Whole Genome Sequencing for Cluster Detection and Investigation
Live Learning Series 2025

Session I
WGS Fundamentals

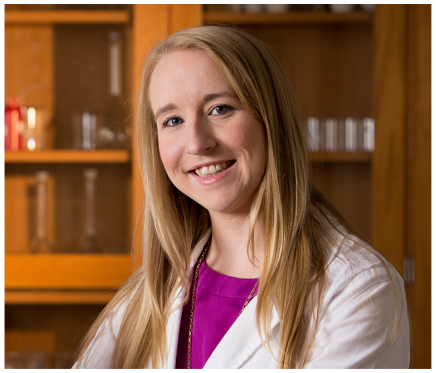
Tuesday, June 17th, 2025

1-2:30 PM ET

Integrated Food Safety
Centers of Excellence

Introductions

Lauren Hudson, PhD



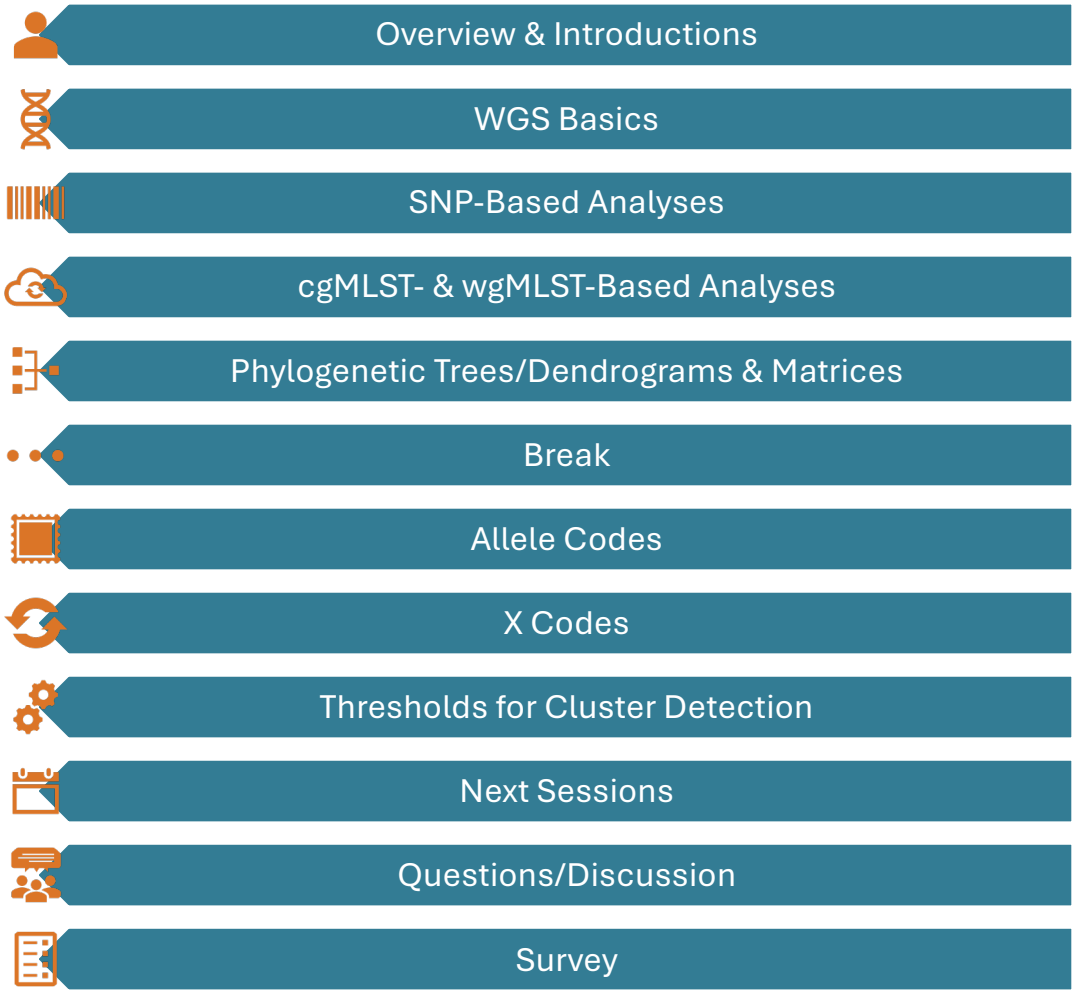
Research Assistant Professor

Dept. of Food Science

University of Tennessee

lkhudson@utk.edu

Overview


Overview & Introductions
WGS Basics
SNP-Based Analyses
cgMLST- & wgMLST-Based Analyses
Phylogenetic Trees/Dendrograms & Matrices
Break
Allele Codes
X Codes
Thresholds for Cluster Detection
Next Sessions
Questions/Discussion
Survey

WGS Basics

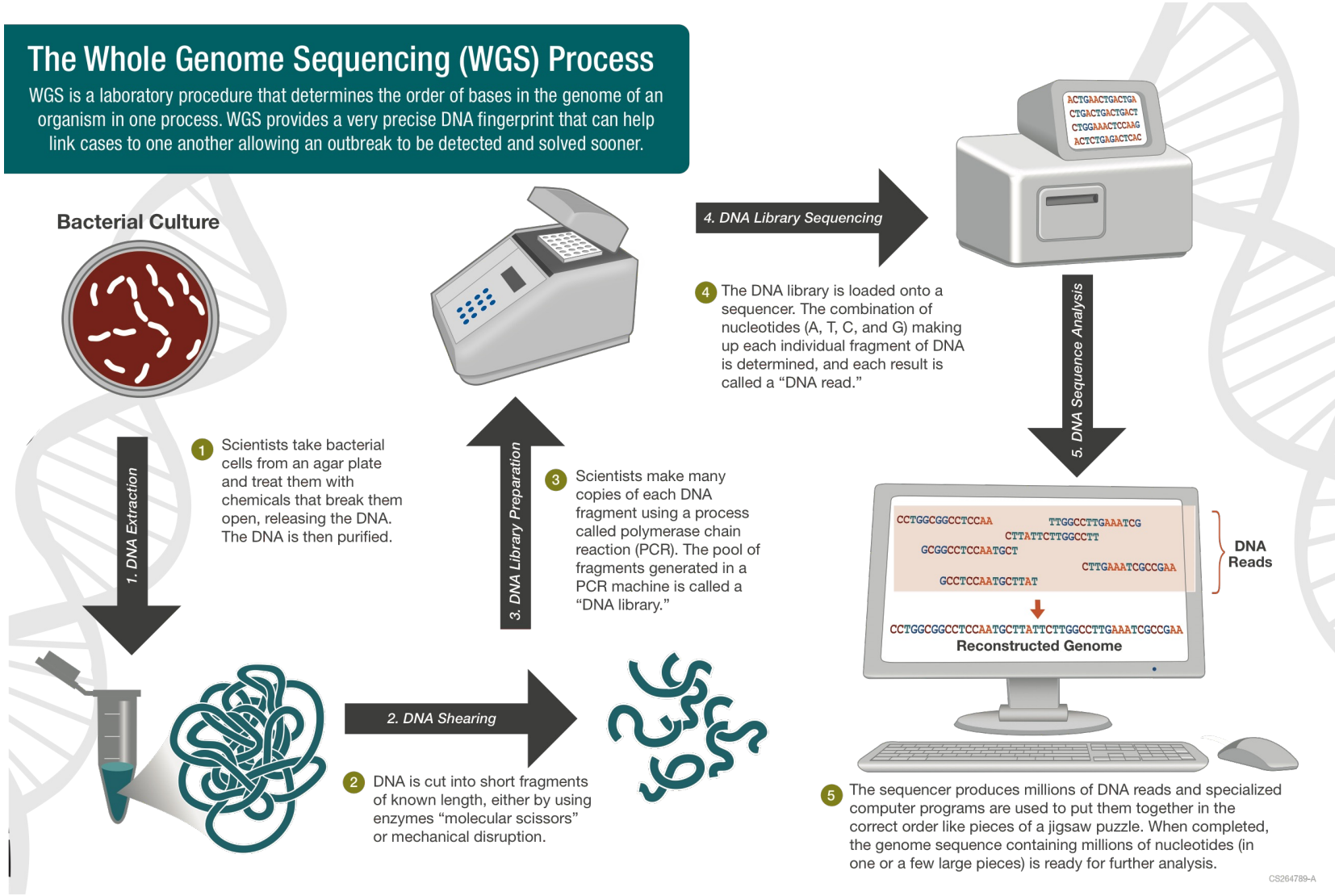
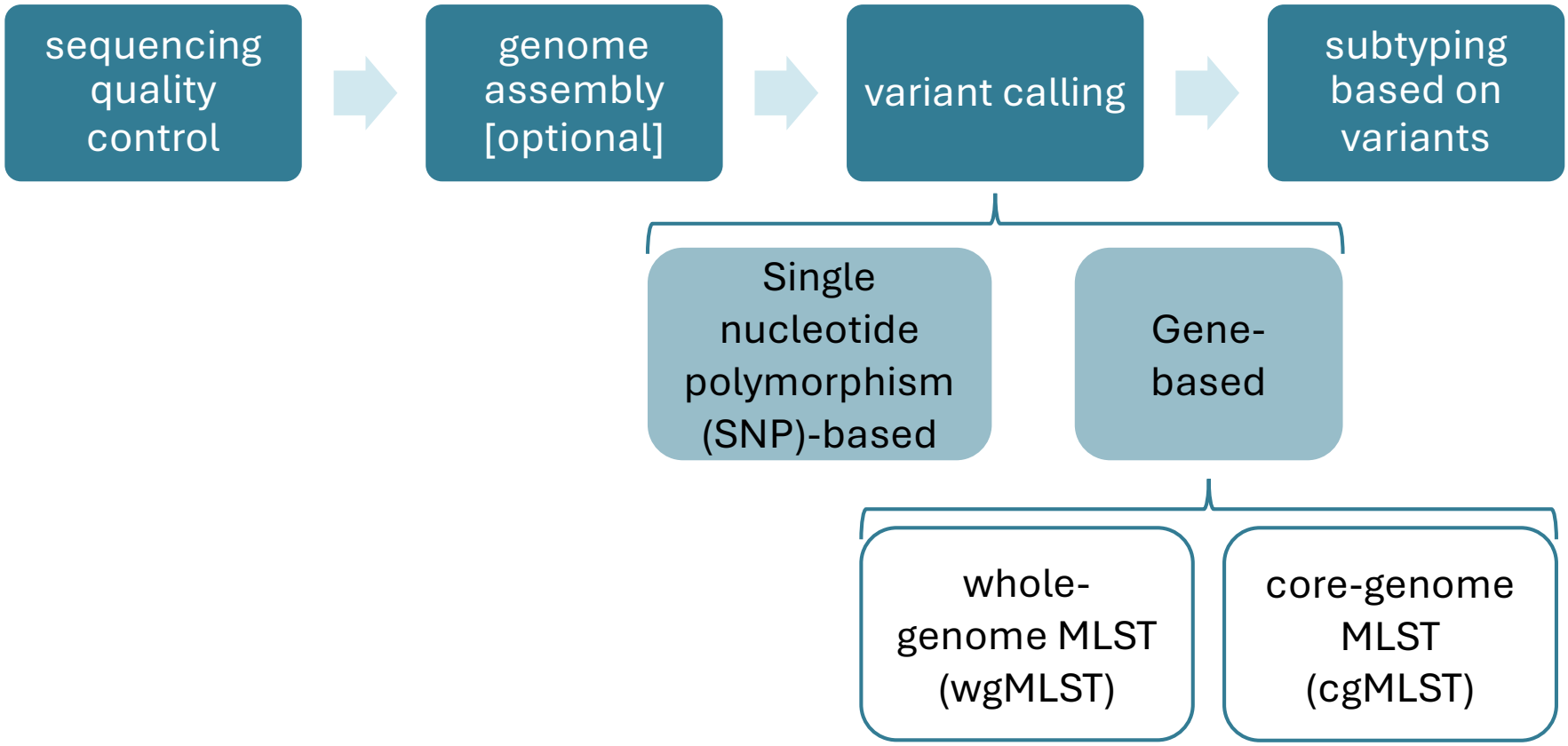


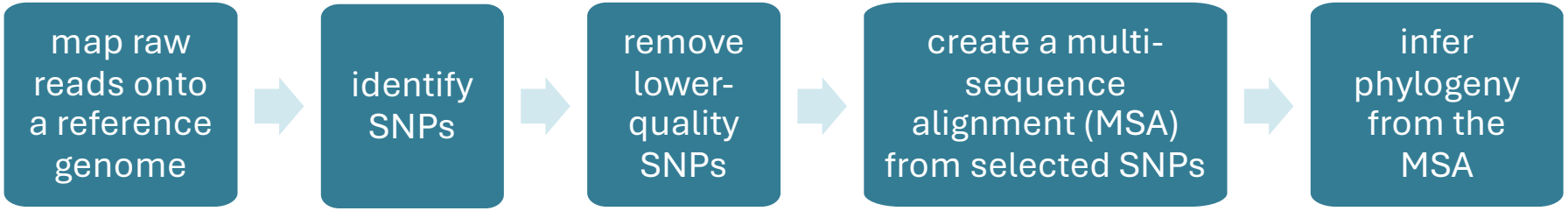
Figure: https://www.cdc.gov/pulsenet/pathogens/wgs.html

WGS Analyses



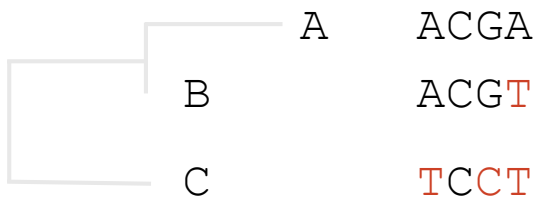
SNP-based Analysis

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




```
reference  ATGTTCCTC
sequence   ATGTTGCTC
```

SNP Calling



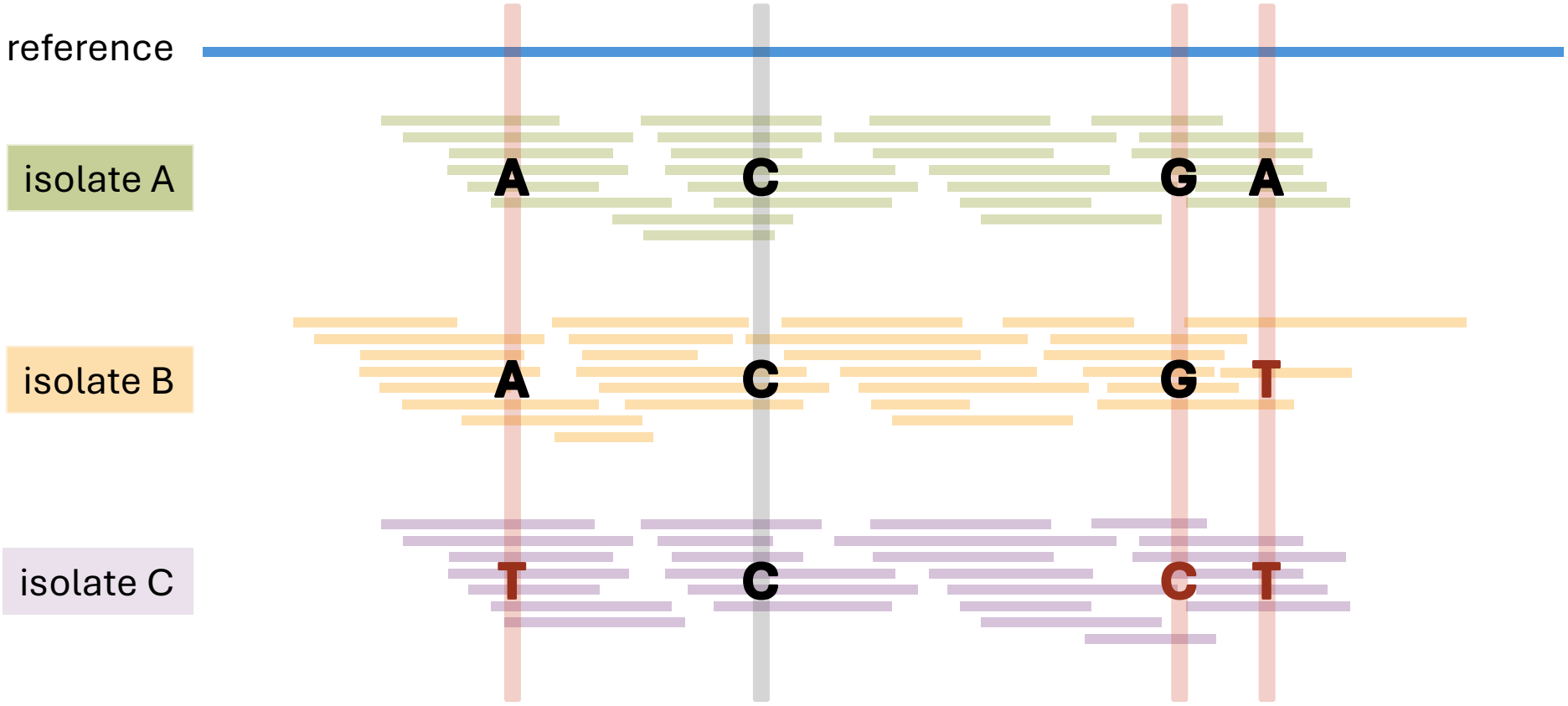
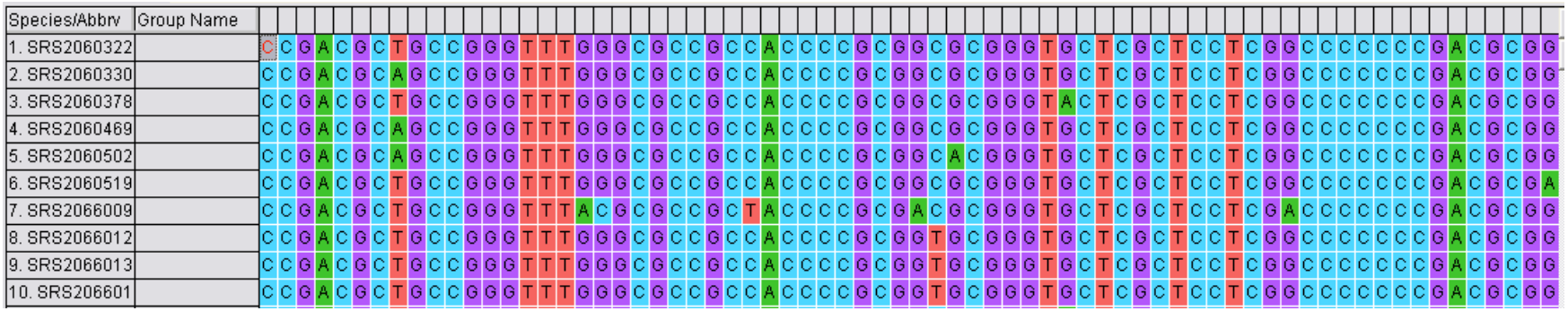


Figure: Adapted from Sevinsky, J. & MacCannell, D. (2017)

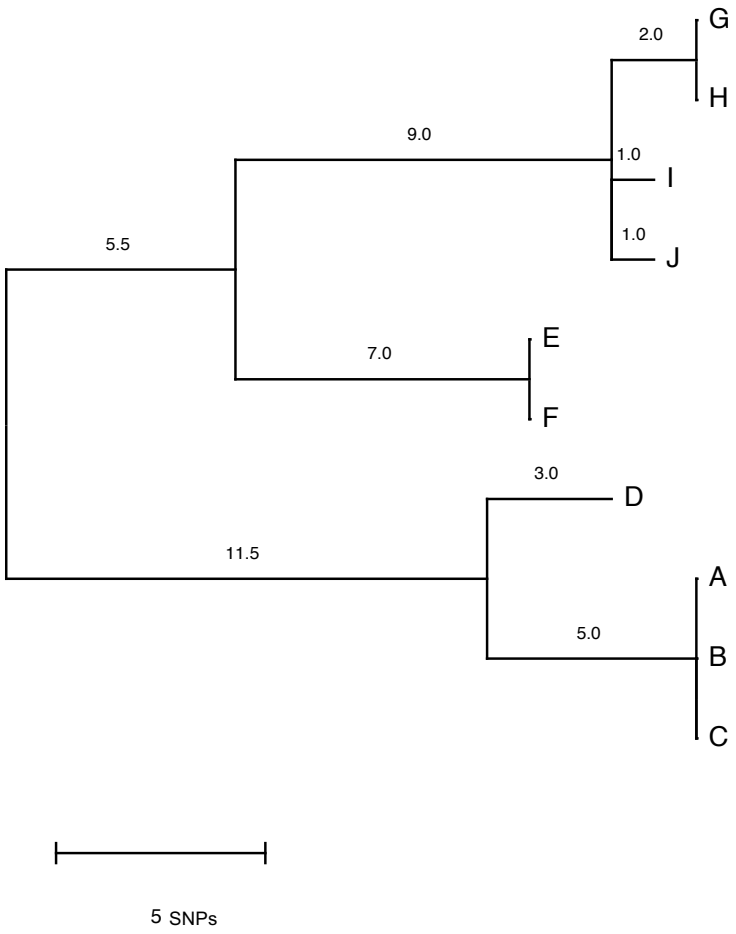
Multi-FASTA Alignment



hqSNP Matrix

	isolate1	isolate2	isolate3	isolate4	isolate5	isolate6	isolate7	isolate8	isolate9	isolate10
isolate1	0	0	0	8	29	29	33	33	32	32
isolate2	0	0	0	8	29	29	33	33	32	32
isolate3	0	0	0	8	29	29	33	33	32	32
isolate4	8	8	8	0	27	27	31	31	30	30
isolate5	29	29	29	27	0	0	18	18	17	17
isolate6	29	29	29	27	0	0	18	18	17	17
isolate7	33	33	33	31	18	18	0	0	3	3
isolate8	33	33	33	31	18	18	0	0	3	3
isolate9	32	32	32	30	17	17	3	3	0	2
isolate10	32	32	32	30	17	17	3	3	2	0

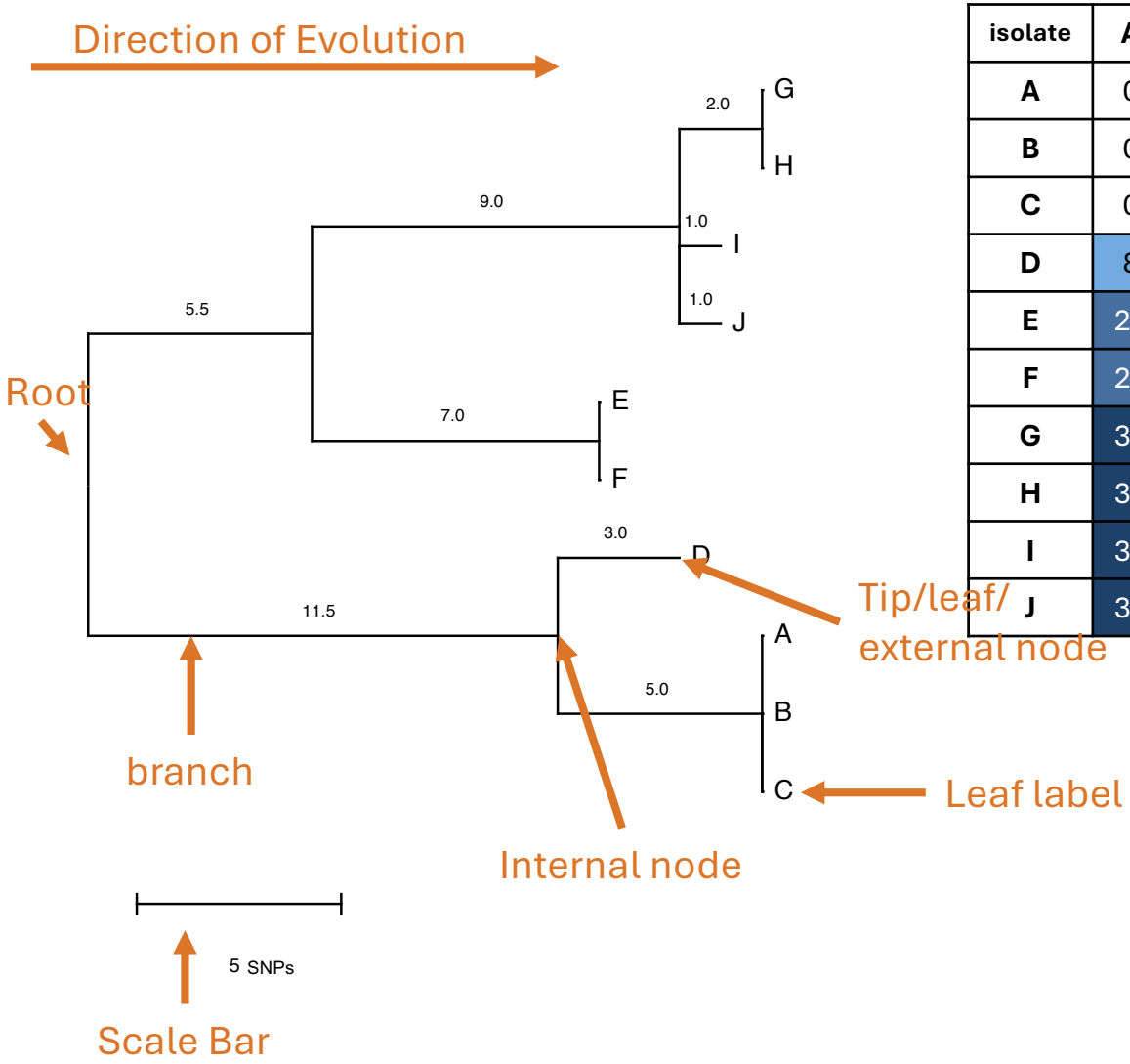
Phylogenetic Tree



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

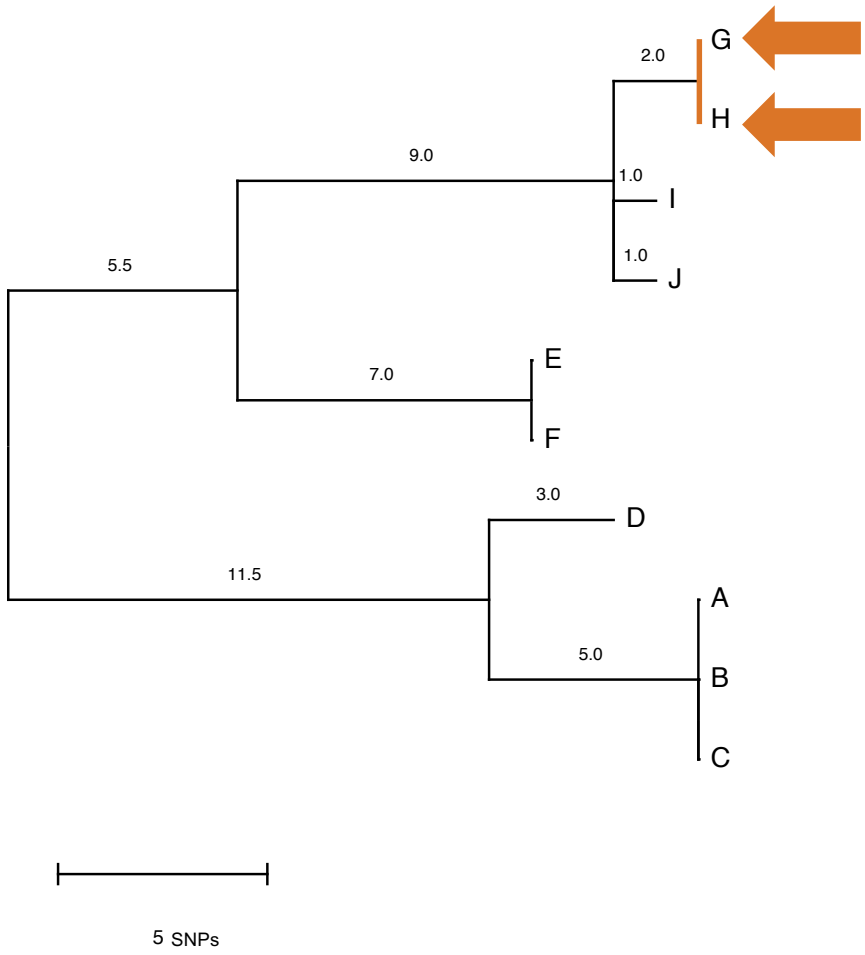
Phylogenetic Tree



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

Phylogenetic Tree



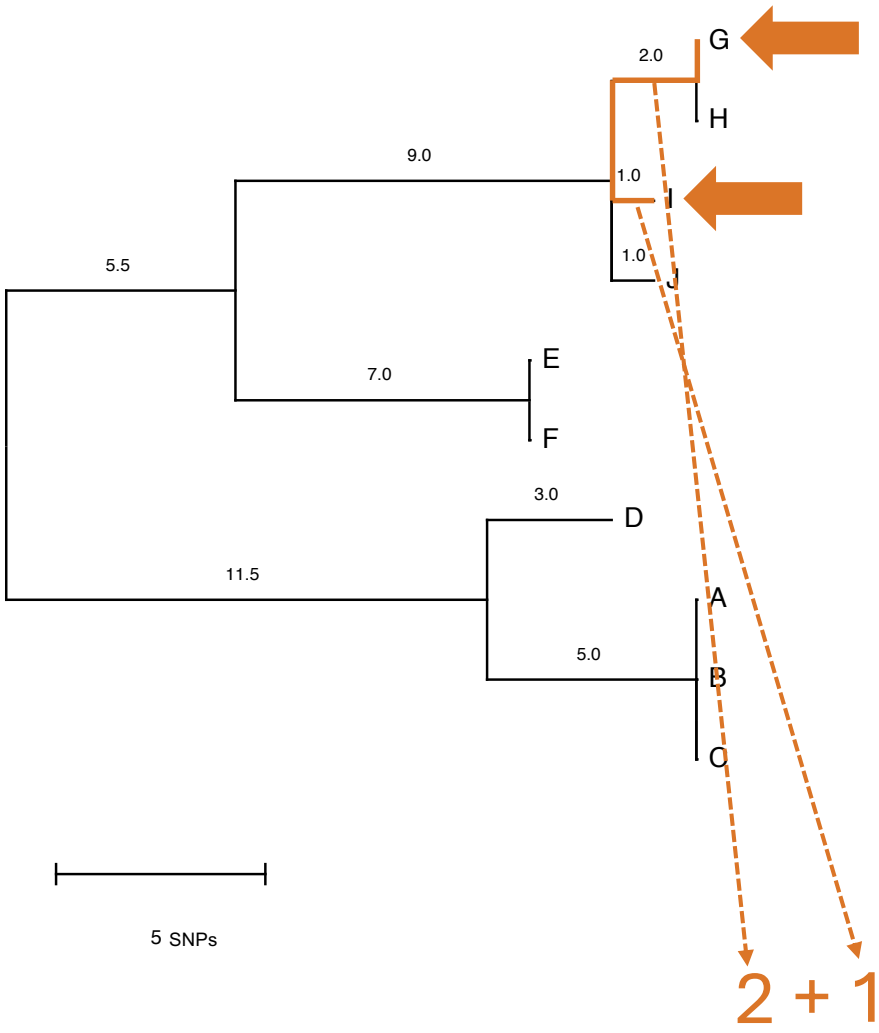
SNP Matrix

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 and isolate H?

0 SNPs

Phylogenetic Tree



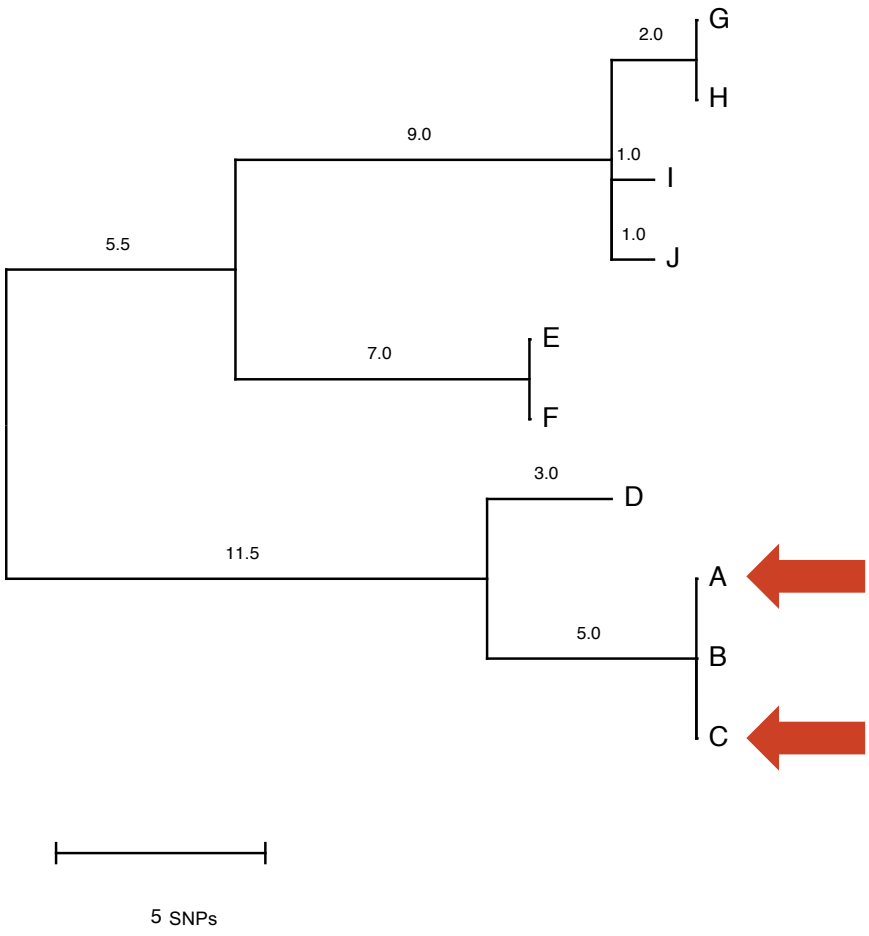
SNP Matrix

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 and isolate I?

3 SNPs

Phylogenetic Tree

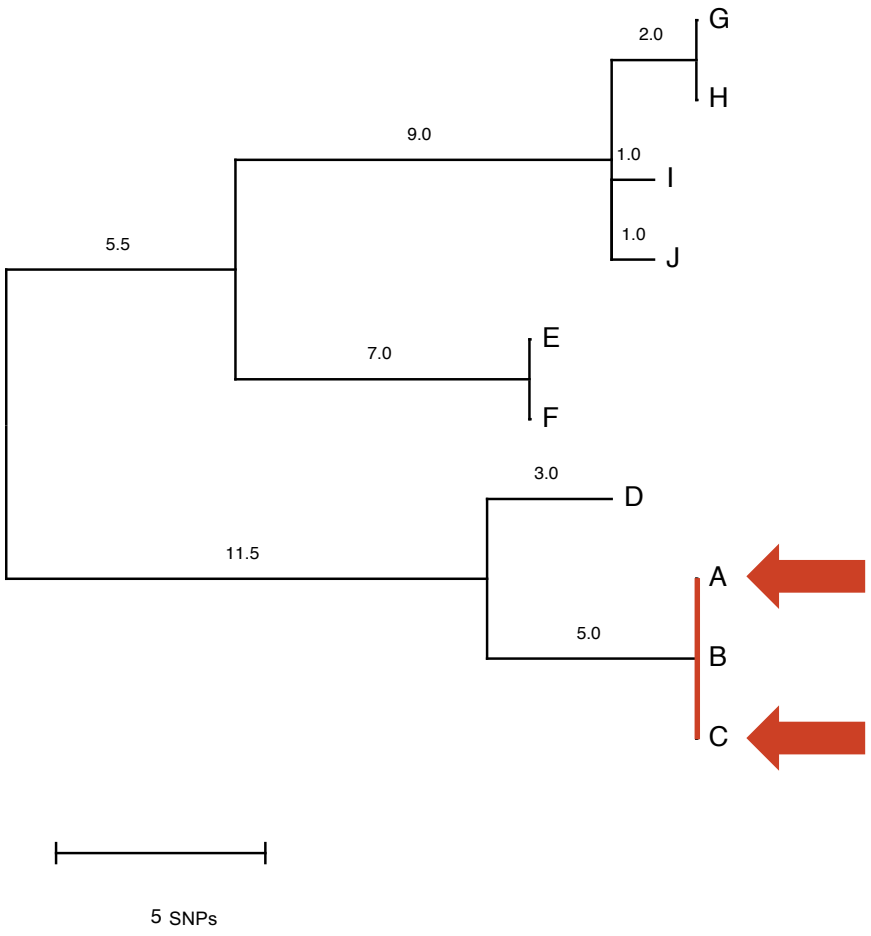


SNP Matrix

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

What is the distance between isolate A and and isolate C?

Phylogenetic Tree



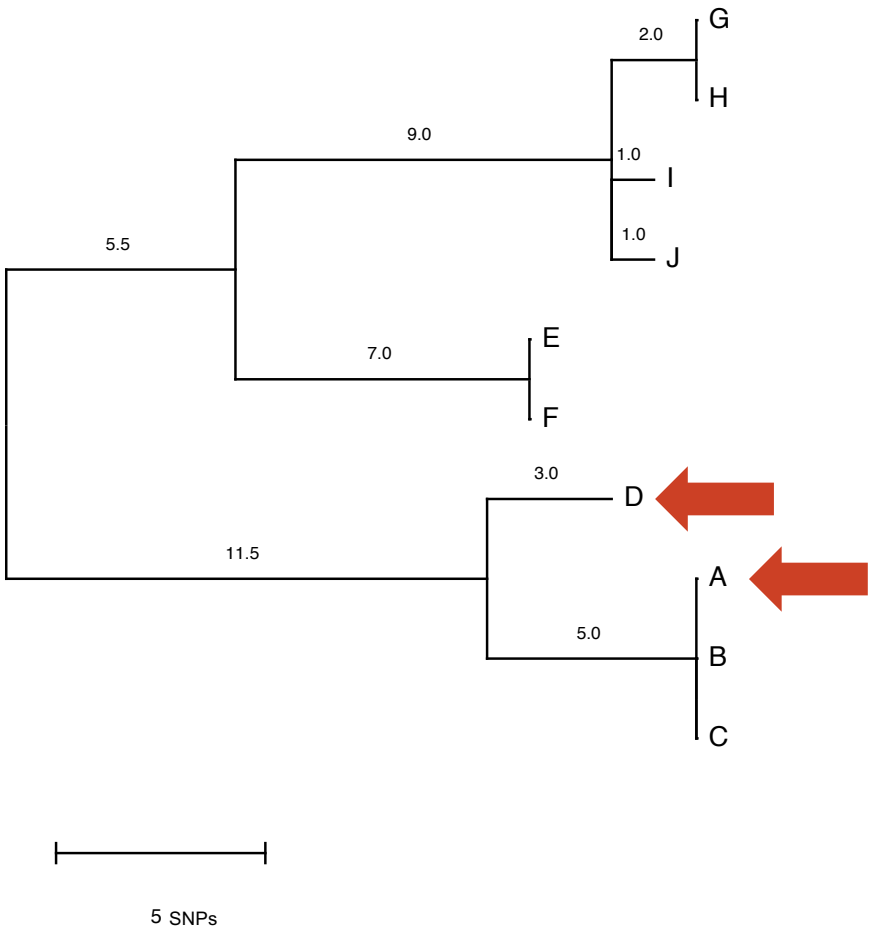
SNP Matrix

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

What is the distance between isolate A and and isolate C?

0 SNPs

Phylogenetic Tree

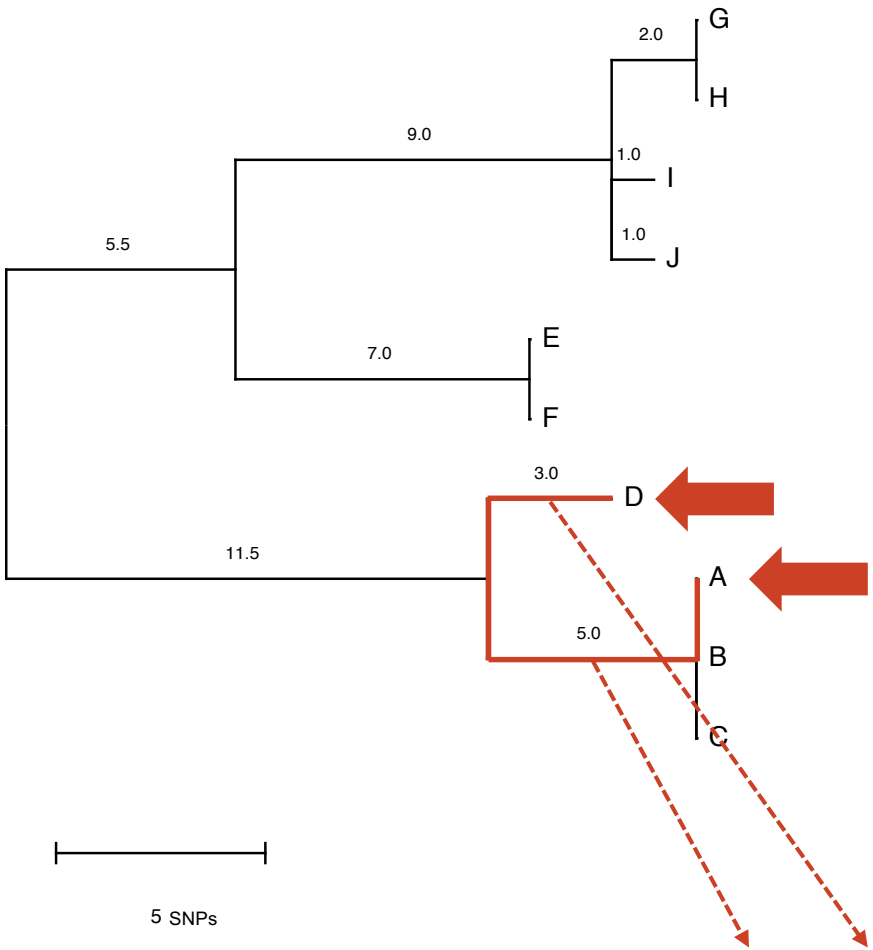


SNP Matrix

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

What is the distance between isolate A and 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

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

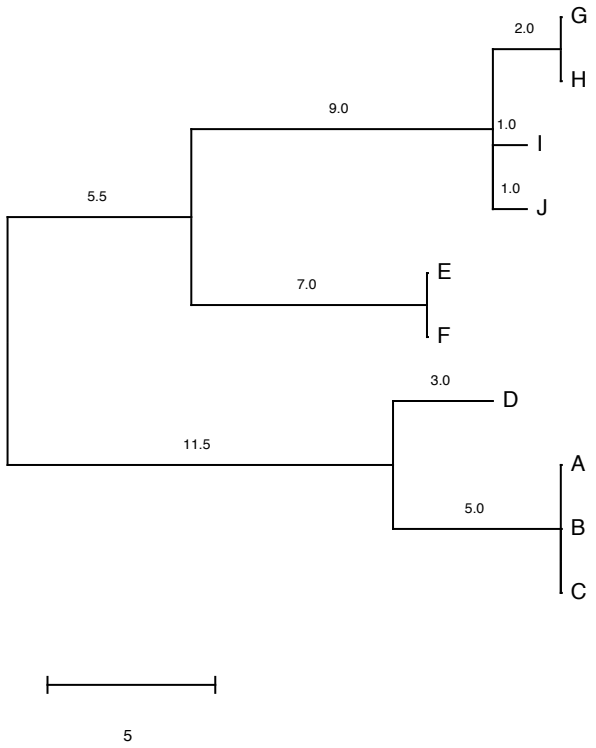
Cladogram

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

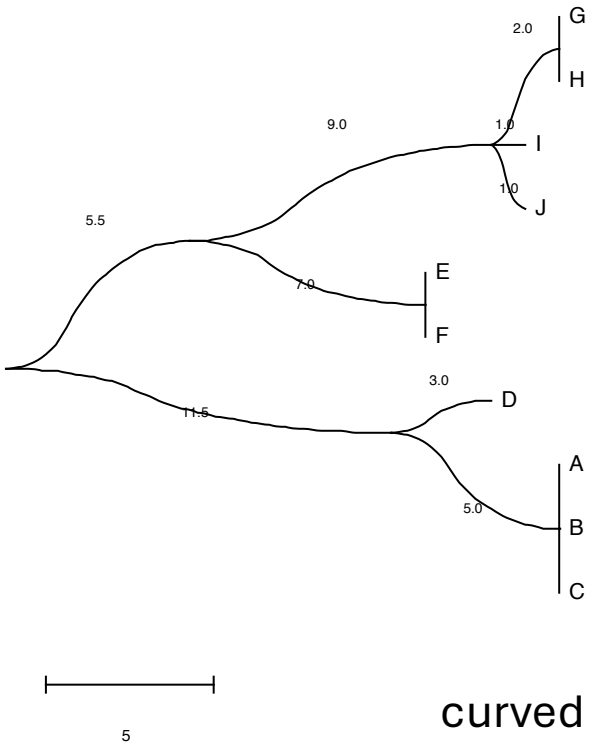
Phylogram

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

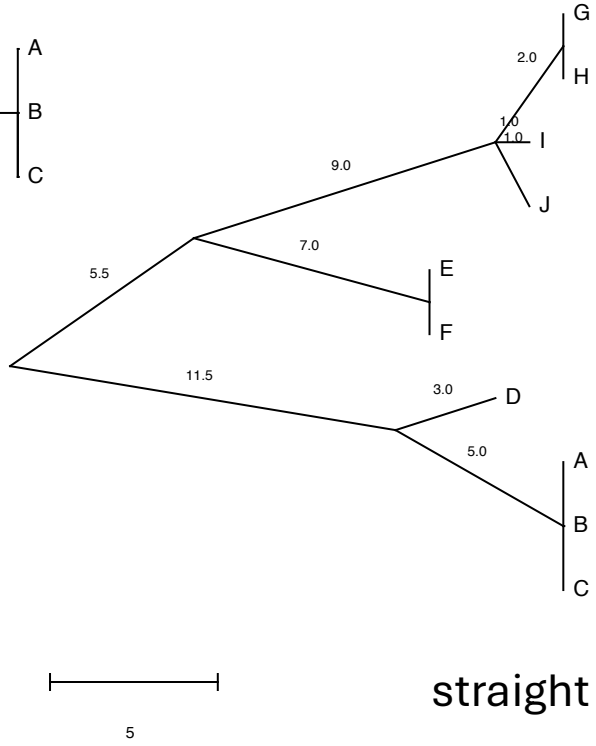
Tree Styles



rectangular

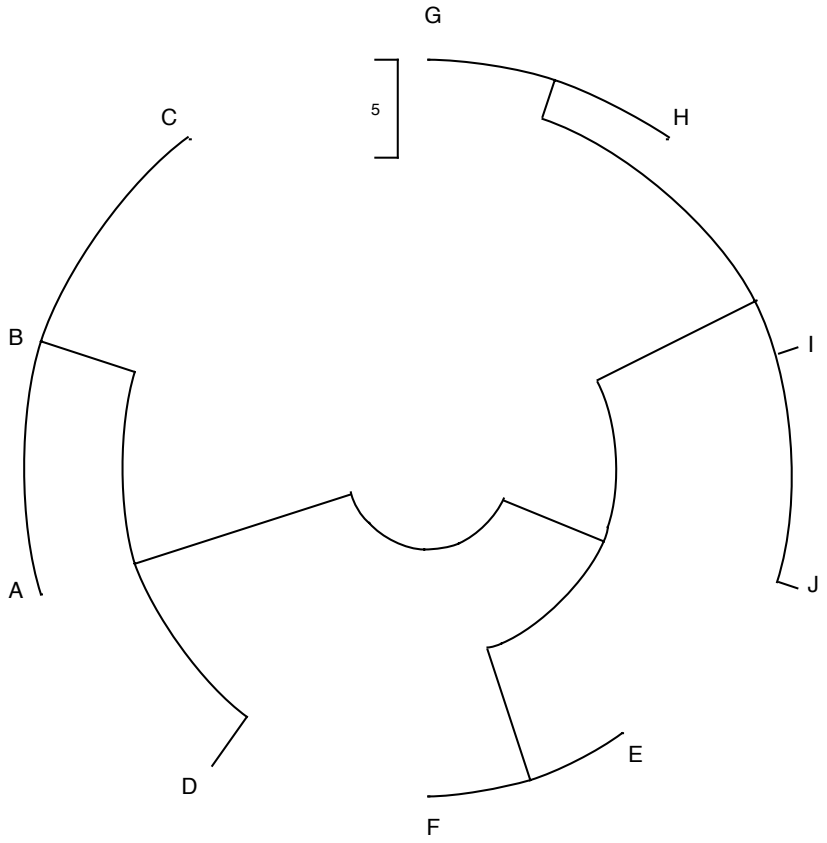


curved

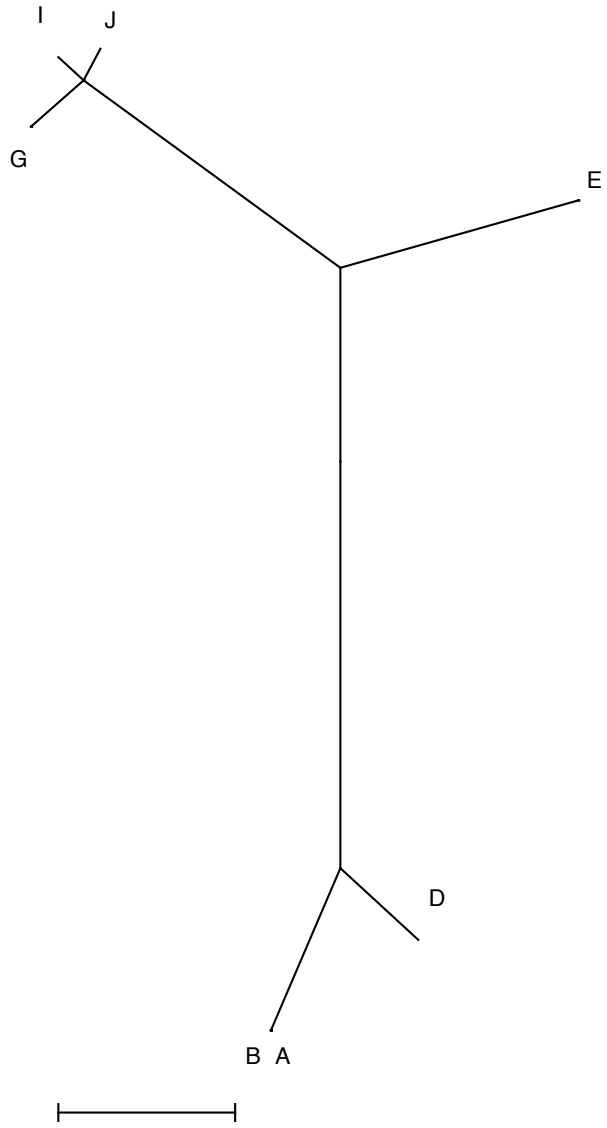


straight

Tree Styles

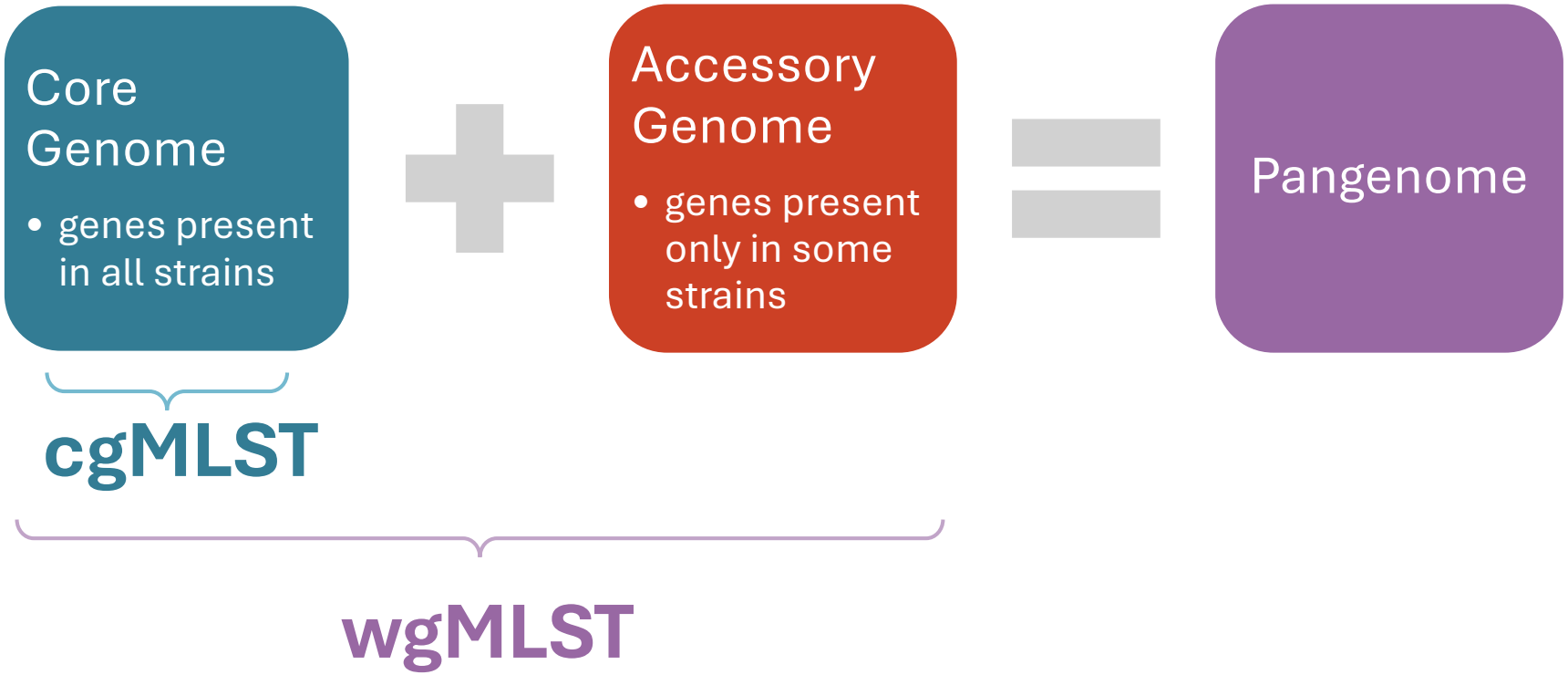


circular



Radiation/unrooted

wgMLST and cgMLST



wgMLST and cgMLST

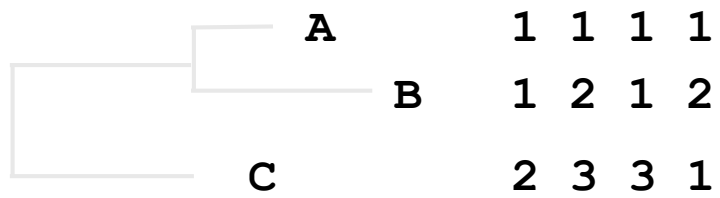
Based on*:

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

* For *L. monocytogenes* specifically

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

wgMLST and cgMLST



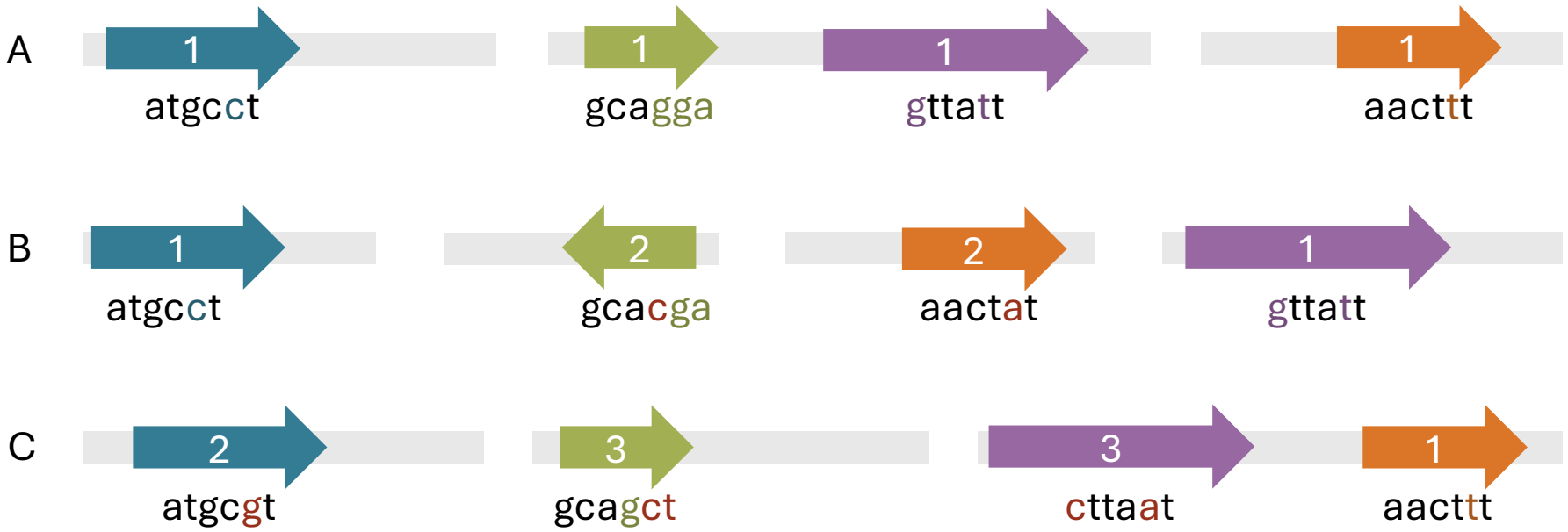
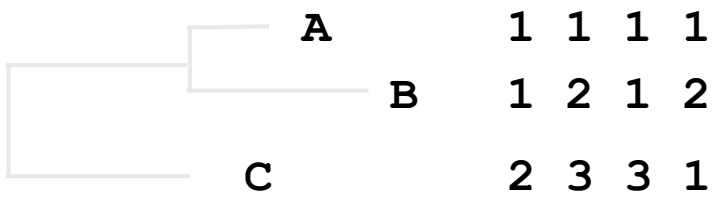


Figure: Adapted from Sevinsky, J. & MacCannell, D. (2017)

wgMLST and cgMLST



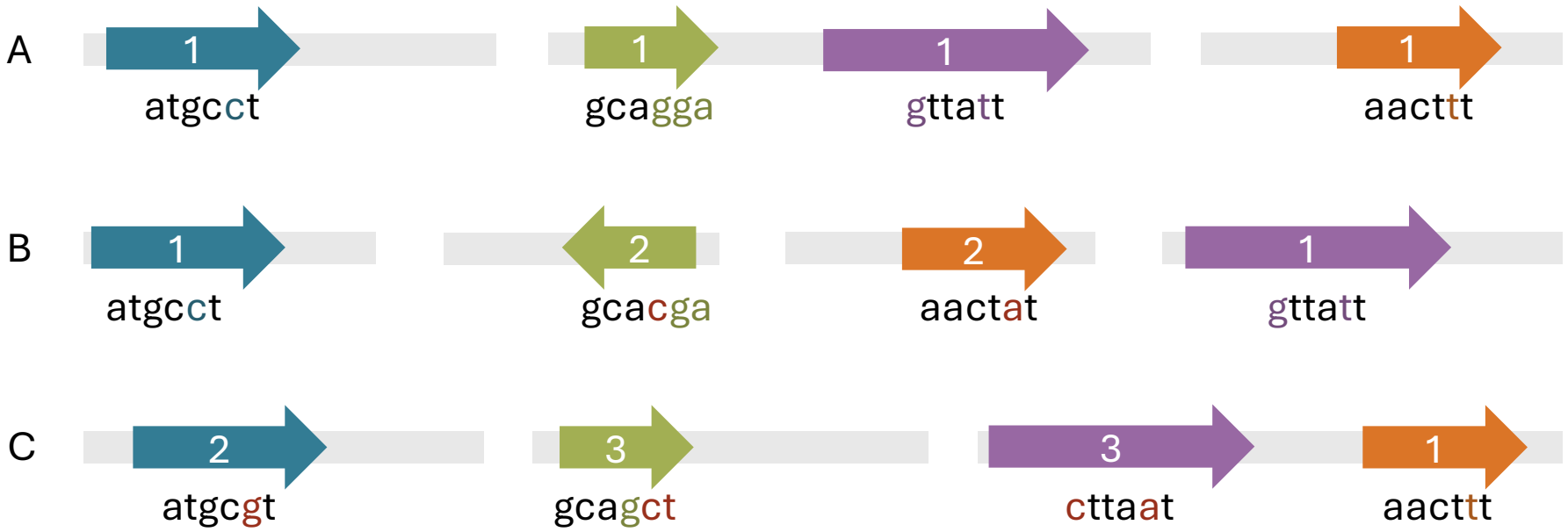
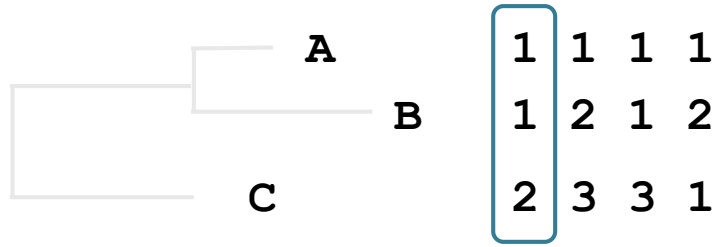


Figure: Adapted from Sevinsky, J. & MacCannell, D. (2017)

wgMLST and cgMLST



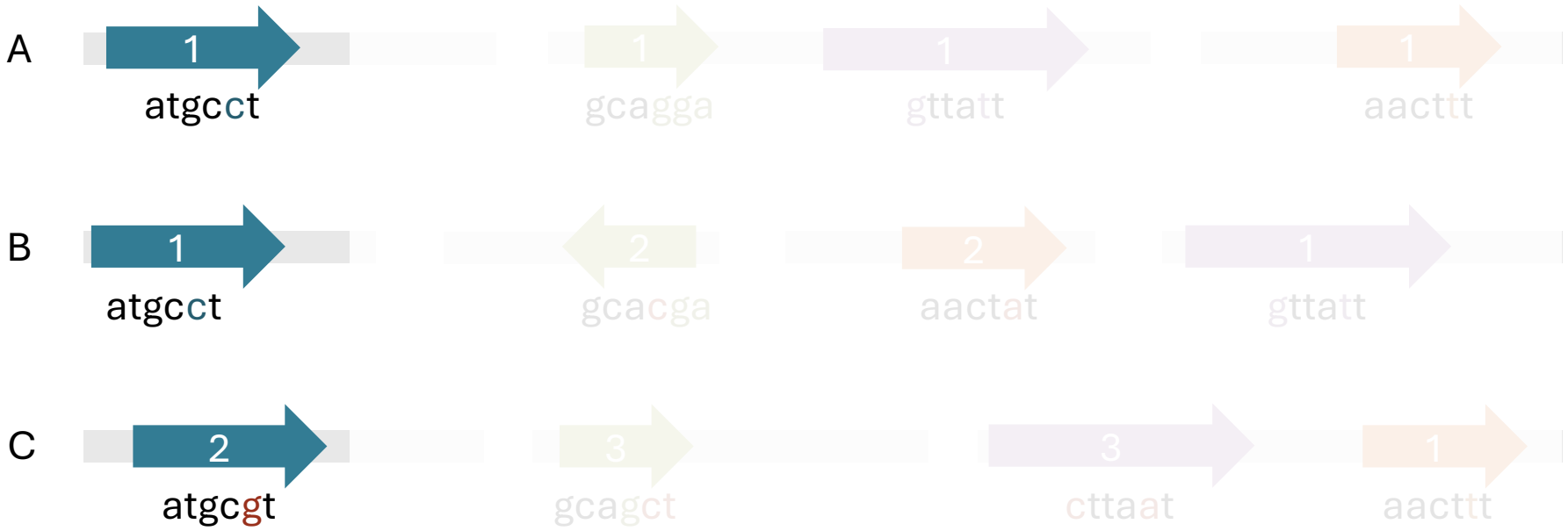
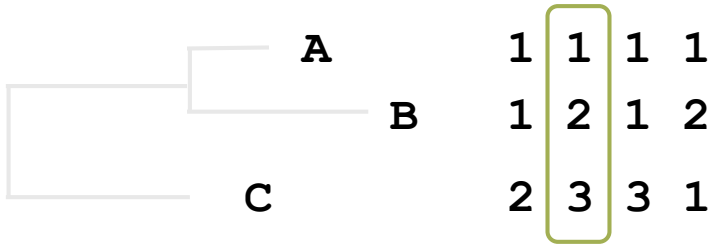


Figure: Adapted from Sevinsky, J. & MacCannell, D. (2017)

wgMLST and cgMLST



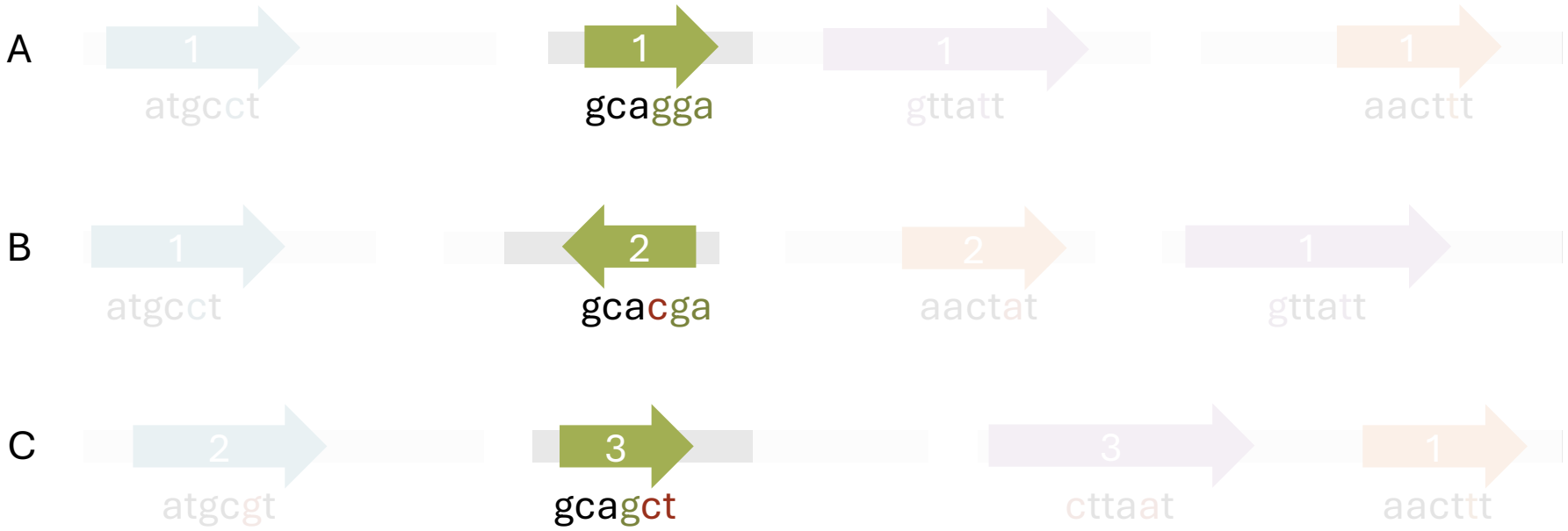
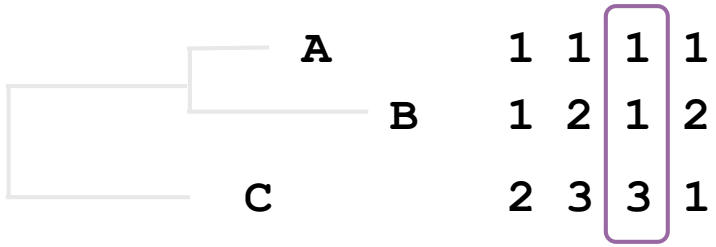


Figure: Adapted from Sevinsky, J. & MacCannell, D. (2017)

wgMLST and cgMLST



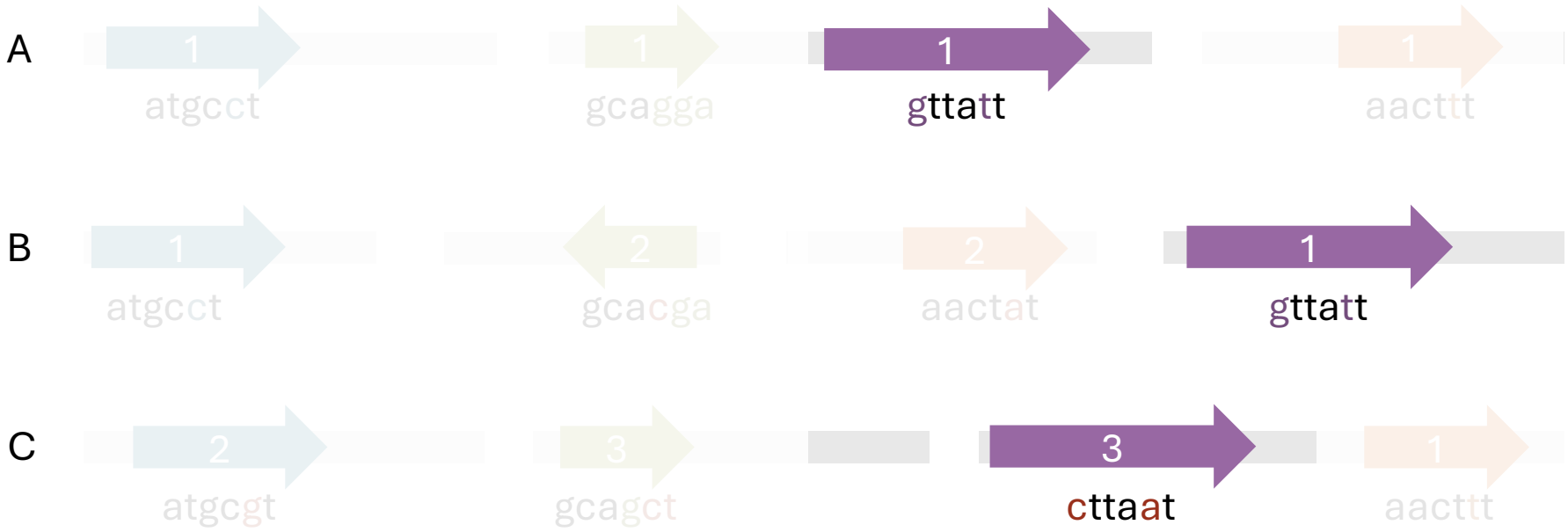
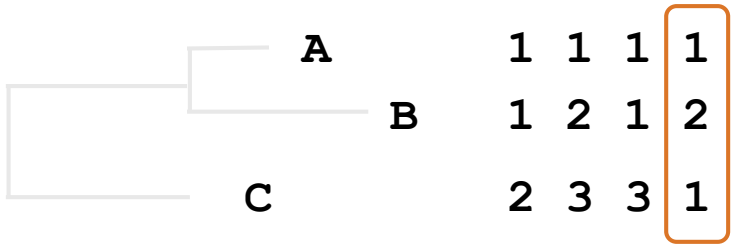


Figure: Adapted from Sevinsky, J. & MacCannell, D. (2017)

wgMLST and cgMLST



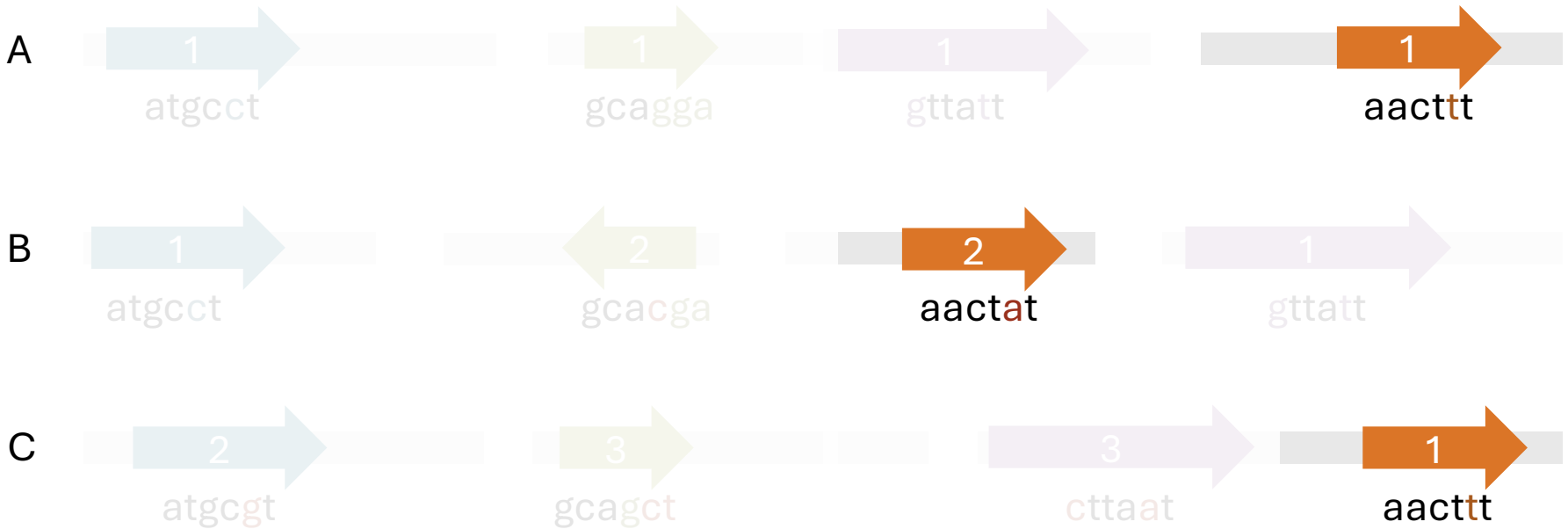
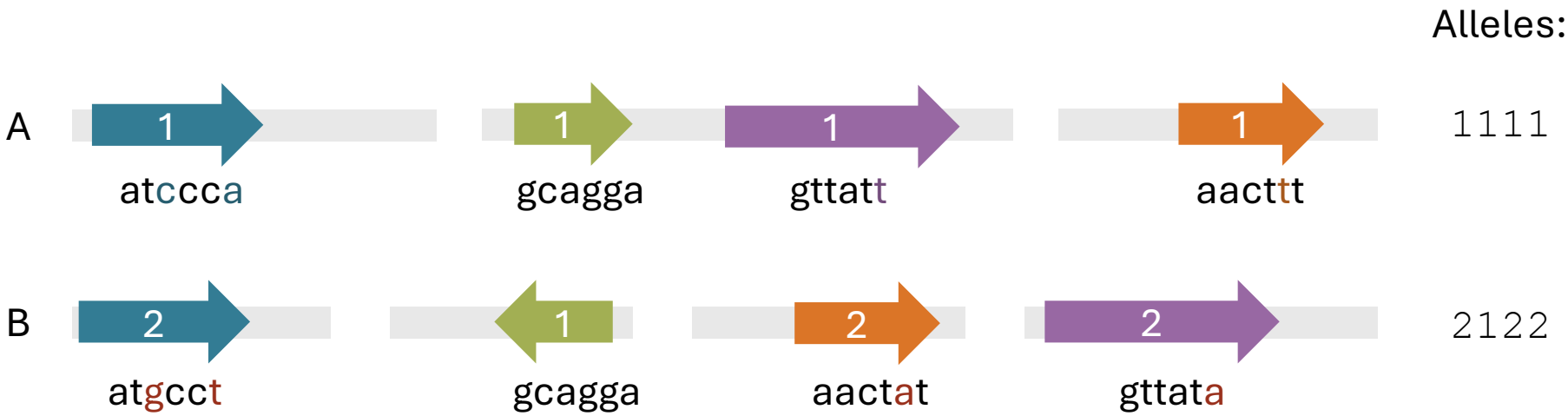


Figure: Adapted from Sevinsky, J. & MacCannell, D. (2017)

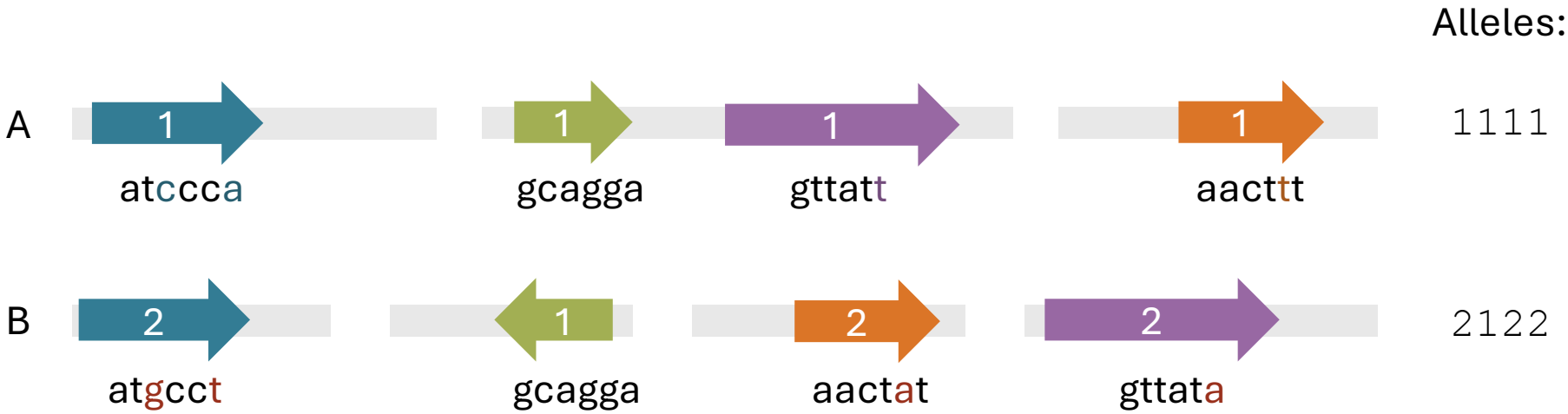
wgMLST and cgMLST



What is the allele distance between isolate 1 and and isolate 2?

Figure: Adapted from Sevinsky, J. & MacCannell, D. (2017)

wgMLST and cgMLST



What is the allele distance between isolate A and and isolate B?

3 alleles

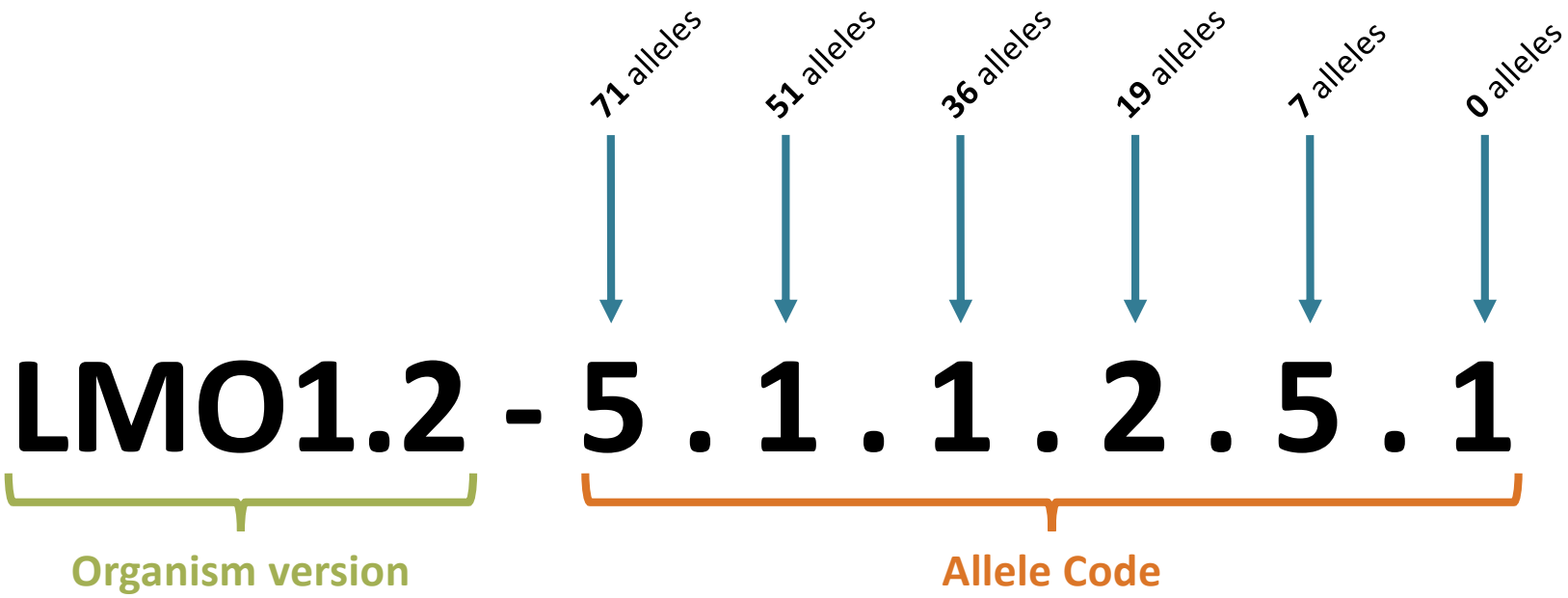
Figure: Adapted from Sevinsky, J. & MacCannell, D. (2017)

Break

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

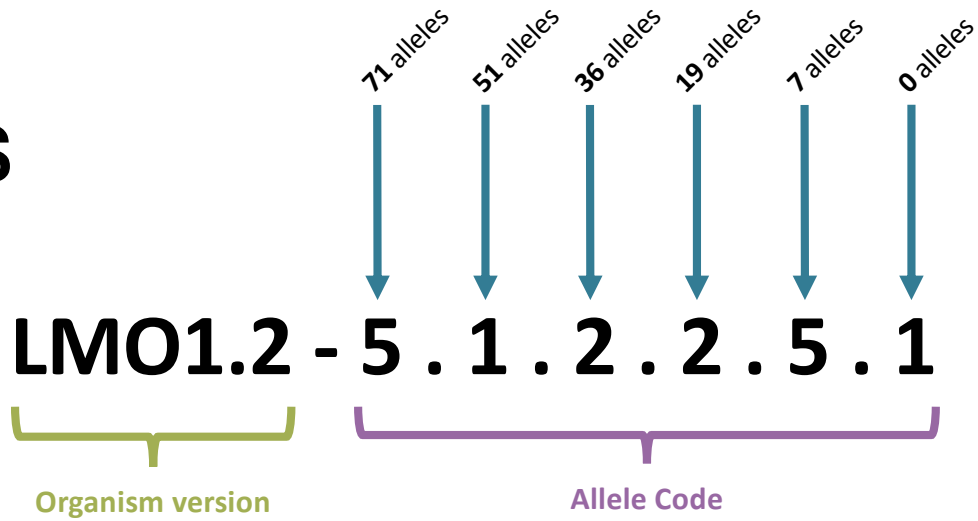
- Allele codes & thresholds of relatedness vary by organism

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

Allele Codes

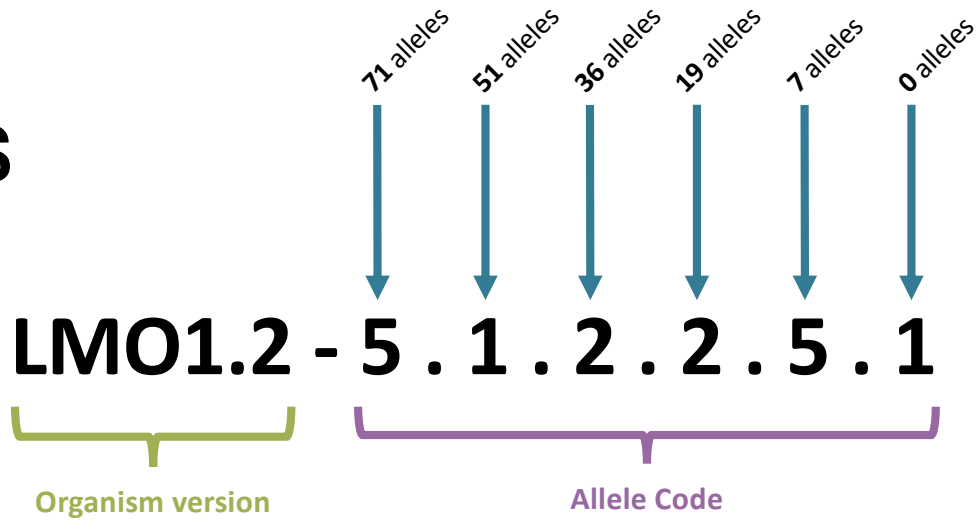
- In general, more numbers in common = more closely related
- Allele code thresholds are approximate for all organisms
- Due to the clonal lineage of some serotypes or strains, the approximate differences may be larger than expected
- Allele differences should always be verified with a dendrogram

Allele Codes



A: **LMO1.2 -** [ ] → Singleton: No close matches, name not assigned.

Allele Codes

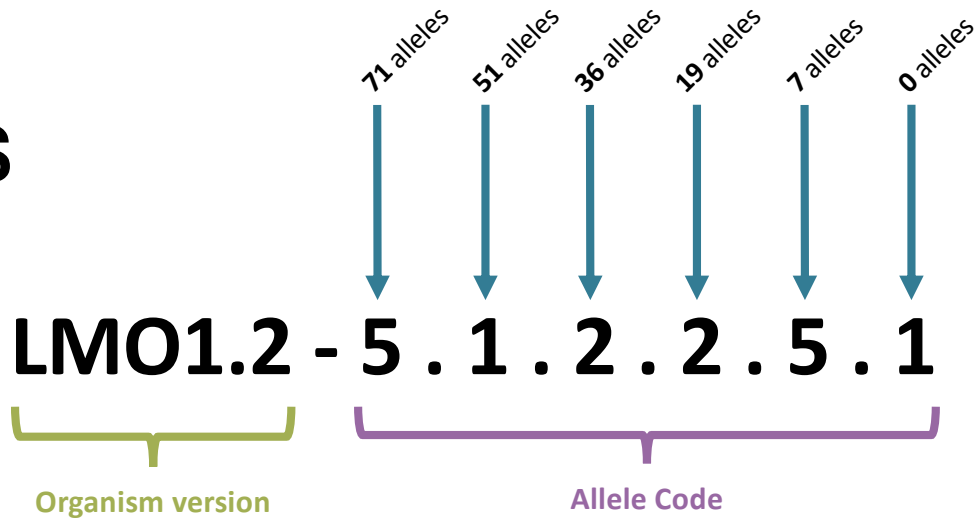


A: **LMO1.2 - 5 . 1 . 2**

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

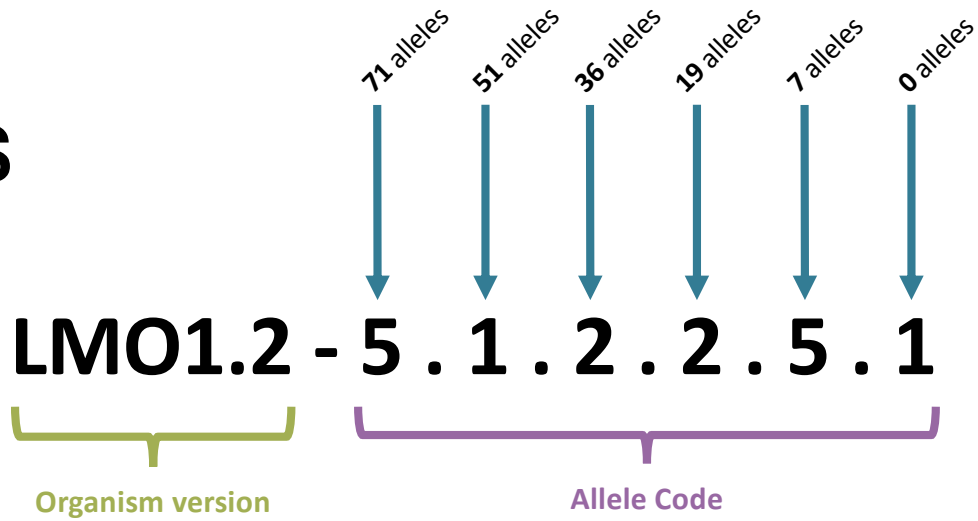


A: **LMO1.2 - 5 . 1 . 2**

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

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

Allele Codes

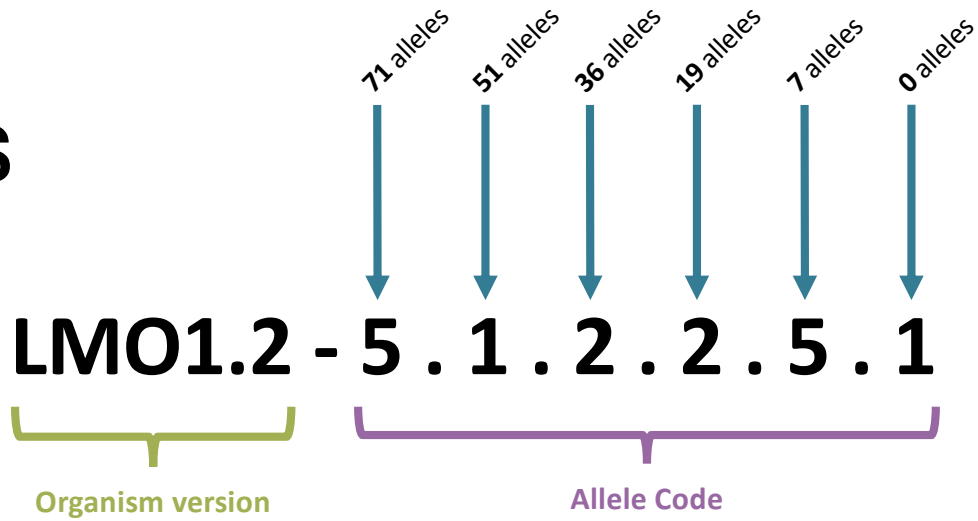


A: **LMO1.2 - 26 . 1 . 1 . 1 . 1 . 1**

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

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

Allele Codes



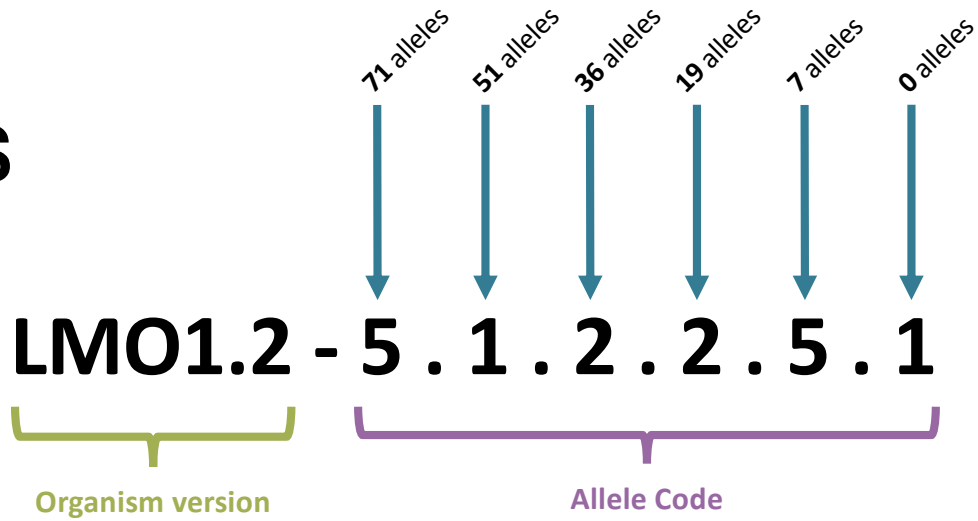
A: **LMO1.2 - 1 . 4 . 64 . 1 . 1 . 2**

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



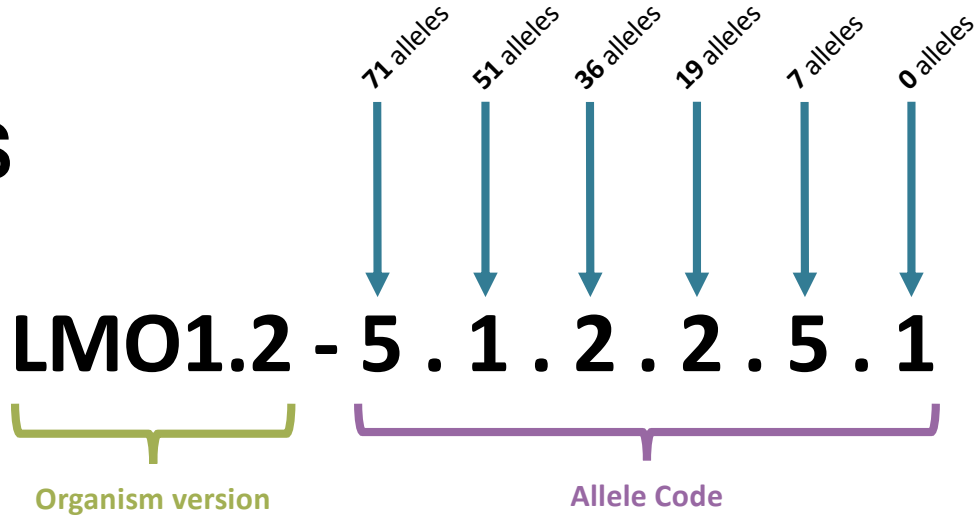
A: **LMO1.2 - 1 . 4 . 64 . 1 . 1 . 2**

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

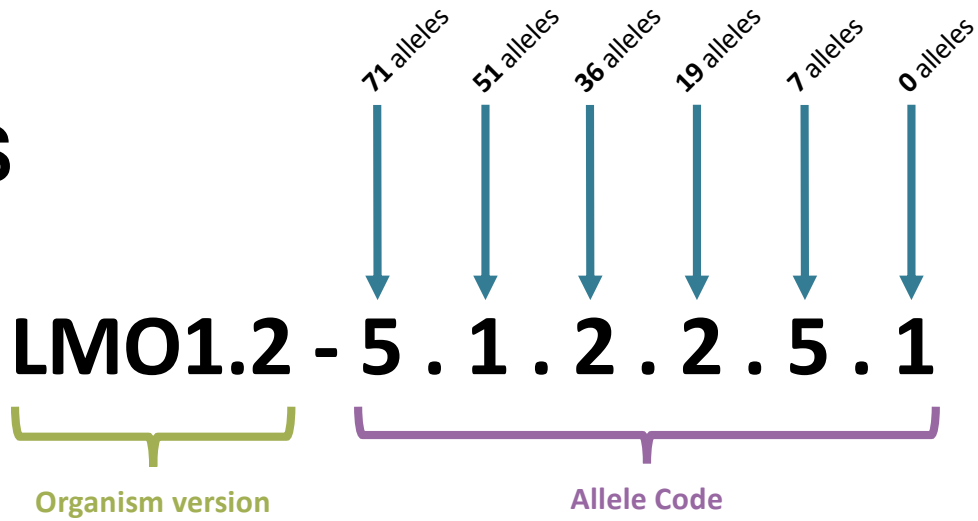


A: **LMO1.2 - 26 . 1 . 1 . 1 . 1 . 1**

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

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**

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

7-19 alleles different

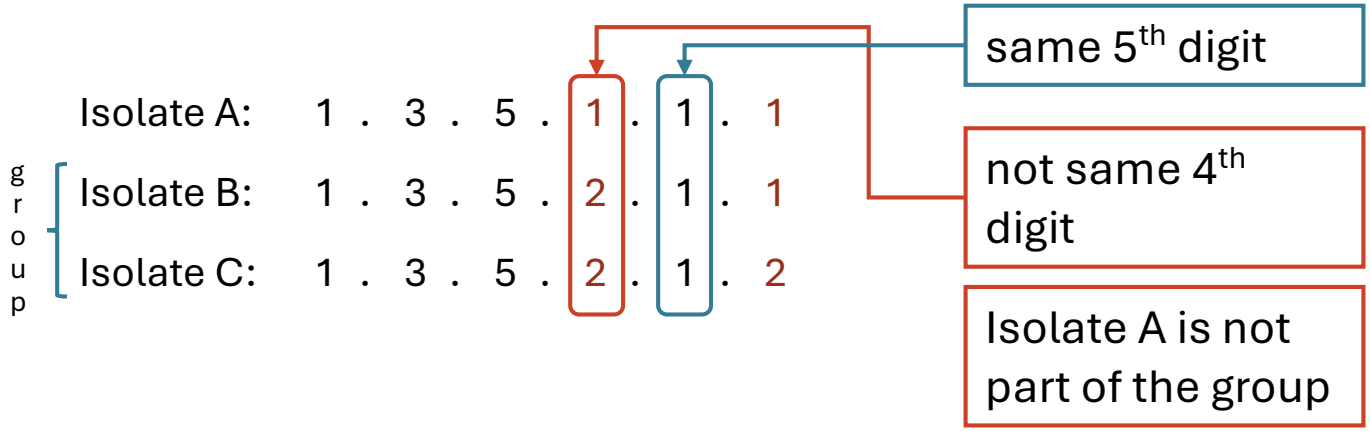
Updated Allele Codes

- 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"

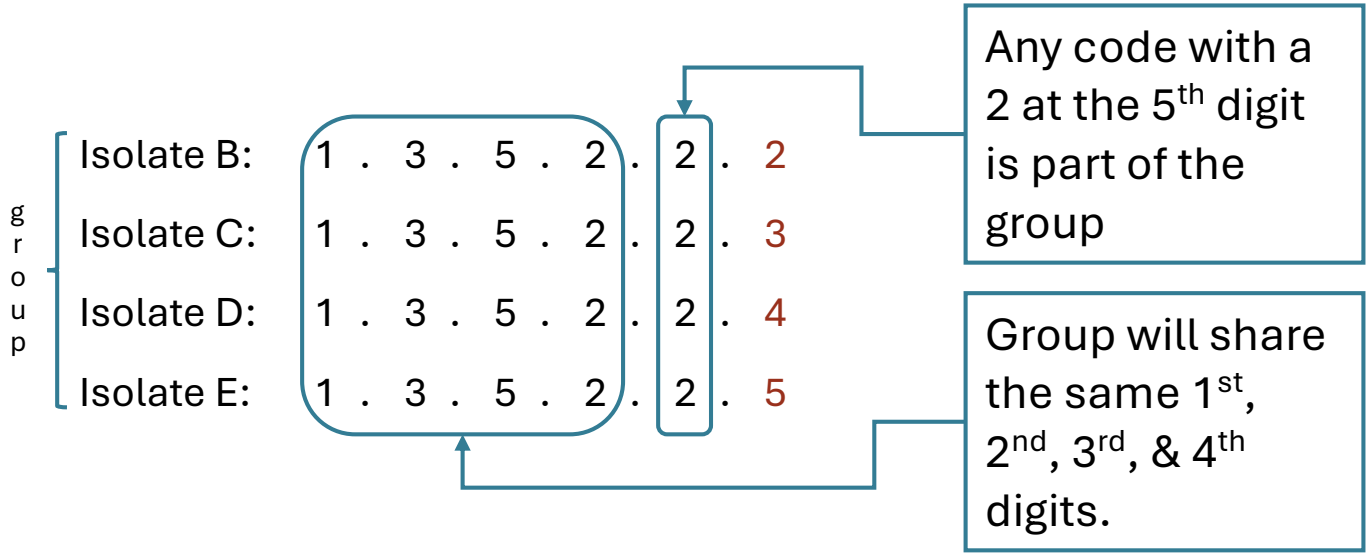
https://www.aphl.org/Materials/PN2.0_comms_partner%20letter.pdf#search=PulseNet%202.0

Updated Allele Codes

Previous Version



New Version



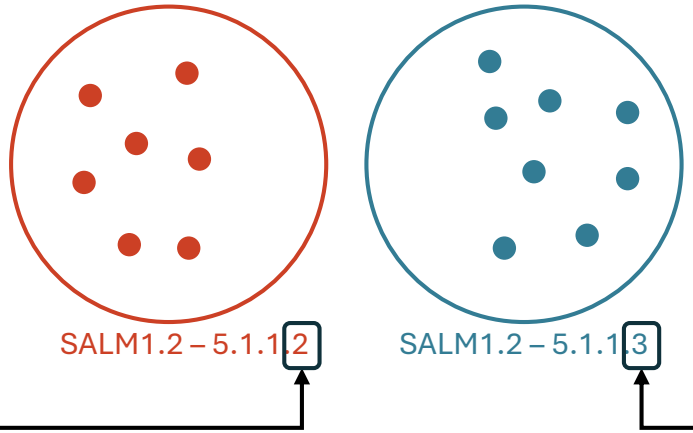
Allele Code Limitations

- 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

Allele code merging/chaining

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

(Same 4th digit)



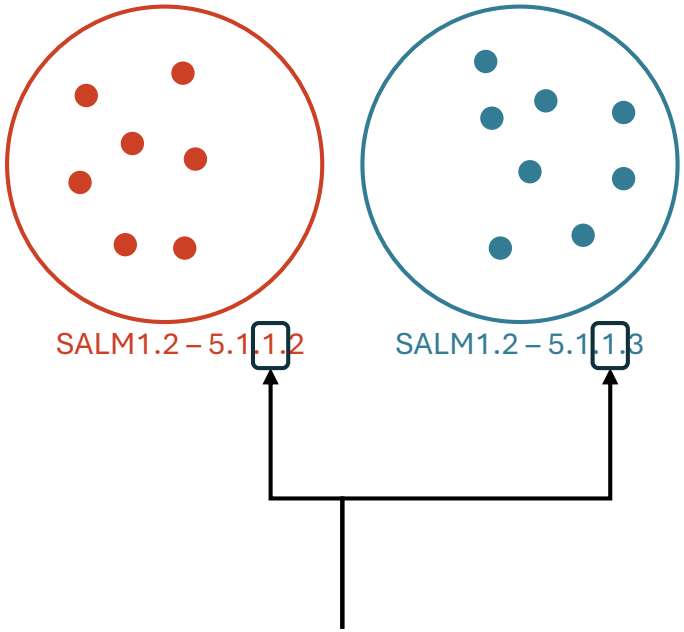
Blue cluster:
isolates up to **7** alleles different

(Same 4th digit)

Allele code merging/chaining

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

(Same 4th digit)



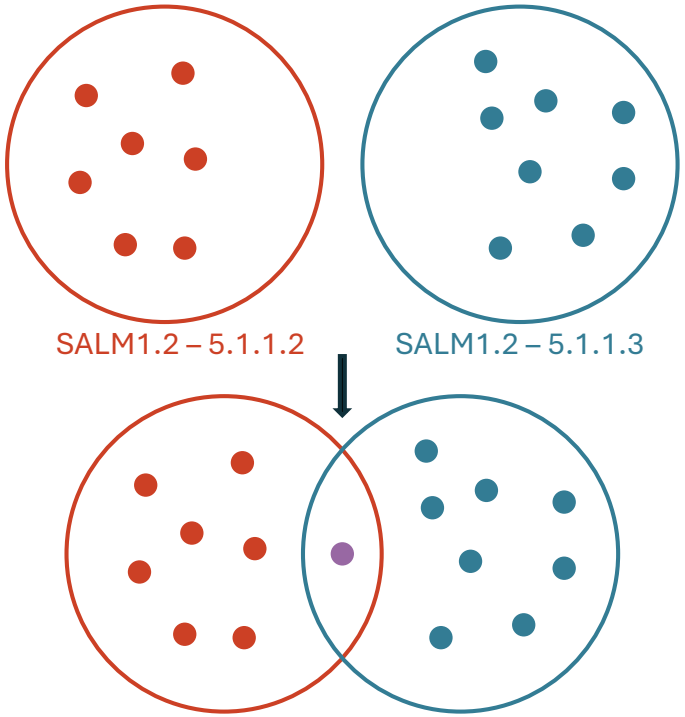
Blue cluster:
isolates up to **7** alleles different

(Same 4th digit)

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

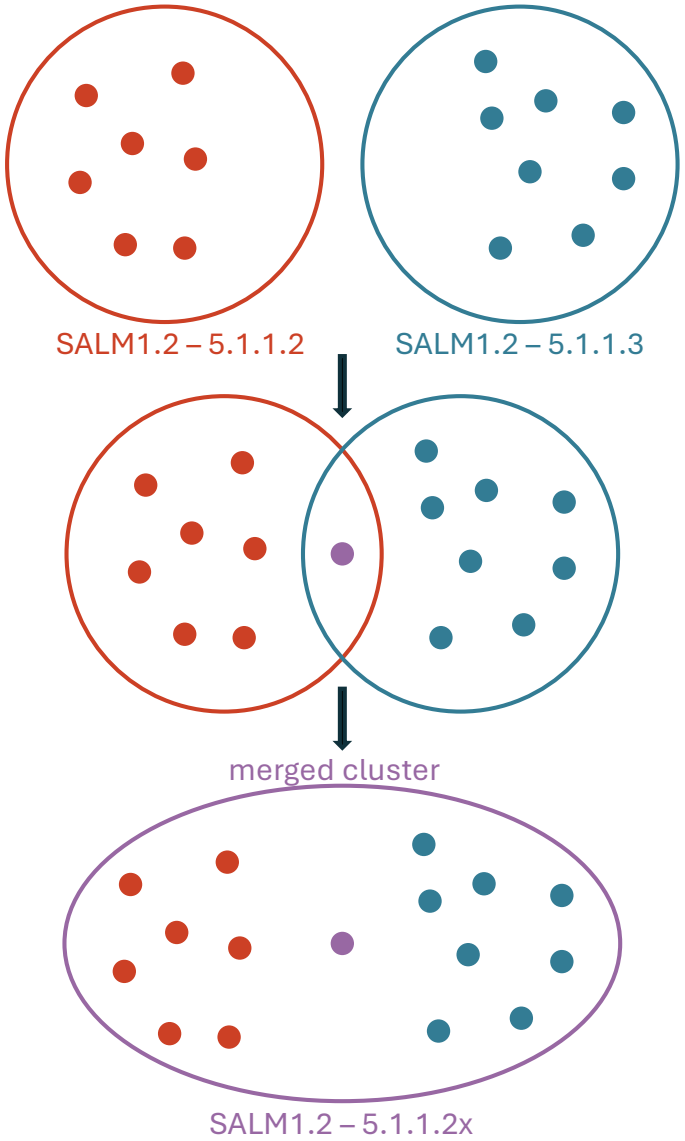
(Same 3rd digit)

Allele code merging/chaining



New **purple isolate** causes a merge of the **red** & **blue** clusters

Allele code merging/chaining



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

contains isolates **7-15** alleles different

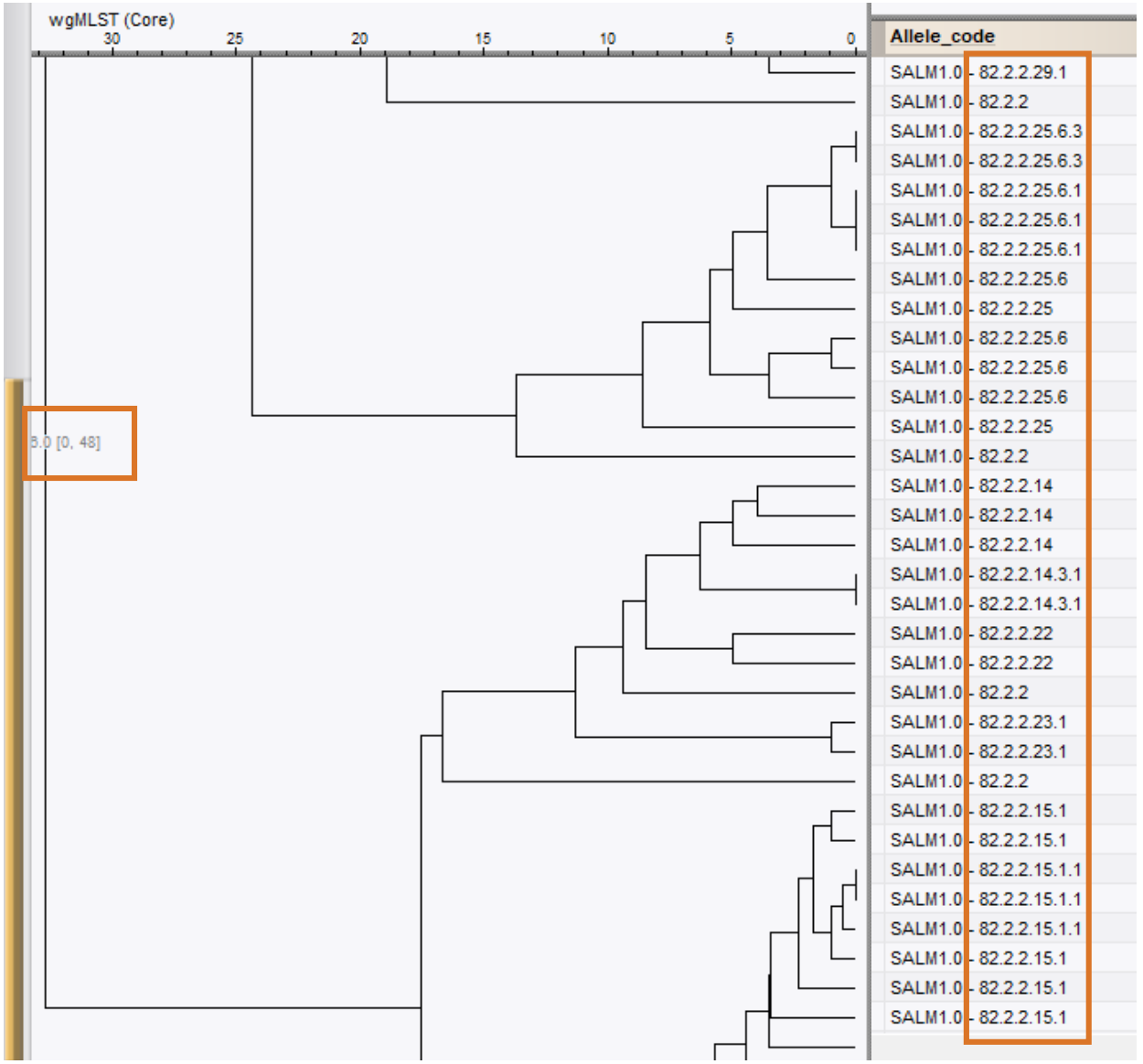
all share 4th digit, which would normally indicate up to **7** alleles different

X codes

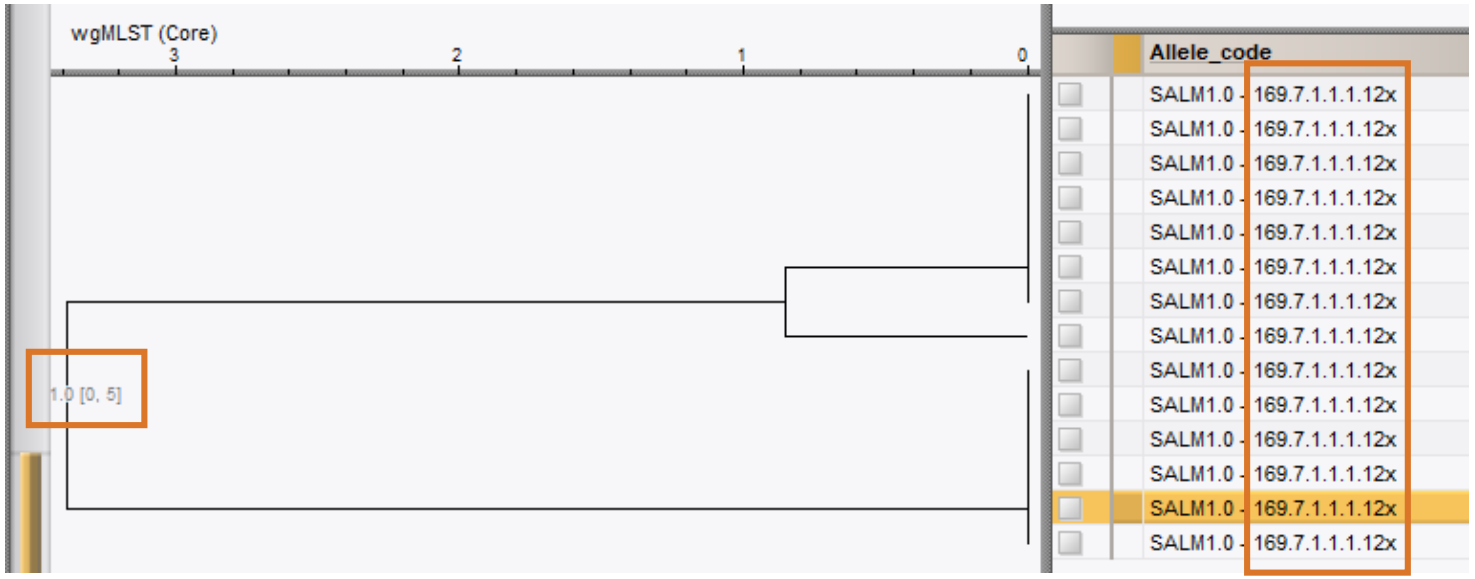
- Groups of isolates with differences >~4x the expected threshold are given X codes
  - Some clonal lineages merge together at a shared digit on the single linkage tree, creating more allele differences than you would expect at that threshold
- E.g., SALM1.1 – 6.6.6.6x.8x.32782
- primarily in *Salmonella* 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

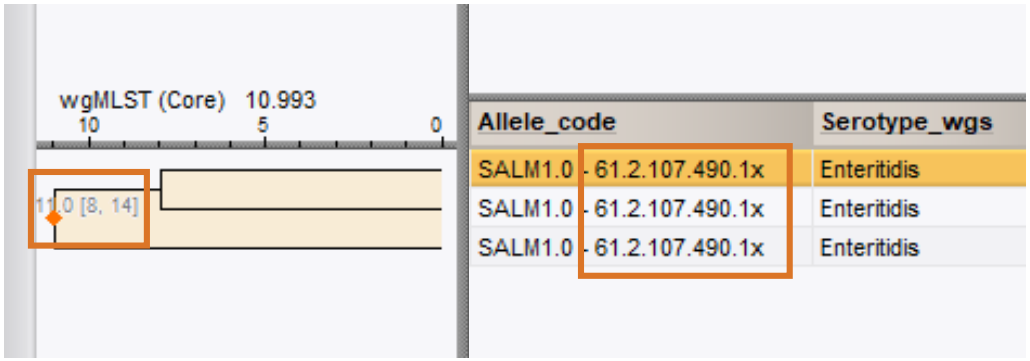


X codes



- 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

- Manually assigned monthly

Only assigned at 4th or 5th digit

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

New Version

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

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

- Automatically assigned

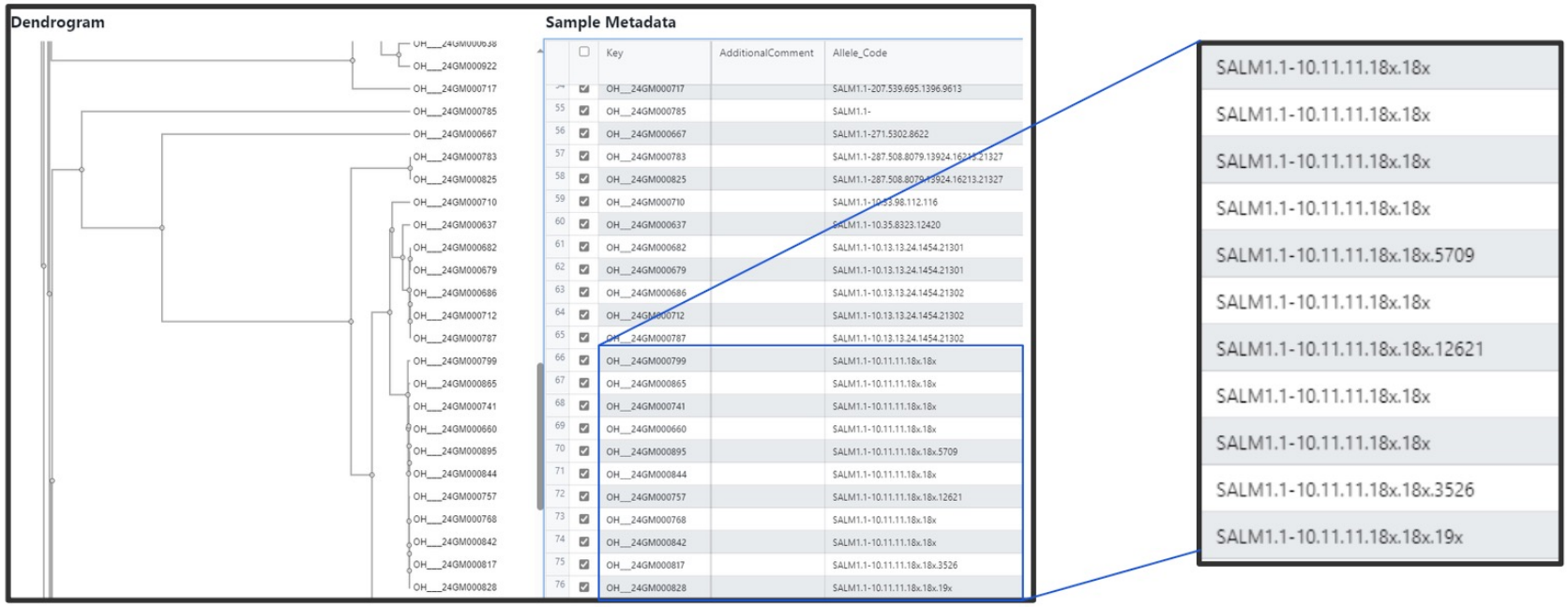
Assigned to all digits >4x their expected allelic threshold

Can continue past the x

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

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



Thresholds for Cluster Detection

	Local	National
Campylobacter	**≥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**
Escherichia	**≥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**.
Salmonella	**≥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**)
Shigella		Same as Escherichia, but may include cases within a wider allele range due to person-to-person transmission
Listeria	**≥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

Next Sessions

2

REP Strains

- REP codes
- Documenting & investigating local REP strains

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)

3

Data Visualization

- Tableau dashboards
- SaTScan

Communication

Communicating WGS results/findings in an outbreak investigation to stakeholders

Standard phrasing

PulseNet 2.0

- how to obtain access

Questions/Discussion

WGS Resources

TN CoE Community of Practice (CoP)

- Discussion-based and collaborative forum for getting information, posing questions, sharing successes, receiving support, sharing tips, and considering best practices.
- TN CoE partners facilitate discussions pertaining to specific topics or go over actual investigations.
- Regional partners can share clusters they are currently investigating or share past investigations that were interesting or informative.
- More info: http://foodsafety.utk.edu/?page_id=90
- Next session: August 4th, 2025

WGS Consultations

- SMEs from the University of Tennessee are available to provide technical assistance, answer questions, and provide information, as needed.
- More info: http://foodsafety.utk.edu/?p=813

Future Trainings and Other Resources

- TN Food Safety CoE website
- CoE Products website
- WGS Resources Sheet