



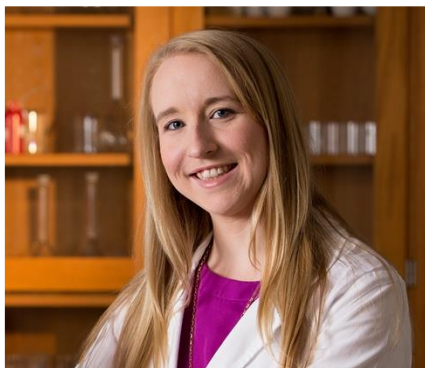
Whole Genome Sequencing for Cluster Detection and Investigation Live Learning Series 2026

Session I
WGS Fundamentals

Friday, May 29th, 2026 12-1:30 PM ET

Introductions

Lauren Hudson, PhD



Research Assistant
Professor

Dept. of Food
Science

University of
Tennessee

lkudson@utk.edu

Overview

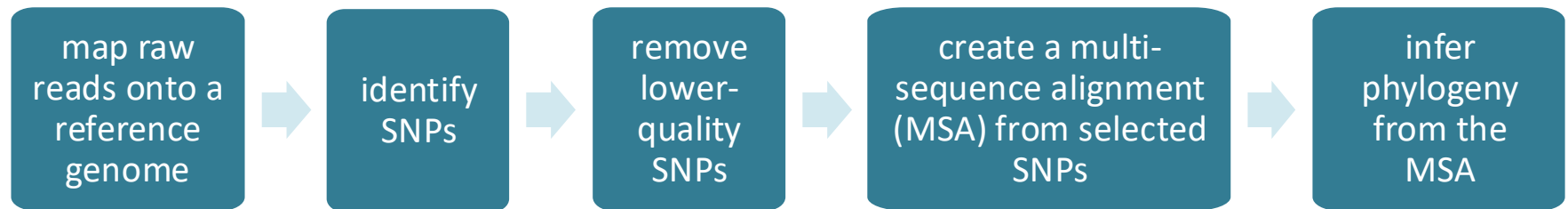
	Overview & Introductions
	SNP & cg/wgMLST Refresher
	Phylogenetic Trees/Dendrograms & Matrices
	Allele Codes
	X Codes
	Break
	Thresholds for Cluster Detection
	REP Codes/Strains
	Next Sessions
	Questions/Discussion
	Survey



SNP-based Analyses

SNP-based Analysis

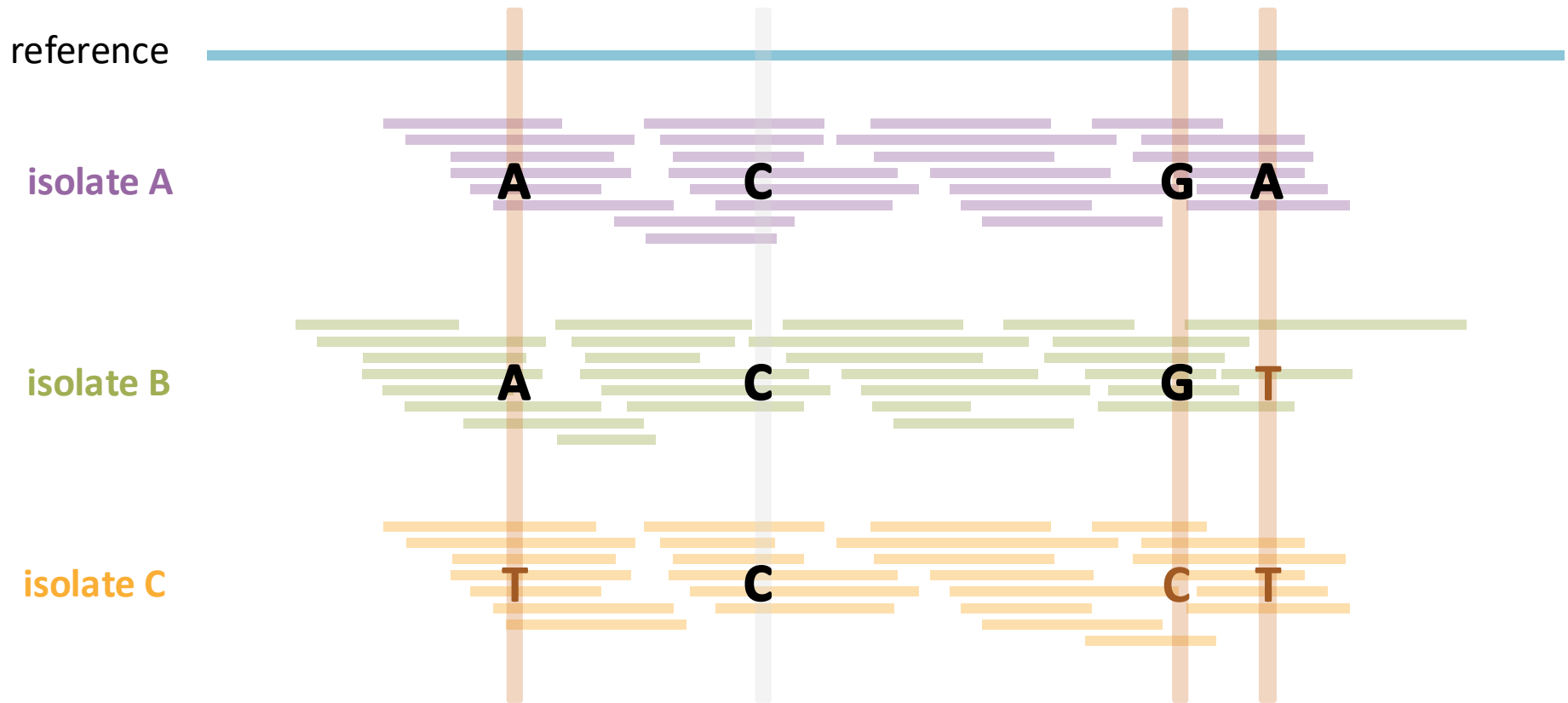
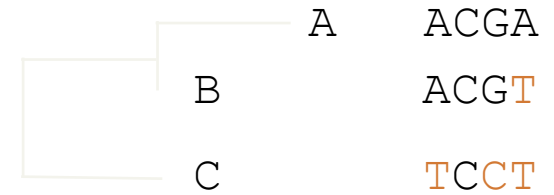
- Single nucleotide changes are used to infer phylogenetic relatedness between isolates, thus they can be used to infer phylogeny and for strain typing



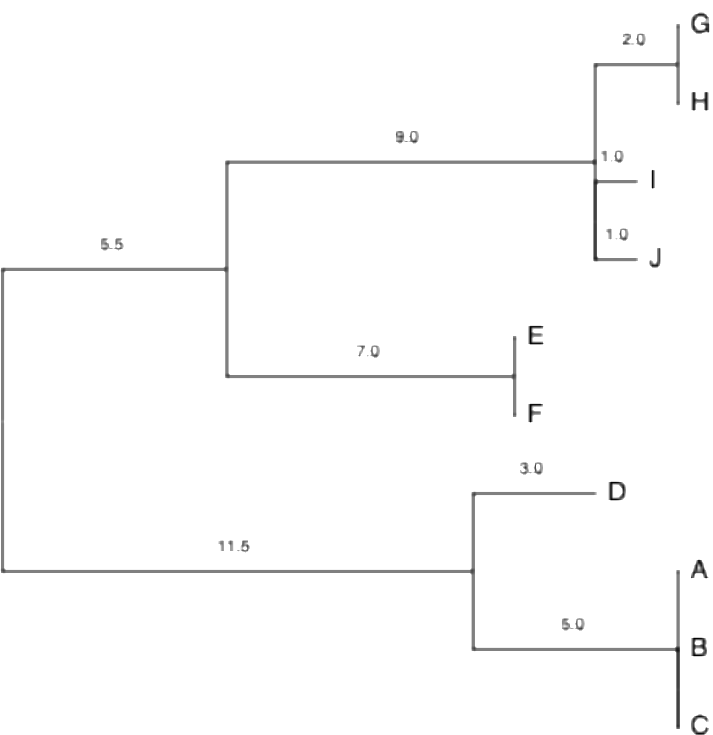
reference
sequence

ATGTT**C**CTC
ATGTT**G**CTC

SNP Calling



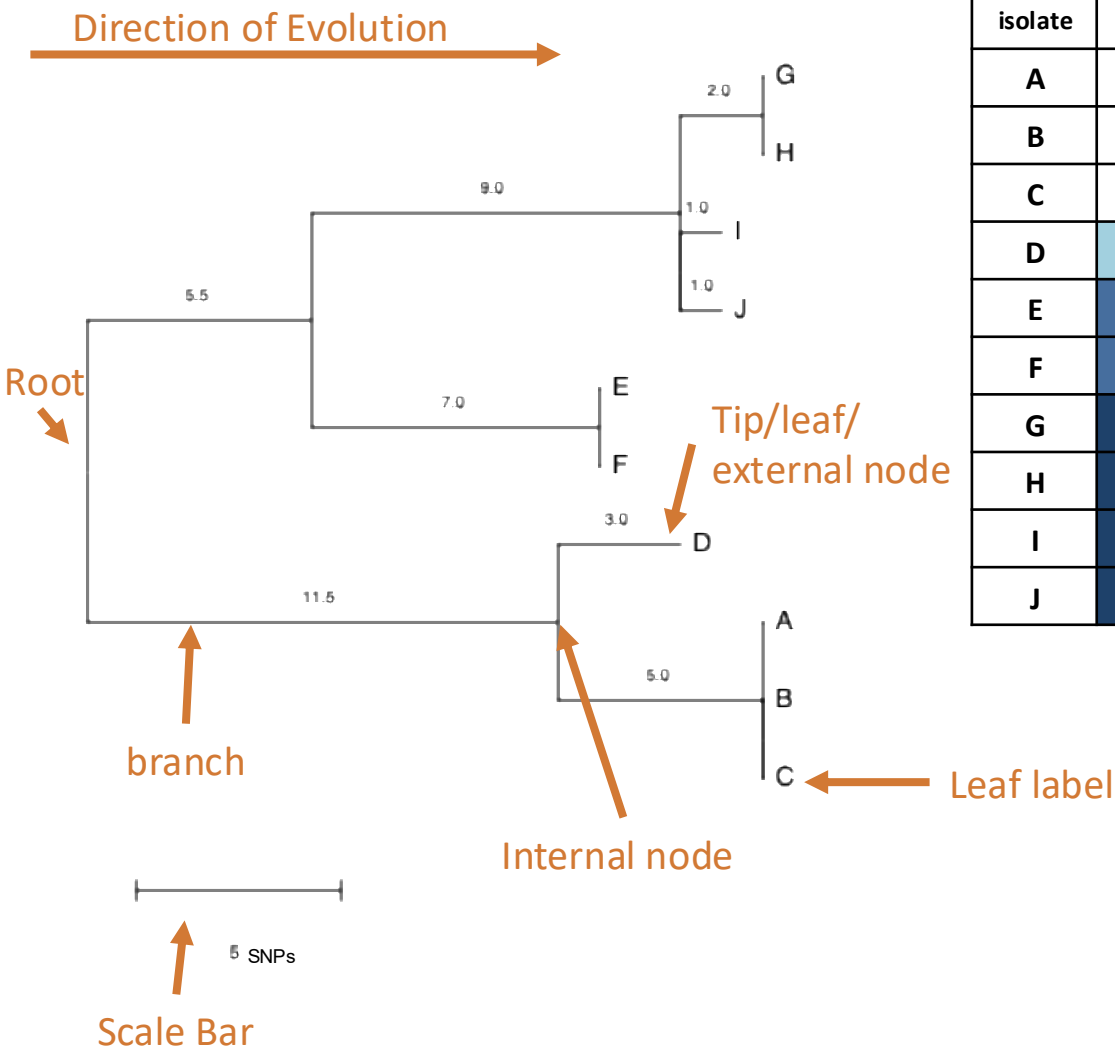
Phylogenetic Tree



SNP Matrix

isolate	A	B	C	D	E	F	G	H	I	J
A	0	0	0	8	29	29	33	33	32	32
B	0	0	0	8	29	29	33	33	32	32
C	0	0	0	8	29	29	33	33	32	32
D	8	8	8	0	27	27	31	31	30	30
E	29	29	29	27	0	0	18	18	17	17
F	29	29	29	27	0	0	18	18	17	17
G	33	33	33	31	18	18	0	0	3	3
H	33	33	33	31	18	18	0	0	3	3
I	32	32	32	30	17	17	3	3	0	2
J	32	32	32	30	17	17	3	3	2	0

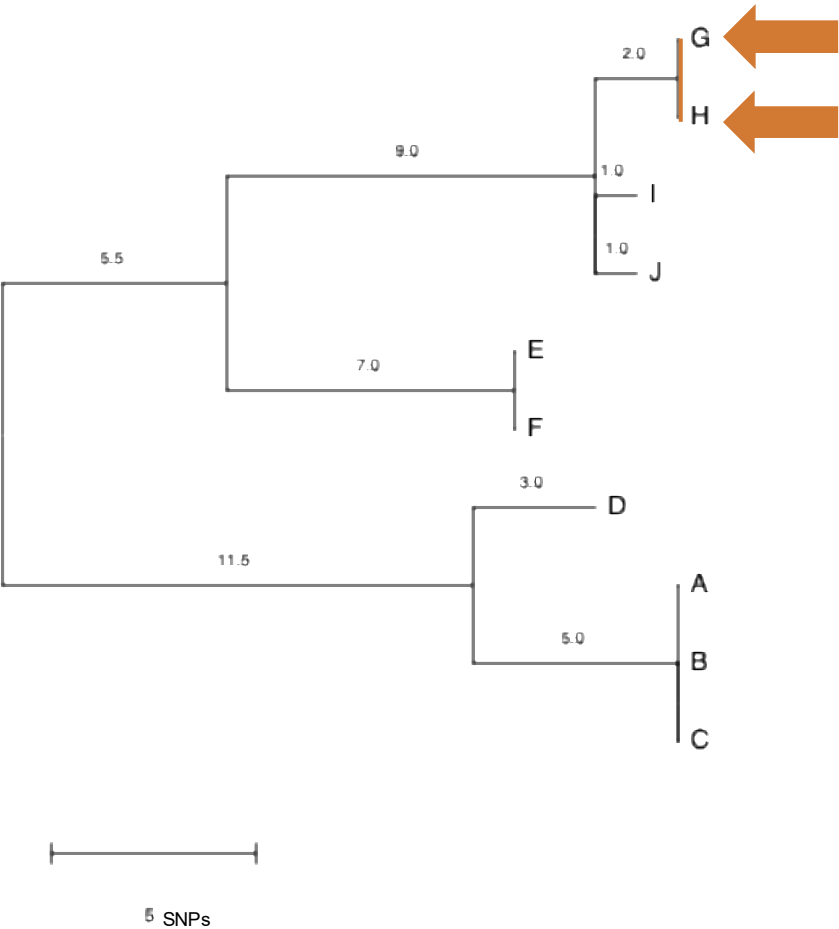
Phylogenetic Tree



SNP Matrix

isolate	A	B	C	D	E	F	G	H	I	J
A	0	0	0	8	29	29	33	33	32	32
B	0	0	0	8	29	29	33	33	32	32
C	0	0	0	8	29	29	33	33	32	32
D	8	8	8	0	27	27	31	31	30	30
E	29	29	29	27	0	0	18	18	17	17
F	29	29	29	27	0	0	18	18	17	17
G	33	33	33	31	18	18	0	0	3	3
H	33	33	33	31	18	18	0	0	3	3
I	32	32	32	30	17	17	3	3	0	2
J	32	32	32	30	17	17	3	3	2	0

Phylogenetic Tree



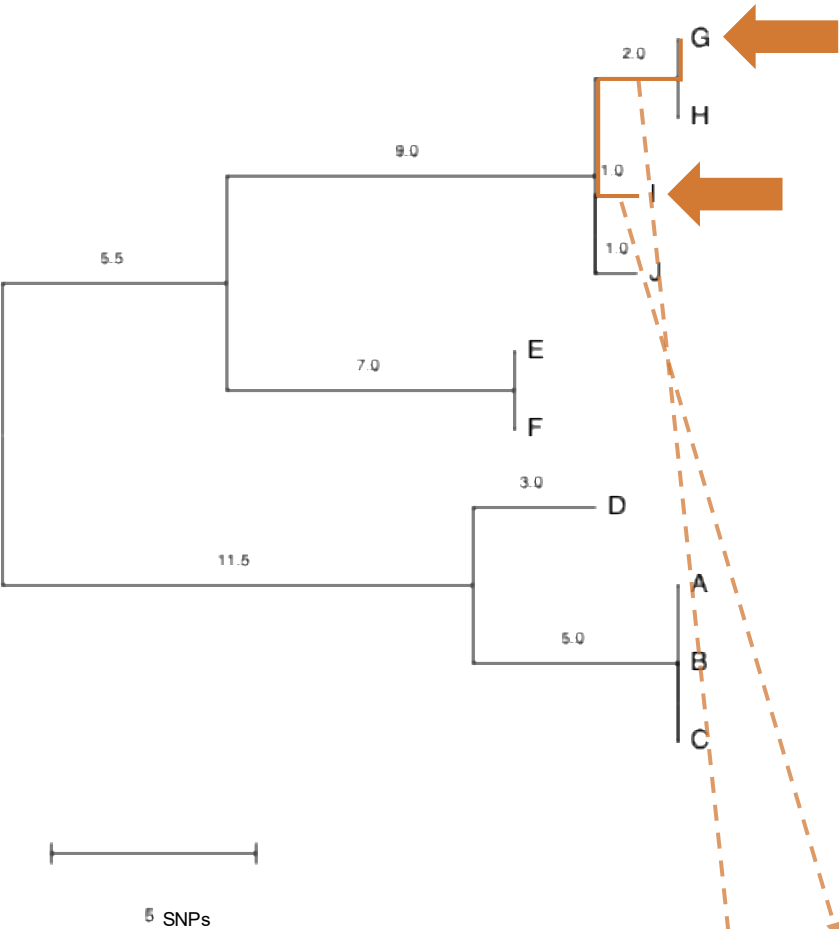
SNP Matrix

isolate	A	B	C	D	E	F	G	H	I	J
A	0	0	0	8	29	29	33	33	32	32
B	0	0	0	8	29	29	33	33	32	32
C	0	0	0	8	29	29	33	33	32	32
D	8	8	8	0	27	27	31	31	30	30
E	29	29	29	27	0	0	18	18	17	17
F	29	29	29	27	0	0	18	18	17	17
G	33	33	33	31	18	18	0	0	3	3
H	33	33	33	31	18	18	0	0	3	3
I	32	32	32	30	17	17	3	3	0	2
J	32	32	32	30	17	17	3	3	2	0

What is the distance between isolate G and isolate H?

0 SNPs

Phylogenetic Tree



SNP Matrix

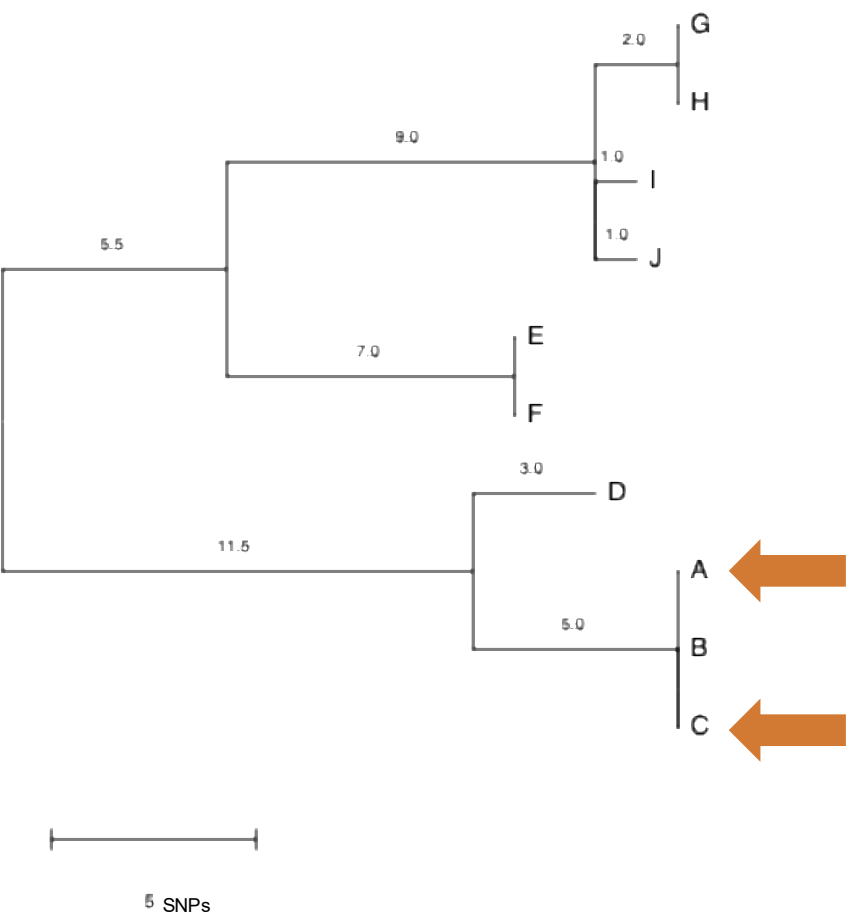
isolate	A	B	C	D	E	F	G	H	I	J
A	0	0	0	8	29	29	33	33	32	32
B	0	0	0	8	29	29	33	33	32	32
C	0	0	0	8	29	29	33	33	32	32
D	8	8	8	0	27	27	31	31	30	30
E	29	29	29	27	0	0	18	18	17	17
F	29	29	29	27	0	0	18	18	17	17
G	33	33	33	31	18	18	0	0	3	3
H	33	33	33	31	18	18	0	0	3	3
I	32	32	32	30	17	17	3	3	0	2
J	32	32	32	30	17	17	3	3	2	0

What is the distance between isolate G and isolate I?

2 + 1 = 3

3 SNPs

Phylogenetic Tree



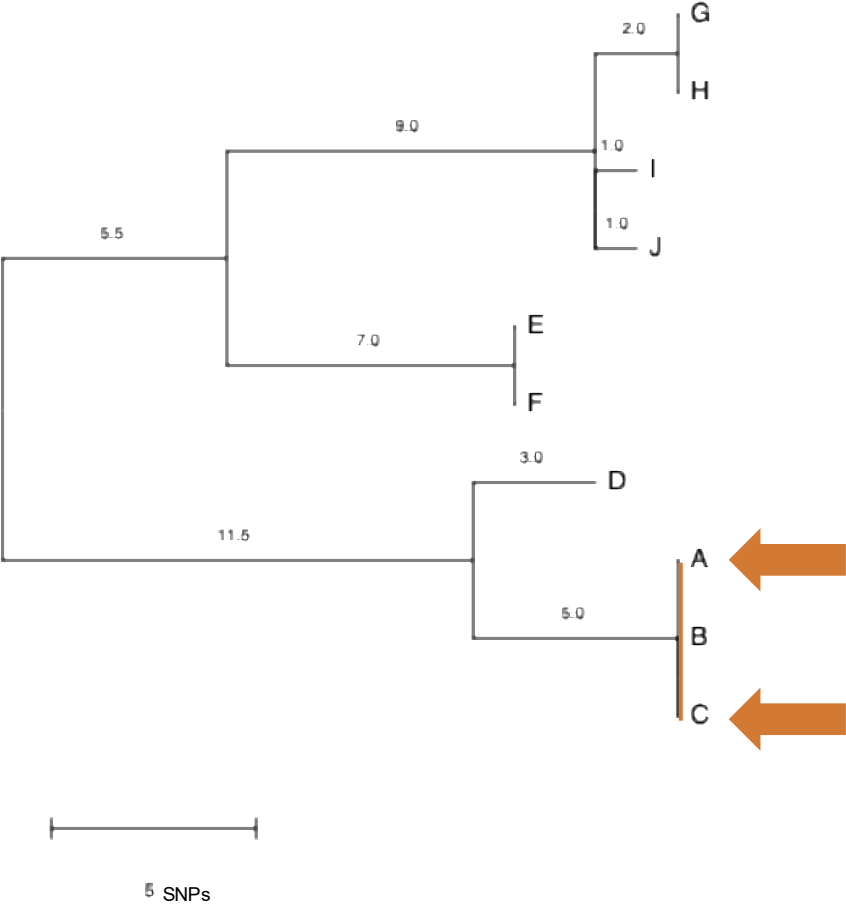
SNP Matrix

isolate	A	B	C	D	E	F	G	H	I	J
A	0	0	0	8	29	29	33	33	32	32
B	0	0	0	8	29	29	33	33	32	32
C	0	0	0	8	29	29	33	33	32	32
D	8	8	8	0	27	27	31	31	30	30
E	29	29	29	27	0	0	18	18	17	17
F	29	29	29	27	0	0	18	18	17	17
G	33	33	33	31	18	18	0	0	3	3
H	33	33	33	31	18	18	0	0	3	3
I	32	32	32	30	17	17	3	3	0	2
J	32	32	32	30	17	17	3	3	2	0



Knowledge Check:
What is the distance between isolate A and isolate C?

Phylogenetic Tree



SNP Matrix

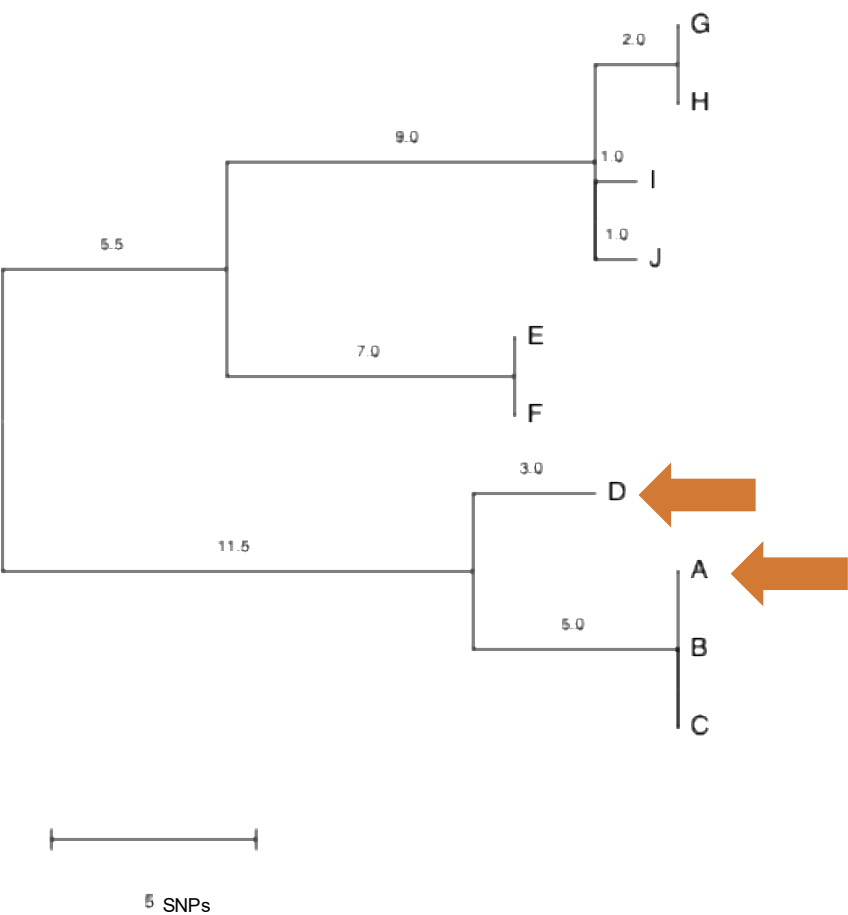
isolate	A	B	C	D	E	F	G	H	I	J
A	0	0	0	8	29	29	33	33	32	32
B	0	0	0	8	29	29	33	33	32	32
C	0	0	0	8	29	29	33	33	32	32
D	8	8	8	0	27	27	31	31	30	30
E	29	29	29	27	0	0	18	18	17	17
F	29	29	29	27	0	0	18	18	17	17
G	33	33	33	31	18	18	0	0	3	3
H	33	33	33	31	18	18	0	0	3	3
I	32	32	32	30	17	17	3	3	0	2
J	32	32	32	30	17	17	3	3	2	0



Knowledge Check:
What is the distance
between isolate A
and and isolate C?

0 SNPs

Phylogenetic Tree



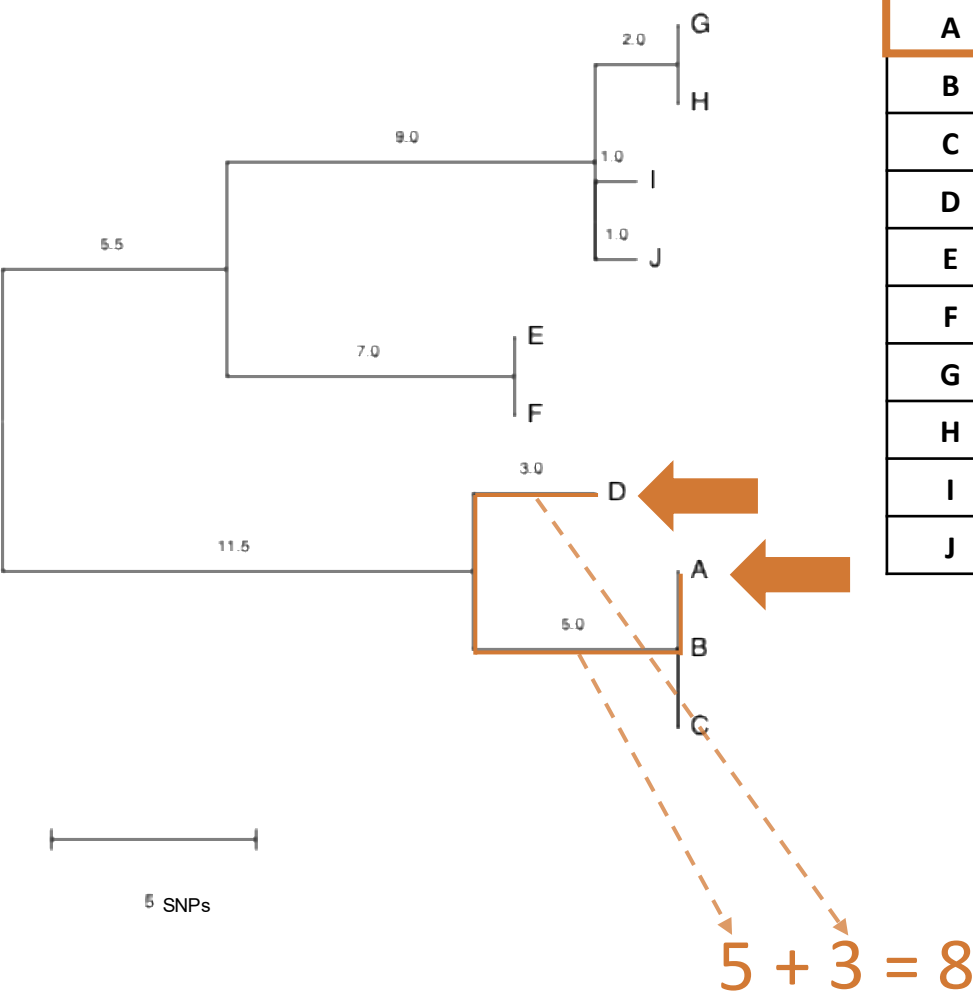
SNP Matrix

isolate	A	B	C	D	E	F	G	H	I	J
A	0	0	0	8	29	29	33	33	32	32
B	0	0	0	8	29	29	33	33	32	32
C	0	0	0	8	29	29	33	33	32	32
D	8	8	8	0	27	27	31	31	30	30
E	29	29	29	27	0	0	18	18	17	17
F	29	29	29	27	0	0	18	18	17	17
G	33	33	33	31	18	18	0	0	3	3
H	33	33	33	31	18	18	0	0	3	3
I	32	32	32	30	17	17	3	3	0	2
J	32	32	32	30	17	17	3	3	2	0



Knowledge Check:
What is the distance between isolate A and isolate D?

Phylogenetic Tree



SNP Matrix

isolate	A	B	C	D	E	F	G	H	I	J
A	0	0	0	8	29	29	33	33	32	32
B	0	0	0	8	29	29	33	33	32	32
C	0	0	0	8	29	29	33	33	32	32
D	8	8	8	0	27	27	31	31	30	30
E	29	29	29	27	0	0	18	18	17	17
F	29	29	29	27	0	0	18	18	17	17
G	33	33	33	31	18	18	0	0	3	3
H	33	33	33	31	18	18	0	0	3	3
I	32	32	32	30	17	17	3	3	0	2
J	32	32	32	30	17	17	3	3	2	0



Knowledge Check:
What is the distance
between isolate A
and and isolate D?

8 SNPs

Terminology

Dendrogram

- general term for any tree-like diagram that shows relationships between entities

Phylogenetic tree

- general term for a diagram representing evolutionary relationships

Chronogram

- branch length represents time, but the length does not provide information on evolutionary change

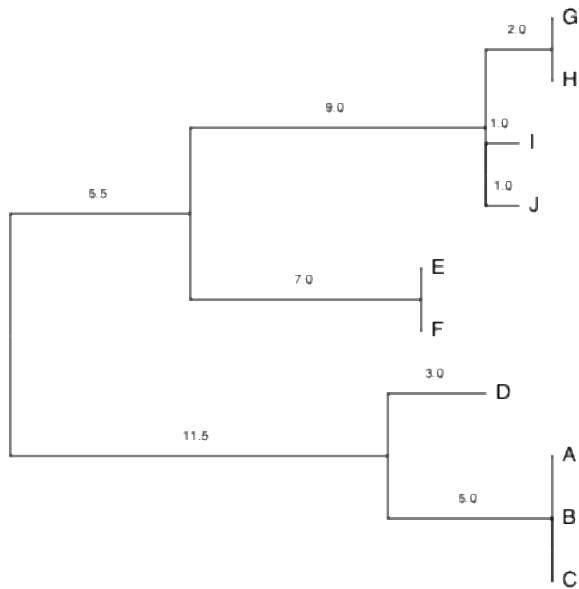
Cladogram

- Unscaled: branch length is arbitrary & does not represent time or evolutionary change
- displays only the branching pattern of evolutionary relationships among organisms

Phylogram

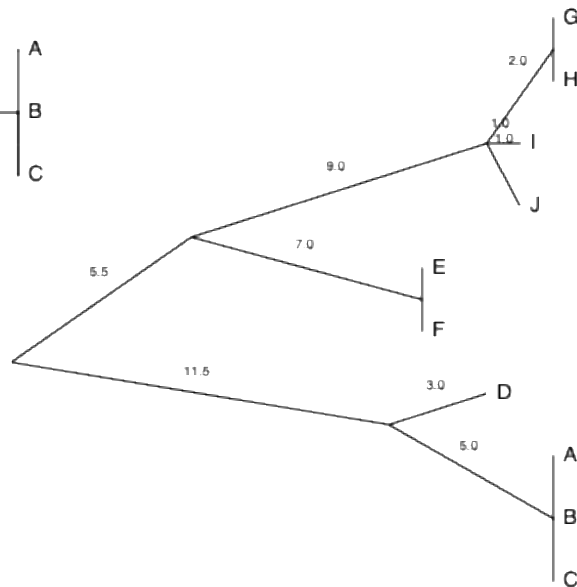
- scaled: branch length represents the amount of evolutionary change, but does not provide any indication of time

Tree Styles



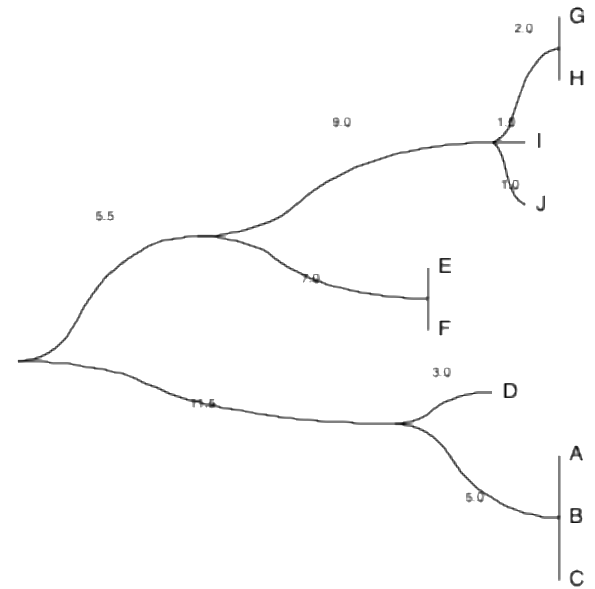
5

rectangular



5

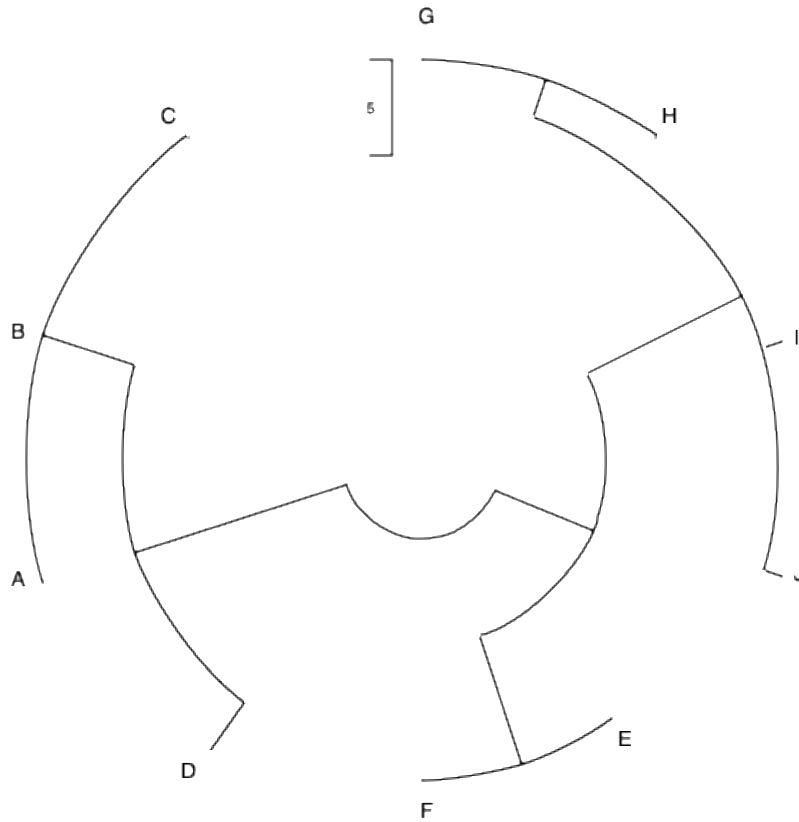
straight



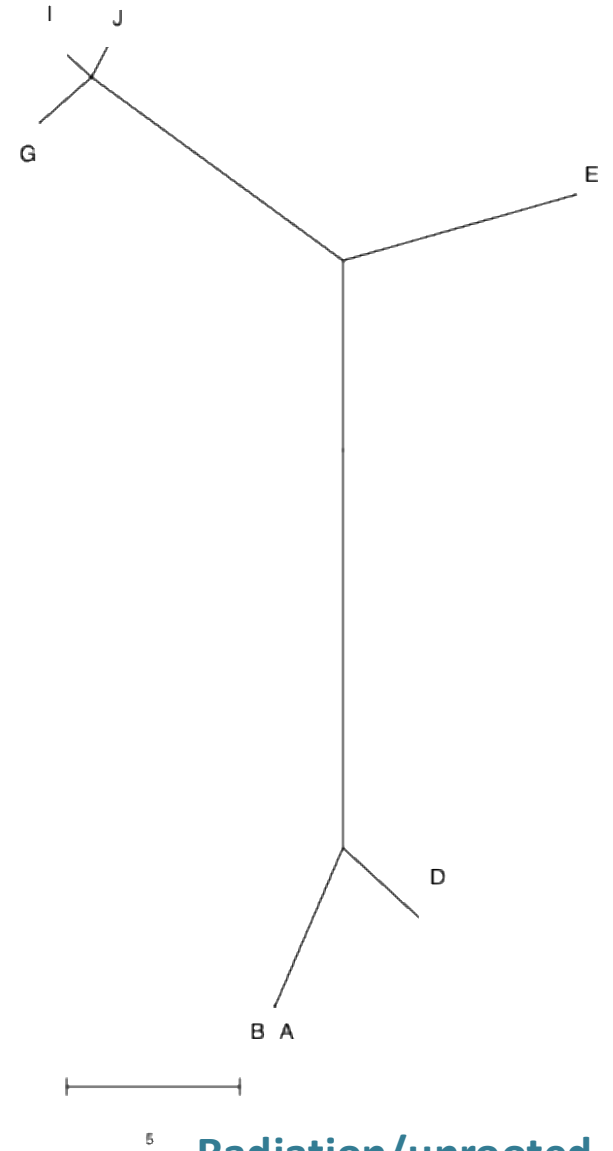
5

curved

Tree Styles



circular

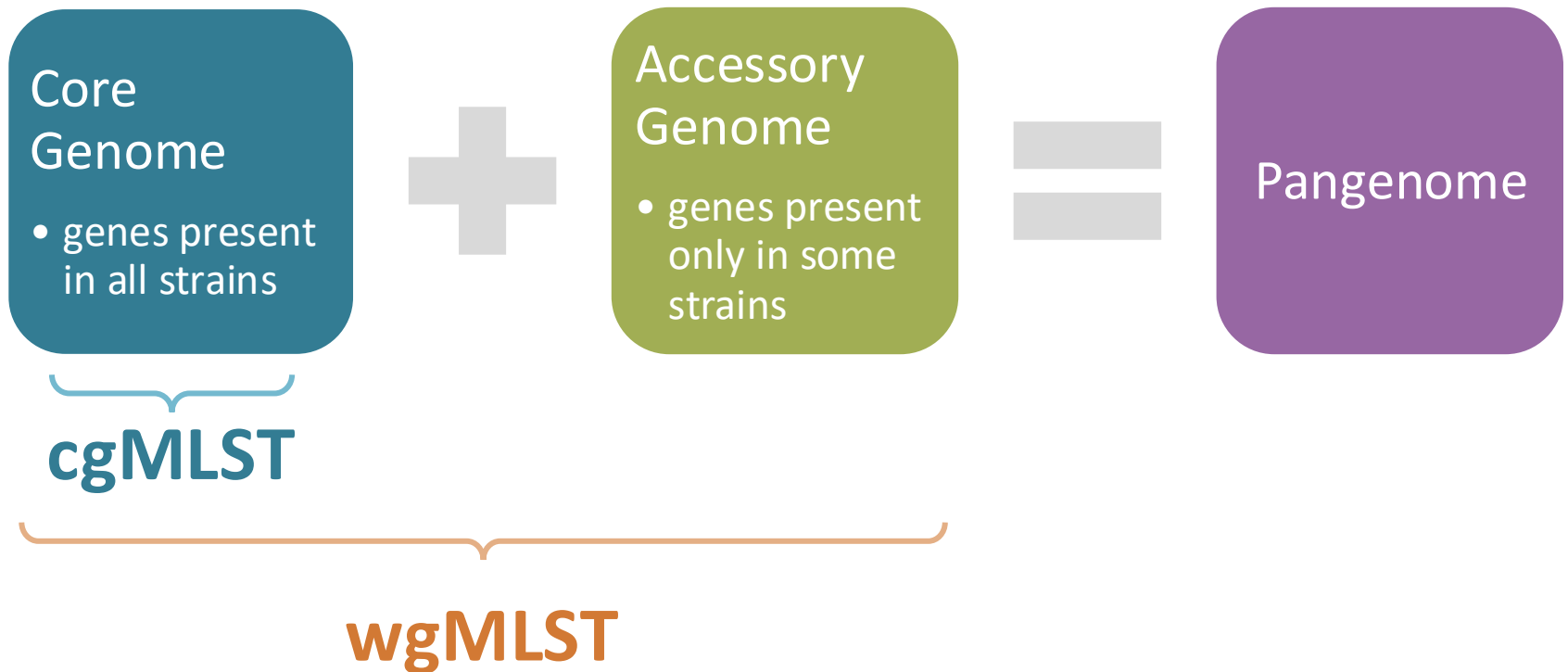


Radiation/unrooted



wgMLST- & cgMLST-based Analyses

wgMLST and cgMLST



wgMLST and cgMLST

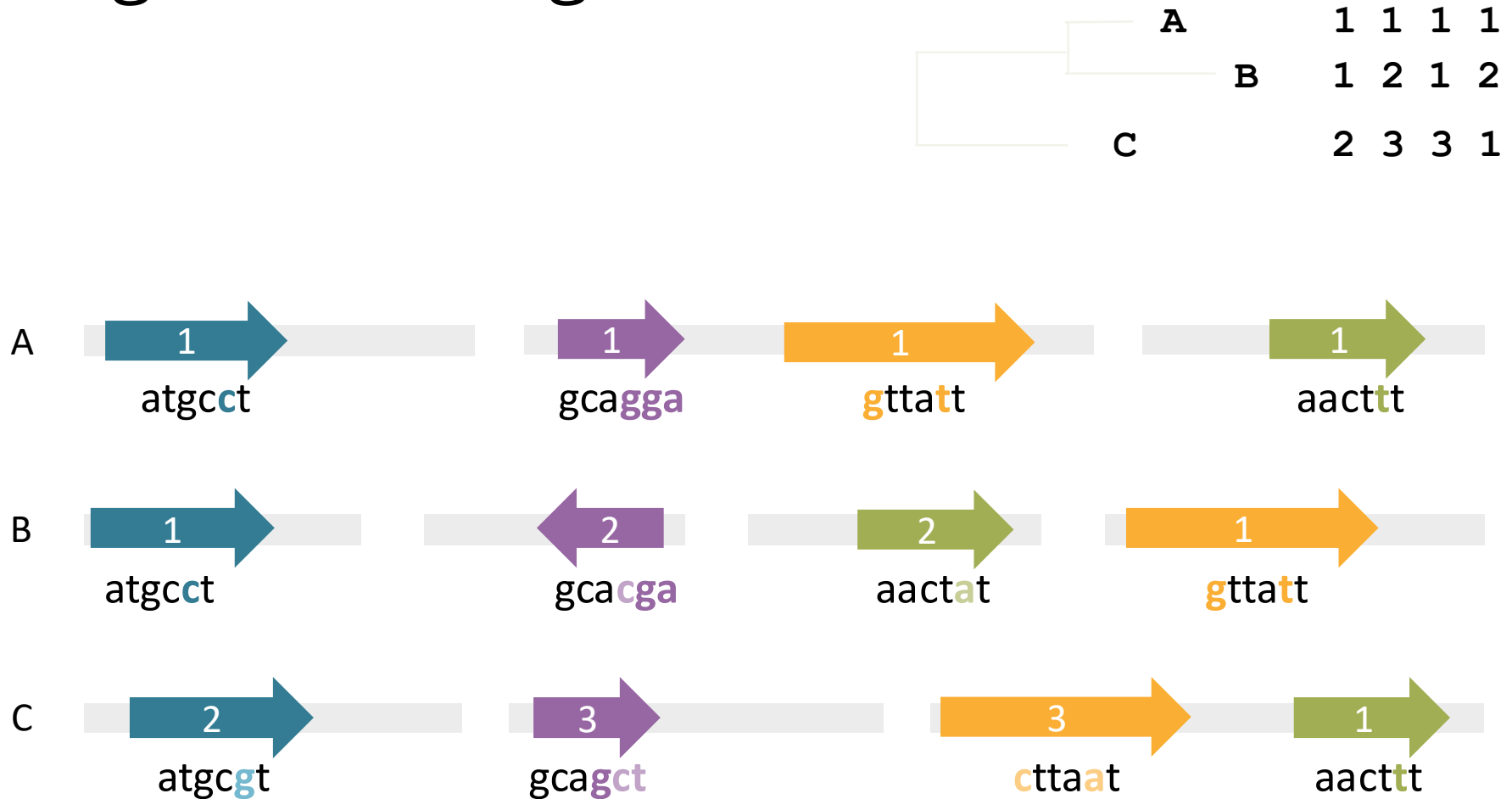
1. All of the loci are compared against a database of known loci
2. For each loci,
 - a. if it **matches** a known allele, it is assigned an identifier
 - b. if it **does not match** a known allele, it is given a new identifier
3. All of the loci between different genomes can be compared and a distance is calculated, and can be used to infer phylogeny

Based on*:

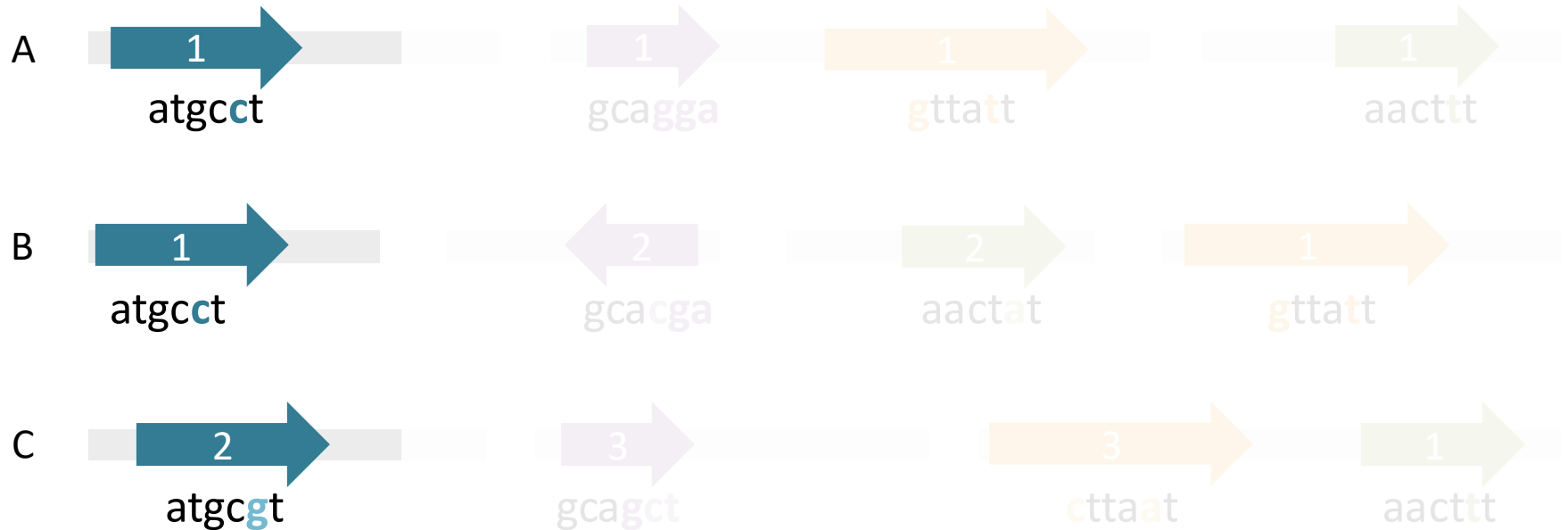
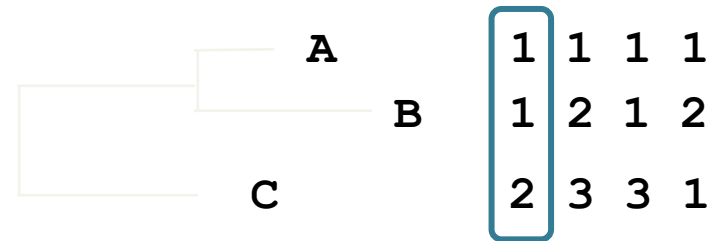
- 4,804 loci (wgMLST)
- 1,791 loci (cgMLST)

* For *L. monocytogenes* specifically

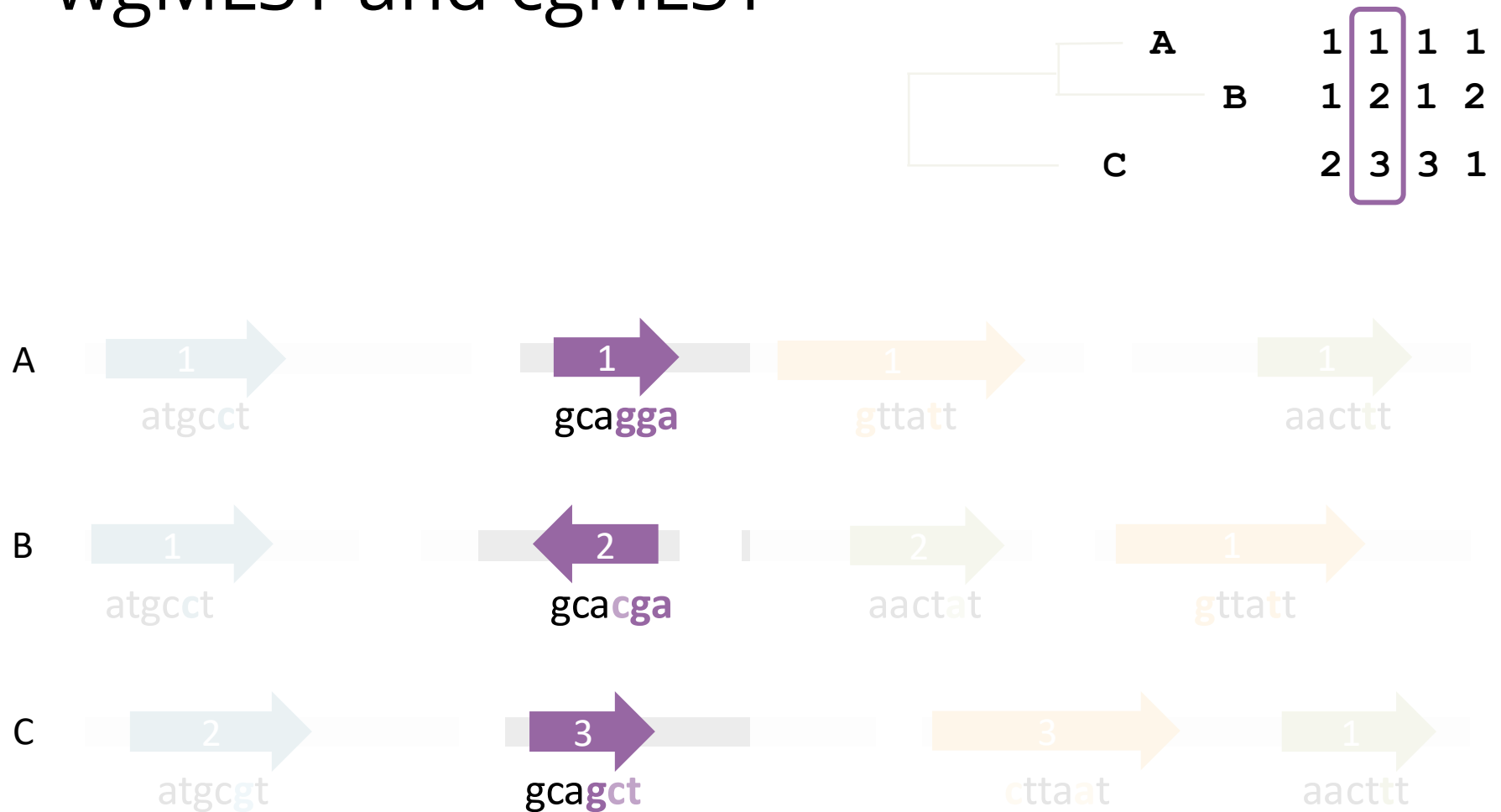
wgMLST and cgMLST



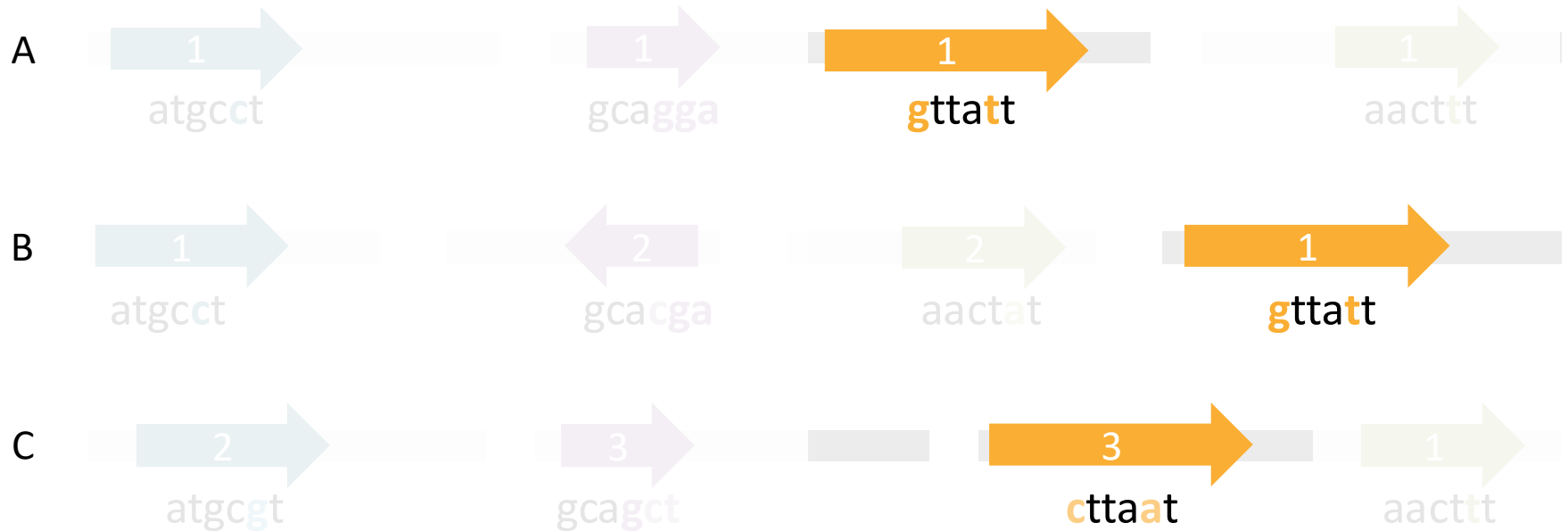
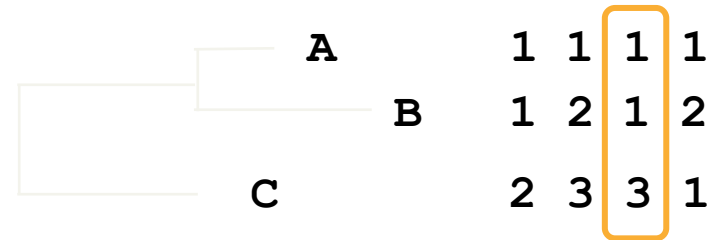
wgMLST and cgMLST



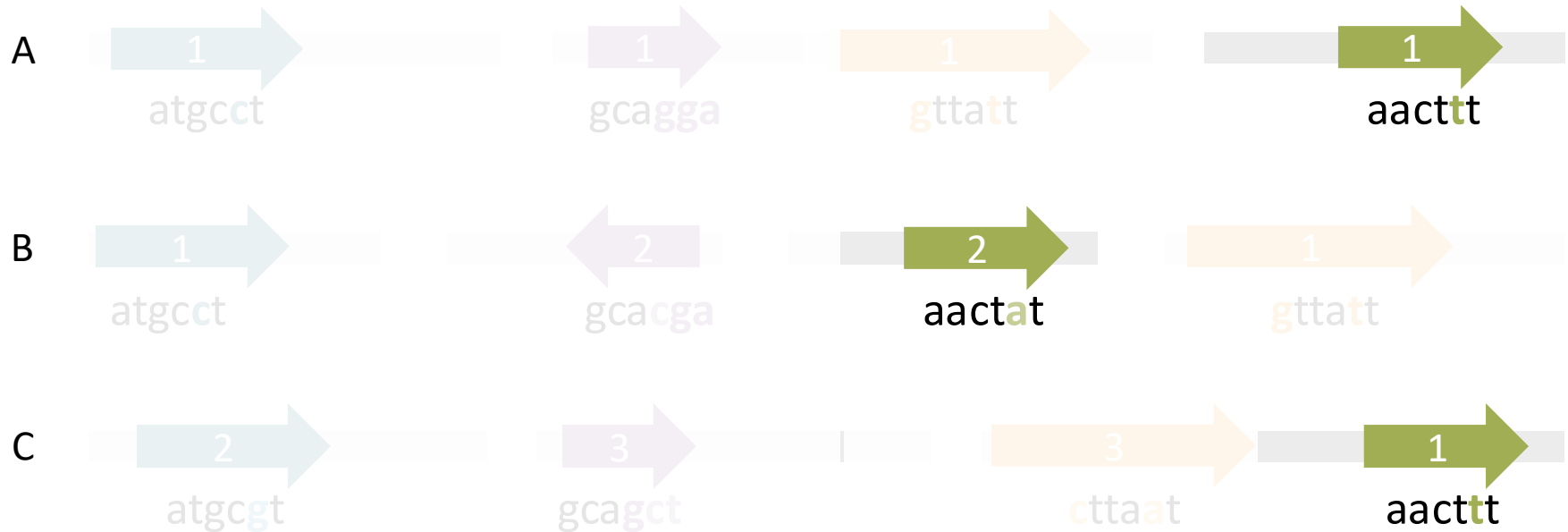
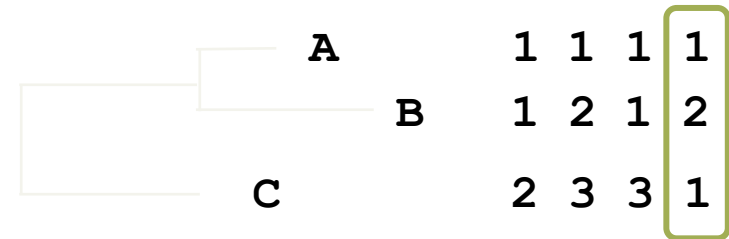
wgMLST and cgMLST



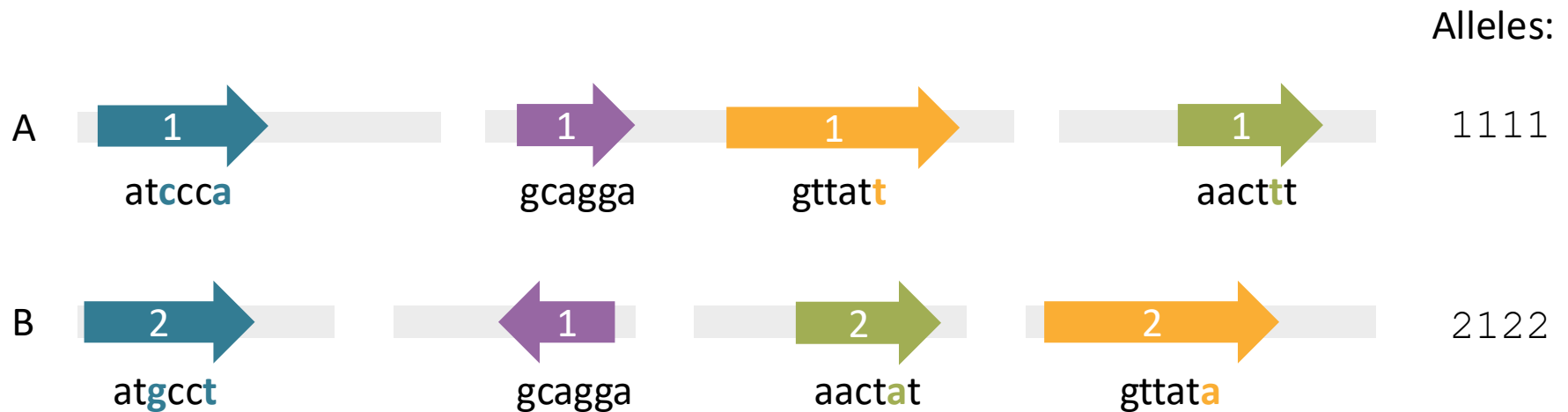
wgMLST and cgMLST



wgMLST and cgMLST

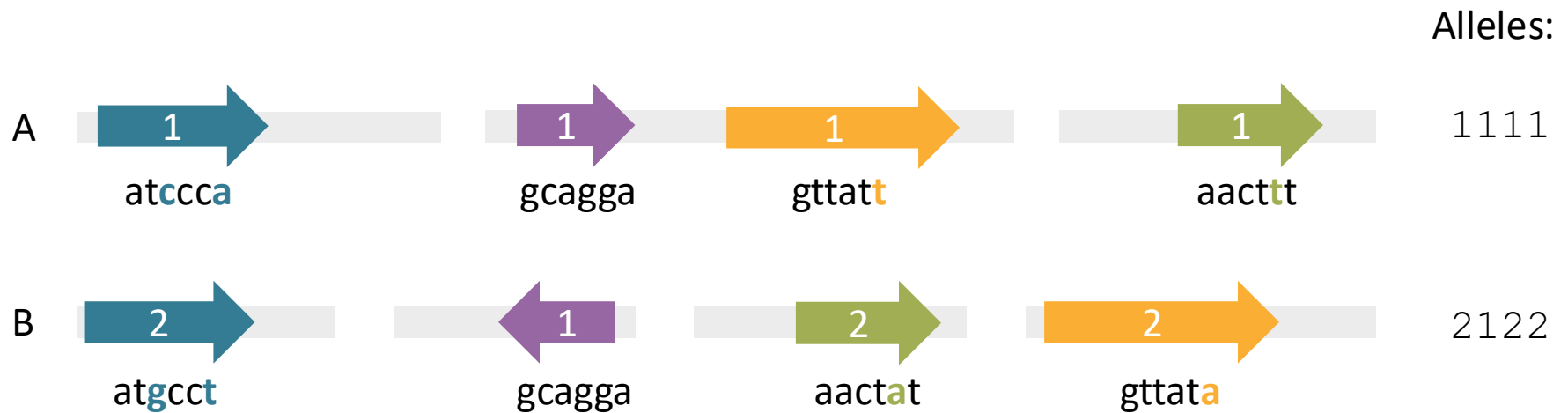


wgMLST and cgMLST



Knowledge Check: What is the allele distance between isolate A and isolate B?

wgMLST and cgMLST



Knowledge Check: What is the allele distance between isolate A and isolate B?

3 alleles

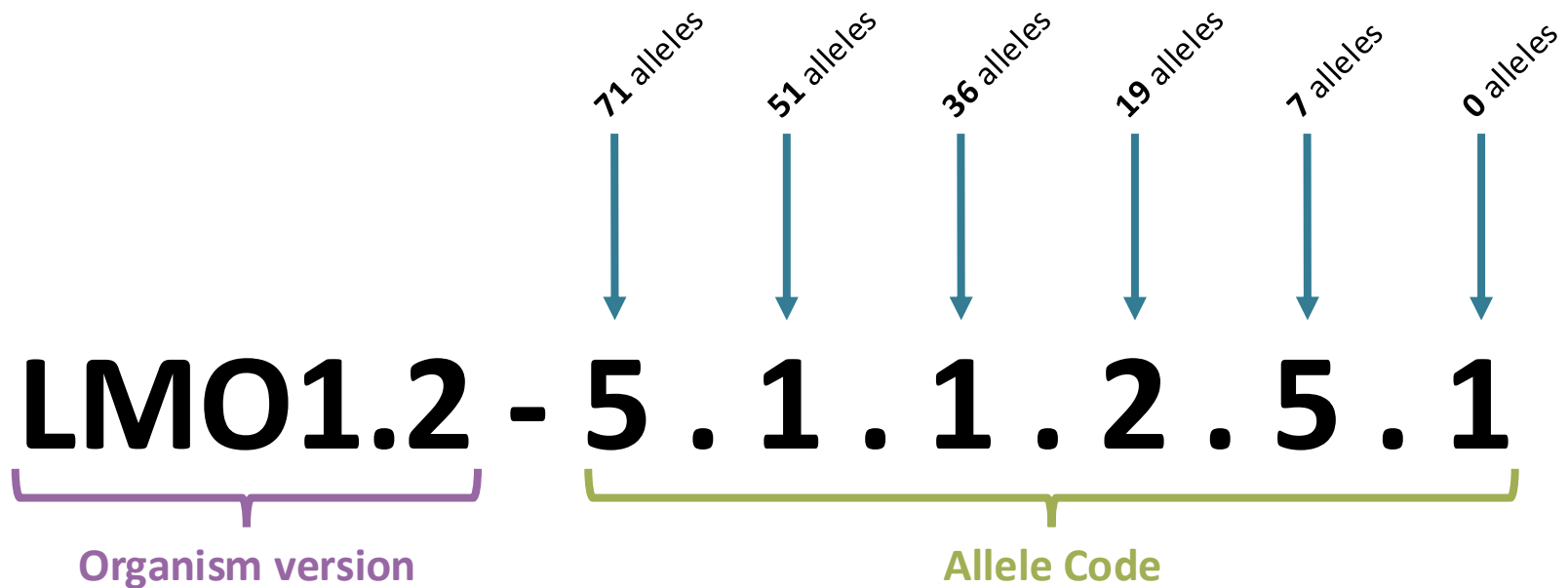


Allele Codes

Allele Codes

- Hierarchical naming system
- Show relatedness between isolates based on cgMLST
- Offer a compact view of the entire population structure for an organism database
 - only *Salmonella*, STEC, *Listeria*, *Campylobacter jejuni*
- must think about allele codes in terms of population, not outbreaks
 - allele codes are not outbreak codes
- May be used to help identify clusters
- May be used as a form of nomenclature for referencing strains
- Can be complete or partial depending on how they relate on the tree from which the nomenclature was built

Allele Codes



Allele Codes

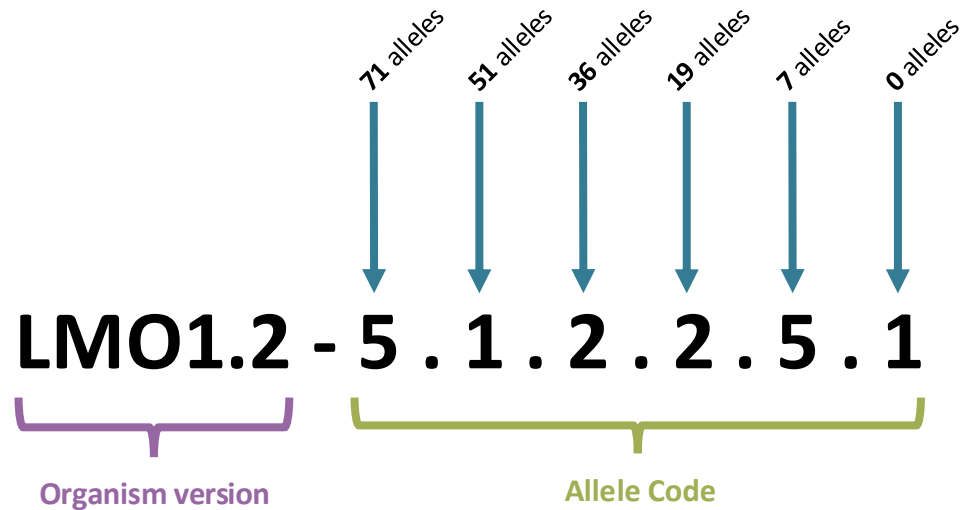
- Allele codes & thresholds of relatedness vary by organism

Organism		Allele Code Format										
<i>Campylobacter</i>	CAMP1.1 -	#	.	#	.	#	.	#	.	#	.	#
		84		61		24		14		5		0
<i>Escherichia</i> (STEC only)	EC1.1 -	#	.	#	.	#	.	#	.	#		
		77		51		16		6		0		
<i>Listeria</i>	LMO1.2 -	#	.	#	.	#	.	#	.	#	.	#
		71		51		36		19		7		0
<i>Salmonella</i>	SALM1.1 -	#	.	#	.	#	.	#	.	#	.	#
		80		28		15		7		4		0

Allele Codes

- In general, more numbers in common = more closely related
- Allele code thresholds are approximate for all organisms
- Due to the clonal lineage of some serotypes or strains, the approximate differences may be larger than expected
- Allele differences should always be verified with a dendrogram

Allele Codes

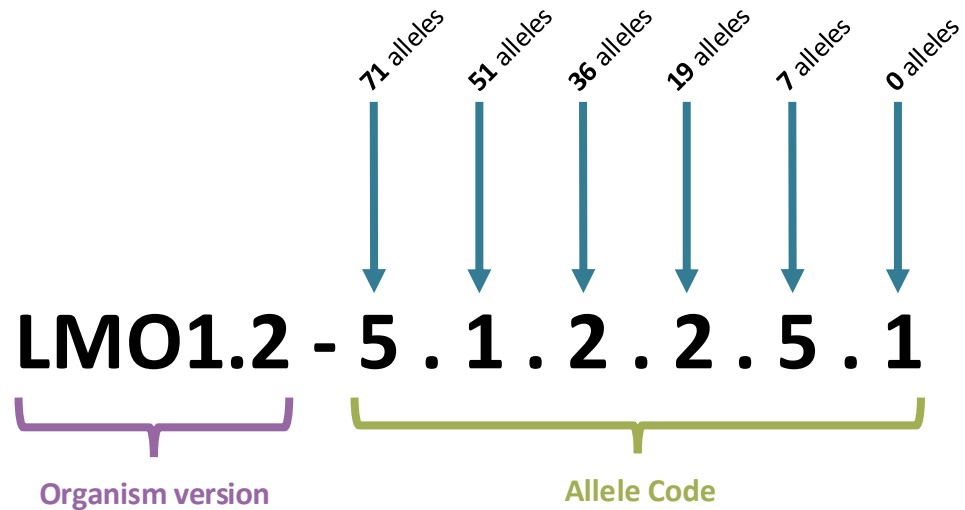


A:

LMO1.2 -

Singleton: No close matches, name not assigned.

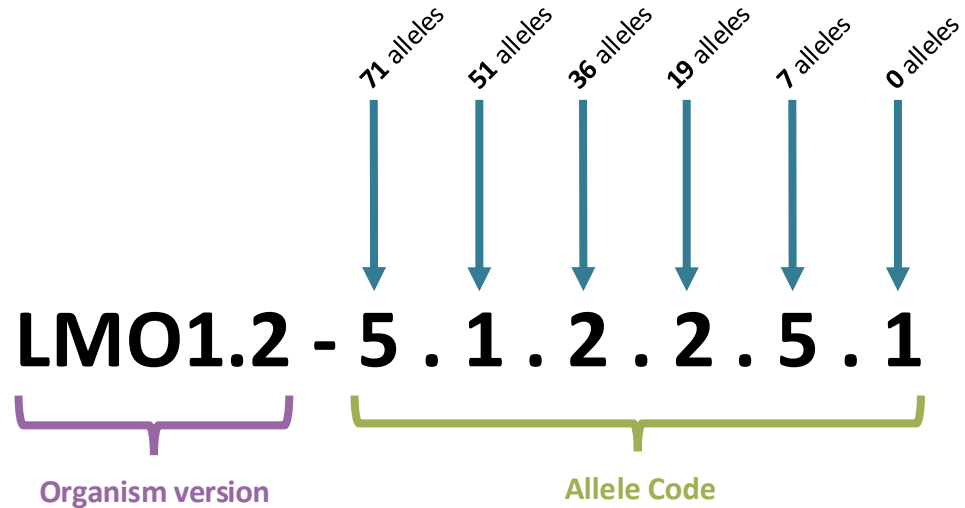
Allele Codes



- A: **LMO1.2 - 5 . 1 . 2**
- B: **LMO1.2 - 5 . 1 . 2 . 2 . 5 . 1**

When sequences have partial names, it means they are singletons in clusters below their last digit.

Allele Codes

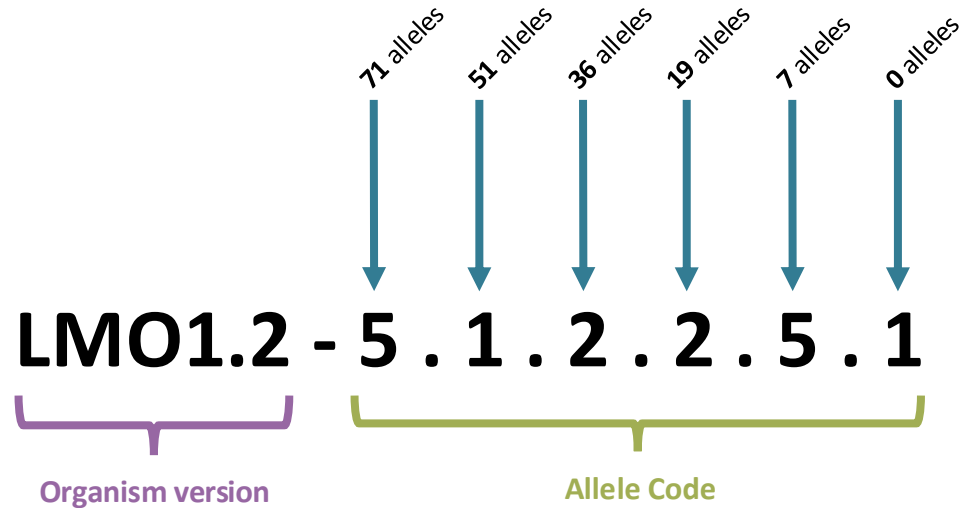


A: **LMO1.2 - 5 . 1 . 2**

B: **LMO1.2 - 5 . 1 . 2 . 2 . 5 . 1**

These sequences are approximately within 36 to 19 alleles of each other.

Allele Codes

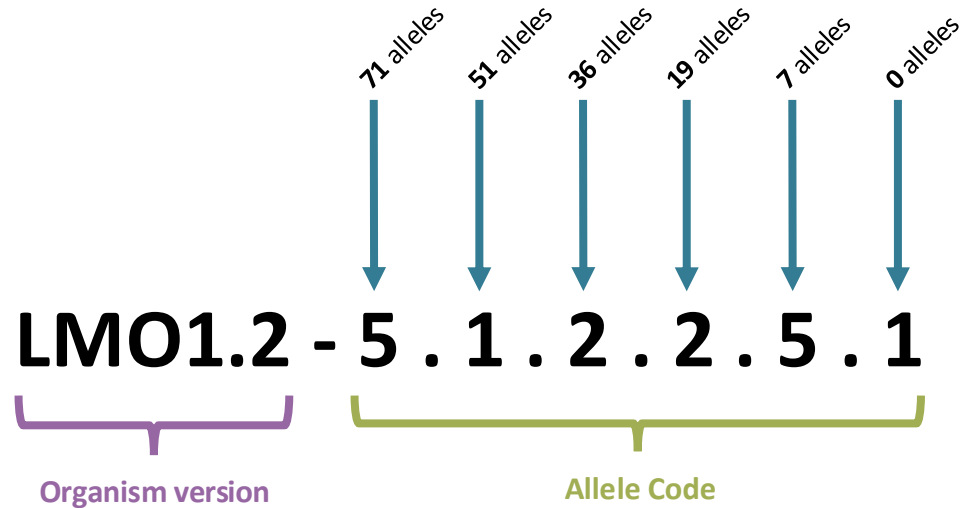


A: **LMO1.2 - 26 . 1 . 1 . 1 . 1 . 1**

B: **LMO1.2 - 26 . 1 . 1 . 1 . 1 . 1**

These two strains are indistinguishable (0 alleles different) based on the core genome.

Allele Codes



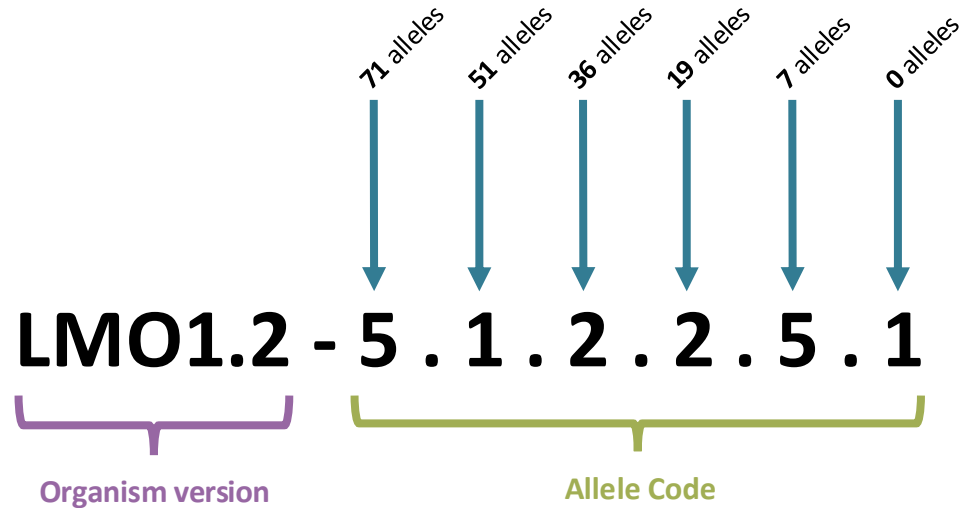
A: **LMO1.2 - 1 . 4 . 64 . 1 . 1 . 2**

B: **LMO1.2 - 1 . 4 . 64 . 1 . 1 . 2**

C: **LMO1.2 - 1 . 4 . 64 . 1 . 1**

The two top strains are indistinguishable with 6 digits matching exactly.

Allele Codes



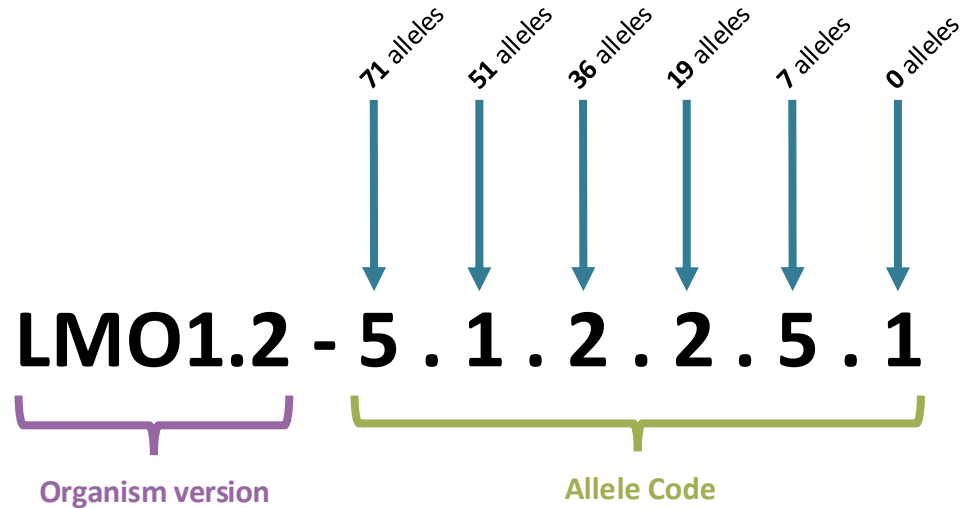
A: **LMO1.2 - 1 . 4 . 64 . 1 . 1 . 2**

B: **LMO1.2 - 1 . 4 . 64 . 1 . 1 . 2**

C: **LMO1.2 - 1 . 4 . 64 . 1 . 1**

The bottom one is missing the 6th digit, so it relates to the other two within 7 alleles.

Allele Codes



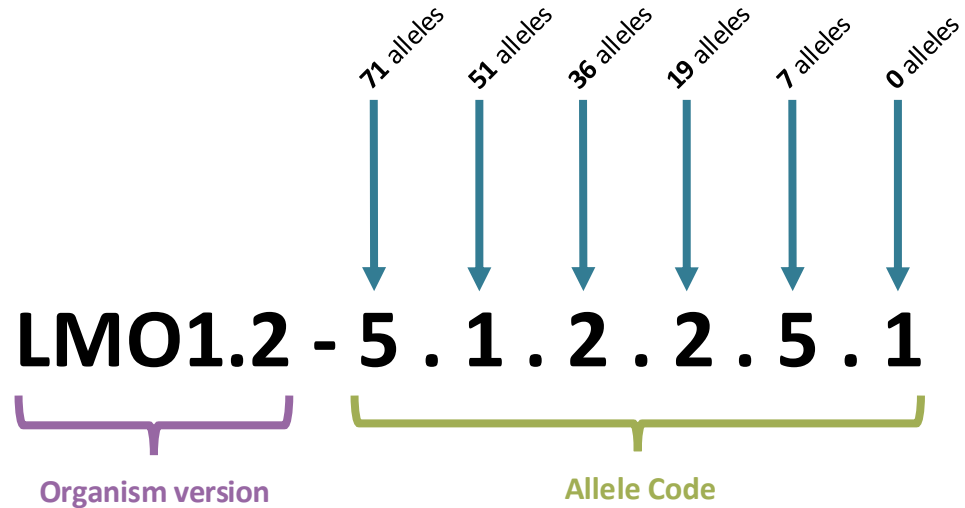
A: **LMO1.2 - 26 . 1 . 1 . 1 . 1 . 1**

B: **LMO1.2 - 26 . 1 . 1 . 1 . 2 . 8**



Knowledge Check: What is the allele difference range between isolate A and and isolate B?

Allele Codes



A: **LMO1.2 - 26 . 1 . 1 . 1 . 1 . 1**

B: **LMO1.2 - 26 . 1 . 1 . 1 . 2 . 8**



Knowledge Check: What is the allele difference range between isolate A and and isolate B?

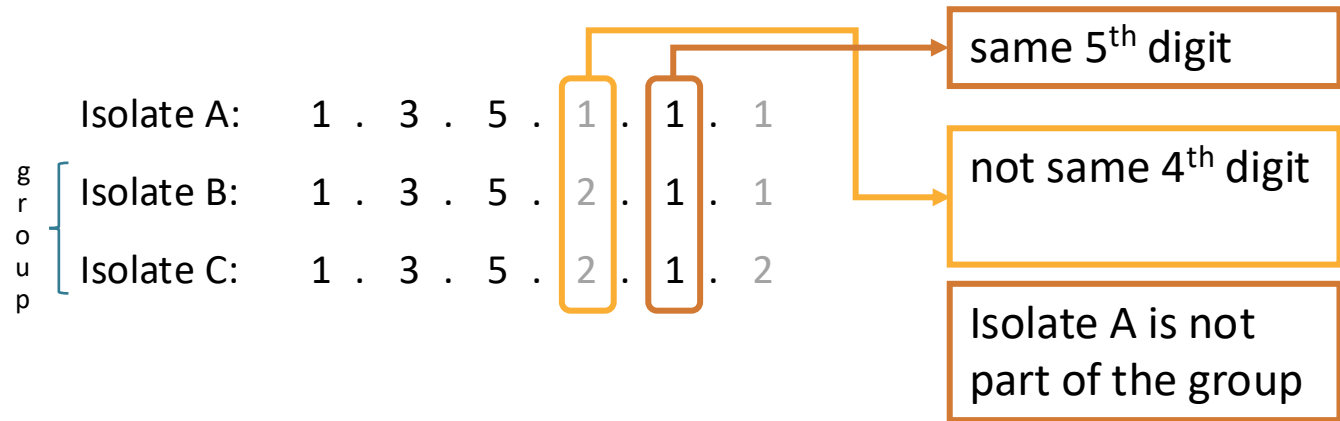
7-19 alleles different

Updated Allele Codes

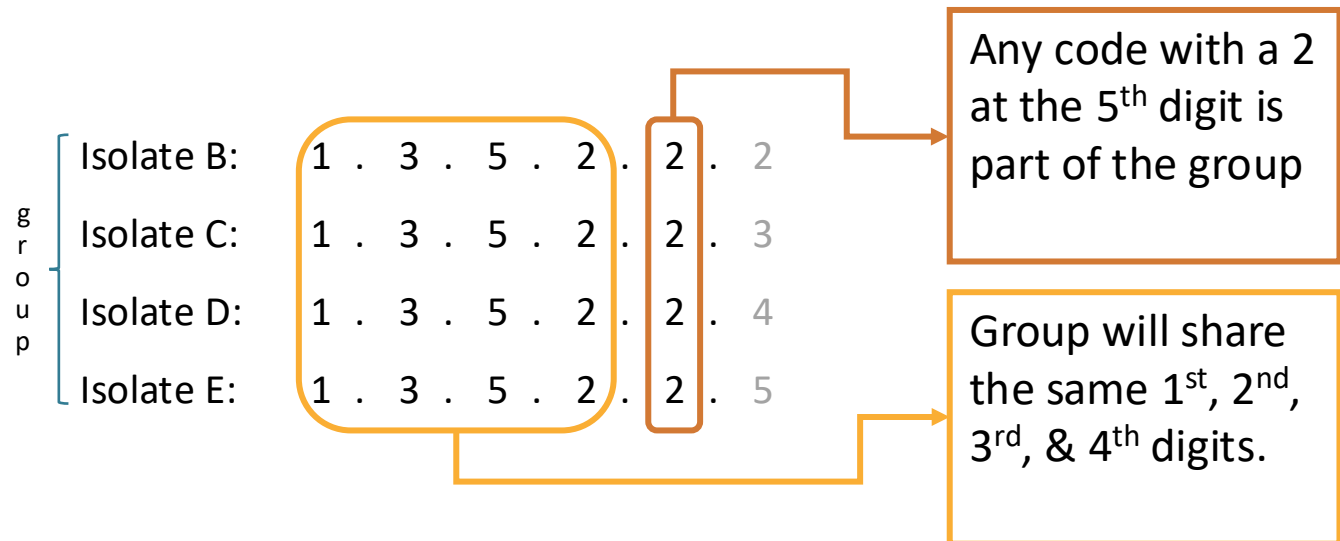
- New version released with PulseNet2.0
- Available for all organisms that currently have an allele code naming system
 - i.e., *Salmonella*, STEC (not *Shigella* or non-STEC *E. coli*), *Listeria*, *Campylobacter jejuni*
- Unique digits at each threshold
 - longer numbers at each digit/threshold
 - E.g., SALM1.1 - 1.32.101.11234.12323
 - Clades for a specific threshold can be referenced by a single digit rather than the entire code, i.e., "*Salmonella* allele code 11234 at the 4th digit"

Updated Allele Codes

Previous Version



New Version



Allele Code Limitations

- No allele codes for *Shigella*, so cannot use the Cluster Detection tool for this pathogen
 - Only *C. jejuni* – cannot do other species (e.g., *C. coli*, etc)
- Certain search parameters result in limited number of clusters available for viewing at once, so you have to narrow down dates, number of isolates in a cluster, and allele thresholds. Can be especially hard to see complete list of clusters during the busy enteric season; may lead to missed clusters/outbreaks
- Have to check each cluster each week to see if new isolates have been added (versus dendrogram can just look at everything at once)
- NCBI may need to be used, especially when there is a clusters with x codes, to check SNP distance and determine which isolates should be included in the cluster and which are genetically too distant

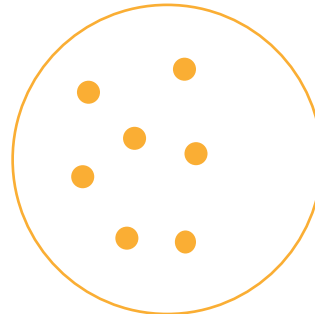


X Codes

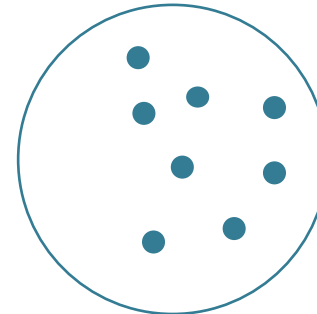
Allele code merging/chaining

Orange cluster:
isolates up to **7**
alleles different

(Same 4th digit)



SALM1.2 – 5.1.1.2



SALM1.2 – 5.1.1.3

Blue cluster:
isolates up to **7**
alleles different

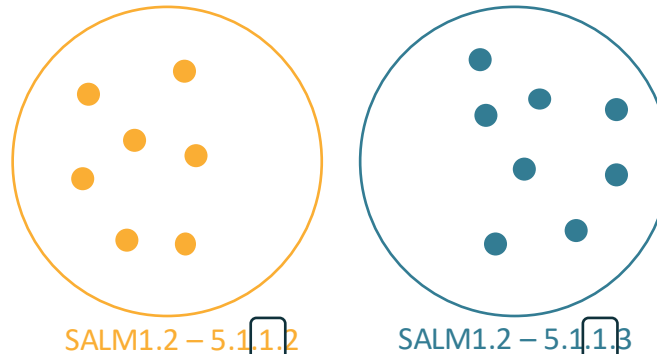
(Same 4th digit)

Allele code merging/chaining

Orange cluster:

isolates up to **7**
alleles different

(Same 4th digit)



Blue cluster:

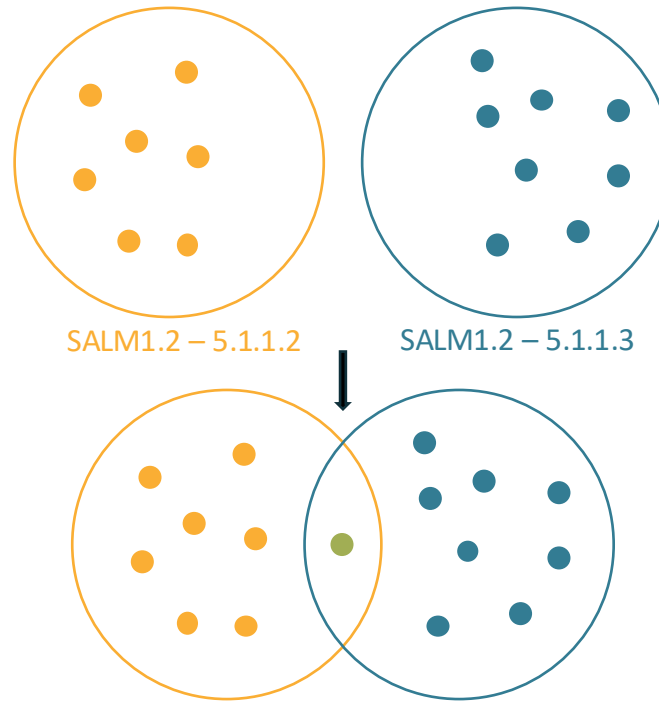
isolates up to **7**
alleles different

(Same 4th digit)

Isolates from **orange cluster** up
to **15** alleles different
from isolates from **blue cluster**
(and vice versa)

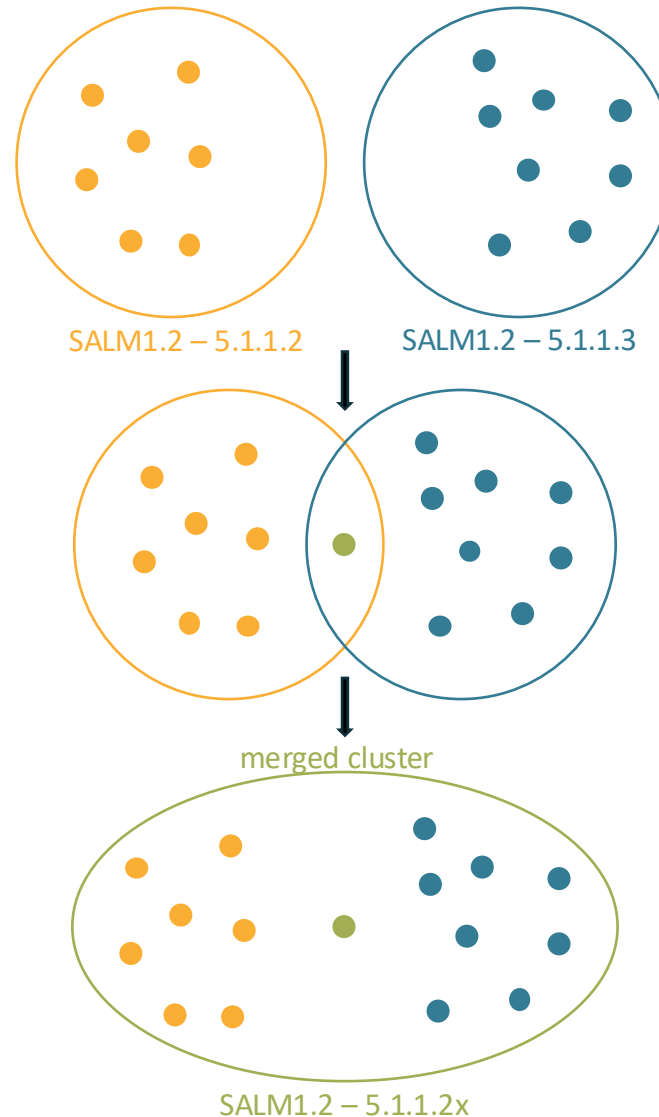
(Same 3rd digit)

Allele code merging/chaining



New **green isolate** causes a merge of the **orange** & **blue** clusters

Allele code merging/chaining



After several merging events (“chaining”), new **combined cluster**:

contains isolates **7-15** alleles different

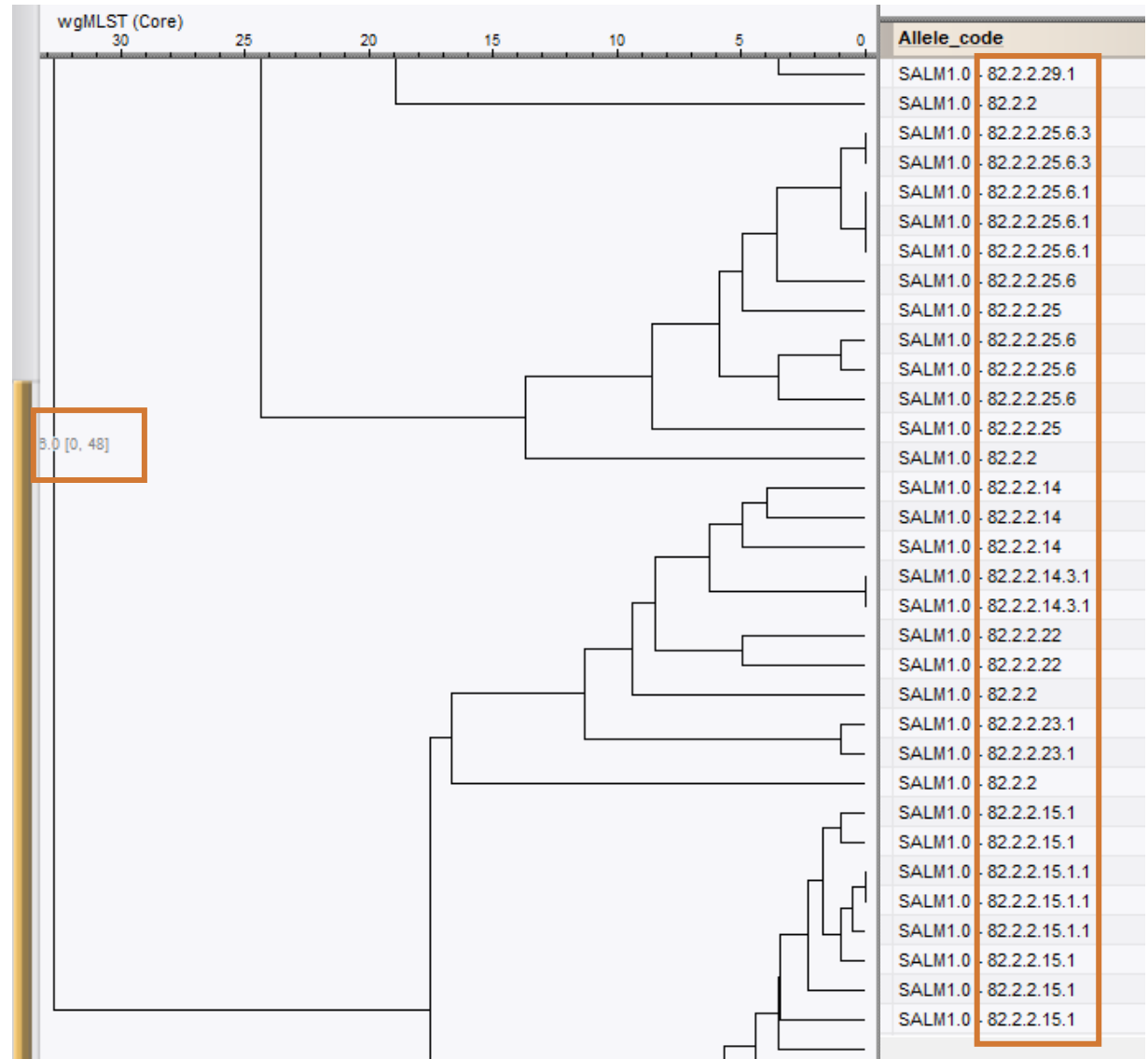
all share 4th digit, which would normally indicate up to **7** alleles different

X codes

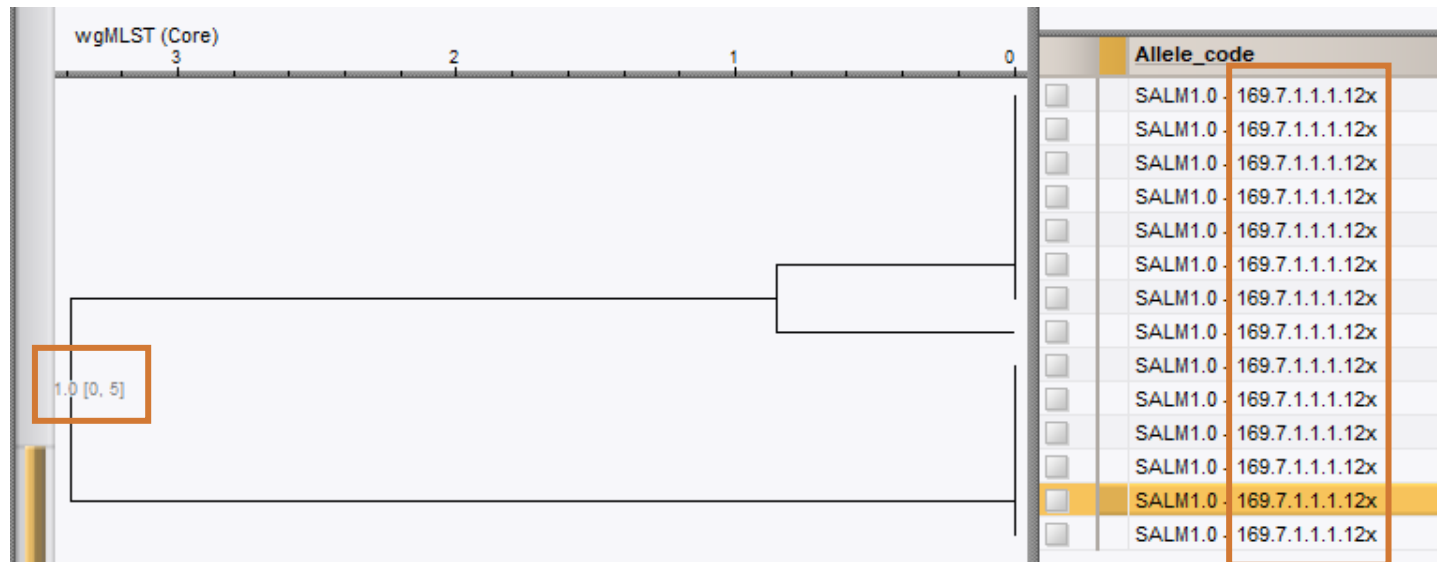
- Groups of isolates with differences >~4x the expected threshold are given X codes
 - Some clonal lineages merge together at a shared digit on the single linkage tree, creating more allele differences than you would expect at that threshold
- E.g., SALM1.1 – 6.6.6.6x.8x.32782
- primarily in *Salmonella* database, some in *Escherichia* (STEC only)
- Due to chaining, allele codes cannot be used to determine exact allele differences
 - Verify allele differences with dendrogram or matrix
 - Use NCBI to determine if isolates are closely related

X codes

- All isolate codes share first 3 digits
- Expected allele differences: up to **15**
- Actual allele differences: **0-48**
- Differences do not meet 4x threshold, not assigned an X code

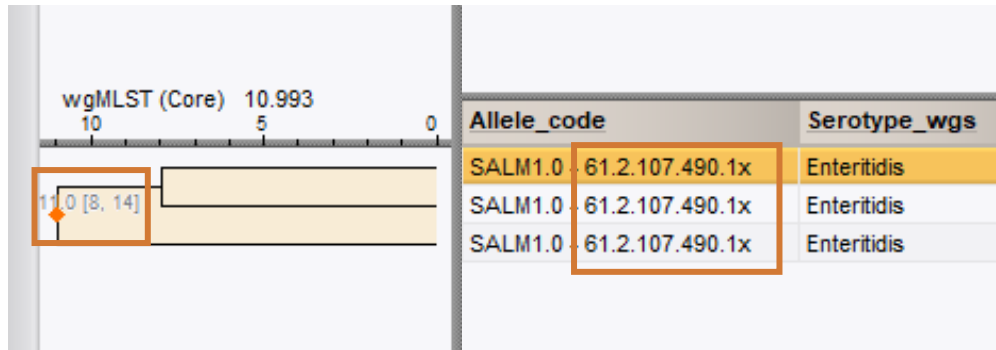


X codes



- All isolate codes share all 6 digits
- Expected allele differences: **0** (indistinguishable)
- Actual allele differences: **0-5**
- Differences >4x expected threshold, “trailing X” added to code

X codes



- All isolate codes share first 5 digits
- Expected allele differences: **1-4**
- Actual allele differences: **8-14**
- Not a cluster!

Allele differences for X codes should always be verified!

Updated X Codes

- Improvements in PulseNet2.0
- may continue past the X with plain numbers or additional numbers with an X.
 - no longer truncated at the first X
- may be multiple Xs in the same code
 - E.g., SALM1.1 - 6778.3.1.1x.16x.16.
- allow users to detect smaller clusters of closely related sequences within an overall X code

Updated X Codes

Previous Version

SALM1.0 - 6778 . 3 . 1 .

1x

Only assigned at 4th or 5th digit

stopped after the x (truncated after the digit to which the x was added)

- Manually assigned monthly

New Version

SALM1.1 - 6778 . 3 . 1 . 1x . 105 . 2

SALM1.1 - 6778 . 3 . 1 . 1x . 16x . 16

Assigned to all digits >4x their expected allelic threshold

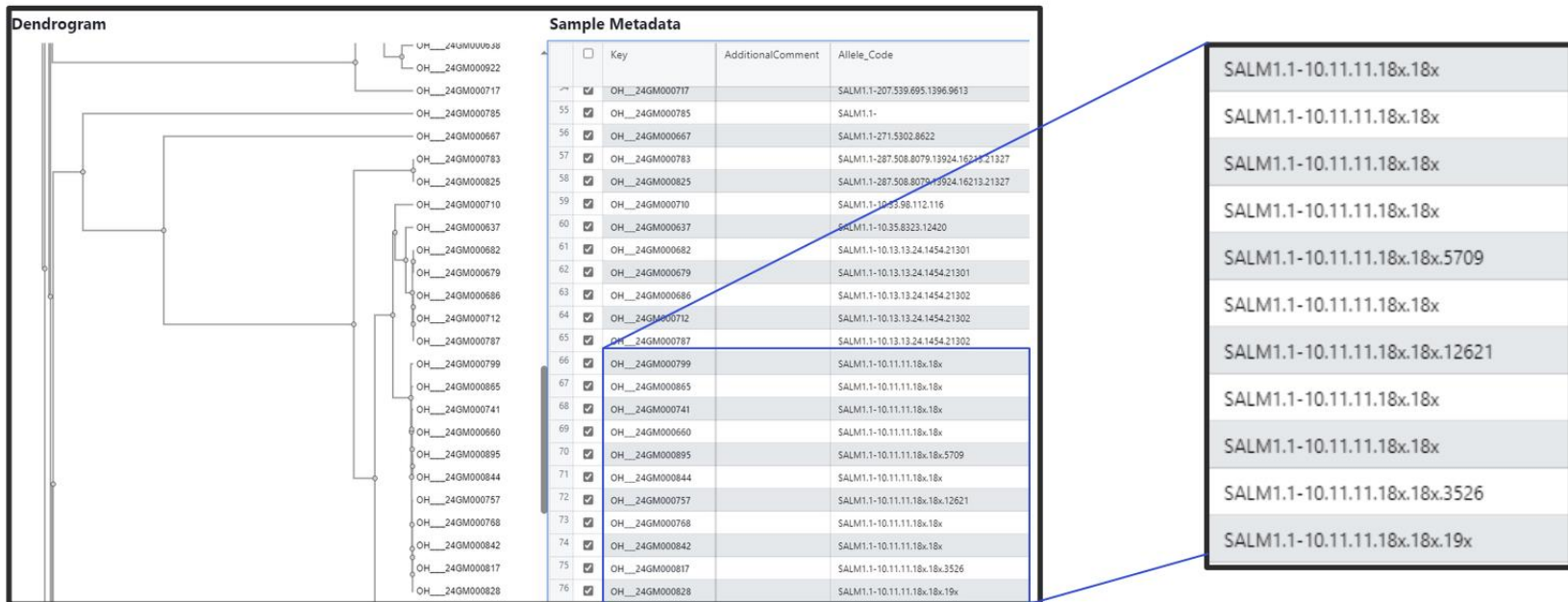
Can continue past the x

Can have more than one x if multiple digits exceed expected threshold

- Automatically assigned

Updated X Codes

- continue after the X, making it easier to find clusters of more closely related sequences within an overall X code
- Allows for more accurate cluster detection at lower thresholds



00:00



Break



Thresholds for Cluster Detection

	Local	National
<i>Campylobacter</i>	≥3 cases differing within 10 alleles (cgMLST), of which 2 differ within 5 alleles. Isolation dates for all should be within the past 60 days of detection	≥5 cases within 10 alleles (cgMLST) uploaded within the past 60 days, ≥50% must have isolation dates within the last 30 days
<i>Escherichia</i>	≥3 cases differing within 10 alleles (cgMLST), of which 2 differ within 5 alleles. Isolation dates for all should be within the past 60 days of detection	≥5 cases differing within 10 alleles (cgMLST), of with ≥3 within 5 alleles uploaded within the past 30 days. ≥30% must have isolation dates within the past 50 days.
<i>Salmonella</i>	≥3 cases differing within 10 alleles (cgMLST), of which 2 differ within 5 alleles. Isolation dates for all should be within the past 60 days of detection	≥10 cases within 10 alleles (cgMLST), with ≥3 within 5 alleles Uploaded within the past 60 days (≥30% must have isolation dates within the past 50 days)
<i>Shigella</i>		Same as <i>Escherichia</i>, but may include cases within a wider allele range due to person-to-person transmission
<i>Listeria</i>	≥3 cases differing within 10 alleles (wgMLST), of which 2 differ within 5 alleles or matching up to the 5th allele code digit	≥3 cases within 25 alleles (wgMLST), with ≥2 within 10 alleles uploaded within the past 120 days. Historical isolates within 25 alleles also included

Thresholds for Cluster Detection

- Allele differences within a cluster may be larger or smaller depending on the organism and epi data
 - There can be similar strains by WGS that may not be epidemiologically linked
 - More clonal species/serotypes may have smaller allele differences
 - Zoonotic outbreaks may have larger allele differences

Thresholds for Cluster Detection



Flour

- Max allele difference: 4
- Isolation date range: ~5 months



Chicken

- Max allele difference: 11
- Isolation date range: ~3 months



REP Strains

What are REP strains?

“Group of bacteria, closely related by whole genome sequencing, that continues to cause illness over time”

- Intentionally vague

Since 2019, CDC has used the PulseNet system to identify and monitor REP strains of several major pathogens, including:

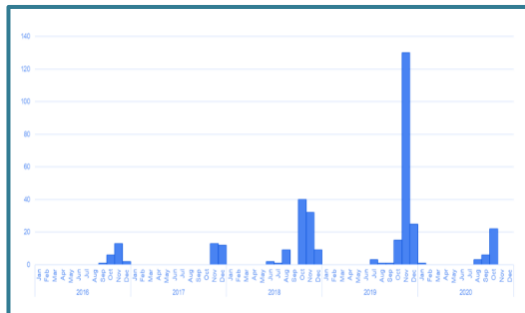
- *Salmonella enterica*
- *E. coli* O157:H7
- *Listeria monocytogenes*
- *Campylobacter jejuni*

What are REP strains?

R

Reoccurring

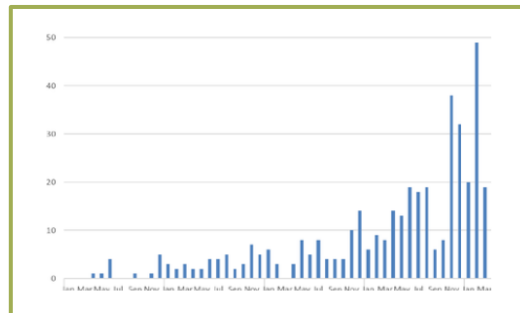
Strains repeatedly causing acute outbreaks separated by periods during which the strain either does not make people sick or causes illnesses at very low levels.



E

Emerging

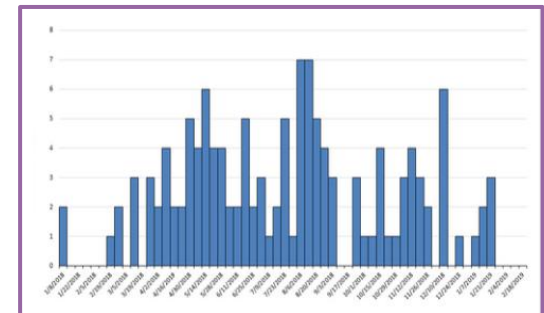
Previously novel or rare strains that increasingly cause illnesses or show the potential to cause increasing illness.



P

Persisting

Strains causing illnesses consistently over a long time.



Identifying & Tracking REP Strains

- PulseNet uses a combination of different WGS tools for identifying and tracking these strains and will develop more as new strains are identified and investigated.
- Identifying a strain as a REP strain is based on many factors, including:
 - the number of illnesses
 - whether illnesses are increasing
 - the characteristics of the strain (e.g., multidrug resistance, high virulence, increased transmissibility)
 - whether the strain is the cause of large or frequent outbreaks

REP Codes

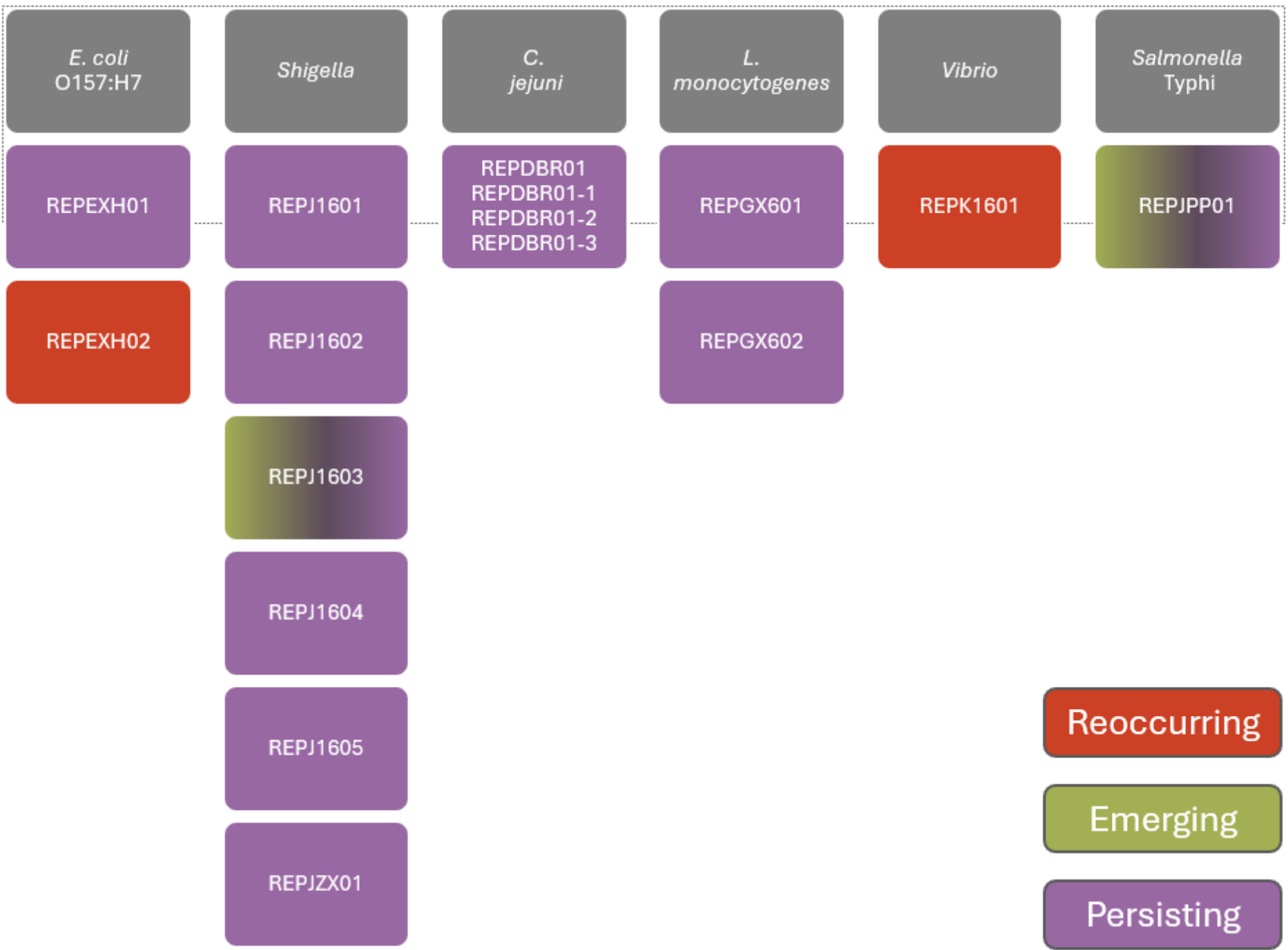
REP EXH 01 - 4

signifies that an isolate is part of a REP strain

three letter/digit serotype code

sequential two digit number for REPs within each serotype

single digit that signifies a sub-group or clade within that REP (optional)



nontyphoidal *Salmonella*

Blockley

Enteritidis

Hadar

Infantis

Kentucky

Newport

Poona

Reading

Stanley

Typhimurium

REPJBG01

REPJEG01

REPTDK01

REPJFX01

REPJGP01

REPJJP01

REPJL601

REPJLG01

REPJNG01

REPJPX01

REPJJP02

REPJLG02

REPJPX02

REPJJP03

REPJPX03

REPJJP04

Reoccurring

Emerging

Persisting

Emerging Illness – REPJPP01

HIGHLIGHTS

- REPJPP01 is an emerging strain of extensively drug-resistant *Salmonella* Typhi, the bacteria that cause typhoid fever. This strain has caused illnesses in the United States since 2018.
- REPJPP01 has caused illness in travelers to Pakistan and neighboring countries and in people who did not leave the United States.
- Getting vaccinated against typhoid fever before international travel – especially to South Asia – and following safe eating and drinking habits while traveling can prevent illness.

What makes this “emerging”?

- ID’ed in Pakistan, now found in more countries including US not related to travel
- Antibiotic resistance – ceftriaxone initially, now some XDR

Did you know: bacteria in REPJPP01 are within 13 alleles?

Persistent Illness – REPEXH01

HIGHLIGHTS

- REPEXH01 is a persistent strain of Shiga toxin-producing *E. coli* O157:H7 bacteria that has caused illnesses and outbreaks in the United States.
- This strain has caused many outbreaks linked to different sources, such as recreational water, ground beef, and romaine lettuce.
- Additional research could help determine where this strain is coming from and how to prevent future illnesses.

What makes this “persistent”?

- Multiple outbreaks
- Variety of sources
- Unclear where strain is coming from

Did you know: REPEXH01 includes 634 PN cases and 14 outbreaks!

REP Strains



Knowledge Check: REP stands for:

- A. Repeating, Emerging, Persisting
- B. Repeating, Enteric, Persisting
- C. Reoccurring, Emerging, Persisting
- D. Reoccurring, Emerging, Problematic



Knowledge Check: REP codes include:

- A. Three letter or digit serotype code
- B. R, E, or P for type of REP code
- C. Two-digit number for strains within each serotype
- D. A and B
- E. A and C



Knowledge Check: REP strains can belong to more than one classification (i.e., reoccurring, emerging, persisting).

- A. True
- B. False

REP Strains



Knowledge Check: REP stands for:

- A. Repeating, Emerging, Persisting
- B. Repeating, Enteric, Persisting
- C. **Reoccurring, Emerging, Persisting**
- D. Reoccurring, Emerging, Problematic



Knowledge Check: REP codes include:

- A. Three letter or digit serotype code
- B. R, E, or P for type of REP code
- C. Two-digit number for strains within each serotype
- D. A and B
- E. **A and C**



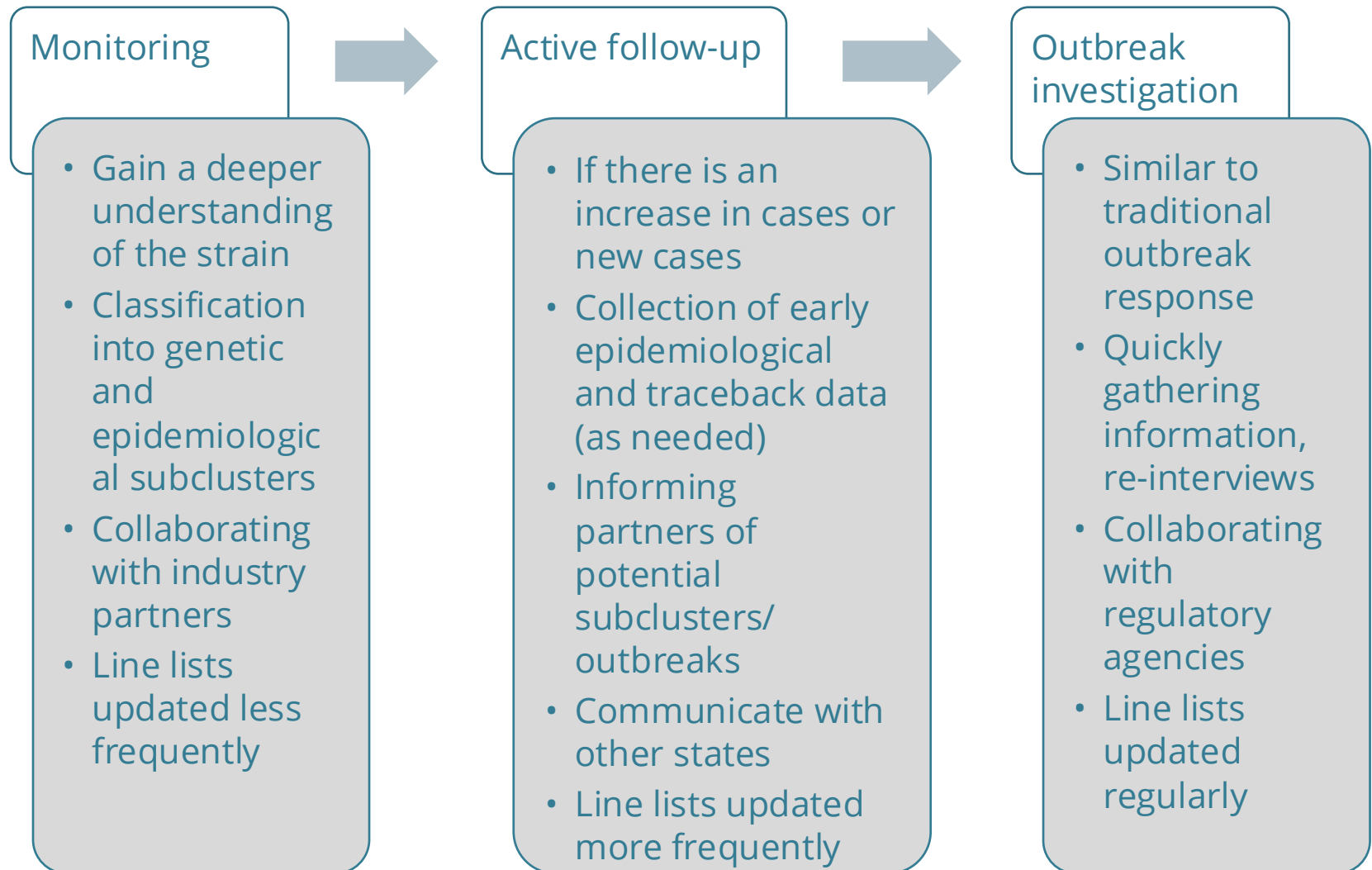
Knowledge Check: REP strains can belong to more than one classification (i.e., reoccurring, emerging, persisting).

- A. **True**
- B. False

Why does CDC track REP Strains?

- CDC uses information gained from investigations of REP strains to:
 - better understand their sources
 - track how they change over time
 - collaborate on measures to reduce their spread

Monitoring & Investigating REP Strains



Why Investigate REP Strains?

- Track how REP strains change over time
- Collaborate on measures to reduce spread
- Better understand transmission routes and sources
 - Prior investigation of settings, other factors
- Streamline future outbreak investigations involving the same strain
 - Guiding hypothesis generation
 - Create focused or supplemental questionnaires and/or to prioritize questions about related exposures during interviews

Documentation and tracking

- Multistate REP codes tracked via internal REDCap
 - With all other local/MS clusters
- TN Local REP codes are not tracked
 - For future local REP codes – will likely implement similar system
 - Do keep records and files in usual secure drive
 - No naming convention yet
- Other state testimonials

Next Sessions



Reminder: please complete survey that will automatically appear in your browser after you leave the Zoom meeting.

2

NCBI Pathogen Detection

- for cluster investigation (e.g., to find other cases)

SEDRIC

- for cluster detection & investigation (e.g., to find other cases, historical info. about rare serotypes)

PulseNet 2.0

- how to obtain access
- benefits

3

Hot Topic: Other Cluster Detection Tools

- SaTScan & TreeScan
- ESSENCE

Communication

- Communicating WGS results/findings in an investigation to stakeholders
- Standard phrasing

Questions/Discussion

