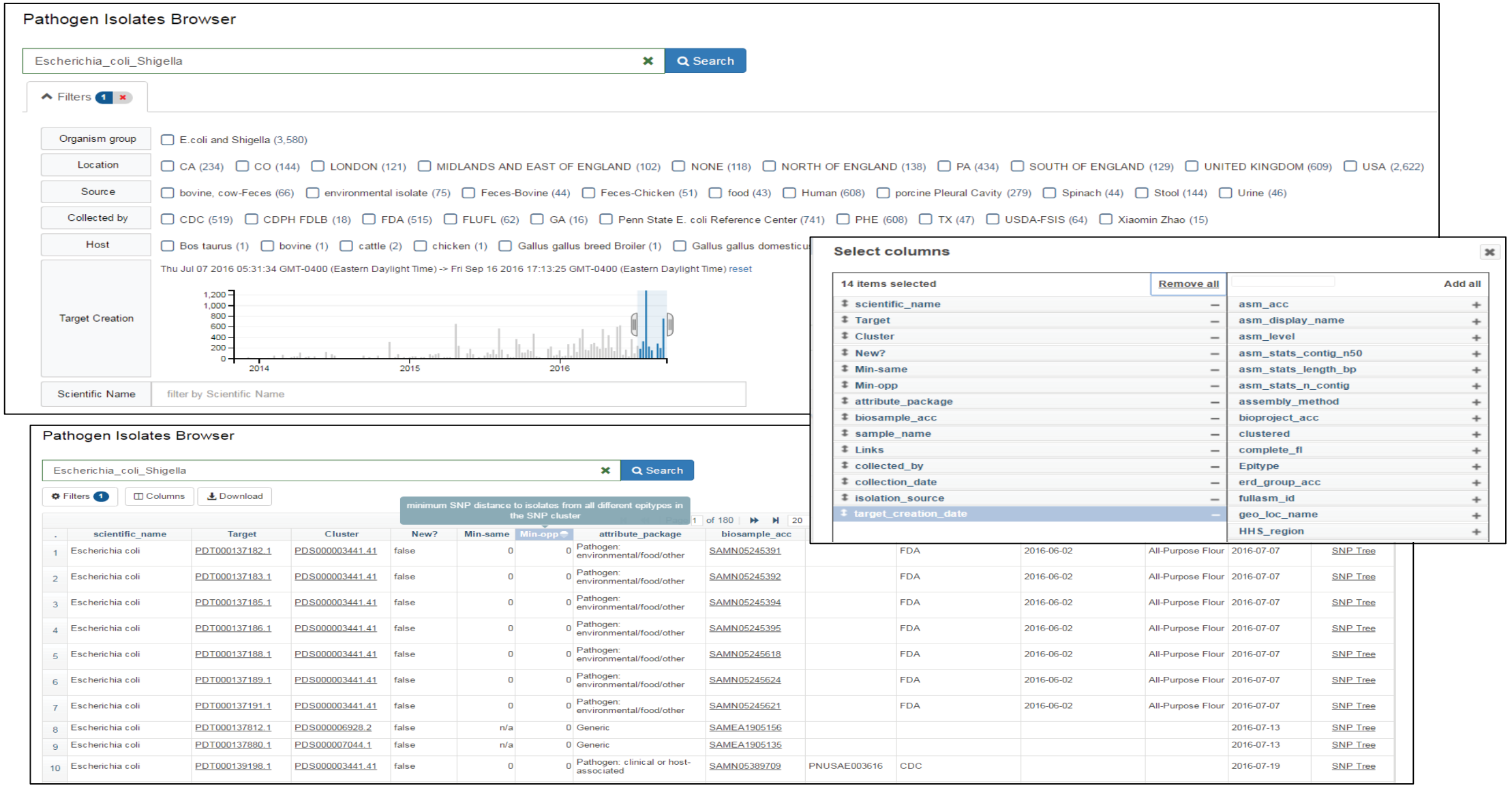


Organism	No. of loci	No. of alleles	No. of genomes	No. of loci for rapid reports
Campylobacter	6823	9813	7173	1175
Escherichia coli/Shigella spp.	12427	14901	21345	3105
Listeria	4992	5834	12771	2150
Salmonella	13362	16721	60840	3340

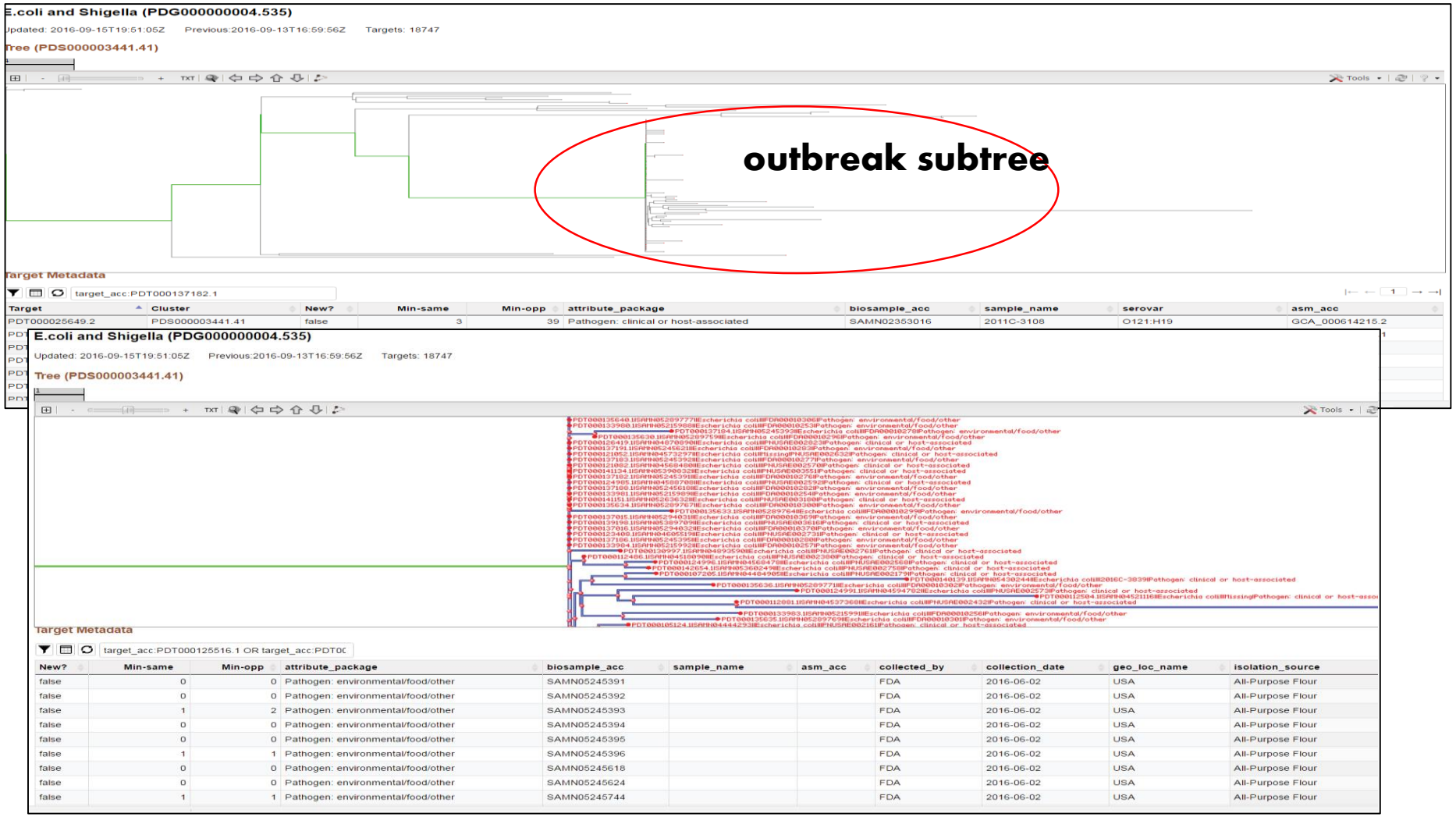
NCBI Pathogen Detection Isolates Browser



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



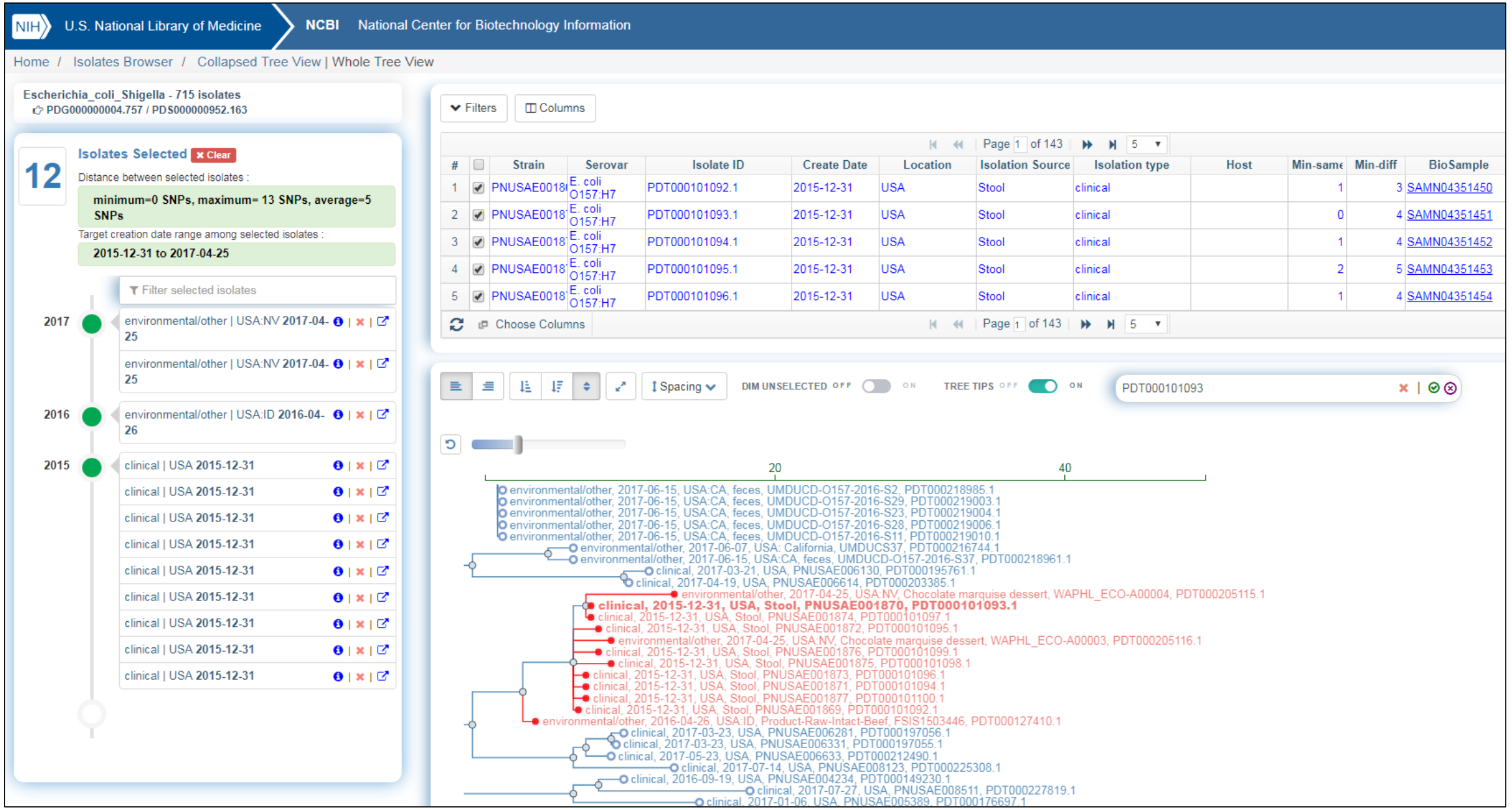
NCBI Isolates Browser (search/filter/sort – links to SNP trees)



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



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 Zasypkin
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.