

## Session 3.2 – SNP matrix and phylogenetics

Sara Monzón Fernández

BU-ISCIII

Unidades Comunes Científico Técnicas – SGSAFI-ISCIII

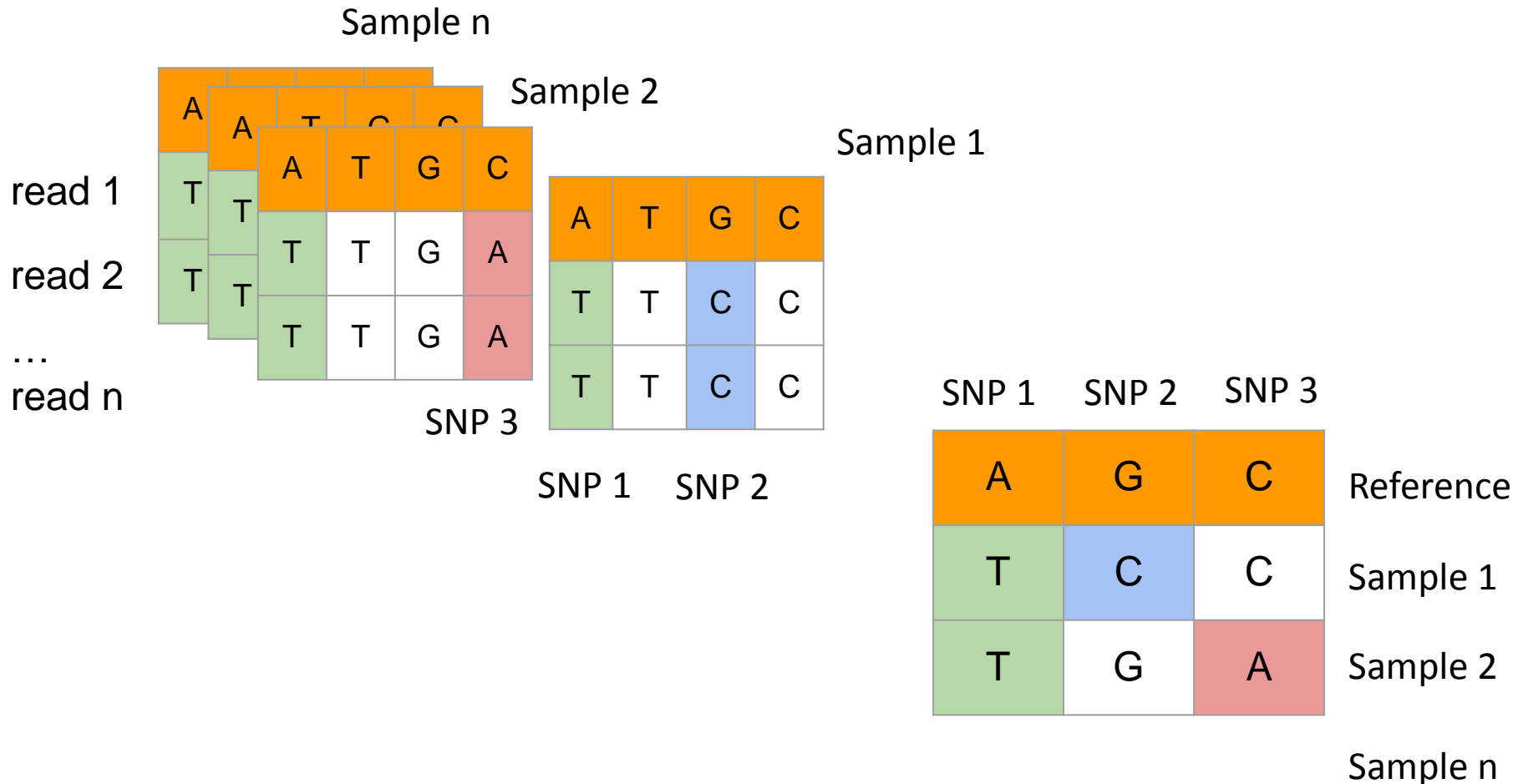
05-09 Noviembre 2018, 1ª Edición  
Programa Formación Continua, ISCIII

# Index

## SNP matrix and Phylogenetics:

- SNP Matrix build
- Phylogeny
  - Maximum Likelihood
  - Parsimony
- WGS-Outbreaker
- iTOL phylogenetic tree visualization.
- How to interpret SNP-based outbreak analysis
- Examples: SnapperDB and GenomeTrakr

# Building a SNP matrix



# Building a SNP matrix

- Once we have our multisample vcf:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	RA-L2073	RA-2805
NC_021827.1	276	.	C	A	291.68	PASS	AC=1;...	GT:AD:DP:..	0:13,0:13:..	1:0,50:30
NC_021827.1	731	.	A	G	2313.68	PASS	AC=1;...	GT:AD:DP:..	0:23,0:23:..	1:0,10:10
NC_021827.1	921	.	C	T	1841.68	PASS	AC=1;...	GT:AD:DP:..	0:53,0:53:..	0:20,0:20

- We can generate the genotype for each sample

#CHROM	POS	RA-L2073	RA-2805
NC_021827.1	276	C	A
NC_021827.1	731	A	G
NC_021827.1	921	C	C

# Building a SNP matrix

- So... now we have a simple multifasta, where each nucleotide represents a SNP.
- This means that even the nucleotide positions are sequentially in the fasta, they don't have to be near each other in the genome!
- The SNP matrix file will look like this:

> RA-2073

CACGAAATTCCATTA



>RA-2805

AGCTCATGCATATTA



Each of this is a SNP:

First one is in position 276 in the genome

Second one is in position 731 in the genome

Third one is in position 921 in the genome

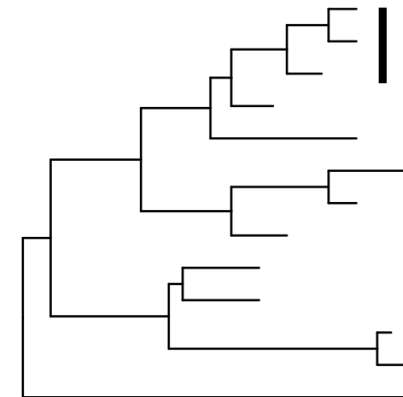
Each SNP is ordered per SNP position. In this sample also first SNP is in position 276 in the genome

# Phylogeny

SNP matrix

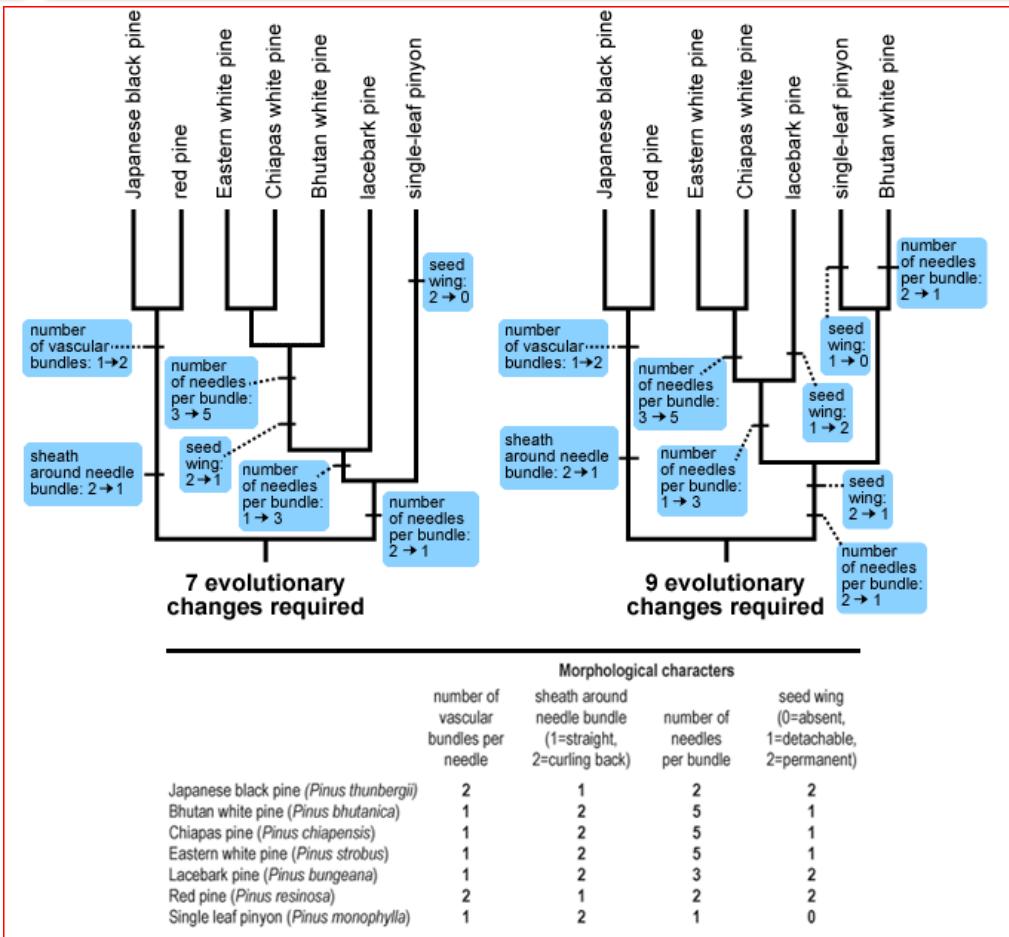
SNP 1	SNP 2	SNP 3	
A	G	C	Reference
T	C	C	Sample 1
T	G	A	Sample 2
			Sample n

Phylogeny  
→



Outbreak!

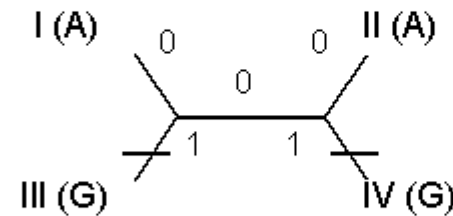
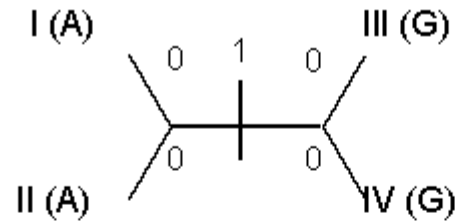
# Maximum parsimony



- Search the most parsimonious tree
- The most simple hypothesis must be the correct.
- Search the tree that explains the relationships with the less changes possible.

# Maximum Likelihood

- Searches the most likely tree given the data and based in a evolutionary model.
- More sophisticated.
- Not prepared a priori for snp matrix.
- RAxML
  - Heterogeneity rate disabled.
  - Branchs indicate the expected number of substitution per site.



- 0,1 = differences along that branch
- Which hypothesis is more likely, given that the change is rare?

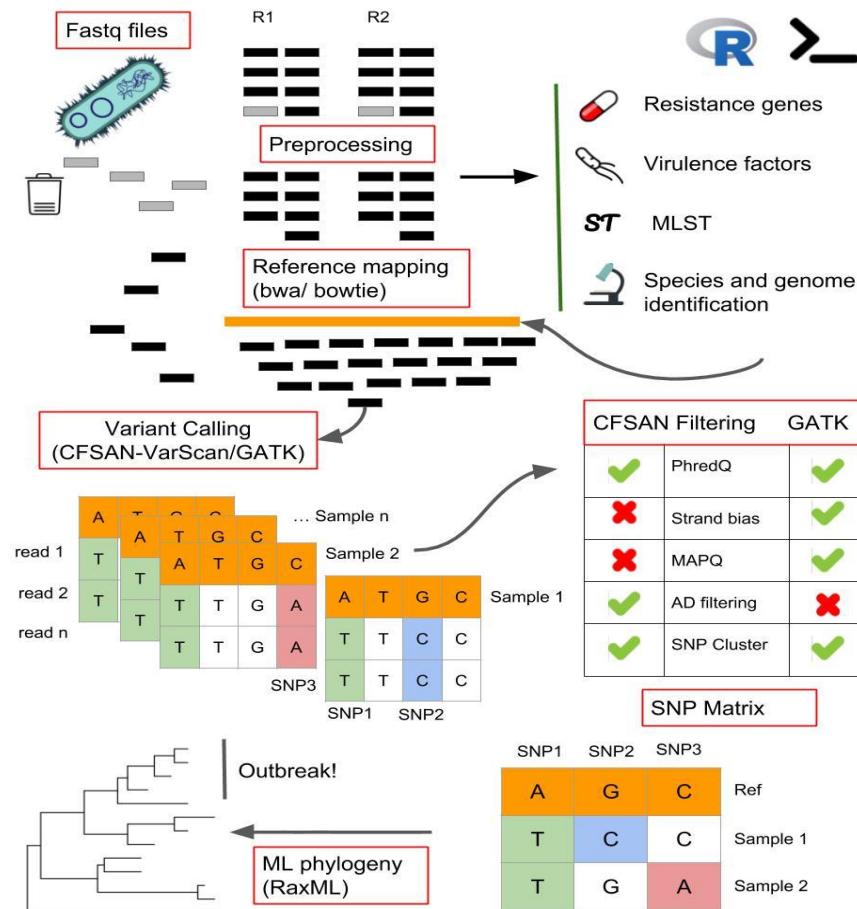


# SNP distance

- SNP distance is calculated with N model. Simply the number of sites that differ between each pair of sequences.
- By default sites with at least one missing data is deleted for all sequences in R (complete deletion option in MEGA).

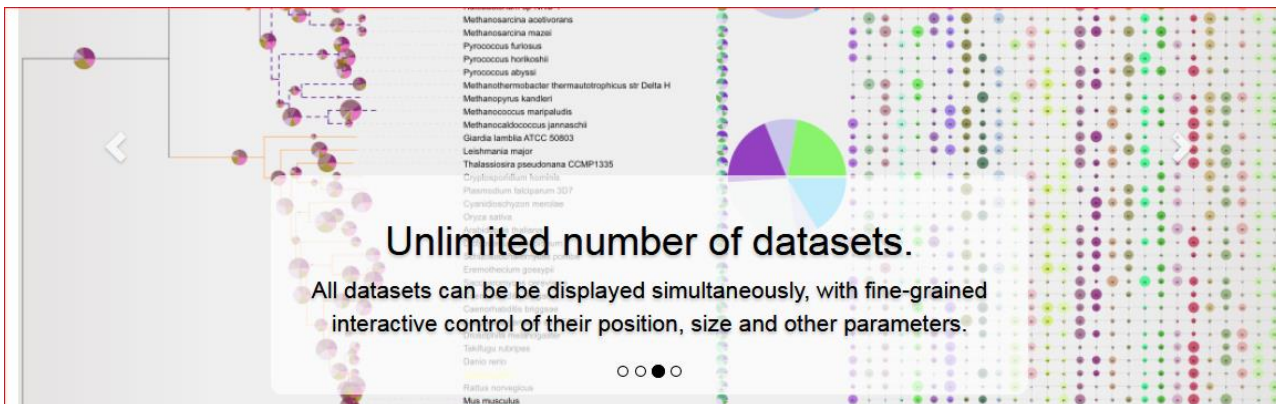
dis_matrix.names	RA.L2073	RA.L2281	RA.L2327	RA.L2391	RA.L2450	RA.L2677	RA.L2701
RA-L2073	0	9403	9028	80	46	46	49
RA-L2281	9403	0	8777	9415	9397	9397	9402
RA-L2327	9028	8777	0	9040	9022	9022	9027
RA-L2391	80	9415	9040	0	74	74	79
RA-L2450	46	9397	9022	74	0	38	45
RA-L2677	46	9397	9022	74	38	0	45
RA-L2701	49	9402	9027	79	45	45	0
RA-L2782	9120	9183	4277	9132	9114	9114	9119
RA-L2805	4	9403	9028	80	46	46	49
RA-L2978	2	9401	9026	78	44	44	47

# WGS - Outbreaker



# iTOL

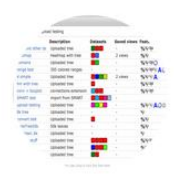
<https://itol.embl.de/>



**Unlimited number of datasets.**

All datasets can be displayed simultaneously, with fine-grained interactive control of their position, size and other parameters.

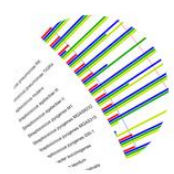
Current changelog: version 4.2.3



### Manage

Organize your trees into workspaces and projects, and access them from any browser. Simply drag and drop multiple tree files onto a project to upload them all at once.

Create an account »




### Annotate

18 dataset types. Full control over branch colors, widths and styles. Individually adjustable label fonts, sizes and styles.

Gallery of user created trees

Upload a tree »



### Export

Create high quality vector or bitmap figures for your publications. Direct WYSIWYG export of what is displayed on the screen.

Explore help »

# How to interpret our phylogeny

- Combination of:
  - SNP counts
  - Tree topologies
  - Bootstrap support

Pightling et al. Frontiers in Microbiology. 2018

# How to interpret our phylogeny

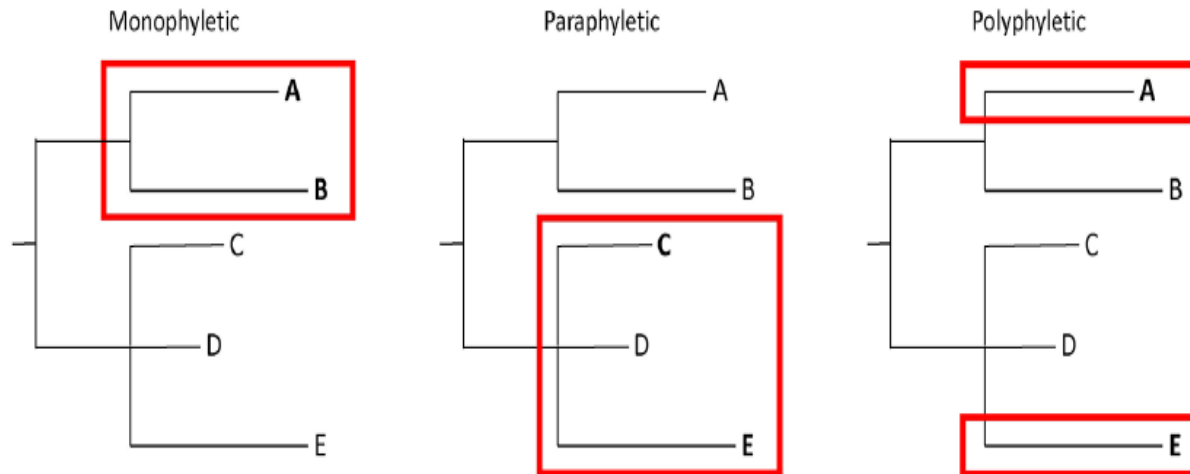
**TABLE 1** | Maximum pairwise SNPs measured during investigations into foodborne illness outbreaks and contamination events.

Organism	Maximum SNP count (number)	Maximum SNP count (range)			Reference
		<21	21–100	> 100	
<i>E. coli</i>	4	X			Underwood et al., 2013
<i>E. coli</i>	15	X			Eppinger et al., 2011
<i>L. monocytogenes</i>	9	X			Chen et al., 2017c
<i>L. monocytogenes</i>	12	X			Chen et al., 2017a
<i>L. monocytogenes</i>	18	X			Li et al., 2017
<i>L. monocytogenes</i>	20	X			Wang et al., 2015
<i>L. monocytogenes</i>	21		X		Nielsen et al., 2017
<i>L. monocytogenes</i>	28		X		Gilmour et al., 2010
<i>L. monocytogenes</i>	29		X		Chen et al., 2017b
<i>L. monocytogenes</i>	42		X		Chen et al., 2016
<i>L. monocytogenes</i>	67		X		Jackson et al., 2016
<i>S. enterica</i>	2	X			Wuyts et al., 2015
<i>S. enterica</i>	3	X			Allard et al., 2016
<i>S. enterica</i>	3	X			Taylor et al., 2015
<i>S. enterica</i>	6	X			Hoffmann et al., 2016
<i>S. enterica</i>	12	X			Octavia et al., 2015
<i>S. enterica</i>	30		X		Leekitcharoenphon et al., 2014

The maximum SNP counts for isolates that were traced back to the same source in the original study are presented. Whether the maximum SNP counts are less than 21 SNPs, 21 to 100 SNP, or greater than 100 SNPs is also indicated.

Pightling et al. Frontiers in Microbiology. 2018

# How to interpret our phylogeny



**FIGURE 2 |** Illustration of monophyletic, paraphyletic, and polyphyletic groupings. A monophyletic topology exists when isolates of interest (e.g., A and B) group together to the exclusion of all others. A paraphyletic topology is one in which isolates of interest (e.g., C and E) group together but not to the exclusion of all others (e.g., D). A polyphyletic topology exists when isolates of interest do not form a group (e.g., A and E).

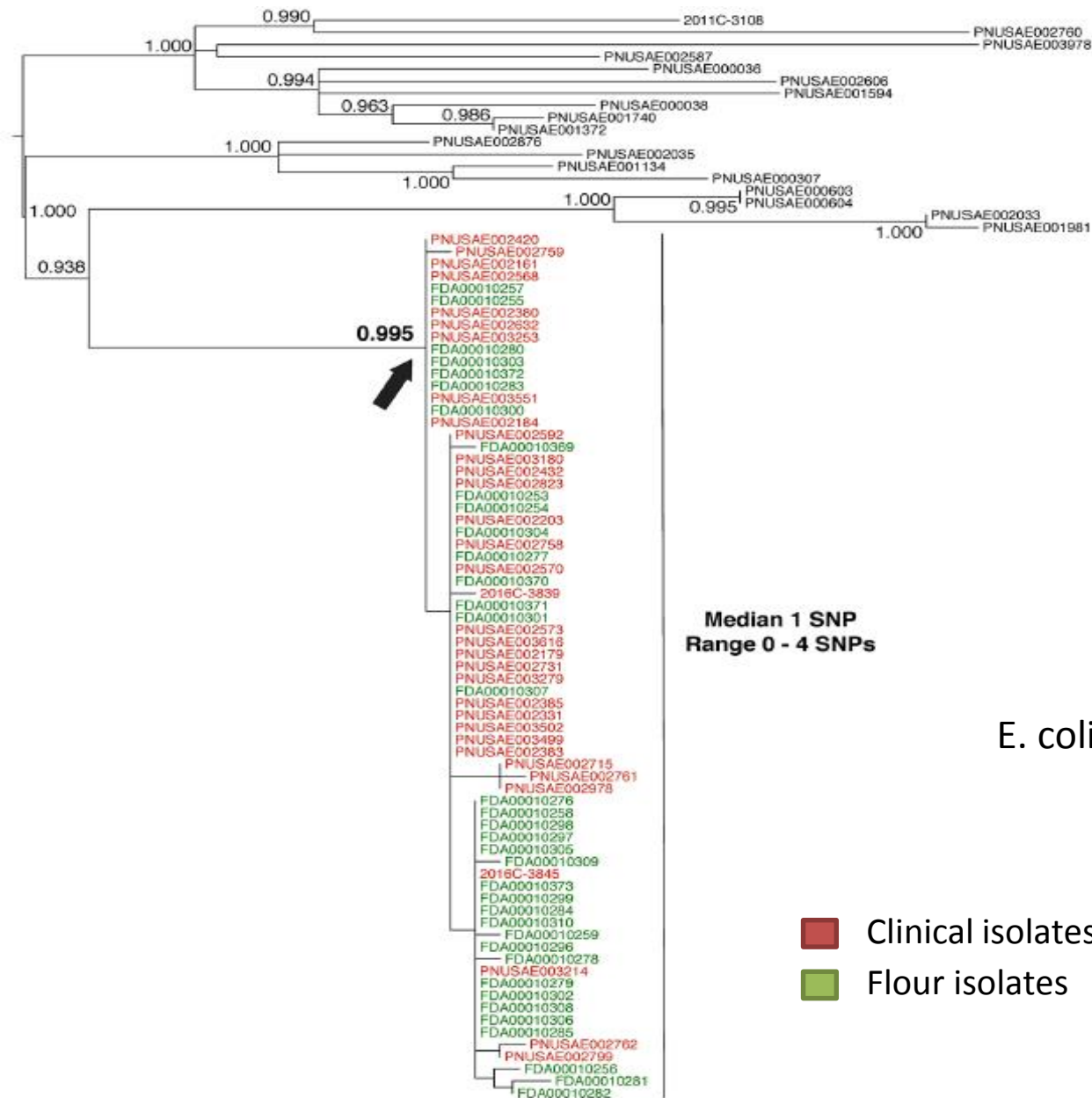
Pightling et al. Frontiers in Microbiology. 2018

# How to interpret our phylogeny

**TABLE 2** | Conditions used to determine whether whole-genome sequence analyses support a match between two or more genomes.

	Supports	Neutral	Does not support
SNP distance	<21	21–100	>100
Bootstrap support	>0.89	0.80–0.89	<0.80
Tree topology	Monophyletic	Paraphyletic	Polyphyletic

Pightling et al. Frontiers in Microbiology. 2018



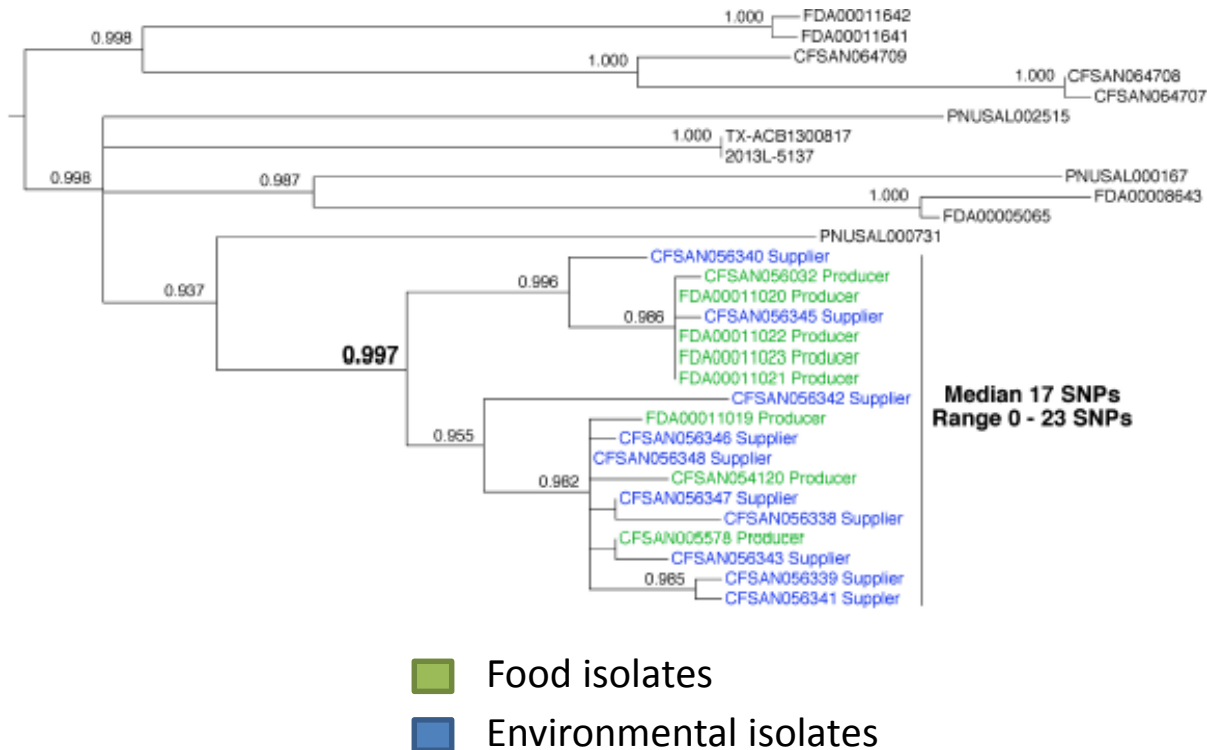
- Bootstrap support
- SNP count support
- Topology support
- Epidemiology support

*E. coli* Clinical isolates - source

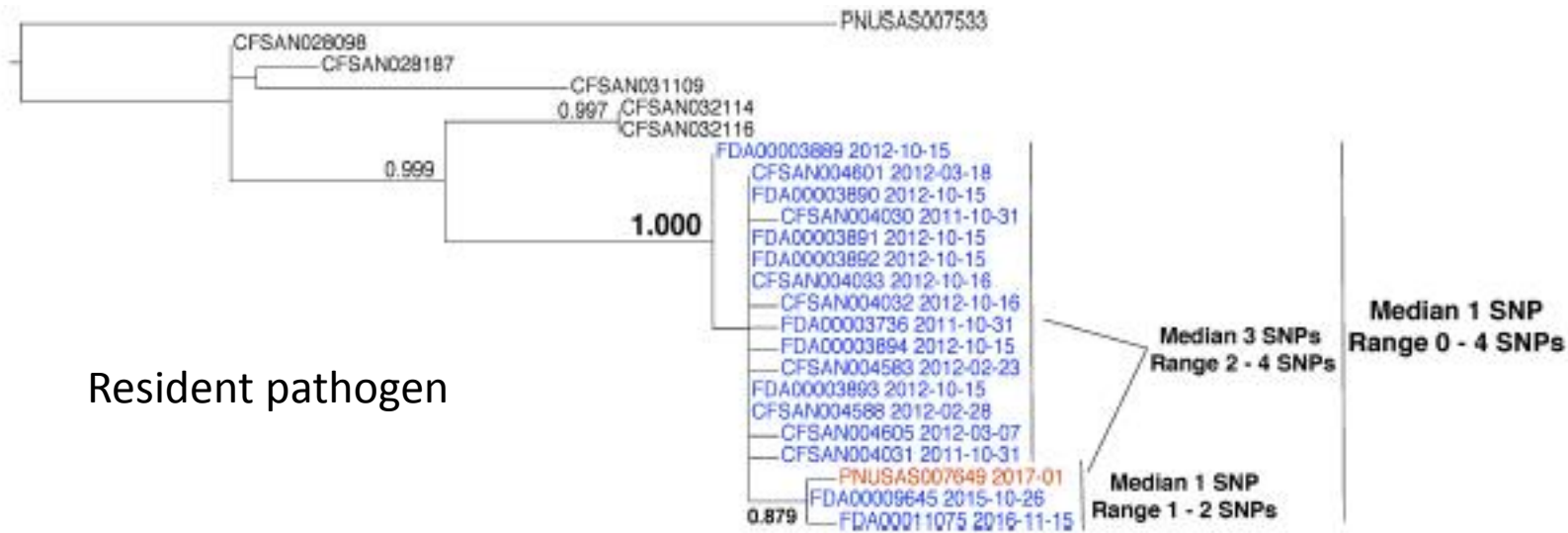
- Clinical isolates
- Flour isolates



## L. monocytogenes ingredient supplier – Ice cream producer

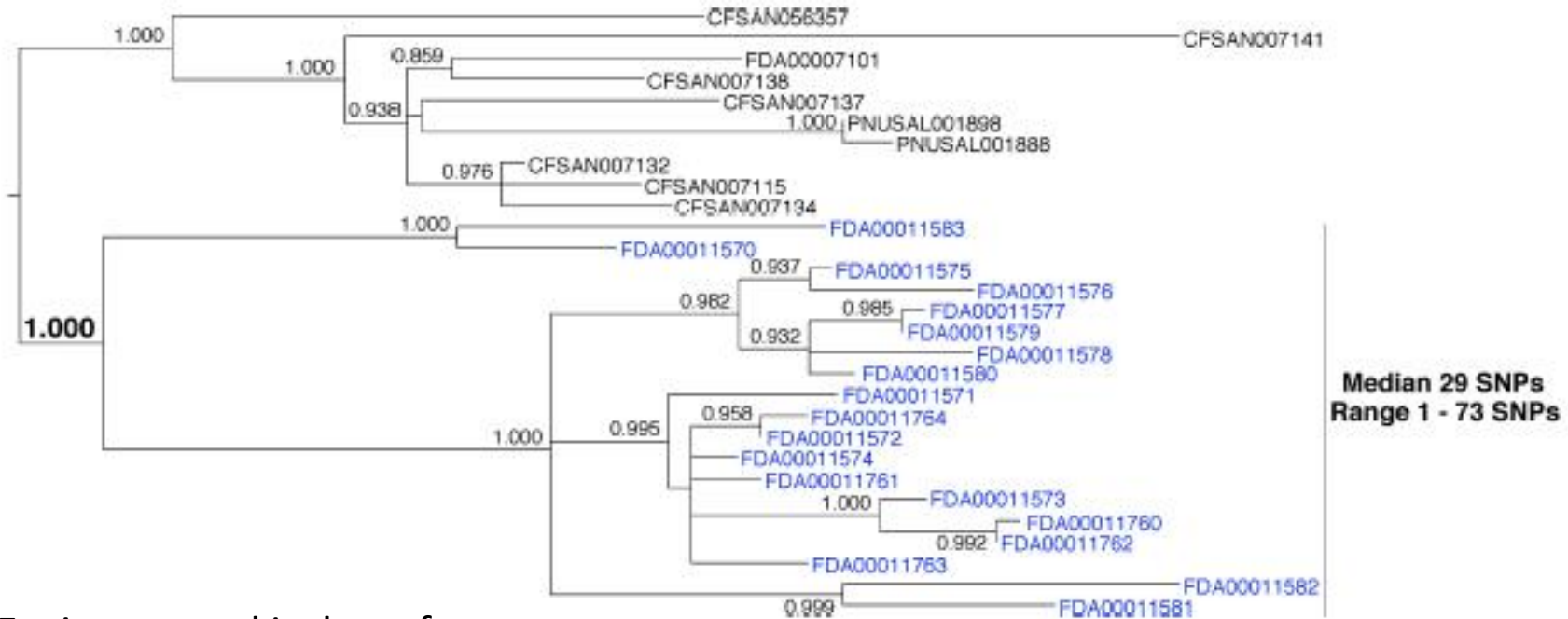


- Bootstrap support
- SNP count support
- Topology support
- Epidemiology support



- Clinical isolates
- Environmental isolates

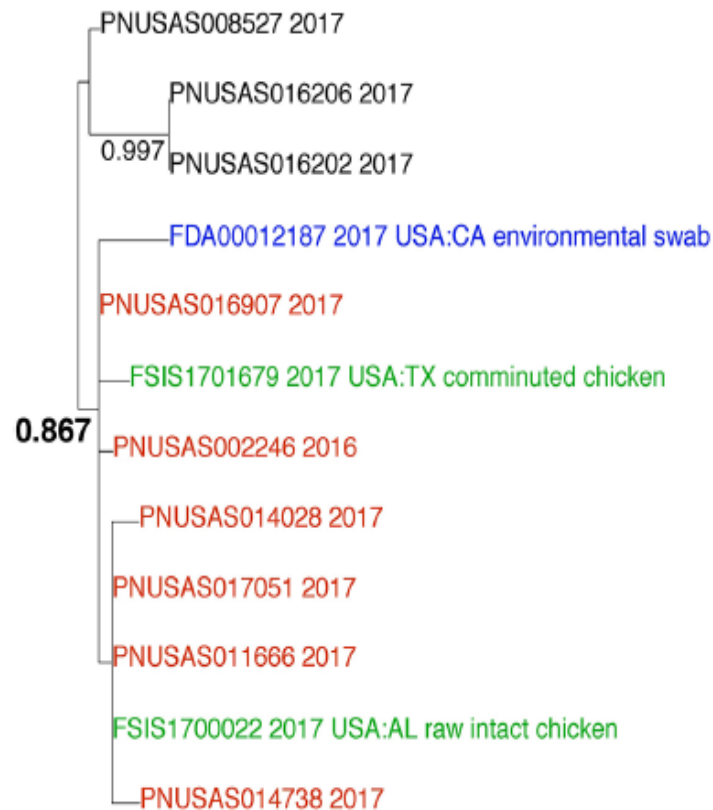
- Bootstrap support
- SNP count support
- Topology support
- Epidemiology not support



Environmental isolates from an  
inspection.

- Clinical isolates
- Environmental isolates

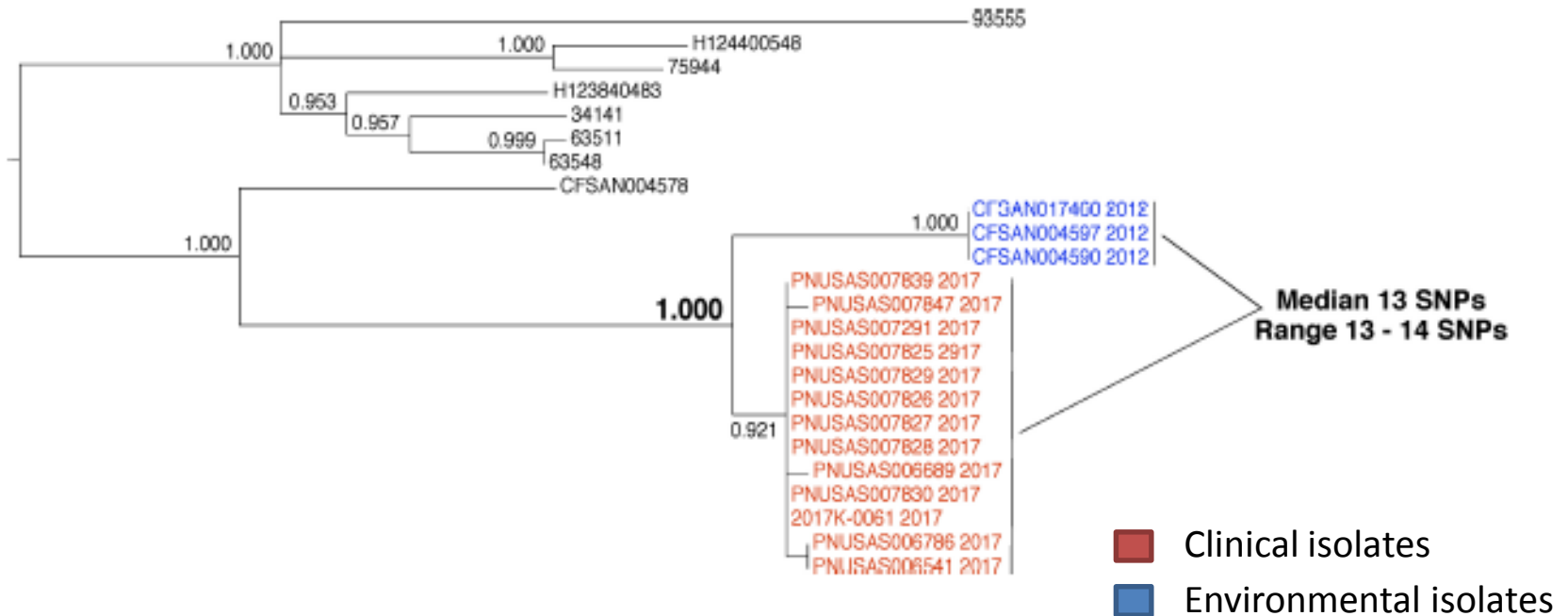
- Bootstrap support
- SNP count neutral
- Topology support



- Bootstrap neutral
- SNP count support
- Topology neutral
- Epidemiology Not support

Median 3 SNPs  
Range 0 - 7 SNPs

- Clinical isolates
- Food isolates
- Environmental isolates



- Bootstrap supports
- SNP count supports
- Topology supports
- Epidemiology does not supports

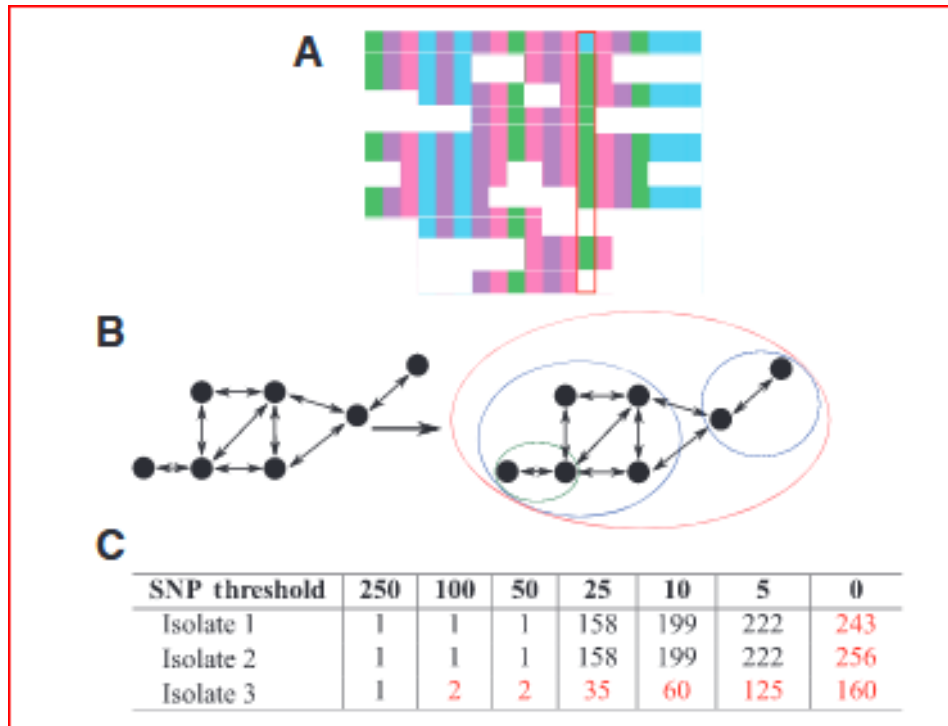
# How to interpret our phylogeny

**TABLE 3** | Characteristics of the examples presented in this paper.

Example	SNP distance	Bootstrap support	Tree topology	Epidemiology, traceback, or compliance findings	Conclusion
Identifying the source of an <i>E. coli</i> outbreak	Supports	Supports	Supports	Supports	Match
Matching food isolates from one firm to environmental isolates from another firm	Supports	Supports	Supports	Supports	Match
Identifying a resident pathogen	Supports	Supports	Supports	Not applicable	Not applicable
Populations of environmental isolates can be very diverse	Neutral	Supports	Supports	Not applicable	Not applicable
Analyzing paraphyletic relationships	Supports	Neutral	Neutral	Does not support	No match
Evidence that isolates arose from the same source by WGS does not necessarily mean that they are linked	Supports	Supports	Supports	Does not support	No match

Pightling et al. *Frontiers in Microbiology*. 2018

# SnapperDB

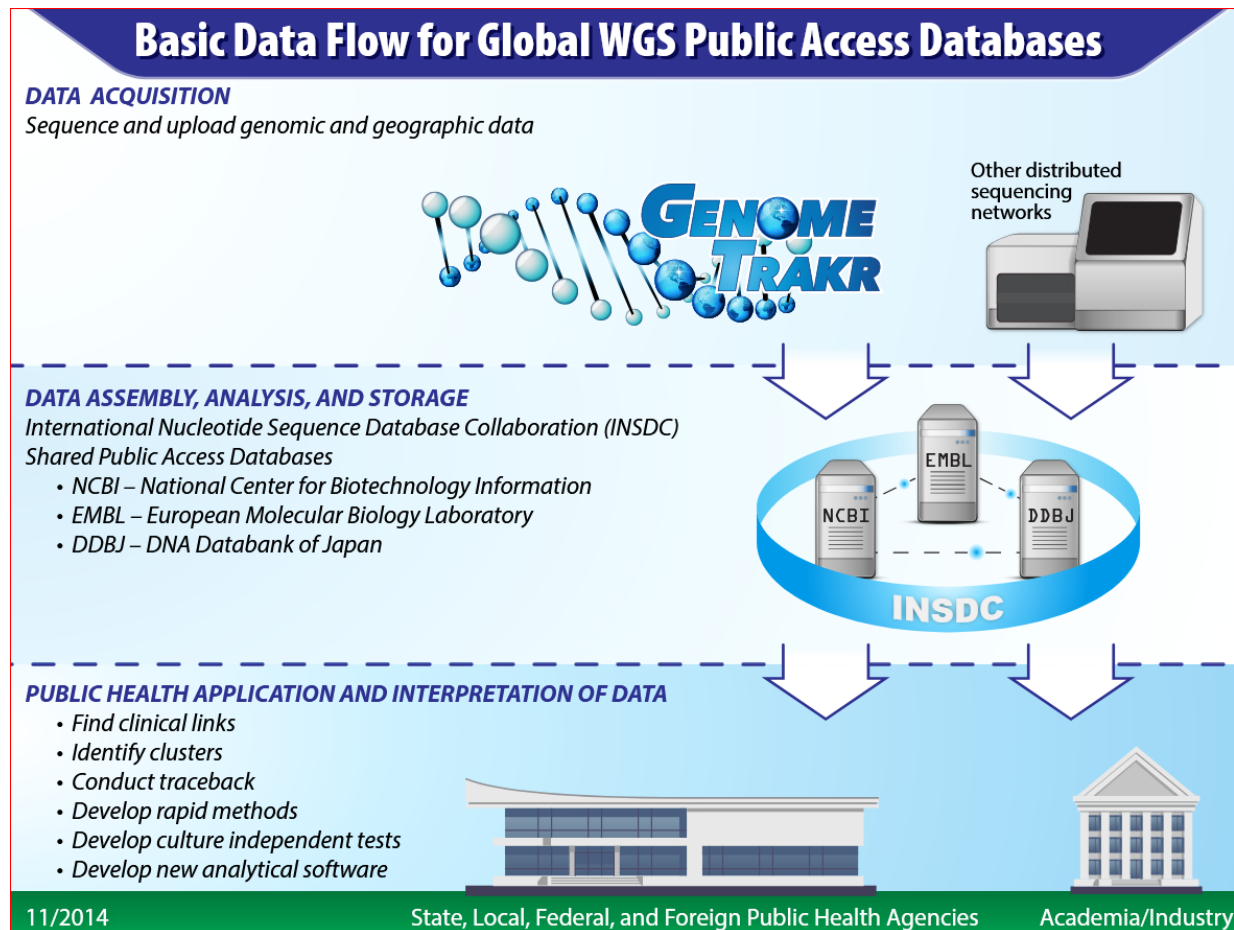


- Hierarchical single linkage clustering of pairwise SNP distances
- Performed at 7 snp thresholds.
- If two strains have the same “address” they have 0 SNP differences.
- If two strains have the same address till 50 threshold, they have less than 50 SNP differences.

Dallman et al. Genome analysis. 2018

# GenomeTrakr

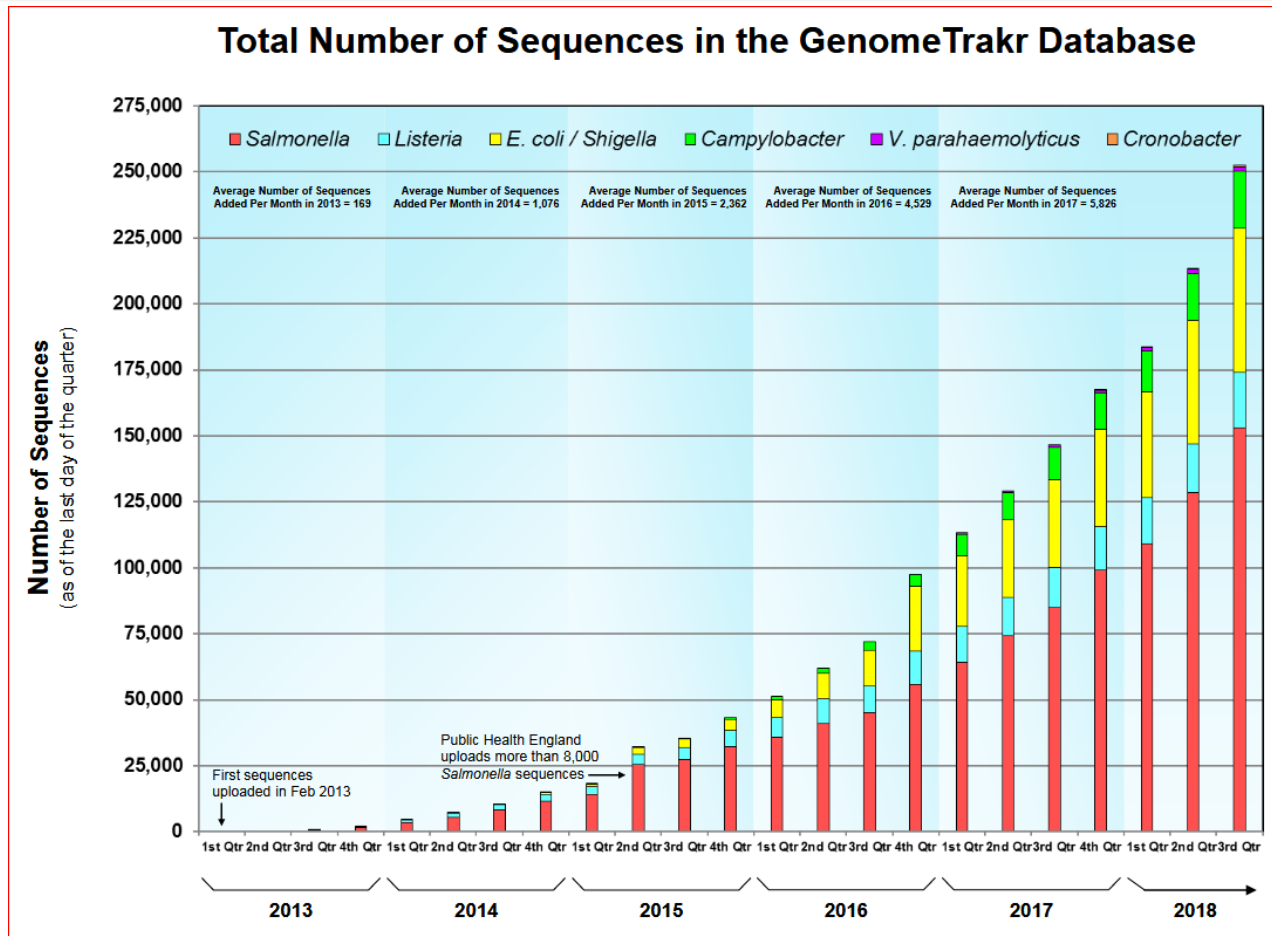
<https://bit.ly/2NNGo37>





# GenomeTrakr

<https://bit.ly/2NNGo37>



# GenomeTrakr

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

Listeria - 1042 isolates

PDG000000001.1023 / PDS000000366.270

**5** Isolates Selected ✕ Clear

Distance between selected isolates:

(minimum=10 SNPs, maximum=30 SNPs, average=22 SNPs)

Target creation date range:

(2013-11-19 to 2018-07-23)

Filter within selected isolates Group ▾

isolates mindiff

**2018 - 1 isolate(s)**

environmental/other USA:PA 2018-07-23 16

**2014 - 2 isolate(s)**

environmental/other USA:CA 2014-01-24 7

clinical USA:NY 2014-01-02 15

**2013 - 2 isolate(s)**

environmental/other USA:WA 2013-11-19 7

environmental/other USA:OR 2013-11-19 17

Filters Share Hide Table

Download

#	Strain	Serovar	Isolate	Create Date	Location	Isolation Source	Isolation type	Host	Min-sam	Min-diff	BioSample	Assembly
1	<input checked="" type="checkbox"/> CFSAN00753		PDT000002448.4	2018-07-23	USA:PA	environmental swab	environmental/other		16	16	<a href="#">SAMN02569870</a>	<a href="#">GCA_003679235.1</a>
2	<input checked="" type="checkbox"/> CFSAN00380		PDT000000909.3	2013-11-19	USA:WA	smoked salmon - cream cheese base	environmental/other		0	7	<a href="#">SAMN02352660</a>	
3	<input checked="" type="checkbox"/> CFSAN00758		PDT000003114.3	2014-01-24	USA:CA	swab	environmental/other		0	7	<a href="#">SAMN02569917</a>	
4	<input checked="" type="checkbox"/> CFSAN00380		PDT000000910.3	2013-11-19	USA:OR	swab	environmental/other		1	17	<a href="#">SAMN02352661</a>	
5	<input checked="" type="checkbox"/> PNUSAL0002.1/2a		PDT000001112.5	2014-01-02	USA:NY	Blood	clinical	Homo sapiens	21	15	<a href="#">SAMN02381991</a>	

Choose Columns Page 1 of 209 5 View 1 - 5 of 1,042

Choose Columns Page 1 of 209 5 View 1 - 5 of 1,042

Export DIM UNSELECTED OFF TREE TIPS OFF

Search & Highlight in Tree

Secuenciación de genomas bacterianos:  
herramientas y aplicaciones

# Thank you for your attention!

---