Publically available tools for WGS analysis at the National Center for Biotechnology Information (NCBI)

William Klimke

Use of Whole Genome Sequence (WGS) Analysis to Improve Food Safety and Public Health

October 26–27, 2017 | Washington, DC

NCBI Resources

Literature

PubMed, PMC,
Books, PMH

~4,000,000 daily users
many authors

Browsing, reading
Small downloads

Molecular and Clinical Data

Gene, Nucleotide, Protein,Structure,
BLAST, PubChem, Clinical Trials

~300,000 daily users
many authors

Batch retrieval
Data analysis
Extensive NCBI curation
Big Data

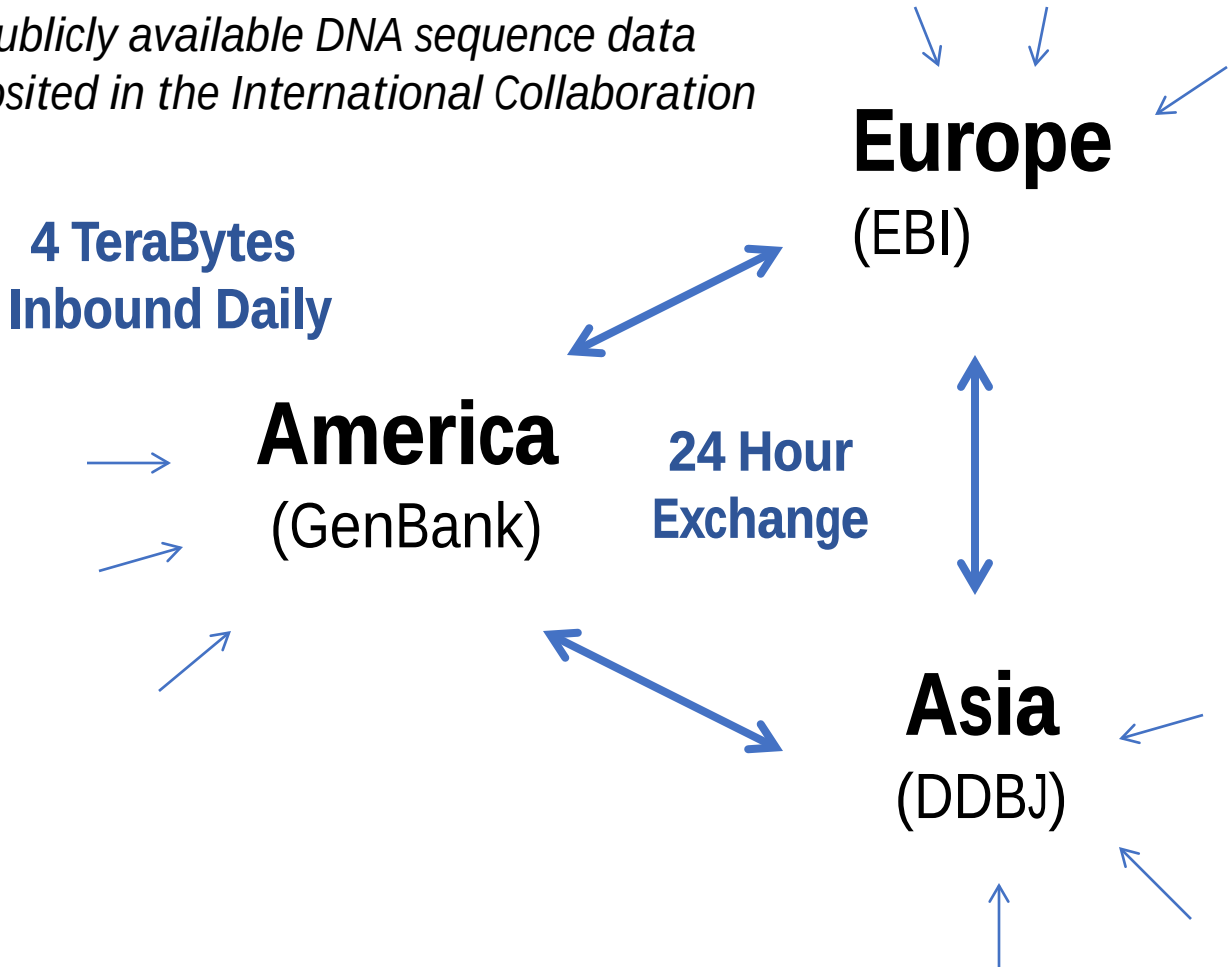
Specialized Datasets

SNP, GEO, SRA, GTR,
dbGaP, ClinVar, etc

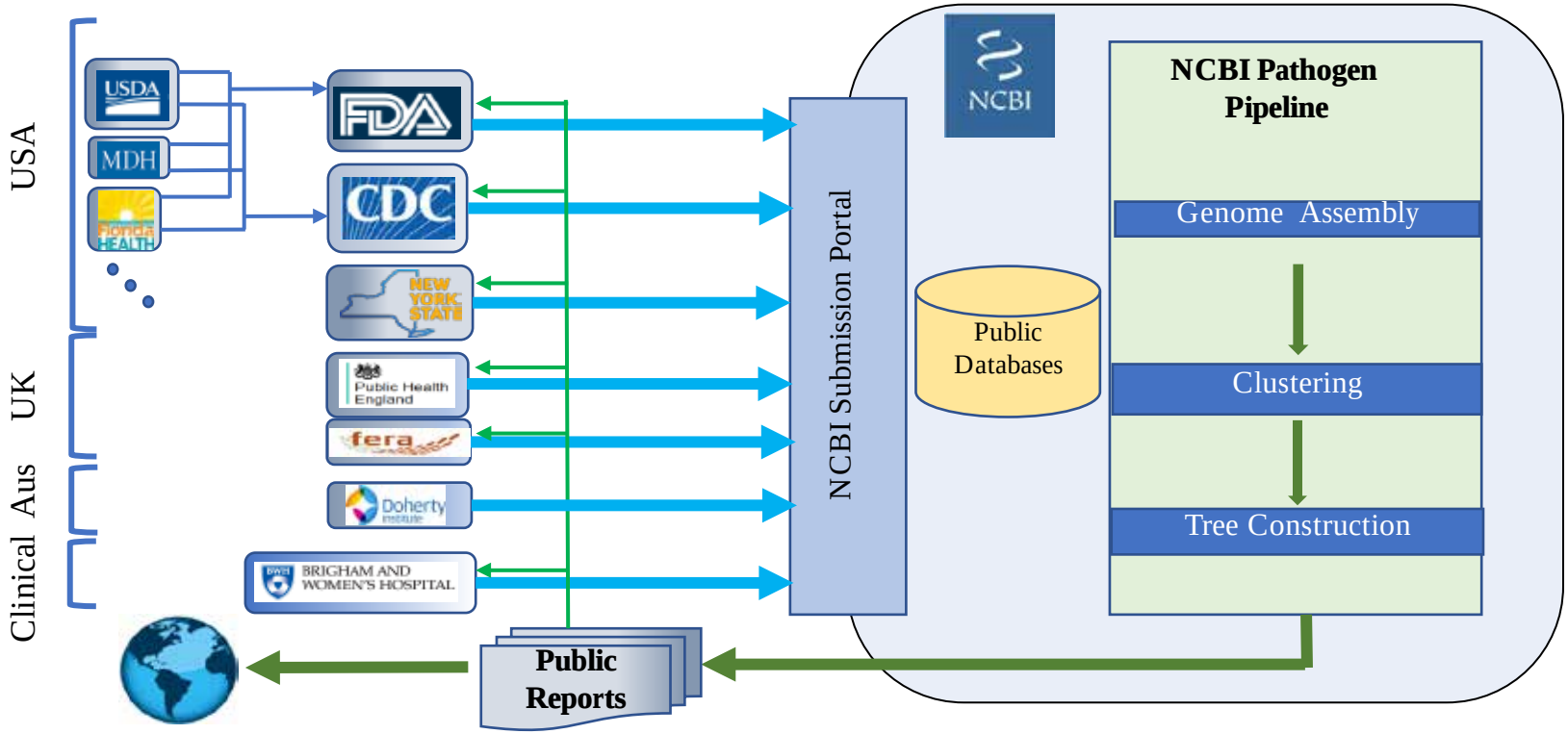
~15,000 daily users

Retrieving whole datasets
Expensive per user
Extensive Curation
Big Data

All publicly available DNA sequence data deposited in the International Collaboration



NCBI/NIH Pathogen Detection Pipeline:

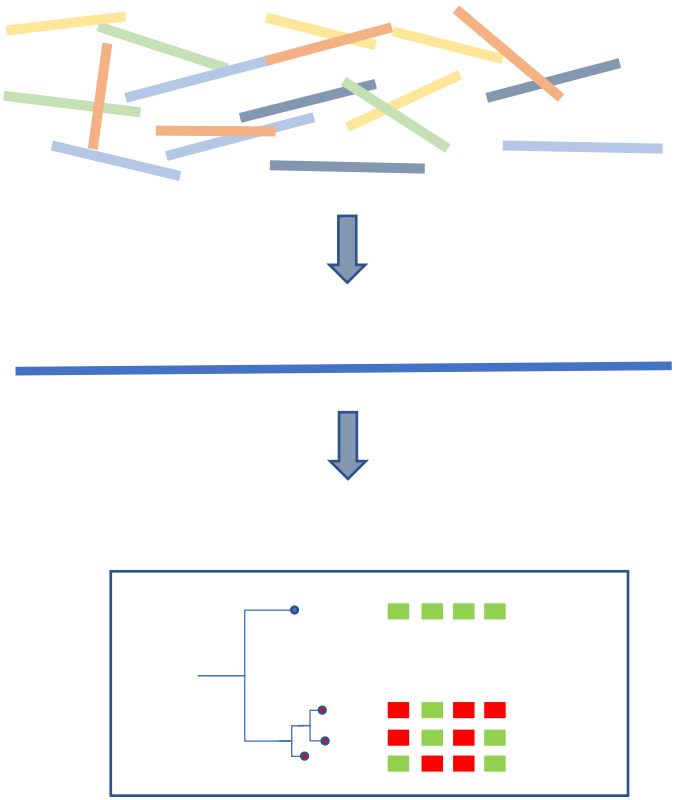


NCBI Biosample – Pathogen Template for Metadata

What sample_name organism strain/isolate Category (attribute_package) 1a) Clinical/Host-associated 1a1) specific_host 1a2) isolation_source 1a3) host-disease OR 1b) Environmental/Food/Other 1b1) isolation_source	**When** collection_date	
	Where Geographic location 6a) geo_loc_name OR 6b) lat_lon	
	Who collected by	minimal metadata

256 198 submissions using the above template
171 540 clinical/host-associated
84 658 food/environmental

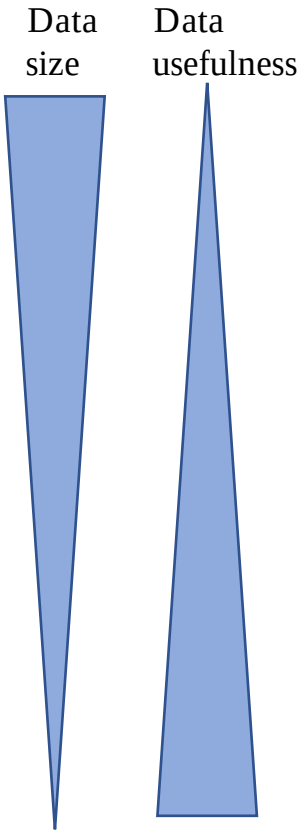
Data Reduction



Raw data: hundreds of millions of base pairs in short reads

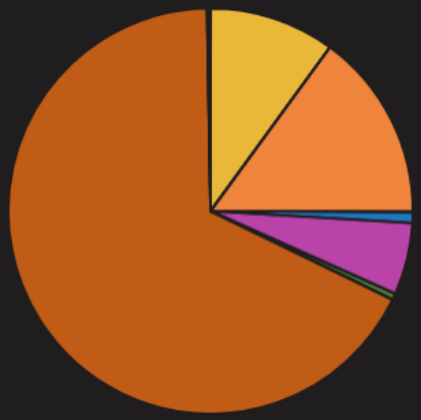
Intermediate data: assembly of a few million bases

Relevant data: a phylogenetic tree clustering isolates, SNP distances, annotated important (antimicrobial resistance) genes

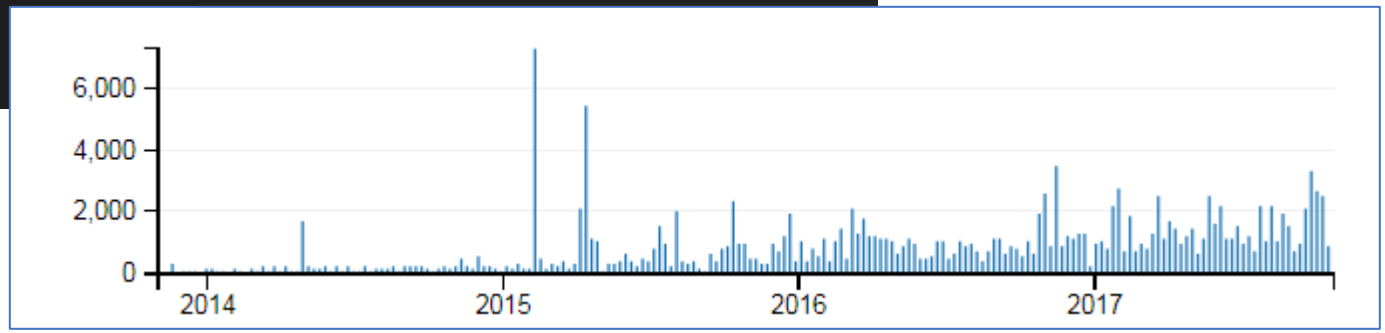


Turn large data into useful data: within 24 hours

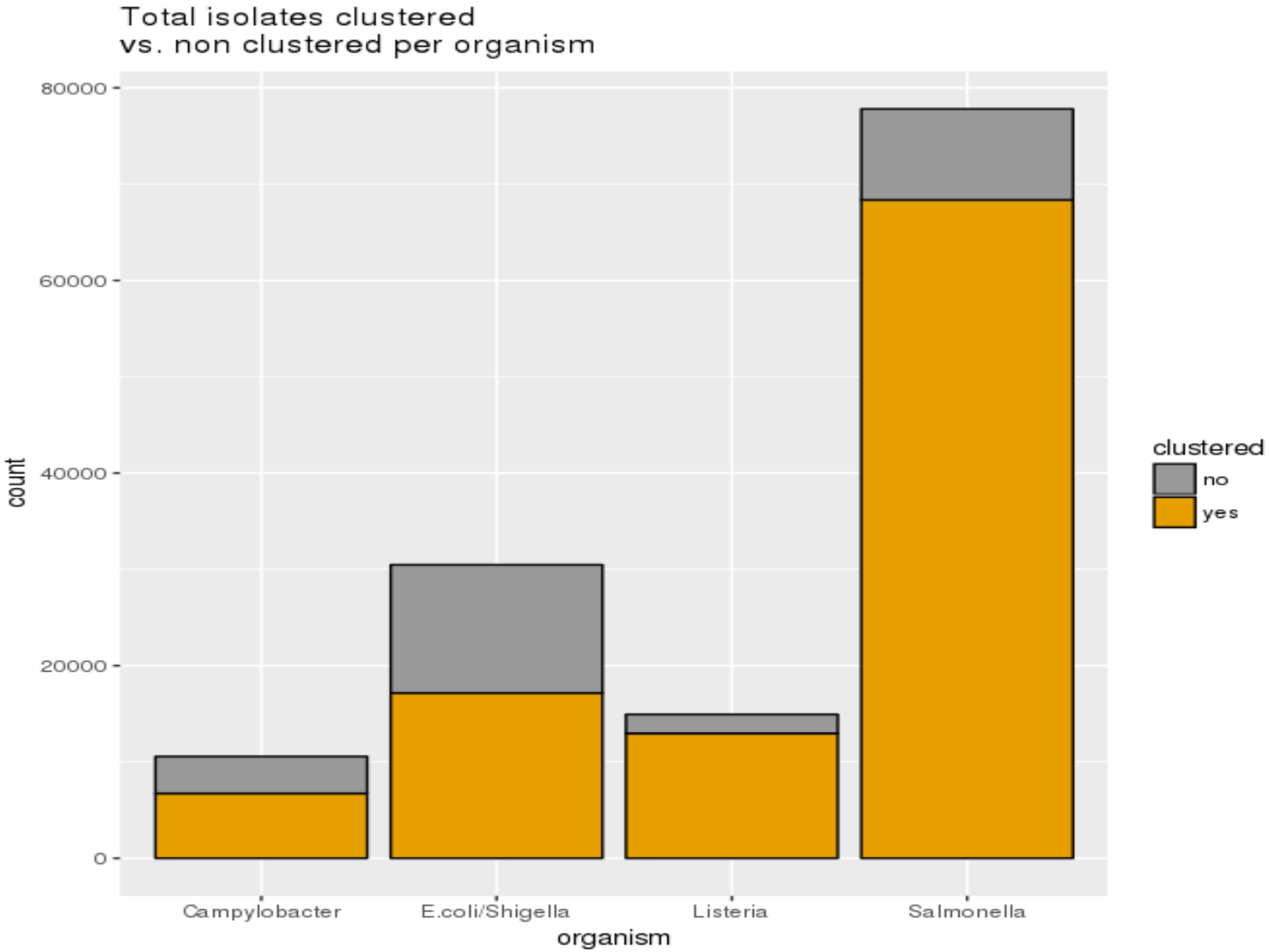
NCBI Pathogen Detection Processed Isolates - Last 6 Months (SRA to Assembly)



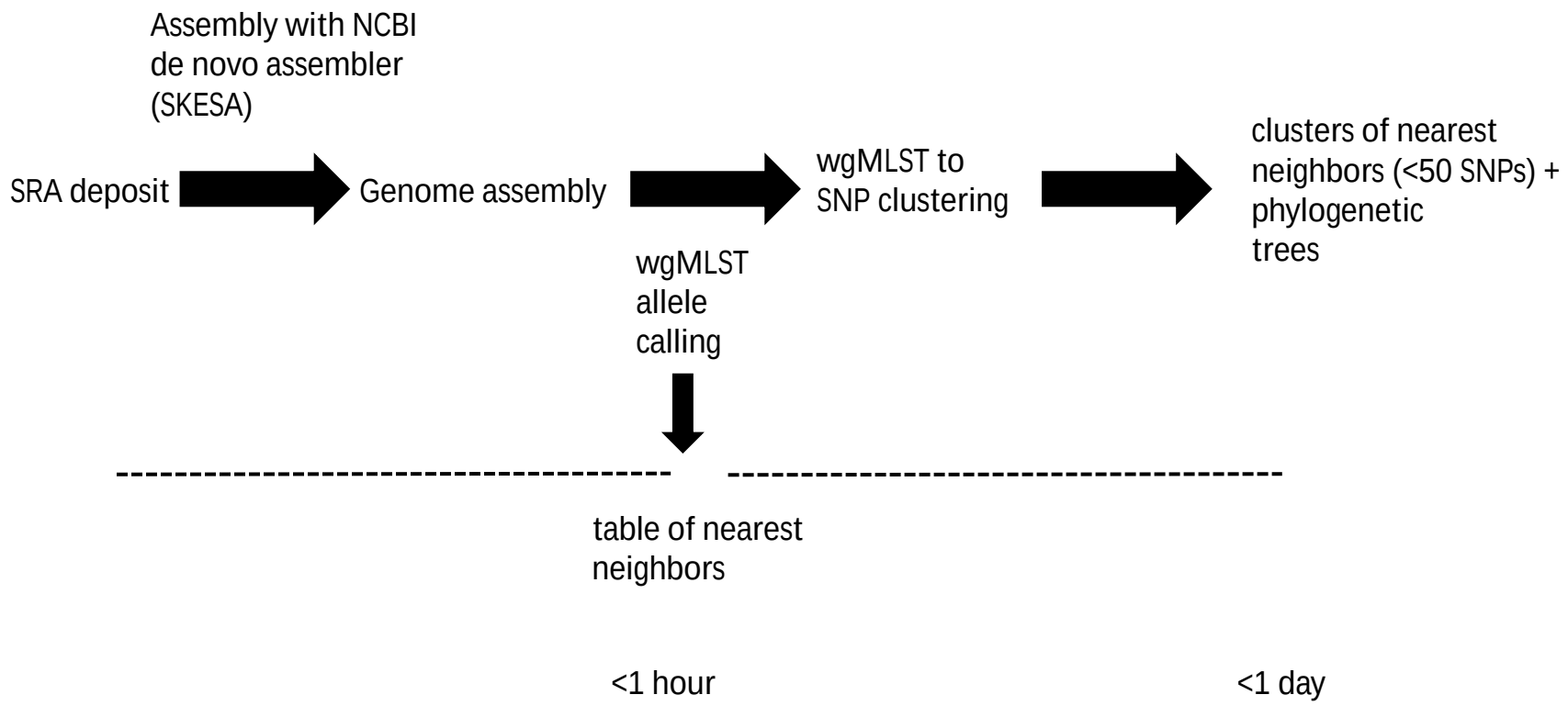
	values
Acinetobacter baumannii	6
Campylobacter jejuni	3.16 K
Citrobacter freundii	6
E.coli and Shigella	4.74 K
Klebsiella oxytoca	5.0
Klebsiella pneumoniae	270
Listeria monocytogenes	1.817 K
Morganella morganii	2
Neisseria	152
Providencia alcalifaciens	1.0
Pseudomonas aeruginosa	3
Salmonella enterica	21.3 K
Serratia marcescens	2



NCBI Pathogen Detection Total Isolates (Foodborne Bacteria)



New Proposed Pipeline based on wgMLST



Nearest neighbors (rapid reports)

Rapid reports are reported per day for a set of Bioprojects for Salmonella and Listeria

Report nearest 5 neighbors and all neighbors <6 allele differences

Report allele differences, loci in common and SNP accession (if exists)

Put in a tab-delimited file per run

ftp://ftp.ncbi.nlm.nih.gov/pathogen/Results/Listeria/Rapid_reports/
ftp://ftp.ncbi.nlm.nih.gov/pathogen/Results/Salmonella/Rapid_reports/

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	223	84,566
E.coli and Shigella	47	33,169
Listeria monocytogenes	16	15,349
Campylobacter jejuni	2	11,514
Mycobacterium tuberculosis	5	5,502
Klebsiella pneumoniae	1	4,313

Learn More

- About
- FAQ
- Factsheet
- Antimicrobial Resistance
- Contributors

Data Resources

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

NCBI Pathogen Detection Isolates Browser

More than 174 00 pathogen isolates

More than 142 000 with an acquired or chromosomal antibiotic resistance gene

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

Analysis of isolation_source for food/environmental samples

What sample_name organism strain/isolate Category (attribute_package) 1a) Clinical/Host-associated 1a1) specific_host 1a2) isolation_source 1a3) host-disease OR 1b) Environmental/Food/Other 1b1) isolation_source	**When** collection_date **Where** Geographic location 6a) geo_loc_name OR 6b) lat_lon **Who** collected by

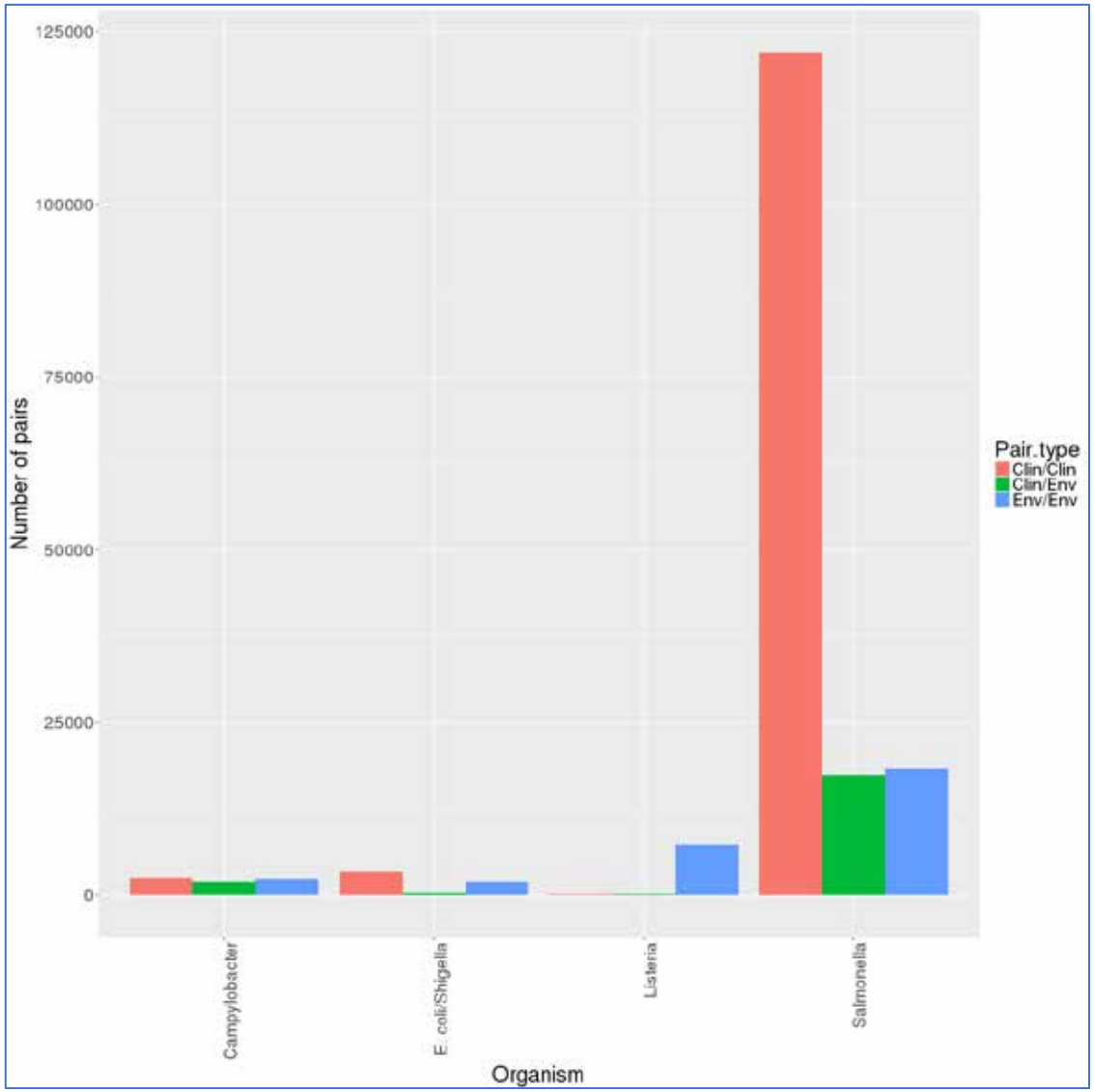
Describes the physical, environmental and/or local geographical source of the biological sample from which the sample was derived.

Analysis of isolation_source for food/environmental samples

- foodborne pathogen submissions in 2017 from CDC/FDA/USDA only
- all pairs of isolates within 10 SNP distances
- each pair is categorized based on sample metadata

1. clinical vs. clinical
2. **clinical vs. food/environmental <- smoking gun?**
3. food/environmental vs. food/environmental

Analysis of isolation_source for food/environmental samples



Analysis of isolation_source for food/environmental samples

Number of isolates					
	< 10 SNPs		All		>50 SNPs
Organism	Clinical	Environmental	Clinical	Environmental	Clinical
Salmonella	8951	4569	13964	6786	2116
Listeria	158	1385	432	1836	104
Campylobacter	514	804	973	1436	168
E. coli/Shigella	1109	538	2559	736	1549

Analysis of isolation_source for food/environmental samples

Salmonella

Source of top ~95% of food/environmental isolates in clinical vs. food/environmental pairs per within 10 SNPs distance

Source	Count of pairs
chicken	9855
beef/bovine	1375
pork/swine	949
water	749
turkey	614
missing	582
feces/stool	527
papaya	504
animal	286
swabs	258
poultry rinse	140
sediment	110
research strain	84
intestine	70
environmental sponge	54
horse/equine	54

Analysis of isolation_source for food/environmental samples

Listeria

Source of isolates in clinical vs. food/environmental for pairs less than 10 SNP distance

Source	Count of pairs
swabs	78
cheese	52
sandwich	12
avocado	15
deli meat	14
environmental	7
produce/salad	4
food	9

Analysis of isolation_source for food/environmental samples

E. coli and Shigella

Source of isolates in clinical vs. food/environmental for pairs within 10 SNP distance

Source	Count of pairs
soy nut butter	268
beef/bovine	29
milk	7
goat	4
flour	3
rabbit	1
stool/feces	5
horse/equine	22

Analysis of isolation_source for food/environmental samples

Campylobacter

Source of isolates in clinical vs. food/environmental for pairs within 10 SNP distance

Source	Count of pairs
milk	972
beef/bovine	712
chicken	238
sheep	16
offal	42
canine	9
missing	4
pig/swine	2
feces/stool	8

Analysis of isolation_source for food/environmental samples

List of Selected Outbreak Investigations, by Year

2017 | 2016 | 2015 | 2014 | 2013 | 2012 | 2011

- Unknown – *Cyclospora* (2017)
- Maradol Papayas – *Salmonella* Urbana
- Maradol Papayas – *Salmonella* Newport and *Salmonella* Infantis
- Maradol Papayas – *Salmonella* Anatum
- Maradol Papayas – *Salmonella* Infections
- Vulto Creamery Soft Raw Milk Cheese – *Listeria monocytogenes*
- I.M. Healthy SoyNut Butter – *E. coli* O157:H7

https://www.cdc.gov/foodsafety/outbreaks/multistate-outbreaks/outbreaks-list.html

Using NCBI Pathogen Detection Isolates Browser

Search for papaya

papaya

Search

Salmonella enterica PDS000003785.8 (23 out of 32)
Salmonella enterica PDS000002071.3 (13 out of 13)
Show all 56 clusters

Salmonella enterica

Filters | Expand All | Columns | Download

225 isolates

Page 1 of 12 | 20 | View 1 - 20 of 225

#	Organism Group	Strain	Serovar	Isolate	Create Date	Location	Isolation Sour	Isolation type	SNP cluster	Min-sam	Min-diff	BioSample	K-mer group
1	Salmonella enterica	FDA000121	enterica	PDT000237170.1	2017-08-28	Mexico	Fresh Papaya	environmental/othe	PDS000015865.4	0	0	SAMN07519491	PDG000000002.1029
2	Salmonella enterica	FDA000121	enterica	PDT000237254.1	2017-08-28	Mexico	Fresh Papaya	environmental/othe	PDS000015865.4	0	0	SAMN07519500	PDG000000002.1029
3	Salmonella enterica	FDA000121	enterica	PDT000245441.1	2017-09-27	Mexico	Fresh Papaya	environmental/othe	PDS000015865.4	0	0	SAMN07519496	PDG000000002.1029
4	Salmonella enterica	MD-DOH-18-000049-1		PDT000228393.1	2017-07-28	USA:MD	Maradol papaya	environmental/othe	PDS000015602.6	0	0	SAMN07374521	PDG000000002.1029
5	Salmonella enterica	FDA000120		PDT000233859.1	2017-08-16	Mexico	papaya	environmental/othe	PDS000015602.6	0	0	SAMN07429352	PDG000000002.1029
6	Salmonella enterica	FDA000120		PDT000233873.1	2017-08-								G000000002.1029
7	Salmonella enterica	VA-WGS-17087		PDT000236342.1	2017-08-								G000000002.1029

papaya

Search

Salmonella enterica PDS000003785.8 (23 out of 32)
Salmonella enterica PDS000002071.3 (13 out of 13)
Salmonella enterica PDS000015602.6 (10 out of 18)
Salmonella enterica PDS000001946.208 (8 out of 1121)
Salmonella enterica PDS000002518.184 (8 out of 710)
Salmonella enterica PDS000003806.110 (7 out of 136)
Salmonella enterica PDS000004194.14 (7 out of 19)

multiple isolates in multiple clusters

Using NCBI Pathogen Detection Isolates Browser

8/710 from papaya search

Salmonella - 710 isolates
PDG000000002.1029 / PDS000002518.184

8 Isolates Selected Clear

Distance between selected isolates :
minimum=0 SNPs, maximum= 103 SNPs, average=51 SNPs

Target creation date range among selected isolates :
2015-02-11 to 2017-05-02

Filter selected isolates

2017 environmental/other | Mexico 2017-05-02

2016 environmental/other | Mexico 2016-03-06
environmental/other | Mexico 2016-03-06
environmental/other | Mexico 2016-03-06

2015 environmental/other | Mexico 2015-02-11
environmental/other | Mexico 2015-02-11
environmental/other | Mexico 2015-02-11
environmental/other | Mexico 2015-02-11

Filters | Columns

Page 1 of 142 5

#	Strain	Serovar	Isolate ID	Create Date	Location	Isolation Sou	Isolation type	Host	Min-sa	Min-dif	BioSample
1	FDA00004	enterica / Agona	PDT000046366.3	2016-03-06	Mexico	papaya	environmental/oth		0	20	SAMN02847490
2	FDA00004	enterica / Agona	PDT000112603.1	2016-03-06	Mexico	papaya	environmental/oth		0	23	SAMN02847438
3	FDA00004	enterica / Agona	PDT000112604.1	2016-03-06	Mexico	papaya	environmental/oth		0	23	SAMN02847437
4	FAR0034	enterica / Agona	PDT000003081.2	2015-02-11	Mexico	papaya	environmental/oth		16	15	SAMN02344910
5	FAR0035	enterica / Agona	PDT000003237.2	2015-02-11	Mexico	papaya	environmental/oth		13	12	SAMN02344911

Choose Columns

only one is from 2017

Spacing | DIM UNSELECTED OFF ON | TREE TIPS OFF ON

Search & Highlight in Tree

only food isolate from 2017 is not near any clinical isolates

Using NCBI Pathogen Detection Isolates Browser

8/1121 from papaya search

three are from 2017

one food isolate from 2017 is extremely close to multiple clinical isolates

Salmonella - 1121 isolates
PDG000000002.1029 / PDS000001946.208

8 Isolates Selected Clear

Distance between selected isolates :
minimum=0 SNPs, maximum= 72 SNPs, average=48 SNPs

Target creation date range among selected isolates :
2014-01-06 to 2017-08-21

Filter selected isolates

2017
- environmental/other | Mexico 2017-08-16
- environmental/other | USA:MD 2017-07-26
- environmental/other | USA:VA 2017-08-21

2016
- environmental/other | Mexico 2016-03-06
- environmental/other | Mexico 2016-03-02

2015
- environmental/other | Mexico 2015-02-11

2014
- environmental/other | Mexico 2014-02-26
- environmental/other | Mexico 2014-01-06

Filters | Columns

Page 1 of 225 | 5

#	Strain	Serovar	Isolate ID	Create Date	Location	Isolation Sou	Isolation type	Host
1	VA-WGS-17091		PDT000234942.1	2017-08-21	USA:VA	papaya	environmental/oth	
2	FDN0010		PDT000002410.1	2014-01-06	Mexico	papaya	environmental/oth	
3	FDN0017	enterica / Thompson	PDT000002482.2	2015-02-11	Mexico	papaya	environmental/oth	
4	FMA0049		PDT000003555.1	2014-02-26	Mexico	papaya	environmental/oth	
5	FDA000040	enterica / Thompson	PDT000111537.1	2016-03-02	Mexico	papaya	environmental/oth	

Choose Columns | Page 1 of 225 | 5

Spacing | DIM UNSELECTED OFF ON | TREE TIPS OFF ON

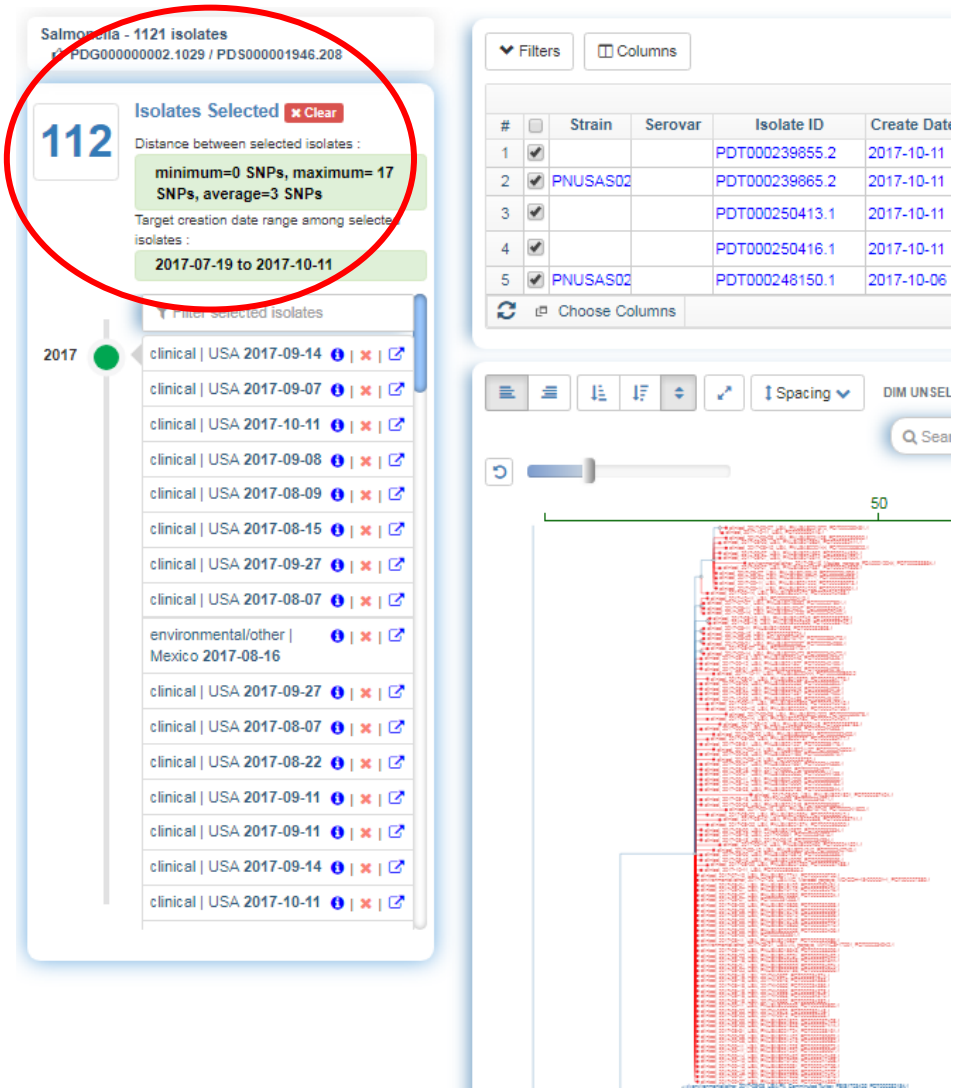
Search & Highlight in Tree

10

clinical, 2017-08-
clinical, 2017-08-
clinical, 2017-08-09, USA, PNUSAS019855, PDT000232410.1
clinical, 2017-08-09, USA, PDT000232361.1
clinical, 2017-08-11, USA, PNUSAS019857, PDT000232986.1
environmental/other, 2017-08-21, USA:VA, papaya, VA-WGS-17091, PDT000234942.1
clinical, 2017-08-14, USA, PNUSAS018643, PDT000233526.1
clinical, 2017-08-22, USA, PNUSAS019751, PDT000235462.1
clinical, 2017-08-21, USA, PNUSAS020925, PDT000234974.1
clinical, 2017-08-22, USA, PNUSAS020736, PDT000235509.1
clinical, 2017-08-18, USA, 2017K-0657, PDT000234374.1
clinical, 2017-08-18, USA, 2017K-0616, PDT000234383.1
clinical, 2017-08-18, USA, 2017K-0655, PDT000234386.1
clinical, 2017-08-18, USA, 2017K-0666, PDT000234378.1
clinical, 2017-08-18, USA, 2017K-0663, PDT000234379.1
clinical, 2017-08-18, USA, 2017K-0659, PDT000234382.1

Using NCBI Pathogen Detection Isolates Browser

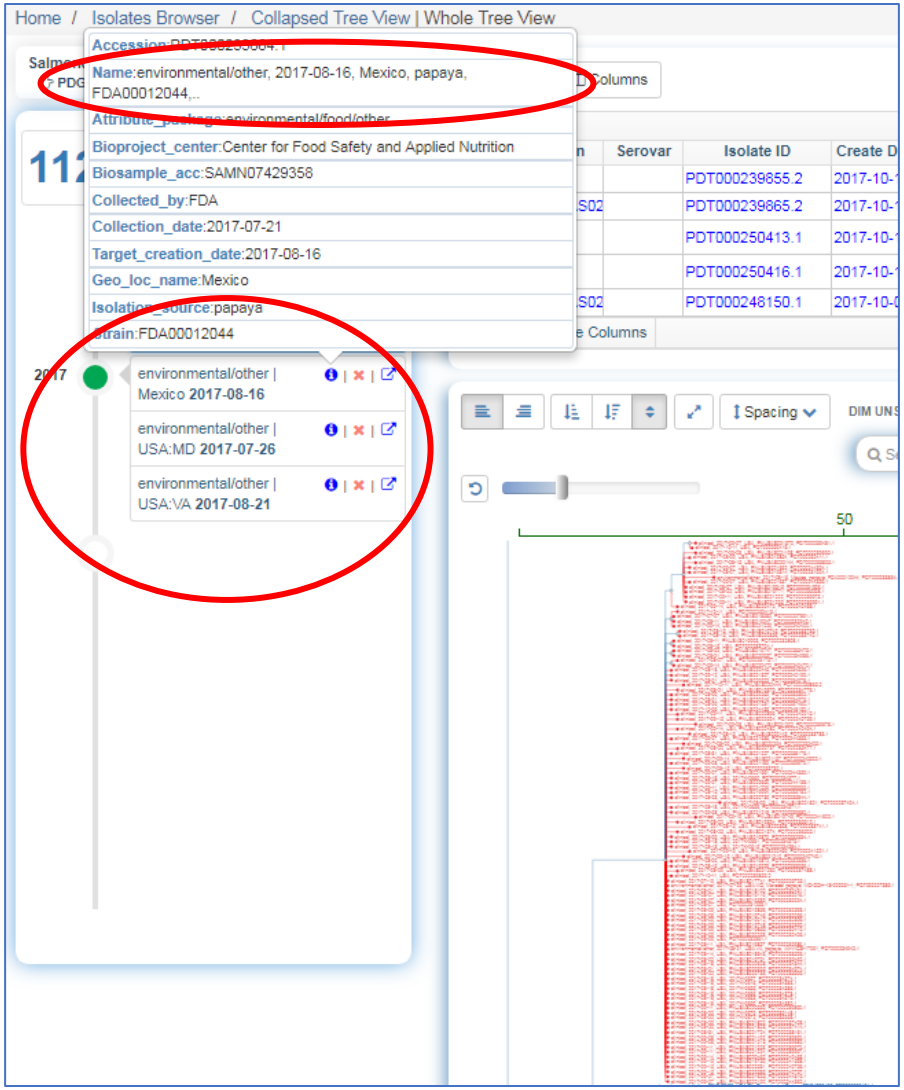
112 isolates in total in clonal subtree
max = 17 SNPs
avg = 3 SNPs



Using NCBI Pathogen Detection Isolates Browser

subtree is filtered for food/environmental samples

one is papya from Mexico in August, 2017



Unlinked isolates

Unlinked clinical isolates (not within 50 SNPs of any food/environmental isolate)

Not just singletons, some are in exceptionally large clusters

Is a new sampling strategy required? Are new food vehicles waiting to be discovered? What does the epidemiology tell us?

Number of isolates					
	< 10 SNPs		All		>50 SNPs
Organism	Clinical	Environmental	Clinical	Environmental	Clinical
Salmonella	8951	4569	13964	6786	2116
Listeria	158	1385	432	1836	104
Campylobacter	514	804	973	1436	168
E. coli/Shigella	1109	538	2559	736	1549

Unlinked isolates

Example:

Cluster of Salmonella Enteritidis
506 isolates
All clinical

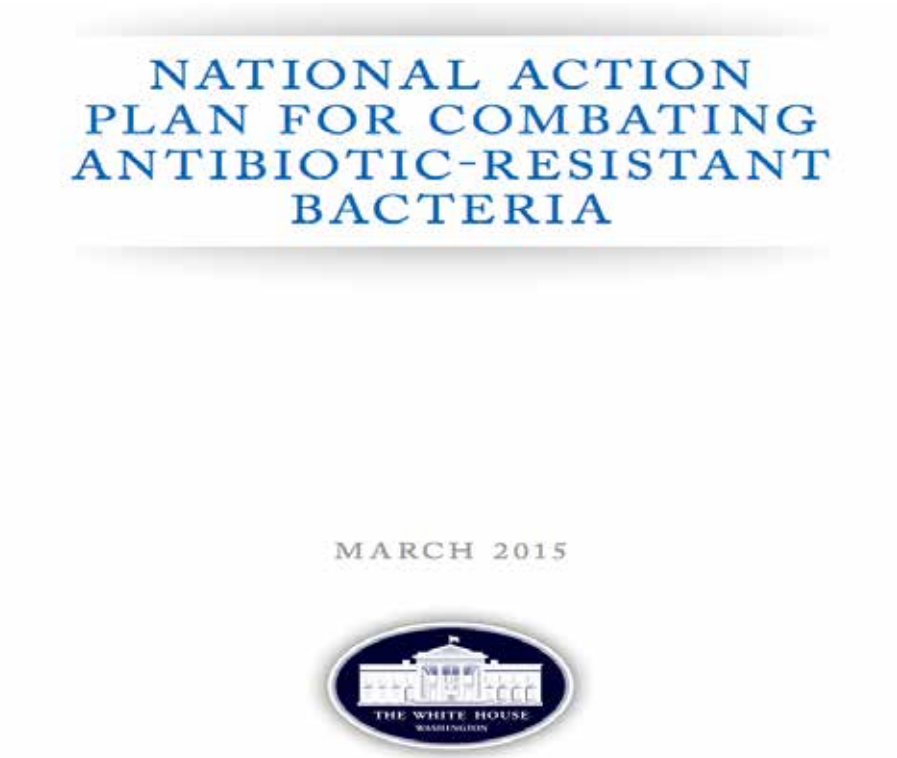
No isolates directly from food/environmental sources

Collected 2013-2017

From CDC, PHE, state labs

What is the cause?

NCBI's Role in Combatting Antibiotic Resistance



"Create a repository of resistant bacterial strains (an "isolate bank") and <u>maintain a well-curated reference database</u> that describes the characteristics of these strains."

"Develop and maintain a national sequence database of resistant pathogens."

NCBI's Role in Combatting Antibiotic Resistance

The Practical Results of AMR efforts at NCBI

With collaborators, build database of sequenced isolates with standardized AST metadata (i.e. accepting submissions of antibiograms) (4928 Samples as of Oct)
https://www.ncbi.nlm.nih.gov/biosample/?term=antibiogram[filter])

Collaborators include: (CDC, WRAIR, FDA, B&W, Broad)

Stable, up-to-date database of AMR genes with standardized nomenclature

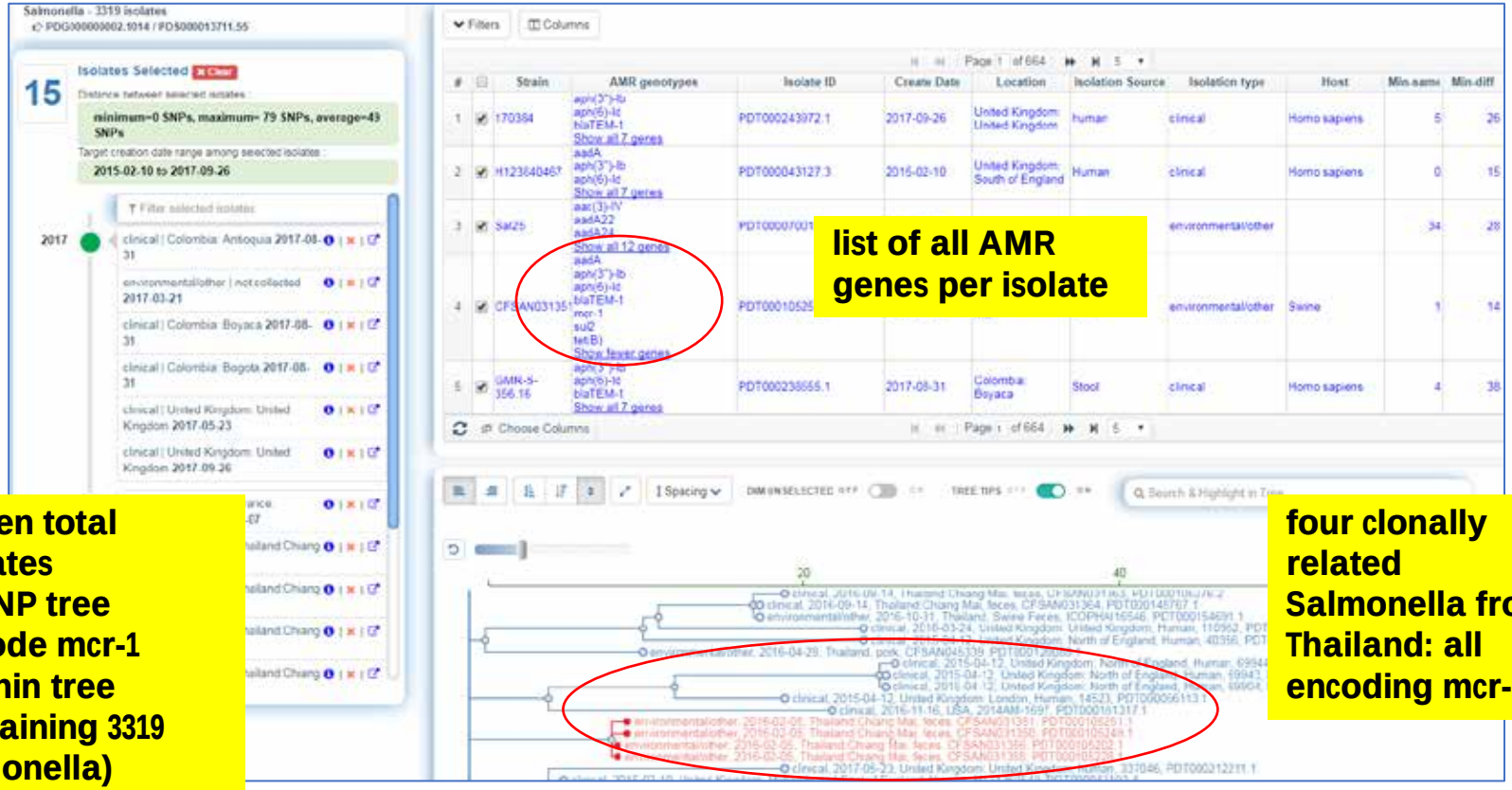
Collaborators (CARD, Lahey, experts)

RefSeq set released in June 2016 (4128 genes/proteins as of Oct)

https://www.ncbi.nlm.nih.gov/bioproject/PRJNA313047

Implement and validate tools for identifying AMR genes in new isolates

Antimicrobial genotypes and phenotypes added to Pathogen Detection interface



Summary

NCBI is enhancing existing analytical pipelines to improve turnaround time to answer two fundamental questions:

1. Are these isolates clonally related?

2. Is there a point source for clinical illnesses?

NCBI is improving publicly accessible web interfaces to the above information and enhancing the information layered onto the phylogenetic trees and list of ioslates

1. To quickly make inclusion/exclusion determinations to aid those responsible for food safety

2. To quickly see the list of genetic determinants that encode antimicrobial resistance

3. In the future other genes responsible for virulence, heavy metal resistance, point mutations, mobile elements

Acknowledgements

Richa Agarwala
Azat Badretdin
Slava Brover
Joshua Cherry
Jinna Choi
Vyacheslav Chetvernin
Robert Cohen
Michael DiCuccio
Boris Fedorov
Mike Feldgarden
Lewis Geer
Dan Haft
Lianyi Han
Avi Kimchi
Michel Kimelman
William Klimke
Alex Kotliarov
Valerii Lashmanov
Aleksandr Morgulis
Eyal Moses
Chris O'Sullivan
Arjun Prasad

Edward Rice
Kirill Rotmistrovskyy
Alejandro A. Schaffer
Stephen Sherry
Sergey Shiryev
Martin Shumway
Oleg Shutov
Alexandre Souvorov
Tatiana Tatusova
Igor Tolstoy
Chunlin Xiao
Leonid Zaslavsky
Alexander Zasypkin
Lukas Wagner
Hlavina Wratko
Eugene Yaschenko

David Lipman
James Ostell
Kim Pruitt

CDC
FDA/CFSAN
GenFS
USDA-FSIS
PHE/FERA
NIHGRI
NIAID
WRAIR
Broad
Wadsworth/MDH
Vendors: PacBio, Illumina, Roche

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