Session 3.2 – SNP matrix and phylogenetics

BU-ISCIII

Unidades Comunes Científico Técnicas – SGSAFI-ISCIII

28-02 Junio 2021, 3ª 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

Secuenciación de genomas bacterianos:
herramientas y aplicaciones
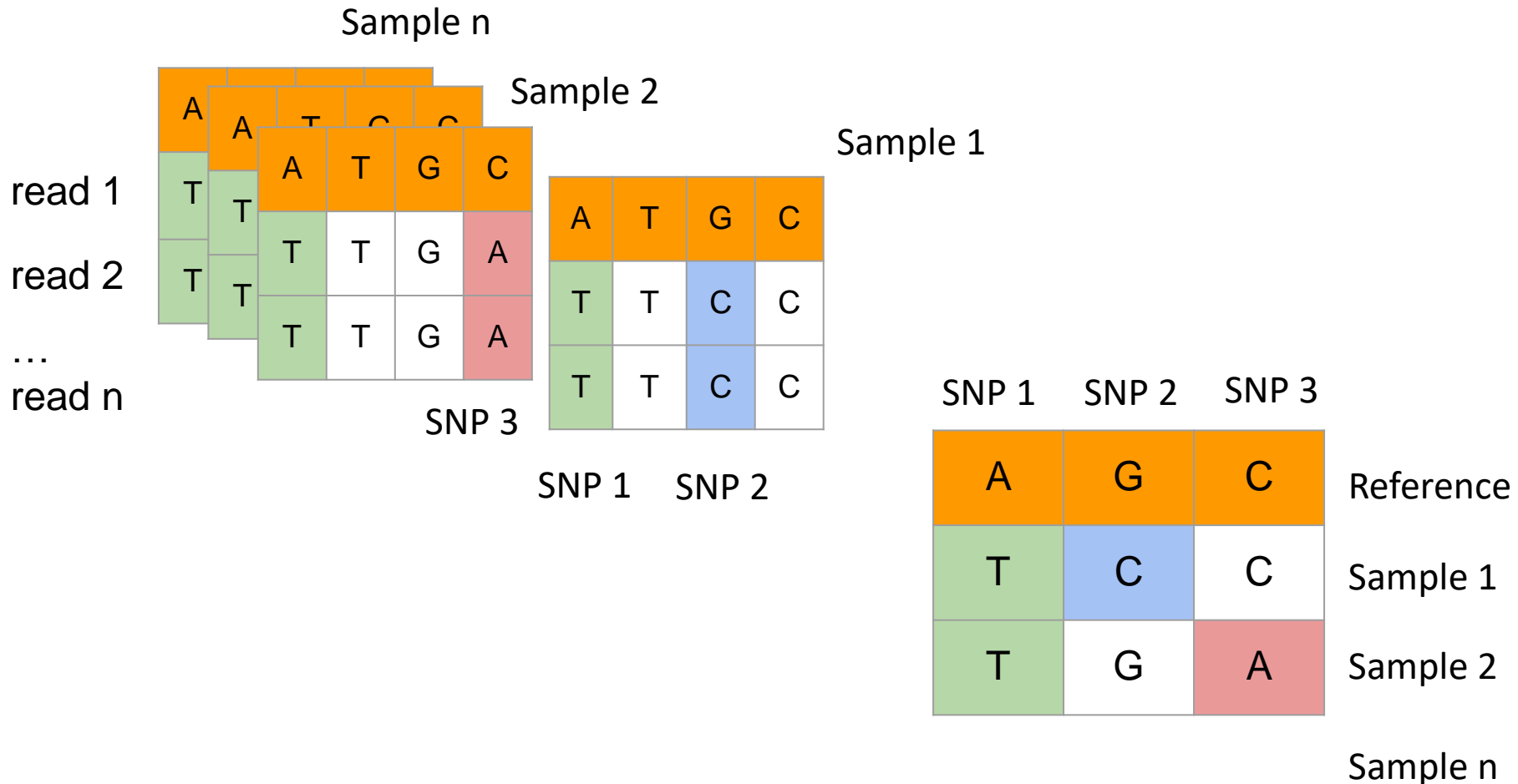
Building a SNP matrix

- 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

Secuenciación de genomas bacterianos:
herramientas y aplicaciones

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
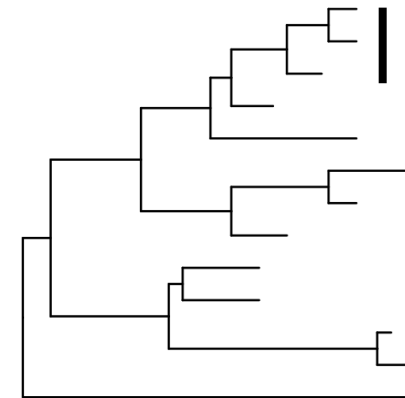Second one is in position 731 in the genome
Third one is in position 921 in the genome

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

Secuenciación de genomas bacterianos: herramientas y aplicaciones

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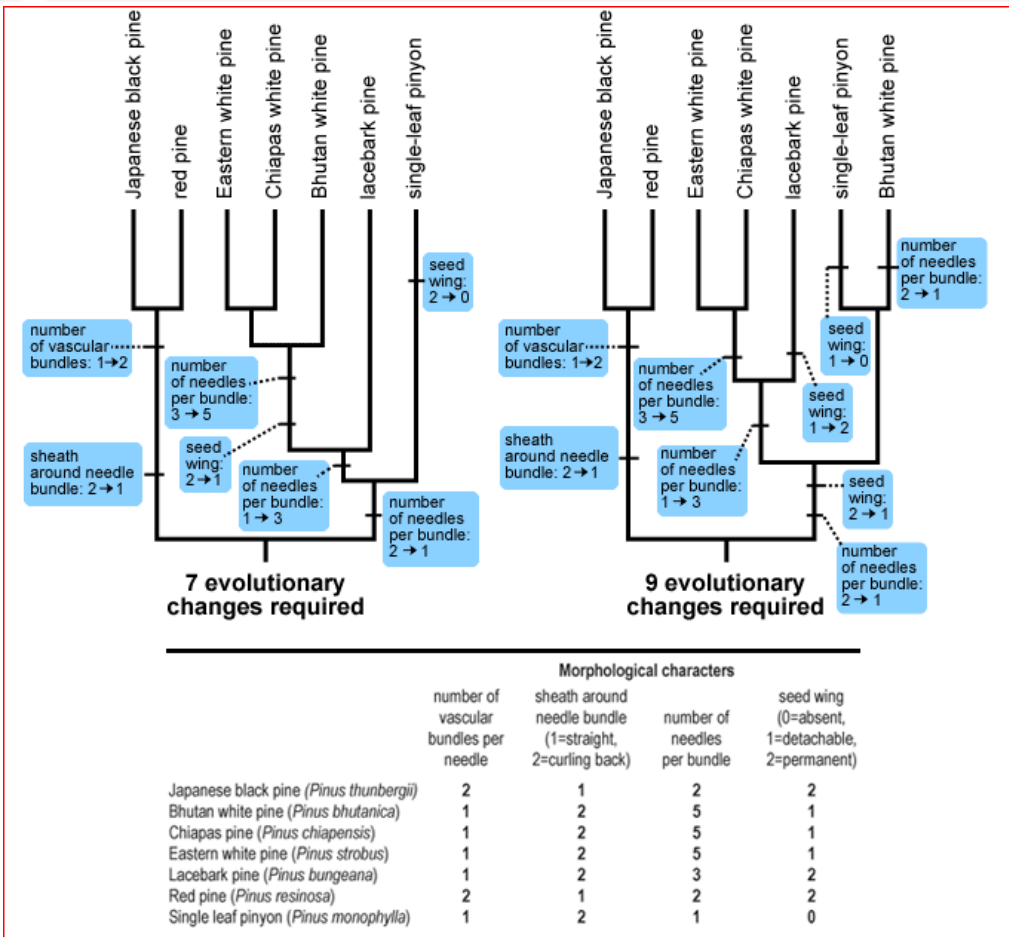
- Search the most parsimonious tree
- The most simple hipothesis must be the correct.
- Search the tree that explains the relationships with the less changes possible.

Secuenciación de genomas bacterianos: herramientas y aplicaciones

Maximum Likelihood

- Searchs the most likely tree given the data and based in a evolutionary model.
- More sofisticated.
- Not prepared a priori for snp matrix.
- RAxML
    - Heterogenity rate disabled.
    - Branchs indicate the expected number of substitution per site.
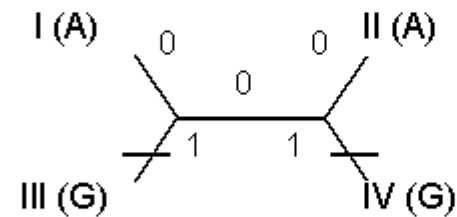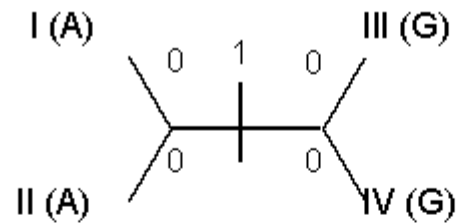
- 0,1 = differences along that branch
- Which hypothesis is more likely, given that the change is rare?

Secuenciación de genomas bacterianos: herramientas y aplicaciones

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).
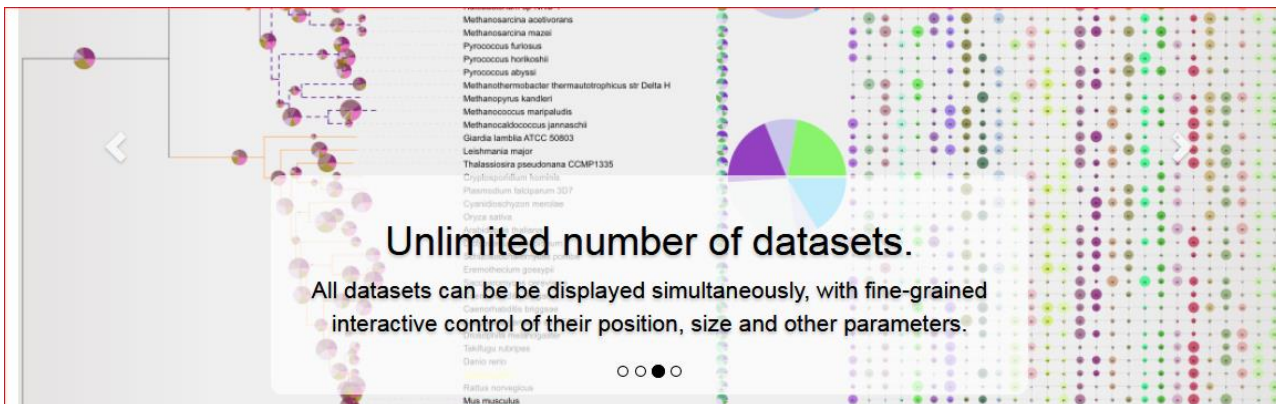
dis_matrix.names	RA.L2073	RA.L2281	RA.L2327	RA.L2391	RA.L2450	RA.L2677	RA.L2701
RA-L2073	0	9403	9028	80	46	46	49
RA-L2281	9403	0	8777	9415	9397	9397	9402
RA-L2327	9028	8777	0	9040	9022	9022	9027
RA-L2391	80	9415	9040	0	74	74	79
RA-L2450	46	9397	9022	74	0	38	45
RA-L2677	46	9397	9022	74	38	0	45
RA-L2701	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

Secuenciación de genomas bacterianos:
herramientas y aplicaciones

iTOL

Secuenciación de genomas bacterianos: herramientas y aplicaciones

How to interpret our phylogeny

- **<u>Combination of:</u>**
  - SNP counts
  - Tree topologies
  - Bootstrap support

Pightling et al. Frontiers in Microbiology. 2018

Secuenciación de genomas bacterianos:
herramientas y aplicaciones

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	
E. coli	4	X			Underwood et al., 2013
E. coli	15	X			Eppinger et al., 2011
L. monocytogenes	9	X			Chen et al., 2017c
L. monocytogenes	12	X			Chen et al., 2017a
L. monocytogenes	18	X			Li et al., 2017
L. monocytogenes	20	X			Wang et al., 2015
L. monocytogenes	21		X		Nielsen et al., 2017
L. monocytogenes	28		X		Gilmour et al., 2010
L. monocytogenes	29		X		Chen et al., 2017b
L. monocytogenes	42		X		Chen et al., 2016
L. monocytogenes	67		X		Jackson et al., 2016
S. enterica	2	X			Wuyts et al., 2015
S. enterica	3	X			Allard et al., 2016
S. enterica	3	X			Taylor et al., 2015
S. enterica	6	X			Hoffmann et al., 2016
S. enterica	12	X			Octavia et al., 2015
S. enterica	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

Secuenciación de genomas bacterianos: herramientas y aplicaciones
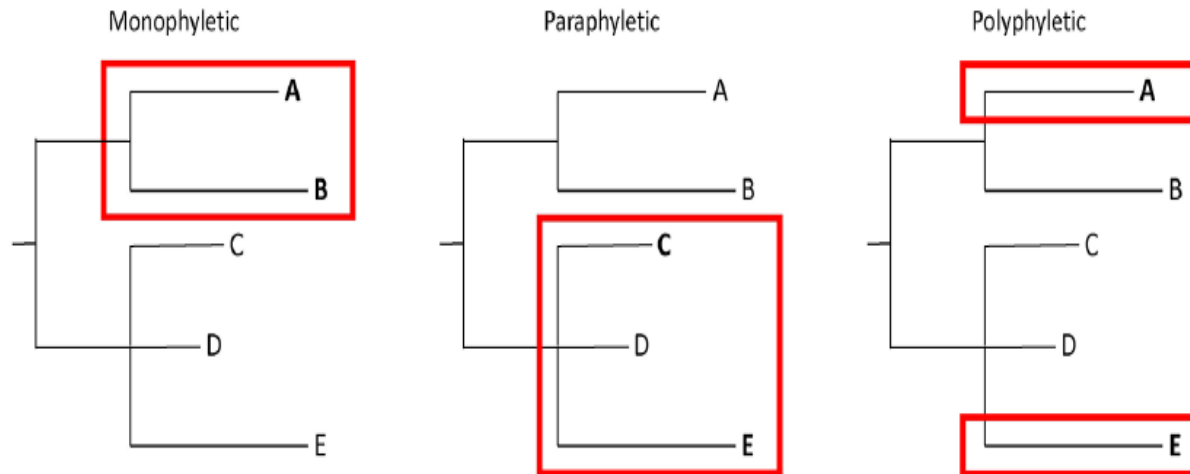
How to interpret our phylogeny

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

Pightling et al. Frontiers in Microbiology. 2018
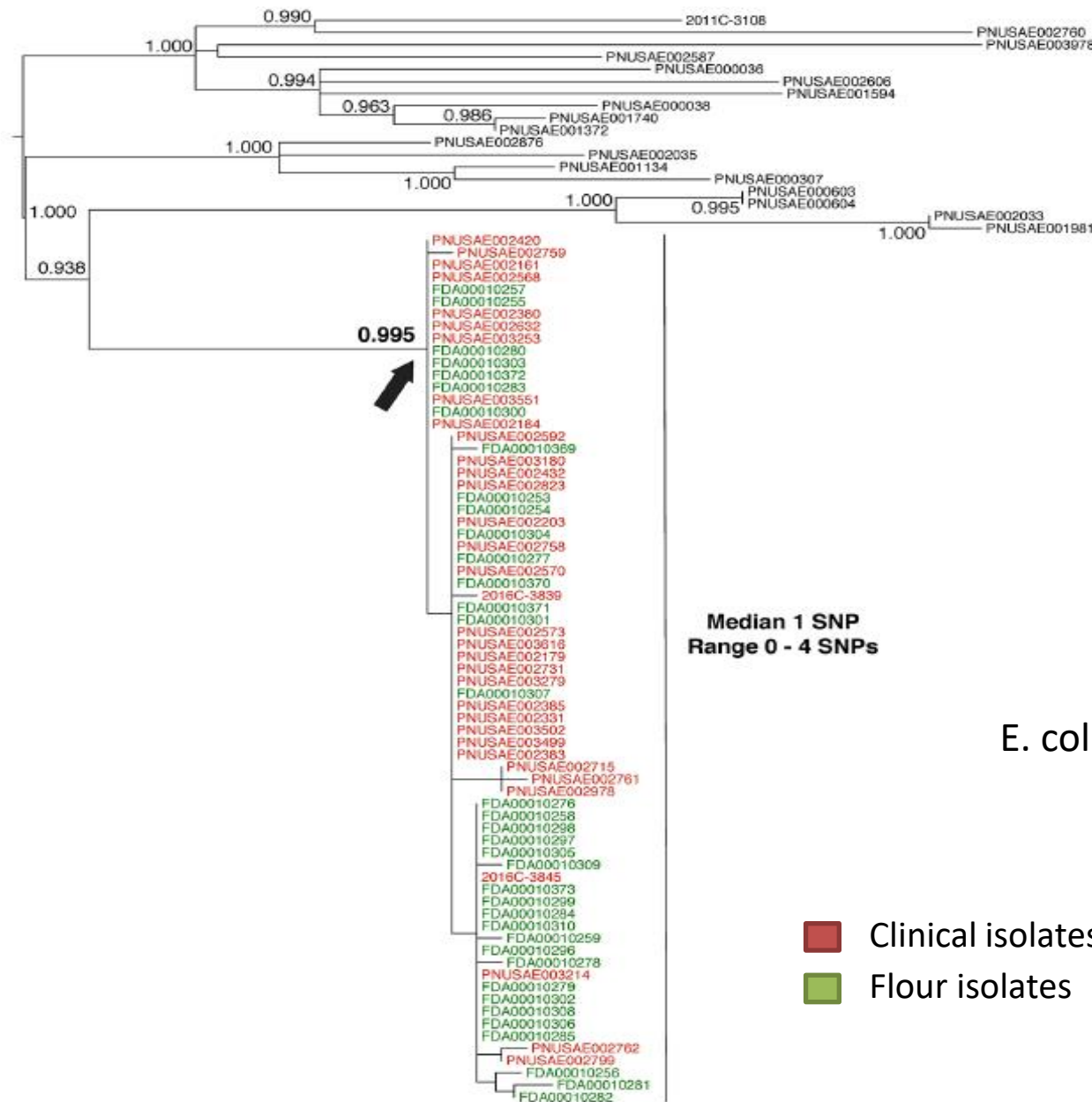
- Bootstrap support
- SNP count support
- Topology support
- Epidemiology support

E. coli Clinical isolates - source

Median 1 SNP
Range 0 - 4 SNPs

Clinical isolates
Flour isolates

L. monocytogenes ingredient supplier – Ice cream producer

- Bootstrap support
- SNP count support
- Topology support
- Epidemiology support
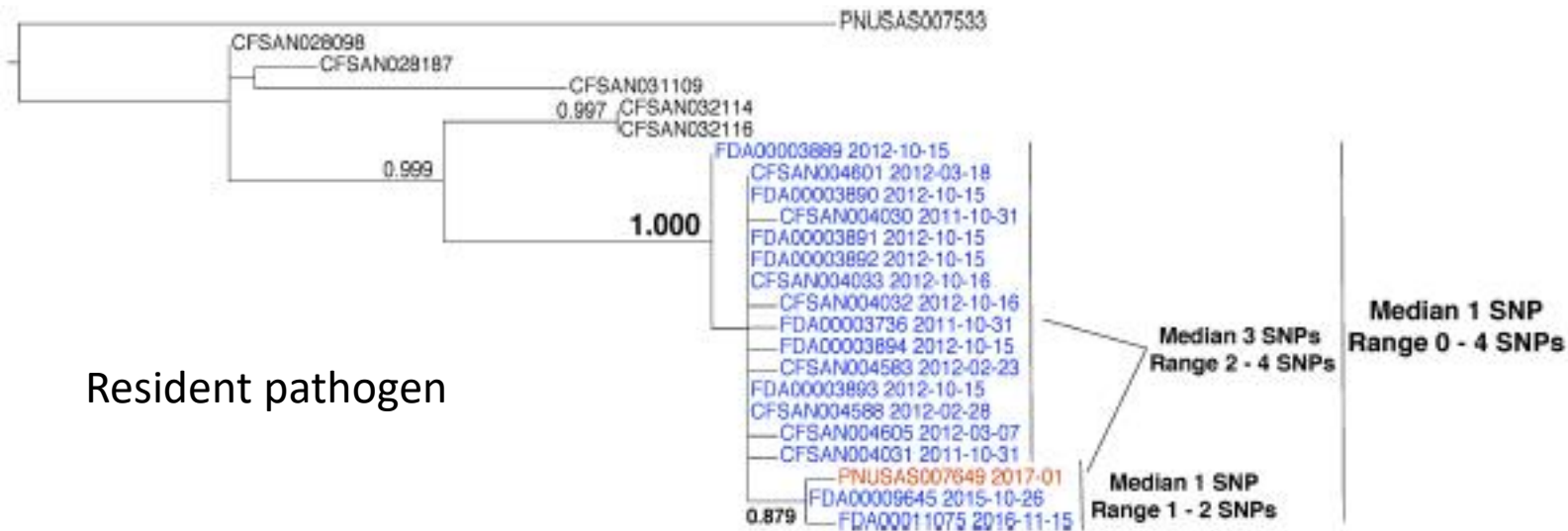
Food isolates
Environmental isolates

Secuenciación de genomas bacterianos: herramientas y aplicaciones

>&_BU-ISCIII

Resident pathogen

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

Clinical isolates
Environmental isolates

Secuenciación de genomas bacterianos:
herramientas y aplicaciones

>ꟸ_BU-ISCIII
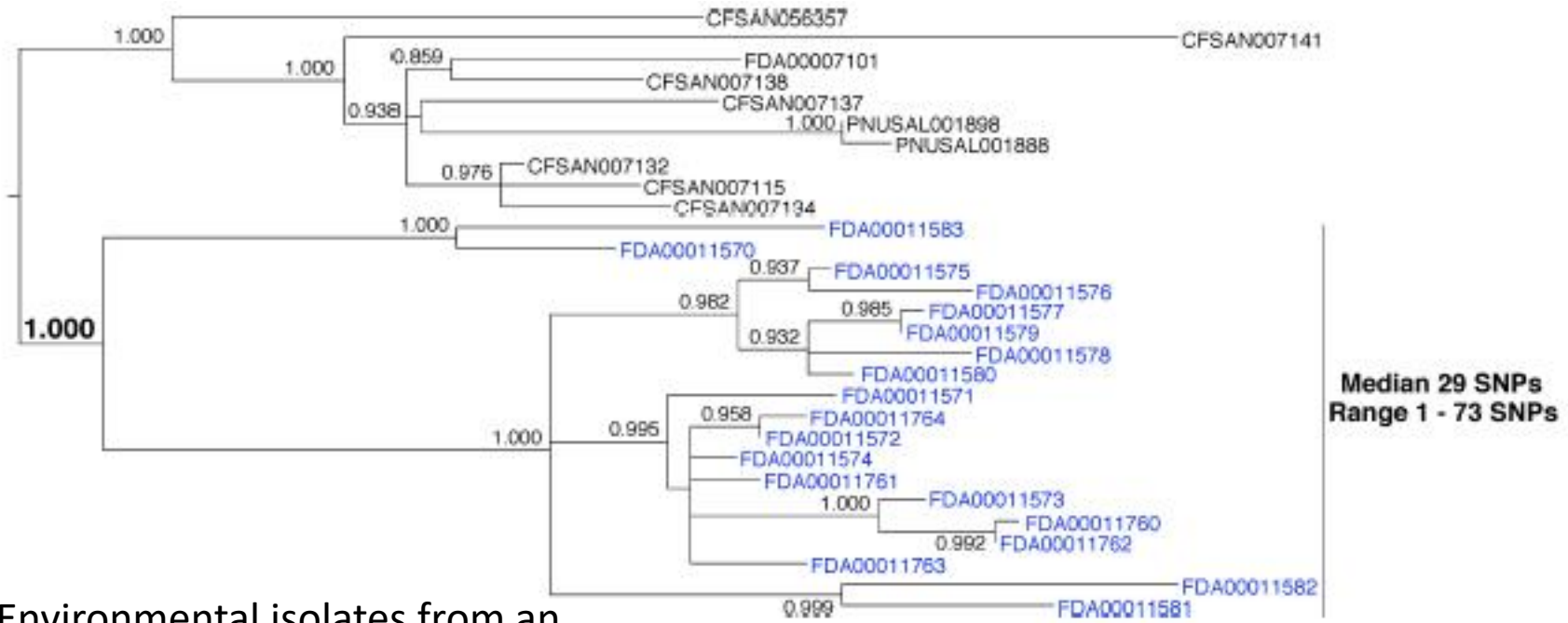
Environmental isolates from an
inspection.

🟥	Clinical isolates	- Bootstrap support
🟦	Environmental isolates	- SNP count neutral
		- Topology support

PNUSAS008527 2017
PNUSAS016206 2017
0.997 PNUSAS016202 2017

FDA00012187 2017 USA:CA environmental swab

PNUSAS016907 2017

FSIS1701679 2017 USA:TX comminuted chicken

0.867 PNUSAS002246 2016

PNUSAS014028 2017

PNUSAS017051 2017

PNUSAS011666 2017

FSIS1700022 2017 USA:AL raw intact chicken

PNUSAS014738 2017
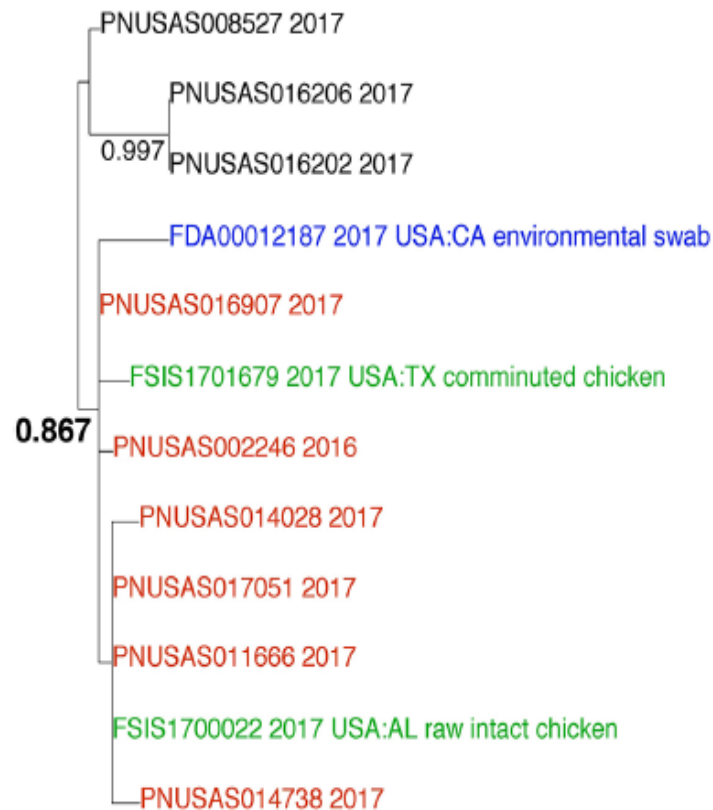
- Bootstrap neutral
- SNP count support
- Topology neutral
- Epidemiology Not support

Median 3 SNPs
Range 0 - 7 SNPs

Clinical isolates
Food isolates
Environmental isolates

Secuenciación de genomas bacterianos:
herramientas y aplicaciones

>ꢀ_BU-ISCIII

Clinical isolates

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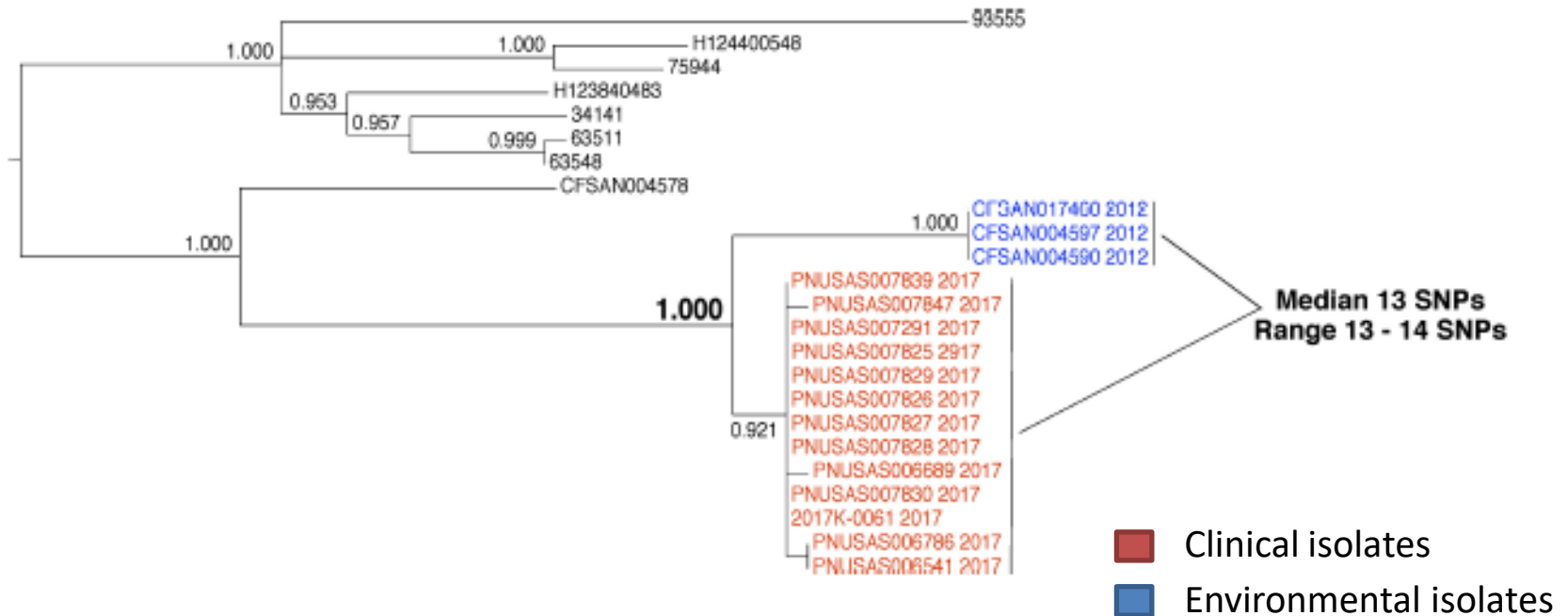
- Bootstrap supports
- SNP count supports
- Topology supports
- Epidemiology does not supports

Secuenciación de genomas bacterianos:
herramientas y aplicaciones

>&_BU-ISCIII

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 *E. coli* 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

Secuenciación de genomas bacterianos:
herramientas y aplicaciones

SnapperDB

SNP threshold	250	100	50	25	10	5	0
Isolate 1	1	1	1	158	199	222	243
Isolate 2	1	1	1	158	199	222	256
Isolate 3	1	2	2	35	60	125	160
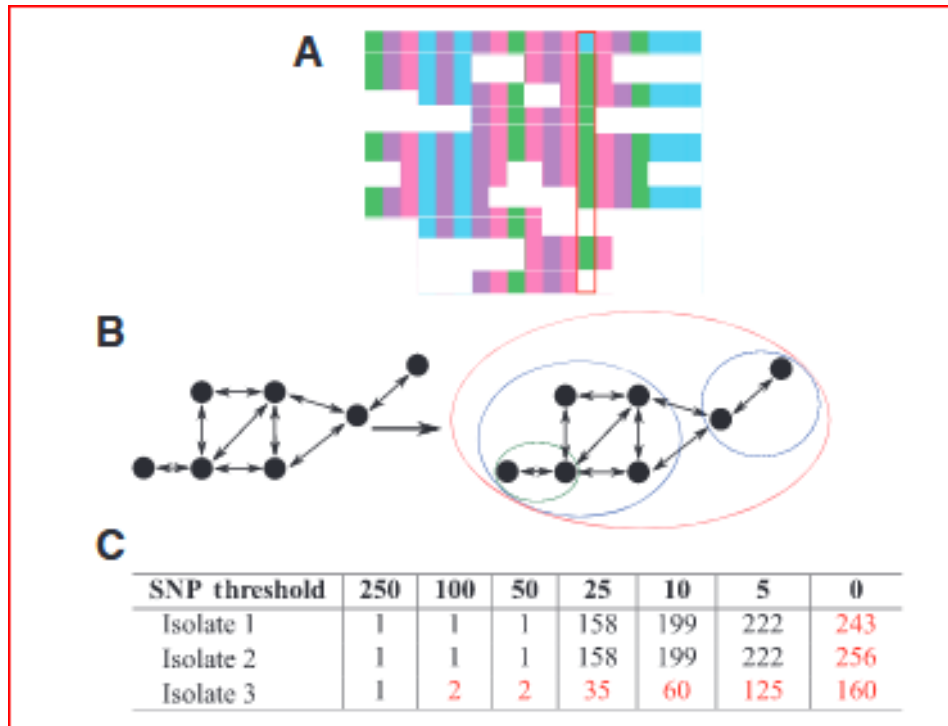
- Hierarchical single linkage clustering of pairwise SNP distances

- 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 anaysis. 2018

Secuenciación de genomas bacterianos:
herramientas y aplicaciones

GenomeTrakr

https://bit.ly/2NNGo37

Secuenciación de genomas bacterianos: herramientas y aplicaciones

GenomeTrakr

Total Number of Sequences in the GenomeTrakr Database

Secuenciación de genomas bacterianos:
herramientas y aplicaciones

GenomeTrakr

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

Thank you for your attention!