



Journal Club

Comparing cgMLST and SNP phylogenies



ELSEVIER

Contents lists available at [ScienceDirect](#)

International Journal of Food Microbiology

journal homepage: www.elsevier.com/locate/ijfoodmicro

Comparative analysis of core genome MLST and SNP typing within a European *Salmonella* serovar Enteritidis outbreak

Madison E. Pearce^{a,b,*}, Nabil-Fareed Alikhan^c, Timothy J. Dallman^d, Zhemin Zhou^c, Kathie Grant^d, Martin C.J. Maiden^{a,b}

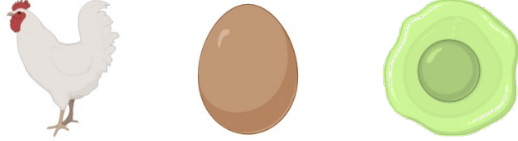
^a Department of Zoology, University of Oxford, Peter Medawar Building for Pathogen Research, South Parks Road, Oxford OX1 3SY, United Kingdom

^b National Institute for Health Research, Health Protection Research Unit, Gastrointestinal Infections, University of Oxford, United Kingdom

^c Warwick Medical School, University of Warwick, Coventry CV4 7AL, United Kingdom

^d Public Health England, Gastrointestinal Bacteria Reference Unit, 61 Colindale Avenue, London NW9 5EQ, United Kingdom

The *Salmonella* genus – a major threat to human health

- Causes >80 mio cases of foodborne gastroenteritis annually
- *Salmonella enterica* serovar Enteritidis accounts for 40-60% of global human *Salmonella* infection cases
- Principal human infection source:

- Food is increasingly distributed across borders, making multi-national outbreaks more common
- Large European outbreak of *Salmonella enterica* serovar Enteritidis in May-Sep 2014
 - > 350 cases in UK, Germany, France, Austria, Luxembourg
 - Caused by contaminated eggs

Public health challenges in **detection**, **tracking**, and **notification** as most disease surveillance occurs at national level

Need for uniform characterisation of disease-associated isolates:
high-resolution, **accessible**, and **replicable** isolate typing schemes



<https://enterobase.warwick.ac.uk/species/index/senterica>

Nucleotide sequence based typing methods – an alternative to serotyping

- Serology based typing method lacks the resolution necessary for outbreak detection and characterisation
- Arrival of high throughput sequencing technologies (HTS) lead to affordable and practical whole genome sequence analyses (WGS)
 - ➔ Highest resolution: Identifying **single nucleotide polymorphisms (SNPs)**
Comparing sequence data to a reference genome and recording varying nucleotides
- Capable of revealing evolutionary histories of homogenous groups, detecting and tracing outbreaks such as *e.g.* the European *Salmonella enteritidis* outbreak in 2014
- Due to inherent inaccuracies in HTS, quality must be assured, *e.g.* minimum coverage, distances allowed between SNPs for accuracy and consistency
- Difficulties in standardisation within and among labs and establishment of consistent nomenclature (Reference genomes used can differ between labs)

Alternative to SNP analysis – core genome MLST (cgMLST)

- Multilocus sequence typing (MLST) is **scalable** in resolution and number of isolates
7-locus MLST < ribosomal MLST (53 loci) < core genome MLST (3002* loci) < whole genome MLST (3258* loci)
- Sequence type (ST) and allele nomenclatures are **internationally available** and readily **standardised** using schemes hosted on **web-based servers** (e.g., Enterobase)
- To distinguish between closely related isolates within an outbreak, a larger number of genes need to be included within a given scheme
- cgMLST balances the number of loci used in a scheme with max. possible resolution by including loci present in majority of isolates (95-99%) in a given grouping of bacteria
- Ideally these genes do not change presence over time
➔ Elements not under strict selection pressures – e.g., repetitive genes and pseudogenes – should be excluded

Congruency between SNPs and cgMLST has been shown previously

cgMLST can form basis of a **stable, reference free, internationally curated** nomenclature scheme accessed via databases allowing **global epidemiology** and other analyses

* As defined in this paper for Salmonella

Aims of this study:

By reanalysing the European *Salmonella enterica* serovar Enteritidis outbreak in 2014:

1. **Validate** the application of **cgMLST** for the **characterisation of international outbreaks**
2. **Comparison** of the results previously obtained by **SNP** analyses

Why was this outbreak database chosen?

- spanned several countries
- occurred over several months
- consisted of three distinct serovar Enteritidis clades associated with primary production
- sub clustering of point source outbreaks

- **Simpson's diversity index**
 - **Identifies** statistically significant **differences between counts of unique profiles** generated by different typing methodologies
 - It has a value between 0 and 1, depending on the number of partitions created by a typing method

- **Adjusted Wallace coefficient***
 - **Quantitative** measure of **congruence**
 - Calculates the statistical **significance of similarities** between partitions generated by different typing methods
 - It has a value between 0 and 1, depending on the ability of a typing method to further subdivide others and accounts for 95% confidence intervals

* For more details check: Severiano A, Pinto FR, Ramirez M, Carriço JA. Adjusted Wallace coefficient as a measure of congruence between typing methods. J Clin Microbiol. 2011 Nov;49(11):3997-4000. doi: 10.1128/JCM.00624-11. Epub 2011 Sep 14. PMID: 21918028; PMCID: PMC3209087.

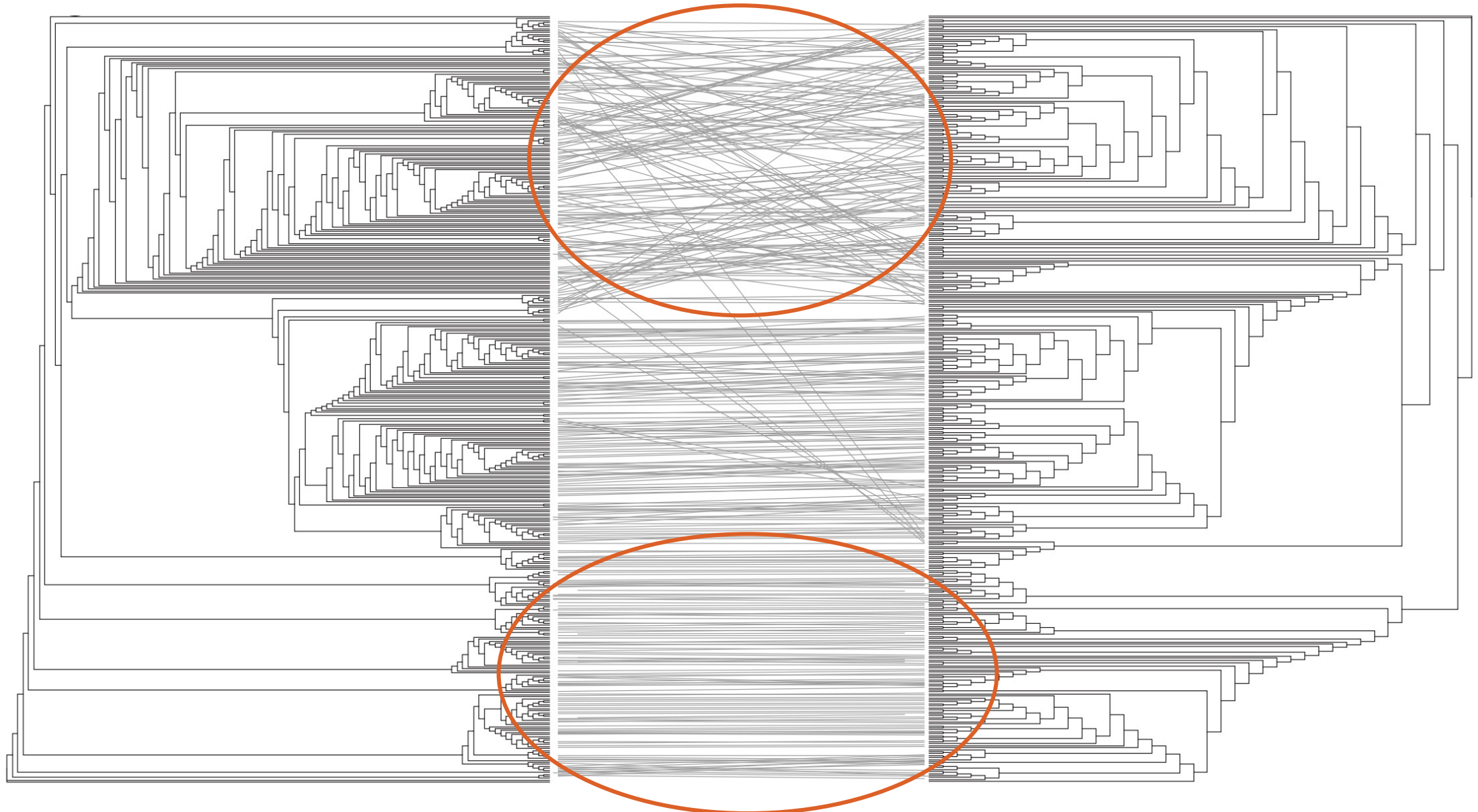
Comparison of SNPs and cgMLST

- **SNP** addresses provided **greater resolution** than cgMLST single linkage clusters
 - # of unique sequence types (SNP vs. cgMLST): 249/527 vs. 229/527
 - Simpson's diversity index (SNP vs. cgMLST): 0.949 vs. 0.901
 - Adjusted Wallace coefficient = 0.874 ; 95% CI: 0.808–0.942; $P < 0.001$
- **cgMLST** will **not include intergenic regions** (unlike SNP analysis) and only one allelic change will be counted if there are multiple SNPs within the same gene
- However, **short insertions or deletions** in the core genes – ignored in many SNP analyses – were **captured** by cgMLST
- A **Tanglegram** was plotted for a **visual** comparison:

cgMLST Tree

Tanglegram

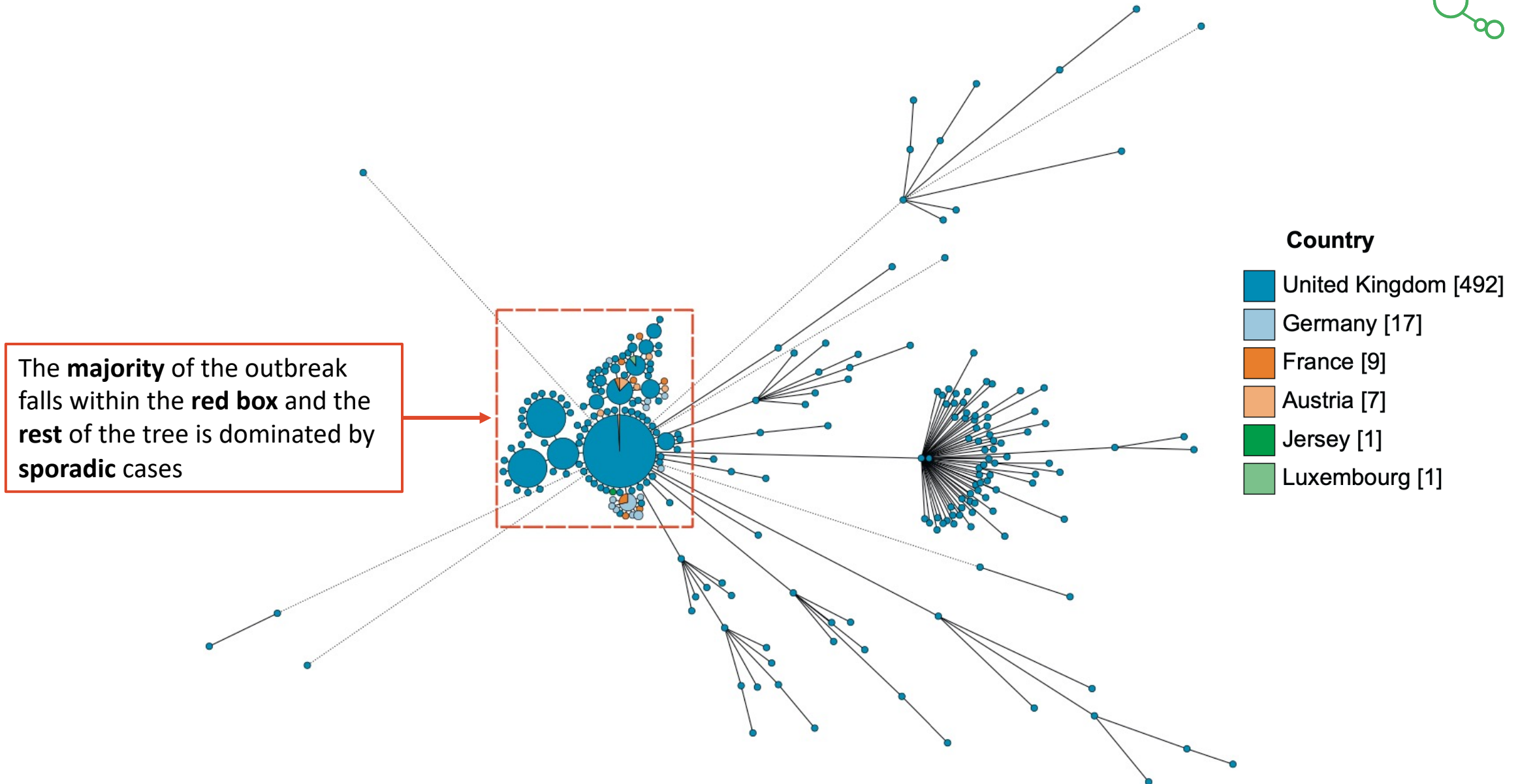
SNP Tree



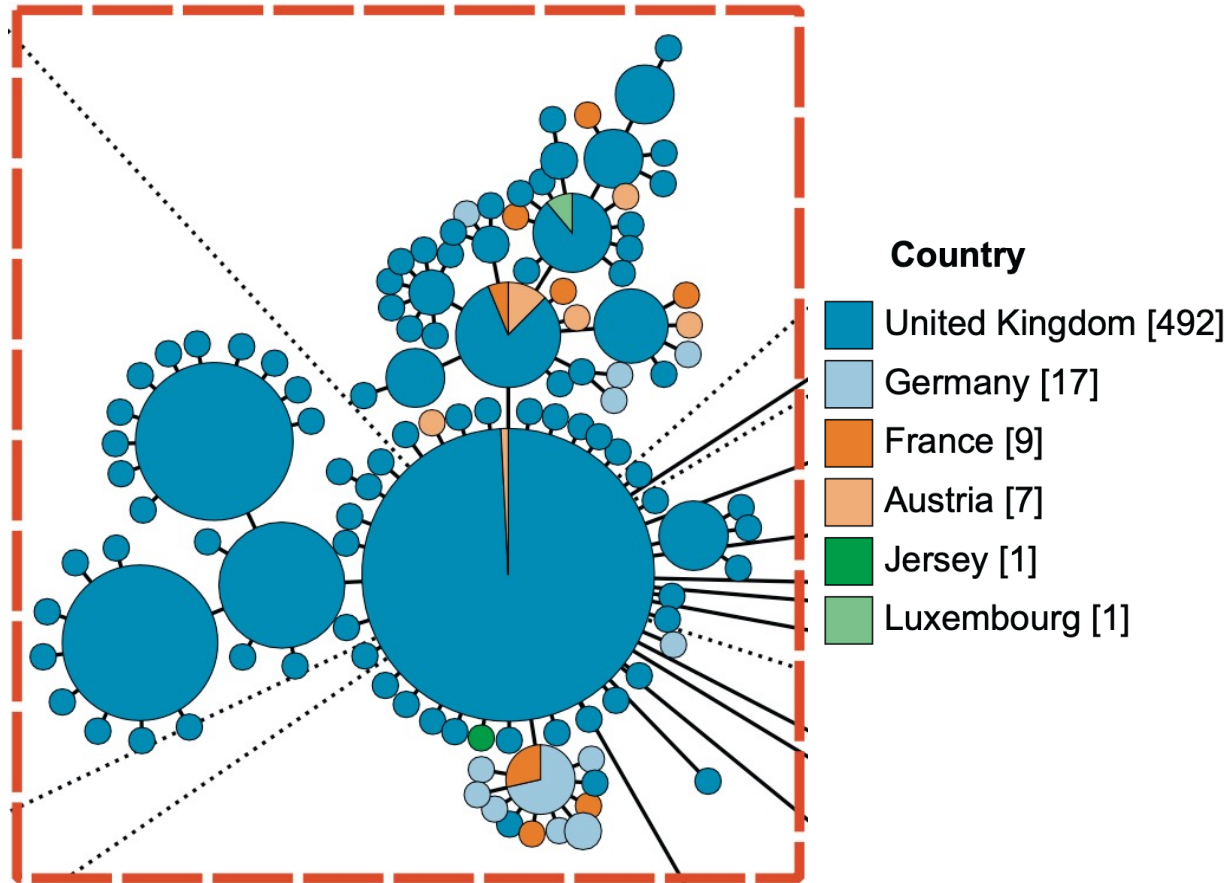
Comparison of SNPs and cgMLST

- **SNP** addresses provided **greater resolution** than cgMLST single linkage clusters
 - # of unique sequence types (SNP vs. cgMLST): 249/527 vs. 229/527
 - Simpson's diversity index (SNP vs. cgMLST): 0.949 vs. 0.901
 - Adjusted Wallace coefficient = 0.874 ; 95% CI: 0.808–0.942; $P < 0.001$
- **cgMLST** will **not include intergenic regions** (unlike SNP analysis) and only one allelic change will be counted if there are multiple SNPs within the same gene
- However, **short insertions or deletions** in the core genes – ignored in many SNP analyses – were **captured** by cgMLST
- A **Tanglegram** was plotted for a **visual** comparison:
 - Good congruence shown between cgMLST and SNP
 - Majority of isolates were grouped into the same clusters
 - Minor differences, predominantly caused by inversions of clusters due to differences between the internal nodes, located deeper within the phylogenies

Core genome minimum spanning tree – by **country** of origin



Core genome minimum spanning tree – by **country** of origin



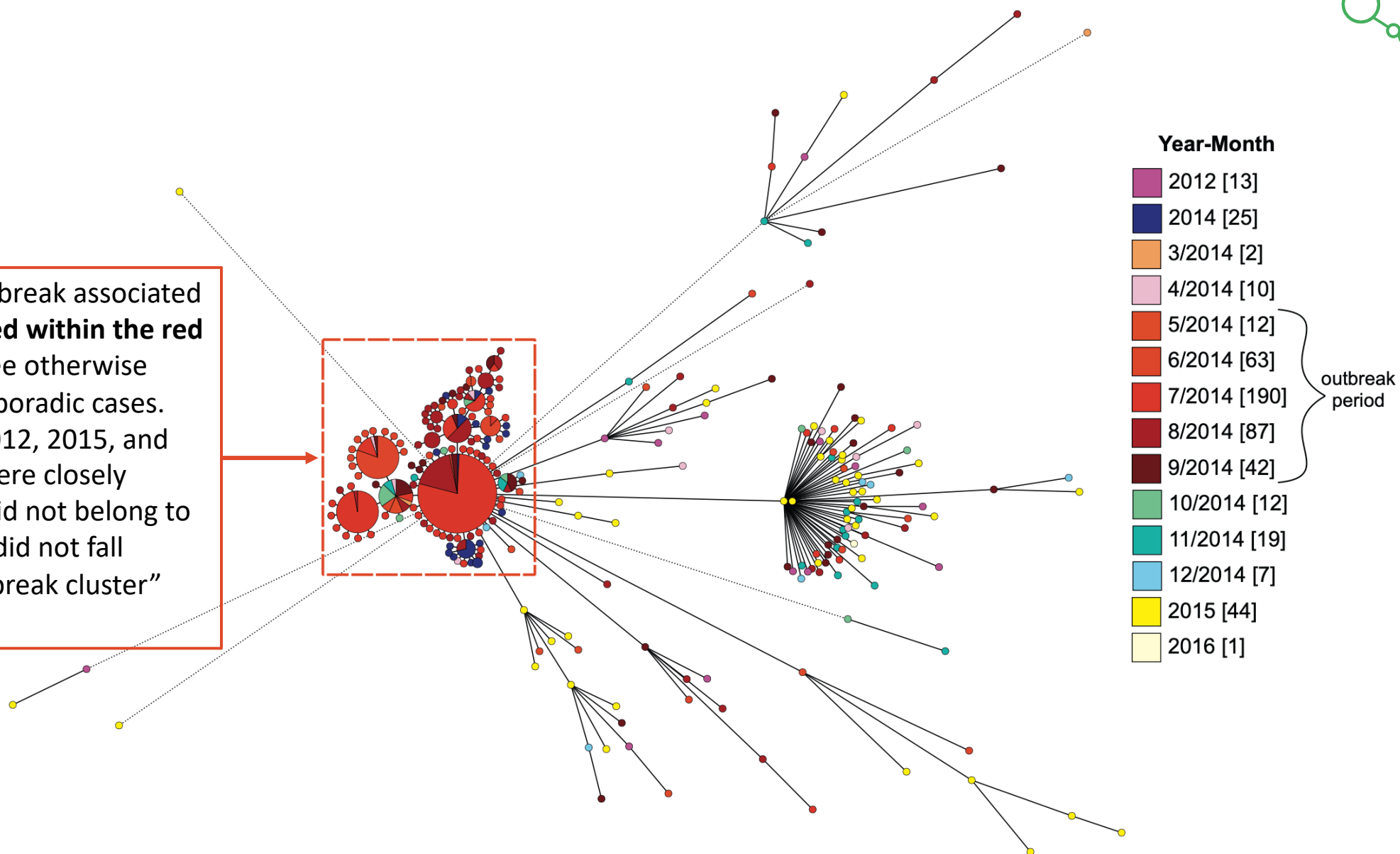
This tree indicates/suggests:

1. There is **no relationship** between the **country** of isolation and the **cgMLST** cluster
2. The **diversity** predated the outbreak

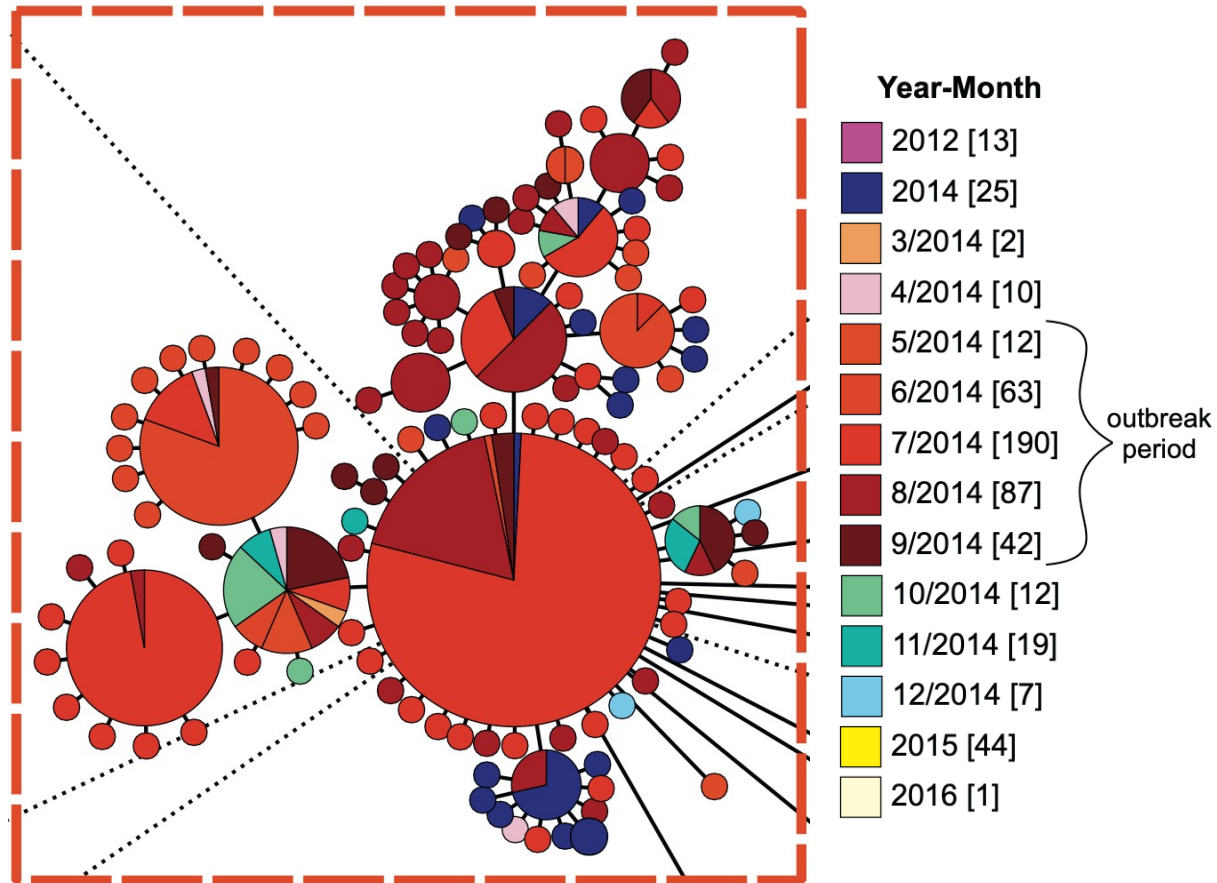
Note: There is no scale for the branch lengths – We don't know the exact allelic distance between the isolates!

Core genome minimum spanning tree – by time of collection

Most of the outbreak associated isolates **clustered within the red box** with the tree otherwise dominated by sporadic cases. Isolates from 2012, 2015, and 2016 – which were closely related to but did not belong to the outbreak – did not fall within the "outbreak cluster" and are diverse.



Relationship between year of infection and genetic diversity



- Most outbreak-associated cases occurred from May to September 2014
- Individual cases persisted until December
- Clusters contain isolates from only one or two months and when a cluster consisted of multiple months they were generally consecutive

This tree indicates/suggests:

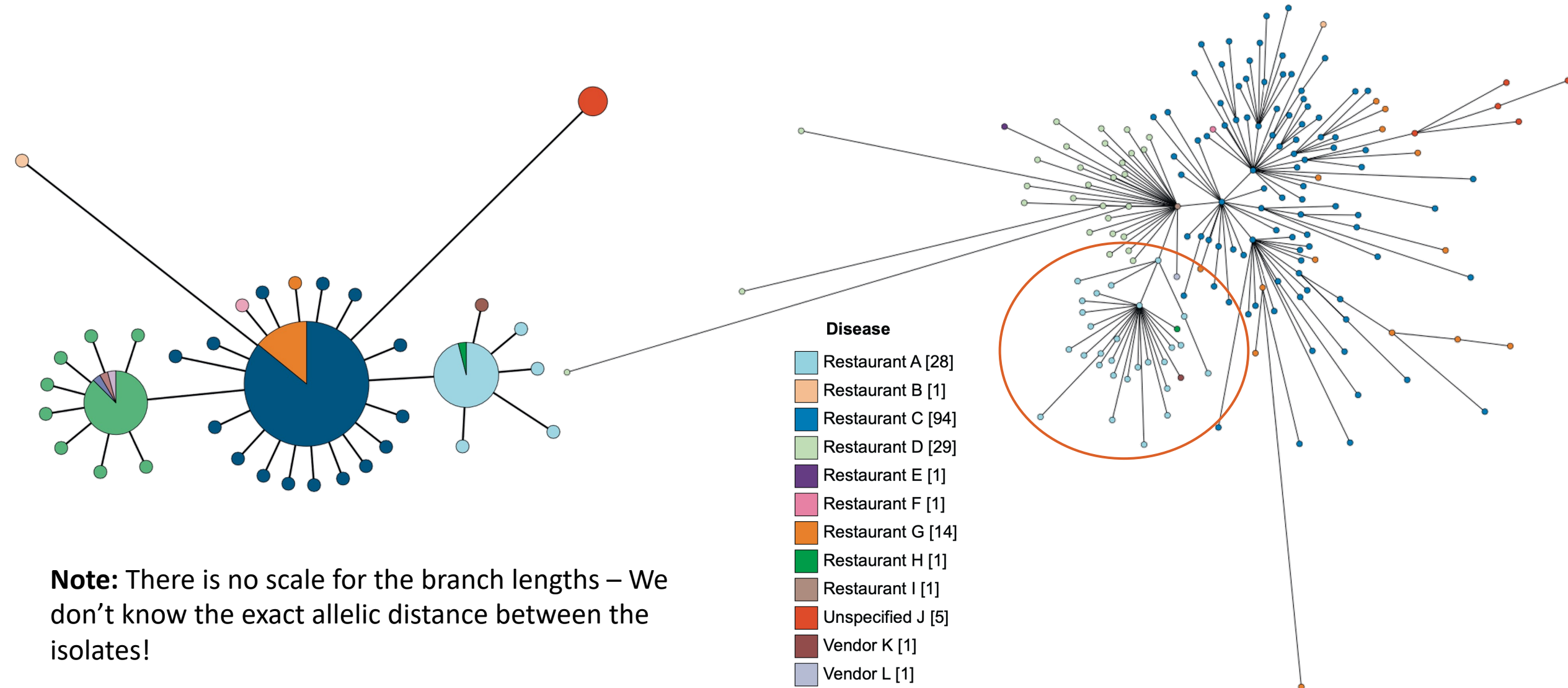
1. Isolates from 2012, 2015, 2016 are more genetically distant from the outbreak group and each other
2. Various closely related subtypes of serovar Enteritidis replaced each other over the course of the outbreak

Note: There is no scale for the branch lengths – We don't know the exact allelic distance between the isolates!

Comparison of cgMLST and wgMLST: food traceback

- **wgMLST** provided **additional resolution** compared to cgMLST due to larger number of distinct profiles
 - # of unique sequence types (wgMLST vs. cgMLST): 177/177 vs. 137/177
 - Simpson's diversity index (wgMLST vs. cgMLST): 1.000 vs. 0.981
- However, there was **no** statistically significant **difference** in the **discriminatory ability**
- Results were **consistent** with **mappings** of available **food traceback** isolates onto minimal spanning trees:

Core (left) vs. Whole (right) genome minimum spanning tree

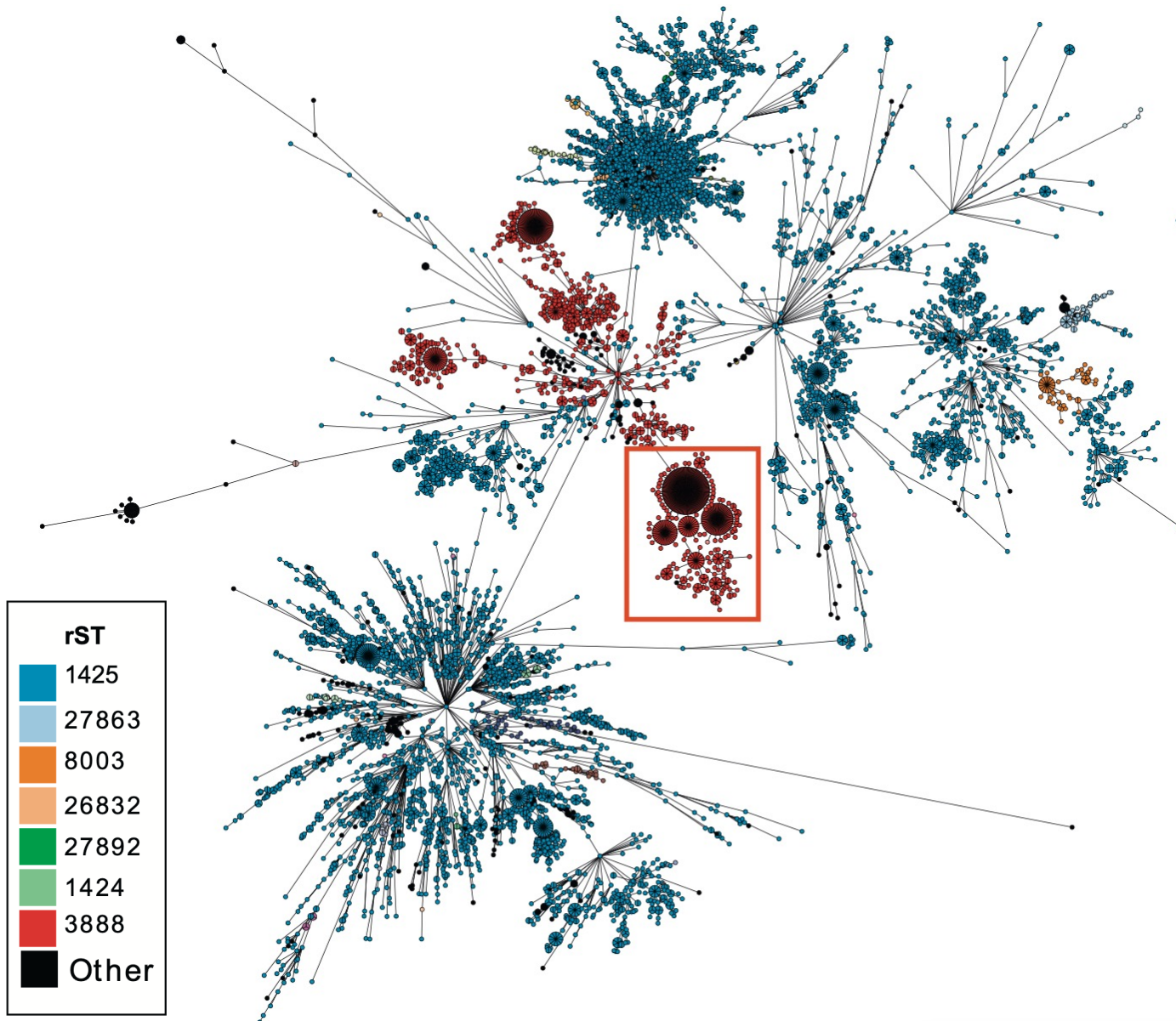


Note: There is no scale for the branch lengths – We don't know the exact allelic distance between the isolates!

Comparison of cgMLST and wgMLST: food traceback

- **wgMLST** provided **additional resolution** compared to cgMLST due to larger number of distinct profiles
 - # of unique sequence types (wgMLST vs. cgMLST): 177/177 vs. 137/177
 - Simpson's diversity index (SNP vs. cgMLST): 1.000 vs. 0.981
- However, there was **no** statistically significant **difference** in the **discriminatory ability**
- Results were **consistent** with **mappings** of available **food traceback** isolates onto minimal spanning trees:
 - Both approaches grouped the isolates into identical clusters (restaurant A + H + vendor K; restaurant D + E + I + vendor L; restaurant B + C + F + G + unspecified J)
 - Neither approach distinguished isolates on geographical source
 - ➔ There was **no relationship** between **place of isolation** and **genetic clustering** of isolates
 - There was **no relationship** between **clusters** and **wholesalers** supplying the sources
 - ➔ Any **diversity** within isolates was **generated at the source** before the outbreak began

Placing the outbreak within the rest of Enteritidis using cgMLST



- All serovar Enteritidis isolates available on EnteroBase (9380) were used to create a minimum spanning tree
 - Node size corresponds to a single sequence type profile based on cgMLST
 - Node colour codes for sequence types defined by rMLST
- ➔ All outbreak isolates belong to one rST
- ➔ Outbreak formed its own cluster, despite increased number of Enteritidis isolates

cgMLST schemes can place an outbreak into a wider context, as they use a predefined set of loci!

Conclusions

- cgMLST has sufficient resolution to:
 - Detect a multi-country disease outbreak caused by very closely related strains
 - Identify substructure within isolates obtained during the outbreak
- cgMLST analyses were congruent with wgMLST & SNP analyses
- Advantages of cgMLST:
 - Can be readily and consistently applied in different labs and jurisdictions as it uses a consistent set of conserved loci
 - Analyses are replicable and forward and backward compatible
 - Rapid location of outbreak isolates into the context of the known diversity of serovar Enteritidis, as the core genome is common to all members of the species
 - Availability of web-based analyses platforms enables these high-resolution analyses to be conducted with minimal requirements for locally installed bioinformatics infrastructure

Improvements to be made

- cgMLST types are too discriminatory and currently include missing alleles
- Implement a formal, standardised clustering system to enable sorting of cgMLST profiles into closely related groups, such as clonal complexes
- Enable easy communication of an outbreak between labs by e.g. creating “MLST addresses” (similar to SNP addresses) based on single linkage clustering to form a hierarchy of relatedness
- In the Figures showing the various trees, there was no scale presented, which shows the allelic differences between the isolates
- There is no discussion on allele distance cut-offs.
 - ➔ At which allelic distance do we decide to add an isolate to one cluster or the other?



Thanks for your attention!

Any questions?



University of
Zurich^{UZH}

Institute of Medical Microbiology



Have a nice weekend!

