How to define “types” when using genome data

Eduardo Taboada
Head, Genomic Epidemiology Research Unit
Surveillance, Outbreak Detection & Response
Division of Enteric Diseases
National Microbiology Laboratory
eduardo.Taboada@canada.ca
The evolving role of genomics in subtyping

Number of *Salmonella* genomes

In one decade, NGS data has become the subtyping data.
WGS-based subtyping...
Using WGS in a molecular epidemiology context

We know how to do this:

Can’t we just do this?

- Of course we can.
- But... **should** we?
For many pathogens, there are far more “sporadic” cases than cases related to outbreaks.

Longitudinal surveillance to try and identify possible sources of infection ➔ long term strain tracking ➔ nomenclature
Using WGS in a molecular epidemiology context

We have some “constraints”

- have to transform WGS data into categorical data

- categories should harness the discriminatory power of WGS data and generate “meaningful” clusters

- categories should be stable over time to permit nomenclature development for strain tracking
Generating categorical data from WGS subtyping

Higher Similarity

# of clusters

31 12 8 4

Those who make many species are the 'splitters' and those who make few are the 'lumpers'...
— Chuck D (1857)
Generating categorical data from WGS subtyping

Higher Similarity

# of clusters

31 12 8 4

Those who make many species are the 'splitters' and those who make few are the 'lumpers'...
– Chuck D (1857)

- Categorical data generated through application of similarity threshold to generate clusters
What is the optimal similarity threshold?

Threshold has generally been set to maximize the formation of clusters that agree with outbreak data.

Those who make many species are the 'splitters' and those who make few are the 'lumpers'...

— Chuck D (1857)
The Sykes-Picot approach to cluster definition
Sykes-Picot Agreement: how to not draw borders
Sykes-Picot Agreement: how to **not** draw borders
The Sykes-Picot approach of arbitrarily drawing borders has led to a very unstable solution.
Sykes-Picot Agreement: how to **not** draw borders

More stable solution could have been found based on other factors: e.g. Topography & Demography
Optimization of WGS clusters

- 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; use this information to help guide the subtype definition
Evidence from population genetics
Optimization of WGS clusters

- A challenge with thresholding of WGS data: small adjustments can radically alter cluster membership

- Changes in cluster membership really mess up subtyping nomenclatures
The Adjusted Wallace Coefficient

Adjusted Wallace Coefficient as a Measure of Congruence between Typing Methods

Ana Severiano,1 Francisco R. Pinto,2 Mário Ramirez,1 and João A. Carriço1 *

Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal,1 and Centro de Química e Bioquímica, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal 2

\[
AW_{A \rightarrow B} = \frac{W_{A \rightarrow B} - W_{i(A \rightarrow B)}}{1 - W_{i(A \rightarrow B)}}
\]

The AWC is used to assess the concordance of cluster partitioning generated by two different subtyping methods.
The nAWC is used to compute the partition concordance generated by adjacent distance thresholds.
We can use the nAWC to identify areas where successive thresholds yield highly stable clusters for nomenclature development.
The synthesis

- Multilevel nomenclature:
  - Level 1: more ‘ephemeral’ → outbreak-centric?
  - Level 2: maximum discriminatory power but high cluster stability
  - Level 3: reflecting lineages with of historical importance to surveillance
Thresholds, clusters, and cluster-exclusive alleles

- We can compute the extent of Allele Exclusivity for a cluster based on within and outside cluster frequencies across all loci.

![Diagram showing cluster 37, threshold = 125 with number of genes and allele exclusivity distribution.]

- More exclusive allele
- Less exclusive allele
Thresholds, clusters, and cluster-exclusive alleles

- And we can compute this for all clusters at a given threshold...
Thresholds, clusters, and cluster-exclusive alleles

- And we can compare this for multiple thresholds...

- As thresholds are adjusted, resulting clusters will have different proportion of exclusive alleles

- Allelic exclusivity can be used to support subtype definitions
Evidence from ecology & epidemiology
Measuring epidemiological concordance

- We wished to develop an approach that could be used to rapidly assess the epidemiologic similarity of sets of bacterial isolates using a systematic and computationally tractable approach.

- These epidemiologic comparisons can then be used to estimate the overall “epidemiologic relevance” of subtyping clusters.
Building a model for epidemiological similarity

“Essentially, all models are wrong, but some are useful.”
George E.P. Box
(1919-2013)
Quantifying epidemiologic relationships

Our proposed approach:
A model for quantifying the epidemiological similarity between bacterial isolates based on three primary factors: source, space, time
Quantifying epidemiologic relationships

The epidemiological address ($\varepsilon$) of an isolate is defined by the vector:

$$
\varepsilon = (g, t, s)
$$

- $g$ = geospatial component
- $t$ = temporal component
- $s$ = (isolation) source component

The epidemiological distance ($\Delta \varepsilon$) between two isolates is defined by:

$$
\Delta \varepsilon = \sqrt{\gamma (\Delta g)^2 + \tau (\Delta t)^2 + \sigma (\Delta s)^2}
$$

- $\gamma$ = geospatial coefficient
- $\tau$ = temporal coefficient
- $\sigma$ = source coefficient
Combining the various components...

\[ \Delta \varepsilon = \sqrt{\gamma((\log\{dist_{ab}\})^2) + \tau \left( \log \left( \sum_{i=1}^{n} (x_i - y_i)^2 \right) \right)^2} + \sigma \left( 1 - \frac{1}{n} \sum_{i=1}^{n} f(u_i, v_i) \right)^2 \]

Geospatial \hspace{2cm} Temporal \hspace{2cm} Source

Epidemiological Similarity between two isolates = 1 - \Delta \varepsilon
Epidemiological clustering of *C. jejuni* isolates
Epidemiological clustering of *C. jejuni* isolates

- Major clusters based on source component
- Subclusters further refined by spatial and temporal component
Assessing epidemiologic relevance of clusters

Epidemiologic cluster cohesion: (ECC)

A measure of the overall epidemiologic homogeneity of isolates within a subtyping cluster

High ECC:
Isolates share highly similar epidemiologic profiles.
e.g. same source, sampling time and geography

Low ECC:
Isolates with moderately related epidemiologic profiles.
e.g. different sampling sources, dates and locations.
Comparing epidemiologic vs. genomic signal

By comparing genomic vs. epi concordance, we can identify clusters that group together significantly stronger via genomic or epidemiologic signal.
Concordance: epidemiological vs. genomic signal

- High concordance
  - Stronger epi similarity
  - Stronger genomic similarity
Concordance: epidemiological vs. genomic signal

- High concordance
- Stronger epi similarity
- Stronger genomic similarity

"Environmental Sink"

"Generalist Genotype"

Source: Julie Arsenault (PhD Thesis): papyrus.bib.umontreal.ca/jspui/handle/1866/4625

© ESCMID eLibrary by author
Epidemiological vs. Genomic discordance

- Genomically heterogeneous strains with high ECC tend to come from environmental sinks.
- Genomically homogeneous strains with low ECC tend to be generalist strains that are widely dispersed.

- High concordance
  - Stronger epi similarity
  - Stronger genomic similarity
Parting thoughts

1. Defining subtypes in the era of genomics-based subtyping is a hard problem
   • minor adjustments in distance thresholds for cluster definition can have major impacts on cluster membership

2. There are various lines of evidence that can be used to guide the process:
   • population structure
   • genetics
   • ecology & epidemiology

3. All available evidence suggests that not all species/lineages/sublineages are equally genetically or epidemiologically cohesive, so single thresholds are unlikely to be adequate
Acknowledgements

- **PHAC-NML**
  - Gary Van Domselaar and bioinformatics core team
  - Morag Graham and genomics core team
  - Division of Enteric Diseases

- **PHAC-CFEZID**
  - FoodNet Canada team
  - CIPARS team

- **Agriculture and Agri-Food Canada**
  - Doug Inglis and team

- **Canadian Food Inspection Agency**
  - Lisa Waddington, Cathy Carrillo, Daniel Leclair

- **INNUENDEO team**
  - Mirko Rossi & Federica Palma
  - João Carriço and team