bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io


Creative Commons
This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska
中文 華語 (台灣) isiZulu
creative commons
Attribution-Share Alike 2.5 Canada
You are free:
to Share — to copy, distribute and transmit the work
to Remix — to adapt the work
Free Cultural
APPROVED FOR
Works
Under the following conditions:
Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.
For any reuse or distribution, you must make clear to others the licence terms of this work.
Any of the above conditions can be waived if you get permission from the copyright holder.
The author's moral rights are retained in this licence.
Disclaimer
Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French
Learn how to distribute your work using this licence

Module 6
WGS-based subtyping

Eduardo Taboada, Jimmy Liu & Guangzhi Zhang

Infectious Disease Genomic Epidemiology

May 15-16, 2024

bioinformatics.ca

SISTR:
the Salmonella
in silico typing
resource

University of Manitoba

Public Health Agency of Canada

Agence de la santé publique du Canada

SFU
SIMON FRASER UNIVERSITY

Introductions…

Who?

- Ed Taboada (eduardo.taboada@phac-aspc.gc.ca)

Affiliation?

- Public Health Agency of Canada
  - National Microbiology Laboratory
    - Division of Enteric Diseases
      - Surveillance, Outbreak Detection & Response Section
        - Genomic Epidemiology Research Unit

What?

- research & methods development for studying the ecology and epidemiology of bacterial foodborne pathogens

Learning Objectives of Module

- A basic introduction to molecular subtyping:
  - role in molecular epidemiology & epidemiological surveillance
- WGS-based subtyping: from sequencing reads to subtypes
  - *In silico* subtyping
  - Multi Locus Sequence Typing (MLST): MLST "classic", wgMLST, cgMLST, MLST pros/ cons
  - Assembly-free approaches: Mash
- Analysis:
  - clustering
  - visualization
  - analysis of clusters & interpretation
- Genomic surveillance:
  - bacterial population structure & clonal complexes
  - what is a nomenclature
  - role of WGS-based subtyping

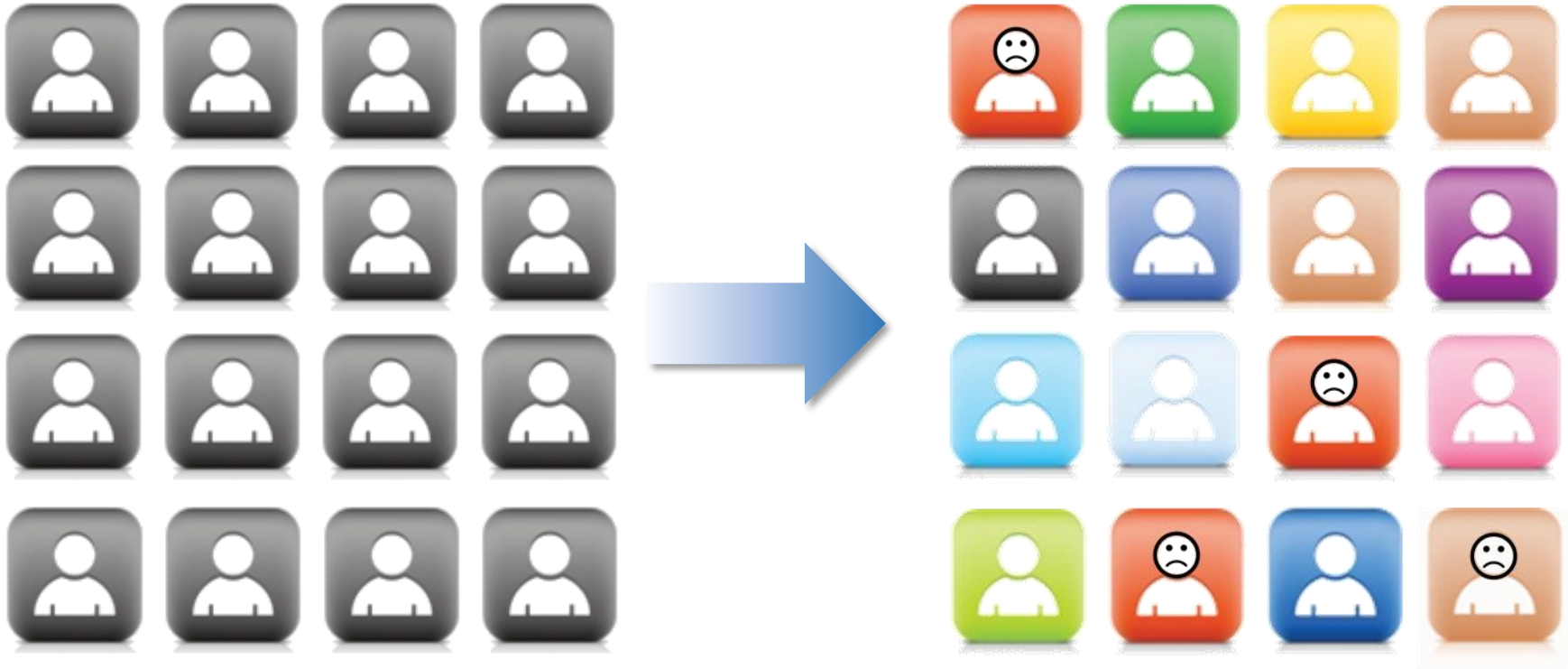
Learning Objectives of Module

- A basic introduction to molecular subtyping:
  - role in molecular epidemiology & epidemiological surveillance
- WGS-based subtyping: from sequencing reads to subtypes
  - *In silico* subtyping
  - Multi Locus Sequence Typing (MLST): MLST "classic", wgMLST, cgMLST, MLST pros/ cons
  - Assembly-free approaches: Mash
- Analysis:
  - clustering
  - visualization
  - analysis of clusters & interpretation
- Genomic surveillance:

- ❑ Please note, the slide deck is a fairly comprehensive overview
- ❑ We may skip some slides to keep to the timeline but they're there for context

A short primer on molecular subtyping in epidemiology…

Infectious disease isn't homogeneously distributed…



- Epidemiology:
  - distribution of exposure to risk factors
  - distribution of disease outcomes

Molecular epidemiology ?

"…addresses epidemiologic problems that cannot be approached or would be more labor intensive, expensive, and/or time consuming to address by conventional techniques."

"…the application of molecular taxonomy, phylogeny, or population genetics to epidemiologic problems."

Foxman and Riley, *Am J Epidemiol* **153**:1135.

- Harnesses molecular approaches to identify and characterize infectious disease agents so that we may examine their distribution and infer their transmission

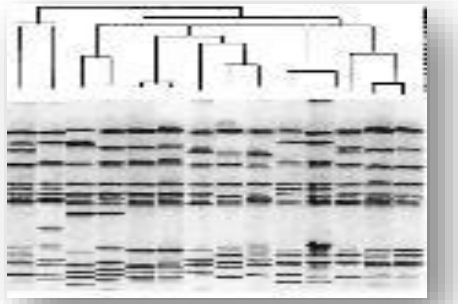
The molecular epidemiology paradigm

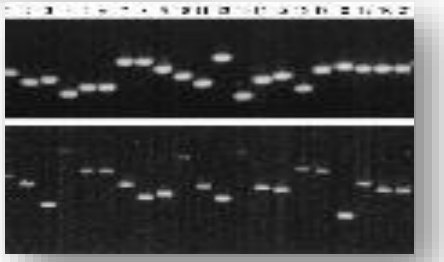
- A fairly simple set of guiding principles:
  - The null hypothesis is to expect agreement between genetic observations on the pathogen and epidemiologic observations on the case
    - If isolates are epidemiologically linked, they should be “genetically identical”
    - If isolates are not epidemiologically linked, they should not be “genetically identical”
  - Use molecular approaches (i.e. **molecular subtyping**) to assess genetic similarity between isolates

The molecular subtyping paradigm

- In the beginning, methods based on phenotypic characteristics
  - Serotyping (i.e. serological)
  - Biotyping (i.e. biochemical)

- In the 90s-00s, emergence of methods based on comparing variation in DNA banding patterns on gels (i.e. “DNA fingerprints”)
  - Pulsed Field Gel Electrophoresis (PFGE)
  - Multiple-Locus VNTR Analysis (MLVA)
  - Amplified Fragment Length Polymorphisms (AFLP)
  - **etc…etc**





The molecular subtyping paradigm

- In the beginning, methods based on phenotypic characteristics
  - Serotyping (i.e. serological)
  - Biotyping (i.e. biochemical)

- In the 90s-00s, emergence of meth
  variation in DNA banding patterns
  fingerprints")
  - Pulsed Field Gel Electrophoresis (P
  - Multiple-Locus VNTR Analysis (ML

July 1996, p. 1870 Vol. 34, No. 7

or Microbiology

A Surfeit of YATMs? By Mark Achtman

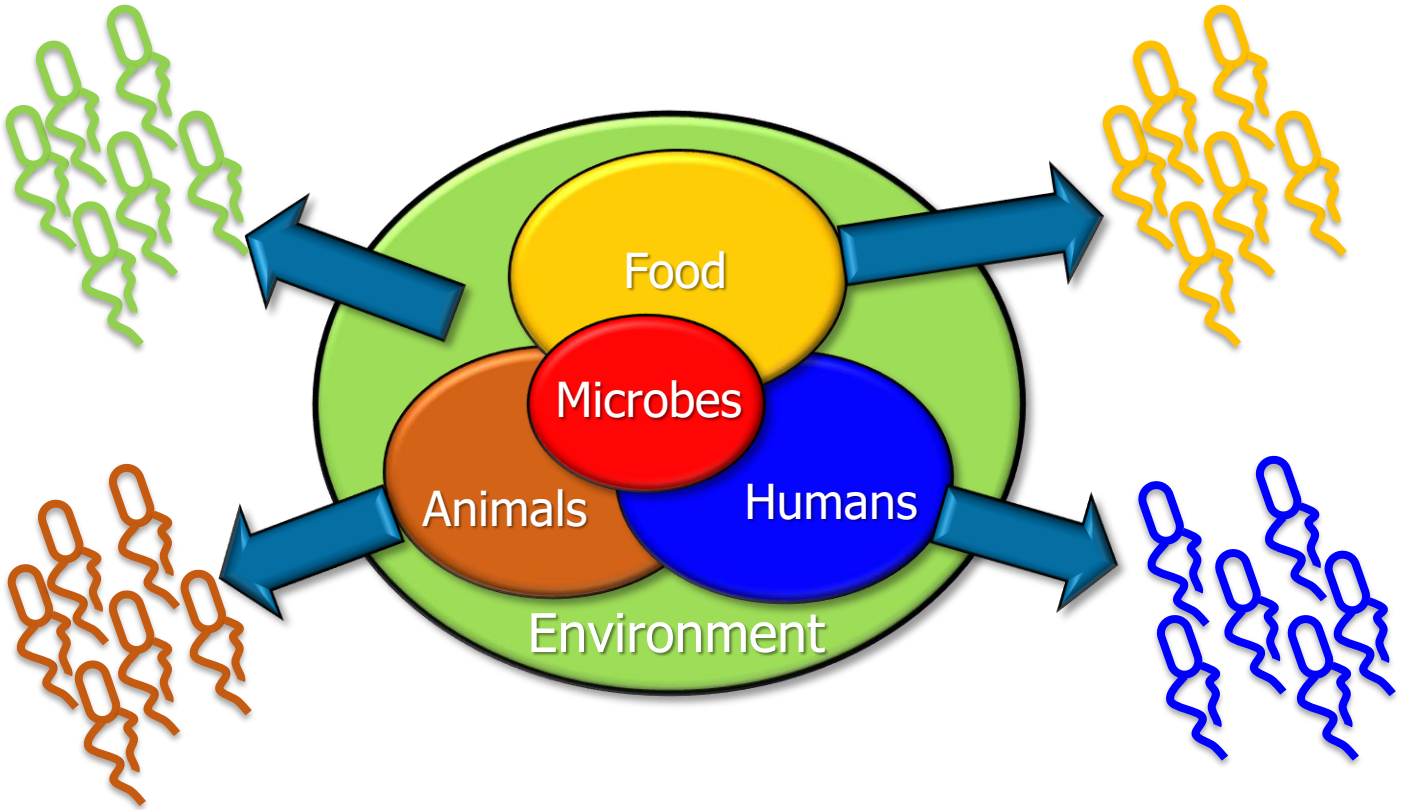
acronym for *Y*et *A*nother

tat' bistem\ *n* acronym for
*M*ethod
acronym for *T*otally *B*oring

YATMs are generally based on DNA technology (e.g., ribotyping, random amplified polymorphic DNA analysis, or pulsed-field gel electrophoresis polymorphism) and are designed to rapidly distinguish clonal epidemic outbreaks from hyperendemic disease levels caused by concurrent multiplication of unrelated strains. In many cases in which the results

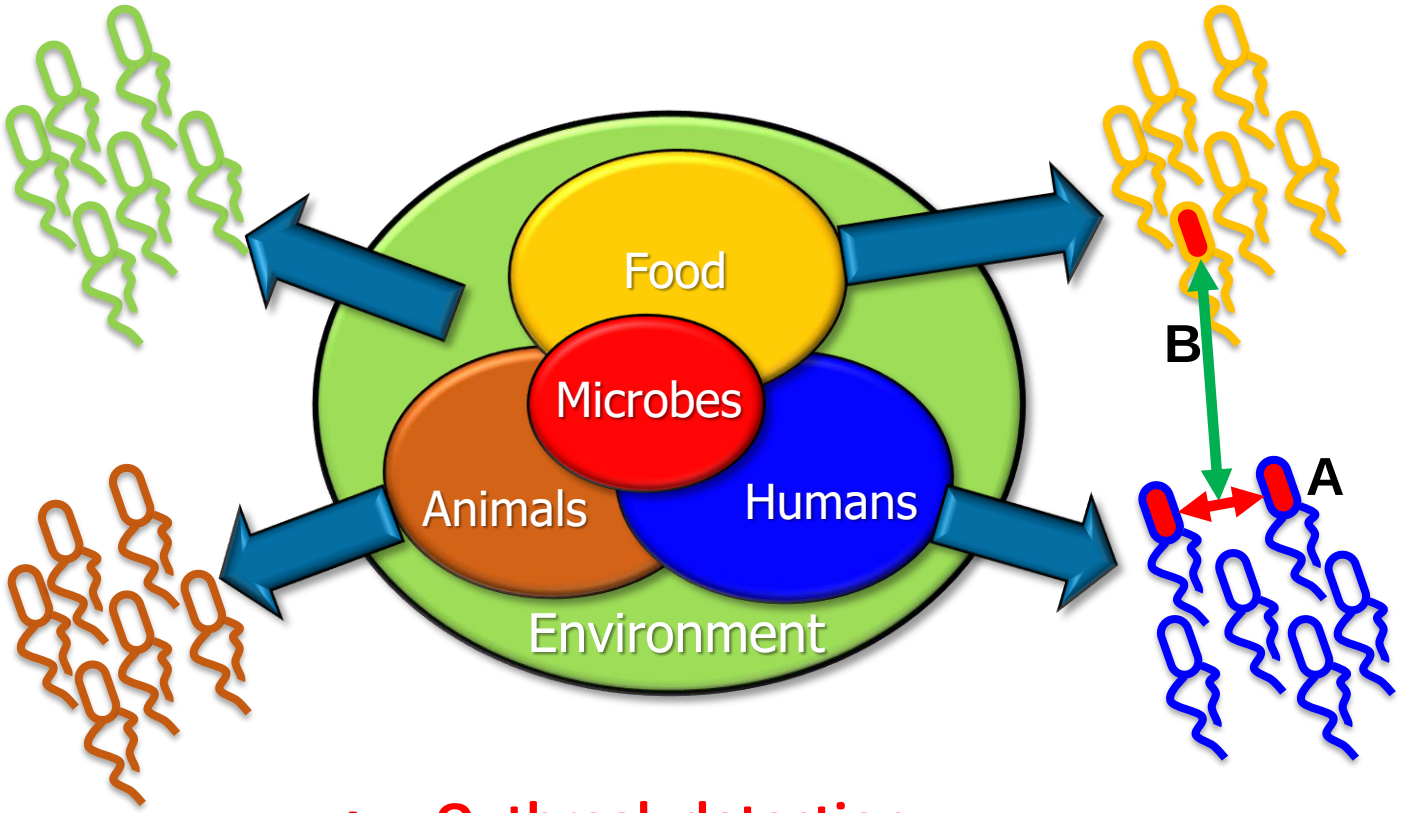
- Molecular subtyping methods were developed to estimate genetic similarity between microbial isolates at a time when genomics was in its infancy
- Proliferation of papers describing novel molecular methods with limited validation hoping for wide adoption by the scientific community: YATMs ("Yet another typing method")

Molecular epidemiology and microbiological surveillance



1) Collection of samples from potential sources of exposure
2) Recovery of microbial isolates and genetic analysis
3) Comparison of genetic data and identification of “matching” isolates
4) Examination of the epidemiology of “matching” isolates

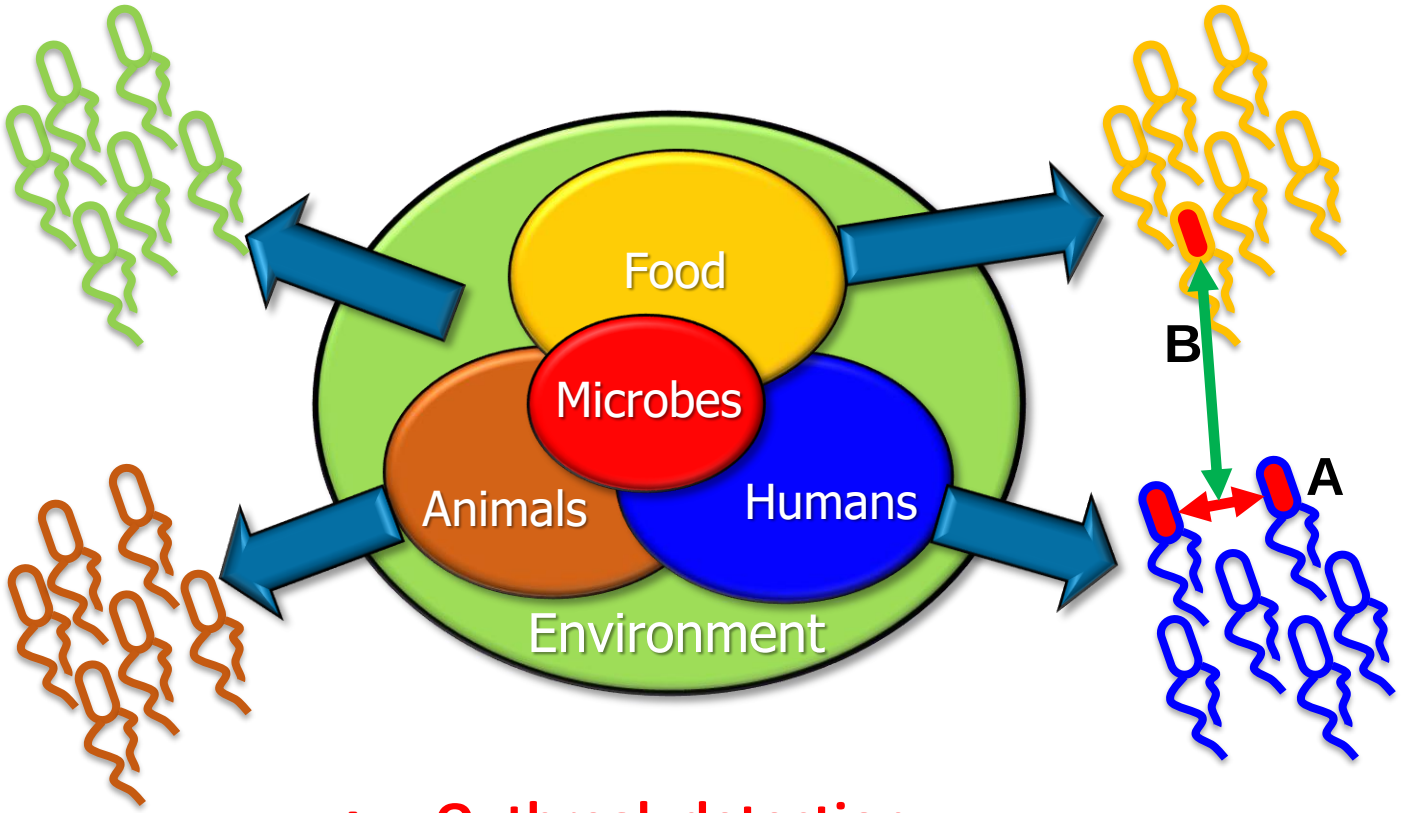
The role of molecular subtyping in molecular epidemiology



A. **Outbreak detection**
   - ➢ **strain from patient X = strain from patient Y?**

B. **Traceback & Source Attribution**
   - ➢ **strain in patient X = strain in source Y?**

The role of molecular subtyping in molecular epidemiology

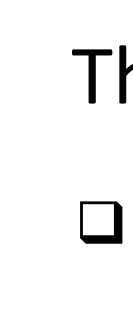


A Outbreak detection

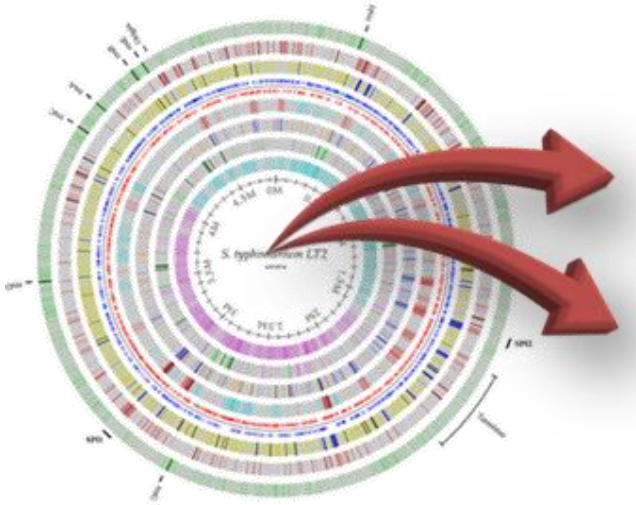
- Significant challenges due to limitations of various molecular subtyping methods in properly estimating genetic similarity
- Easy to **over-** or **under-**estimate genetic similarity and misinterpret the significance of matches or mismatches when using conventional subtyping approaches

WGS-based subtyping…

The role of WGS-based subtyping…

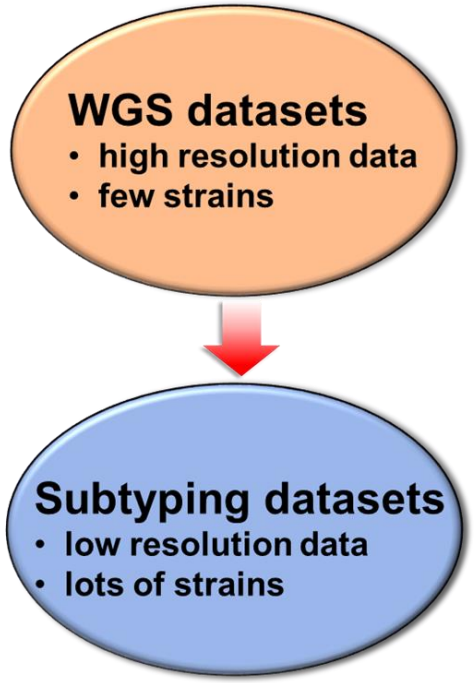


- 10 years ago, the major challenge was a scarcity of WGS data compared to the subtyping data in the major surveillance databases



Phenotype prediction
(e.g. serovar, AMR, etc…)

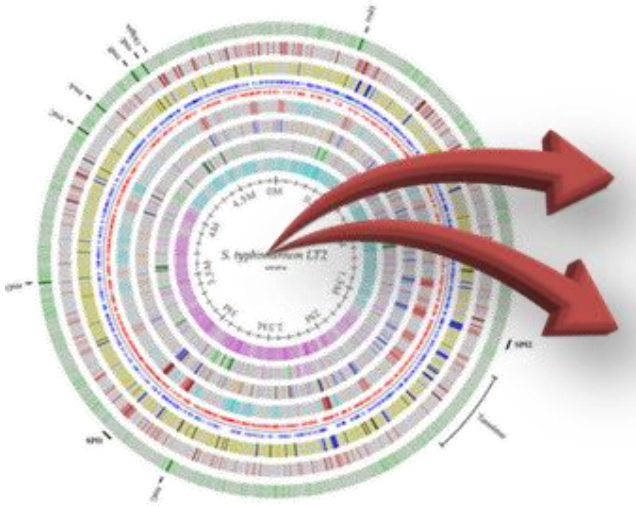
Genotype prediction
(e.g. MLST, MLVA, etc…)



- Typing by searching for the gene(s) or target(s) responsible for the subtype,
  → simultaneously look for genes/targets from multiple typing methods
- *In silico* typing to provide a link between legacy subtyping databases and WGS databases

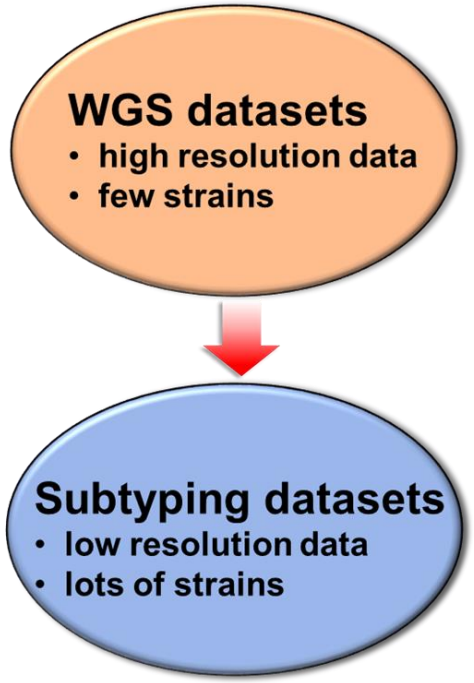
The role of WGS-based subtyping…

- 10 years ago, the major challenge was a scarcity of WGS data compared to the subtyping data in the major surveillance databases



Phenotype prediction
(e.g. serovar, AMR, etc…)

Genotype prediction
(e.g. MLST, MLVA, etc…)

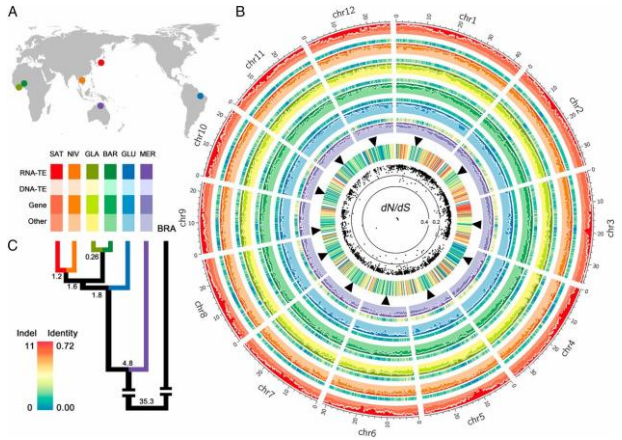


- Typing by searching for the gene(s) or target(s) responsible for the subtype,

- …Then Public Health England announced that they would be sequencing **every single** *Salmonella* isolate collected through clinical surveillance at the national level (and everyone else joined in)

The new WGS-based subtyping paradigm

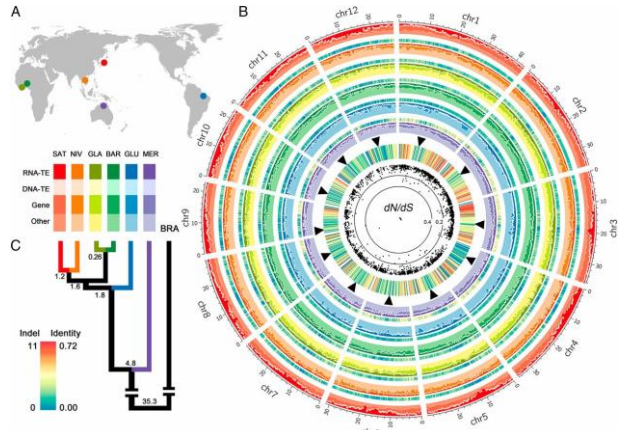
- ❑ Goal is still to estimate genetic similarity between isolates obtained through surveillance sampling
- ❑ The difference: the ability to use the full weight of WGS data to infer better estimates
- ❑ "Comparative genomics"
  - **Regional-level** variation
    - Strain-specific chromosomal regions (e.g. genes, non-coding regions, genomic islands)
    - Strain-specific extra-chromosomal regions (e.g. plasmids)
    - VNTRs (tandem repeats)
  - **Gene-level** variation
    - Allelic differences
  - **Single nucleotide-level** variation
    - Single nucleotide variants (SNVs) & Single nucleotide polymorphisms (SNPs)



https://doi.org/10.1073/pnas.1418307111

The new WGS-based subtyping paradigm

- ❑ Goal is still to estimate genetic similarity between isolates obtained through surveillance sampling
- ❑ The difference: the ability to use the full weight of WGS data to infer better estimates
- ❑ "Comparative genomics":
  - **Region-level** variation
    - Strain-specific chromosomal regions (e.g. genes, non-coding regions, genomic islands)
    - Strain-specific extra-chromosomal regions (e.g. plasmids)
    - VNTRs (tandem repeats)
  - **Gene-level** variation
    - Allelic differences
  - **Single nucleotide-level** variation



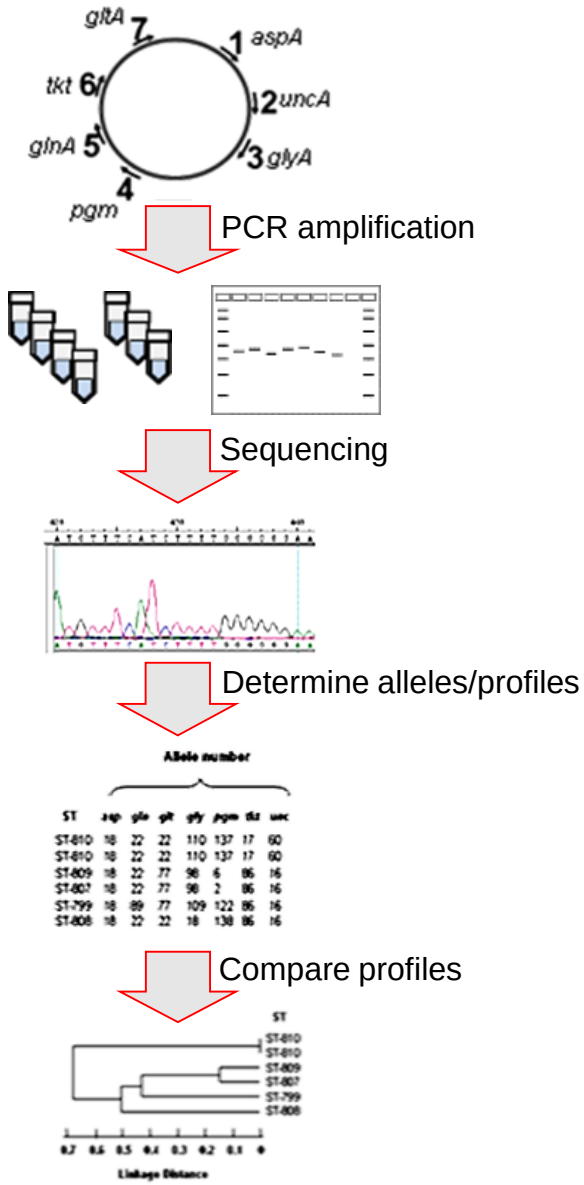
https://doi.org/10.1073/pnas.1418307111

- ❑ Development of methods for extracting & indexing various types of sequence variation information in WGS data in order to compute higher resolution genetic similarity estimates

Sequence-based typing: Multi Locus Sequence Typing (MLST)

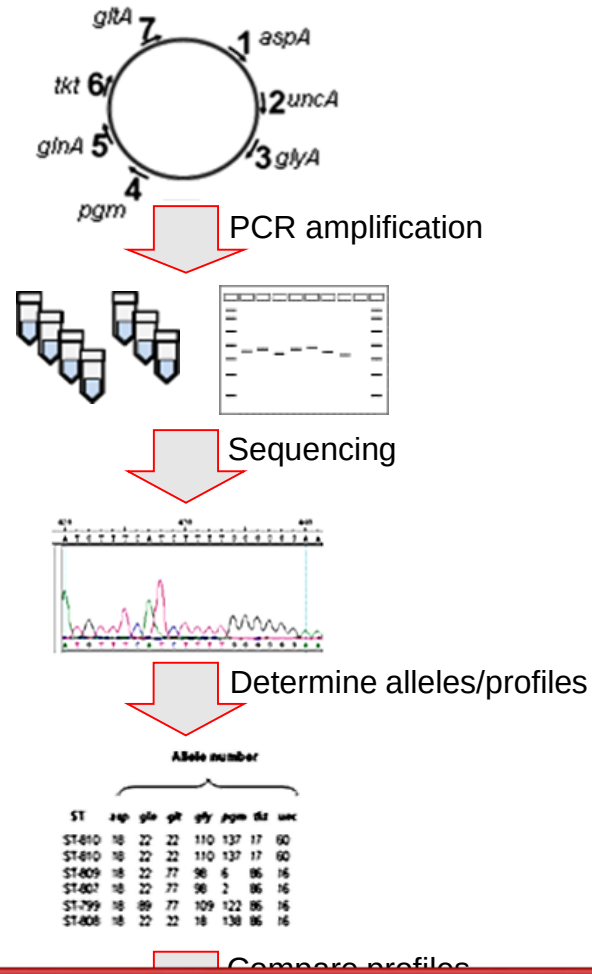
Multi **L**ocus **S**equence **T**yping

- Described by Maiden et al. (1998) *PNAS USA* **95**: 3140.
  - Analysis of 7 to 9 loci
    - “housekeeping” genes (i.e. core genes)
    - loci are distributed around the genome
  - Each locus is PCR amplified and sequenced
    - 450-500 bp gene fragments
    - Each locus is PCR amplified and sequenced
  - Analysis of allelic data
    - for each locus we determine the **allele** based on finding a match in a centralized database (MLSTdb)
    - a **Sequence Type** (ST) is assigned based on the **combination of alleles** at all the loci
    - groups of related STs are assigned to a **Clonal Complex**
  - The MLSTdb keeps track of all known alleles and STs
    - any **novel allele** is assigned a new (n+1) allele number
    - any **novel ST** is assigned a new (n+1) ST number



Multi **L**ocus **S**equence **T**yping

- Described by Maiden et al. (1998) *PNAS USA* **95**: 3140.
  - Analysis of 7 to 9 loci
    - "housekeeping" genes (i.e. core genes)
    - loci are distributed around the genome
  - Each locus is PCR amplified and sequenced
    - 450-500 bp gene fragments
    - Each locus is PCR amplified and sequenced
  - Analysis of allelic data
    - for each locus we determine the **allele** based on finding a match in a centralized database (MLSTdb)
    - a **Sequence Type** (ST) is assigned based on the **combination of alleles** at all the loci
    - groups of related STs are assigned to a **Clonal Complex**

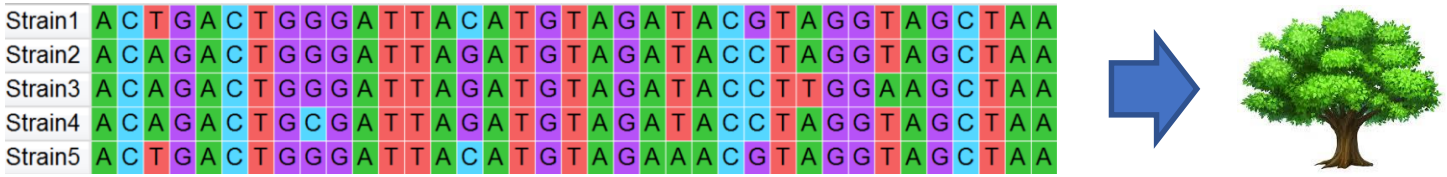


- MLST became the gold standard method in molecular epidemiology
  e.g. sequence data vs. fragments on gels
- Over 50 schemes were developed for different species & used in hundreds of published studies

There is something curious about MLST…

MLST isn't technically a "proper" phylogenetic analysis

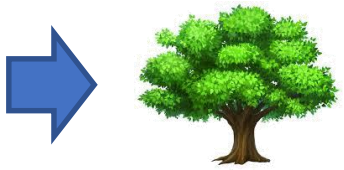
- Phylogenetic analysis of **one** MLST locus
  - ≈ 450 bp → analysis of ≈ 450 data points **with** evolutionary modelling



- Phylogenetic analysis of **seven** MLST loci
  - ≈ 450 bp x 7 → analysis of ≈ 3,150 data points **with** evolutionary modelling

- (Actual) MLST analysis
  - comparison of **seven** data points **without** evolutionary modelling

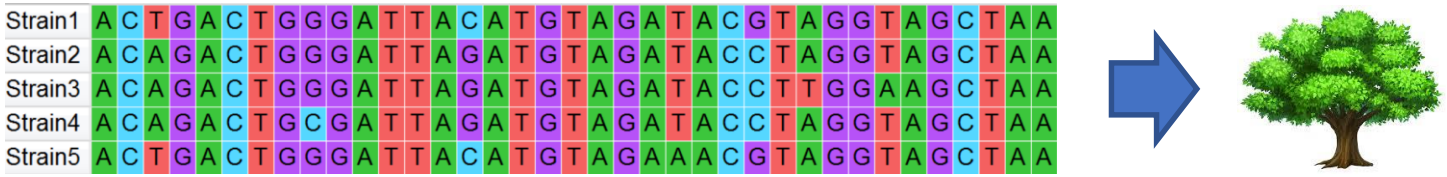
	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
Strain1	2	3	7	4	1	14	1
Strain2	1	12	13	2	5	9	2
Strain3	1	12	5	2	5	2	10
Strain4	1	12	13	2	5	7	2
Strain5	2	5	7	8	1	14	1



There is something curious about MLST…

MLST isn't technically a "proper" phylogenetic analysis

- Phylogenetic analysis of **one** MLST locus
  - ≈ 450 bp → analysis of ≈ 450 data points **with** evolutionary modelling



- Phylogenetic analysis of **seven** MLST loci
  - ≈ 450 bp x 7 → analysis of ≈ 3,150 data points **with** evolutionary modelling

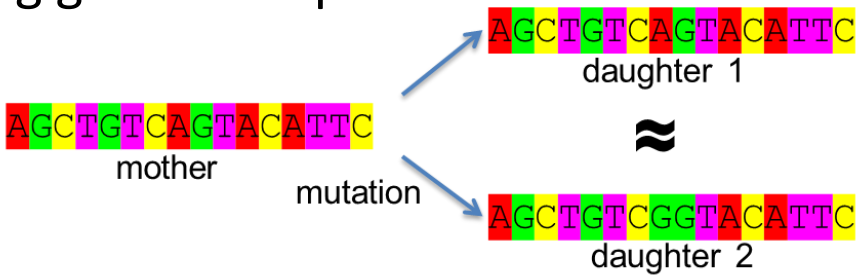
- (Actual) MLST analysis
  - comparison of **seven** data points **without** evolutionary modelling

- In MLST analysis, the sequence at a locus is reduced to a single allele type: **no weight is given to the number of nucleotide differences, no evolutionary model is used to analyze substitution patterns**
- Strains are compared on the basis of the number of matching alleles at 7 loci

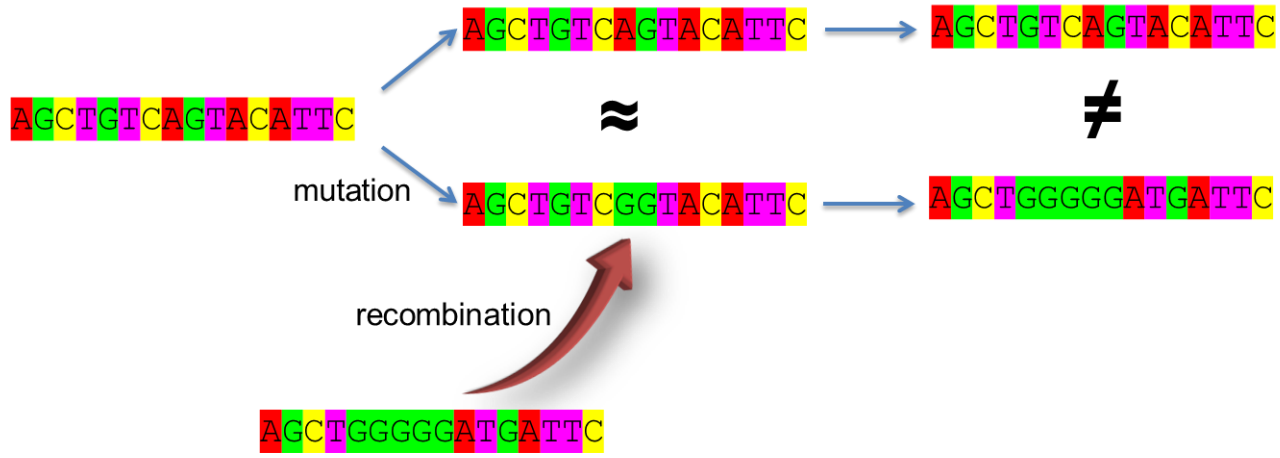
A short detour into microbial (sequence) evolution…

Mutation and recombination in phylogenetic analysis

- Phylogenetic analysis assumes that sequences primarily evolve via the accumulation of mutations during genome replication

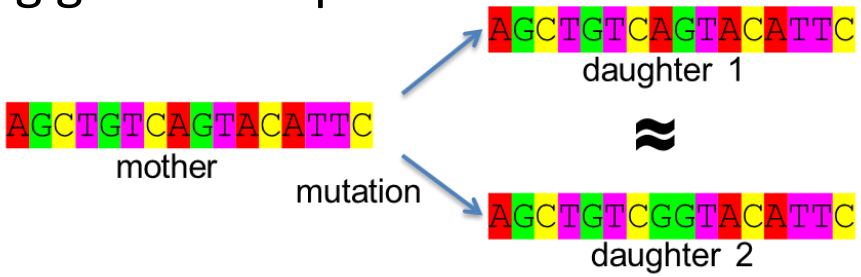


- In recombinogenic species, recombination can lead to erroneous estimates of phylogenetic distance due to allelic replacement



Mutation and recombination in phylogenetic analysis

- Phylogenetic analysis assumes that sequences primarily evolve via the accumulation of mutations during genome replication



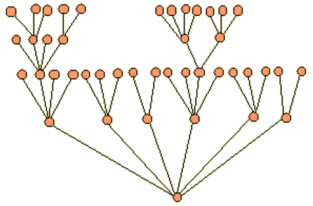
- In recombinogenic species, recombination can lead to erroneous estimates of phylogenetic distance due to allelic replacement

- Mutation during DNA replication is a process for <u>vertical</u> transfer of genetic information → phyogenetic signal
- Recombination is a process for <u>lateral</u> transfer of genetic information (i.e. no need for common ancestry) → **distorted** phylogenetic signal

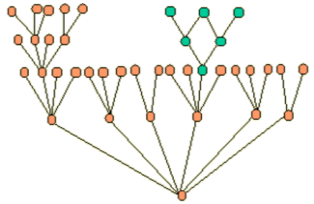
Microbial Population Structure (part 1)

"…Differences in the ratio of genetic change caused by recombination relative to *de novo* mutation leads to a spectrum of population structures, from the extremes of strictly ***clonal***, where effectively no recombination has occurred in the evolutionary history of the species, to non-clonal, or ***panmictic***, where recombinational exchanges are sufficiently frequent to randomize the alleles in the population and to prevent the emergence of stable clones…."

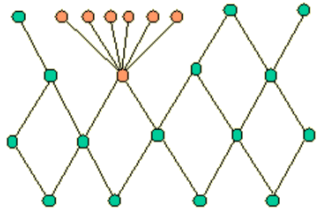
Spratt and Maiden (1999) *Phil Trans R Soc Lond* **354**: 701.



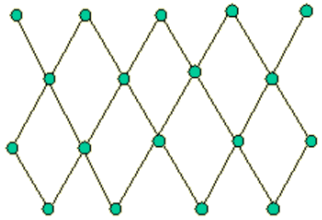
Clonal



Weakly Clonal



Epidemic



Panmictic

Microbial Population Structure (part 1)

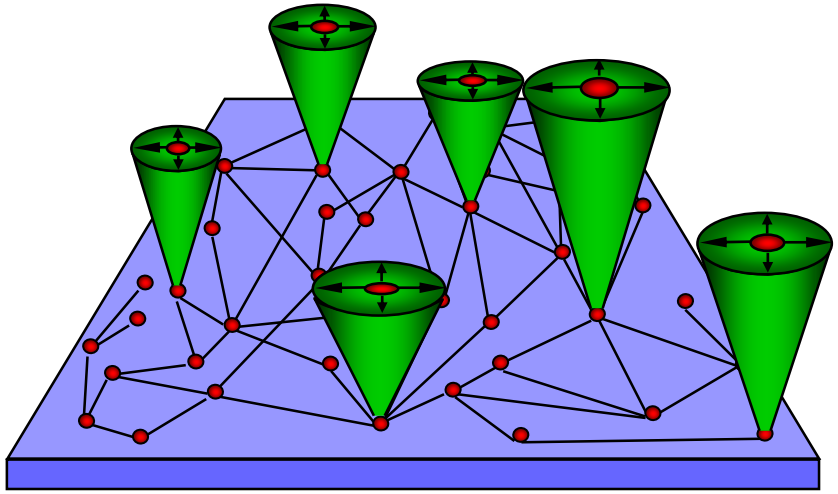
"…Differences in the ratio of genetic change caused by recombination relative to *de novo* mutation leads to a spectrum of population structures, from the extremes of strictly ***clonal***, where effectively no recombination has occurred in the evolutionary history of the species, to non-clonal, or ***panmictic***, where recombinational exchanges are sufficiently frequent to randomize the alleles in the population and to prevent the emergence of stable clones…."

Spratt and Maiden (1999) *Phil Trans R Soc Lond* **354**: 701.

- ❑ Microbial populations can be composed of clonal lineages that slowly diversify through mutation, recombination, or both
- ❑ Contribution of recombination vs. *de novo* mutation varies by species
- ❑ High recombination has a distorting effect on evolutionary relationships and their inference

Microbial Population Structure (part 2)

- Many pathogens exhibit an “epidemic” population structure
  - Lots of “rare” genotypes in circulation
    - Significant genetic exchange through recombination (i.e. allelic replacement)
    - More network-like
  - Distinct “clones” (i.e. strains) rise in prominence via “clonal expansion”
    - Clones undergo genetic diversification through recombination and mutation
    - More tree-like

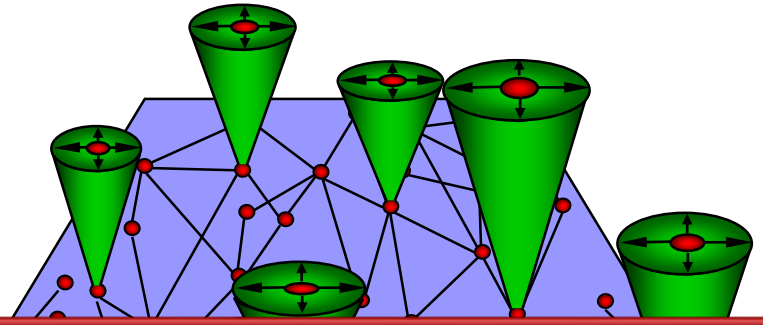


“…the background population is composed of a large number of relatively rare and unrelated genotypes (small circles) that are recombining at a high frequency…most accurately represented as a **network**, rather than a bifurcating tree, as *recombination has overwhelmed the phylogenetic signal*…a limited number of very frequent genotypes, or clusters of closely related genotypes, illustrated as cones. These are **clonal complexes**, and typically emerge from a single, highly adaptive, ancestral genotype (the large circles).”

adapted from Maynard Smith (2000) *BioEssays* **22**: 1115.

Microbial Population Structure (part 2)

- Many pathogens exhibit an "epidemic" population structure
  - Lots of "rare" genotypes in circulation
    - Significant genetic exchange through recombination (i.e. allelic replacement)
    - More network-like
  - Distinct "clones" (i.e. strains) rise in prominence via "clonal expansion"
    - Clones undergo genetic diversification through recombination and mutation
    - More tree-like



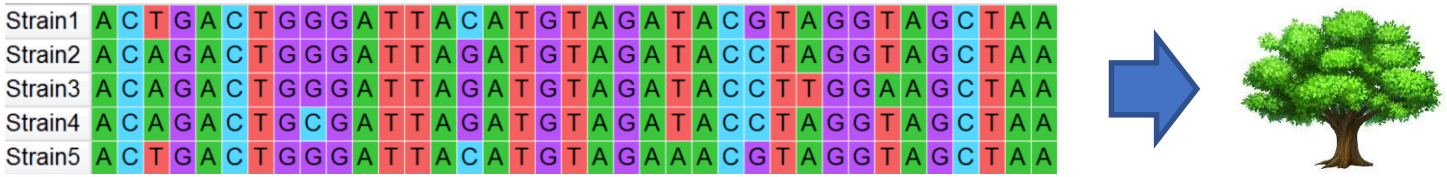
"…the background population is composed of a large number of relatively rare and unrelated genotypes (small circles) that are recombining at a high frequency…most accurately represented as a **network**, rather than a bifurcating tree, as *recombination has overwhelmed the phylogenetic signal*…a limited number of very frequent genotypes or clusters of closely related genotypes

- For many pathogens, phylogenetic relationships between these clones are difficult to elucidate because of the distorting effects of recombination
  → More important to be able to identify these clones

There is something curious about MLST…

MLST isn't technically a "proper" phylogenetic analysis

- Phylogenetic analysis of **one** MLST locus
  - ≈ 450 bp → analysis of ≈ 450 data points **with** evolutionary modelling



- Phylogenetic analysis of **seven** MLST loci
  - ≈ 450 bp x 7 → analysis of ≈ 3,150 data points **with** evolutionary modelling

- (Actual) MLST analysis
  - comparison of **seven** data points **without** evolutionary modelling

- In MLST analysis, the sequence at a locus is reduced to a single allele type: **no weight is given to the number of nucleotide differences, no evolutionary model is used to analyze substitution patterns**
- Strains are compared on the basis of the number of matching alleles at 7 loci

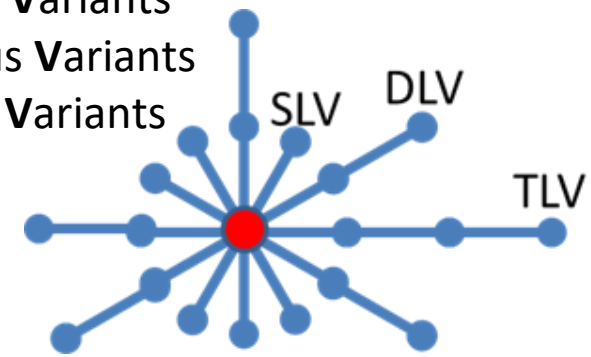
Bacterial Population Structure and MLST

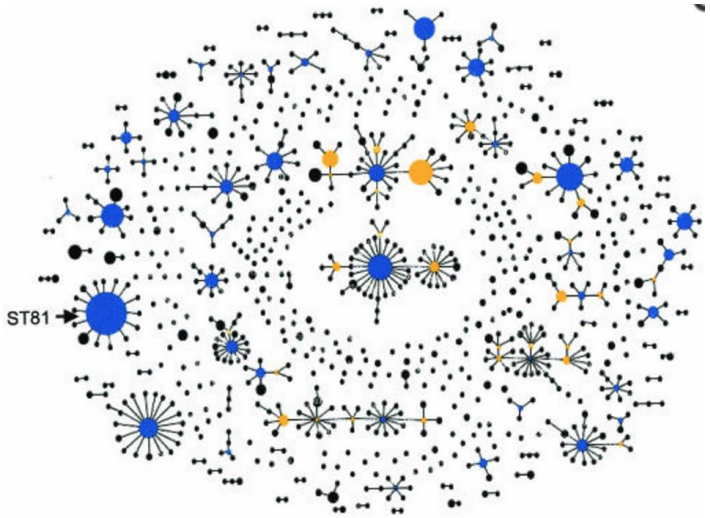
- BURST (**B**ased **U**pon **R**elated **S**equence **T**ypes): clustering algorithm for MLST developed by Feil *et al.* (2004) based on epidemic clone model.
  - Identifies groups of related Sequence Types defined by a certain number of shared alleles (e.g. 4 out of 7)
  - **eBURST**: advanced implementation with Minimum-Spanning Tree visualization

    → eBURST group ≈ Clonal Complex
  - **goeBURST**: globally optimized eBURST developed by Francisco *et al.* (2009)

<u>eBURST group</u>

- A founding (ancestral) Sequence Type
- Related Sequence Types
  - **S**ingle **L**ocus **V**ariants
  - **D**ouble **L**ocus **V**ariants
  - **T**riple **L**ocus **V**ariants





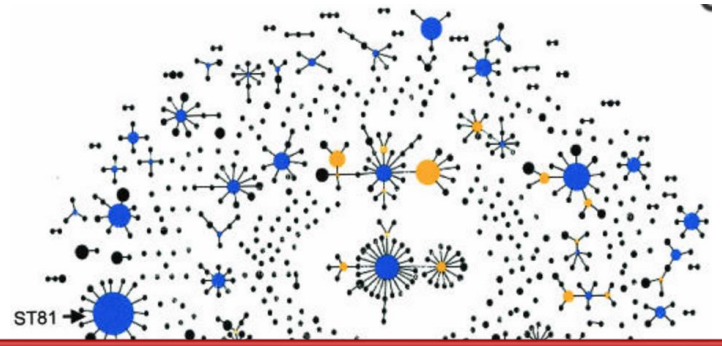
eBURST analysis of *S. pneumoniae* population structure. Feil et al. *J Bacteriol* **186**(5): 1518.

Bacterial Population Structure and MLST

- BURST (**B**ased **U**pon **R**elated **S**equence **T**ypes): clustering algorithm for MLST developed by Feil *et al.* (2004) based on epidemic clone model.
  - Identifies groups of related Sequence Types defined by a certain number of shared alleles (e.g. 4 out of 7)
  - **eBURST**: advanced implementation with Minimum-Spanning Tree visualization
    → eBURST group ≈ Clonal Complex
  - **goeBURST**: globally optimized eBURST developed by Francisco *et al.* (2009)

<u>eBURST group</u>

- A founding (ancestral) Sequence Type
- Related Sequence Types
  - **S**ingle **L**ocus **V**ariants
  - **D**ouble **L**ocus **V**ariants
  - **T**riple **L**ocus **V**ariants



- Many bacterial species do not generate tree-like phylogenies because different areas of the genome may generate incongruent phylogenetic signal due to recombination
- For species with an epidemic population structure eBURST analysis of MLST data is very useful because it is not impacted by recombination

The MLST nomenclature

- ❑ Each locus is assigned an allele number by finding its match in the central MLST database (MLSTdb)
- ❑ A Sequence Type is assigned based on the combination of alleles at all loci, also in the MLSTdb
- ❑ Clonal Complexes are assigned based on eBURST analysis
- ❑ Each novel allele/Sequence Type is assigned a number in order of discovery (i.e. n+1)

Clonal Complex	ST	Asp	Gln	Glt	Gly	Pgm	Tkt	Unc
ST_21	21	2	1	1	3	2	1	5
ST_21	8	2	1	1	3	2	1	6
ST_21	50	2	1	12	3	2	1	5
ST_21	141	2	1	10	3	2	1	5
ST_21	262	2	1	1	3	2	1	3
ST_21	917	2	21	1	3	2	1	5
ST_21	982	2	1	2	3	2	1	5
ST_21	806	2	1	1	3	140	3	5
ST_21	2513	2	2	27	3	2	1	5
ST_21	3857	2	1	2	10	2	1	5

→ This allelic profile = Sequence Type 21 (ST 21)

→ ST 50 is a **Single Locus Variant** (SLV) of ST 21 because it is different at one locus with respect to ST 21 (Glt-12)

→ The allele at the Tkt locus for ST 982 is Tkt-1

→ ST 2513 is a **Double Locus Variant** (DLV) of ST 21 because it is different at two loci with respect to ST 21 (Gln-2 & Glt-27)

→ The **Clonal Complex** ST-21 (i.e. the ST-21 Complex) includes the hypothetical founder (ST-21) and various related STs (8, 50, 141, 262, etc…)

The MLST nomenclature

- Each locus is assigned an allele number by finding its match in the central MLST database (MLSTdb)
- A Sequence Type is assigned based on the combination of alleles at all loci, also in the MLSTdb
- Clonal Complexes are assigned based on eBURST analysis
- Each novel allele/Sequence Type is assigned a number in order of discovery (i.e. n+1)

Clonal Complex	ST	Asp	Gln	Glt	Gly	Pgm	Tkt	Unc
ST_21	21	2	1	1	3	2	1	5
ST_21	8	2	1	1	3	2	1	6
ST_21	50	2	1	12	3	2	1	5
ST_21	141	2	1	10	3	2	1	5
ST_21	262	2	1	1	3	2	1	3
ST_21	917	2	21	1	3	2	1	5
ST_21	982	2	1	2	3	2	1	5
ST_21	806	2	1	1	3	140	3	5

→ This allelic profile = Sequence Type 21 (ST 21)

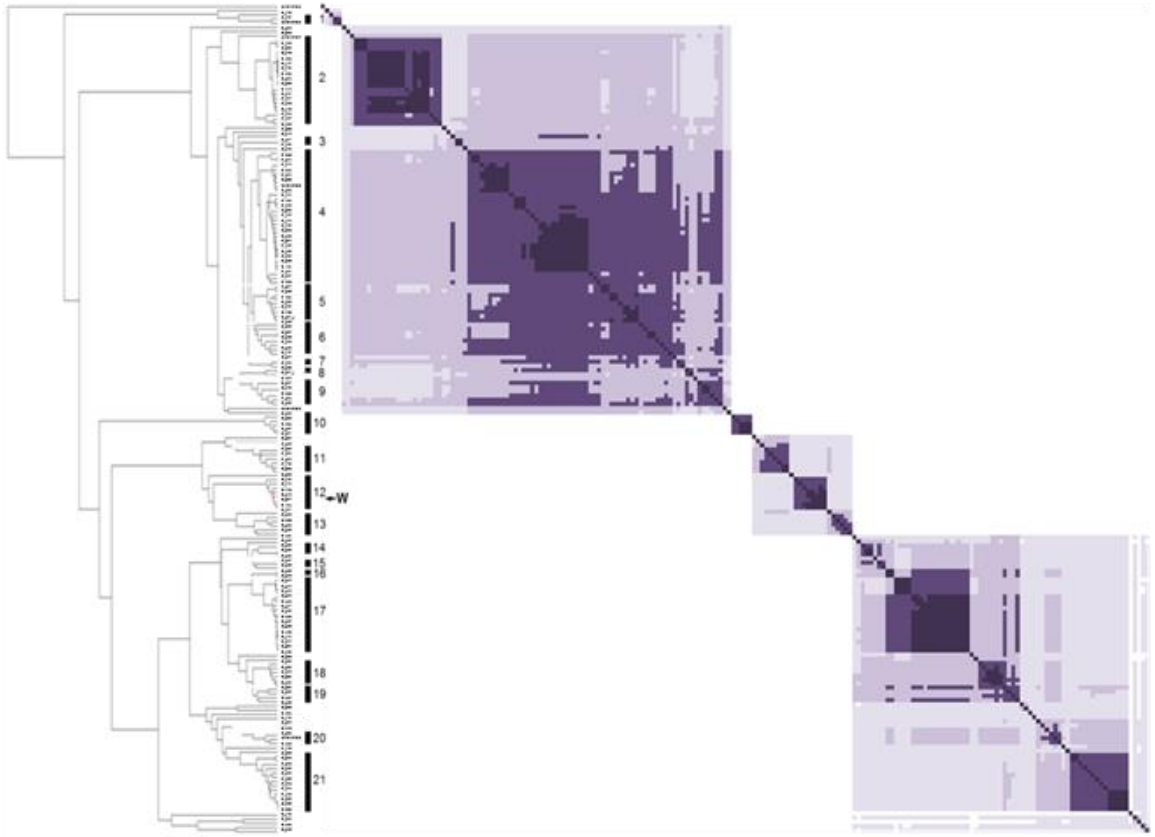
→ ST 50 is a **Single Locus Variant** (SLV) of ST 21 because it is different at one locus with respect to ST 21 (Glt-12)

→ The allele at the Tkt locus for ST 982 is Tkt-1

- Nomenclatures are an important concept because they represent a systematic approach to unambiguously defining and naming various genetic lineages and their sublineages
- Nomenclatures are what can make a subtyping approach “portable”: everyone knows what the various subtypes represent and the subtype “names” are all that is needed to compare across labs

The problem with MLST…

- Strains that are “identical” by MLST may be very genomically different
- 7 genes represents a tiny fraction of the data present in the whole genome of a strain → limited information content
- Most MLST datasets are overrepresented with a small number of prevalent STs → limited epidemiological usefulness

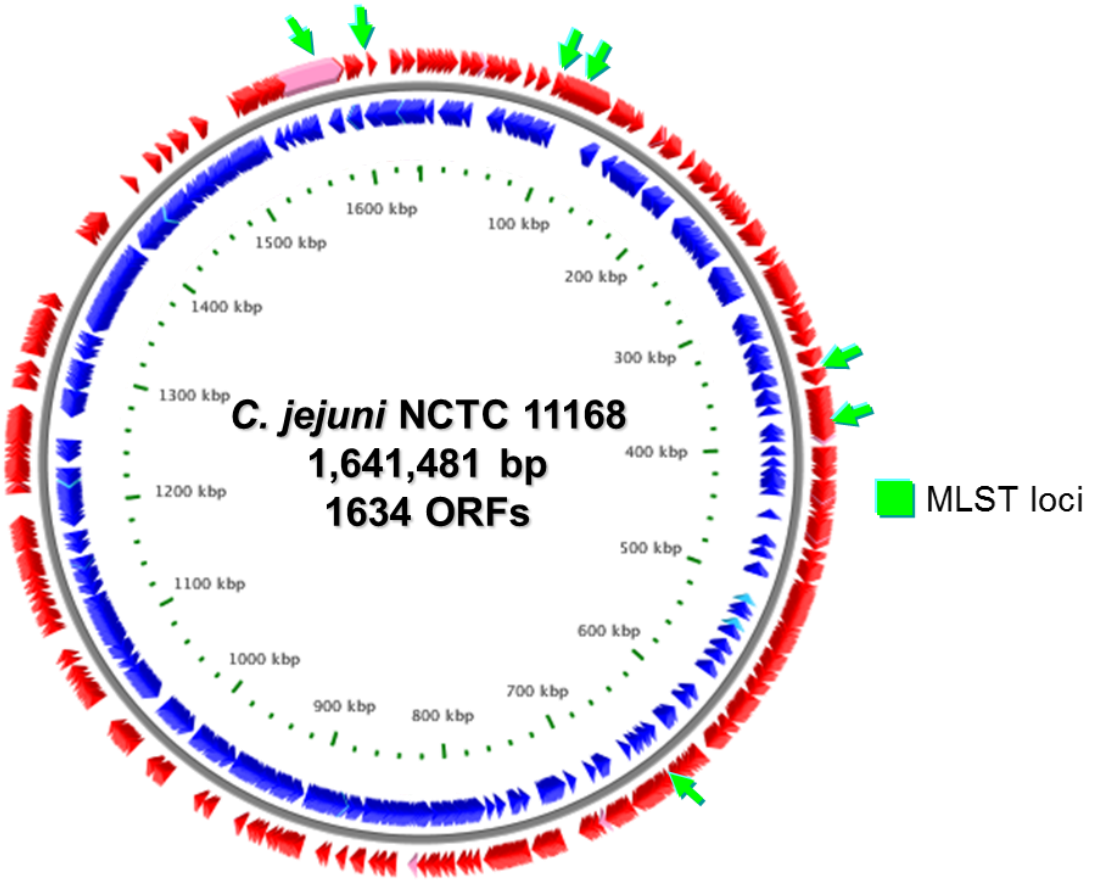


Analysis of *C. jejuni* isolates that share identical MLST profiles using an expanded set of 697 loci (Barker and Taboada, unpublished).

- MLST tends to be good for long-term tracking of lineages but often lacks sufficient power for certain epidemiological investigations, including outbreaks

The solution? genome-scale MLST

- Fast and inexpensive high-throughput sequencing platforms allow us to generate massive amounts of sequence data
- At a genome-scale, could extend the MLST concept to hundreds/thousands of loci
- Increasing availability of WGS data allows development of *in silico* prototype MLST schemes
- Increased resolution of WGS + portability of MLST



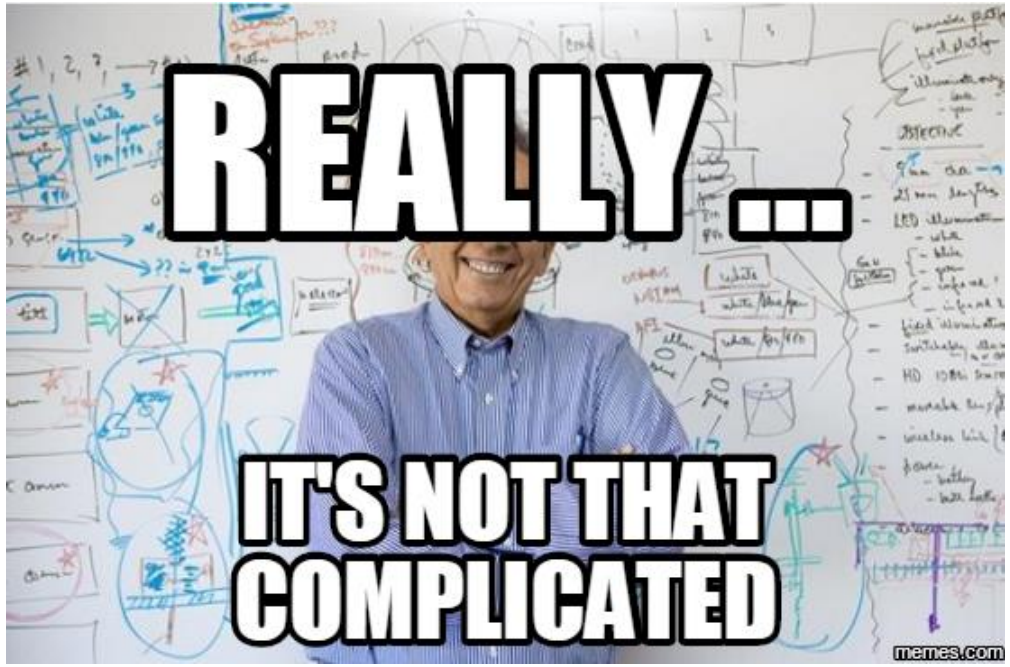
- Easier and cheaper to perform draft whole-genome sequencing and to extract the MLST data (with the possibility to expand on the number of loci by a couple of orders of magnitude)

How to scale-up MLST…

Scaling up MLST to the whole-genome level…

- ❑ Traditional MLST uses 7 to 9 loci
- ❑ A typical bacterial genome contains several thousand genes!!!
- ❑ More genes = more analytical power
- ❑ Can't we just build an MLST scheme with all of those genes?

→ i.e. whole-genome MLST (wgMLST)

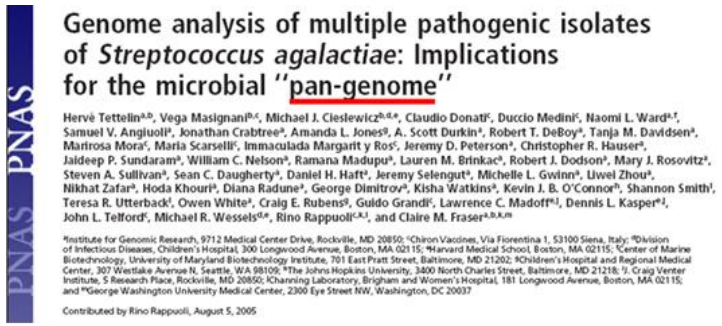


or is it???

The challenges of scaling-up MLST (Part 1)

- wgMLST will require fishing out the loci out of the WGS data, this is not trivial:
  - → Not all genes are present in all strains (Core genes vs Accessory genes)

"Core genes" vs. "Accessory genes"


PNAS

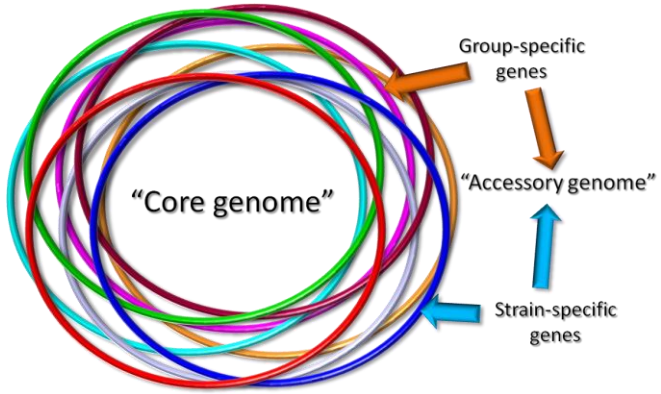
Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome"

Hervé Tettelin[a,b], Vega Masignani[b,c], Michael J. Cieslewicz[b,d,e], Claudio Donati[c], Duccio Medini[c], Naomi L. Ward[a,f], Samuel V. Angiuoli[a], Jonathan Crabtree[a], Amanda L. Jones[g], A. Scott Durkin[a], Robert T. DeBoy[a], Tanja M. Davidsen[a], Marirosa Mora[c], Maria Scarselli[c], Immaculada Margarit y Ros[c], Jeremy D. Peterson[a], Christopher R. Hauser[a], Jaideep P. Sundaram[a], William C. Nelson[a], Ramana Madupu[a], Lauren M. Brinkac[a], Robert J. Dodson[a], Mary J. Rosovitz[a], Steven A. Sullivan[a], Sean C. Daugherty[a], Daniel H. Haft[a], Jeremy Selengut[a], Michelle L. Gwinn[a], Liwei Zhou[a], Nikhat Zafar[a], Hoda Khouri[a], Diana Radune[a], George Dimitrov[a], Kisha Watkins[a], Kevin J. B. O'Connor[h], Shannon Smith[h], Teresa R. Utterback[i], Owen White[a], Craig E. Rubens[g], Guido Grandi[c], Lawrence C. Madoff[e,j], Dennis L. Kasper[e,j], John L. Telford[c], Michael R. Wessels[d,e], Rino Rappuoli[c,k,l], and Claire M. Fraser[a,b,k,m]

[a]Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850; [c]Chiron Vaccines, Via Fiorentina 1, 53100 Siena, Italy; [d]Division of Infectious Diseases, Children's Hospital, 300 Longwood Avenue, Boston, MA 02115; [e]Harvard Medical School, Boston, MA 02115; [f]Center of Marine Biotechnology, University of Maryland Biotechnology Institute, 701 East Pratt Street, Baltimore, MD 21202; [h]Children's Hospital and Regional Medical Center, 307 Westlake Avenue N, Seattle, WA 98109; [g]The Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218; [i]J. Craig Venter Institute, 5 Research Place, Rockville, MD 20850; [j]Channing Laboratory, Brigham and Women's Hospital, 181 Longwood Avenue, Boston, MA 02115; and [m]George Washington University Medical Center, 2300 Eye Street NW, Washington, DC 20037

Contributed by Rino Rappuoli, August 5, 2005

Tettelin et al. *PNAS USA* **102**: 13950.

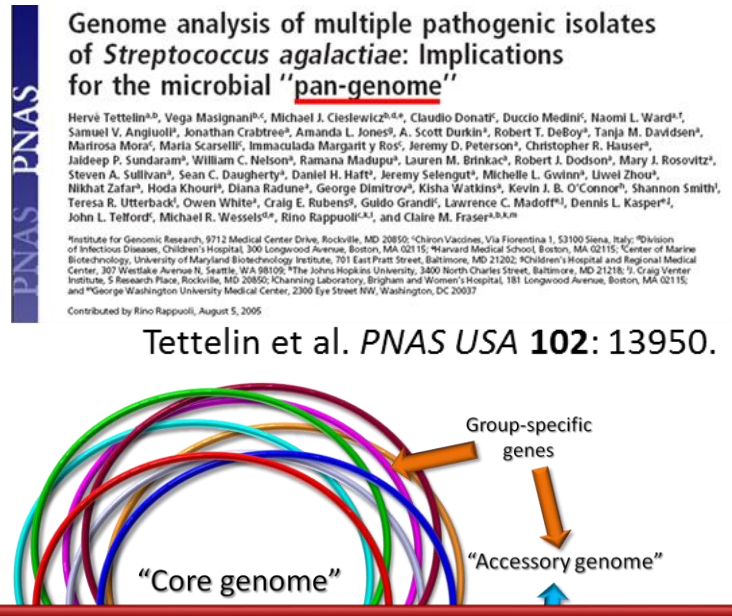


- **Core** genes
  - "Housekeeping" genes
  - Essential and found on every member of the species
- **Accessory** genes
  - AMR genes, "virulence" genes, carbon source utilization, etc…
  - generally clustered in "plasticity regions"
  - carriage is highly variable in the population
- **Pan-genome = Core + Accessory**

The challenges of scaling-up MLST (Part 1)

- wgMLST will require fishing out the loci out of the WGS data, this is not trivial:
  → Not all genes are present in all strains (Core genes vs Accessory genes)

"Core genes" vs. "Accessory genes"



- **Core** genes
  - "Housekeeping" genes
  - Essential and found on every member of the species
- **Accessory** genes
  - AMR genes, "virulence" genes, carbon source utilization, etc…
  - generally clustered in "plasticity regions"
  - carriage is highly variable in the population

- In the context of MLST, any accessory gene is problematic because we have no *a priori* knowledge on whether the gene should be present or absent from a given genome

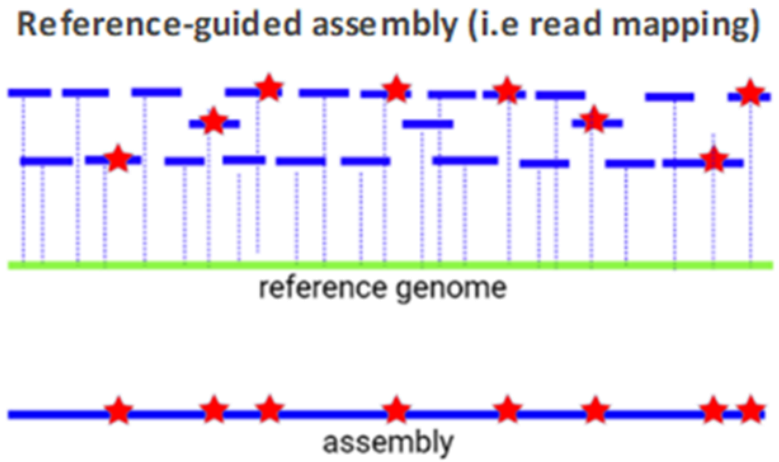
The challenges of scaling-up MLST (Part 2)

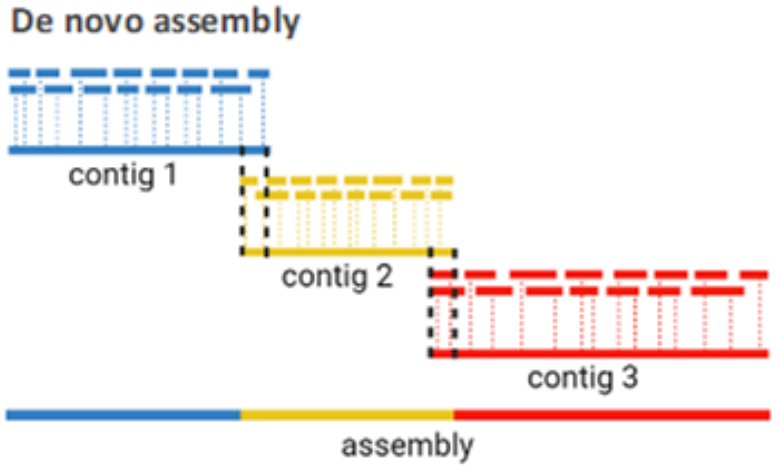
- wgMLST will require fishing out the loci out of the WGS data; this is not trivial:

  → variable quality and completeness of **genome assemblies**

Reference-guided assembly: assembly of reads into a draft genome by performing mapping of sequencing reads to a reference genome → reference genome not always available

De novo assembly: assembly of reads into a draft genome based on the sequence information of the sequencing reads using computationally efficient algorithms to look for overlapping reads and extending them into longer contiguous sequences (i.e. contigs).

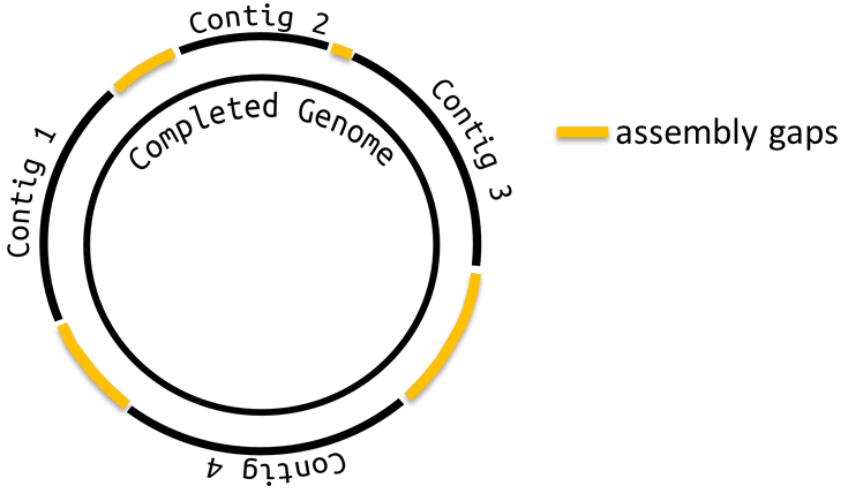




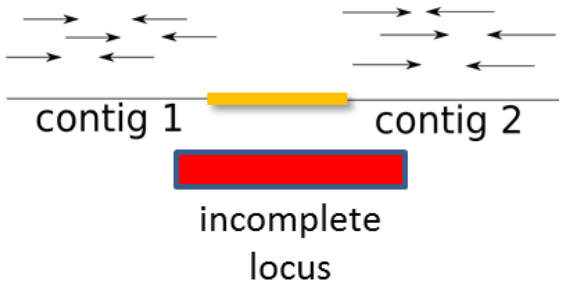
The challenges of scaling-up MLST (Part 2)

- wgMLST will require fishing out the loci out of the WGS data; this is not trivial:
  → variable quality and completeness of **genome assemblies**

  - WGS projects don't generally generate complete genomes
    → "Genome Assemblies" with gaps



  - unassigned alleles generally due to "collision" between a locus and a gap in the assembly
    → incomplete (i.e. truncated) allele



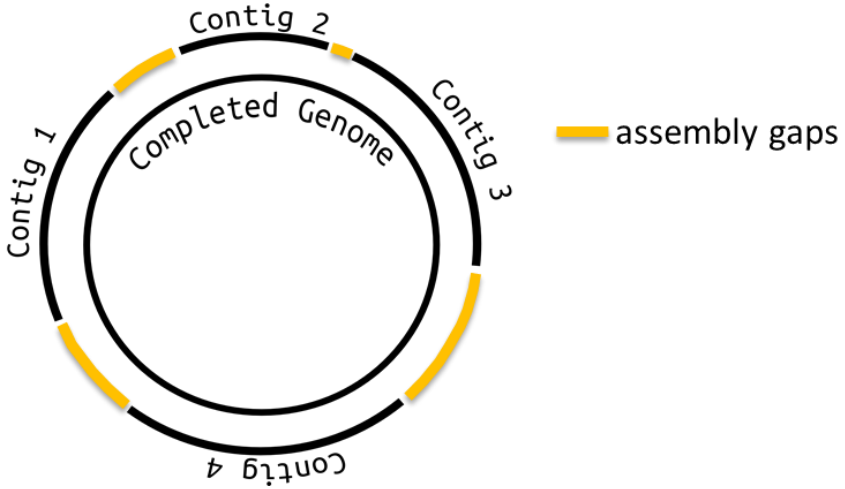
The challenges of scaling-up MLST (Part 2)

- wgMLST will require fishing out the loci out of the WGS data; this is not trivial:

  → variable quality and completeness of **genome assemblies**

  - WGS projects don't generally generate complete genomes

    → "Genome Assemblies" with gaps



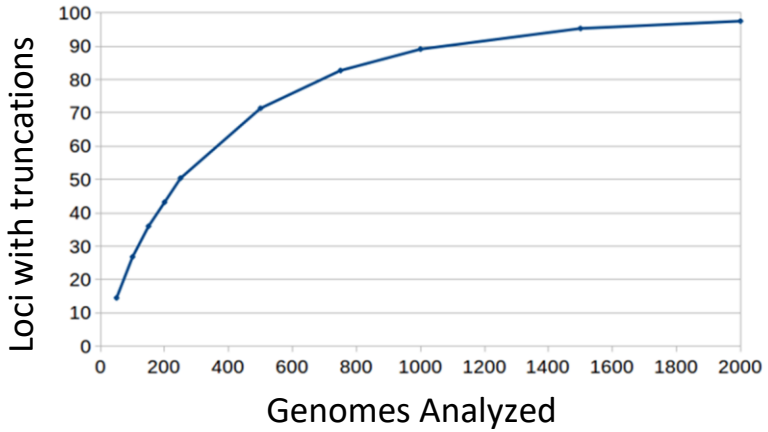
  - unassigned alleles generally due to "collision" between a locus and a gap

- Accessory genes are problematic: a gene could be absent because it is an accessory gene absent from the strain or because it is "missing" from the incomplete assembly

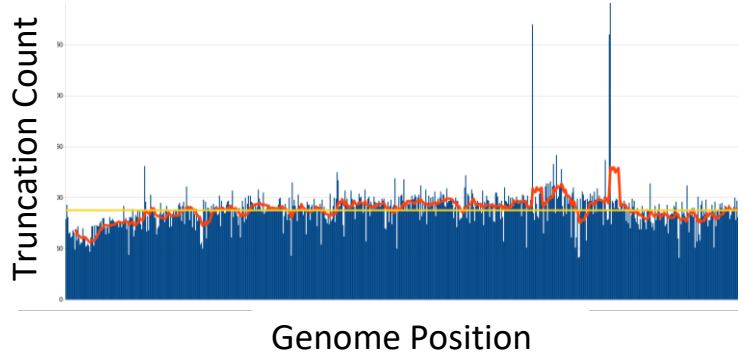
The challenges of scaling-up MLST (Part 2)

- wgMLST will require fishing out the loci out of the WGS data; this is not trivial:

  → variable quality and completeness of **genome assemblies**



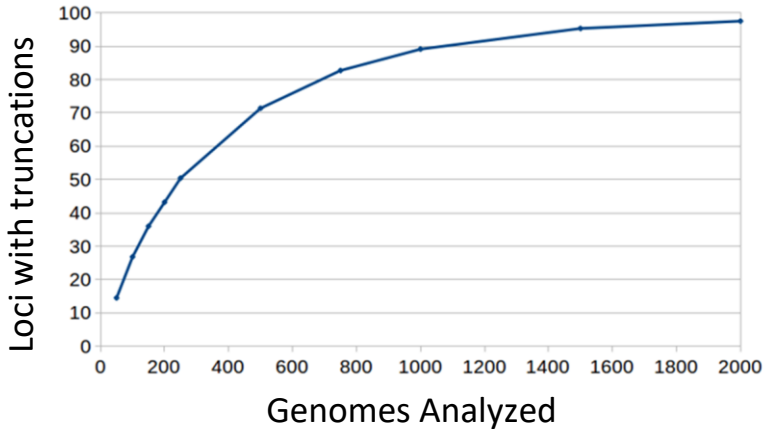
- On a large enough dataset, all genes will produce missing data for at least one genome
- A small proportion of genomes tends to be the worst offenders



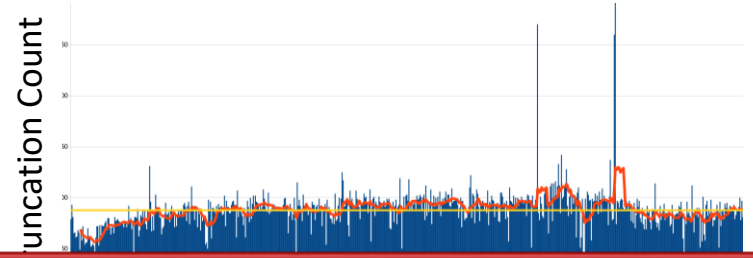
- Most loci have a background level of missing data due to assembly gaps
- Some loci are far more likely to reside in regions with assembly gaps

The challenges of scaling-up MLST (Part 2)

- wgMLST will require fishing out the loci out of the WGS data; this is not trivial:
  → variable quality and completeness of **genome assemblies**



- On a large enough dataset, all genes will produce missing data for at least one genome
- A small proportion of genomes tends to be the worst offenders

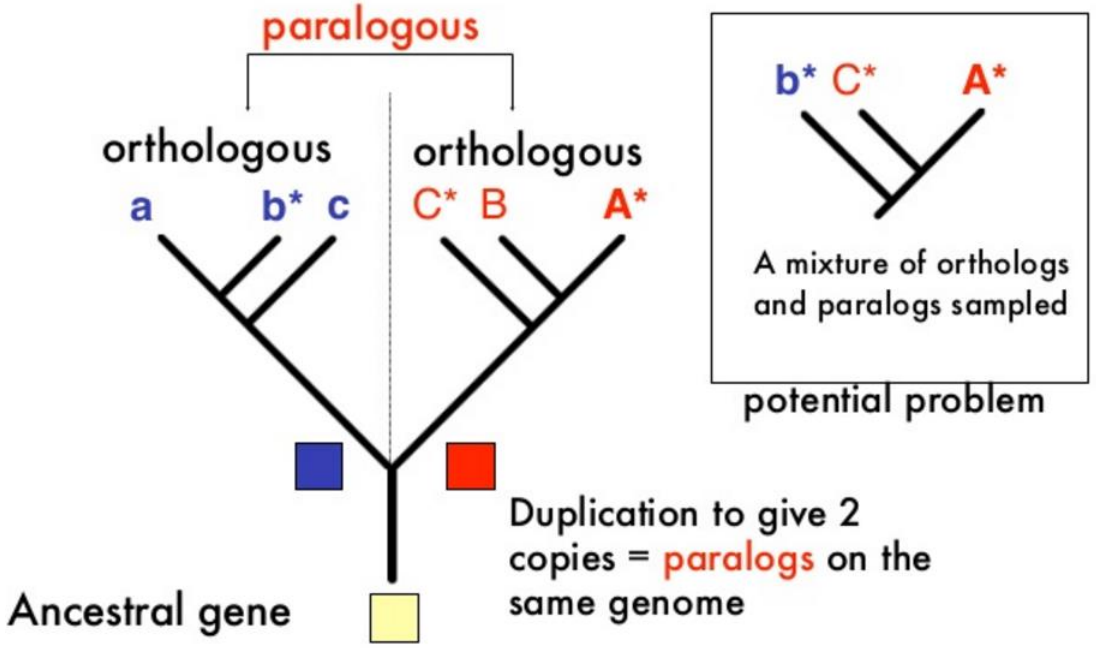


- Most loci have a background level of missing data due to assembly gaps
- Some loci are far more likely to reside in regions with assembly gaps

- Missing data is impossible to escape, especially as datasets get larger
- Must be willing to sacrifice certain genomes / loci for quality control

The challenges of scaling-up MLST (Part 3)

- wgMLST will require fishing out the loci out of the WGS data, this is not trivial:
  → some genes have multiple copies and number of copies can vary by strain



- **Orthologous** genes represent the "same" version of a multi-copy gene
- **Paralogous** genes are distinct versions and are likely evolving differently
- Identification of orthologs requires careful analysis

- Duplicated genes present a problem because it can be difficult to ascertain ortholog/paralog status; only comparison of orthologs is apples to apples

The challenges of scaling-up MLST (Part 4)

- wgMLST will require fishing out the loci out of the WGS data, this is not trivial:

  → some genes show significant variation in sequence and in length

	1	2	3	4
1	100%	90%	83%	?
2	90%	100%	91%	?
3	83%	91%	100%	?
4	?	?	?	100%

- Highly variable genes are problematic because identification of loci requires some form sequence homology searching (e.g. BLAST)
- Length variability makes it difficult to "map" where the gene begins and ends
- High sequence variability makes it difficult to define an appropriate sequence similarity threshold

- Length and sequence variability poses a problem for gene identification via homology searching as the allele database grows in size and strange alleles "pollute" it

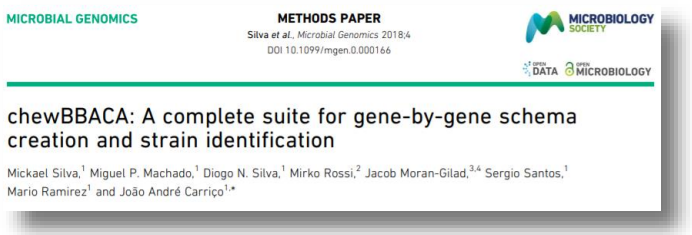
core genome Multi Locus Sequence Typing (cgMLST)

Core Genome MLST (cgMLST)

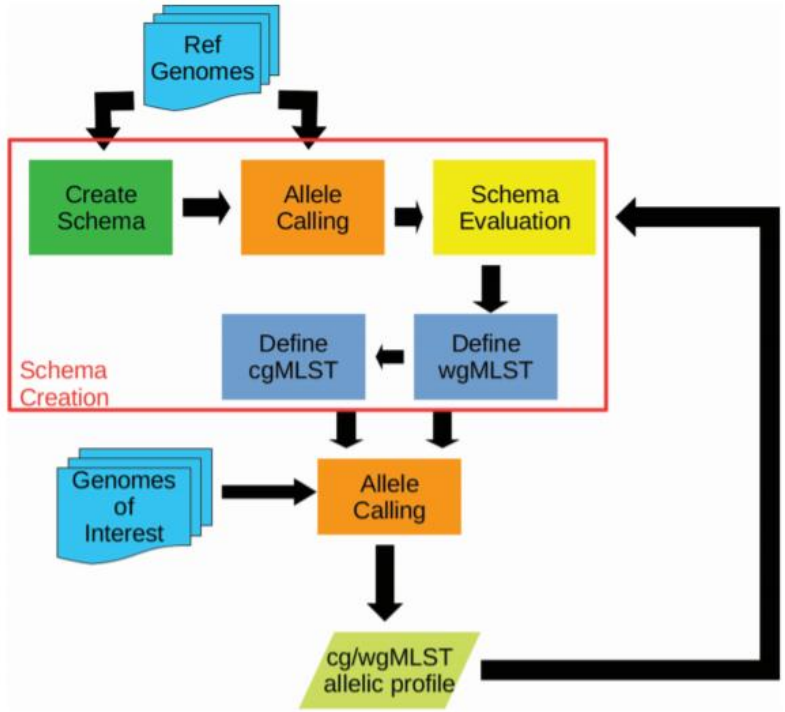
- Focusing on **core genes** solves several problems because manual curation of alleles and STs is not going to be possible (i.e. assignment of novel alleles and STs will have to be done automatically and without manual supervision)
- Because curation of a schema will be limited, significant effort should be placed up-front to ensure that it will scale up properly.
- Core Genome MLST (cgMLST) has been proposed as a possible approach for generating "well-behaved" schemas:
  - Core genes are shared by all members of the species → they should be present; you don't have to wonder whether it's a missing accessory gene or a missing gene in an incomplete assembly
  - Core genes display mostly SNV-level genetic variation → we should not have much trouble identifying the locus in an assembly of decent quality

- cgMLST provides a robust foundation for standardizing the transformation of WGS data into subtyping data suitable for outbreak response and longitudinal surveillance.

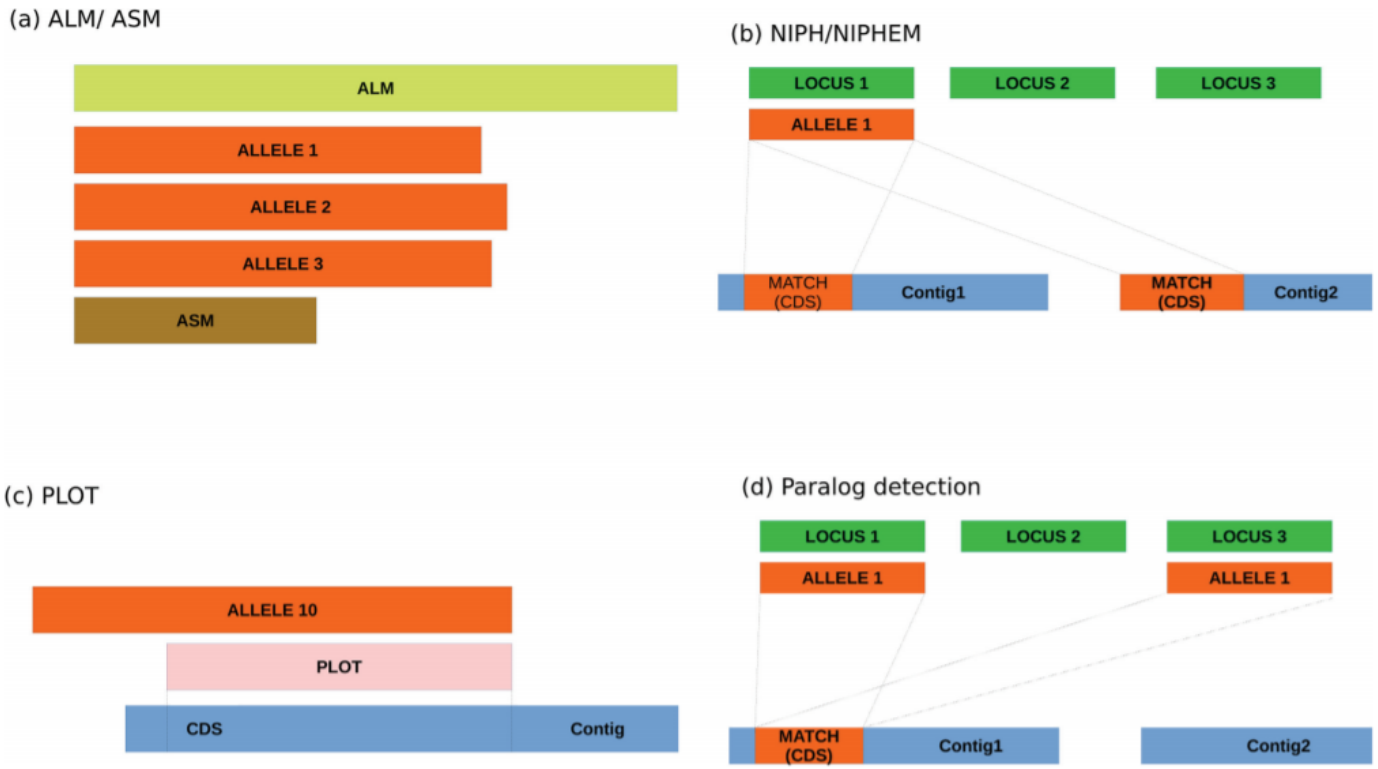
Designing a cgMLST schema the chewBBaca way



Silva et al. *Microbial Genomics* **2018**: 4.

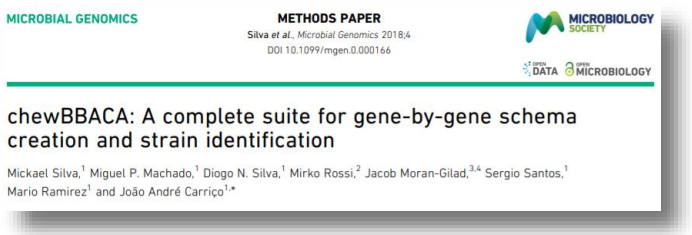


chewBBaca workflow: from schema creation to schema evaluation to allele calling

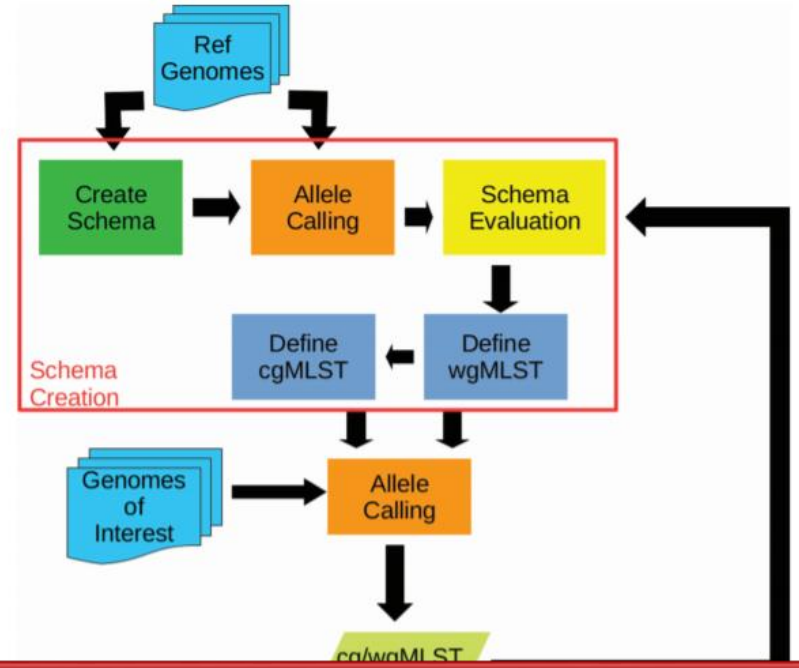


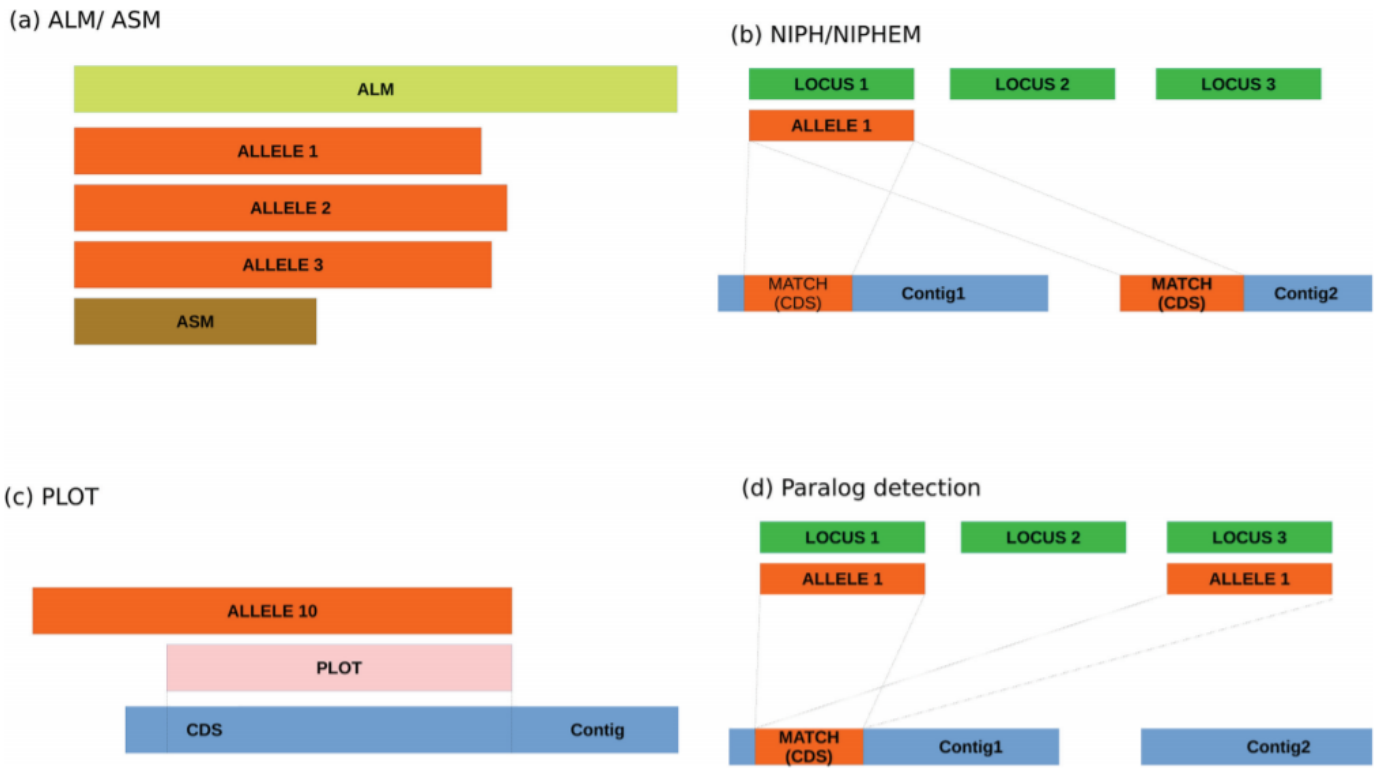
Exclusion of loci by size, duplicated genes, truncated genes and paralogous genes

Designing a cgMLST schema the chewBBaca way



Silva et al. *Microbial Genomics* **2018**: 4.





Exclusion of loci by size, duplicated genes, truncated genes

- Tools like chewBBACA can help standardize the development of cgMLST schema by applying basic principles to identify loci with robust performance.

Designing a cgMLST schema (part 2)

Sanity checks & Common Pitfalls:

- ❑ Core genome size → Do your numbers make sense based on the literature?
  - If core genome is too big
    - Likely included accessory genes in schema
    - Will generate lots of missing data
  - If core genome is too small
    - Likely missing true core genes from the schema
    - Fewer loci will reduce discriminatory power
  - Core genome definition should
    - Incorporate as much genetic diversity for the species as possible
    - Avoid poor quality genomes
    - Avoid genomes outside the species of interest

      → Species in repositories are often mislabelled: **Trust no one!**
  - Inspection of patterns of missing data
    - Some genomes may be problematic → Disproportionate number of truncated/absent putative core genes
    - Some genes may be problematic: present, but incomplete in many genomes

Designing a cgMLST schema (part 2)

Sanity checks & Common Pitfalls:

- ❑ Core genome size → Do your numbers make sense based on the literature?
  - If core genome is too big
    - Likely included accessory genes in schema
    - Will generate lots of missing data
  - If core genome is too small
    - Likely missing true core genes from the schema
    - Fewer loci will reduce discriminatory power
  - Core genome definition should
    - Incorporate as much genetic diversity for the species as possible
    - Avoid poor quality genomes
    - Avoid genomes outside the species of interest

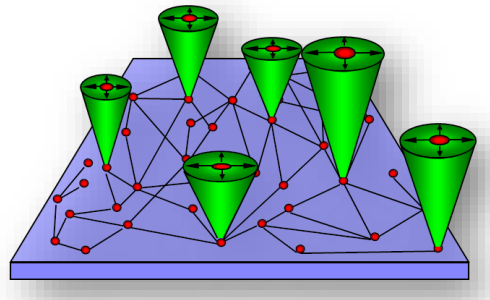
→ Species in repositories are often mislabelled: **Trust no one!**

- ❑ You have to be ruthless: a gene found in 95% of genomes is **not** a core gene
- ❑ Drop some genomes, drop some genes or both because fewer high quality genomes & loci are better than a larger dataset with lots of unassigned alleles

cgMLST: clustering

Clustering of cgMLST data

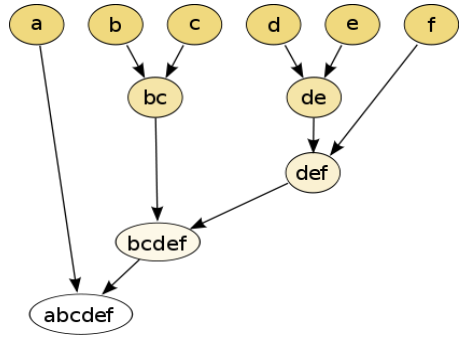
- A lot of the focus of WGS-based typing is on identifying potential outbreaks and circulating high-risk sublineages of the pathogen in question (i.e. *clone identification* vs. *phylogenetic relationships between clones*).



- Allelic profiles are compared and pairwise similarities are computed using the ***Hamming Distance*** *(i.e. the number of positions at which corresponding alleles are different, ignoring any that are "missing").*

$$d(x,y) = \frac{1}{n}\sum_{n=1}^{n=n} |x_i - y_i|$$

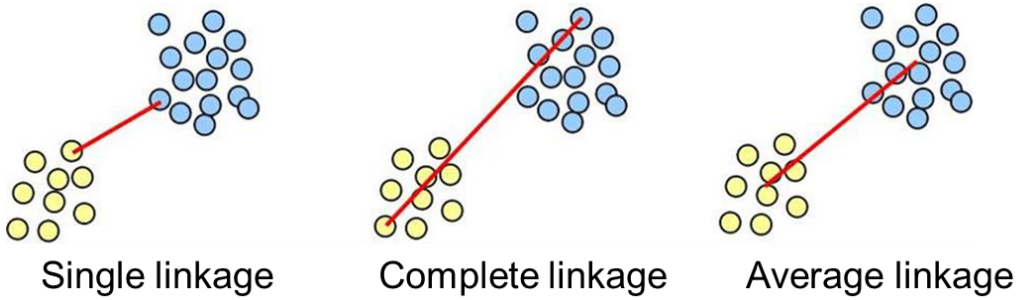
- Hierarchical clustering of the matrix of pairwise Hamming distances



Clustering of cgMLST data

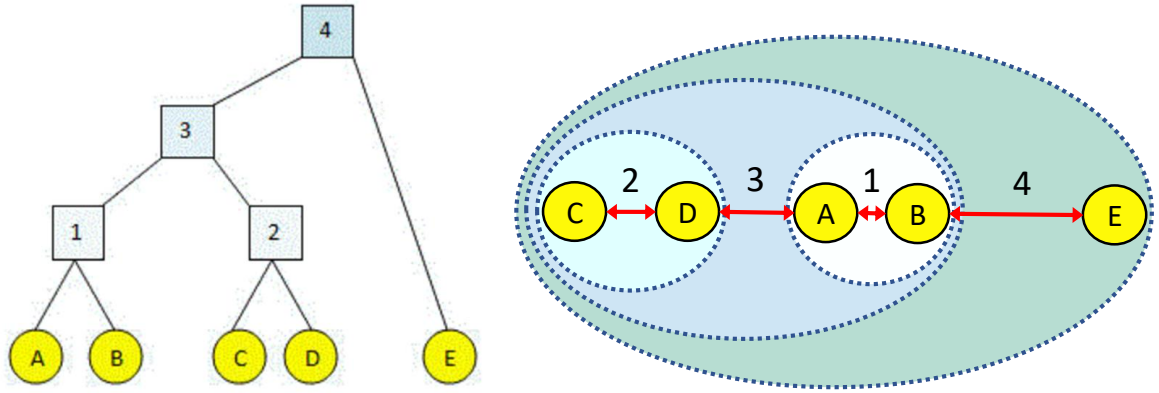
Single Linkage Clustering: (1) Each genome is its own cluster, and (2) Clusters are sequentially combined into larger clusters in step-wise fashion (i.e. agglomerative clustering), until all genomes are in the same cluster. At each step, the **two clusters separated by the shortest distance** are combined

Linkage functions



https://tinyurl.com/5h5vhcxb

Single Linkage Clustering

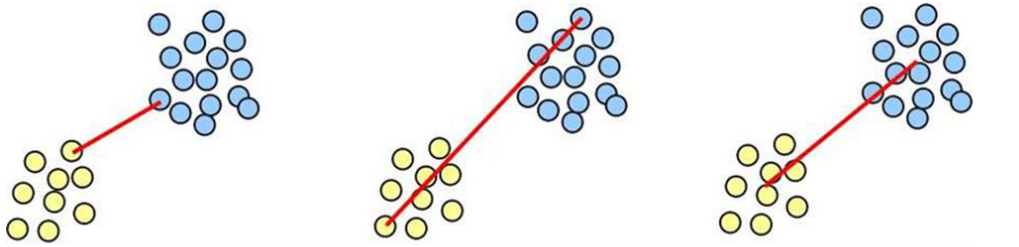


Adapted from https://drive5.com/usearch/manual/agg.html

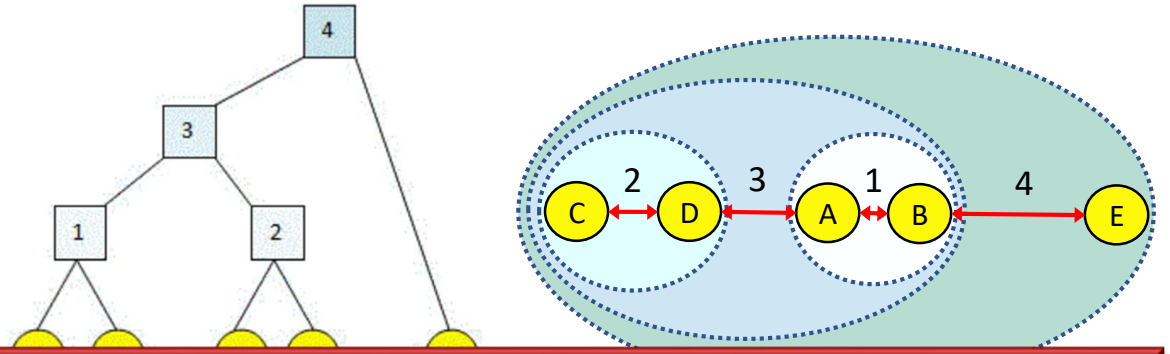
Clustering of cgMLST data

Single Linkage Clustering: (1) Each genome is its own cluster, and (2) Clusters are sequentially combined into larger clusters in step-wise fashion (i.e. agglomerative clustering), until all genomes are in the same cluster. At each step, the **two clusters separated by the shortest distance** are combined

Linkage functions



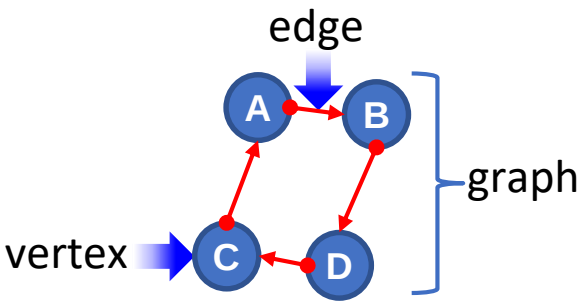
Single Linkage Clustering



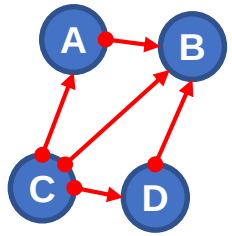
- The resulting dendrogram isn't phylogenetically "correct" because the molecular data isn't analyzed using an evolutionary model but robust clusters and broad inter-lineage relationships can be identified in the data

Clustering of cgMLST data

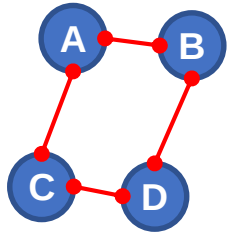
❑ Spanning Trees



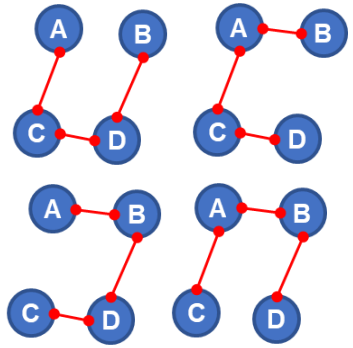
Directed Cyclic graph



Directed Acyclic graph



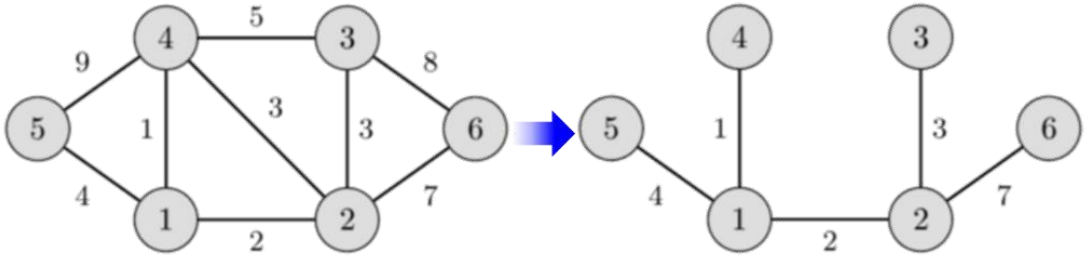
Undirected Cyclic graph



"Spanning" trees

# of edges = (# vertices – 1)

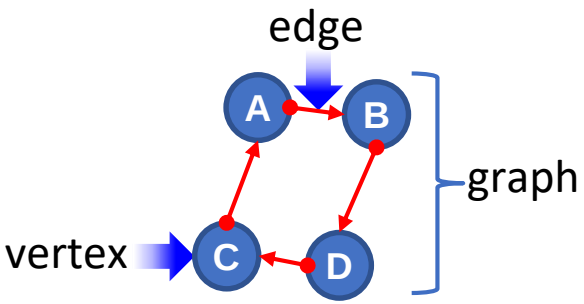
❑ Kruskal's Minimum Spanning Tree (MST)



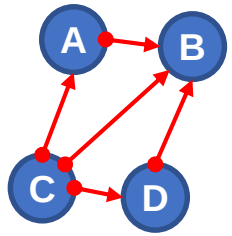
1. Sort all edges from the lowest weight to the highest.
2. Take edge with the lowest weight and add it to the spanning tree.
3. Take edge with next lowest weight and add it to the tree; if a cycle is created, discard the edge.
4. Keep adding edges like in step 1 until all the vertices are considered.

Clustering of cgMLST data

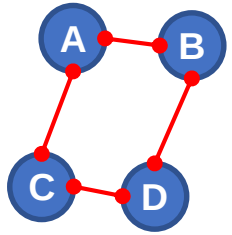
- Spanning Trees



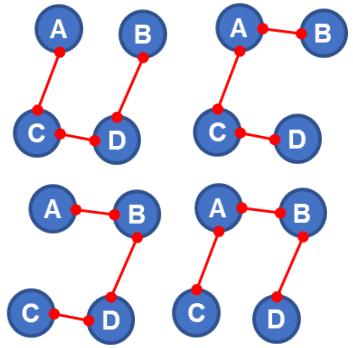
Directed Cyclic graph



Directed Acyclic graph



Undirected Cyclic graph



"Spanning" trees

# of edges = (# vertices – 1)

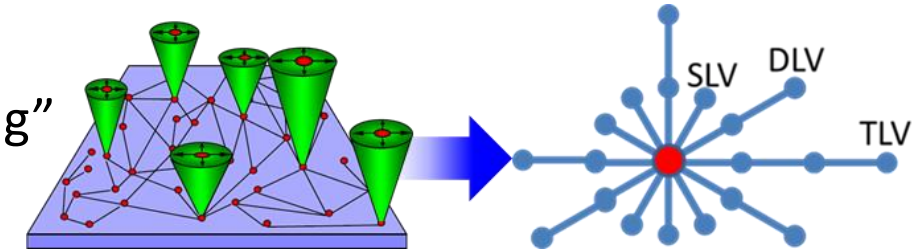
- Kruskal's Minimum Spanning Tree (MST)

1. Sort all edges from the lowest weight to the highest.
2. Take edge with the lowest weight and add it to the spanning tree.

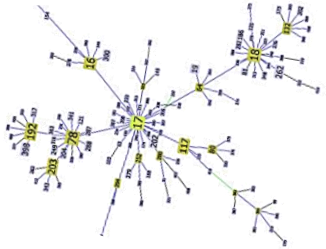
- MSTs are a popular way of clustering MLST data in which the tree is constructed so that total edge weight is minimized
- In an MST, all vertices in the graph are connected and the graph is acyclic

GrapeTree: a new approach to MST building

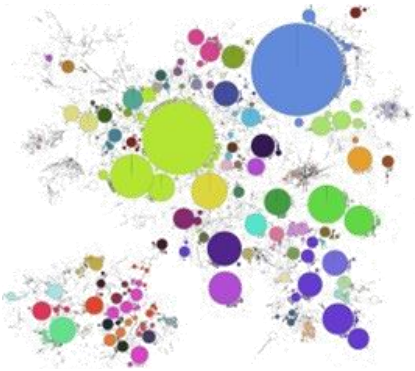
- Recall that **eBURST** algorithm used to cluster 7-locus MLST generates clusters (Clonal Complexes) comprising a "founding" Sequence Type and its "cloud" of single, double, and triple locus variants, a simple model for epidemic clone expansion



- Francisco et al. (2009) *BMC Bioinformatics* 10:152 propose a "globally optimized" eBURST algorithm (**goeBURST**) that generates an MST-like output

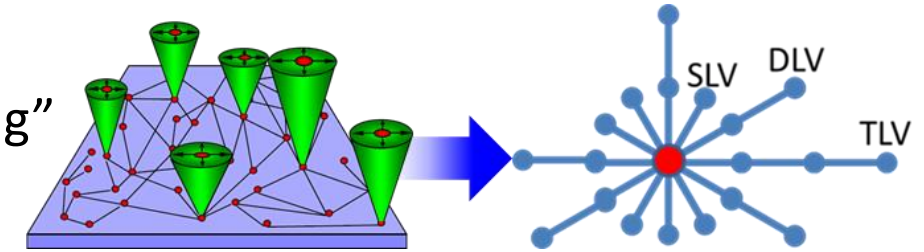


- Zhou et al. (2018) *Genome Res* 28(9):1395 propose **MSTree v2**, a novel MST algorithm better suited for handling missing data than classical MSTs and capable of generating visualizations with >100,000 nodes

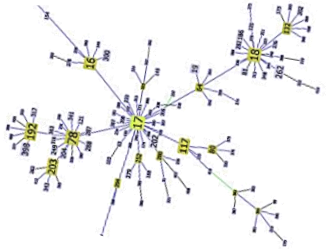


GrapeTree: a new approach to MST building

- Recall that **eBURST** algorithm used to cluster 7-locus MLST generates clusters (Clonal Complexes) comprising a "founding" Sequence Type and its "cloud" of single, double, and triple locus variants, a simple model for epidemic clone expansion



- Francisco et al. (2009) *BMC Bioinformatics* 10:152 proposed a "globally optimized" eBURST algorithm (**goeBURST**) that generates an MST-like output



- MST V2 is preferable when clustering allelic profiles that may contain varying levels of missing data, which can create "phantom" STs that differ from nearly identical STs due to the missing alleles and represent unique nodes in the MST
- MST V2 also uses an approach for identifying Clonal Complex founders adapted from goeBURST that is more appropriate when scaling up MLST to hundreds/thousands of loci

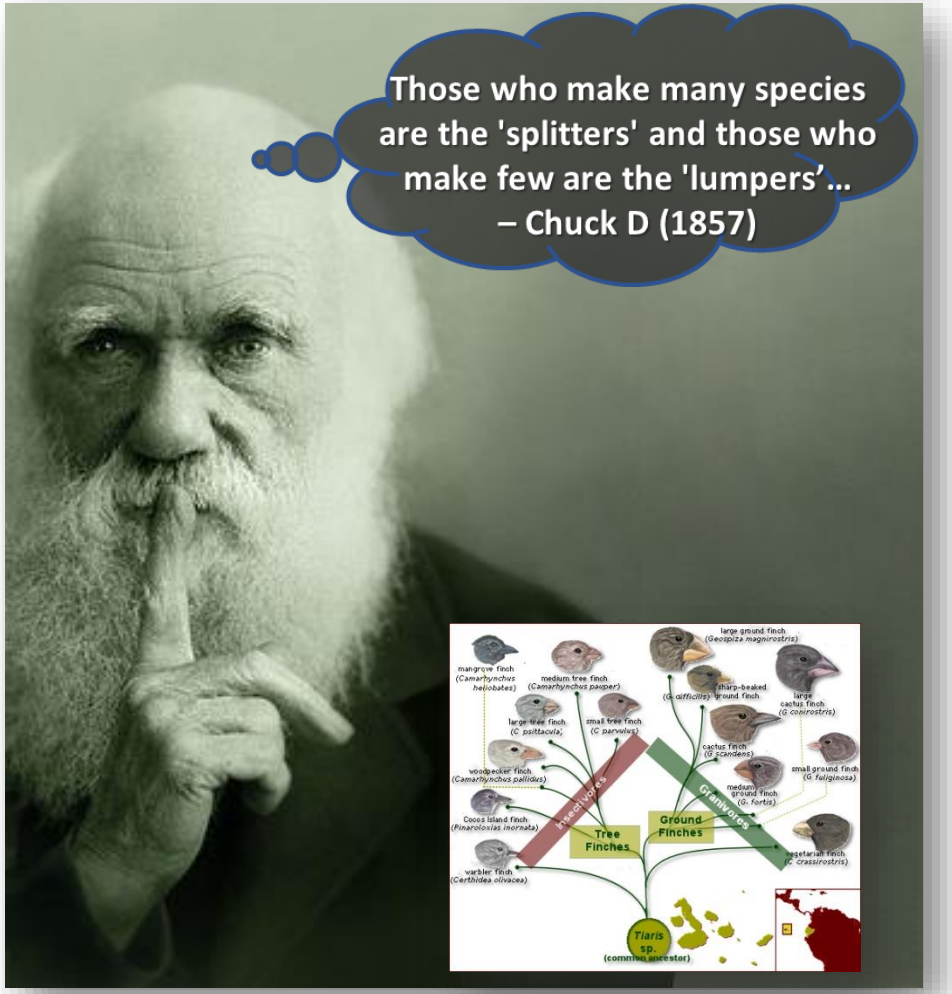
cgMLST: epidemiological interpretation

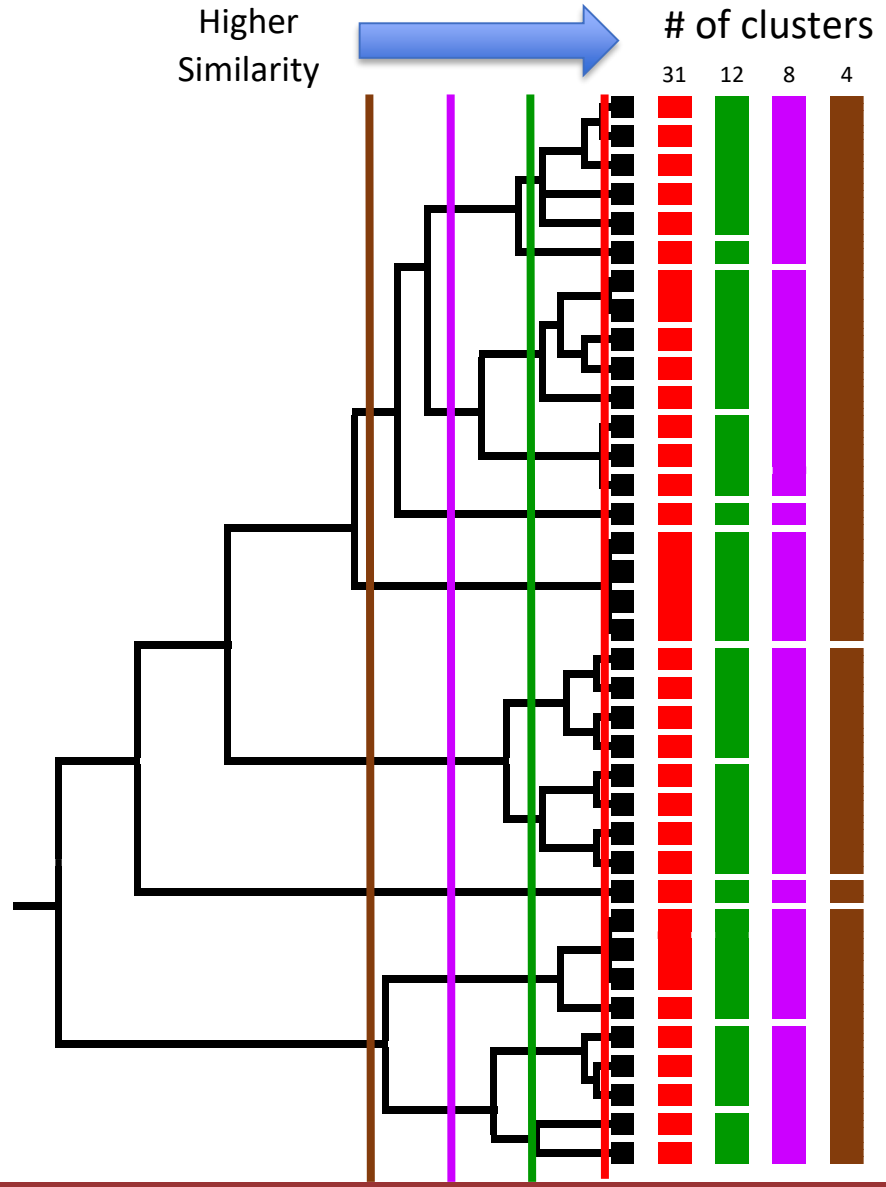
Generating “genomic clusters” from WGS data

- Before we can examine trends among genomes in a dataset, we must define clusters of genomes that are highly related and that can be defined as analytical units (i.e. genomic subtypes)
- This is generally achieved through the application of **distance thresholds**

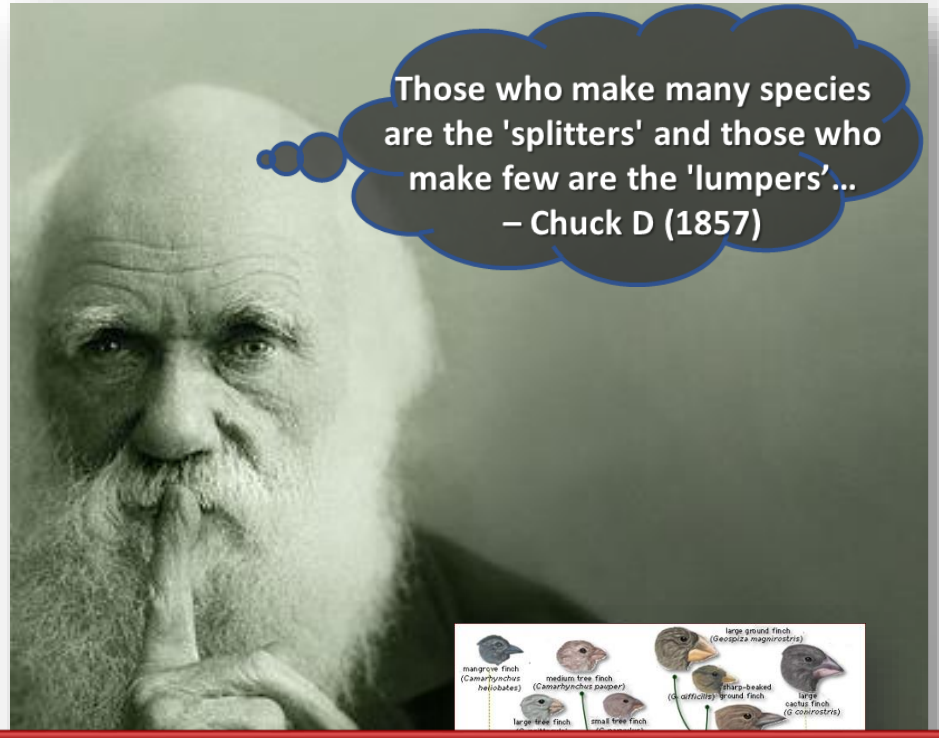
A short aside on distance thresholds…

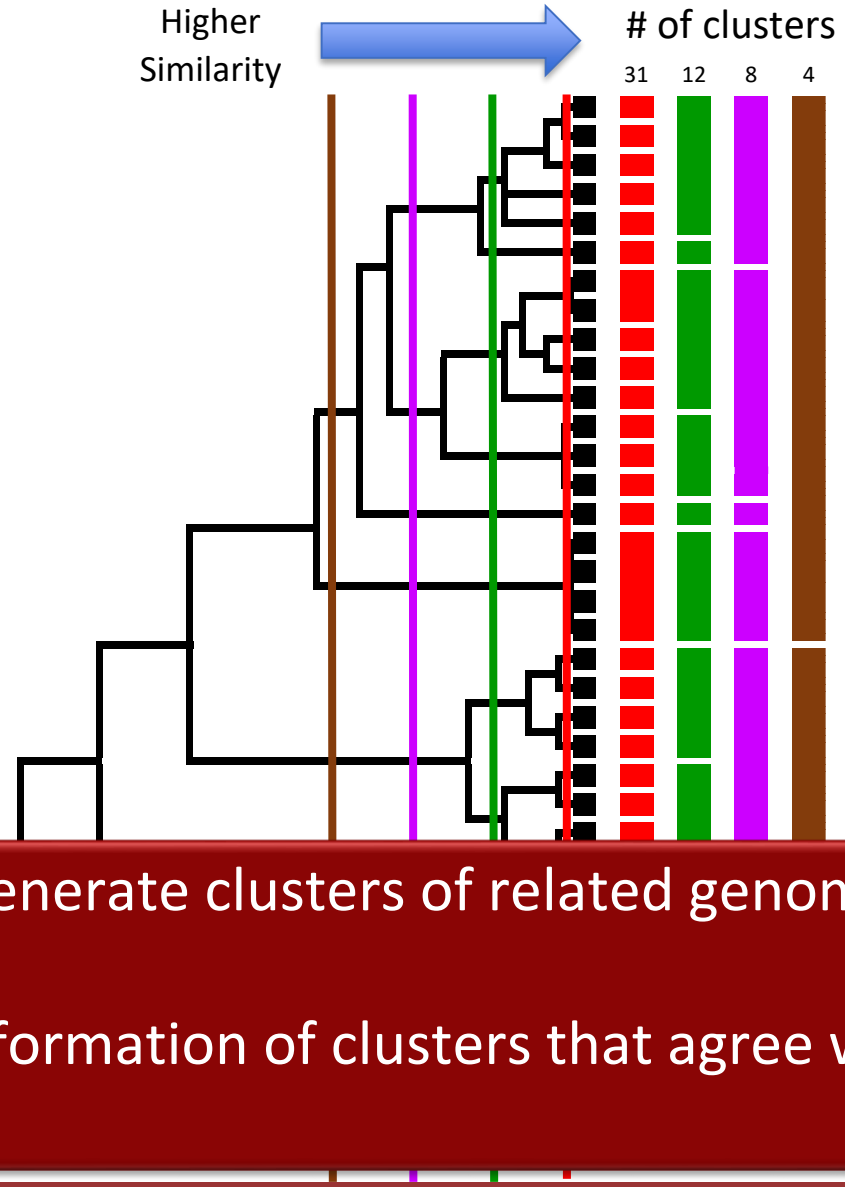
Lumping vs. splitting in the genomic era





Lumping vs. splitting in the genomic era



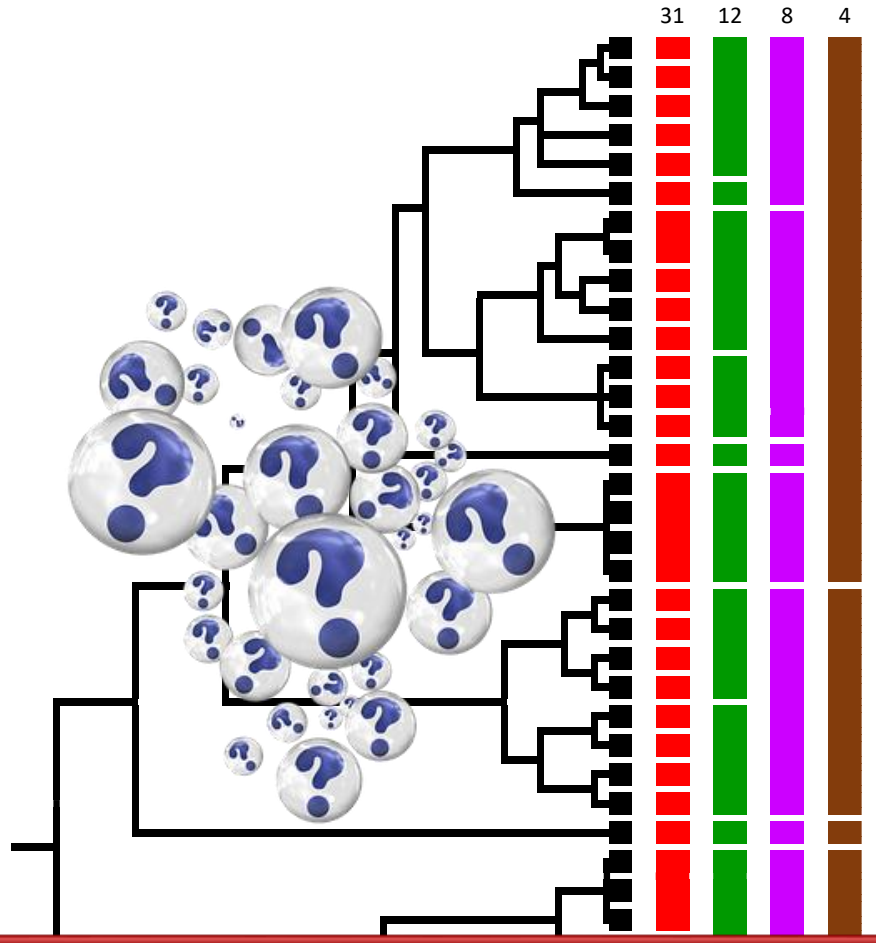


- Application of distance threshold(s) can be used to generate clusters of related genomes at varying levels of similarity
- Thresholds have generally been set to maximize the formation of clusters that agree with retrospective outbreak data

Optimization of “genomic clusters” from WGS data

- Cluster definition for bacterial populations is analogous to defining country borders:
  - Demography → Population structure
  - Topography → Ecology/Epidemiology

- Ecology/Epidemiology influences the population structure:
  - Strains that share provenance or environmental niche will look more like one other though common descent or adaptation



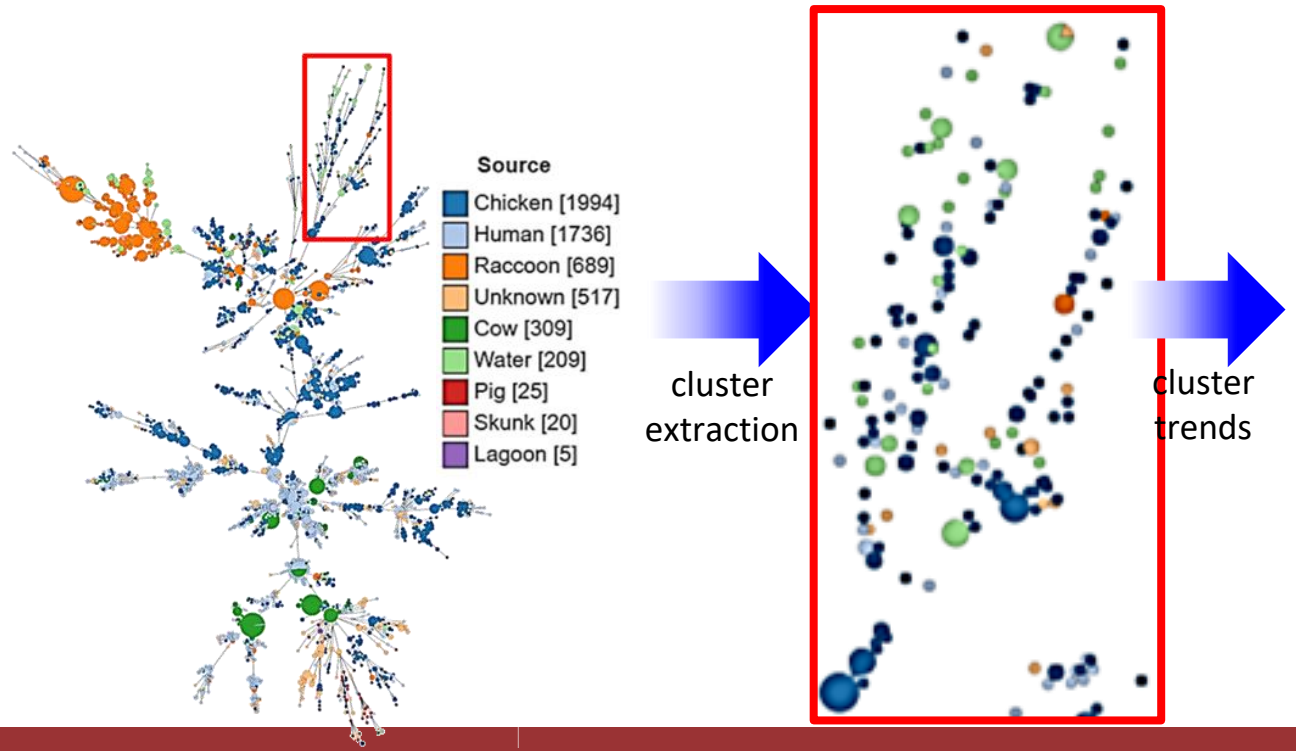
- Ecology, epidemiology and population genetics are correlated; this information can inform cluster definitions and interpretation criteria

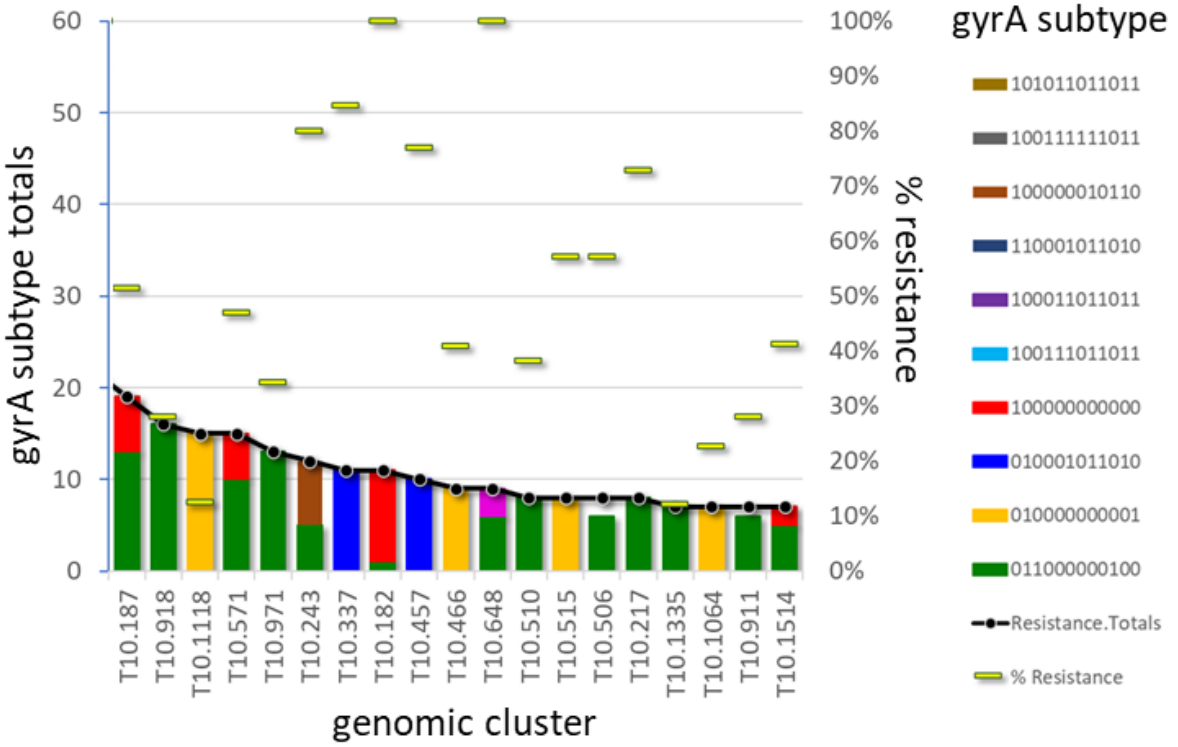
Generating “genomic clusters” from WGS data

- Before we can examine trends among genomes in a dataset, we must define clusters of genomes that are highly related and that can be defined as analytical units (i.e. genomic subtypes)
- This is generally achieved through the application of **distance thresholds**

Generating “genomic clusters” from WGS data

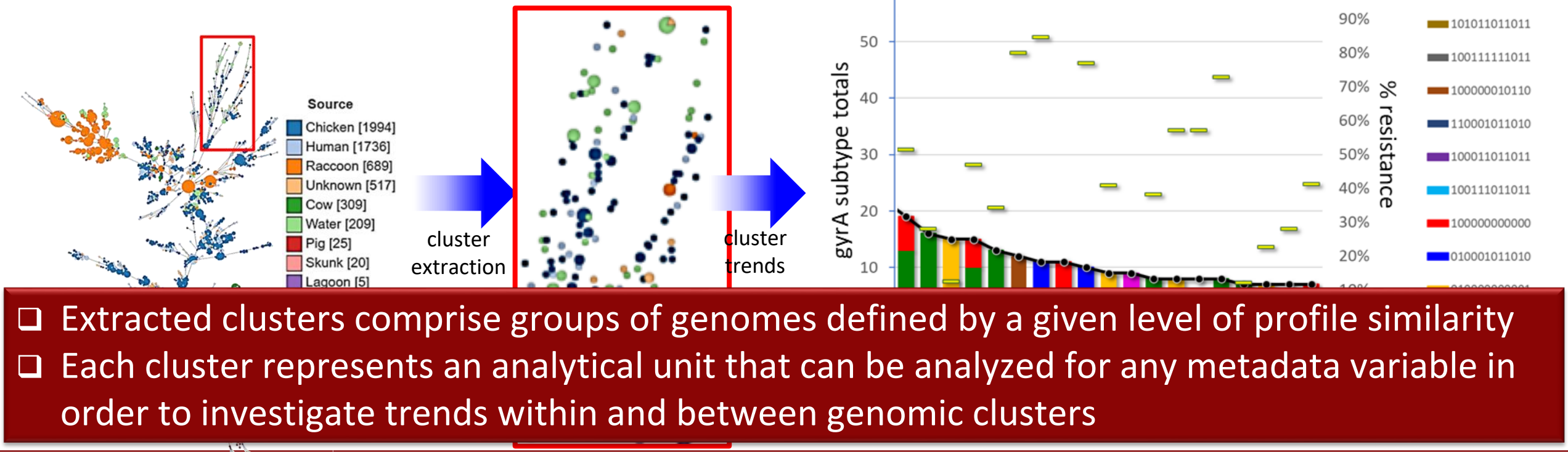
- **Cluster extraction** is the process of applying a distance threshold so that groups comprising sets of similar genomes sharing below-threshold intracluster distances can be defined as analytical units
- Cluster trends can be computed for any associated metadata (e.g. country of origin, sample source, temporal distribution, AMR, …)





Generating “genomic clusters” from WGS data

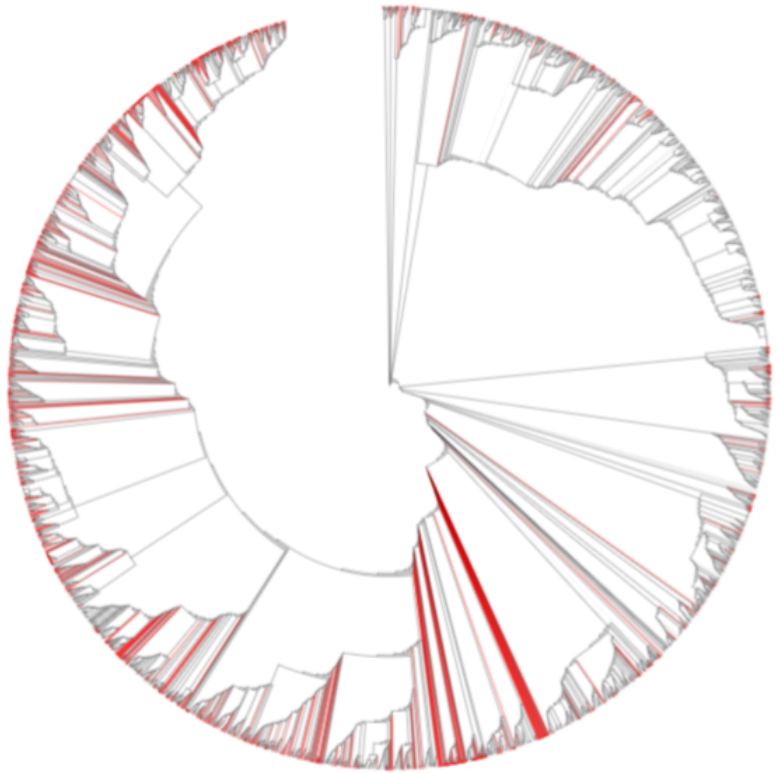
- **Cluster extraction** is the process of applying a distance threshold so that groups comprising sets of similar genomes sharing below-threshold intracluster distances can be defined as analytical units
- Cluster trends can be computed for any associated metadata (e.g. country of origin, sample source, temporal distribution, AMR, …)



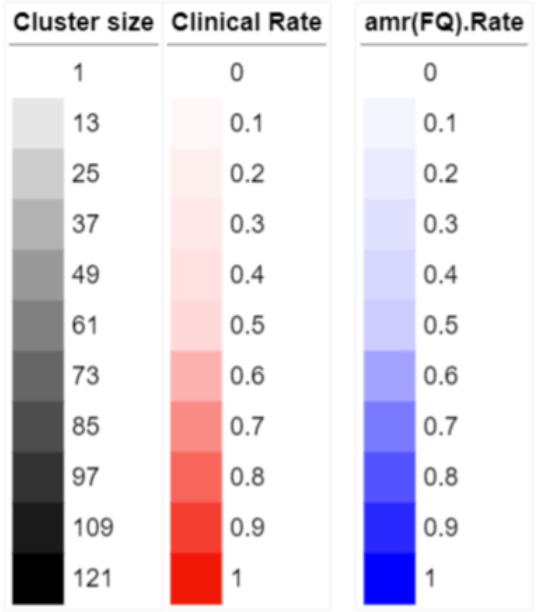
- Extracted clusters comprise groups of genomes defined by a given level of profile similarity
- Each cluster represents an analytical unit that can be analyzed for any metadata variable in order to investigate trends within and between genomic clusters

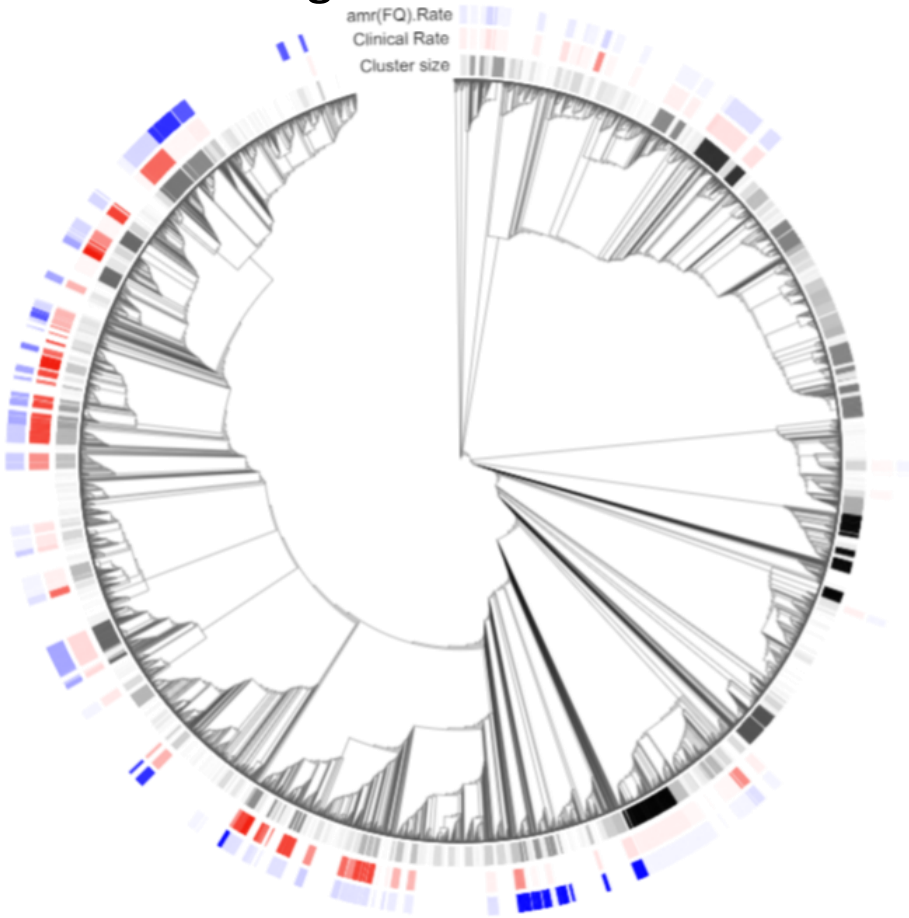
Visualizing genomic cluster trends

- Computed cluster trends can be visualized for any associated metadata (e.g. country of origin, sample source, temporal distribution, AMR, …) in the context of a dendrogram

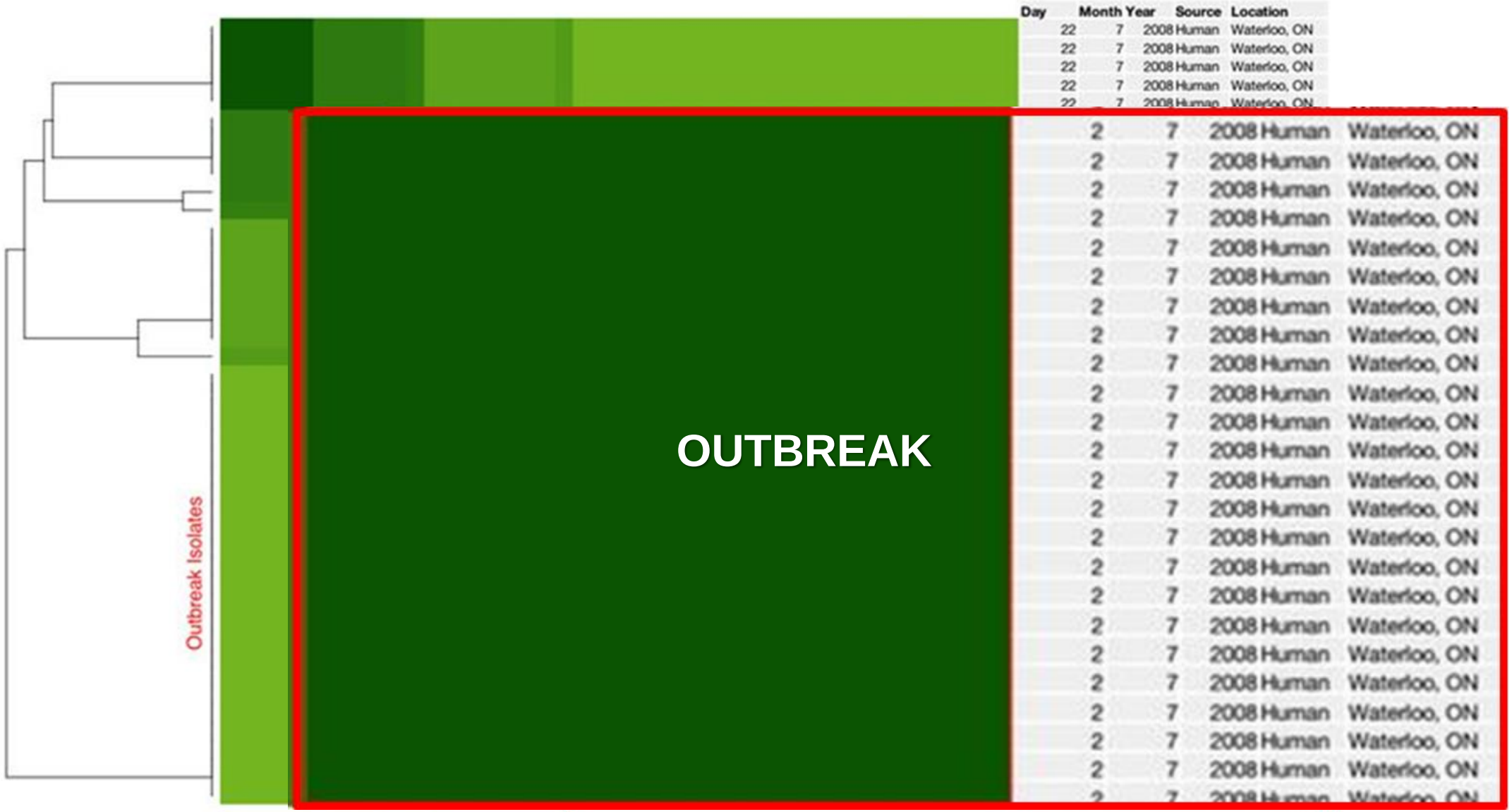


trend visualization



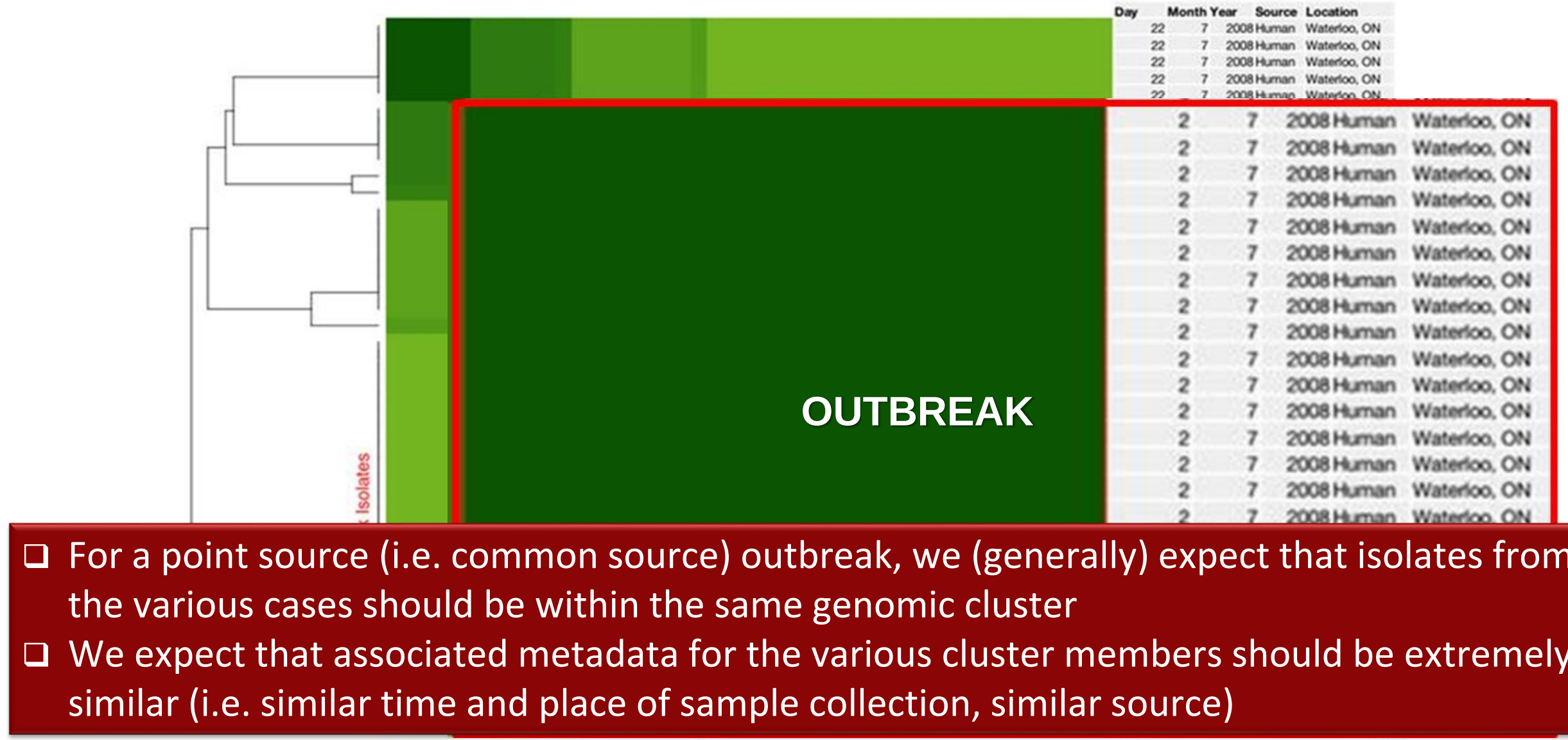


(manual) Inspection of epidemiological contextual data



Day	Month	Year	Source	Location
22	7	2008	Human	Waterloo, ON
22	7	2008	Human	Waterloo, ON
22	7	2008	Human	Waterloo, ON
22	7	2008	Human	Waterloo, ON
22	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON
2	7	2008	Human	Waterloo, ON

(manual) Inspection of epidemiological contextual data



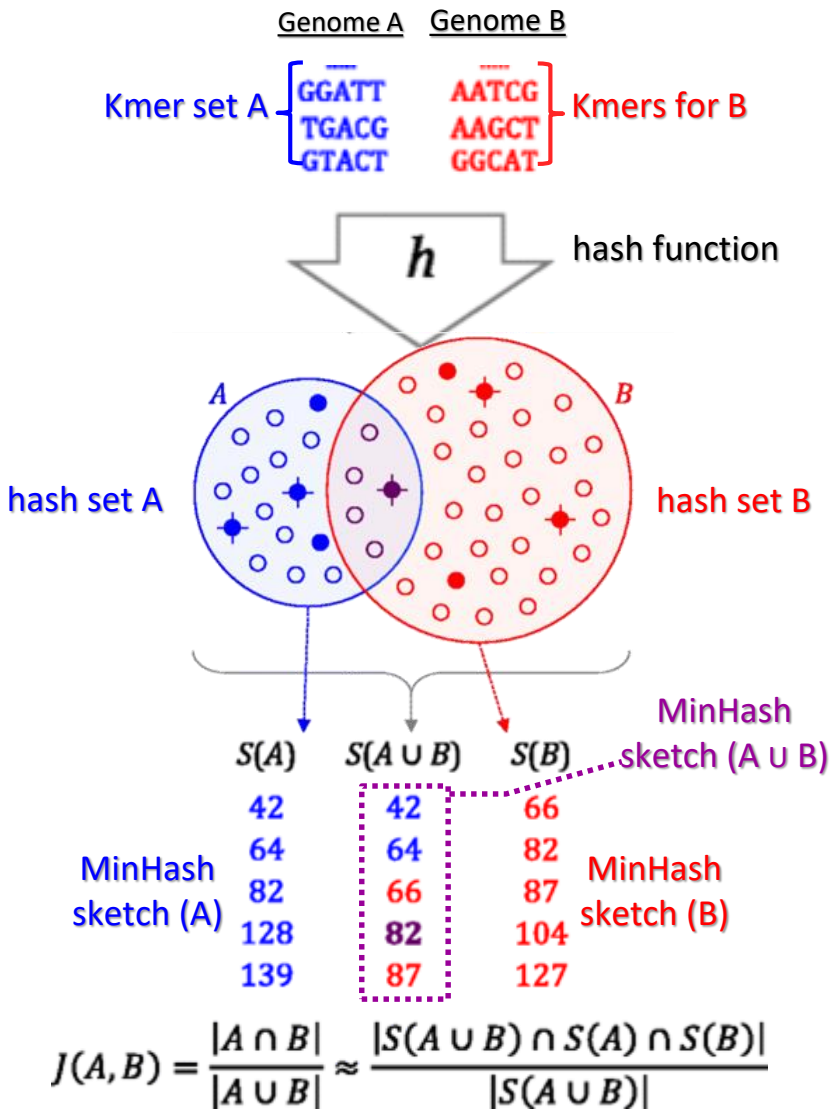
- For a point source (i.e. common source) outbreak, we (generally) expect that isolates from the various cases should be within the same genomic cluster
- We expect that associated metadata for the various cluster members should be extremely similar (i.e. similar time and place of sample collection, similar source)

WGS-based subtyping: novel approaches…

Mash: fast genome and metagenome distance estimation using MinHash

- **MinHash:** dimensionality-reduction technique that reduces sets of data to "sketches" that can be efficiently compared in order to generate similarity estimates; originally developed to compare webpages in order to detect (near-) duplicates in web searching

- **Mash**: described by Ondov et al. (2016), extends MinHash to use k-mers and sampling to estimate pairwise mutation distances between sequence sets, enabling the efficient comparison and clustering of massive sequence collections in an alignment-free manner.

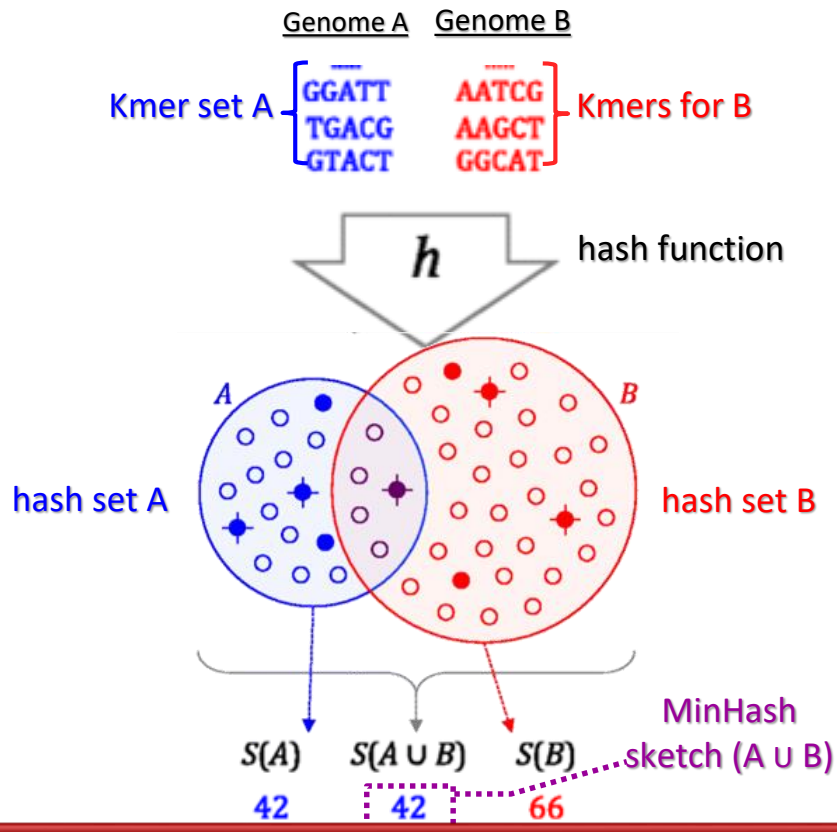


Adapted from Ondov et al. (2016) *Genome Biol* **17**, 132.

Mash: fast genome and metagenome distance estimation using MinHash

- **MinHash:** dimensionality-reduction technique that reduces sets of data to "sketches" that can be efficiently compared in order to generate similarity estimates; originally developed to compare webpages in order to detect (near-) duplicates in web searching

- **Mash**: described by Ondov et al. (2016), extends MinHash to use k-mers and sampling to estimate pairwise mutation distances between sequence



- Because Mash analysis is performed at the k-mer level, inputs can be whole genomes, metagenomes, nucleotide sequences, raw sequencing reads, etc.
- Mash has been used to perform rapid and alignment-free comparison and clustering of the entire NCBI RefSeq genome database

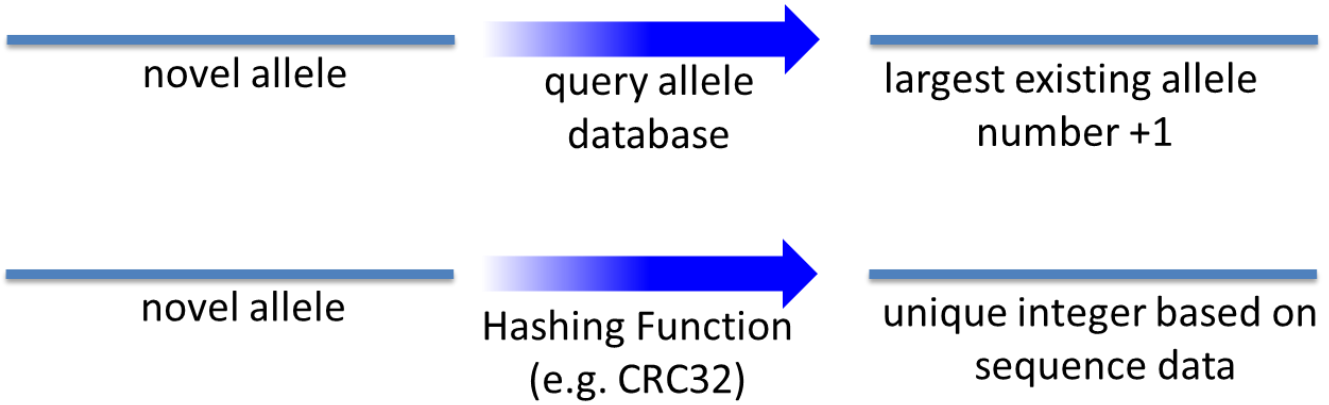
Additional considerations…

Complementing cgMLST…

- MLST analysis indexes variation at the level of a locus; multiple SNPs are treated the same as a single SNP, reducing discriminatory power.
- Because it uses fewer loci, cgMLST represents a compromise between robust performance and a loss of discriminatory power.
- **Problem:** For species, lineages, or sub-lineages with limited diversity, cgMLST may not provide sufficient discriminatory power to differentiate genomes that are distinct.
- **Solution**: two primary options have been proposed to deal with such cases:
    1. Perform SNV-based analysis on a smaller group of strains known to be highly similar based on cgMLST
    2. Perform MLST with an expanded set of loci by including accessory genes shared by the smaller group of strains → Shared Genome MLST (sgMLST)

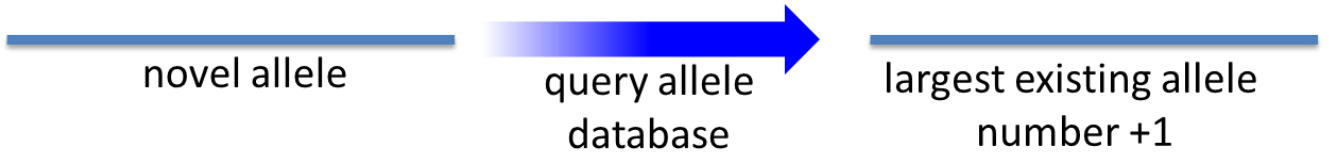
Coordinating a global centralized cgMLST database

- In a global community that is continually generating data to solve "local" investigations, how do we ensure that the allele and allelic profile definitions are completely up to date?
- **Problem:** potential "collisions" if allele calling software running locally needs to assign novel alleles or allelic profiles that need to be synchronized internationally
- **Solution:** a proposed workaround is "allele hashing" → hash-cgMLST
  - Eyre et al. 2019 *J Clin Microbiol* **58**(1): e01037-19
  - Deneke et al. 2021 *Front Microbiol* **12**: 649517



Coordinating a global centralized cgMLST database

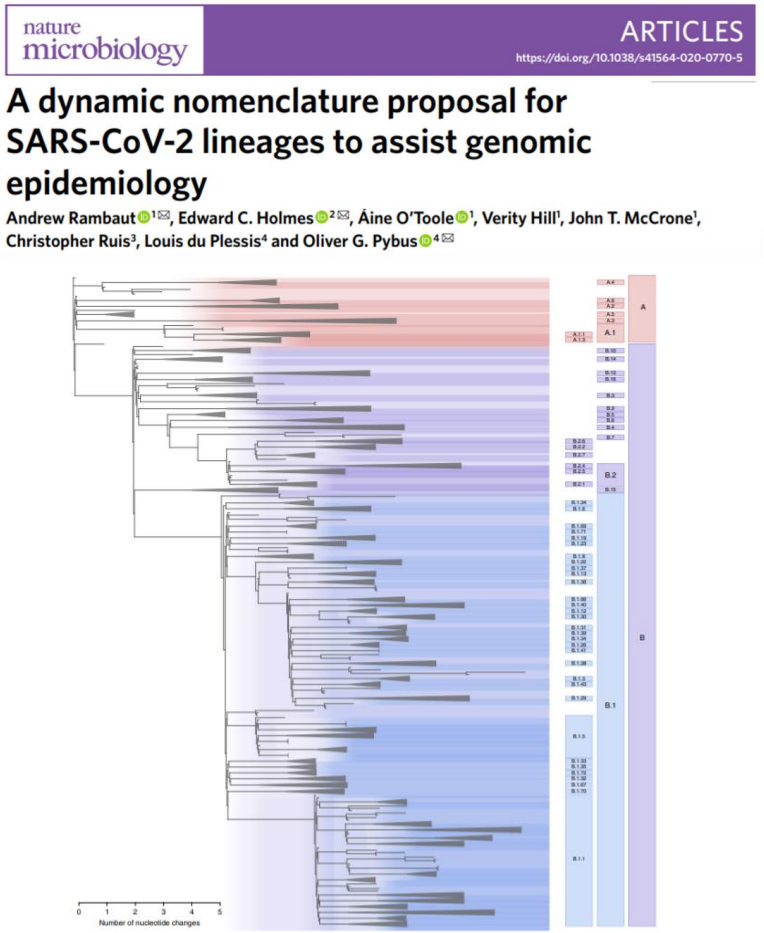
- In a global community that is continually generating data to solve "local" investigations, how do we ensure that the allele and allelic profile definitions are completely up to date?
- **Problem:** potential "collisions" if allele calling software running locally needs to assign novel alleles or allelic profiles that need to be synchronized internationally
- **Solution:** a proposed workaround is "allele hashing" → hash-cgMLST
  - Eyre et al. 2019 *J Clin Microbiol* **58**(1): e01037-19
  - Deneke et al. 2021 *Front Microbiol* **12**: 649517



- A lot of labs are performing cgMLST without "being connected" to global databases in real time (i.e. without synchronization of allele and ST definitions).
- Hashing can overcome this by generating a stable definition for a given allele/ST

Nomenclatures for genomic surveillance

- Nomenclatures are used to systematically describe the various subtypes/genetic lineages generated by a subtyping system
- Ideally, a set of rules that can be implemented algorithmically in software that can automatically assign the subtype/linage of novel genomes with minimal curation required



- Nomenclatures provide a means of efficiently communicating subtype information and facilitating tracking and monitoring of subtypes of interest
- Can facilitate data sharing in cases where the WGS data itself may not be freely shareable

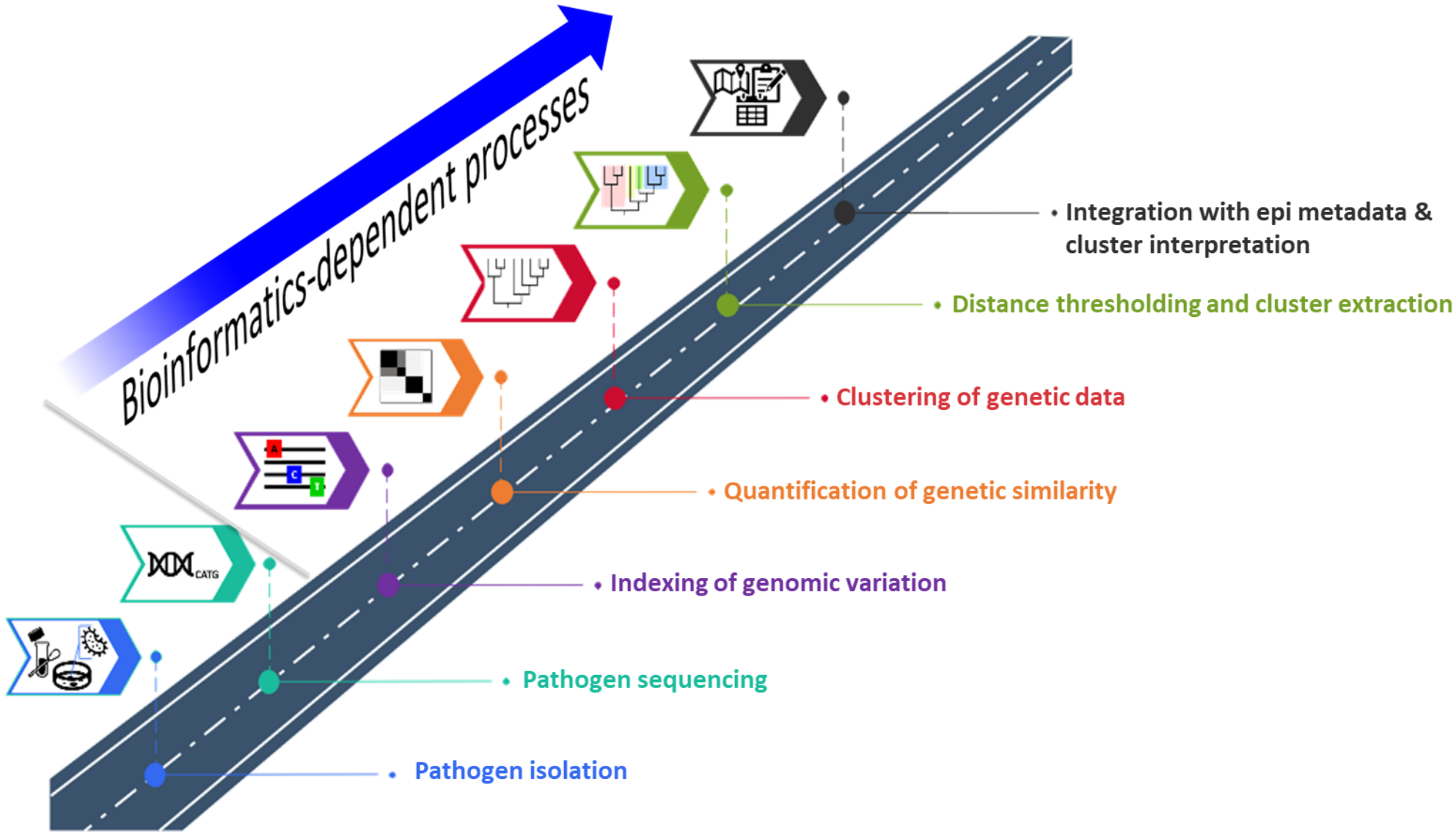
Nomenclatures for genomic surveillance

- A common theme in proposed nomenclatures is the use of several hierarchical levels reflecting different degrees of strain relatedness (i.e. lineages, sub-lineages, etc…)

Dallman et al. *Bioinformatics* **34**(17): 3028. →	**VTEC:** Seven SNP thresholds of Δ250, Δ100, Δ50, Δ25, Δ10, Δ5, Δ0.
Moura et al. *Nature Microbiol* **2**: 16185. →	***Listeria* cgMLST types:** 7 allele differences ***Listeria* cgMLST sub-lineages:** 150 allele differences
Tolar et al. *Foodborne Pathogens Dis* **16**(7). →	***Listeria*:** six cgMLST thresholds of Δ71, Δ51, Δ36, Δ19, Δ7, Δ0.
Rossi et al. INNUENDO project. →	**Level 1:** outbreak-centric **Level 2:** maximum discriminatory power but high cluster stability **Level 3:** reflecting lineages with of historical importance to surveillance

- Ongoing research to further establish optimal organism-specific ranges of strain relatedness that should be used in developing hierarchical nomeclatures that “make sense”

Recap of basic steps in a genomic epidemiology analysis workflow



Parting thoughts

- Whole-genome phylogenetic analysis can be viewed as the “gold standard” analysis of any pathogen of interest; however, such an analysis may neither be feasible nor desirable depending on the pathogen being studied
  - considerations: genome size; population structure (i.e. highly clonal vs. panmictic…); mutation vs. recombination rates; genome rearrangements; pangenome dynamics (core vs. accessory genome size); availability of other WGS data (i.e. reference genomes)
- WGS-based subtyping is the process of cataloging, tabulating and quantifying genomic variation to assess genome similarity using methods that emphasize cluster detection, robustness, and scalability
  - generally requires sacrificing phylogenetic accuracy
  - need to define criteria for data inclusion/exclusion because it may be unreliable biologically, analytically, or both.
  - Requires: methodology for computing genetic similarity; approach for clustering genetic profiles; lineage definitions (i.e. “cluster extraction) & nomenclature
  - “One-off” analysis of a dataset *does not require* WGS-based typing; using a WGS-based typing approach facilitates comparison to *other* datasets separated in time and space and genomic surveillance in a global context

Parting thoughts

- MLST-based analysis:
  - extension of the classical Multi-Locus Sequence Typing approach of Maiden et al. to hundreds or thousands of loci.
  - indexes genetic variation at the allelic level (i.e. locus variants); only allelic differences are considered, not the number of sequence differences
  - Pros: works well for investigating species that are prone to high levels of recombination and/or that have an epidemic population structure; scale well to analysis of large number of genomes (1000s of genomes); the current approach being applied for global genomic surveillance of some important bacterial pathogens
  - Cons: generally unsuitable for highly clonal or monomorphic organisms because of insufficient discriminatory power; requires availability of an MLST schema
- WGS-based subtyping analysis works best as a hybrid approach best deployed at the population-level and complemented with higher resolution methods (sgMLST or SNV-analysis) to zoom in on specific clusters

After the break…

- cgMLST analysis exercise with Guangzhi Zhang (PHAC-NML)
  - Cleaning up noisy allele data
  - Clustering cgMLST data
  - Visualizing dendrograms generated from cgMLST data
  - Extracting genomic clusters for downstream analysis
  - Coupling genomic cluster data with epidemiological data for outbreak interpretation

We are on a Coffee Break & Networking Session

Workshop Sponsors:

Canadian Centre for Computational Genomics