Recueil et utilisation des métadonnées issues du séquençage

Sylvain Brisse

Institut Pasteur,
Biodiversity and Epidemiology of Bacterial Pathogens Unit,
National Reference Center for Whooping Cough
National Reference Center for Diphtheria

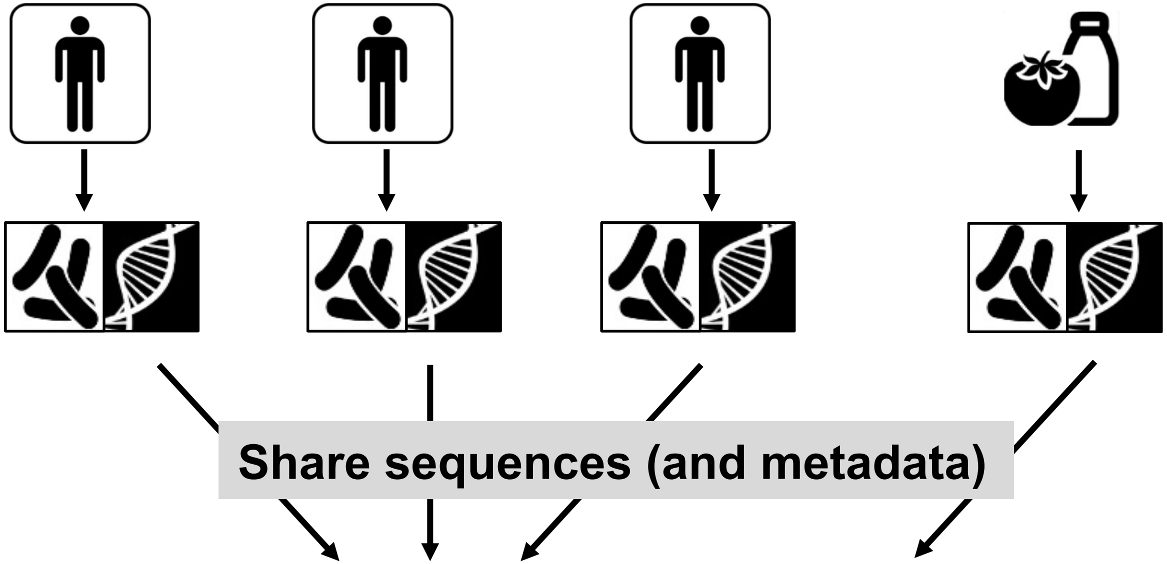
Paris, France

Institut Pasteur

La taxonomie génomique des souches, Ou comment assurer la surveillance sans échanger les données de séquençage génomique

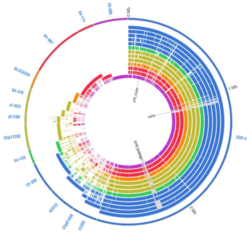
The sequence centralization model

Local



Share sequences (and metadata)

Central

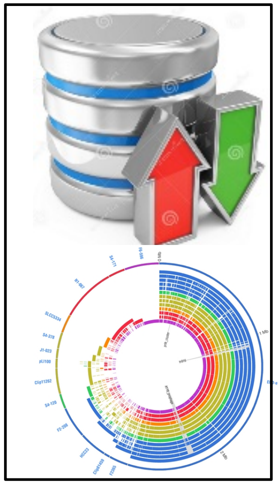


- Sequence comparison
- Subtype increase detection
- Alert
- *Optionally*, work with local labs

The sequence centralization model

Advantages:

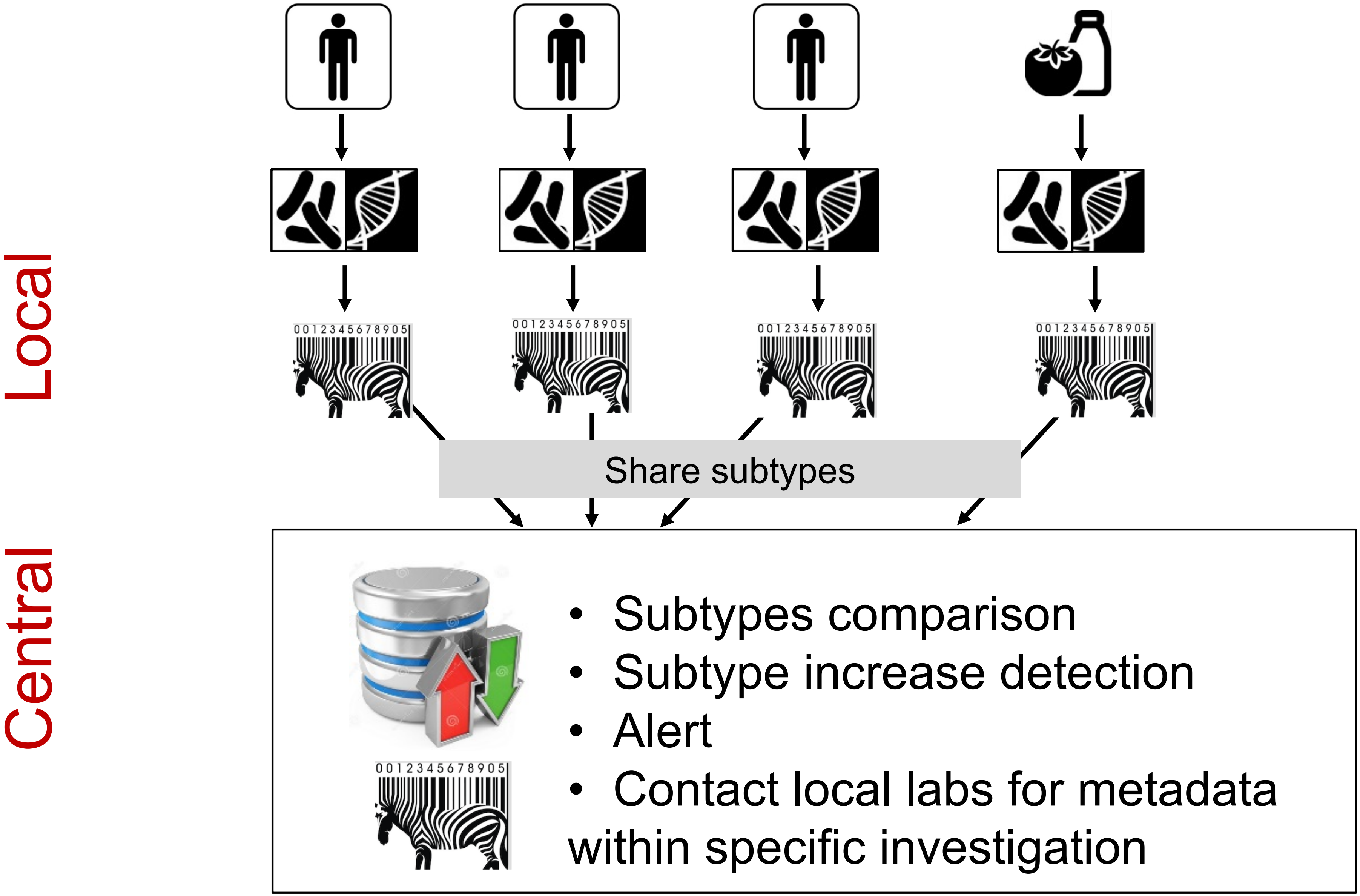
Flexibility of sequence data analysis



Challenges:

Data volume (Storage and transmission)
Data proprietary issues
Metadata confidentiality issues
Future use of data is hard to define and agree upon

The public nomenclature model



Local

Share subtypes

Central

- Subtypes comparison
- Subtype increase detection
- Alert
- Contact local labs for metadata within specific investigation

The public nomenclature model

Advantages:

- Share subtypes, not sequence nor metadata: solves multiple property & confidentiality issues

Challenges:

- Reproducibility of sequence-to-barcode algorithm
- Requires a public nomenclature server:
  - Quality of service?
  - Sustainability?

Public MLST databases

PubMLST

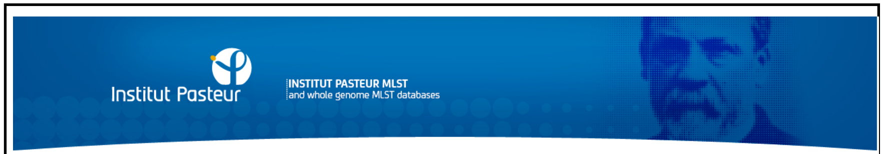
pubmlst.org

~ 100 pathogens, incl.:
Neisseria meningitidis
Campylobacter
Acinetobacter baumannii
…

EnteroBase

enterobase.warwick.ac.uk

~ 7 pathogens, incl.:
Salmonella
Escherichia/Shigella
Vibrio
Yersinia
…

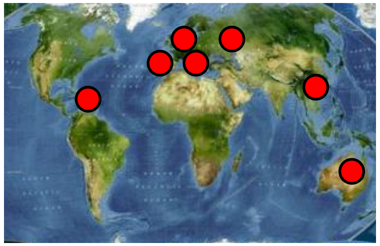


bigsdb.pasteur.fr

~ 10 pathogens, incl.:
Klebsiella pneumoniae
Listeria monocytogenes
Mycobacterium abscessus

http://bigsdb.pasteur.fr/klebsiella

- 4500 STs
- 9000 isolates
- 450 submitters



Molecular dissection of the evolution of carbapenem-resistant multilocus sequence type 258 *Klebsiella pneumoniae*

Tracking microevolution events among ST11 carbapenemase-producing hypervirulent *Klebsiella pneumoniae* outbreak strains

~20 years of service for more than 150 species of bacterial and fungal pathogens

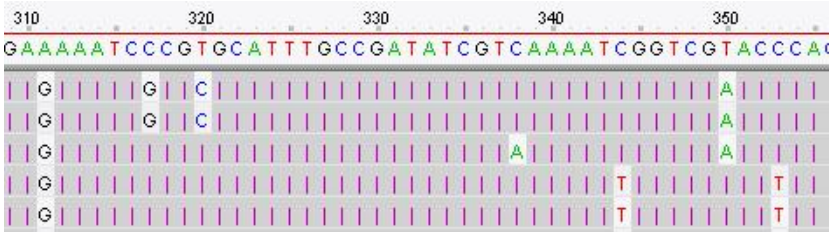
The Multilocus Sequence Typing standard

Bacteria
Fungi
Parasites

Maiden, Achtman, Spratt *et al.*

Extension to
Core genome:
Improves
discrimination

- Sequence of one gene (*rpoB*, 501 bp)
- Allele

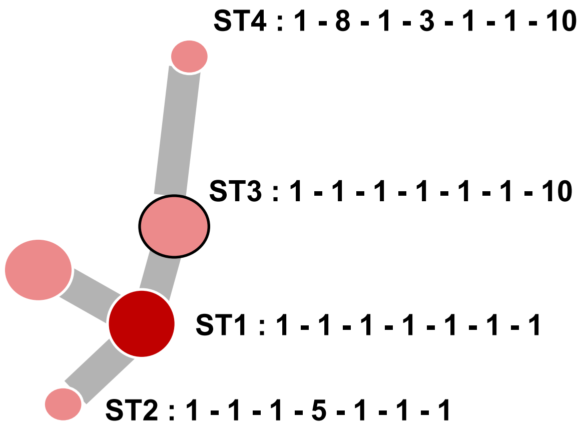


1
1
2
3
3

gapA	*infB*	*mdh*	*pgi*	*phoE*	*rpoB*	*tonB*
3	3	1	1	1	1	79

Sequence type (ST)	Alleles
1	4 4 1 1 7 4 10
11	3 3 1 1 1 1 4
258	3 3 1 1 1 1 79

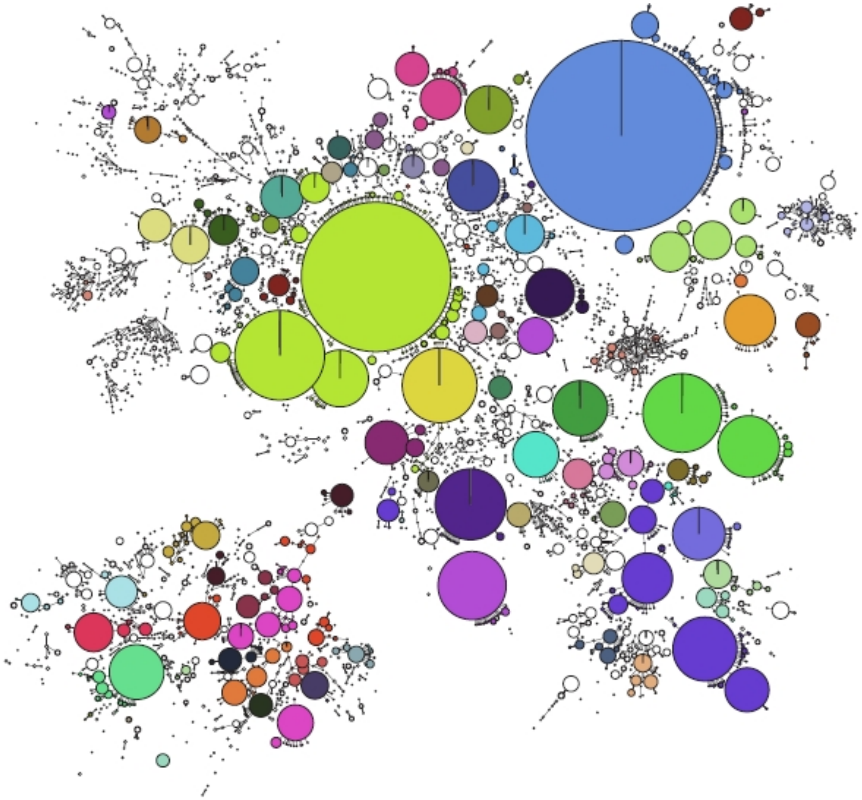
Bacterial population structure

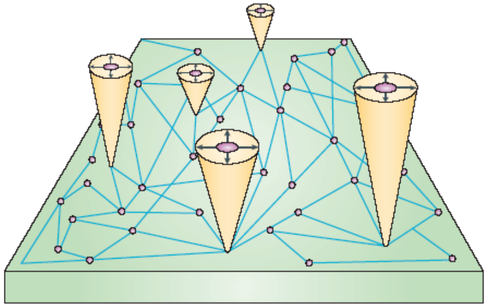


Single-linkage clustering

Classification:
Clonal complexes,
Clonal groups,…

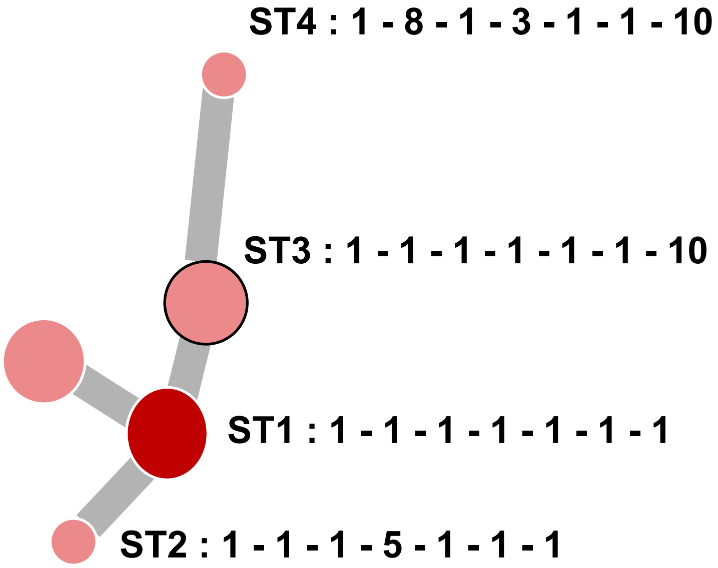
99,722 *Salmonella* (Alikhan et al. 2018)





MODEL:
Clonal expansion – diversification

MLST is human-interpretable



ST4 : 1 - 8 - 1 - 3 - 1 - 1 - 10

ST3 : 1 - 1 - 1 - 1 - 1 - 1 - 10

ST1 : 1 - 1 - 1 - 1 - 1 - 1 - 1

ST2 : 1 - 1 - 1 - 5 - 1 - 1 - 1

Santé publique France

Numbers of allelic differences

↕

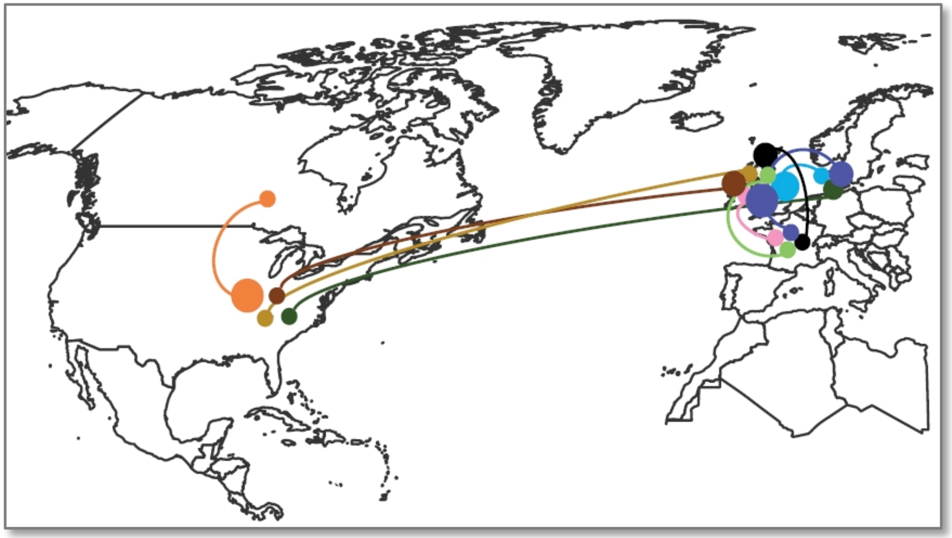
The time to last ancestor of isolates

Unified strain identifiers

From: MyInstitute@foodNetwork.org
Subject: type CT131 outbreak

Body: Are you observing this genotype in your country?

Detecting
outbreaks & emergence



Understanding
transmission & biology

Strain taxonomy: Why should Public Health agencies care?

- They would be the main end-users
- Can influence/define specifications
- Huge benefits expected in a distributed sequencing world
- Technically proven technology, low risk
- No one else cares

Models for multicentric genome-based surveillance

Sequence centralization

- Flexible analysis of data at hand
- Closed, tighlty controled centralized system
- Data transfer & storage
- Delegation of analysis

Public nomenclature

- Distributed, open system
- Local storage, data compression
- Joint collaborative data analysis
- Needs a robust, predefined barcoding algorithm