Public Health England

Whole genome sequencing of foodborne pathogens: experiences from the Reference Laboratory

Kathie Grant
Gastrointestinal Bacteria Reference Unit

16th June 2014

Public Health England

Planning for Implementation of WGS 2011- 2014

- PHE investment: financial, laboratory, bioinformatics, staff, training
- Prioritise organisms for routine WGS
- Practical implementation

Public Health England

PHE WGS Sequencing Service Hardware

- Two HiSeq 2500 high-throughput sequencers
- Two MiSeq machines

Capacity ~ 3,000 genomes per week

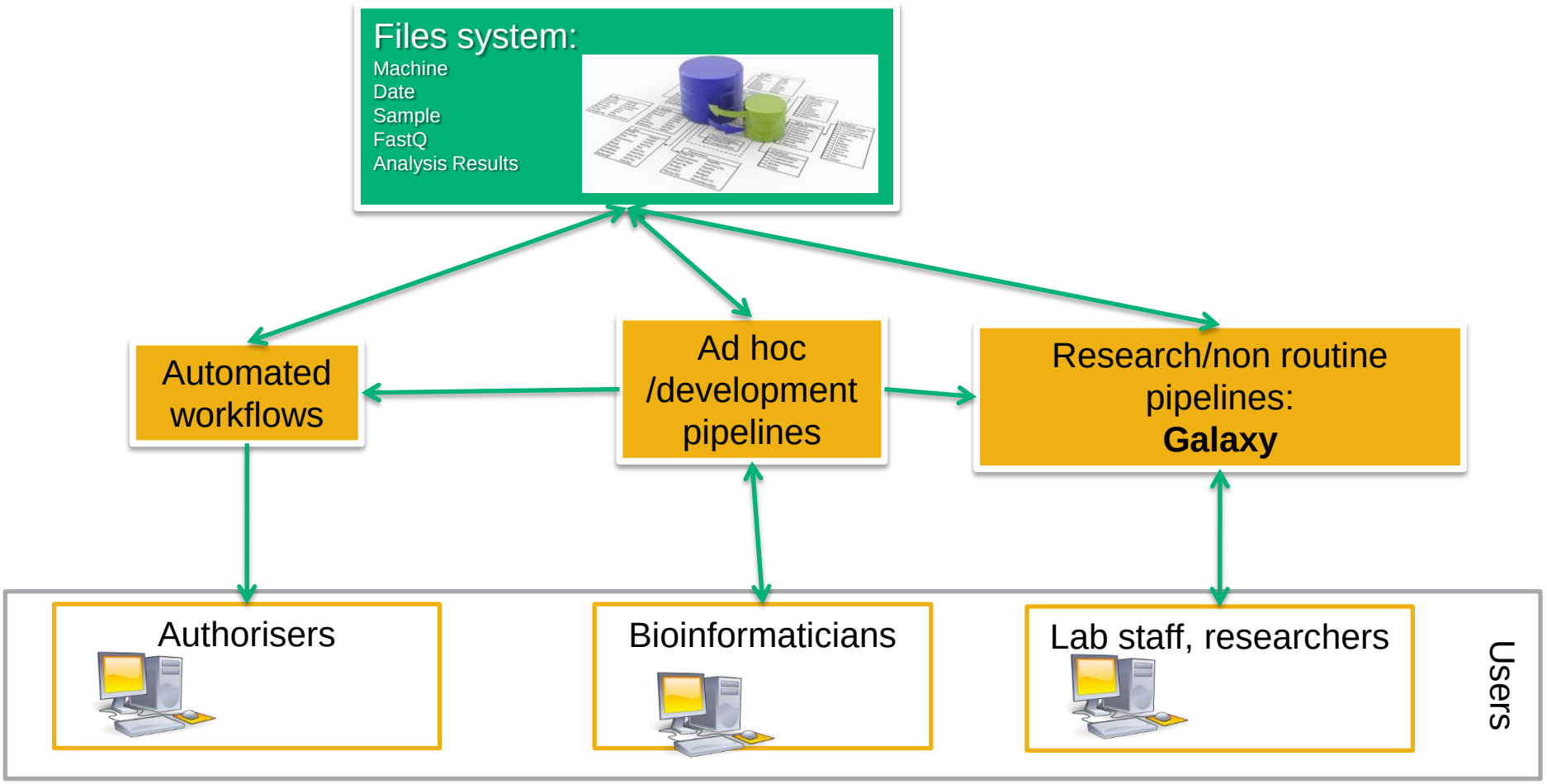
Public Health England

Infrastructure

Data storage warehouse
Generators & Coolers

Public Health England

Skills Developing a Bioinformatics Capability



Data and Bioinformatics Workflow

Sample Workflow

1) Nucleic Acid sent to NGSS along with sample Info.

3) Batch of samples pre-processed (ROBOTICS)

5) Sample batch run on Machine

Information Flow

2) Sample Info to NGS LIMS

Metrics imported in NGS LIMS

4) NGS LIMS exports sample sheet for HiSeq

- QC metrics imported to NGS LIMS
- Bulk Data temp stored

6) Bulk data processed

7) QC data stored

8) Sample Fastq's stored

9) Detailed sample Info linked with NGS data and metrics

10) Requestor accesses files, data metrics and results in web interface

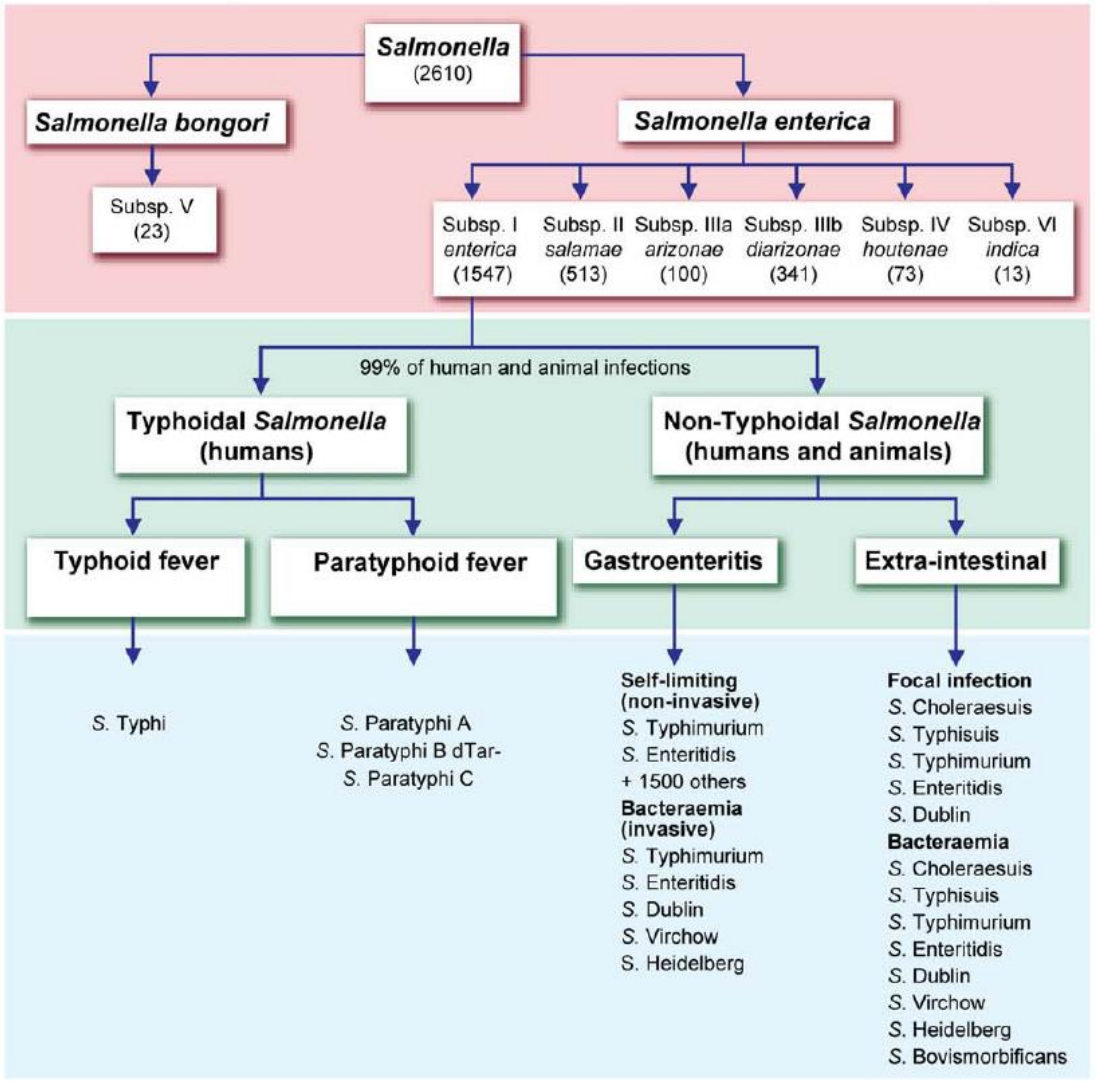
Public Health England

Sequencing Service Validation

- reliable sample handling processes through the robotics
- reproducible high-quality data
- consistent linking of meta-data through the whole sample workflow.
- reliable capture of quality metrics into NGS LIMS
- ISO15189 Accreditation

Public Health England

Salmonella classification is complicated – 20th Century



Species and subspecies were originally defined by DNA-DNA hybridisation, confirmed by MLEE and MLST and are currently differentiated by biochemistry and serology.

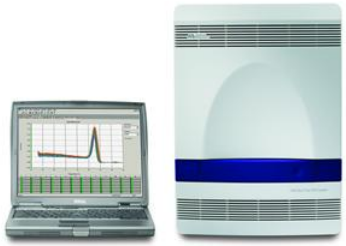
The split in typhoidal and non-typhoidal is based on the disease syndrome. Typhoid and paratyphoid fever is prolonged, whilst extra-intestinal infection is usually acute and metastatic. Gastroenteritis is characterised by diarrhoea.

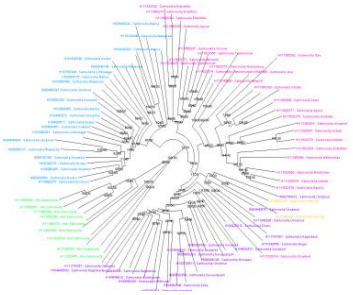
Differentiation of serovars is by agglutination with specific antisera against LPS (O), two phases of flagella (H1 and H2). There are 46 O, 85 H and 1 capsule (Vi) antigen which have been described in about 1,500 combinations within subspecies I.

Public Health England

Current GBRU Typing Methods for *Salmonella*

Subspeciation





Real time TaqMan® PCR assays
- target three different genes

OmniLog® ID System (Biolog)
- phenotypic microarray

Serotyping

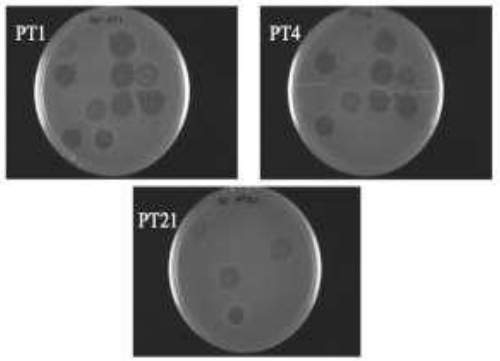
Agglutination with specific antisera against LPS & flagella (O & H antigens)

- **Slide agglutination**
- **Microtitre plates**
- **Dreyer's tubes**

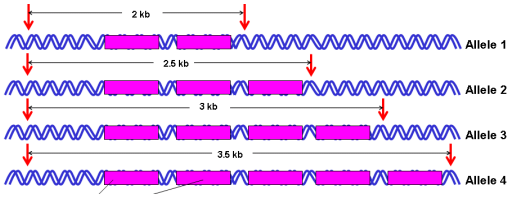
Public Health England

Current Sub-typing Methods for *Salmonella*

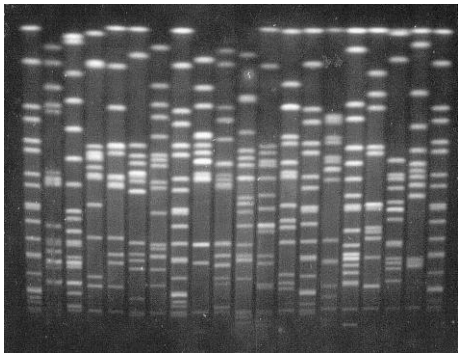
- **Phage typing**
  - **e.g. Typhimurium DT1, DT193**



- **Multi-locus Variable Number Tandem Repeat Analysis (MLVA)**
  - **e.g. 4-13-13-10-0211**



- **Pulsed-field gel electrophoresis (PFGE) - e.g. SNWPXB.0010**



Public Health England

Issues with existing Salmonella typing methods

Turn around times too long

- Serotyping:
  - Originally 25 days
  - Reduced this year to 17 days
- Phage typing:
  - Originally 20 days
  - Reduced this year to 10 days
- PFGE: 4 days, VNTR: 2 days

Biological

- Not a true classification compared with sequence based typing

Safety problems

- Isolates identified local clinical lab as CL2 serovar
- Referred to SRS and handled at CL2
- Identified by reference lab as CL3

Quality

- Typing methods can be difficult to standardise – including existing molecular methods

Public Health England

Salmonella NGS Project

Salmonella identified as a priority organism

- Use of whole genome sequencing to replace lengthy laboratory methods and improve safety, quality
  - Serotying
  - Phage typing
  - PFGE
  - MLVA

WGS provides opportunity for identification and typing

using a single method

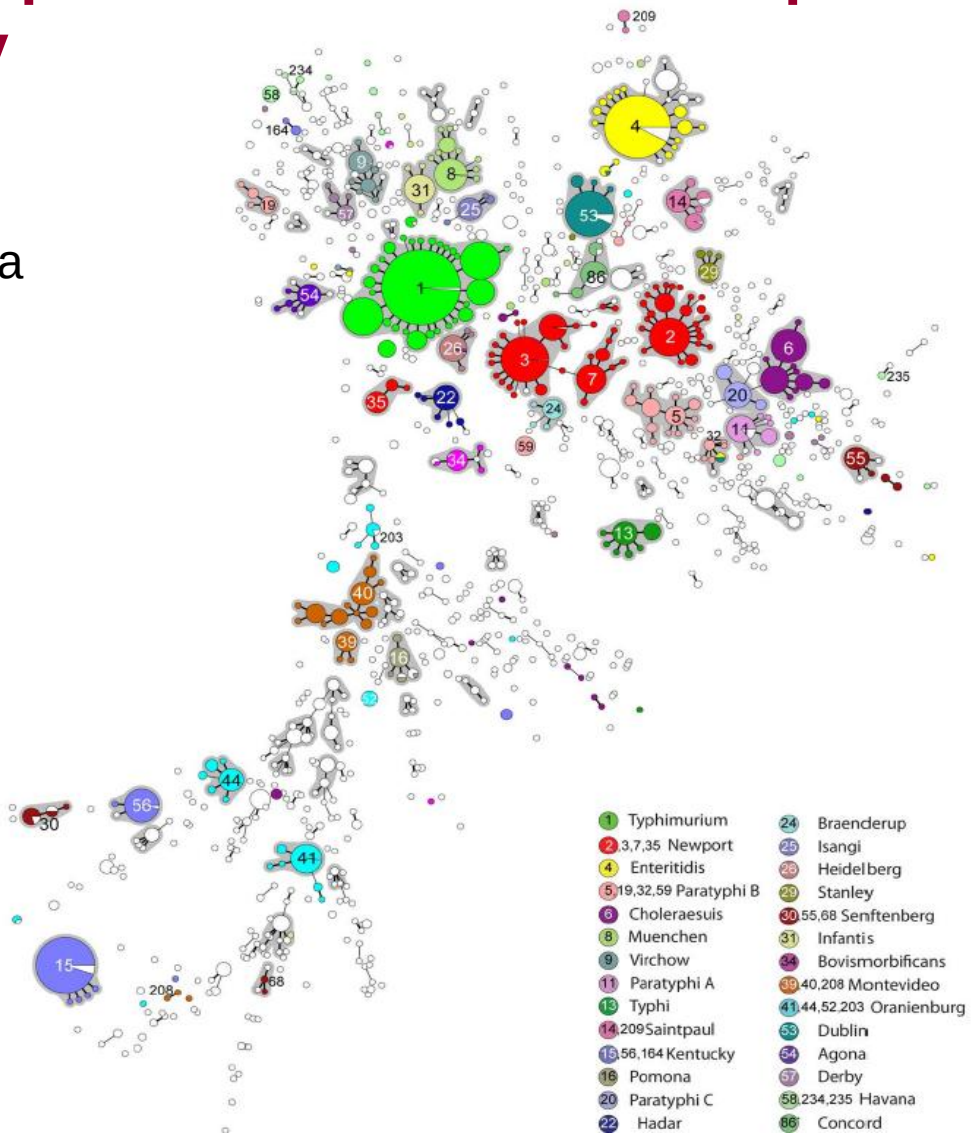
WGS = MLST data + SNP detection + lots of other interesting data

Public Health England

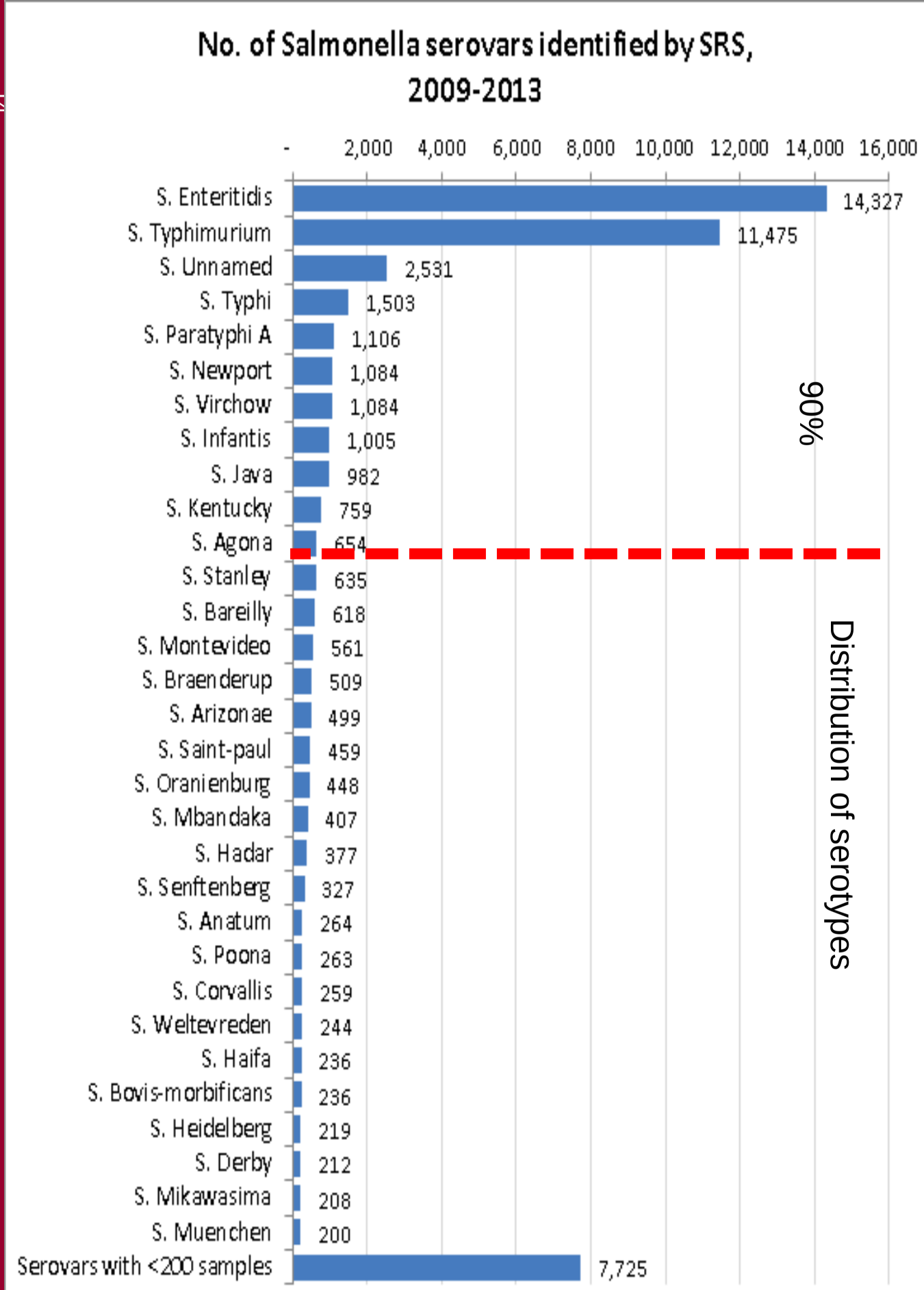
Salmonella population structure is complicated – 21st Century

Minimal spanning tree of MLST data for *S. enterica* subspecies *enterica*

- Each circle corresponds to a sequence type (ST)
- The size is proportional to the number of isolates
- eBGs are natural clusters of genetically related isolates
- Increasing distance equates to fewer shared alleles
- MLST STs correlate with serotypes



Achtman et al., 2012


No. of Salmonella serovars identified by SRS, 2009-2013
- 2,000 4,000 6,000 8,000 10,000 12,000 14,000 16,000
S. Enteritidis 14,327
S. Typhimurium 11,475
S. Unnamed 2,531
S. Typhi 1,503
S. Paratyphi A 1,106
S. Newport 1,084
S. Virchow 1,084
S. Infantis 1,005
S. Java 982
S. Kentucky 759
S. Agona 654
S. Stanley 635
S. Bareilly 618
S. Montevideo 561
S. Braenderup 509
S. Arizonae 499
S. Saint-paul 459
S. Oranienburg 448
S. Mbandaka 407
S. Hadar 377
S. Senftenberg 327
S. Anatum 264
S. Poona 263
S. Corvallis 259
S. Weltevreden 244
S. Haifa 236
S. Bovis-morbificans 236
S. Heidelberg 219
S. Derby 212
S. Mikawasima 208
S. Muenchen 200
Serovars with <200 samples 7,725
90%
Distribution of serotypes

Public Health England

Salmonella NGS Project - 2013

Validation set

- 1500 strains selected for sequencing
- 1000 common strains representative of 2012 (10%)
  - *>50% Salmonella* Enteritidis & *Salmonella* Typhimurium
  - different phage types
- 500 strains of less common serovars
  - proportional representation of 2012

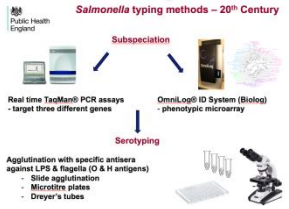
Common Serovar	No of Isolates
Salmonella Enteritidis	364
Salmonella Typhimurium	337
Salmonella Infantis	36
Salmonella Typhi	36
Salmonella Java	33
Salmonella Paratyphi A	33
Salmonella Newport	32
Salmonella Virchow	31
Salmonella Kentucky	22
Salmonella Stanley	20
Salmonella Braenderup	19
Salmonella Montevideo	19
Salmonella Agona	18
	1000

PHE in the 21st Century – 1st phase validation

Public Health England

Salmonella sample from 2012

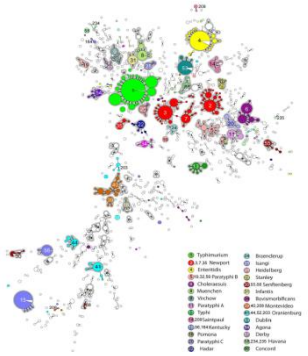
Back to the 20th Century



PHE Bioinformatics pipeline

- Quality trim
- KmerID to check purity
- Short read sequence typing to determine MLST

Sequencing of 1500 representative Salmonella – results compared with serotyping

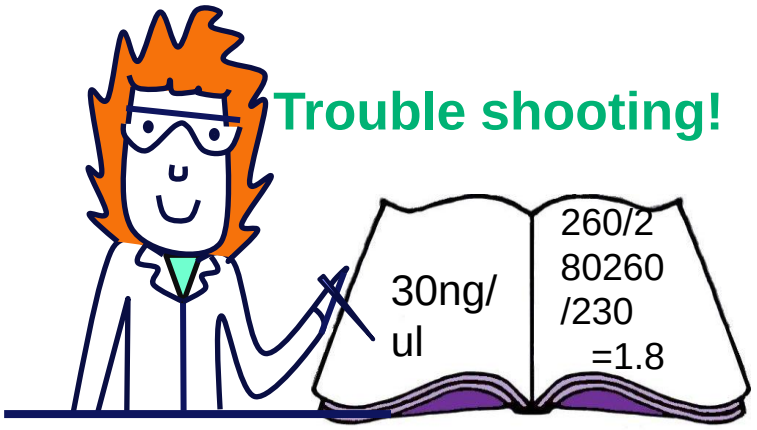


MLST/EBG-serotype

Public Health England

Salmonella NGS project - workflow

- Innoculate broth culture (overnight growth) or use growth on slopes
- Automated Genomic Extraction – QiaSymphony 96 well
- Measure DNA quantity & quality
  - Glomax/Labchip



- Automated Library preparation (Nextera)
- Sequence on HiSeq2500 (Rapid run)
- Automated Bioinformatics Analysis – pipeline development, analysis tools

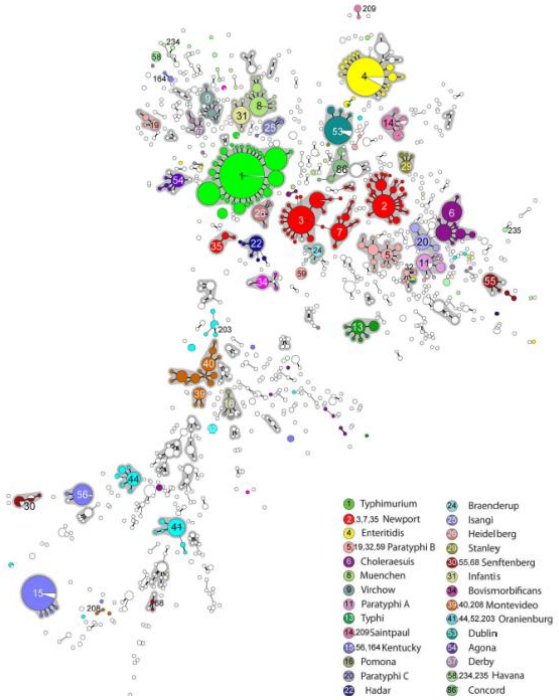
Public Health England

Results - MLST derivation

- WGS MLST derived grouping correlated with traditional serogroup > 94% for the common serovars (Common serovars make up to 90% of the workload)

- But lower correlation with rarer serovars

- Current MLST database
  - Only 900 out of 2600 serotypes have be assigned MLST profiles
  - Mis-matches between serotypes and MLST serogroups

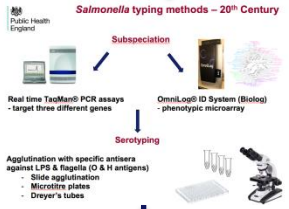


Public Health England

PHE in the 21st Century – 2nd phase validation
Routine use of sequencing

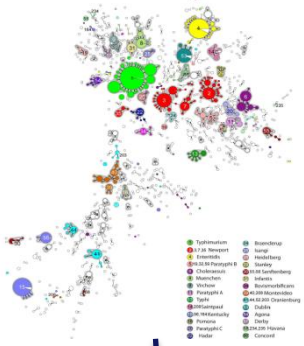
Sample received by reference lab

Back to the 20th Century



Reported to customer

MLST/serotype



If the strain belongs to serotype in our 'Top 14' it goes for SNP analysis

Reporting to customer

Public Health England

Detection of *Salmonella* outbreaks

At PHE, laboratory and epidemiological staff work closely together to detect and investigate outbreaks

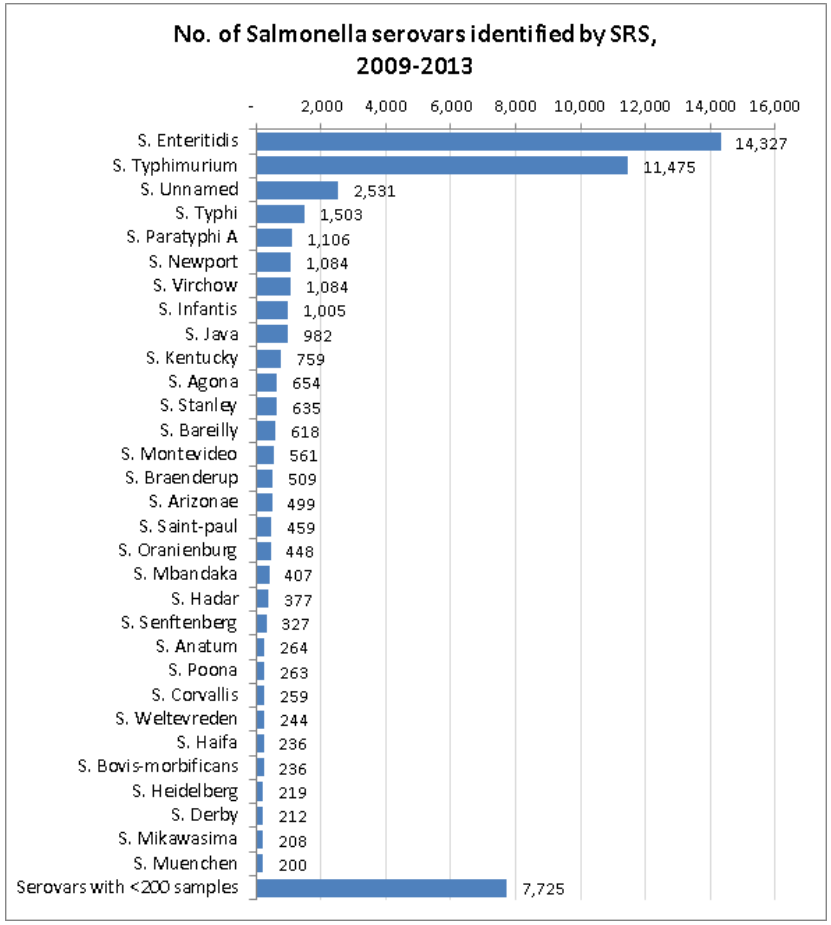
Currently, this is done on the basis of serotype, phage type, MLVA and PFGE - these techniques have varying resolution and molecular typing not performed on every isolate

Use an ‘exceedance’ above what we would expect to see as background, before an outbreak investigation is triggered

The more common a serotype is, the harder it is to spot an outbreak

Public Health England

Top 14 serotypes – SNP typing



Serotype	Total number EBGs	Number of EBGs seen >5 times in phase I validation	Number STs
Enteritidis	3	1	9
Typhimurium	4	2	16
Typhi	1	1	1
Paratyphi-A	1	1	1
Newport	5	2	25
Virchow	2	2	4
Infantis	1	1	1
Paratyphi-B/Java	5	2	25
Kentucky	3	1	9
Agona	3	1	9
Stanley	1	1	1
Montevideo	3	1	9
Braenderup	1	1	1
Oranienburg	6	0	36
Total	**39**	**17**	**147**

IPython Notebook - bit.ly/1t2g5kl

David Powell

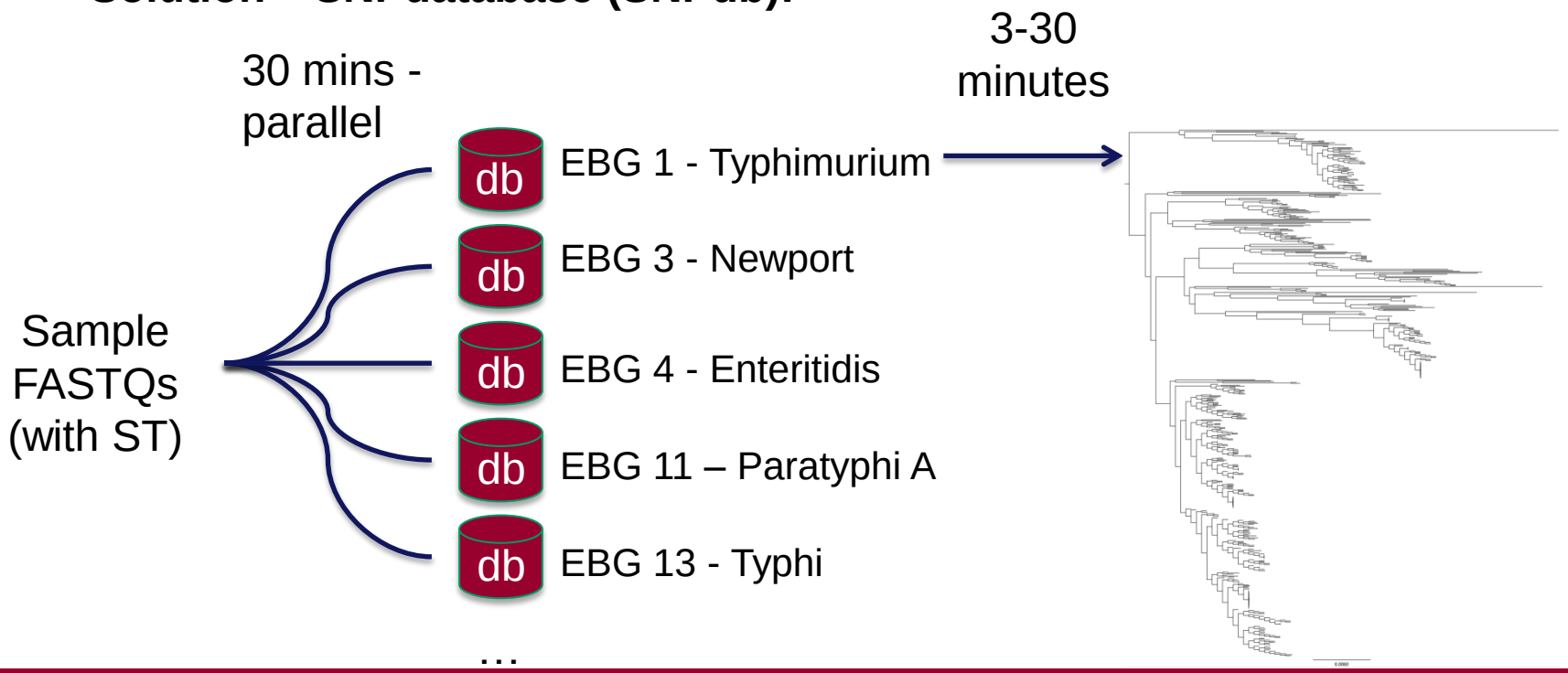
Public Health England

Top 14 serotypes – SNP typing

Challenges:

- Many EBGs
- Hundreds of strains a week
- Rapid, hands-off analysis

Solution – SNPdatabase (SNPdb):



Public Health England

Uploading data into Short Read Archive

Public Health England Pathogen Sequencing

Accession: PRJNA248064 ID: 248064

Whole genome sequencing data from Public Health England.

Project Type: Umbrella project

Relevance: Medical

▾ SRA Data Details

Parameter	Value
Data volume, Gbases	1
Data volume, Mbytes	248

This project encompasses the following sub-project:

Project Type	Number of Projects
Genome sequencing *Highest level of assembly :* SRA or Trace	1

BioProject accession	Assembly level	Name	Title
PRJNA248792	SRA or Trace	Public Health England - Gastrointestinal Bacteria Reference Unit pathogens Genome sequencing	Public Health England - Gastrointestinal Bacteria Reference Unit pathogens Genome sequencing (Public Health England)

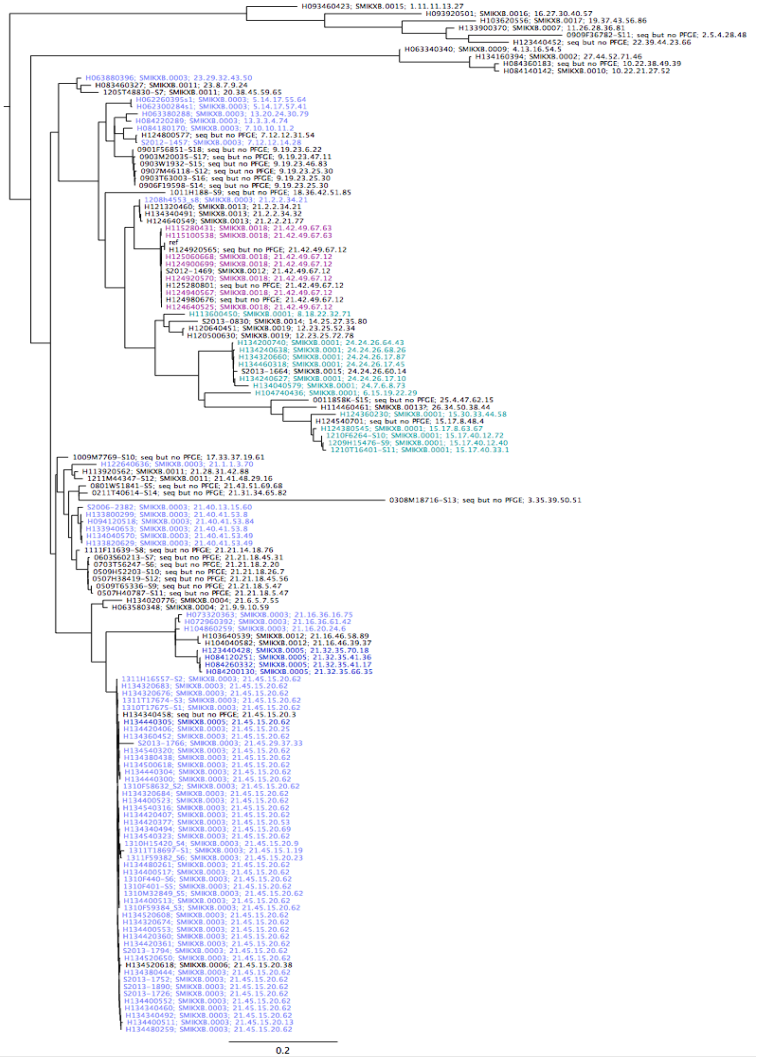
Submission:

Registration date: 19-May-2014

Public Health England

NCBI BioProject accession: PRJNA248064

Salmonella Mikawasima Outbreak WGS Analysis Dec 2013



- Dec 2013 increase in Salmonella Mikawasima in England, Wales, Scotland
- Several different PFGE profiles but 2 predominant ones
- Sequenced 109 isolates England & Wales, 11 Scotland and included in analysis 38 sequenced in Denmark (SSI, DTU)
- 80 from 2013, 28 2012
- 44 isolates with OB PFGE profile clustered <10 SNPs (31 E, 10 D, 3 S) also 3 isolates with different PFGE profile
- 4 with this PFGE profile formed distinct cluster (<10SNPs) with isolate from 2009
- 6 isolates with 2nd OB profile clustered with Scottish isolate with different profile
- Colours represent different PFGE profiles

Public Health England

Acknowledgements

This presentation is the result of work across PHE particularly Genomic Services Unit, Bioinformatics Unit and the Gastrointestinal Bacteria Reference Unit

GBRU

Elizabeth de Pinna, Tansy Peters, Satheesh Nair, Tim Dallman, Phil Ashton and other staff in the lab

Genomic Services Unit

Cath Arnold and team

Bioinformatics Unit

Jonathon Green, Anthony Underwood, Rediat Tewolde