bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io

Supported by

McGill UNIVERSITY


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 3
WGS-based subtyping

Ed Taboada & Dillon Barker
Infectious Disease Epidemiology Analysis
October 4-6, 2021

bioinformatics.ca

SISTR:
the Salmonella
in silico typing
resource

IRIDA

Public Health Agency of Canada
Agence de la santé publique du Canada

Who we are:

- Ed Taboada (ed.taboada@phac-aspc.gc.ca)
- Dillon Barker (dillon.barker@phac-aspc.gc.ca)

Our affiliation:

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

What we do:

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

Learning Objectives of Module

- A basic introduction to infectious disease epidemiology:
  - the role of molecular epidemiology; the role of molecular subtyping

- Variant analysis: from sequencing reads to subtypes
  - SNV analysis
    - Basic how-to
  - Multi Locus Sequence Typing (MLST)
    - MLST "classic", wgMLST, cgMLST, sgMLST
  - SNV and MLST pros and cons

- Genomic surveillance:
  - bacterial population structure & clonal complexes
  - what is a nomenclature
  - why WGS-based subtyping is necessary for genomic surveillance

Populations aren't homogeneous…

- Population distribution of:
  - exposure to risk factors
  - distribution of disease

} epidemiology

What's 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."

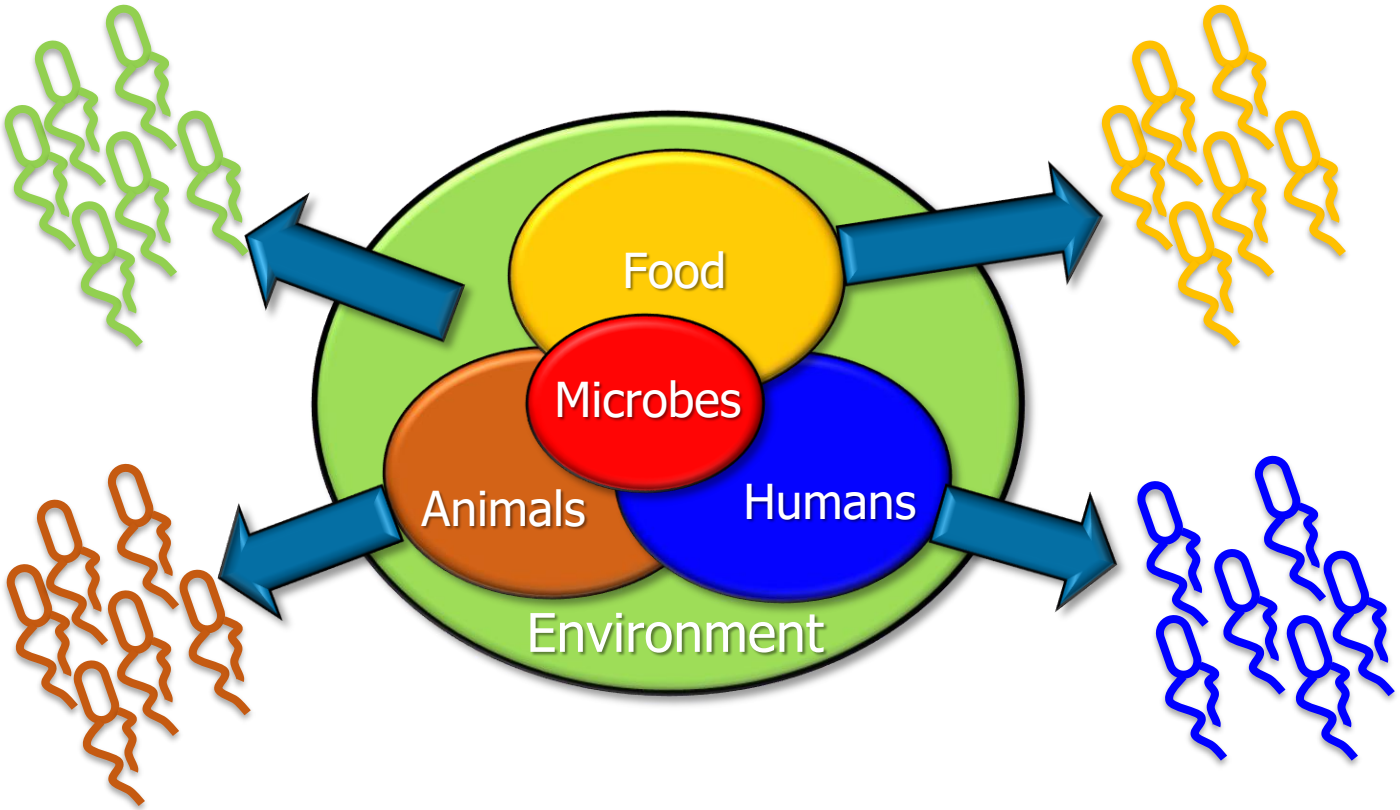
Foxman and Riley, *Am J Epidemiol* **153**:1135.

- 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

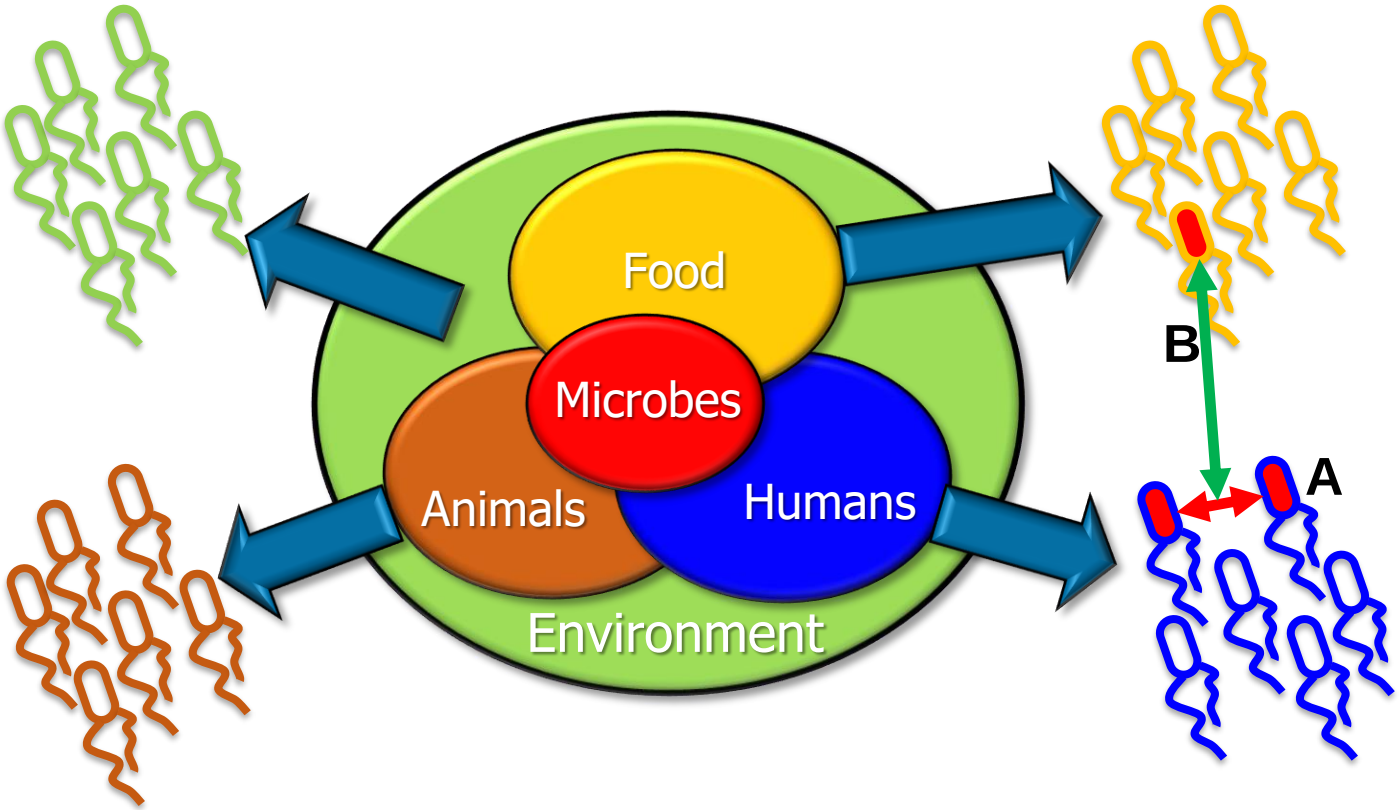
- A fairly simple set of guiding principles:
  - If isolates are epidemiologically linked, they should be "genetically identical"
  - If isolates are not epidemiologically linked, they should not be "genetically identical"
  - We expect agreement between genetic and epidemiologic observations
  - Use **molecular subtyping** to assess genetic similarity between isolates

Molecular surveillance and epidemiology



1) Sampling of potential sources of exposure
2) Molecular subtyping of isolates
3) Comparison of the isolate subtyping data
4) Examine epidemiology of matching isolates

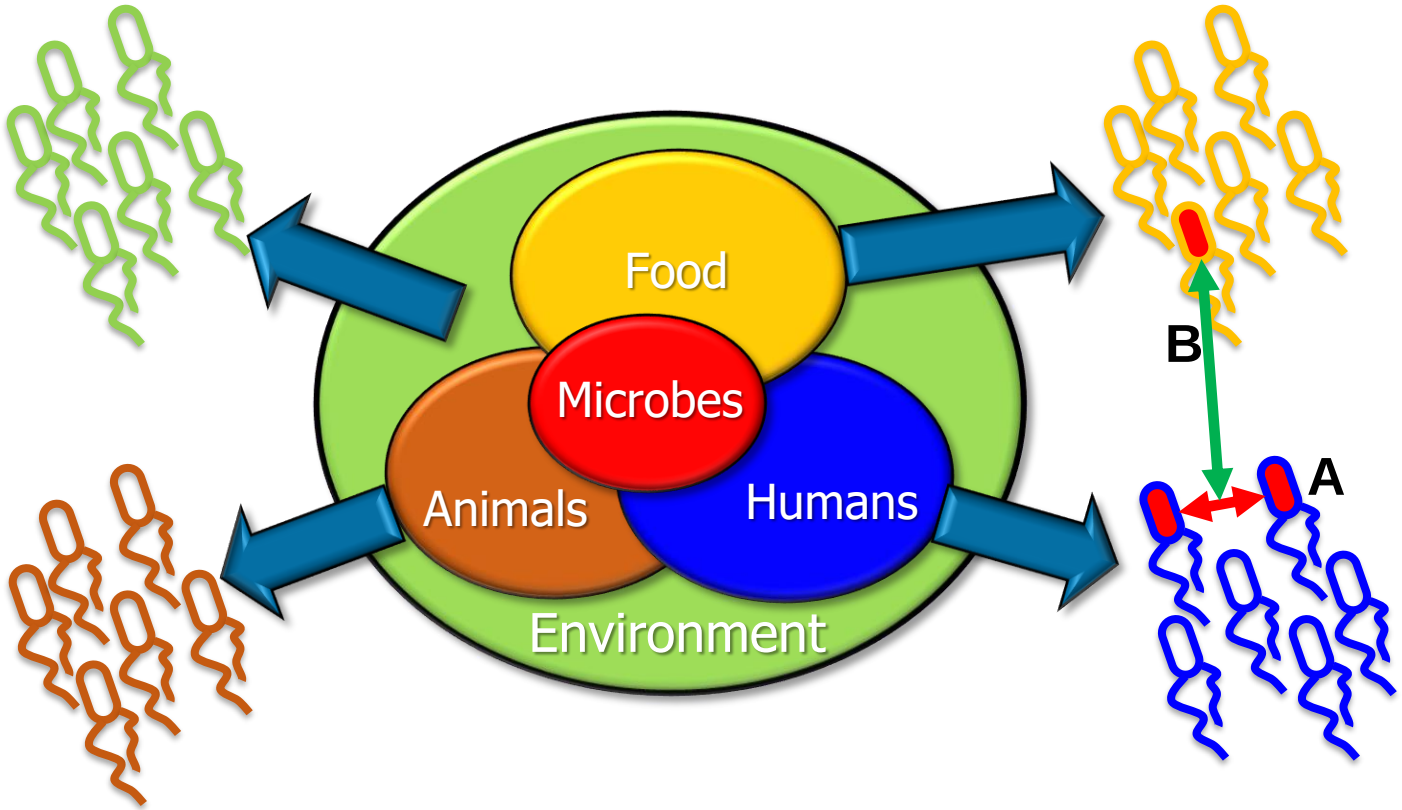
Molecular surveillance and epidemiology



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

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

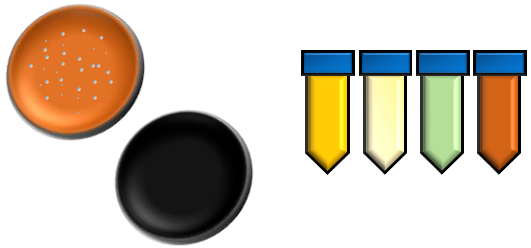
Molecular surveillance and epidemiology



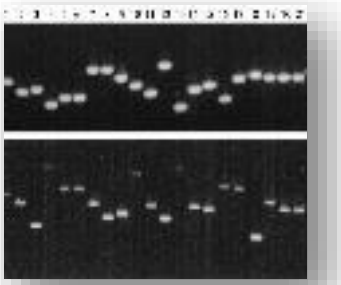
- The challenge has always been the limitations of molecular subtyping methods in assessing genetic similarity
- Easy to **over-** or **under-**estimate genetic similarity and misinterpret the significance of matches/mismatches

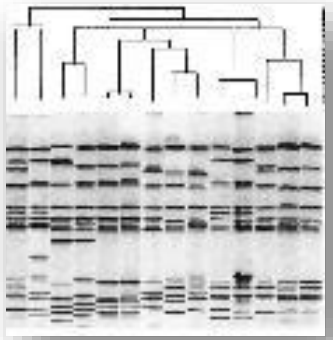
The molecular subtyping paradigm

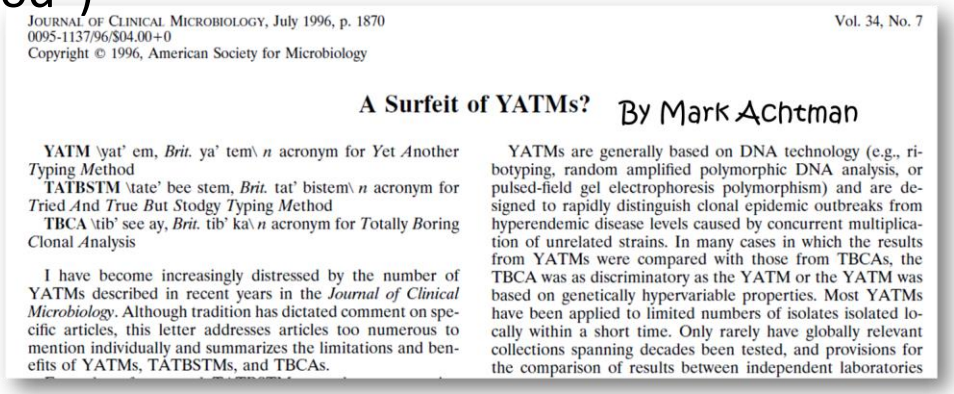
- In the beginning, methods based on phenotypic characteristics
  - Serotyping (i.e. serological)
  - Biotyping (i.e. biochemical)



- 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 Variable number tandem repeat Analysis (MLVA)
  - Amplified Fragment Length Polymorphisms (AFLP)
  - **YATMs** (“Yet Another Typing Method”)







JOURNAL OF CLINICAL MICROBIOLOGY, July 1996, p. 1870 Vol. 34, No. 7
0095-1137/96/$04.00+0
Copyright © 1996, American Society for Microbiology

A Surfeit of YATMs? By Mark Achtman

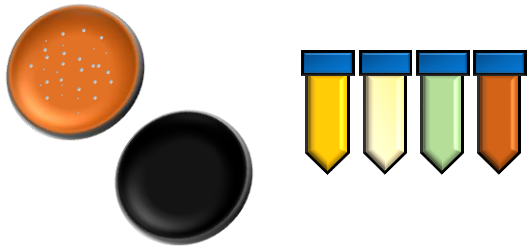
YATM \yat' em, *Brit.* ya' tem\ *n* acronym for *Y*et *A*nother *T*yping *M*ethod
TATBSTM \tate' bee stem, *Brit.* tat' bistem\ *n* acronym for *T*ried *A*nd *T*rue *B*ut *S*todgy *T*yping *M*ethod
TBCA \tib' see ay, *Brit.* tib' ka\ *n* acronym for *T*otally *B*oring *C*lonal *A*nalysis

I have become increasingly distressed by the number of YATMs described in recent years in the *Journal of Clinical Microbiology*. Although tradition has dictated comment on specific articles, this letter addresses articles too numerous to mention individually and summarizes the limitations and benefits of YATMs, TATBSTMs, and TBCAs.

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 from YATMs were compared with those from TBCAs, the TBCA was as discriminatory as the YATM or the YATM was based on genetically hypervariable properties. Most YATMs have been applied to limited numbers of isolates isolated locally within a short time. Only rarely have globally relevant collections spanning decades been tested, and provisions for the comparison of results between independent laboratories

The molecular subtyping paradigm

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

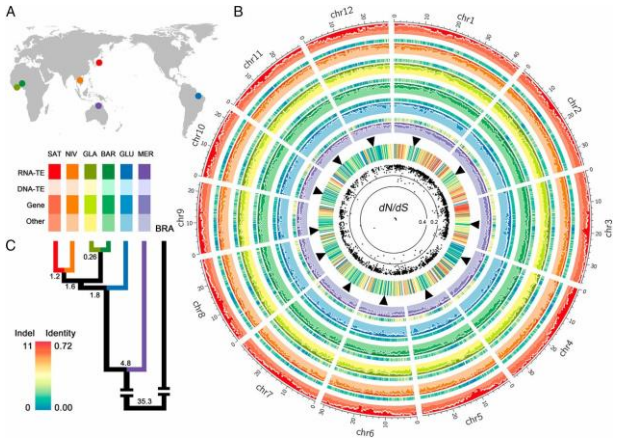


- In the 90s-00s, emergence of methods based on comparing variation in DNA banding patterns on gels (i.e. “DNA fingerprints”)
  - Pulsed Field Gel Electrophoresis (PFGE)
  - Multiple-Locus Variable number tandem repeat Analysis (MLVA)
  - Amplified Fragment Length Polymorphisms (AFLP)
  - **YATMs** (“Yet Another Typing Method”)

- What we <u>needed</u>: genetic data on isolates from surveillance sampling to identify those that were genetically similar
- What we <u>had</u>: a bunch of proxy methods to estimate genetic similarity
- YATM phenomenon due to need to find methods with “improved” convenience and accuracy

The WGS-based Subtyping Paradigm

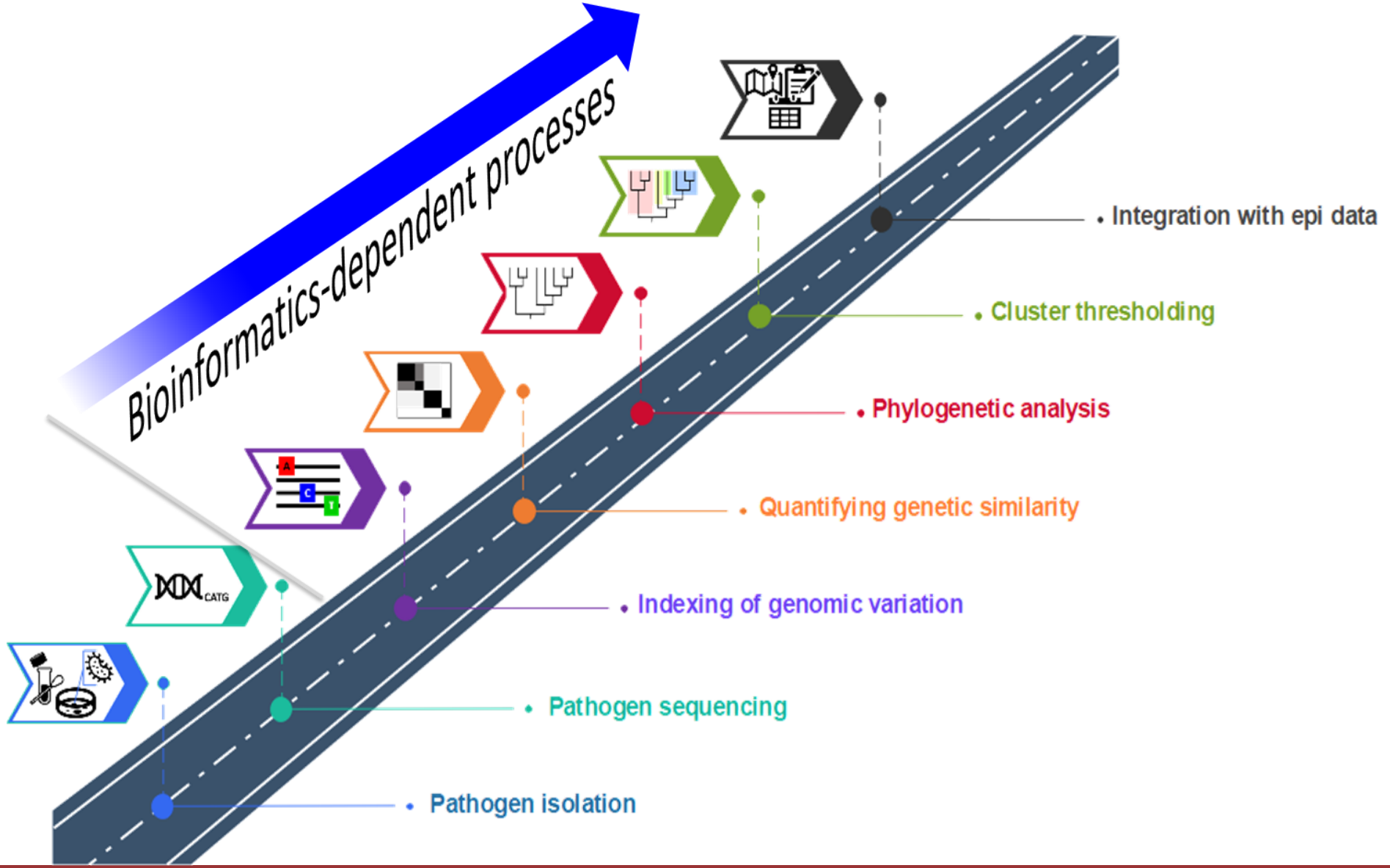
- Goal is still to estimate genetic similarity between isolates obtained through surveillance sampling
- The difference: ability to use the full weight of WGS data to make these estimates
- "Comparative genomics" & "variant analysis":
  - **Region-level** variation
    - Strain-specific chromosomal regions (e.g. genes, non-coding regions)
    - 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

- Development of methods for extracting & cataloging variant information from WGS data and for computing genetic similarity estimates

Basic steps in a genomic epidemiology workflow



Variant Analysis (part I): SNV analysis

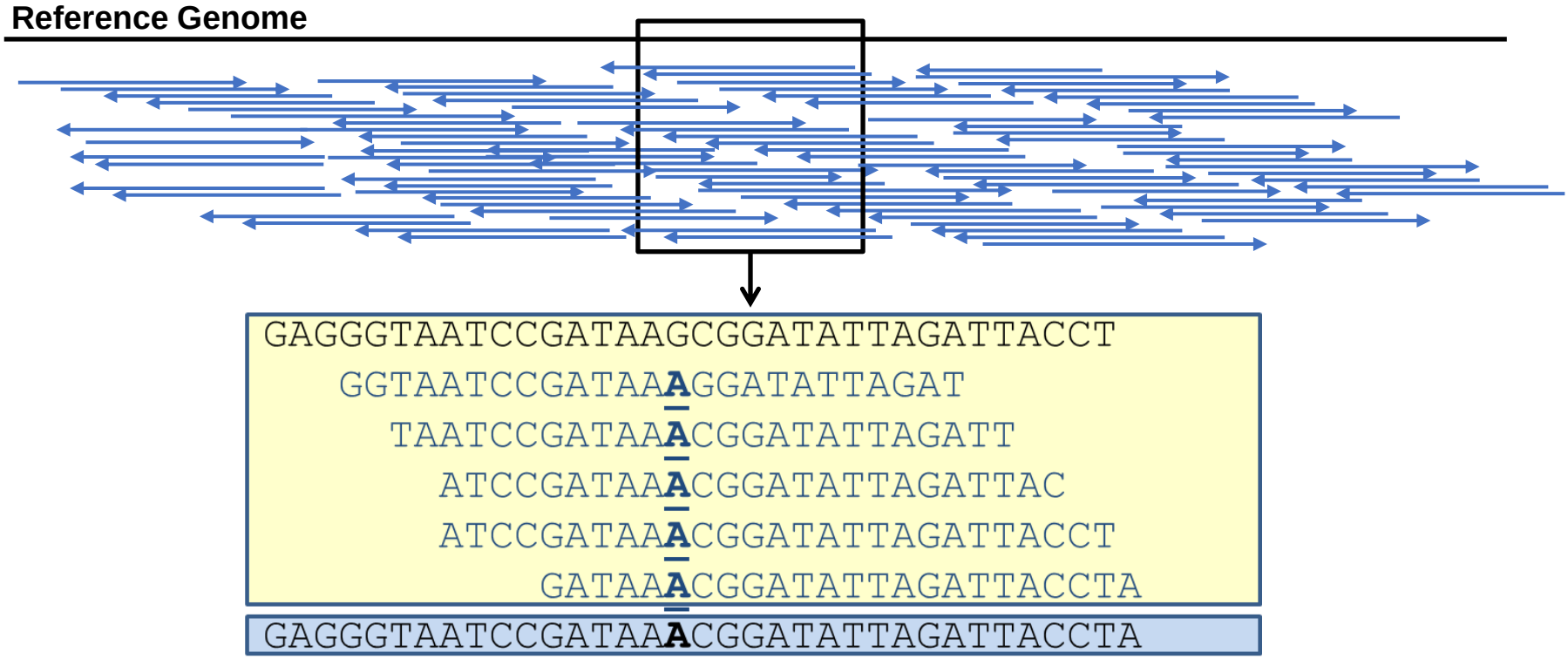
SNV analysis

- Variant analysis using SNVs is a character-based method based on comparing single nucleotide variants in each genome

- Single Nucleotide Variant (SNV): DNA sequence variation that occurs when a single nucleotide is altered in a genome sequence; SNVs that are observed in >1% of the population are considered SNPs

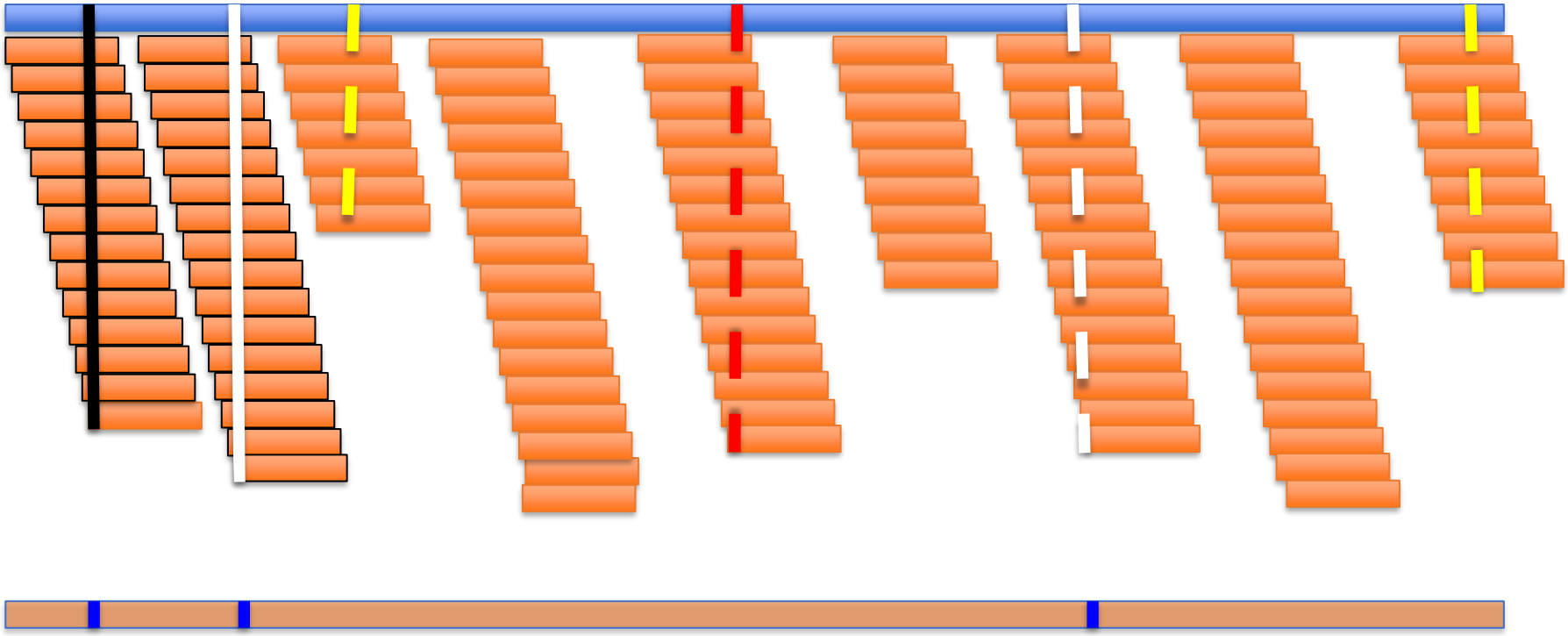
- Most popular approach is based on **reference mapping** to extract the single nucleotide variants (SNVs) from each genome.
  - (suitable) SNVs are identified in each genome via comparison to a reference genome
  - SNVs are tabulated into a multiple sequence alignment (MSA).
  - MSA is used to infer a maximum likelihood phylogeny.

SNV analysis: reference mapping



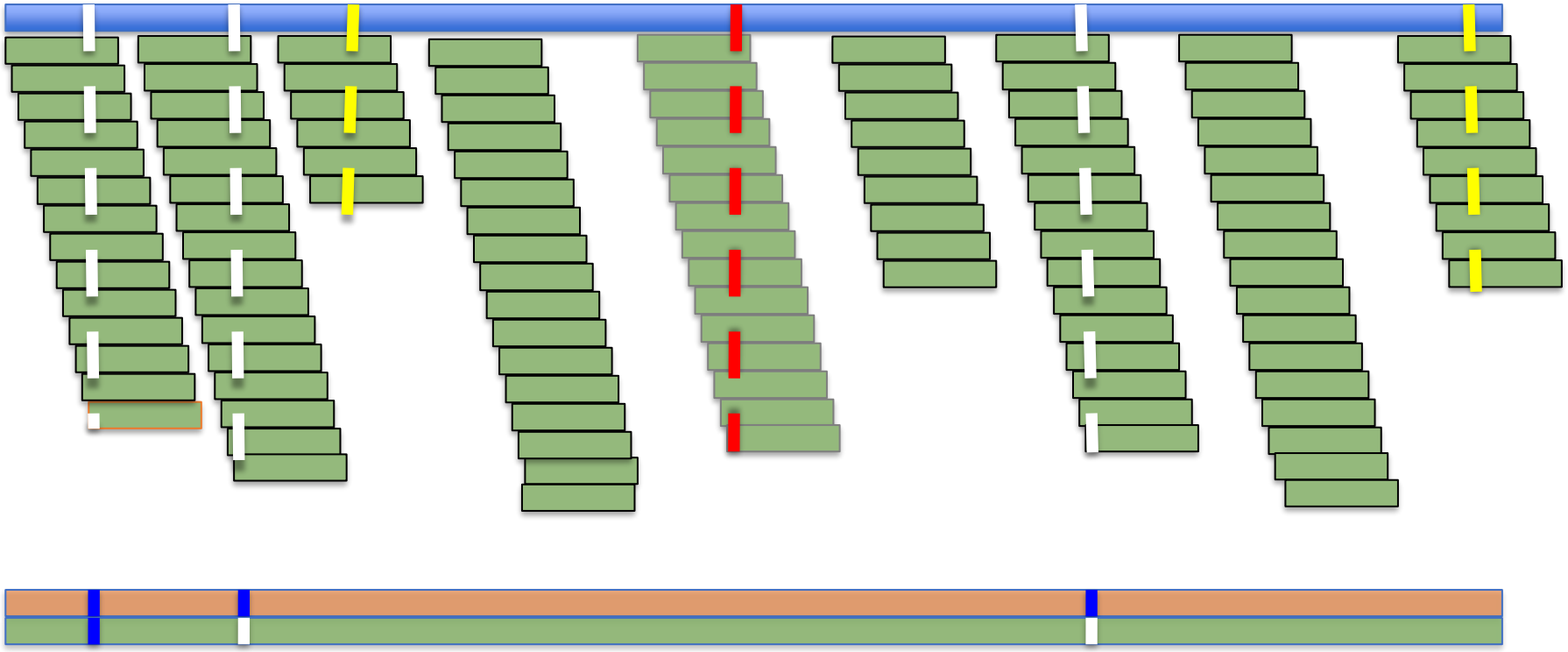
- Reference mapping assembles sequencing reads by aligning them to a **highly similar** reference genome
- Aligned reads are referred to as a '**pileup**'
- Main application is in SNV discovery

SNV analysis: reference mapping



- Map reads from first genome to reference genome and collect SNVs (in blue) → unambiguous & good coverage
- Avoid poor quality SNVs (red) and low coverage SNVs (yellow).

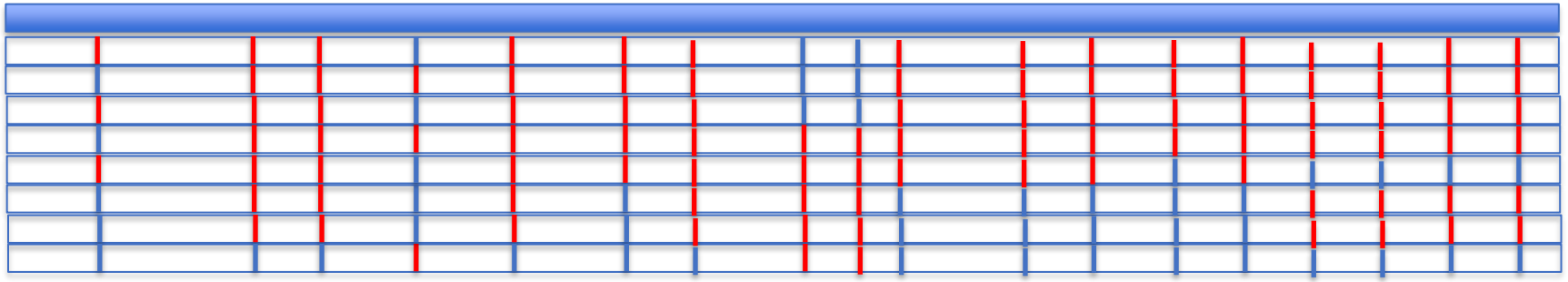
SNV analysis: reference mapping



- Map reads from 2nd genome to reference and collect SNVs (in blue; black lines represent wildtype)

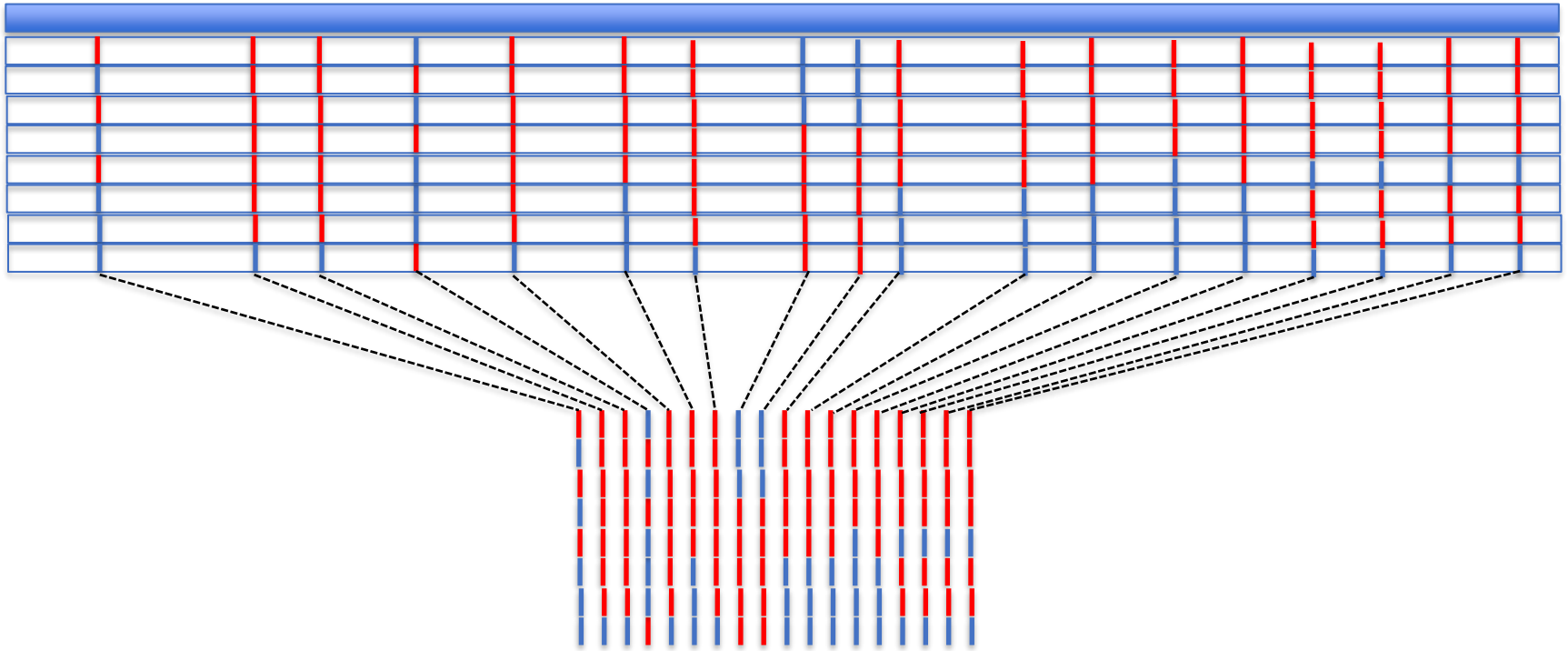
SNV analysis: reference mapping

- Continue until all SNVs have been identified for all genomes.



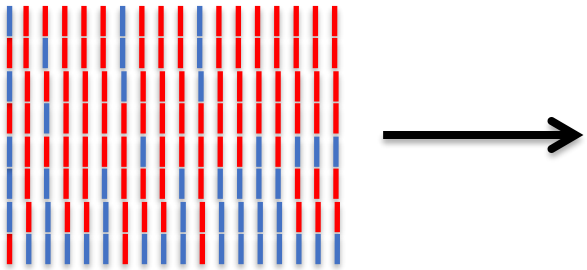
SNV analysis: SNV alignment

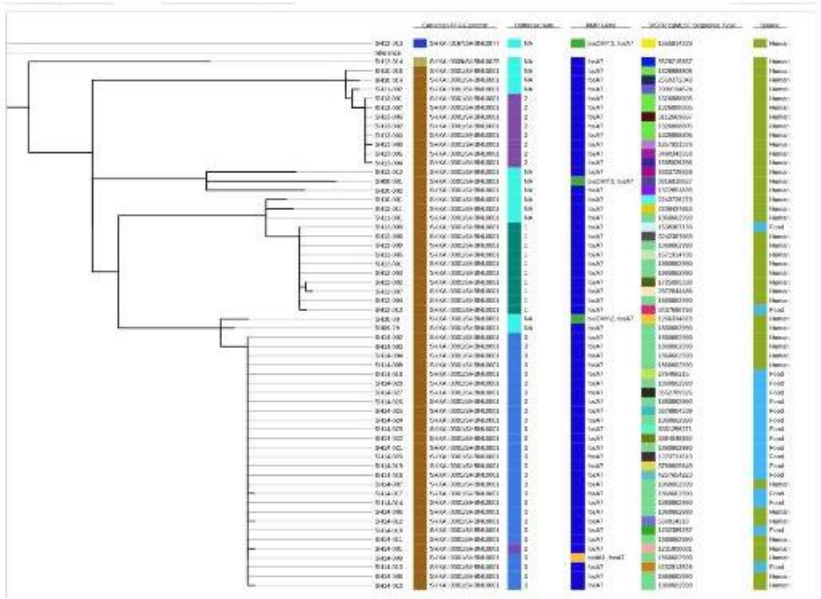
- Compress into a ‘SNV-alignment’.



SNV analysis: phylogeny

- Build a phylogenetic tree from the SNV alignment.





SNV analysis: phylogeny

- Build a phylogenetic tree from the SNV alignment.

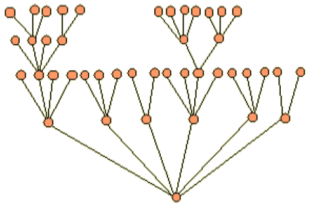
- Pros:
  - replacement of subtyping with “true” phylogenetic analysis
  - well-suited for highly similar genomes where most of the variation is contained within SNVs.
- Cons:
  - Read mapping works best if there is a **good** reference genome
  - Issues with possible inclusion of **bad** SNVs: homoplastic SNVs due to recombination; paralogous genes not present in reference genome

A detour into microbial population structure…

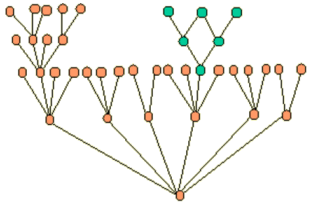
Microbial Population Structure (part 1)

"…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…."

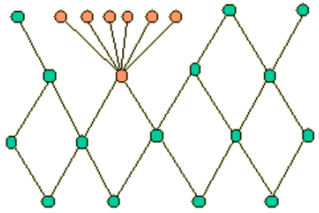
Spratt and Maiden (1999) *Phil Trans R Soc Lond* **354**: 701.



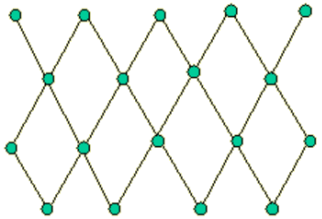
Clonal



Weakly Clonal



Epidemic

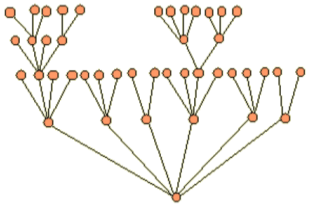


Panmictic

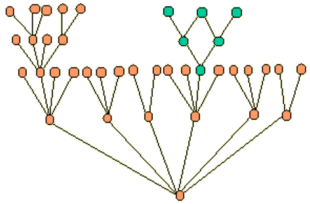
Microbial Population Structure (part 1)

"…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…."

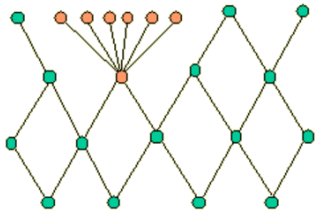
Spratt and Maiden (1999) *Phil Trans R Soc Lond* **354**: 701.



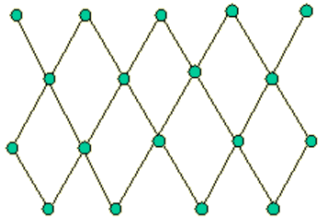
Clonal



Weakly Clonal



Epidemic

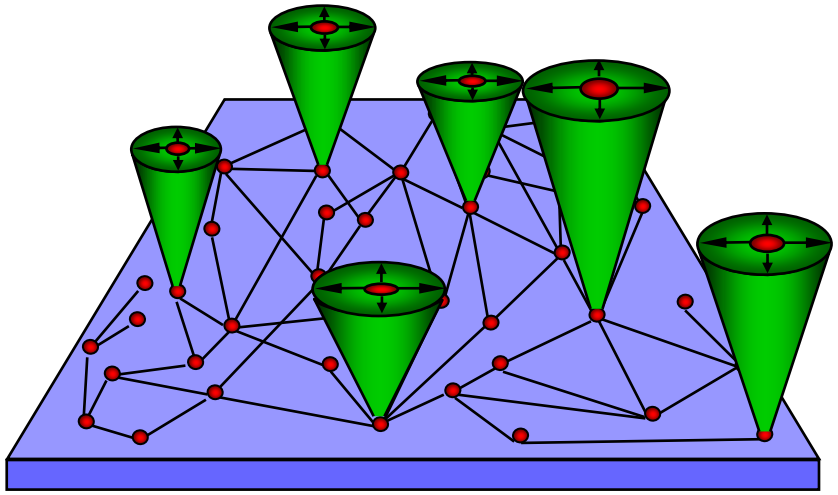


Panmictic

- 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
- More recombination distorts evolutionary relationships

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 (i.e. "clonal expansion")
    - Clones undergo genetic diversification through recombination and mutation
    - More tree-like

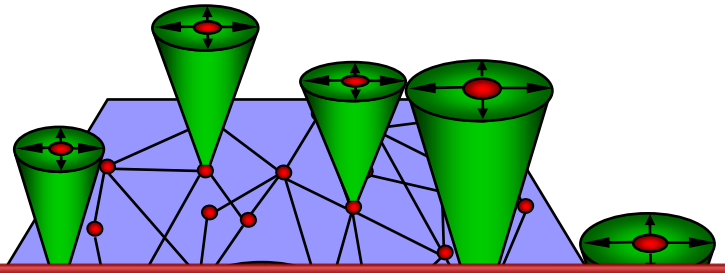


"…the background population is composed of a large number of relatively rare and unrelated genotypes (small circles) that are recombining at a high frequency…most accurately represented as a **network**, rather than a bifurcating tree, as *recombination has overwhelmed the phylogenetic signal*…a limited 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 (i.e. "clonal expansion")
    - Clones undergo genetic diversification through recombination and mutation
    - More tree-like

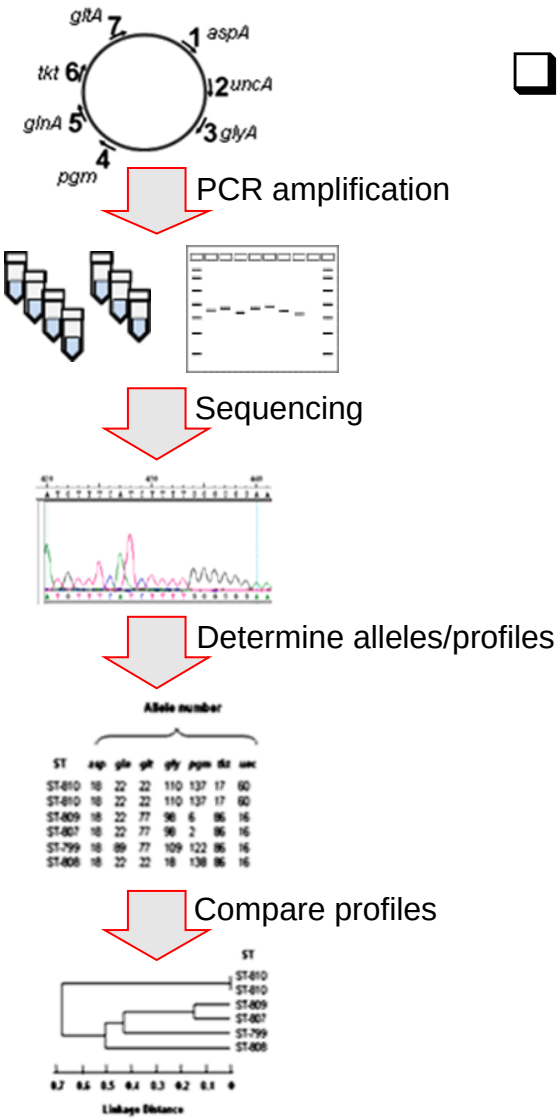


"…the background population is composed of a large number of relatively rare and unrelated genotypes (small circles) that are recombining at a high frequency…most accurately represented as a **network**, rather than a bifurcating tree, as *recombination has overwhelmed the phylogenetic signal*…a limited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

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

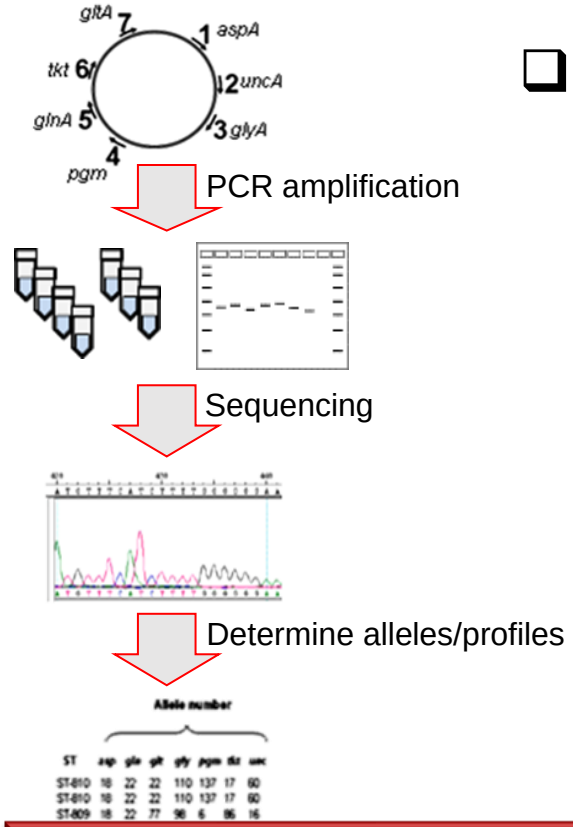
Variant Analysis (part II): MLST analysis

Multi Locus Sequence Typing



- Described by Maiden et al. (1998) *PNAS USA* **95**: 3140.
  - Analysis of 7 to 9 loci
    - "housekeeping" genes (i.e. core genes)
    - 450-500 bp gene fragments
    - Genes distributed around the genome
  - Each locus is PCR amplified and sequenced
    - For each locus we determine allele based on finding a match in central database
    - **Sequence Type** is assigned based on the combination of alleles at all the loci, also held in central database
  - Central database also keeps track of novel alleles and novel combinations of alleles

Multi Locus Sequence Typing



- Described by Maiden et al. (1998) *PNAS USA* **95**: 3140.
  - Analysis of 7 to 9 loci
    - "housekeeping" genes (i.e. core genes)
    - 450-500 bp gene fragments
    - Genes distributed around the genome
  - Each locus is PCR amplified and sequenced
    - For each locus we determine allele based on finding a match in central database
    - **Sequence Type** is assigned based on the combination of alleles at all the loci, also held in central database

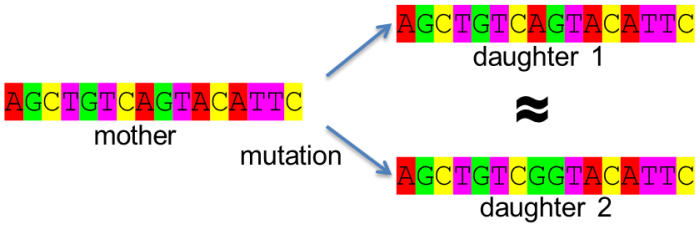
- MLST became the gold standard method in molecular epidemiology
  e.g. sequence data vs. fragments on gels
- Over 50 schemes developed for different species & used in hundreds of published studies

Mutation and recombination in phylogenetics

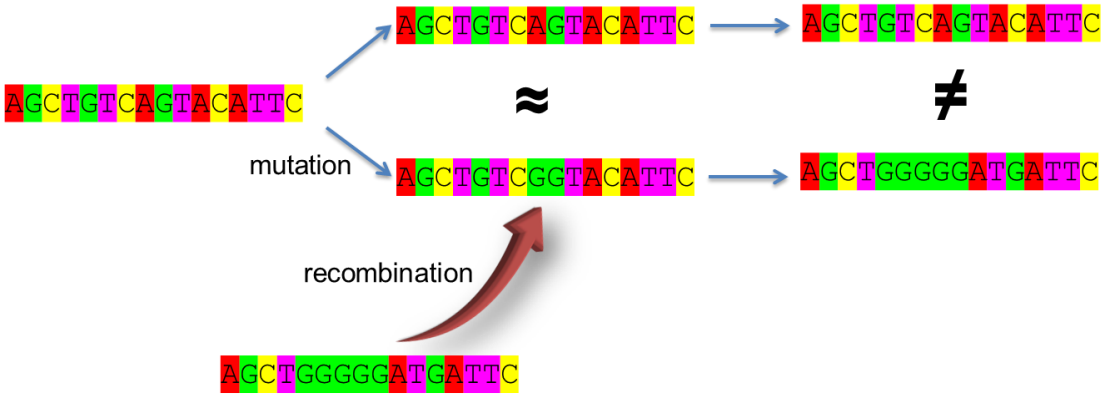
- Wait….isn't collapsing sequence data to an allele type a waste of perfectly good sequence data for phylogenetic analysis???

http://www.istockphoto.com/stock-photo-378740-money-down-the-toilet.php

- Phylogenetic analysis assumes that sequences 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

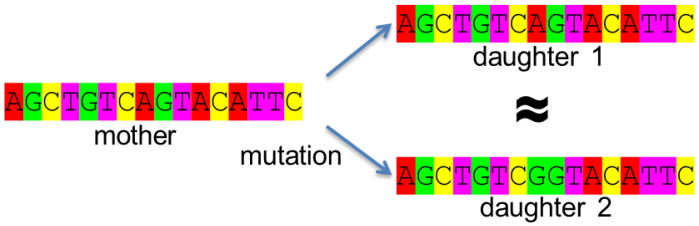


Mutation and recombination in phylogenetics

- Wait….isn't collapsing sequence data to an allele type a waste of perfectly good sequence data for phylogenetic analysis???

http://www.istockphoto.com/stock-photo-378740-money-down-the-toilet.php

- Phylogenetic analysis assumes that sequences 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

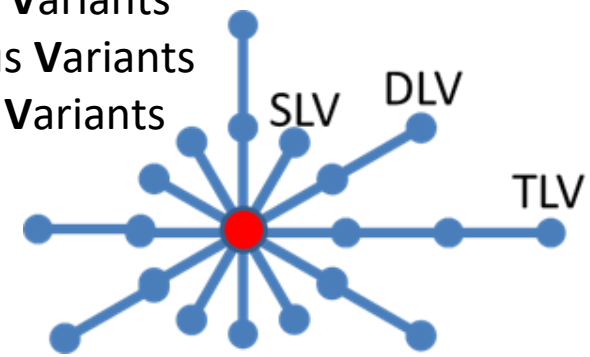
- Mutation during DNA replication is a process for <u>vertical</u> transfer of genetic information → phyogenetic signal
- Recombination is a process for <u>lateral</u> transfer of genetic information (i.e. no need for common ancestry) → distorted phylogenetic signal

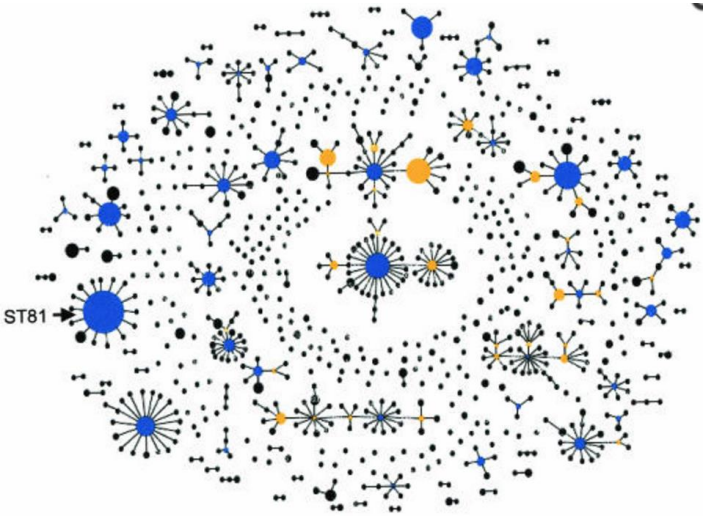
Bacterial Population Structure and MLST

- BURST (**B**ased **U**pon **R**elated **S**equence **T**ypes): original clustering algorithm for MLST developed by Feil et al. 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 group ≈ Clonal Complex
  - eBURST: advanced implementation with Minimum-Spanning Tree visualization
  - 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





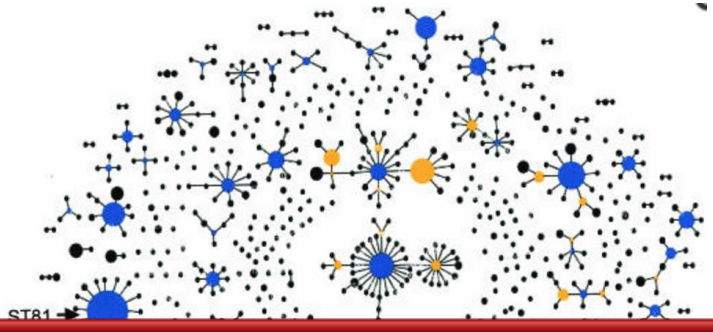
eBURST analysis of *S. pneumoniae* population structure. Feil et al. *J Bacteriol* **186**(5): 1518.

Bacterial Population Structure and MLST

- BURST (**B**ased **U**pon **R**elated **S**equence **T**ypes): original clustering algorithm for MLST developed by Feil et al. 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 group ≈ Clonal Complex
  - eBURST: advanced implementation with Minimum-Spanning Tree visualization
  - 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



- Many bacterial species do not generate tree-like phylogenies
  → Different genes generate incongruent phylogenetic signal
- For non-clonal, epidemic species, eBURST analysis and the Clonal Complex concept are particularly useful

MLST nomenclature

- Each locus is assigned an allele by finding a match in the central MLST database
- Sequence Type is assigned based on the combination of alleles at all loci, also held in central MLST database
- Clonal Complexes are based on eBURST analysis
- Each novel allele/Sequence Type is assigned a number in order of discovery

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

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

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

MLST nomenclature

- Each locus is assigned an allele by finding a match in the central MLST database
- Sequence Type is assigned based on the combination of alleles at all loci, also held in central MLST database
- Clonal Complexes are based on eBURST analysis
- Each novel allele/Sequence Type is assigned a number in order of discovery

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

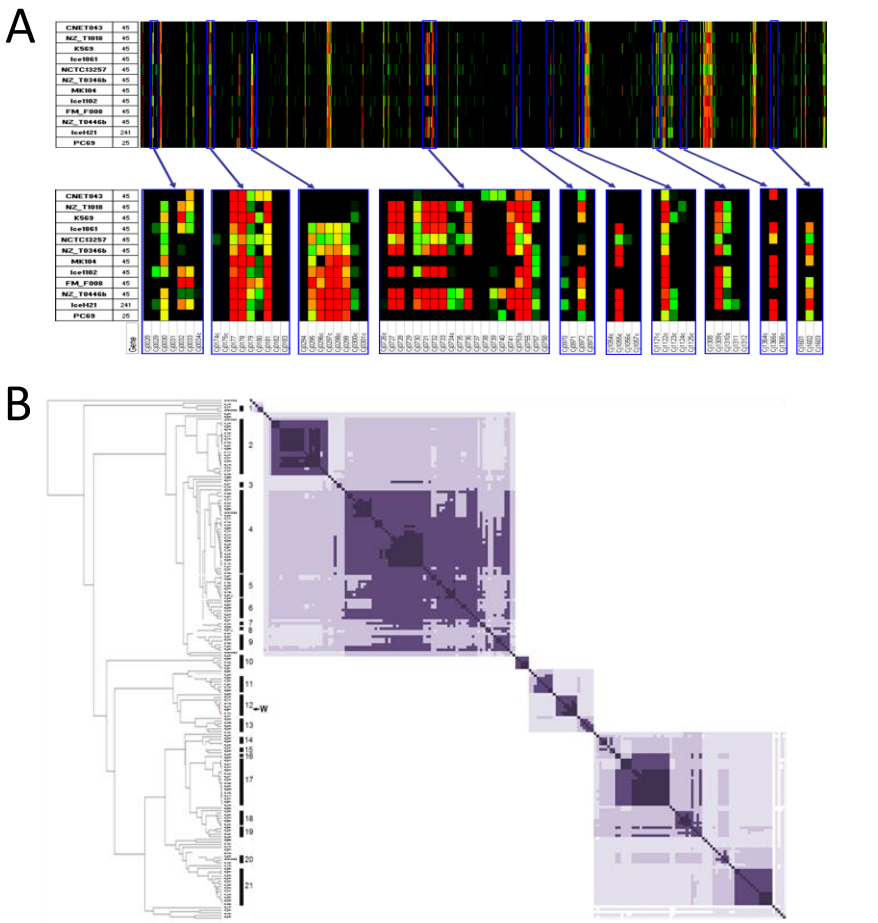
→ This allelic profile = Sequence Type 21 (ST 21)

→ ST 50 is a **Single Locus Variant** (SLV) of ST 21 because

- Allele and Sequence Type assignment is performed by comparison against central MLST database
  - Allele and ST definitions are universal and centralized
- MLST databases are meticulously curated
  - All novel alleles are verified manually before inclusion

The problem with MLST…

- Strains that are “identical” by MLST may be very genomically distinct
- 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

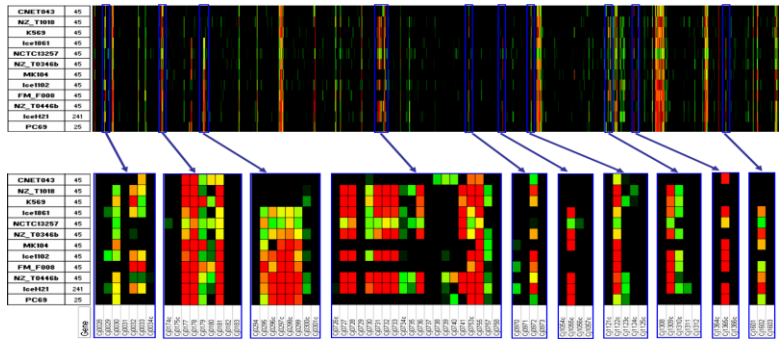


Analysis of *C. jejuni* isolates identical by MLST using (A) whole-genome microarrays [Taboada et al. *BMC Evol Biol* **8**: 229.] and (B) MLST using expanded set of 697 loci (Barker and Taboada, unpublished).

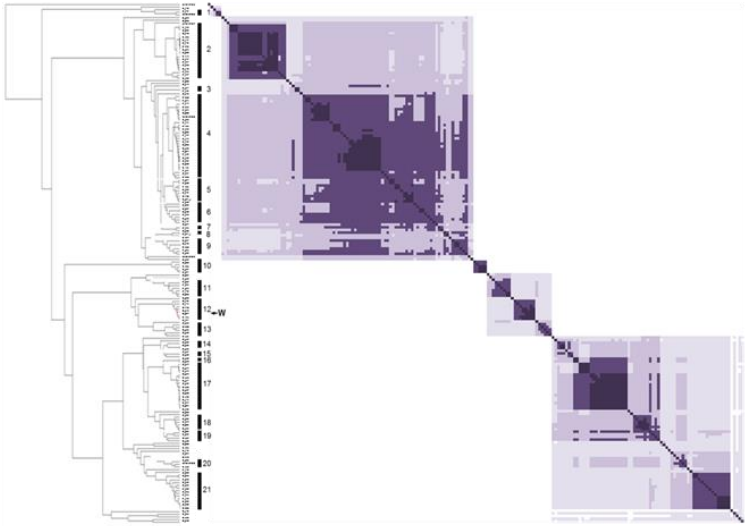
The problem with MLST…

- Strains that are “identical” by MLST may be very genomically distinct
- 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

A



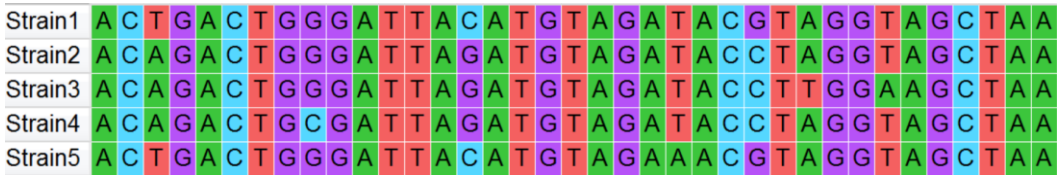
B

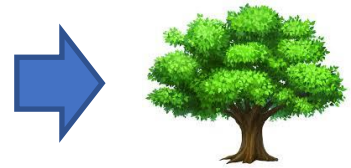


- MLST tends to be good for long-term tracking of lineages but lacks power for outbreak investigations

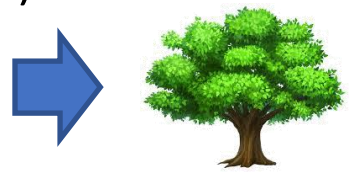
Why MLST analysis fails in outbreak analysis

- Phylogenetic analysis of **one** MLST locus
  - (≈ 450 bp → comparison of ≈ 450 data points)



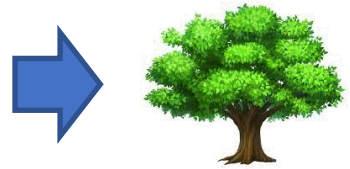


- Phylogenetic analysis of **seven** MLST loci
  - ≈ 450 bp x 7 → comparison of ≈ 3,150 data points)



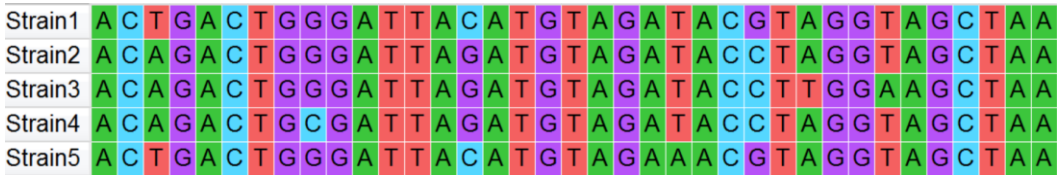
- MLST analysis
  - comparison of 7 data points!!!

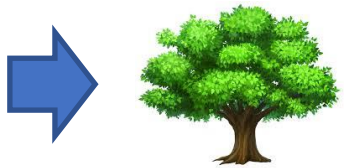
	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
Strain1	2	3	7	4	1	14	1
Strain2	1	12	13	2	5	9	2
Strain3	1	12	5	2	5	2	10
Strain4	1	12	13	2	5	7	2
Strain5	2	5	7	8	1	14	1



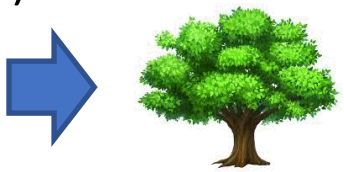
Why MLST analysis fails in outbreak analysis

- Phylogenetic analysis of **one** MLST locus
  - (≈ 450 bp → comparison of ≈ 450 data points)





- Phylogenetic analysis of **seven** MLST loci
  - ≈ 450 bp x 7 → comparison of ≈ 3,150 data points)

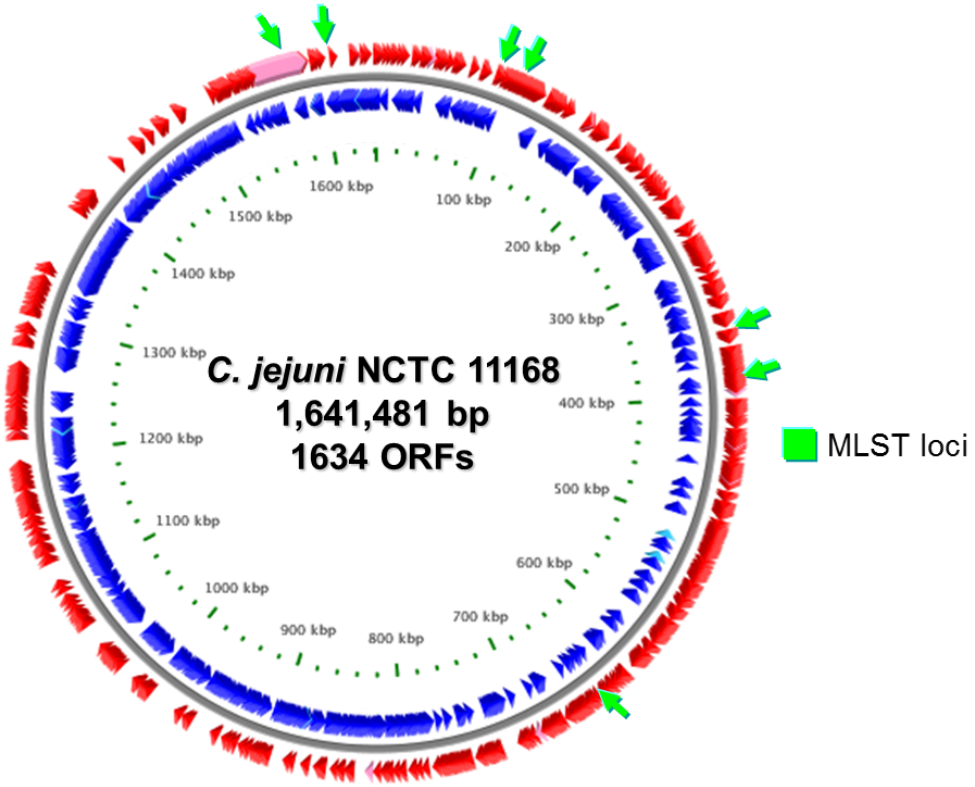


- MLST analysis
  - comparison of 7 data points!!!

- 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
- Strains are compared by the number of matching alleles at all 7 loci

The solution? genome-scale MLST

- Fast, cheap high-throughput sequencing platforms allow us to generate massive amounts of sequence data
- Genome-scale MLST extends the concept to hundreds or thousands of loci
- Increasing availability of WGS data allows development of in silico prototype schemes
- Increased resolution of WGS + portability of MLST

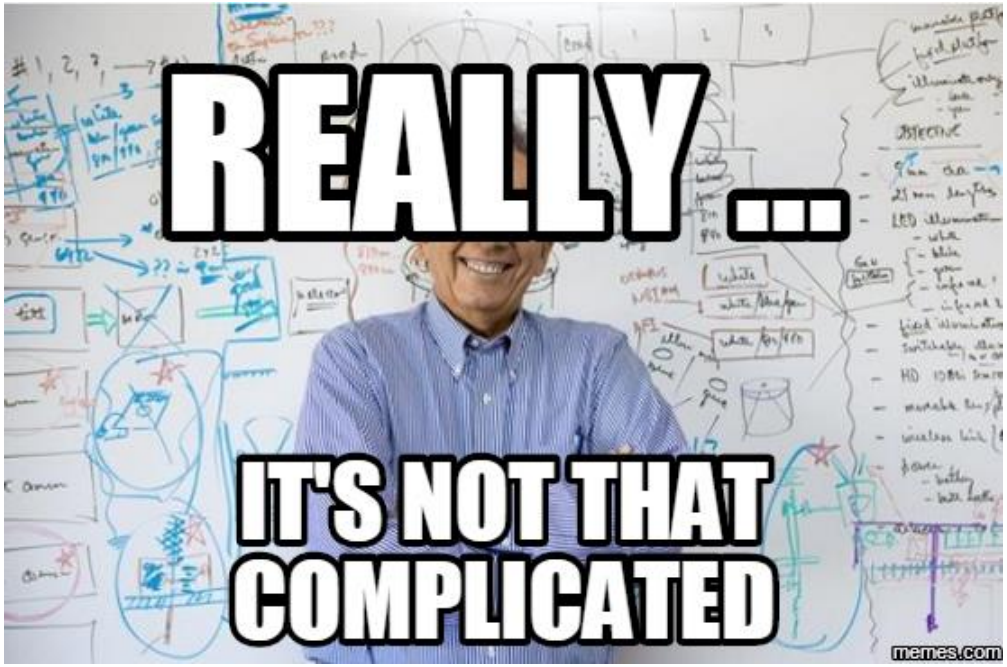


- Easier and cheaper to perform draft whole-genome sequencing and to extract MLST data from WGS data

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)

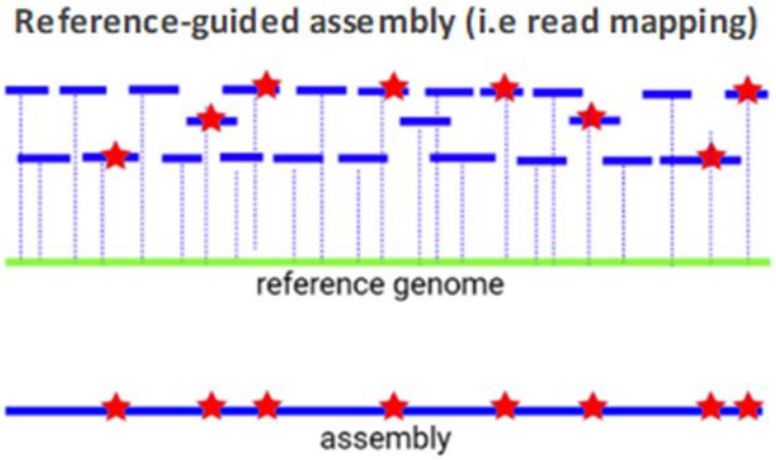


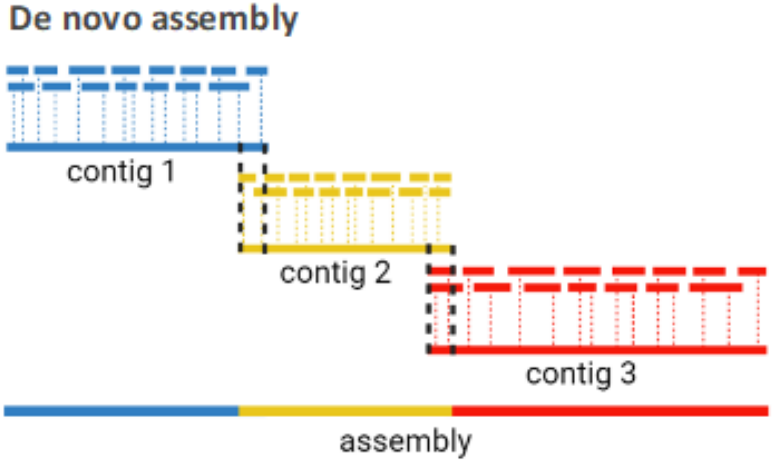
The challenges of scaling-up MLST (Part 1)

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

Reference-guided assembly: assembly of reads into a draft genome by performing mapping of sequencing reads to a reference genome.

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 extend them into longer contiguous sequences (i.e. contigs).

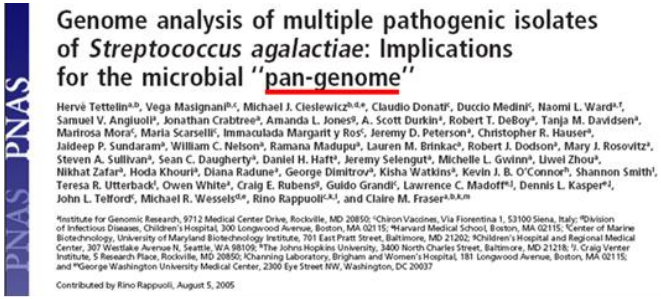




The challenges of scaling-up MLST (Part 2)

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

What are "Core genes"? "Accessory genes"?



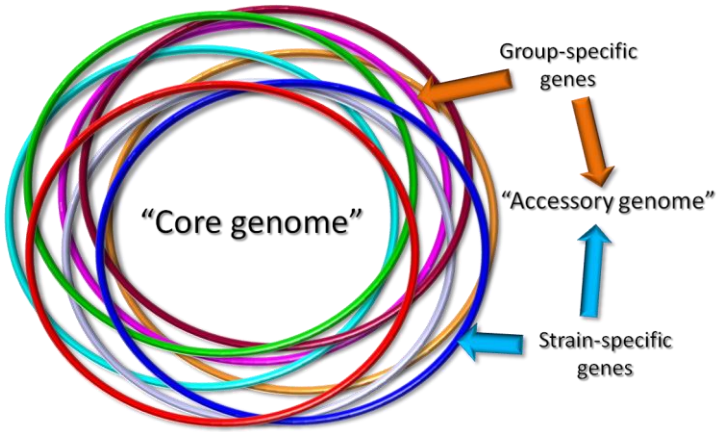
Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome"

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[a], Owen White[a], Craig E. Rubens[g], Guido Grandi[c], Lawrence C. Madoff[e,i], 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] 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.

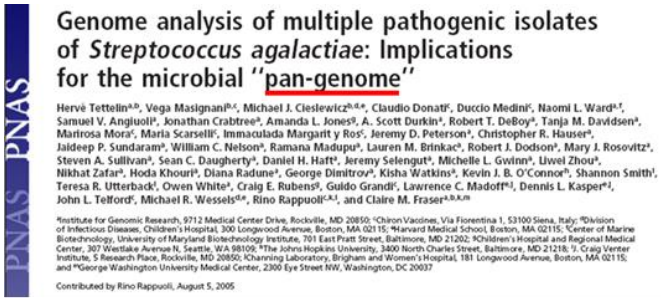


- **Core** genes
  - "Housekeeping" genes
  - Essential for the species
- **Accessory** genes
  - AMR genes, "virulence" genes, carbon source utilization, etc…
  - generally clustered in "plasticity regions"
  - carriage is highly variable in the population
- Pan-genome = **Core** + **Accessory**

The challenges of scaling-up MLST (Part 2)

- 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)

What are "Core genes"? "Accessory genes"?



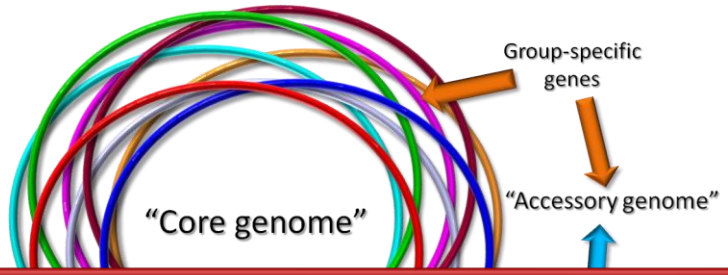
Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome"

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[a], Owen White[a], Craig E. Rubens[g], Guido Grandi[c], Lawrence C. Madoff[e,i], 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.



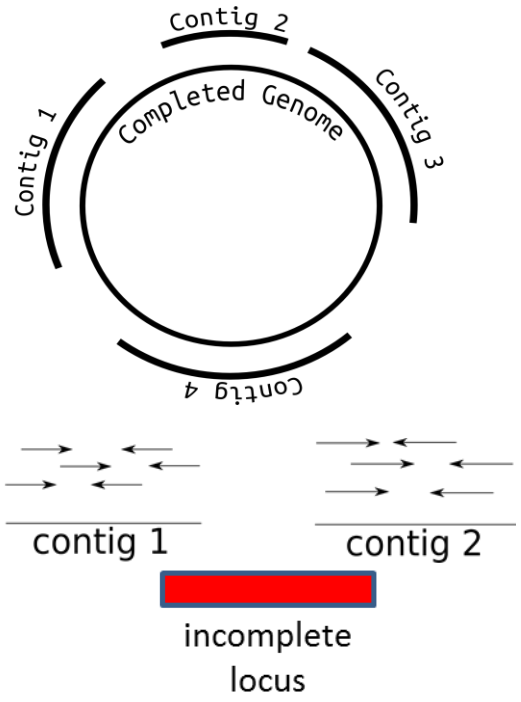
- **Core** genes
  - "Housekeeping" genes
  - Essential for the species
- **Accessory** genes
  - AMR genes, "virulence" genes, carbon source utilization, etc…
  - generally clustered in "plasticity regions"
  - carriage is highly variable in the population

- Accessory genes are problematic because it is difficult to know if a gene should be present or absent

The challenges of scaling-up MLST (Part 3)

- wgMLST will require fishing out the loci out of the WGS data, this is not trivial:
  → accessory genes may or may not be present in a genome

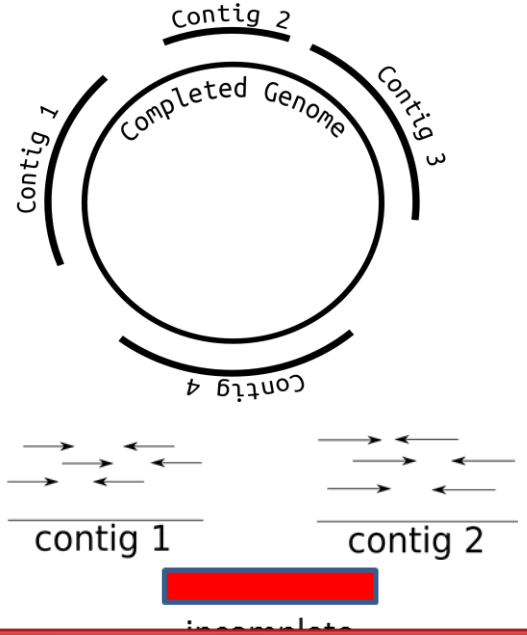
- WGS projects don't usually generate complete genomes → "Genome Assemblies"
  - A significant proportion of genome assemblies (30-40%) do not have complete data at all loci
  - Incomplete data generally due to "collision" between a locus and a gap in the assembly
    → incomplete (i.e. truncated) allele
  - *Mostly* randomly distributed



The challenges of scaling-up MLST (Part 3)

- wgMLST will require fishing out the loci out of the WGS data, this is not trivial:
  → accessory genes may or may not be present in a genome

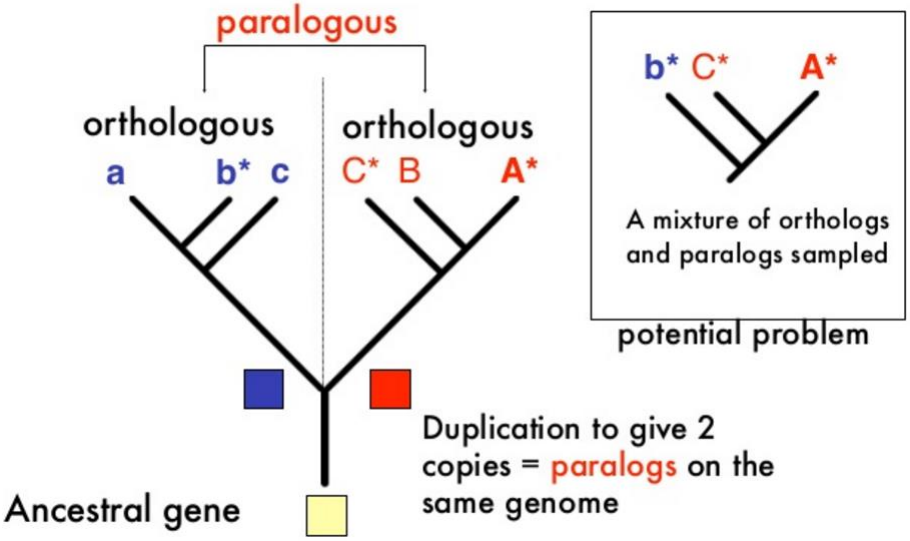
- WGS projects don't usually generate complete genomes → "Genome Assemblies"
  - A significant proportion of genome assemblies (30-40%) do not have complete data at all loci
  - Incomplete data generally due to "collision" between a locus and a gap in the assembly
    → incomplete (i.e. truncated) allele
  - *Mostly* randomly distributed



- Accessory genes are problematic because it is difficult to know if a gene is absent from the strain because it is accessory or just "missing" from the incomplete assembly

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 have multiple copies and 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 evaluation

- Duplicated genes present a problem because paralogs should not be compared, only orthologs; difficult to ascertain ortholog status

The challenges of scaling-up MLST (Part 5)

- 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

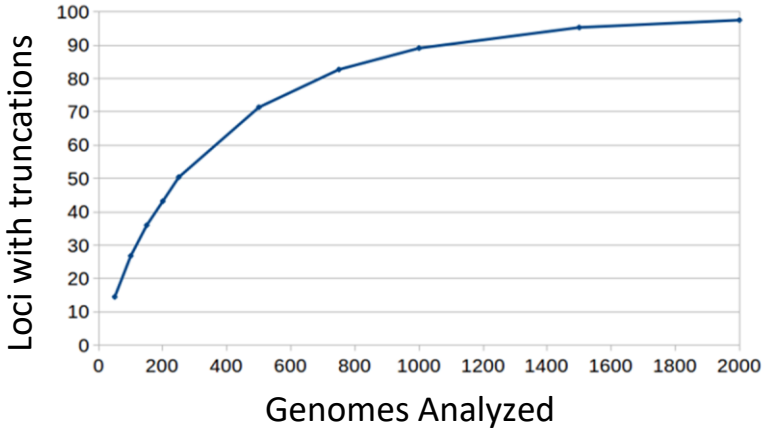
	1	2	3	4
1	100%	90%	83%	?
2	90%	100%	91%	?
3	83%	91%	100%	?
4	?	?	?	100%

- 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 know appropriate sequence similarity threshold

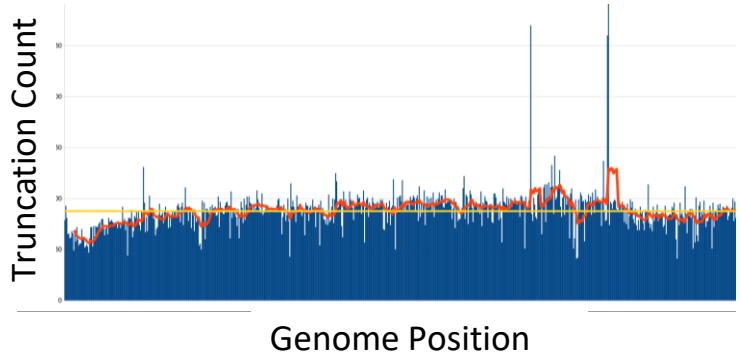
- Length and sequence variability poses a problem for gene identification via homology searching as database grows in size and strange alleles "pollute" your database

The challenges of scaling-up MLST (Part 6)

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



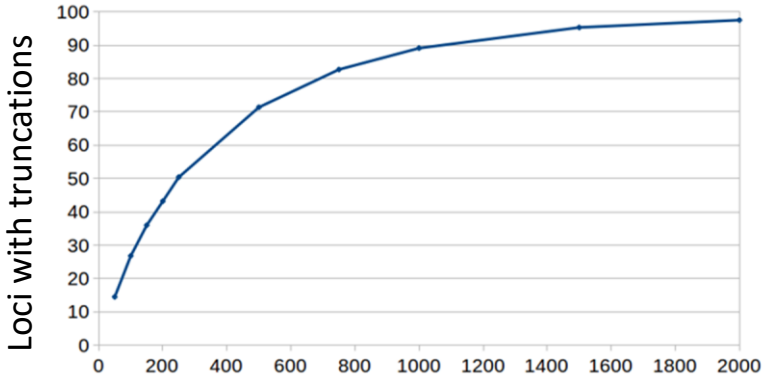
- On a large enough dataset, all genes will produce missing data for at least one genome
- A small proportion of genomes tends to be the worst offenders



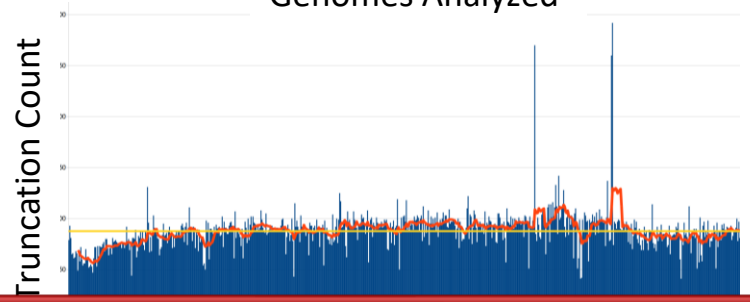
- Most loci have a background level of missing data due to assembly gaps
- Some loci are far more likely to reside in regions with assembly gaps

The challenges of scaling-up MLST (Part 6)

- wgMLST will require fishing out the loci out of the WGS data, this is not trivial:
  → Draft WGS data, variable completeness and quality of 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

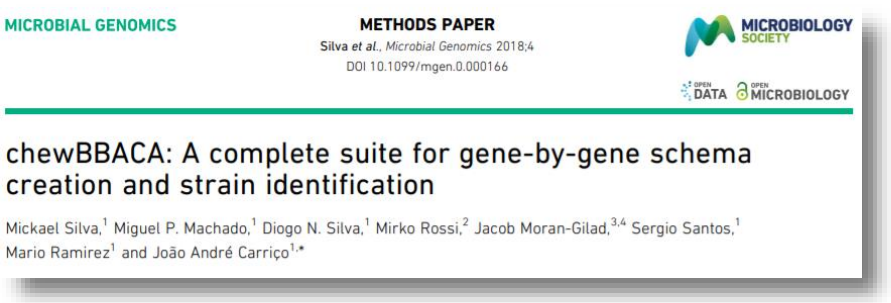
- Missing data is impossible to escape, especially as datasets get larger, given that most WGS data produced is of draft quality
- Must be willing to sacrifice certain genomes / loci for quality control

Core Genome MLST (cgMLST)

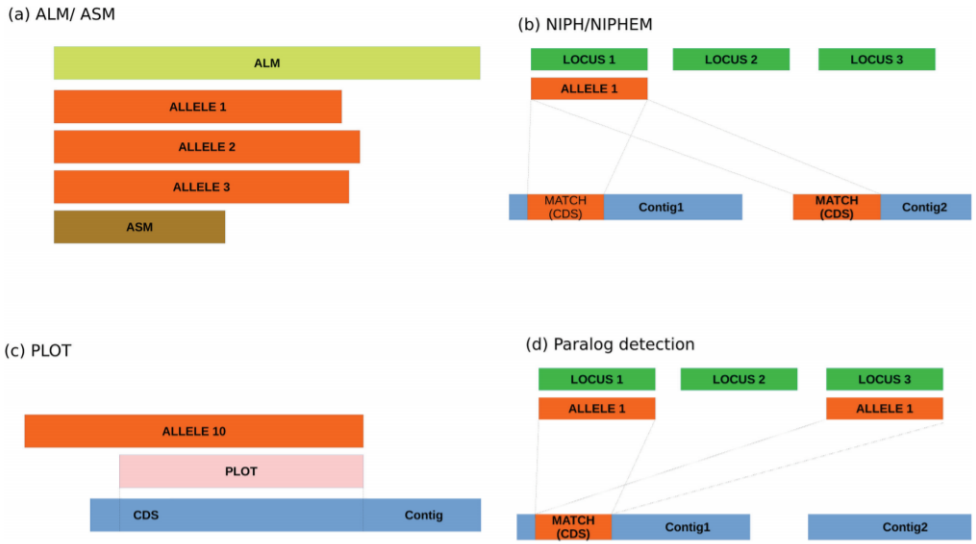
- Manual curation of alleles and STs is not going to be possible
- Assignment of novel alleles and STs will have to be done automatically and without significant manual supervision
- Because curation of a schema will be limited, significant effort should be placed up-front to ensure that it will scale 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 → don't have to wonder whether it's a missing accessory gene or an incomplete gene from 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 data suitable for outbreak response and longitudinal surveillance.

Designing a cgMLST schema (part 1)



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



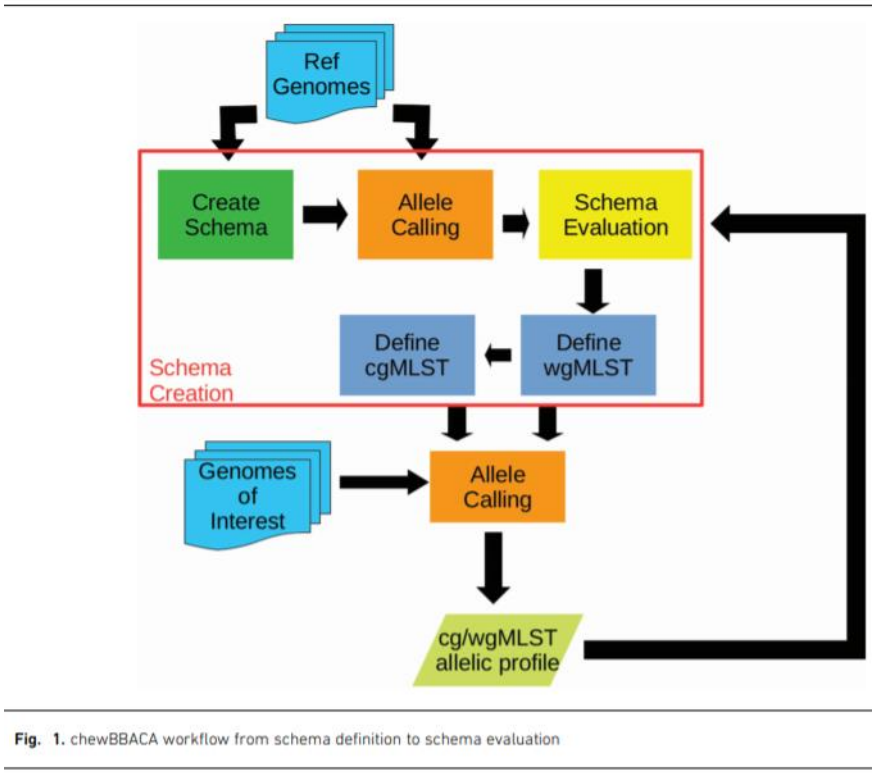
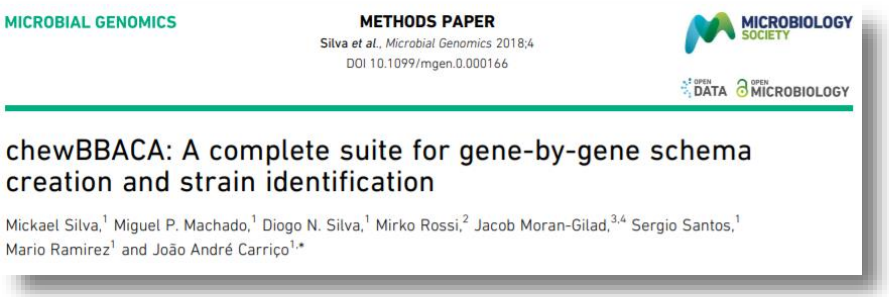


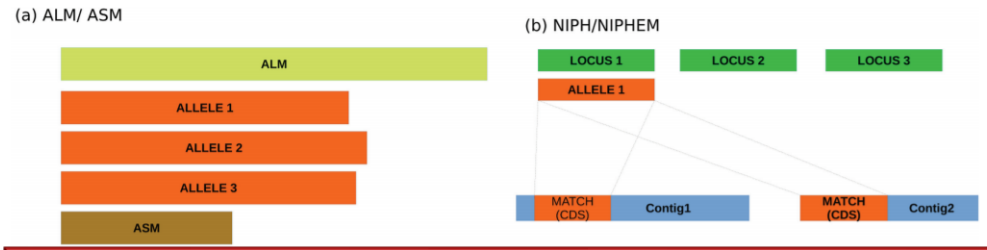
Fig. 1. chewBBACA workflow from schema definition to schema evaluation

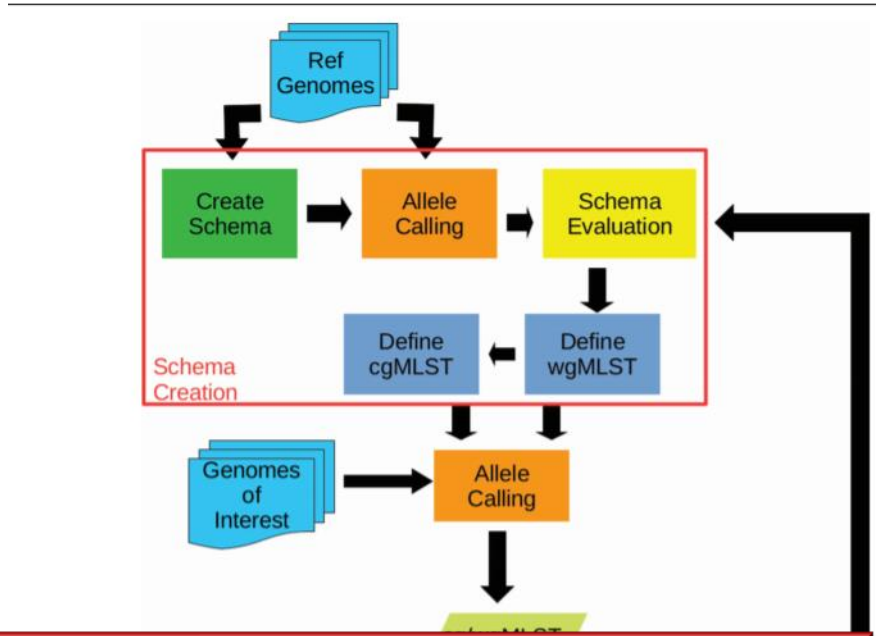
Exclusion of loci by size, duplicated genes, truncated genes and paralogous genes

Designing a cgMLST schema (part 1)



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





- If you remove core genes displaying large size/sequence variation and duplicated genes in the core, the remaining genes will tend to generate extremely good data
- Tools like chewBBACA can help standardize the development of cgMLST schema by applying a set of basic principles to ensure 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

http://combiboilersleeds.com/picaso/less-is-more/less-is-more-5.html

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

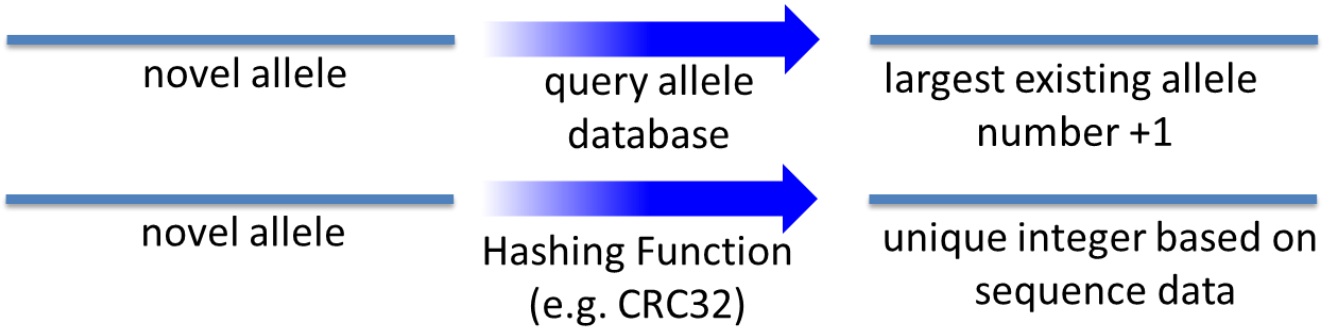
→ Species in repositories are often mislabelled: **Trust no one!**

http://combiboilersleeds.com/picaso/less-is-more/less-is-more-5.html

- Be ruthless!!! A gene found in 95% of genomes is not a core gene; drop some genomes, drop some genes or both
- Fewer high quality loci are better than many unreliable loci

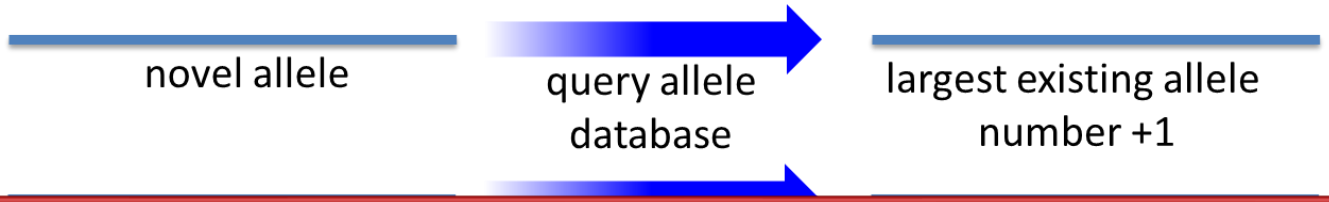
Hash MLST and database decentralization

- In a global community that is continually generating data, 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



Hash MLST and database decentralization

- In a global community that is continually generating data, 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

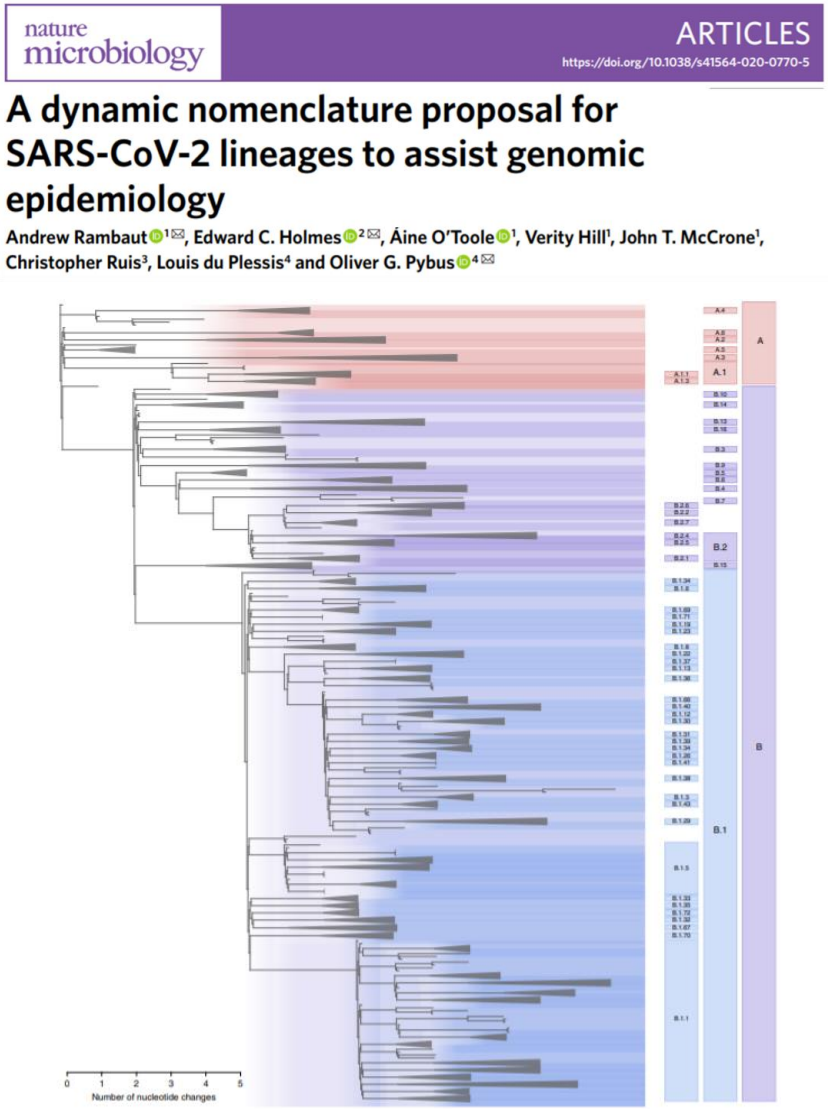


- Hashing of alleles and subtype information ensures that the same allele code and sequence type will be generated by anyone on earth

Nomenclatures for global genomic surveillance

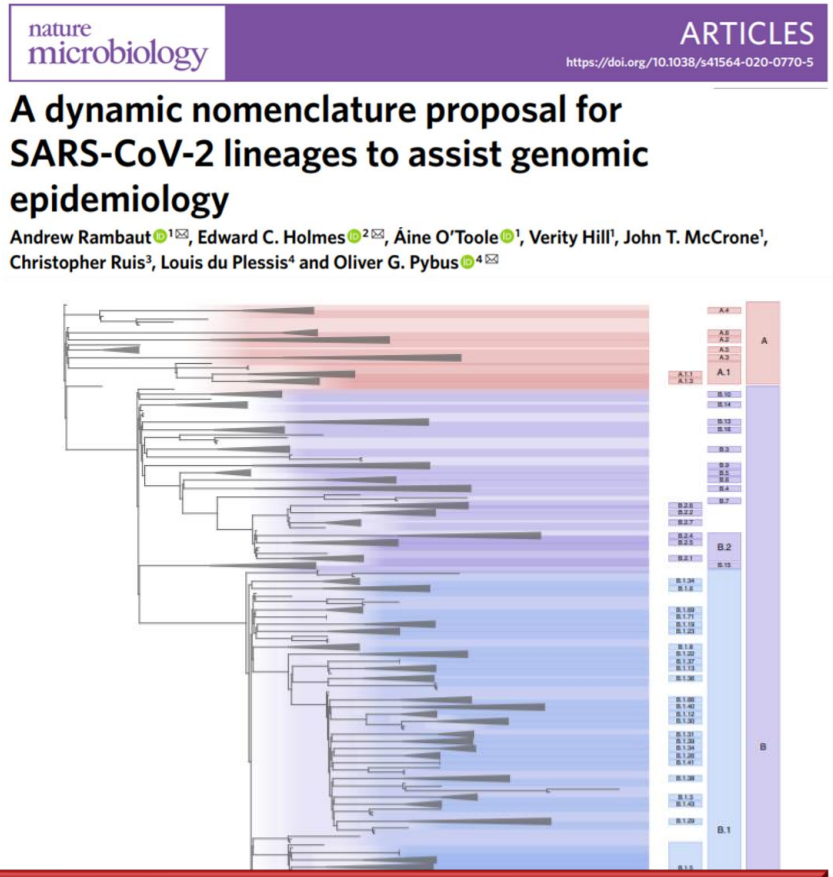
Nomenclatures and global surveillance

- Nomenclatures are used to systematically describe the various subtypes generated by a subtyping system
- A set of rules that can be implemented algorithmically in software that can automatically assign the linage of novel sequences with minimal curation needed



Nomenclatures and global surveillance

- Nomenclatures are used to systematically describe the various subtypes generated by a subtyping system
- A set of rules that can be implemented algorithmically in software that can automatically assign the linage of novel sequences with minimal curation needed



- Nomenclatures provide a means of efficiently communicating subtype information and facilitating tracking and monitoring of subtypes of interest.

Nomenclatures and global 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

Nomenclatures and global 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

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

The home stretch…

To MLST or to SNV?

- Each approach presents its own analytical advantages:
  - SNPs:
    - highest analytical power (i.e. discriminatory power)
    - susceptible to the effects of horizontal transfer
    - more difficult to standardize across large number of strains, particularly if there is significant genetic variation
  - MLST:
    - easier to standardize large-scale analysis and outputs → **nomenclature**
    - less susceptible to effects of horizontal transfer
    - Less discriminatory power than SNV analysis

- SNVs are particularly useful for clonal organisms that do not show a lot of recombination; MLST is better for recombinogenic organisms
- MLST analysis scales up better for analysis of thousands of genomes

To MLST or to SNV?

- 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 further loss of discriminatory power.
- **Problem:** For highly monomorphic species, lineages, or sub-lineages, cgMLST may provide insufficient discriminatory power to differentiate strains that should be distinct.
- **Solution**: two primary options have been proposed to deal with such cases:
  1. Perform SNV-based analysis on the strains of concern
  2. Perform MLST analysis in which set of loci is expanded include accessory genes shared by the strains of concern → Shared Genome MLST (sgMLST)

Parting thoughts

- Whole-genome phylogenetic analysis can be viewed as the "gold standard" analysis of the pathogen of interest; however, such an analysis may neither be feasible nor desirable depending on the pathogen being studied
  - considerations on genome size; considerations on population structure and mutation vs. recombination (i.e. highly clonal vs. panmictic…); considerations for pangenome dynamics (core vs. accessory genome); availability of other WGS data (i.e. reference genomes)

- WGS-based typing is the process of cataloging, tabulating and quantifying genetic variability using a well-defined analytical process that emphasizes robustness and replicability of analysis in favour of sheer discriminatory power.
  - criteria for variant data inclusion/exclusion; methodology for computing genetic similarity; genetic profile clustering approach; lineage definitions & nomenclature
  - much of the emphasis on WGS-based typing is on what data to keep and which to "throw away" because it may be unreliable biologically, analytically, or both.
  - "One-off" analysis of a dataset *does not require* WGS-based typing; using a WGS-based typing approach can facilitate comparison to *other* datasets

Parting thoughts

- Two primary methods are currently being used for WGS-based subtyping in the context of global genomic surveillance: SNV-based and MLST-based variant analyses
  - SNV-based analysis; MLST-based analysis

- SNV-based analysis:
  - involves identifying single nucleotide variation; Involves mapping of sequencing reads against a closely related reference genome and extracting high quality SNV positions into an alignment that can be analyzed using traditional phylogenetic approaches
  - Pros: closest to full genome phylogenetics; does not require genome assembly; works well on highly similar genomes (e.g. highly clonal species; possible outbreak clusters); well-suited to the analysis of viral genomes, which are small
  - Cons: lack of availability of reference genomes for read mapping; possible to include SNVs that *should* be excluded from analysis; best suited to the analysis of 10s to 100s of genomes depending on computational resources

Parting thoughts

- MLST-based analysis:
  - involves an extension of the classical Multi-Locus Sequence Typing approach of Maiden et al. in which the analysis is extended 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 interspecific recombination and/or that have an epidemic population structure; scale well to analysis of large number of genomes (1000s of genomes); until the current pandemic had been more successfully adapted to global genomic surveillance
  - Cons: generally unsuitable for highly clonal organisms because of insufficient discriminatory power; requires availability of an MLST schema

- WGS-based subtyping analysis works best as a hybrid approach where you can use Core Genome MLST (cgMLST) at the population-level that can be complemented with SNV-analysis to zoom in on specific clusters

After the break…

- Analysis exercise with Dillon Barker
  - Interpreting genome-scale MLST results
  - Comparing allelic typing methods
  - Visualizing phylogenetic trees using R-lang and ggtree
  - Cleaning noisy data
  - Investigating an outbreak detected via routine surveillance

We are on a Coffee Break & Networking Session

Workshop Sponsors:

Canadian Centre for Computational Genomics