



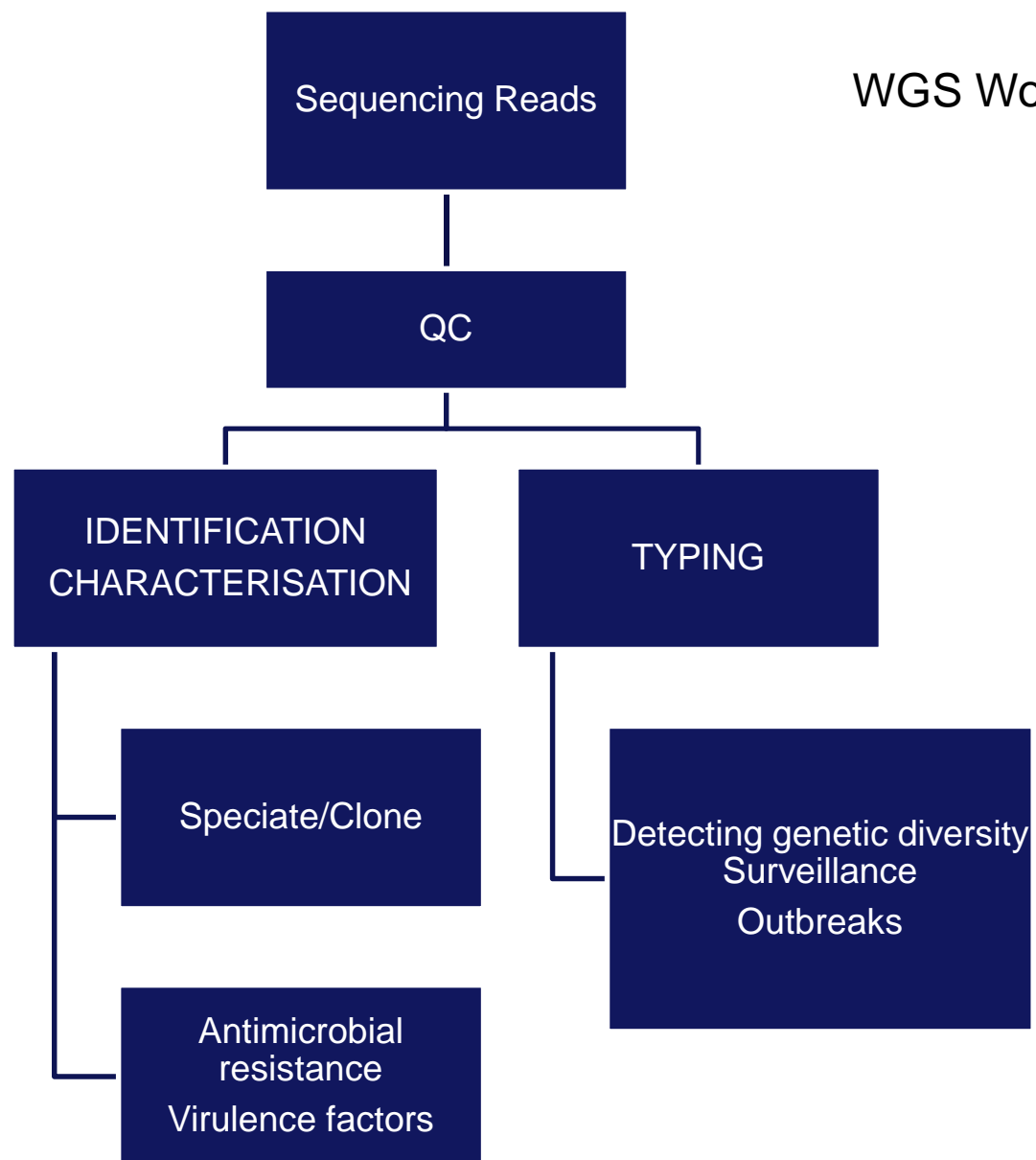
Public Health  
England

# Routine WGS Analysis of GI Pathogens

Dr Tim Dallman

Gastrointestinal Bacteria Reference Unit

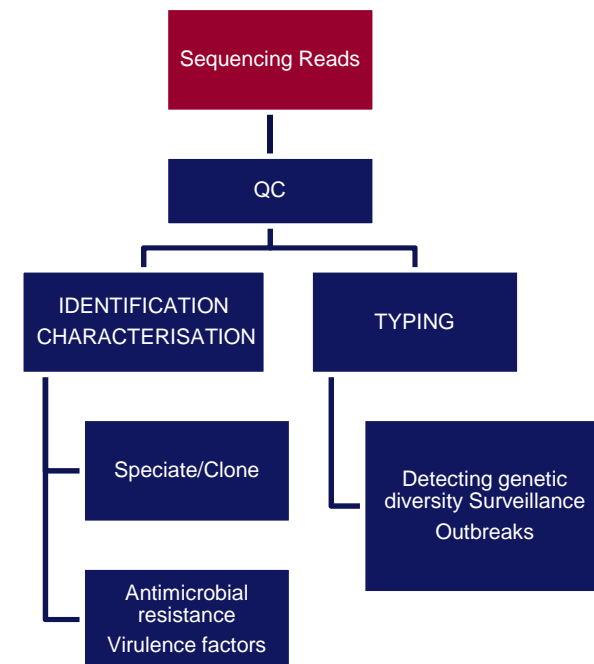
29<sup>th</sup> January 2015





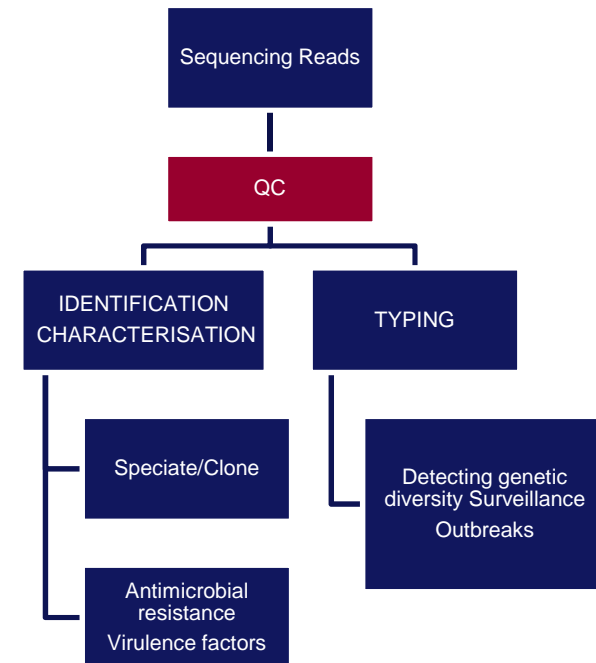
# Sequencing Platform

- Illumina HiSeq 2500
  - Rapid 27hr run
  - 2x100 bp read length
  - >150Mb of sequence





- Deplex - CASAVA
- Adapter trimming - CASAVA
- Quality trimming - Trimomatic
- Deduplication – Diginorm
- Error correction - Quiver



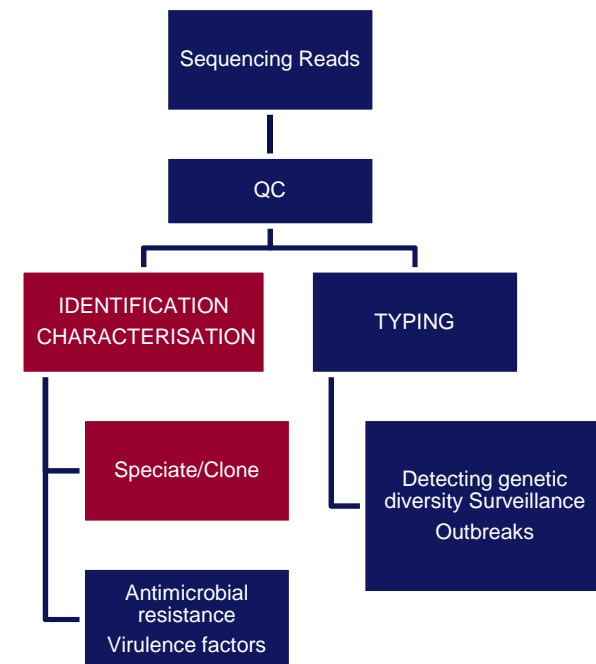


# Identification / Mixed

- K-mer Gateway

- Have I sequenced what I thought I have?
- Is my sequence mixed?

A k-mer is a nucleotide sequence of length k.





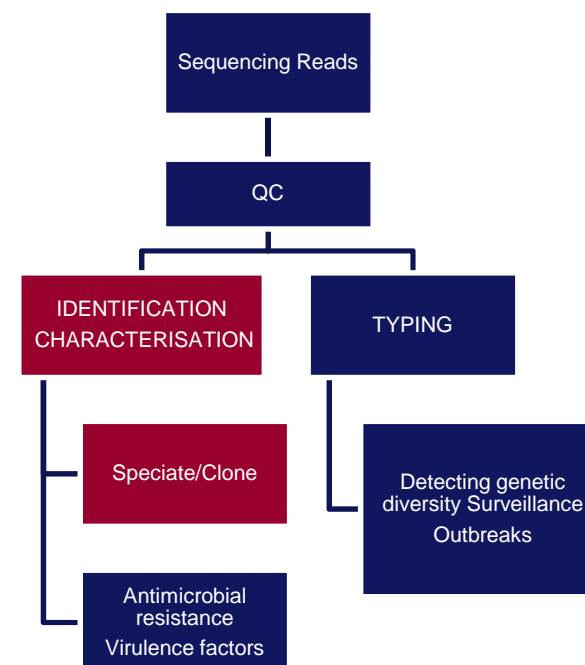
# Identification / Mixed

## •K-mer Gateway

Acetobacter	Klebsiella
Acinetobacter	Lactobacillus
Actinomyces	Legionella
Aeromonas	Leptospira
Aggregatibacter	Leuconostoc
Bacillus	Listeria
Bacteroides	Morganella
Bartonella	Mycobacterium
Bifidobacterium	Mycoplasma
Bordetella	Neisseria
Borrelia	Nocardia
Brucella	Paenibacillus
Burkholderia	Prevotella
Campylobacter	Propionibacterium
Chlamydia	Pseudomonas
Chlamydophila	Rhizobium
Clostridium	Rhodococcus
Corynebacterium	Rickettsia
Desulfovibrio	Salmonella
Enterobacter	Shewanella
Enterococcus	Shigella
Escherichia	Staphylococcus
Francisella	Streptococcus
Fusobacterium	Streptomyces
Gardnerella	Treponema
Gordonia	Ureaplasma
Haemophilus	Vibrio
Helicobacter	Yersinia

Off all the k-mers of length 18 in each reference genome what percentage are in our sequencing reads?

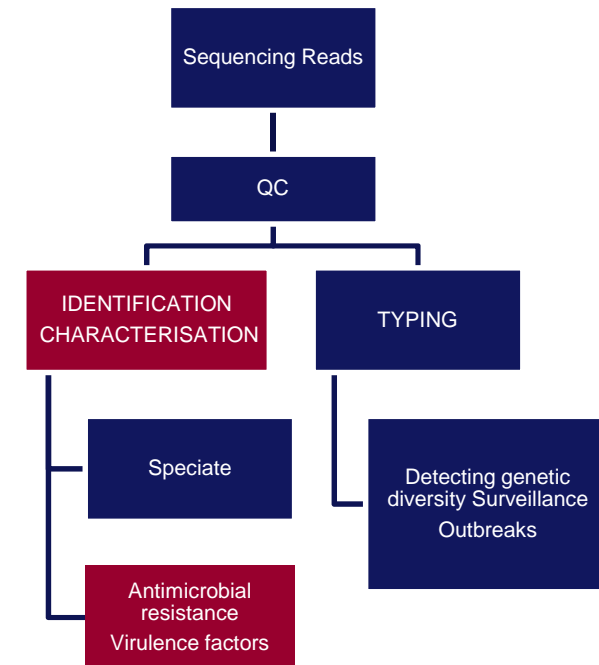
Can be used to identify cross species contamination





# Strain Characterisation

- GeneFinder
  - Antimicrobial / Toxin / Plasmid
- Stx Subtyping
  - Ashton, P. *et al. Insight into Shiga toxin genes encoded by Escherichia coli O157 from whole genome sequencing.* (PeerJ PrePrints, 2014)
- Geno/Pheno
  - Monophasic identification
  - Serotype identification in *E. coli*





Public Health  
England

# AMS :Genotype vs phenotype

Validation : 642 Salmonella strains

Resistance : 57.5 % susceptible  
24.7 % multi-resistant (> 2 classes)

	Phenotype S		Phenotype R		Error%
	Genotype S	Genotype R	Genotype R	Genotype S	
CHL	580	2	39	3	0.78
SUL	466	1	149	8	<b>1.40</b>
TET	467	3	151	3	0.93
TMP	562	3	57	2	0.78
AMP	484	5	134	1	0.93
CTX/CAZ	618	1	5	0	0.16
CPR	619	1	4	0	0.16
FOX	612	0	7	5	0.78
CIP	478	4	138	4	<b>1.25</b>
NAL	485	2	127	10	<b>1.87</b>
GEN	608	1	14	1	0.31
TOB	613	2	8	1	0.47
AMK	622	0	1	1	0.16
STR	486	18	119	1	2.96
	7700	43	953	40	<b>0.95</b>

Michel Doumith/Martin Day

Major errors

Very major errors

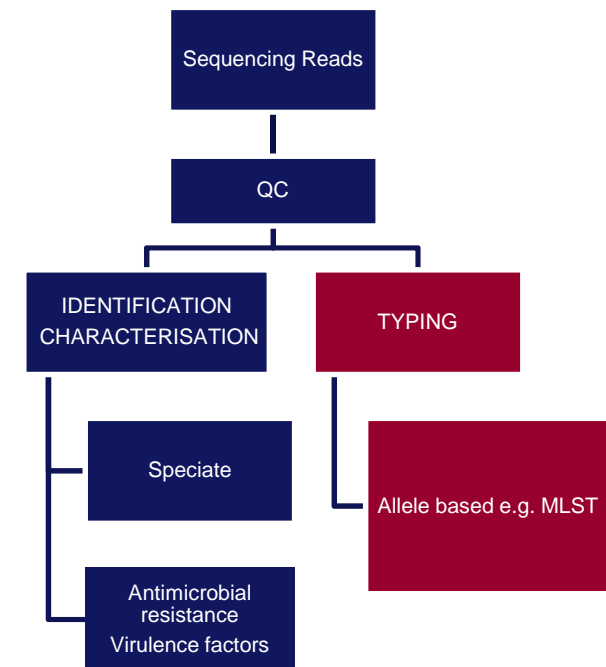




# Typing - Allele

**SRST** – Inouye, M *et al.* Short read sequence typing (SRST): multi-locus sequence types from short reads. *BMC Genomics* **13**, 338 (2012).

- Provides per base quality of allele call





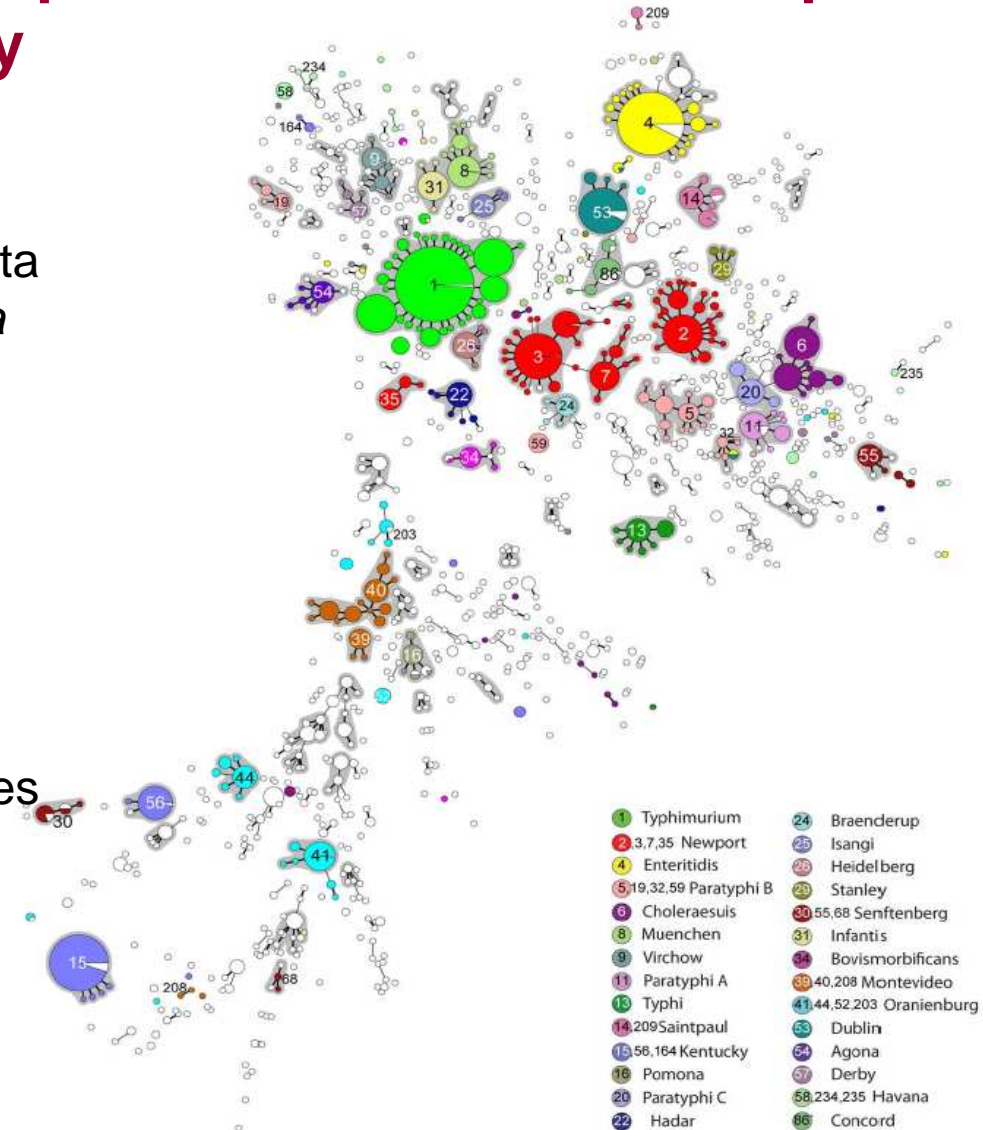
Public Health  
England

# *Salmonella* population structure is complicated

## – 21<sup>st</sup> Century

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

- Each circle corresponds to a sequence type (ST)
- eBGs are natural clusters of genetically related isolates
- MLST STs correlate with serotypes

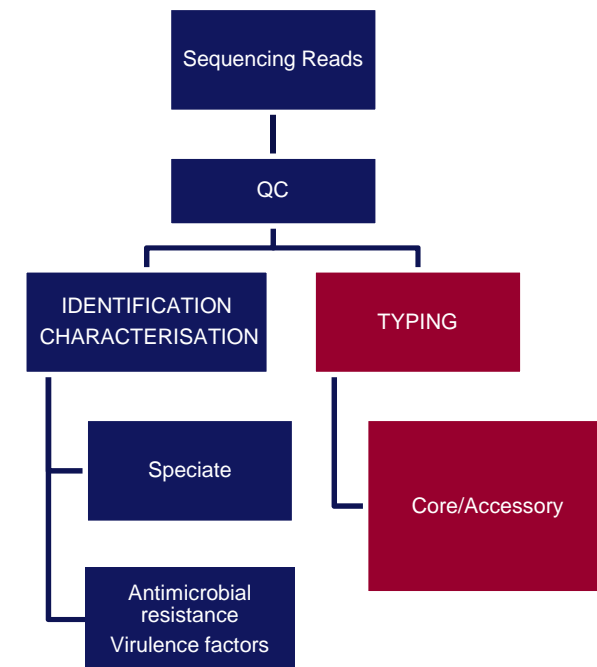


Achtman et al., 2012

# Typing – Core / Accessory

## Chimera

- Pan – Genome Species analysis
- Hidden Markov Models of Gene Families
- Rapid method of finding related isolates
- Cross compatible with wgMLST schemes

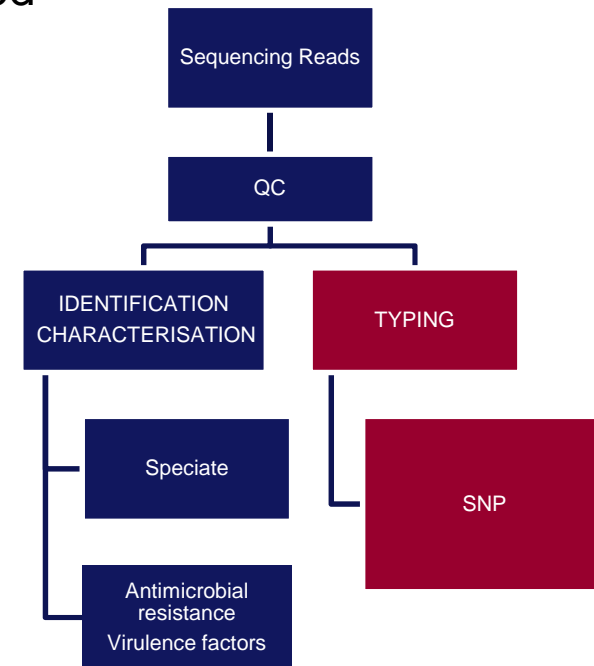
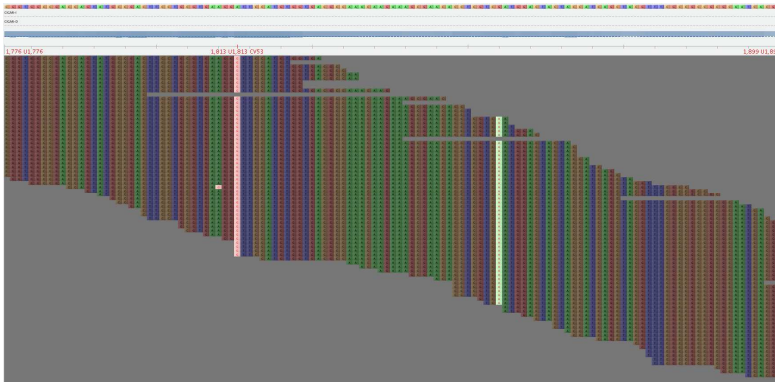




# Typing – SNP

## SnapperDb

- Mapping reference genome (BWA-MEM)
- Variants identified (GATK2)
- Variants and Uncertain, Recombinant positions stored
- Parallelisation of SNP analysis

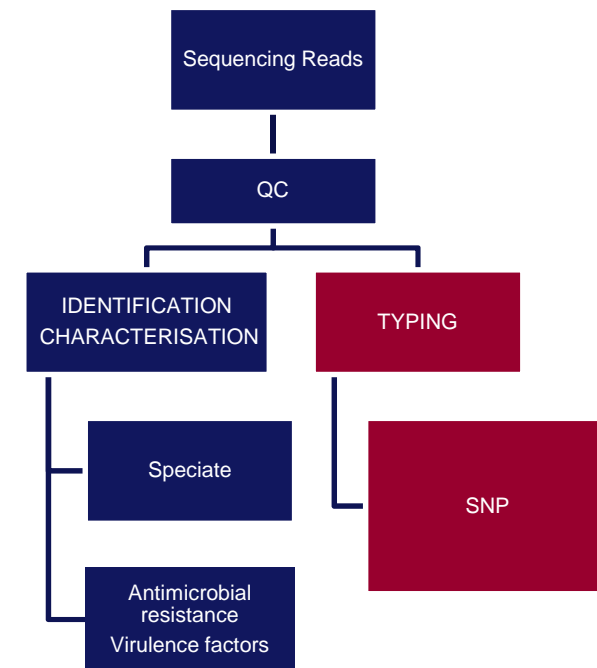
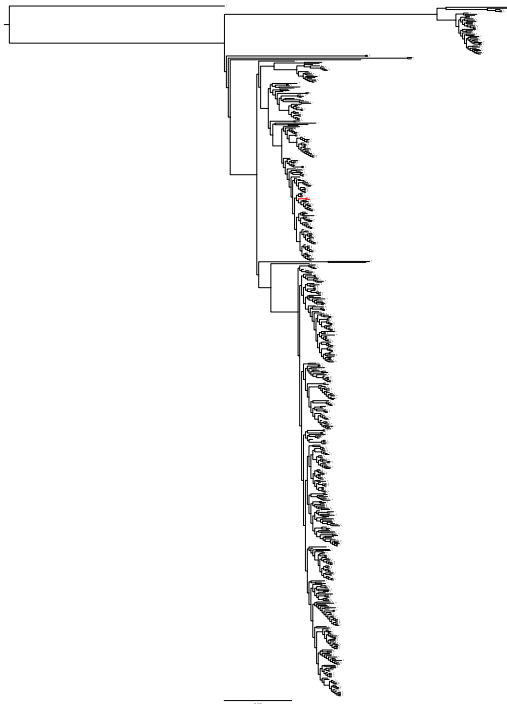




# Typing – SNP

## SnapperDb - Outputs

- SNP alignments
- Annotated variants

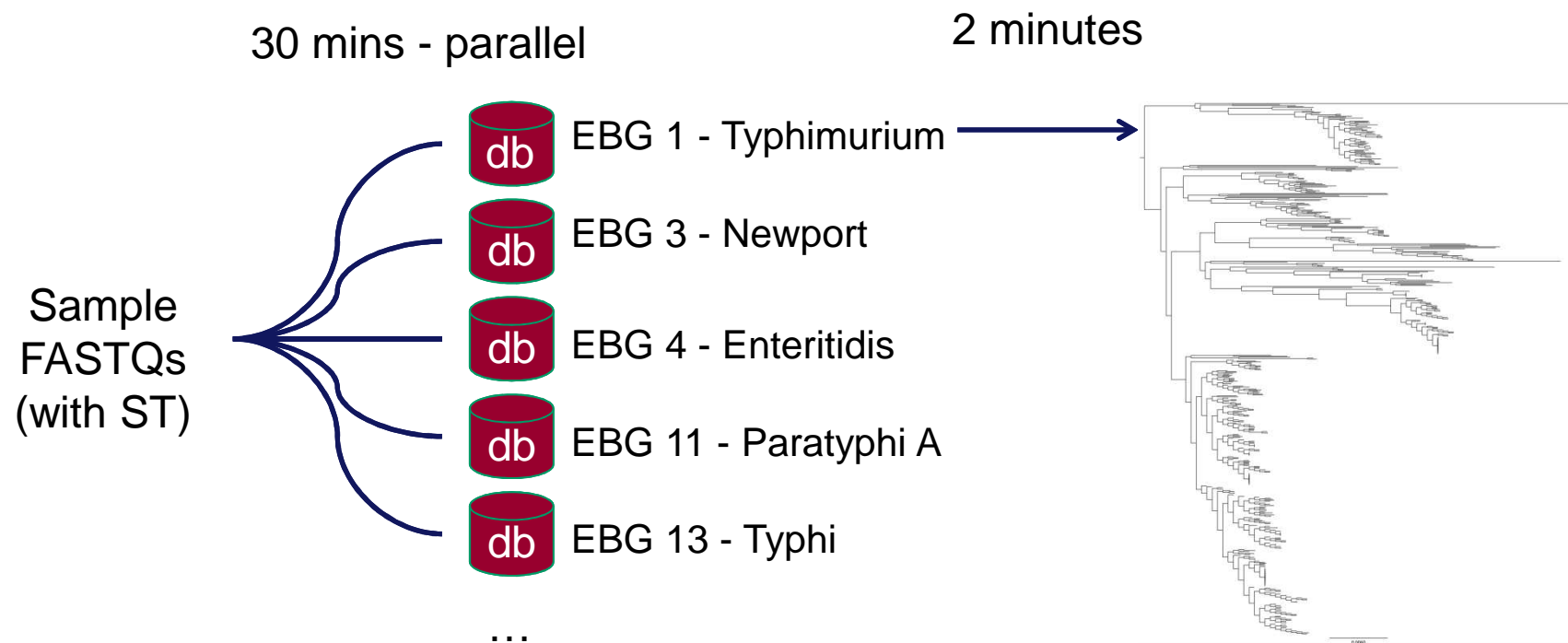




# Salmonella SnapperDb

## Challenges:

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

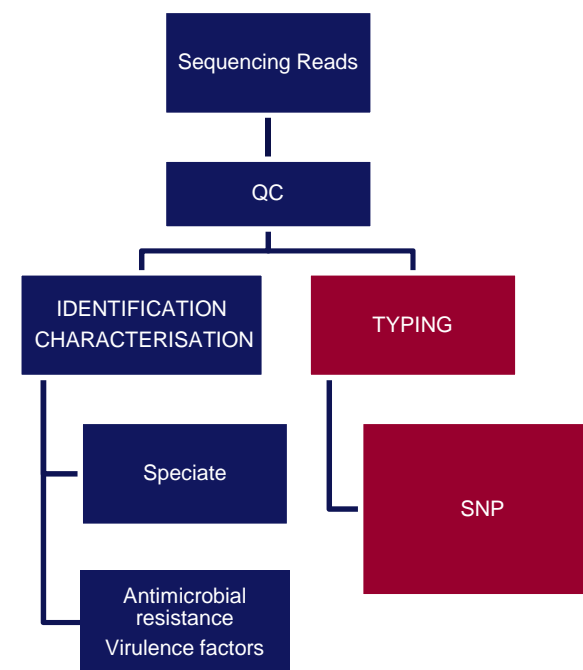




# Typing – SNP

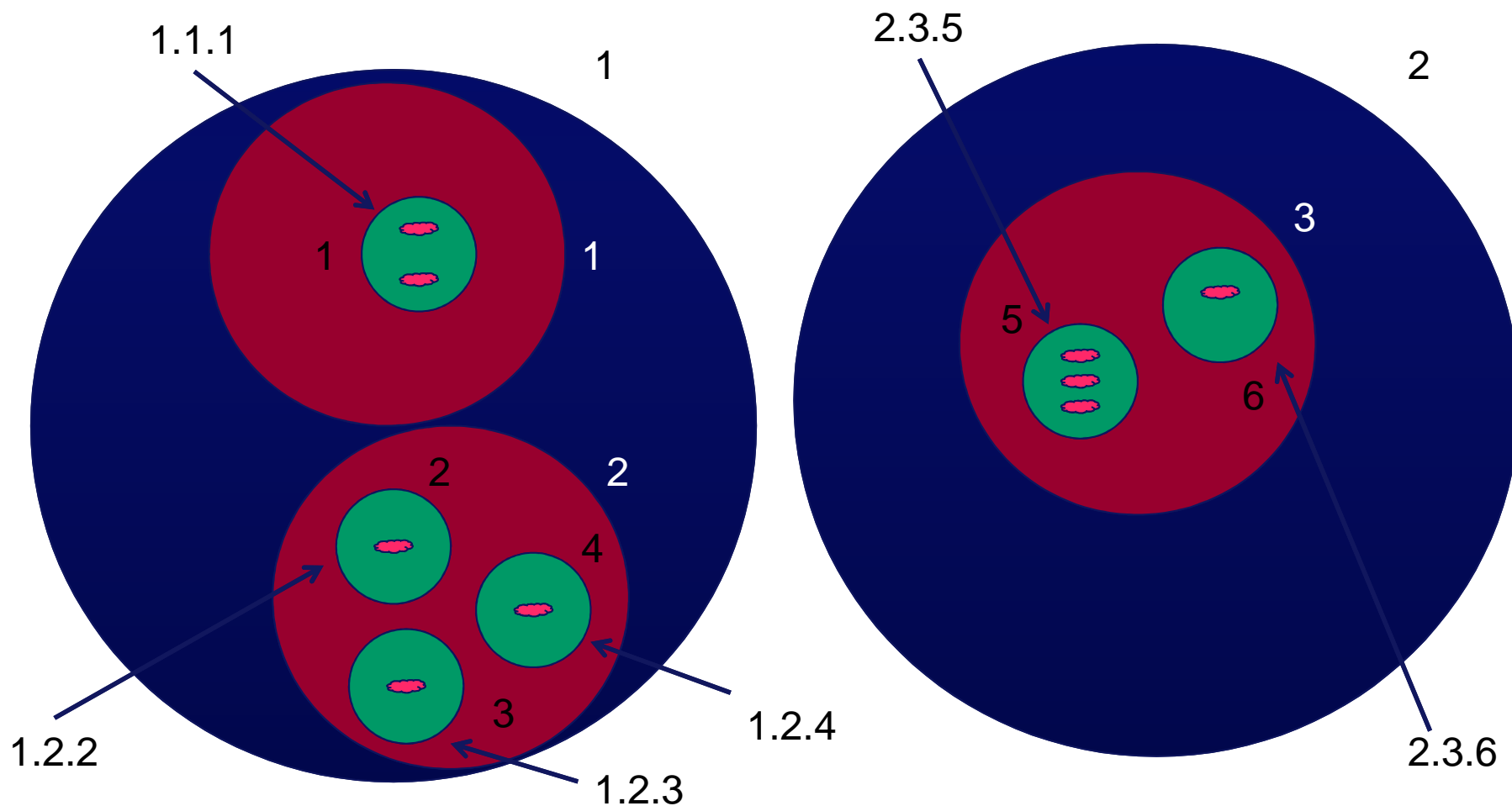
## SnapperDb – Clustering

- Maintain a SNP distance matrix
- Hierarchical Clustering

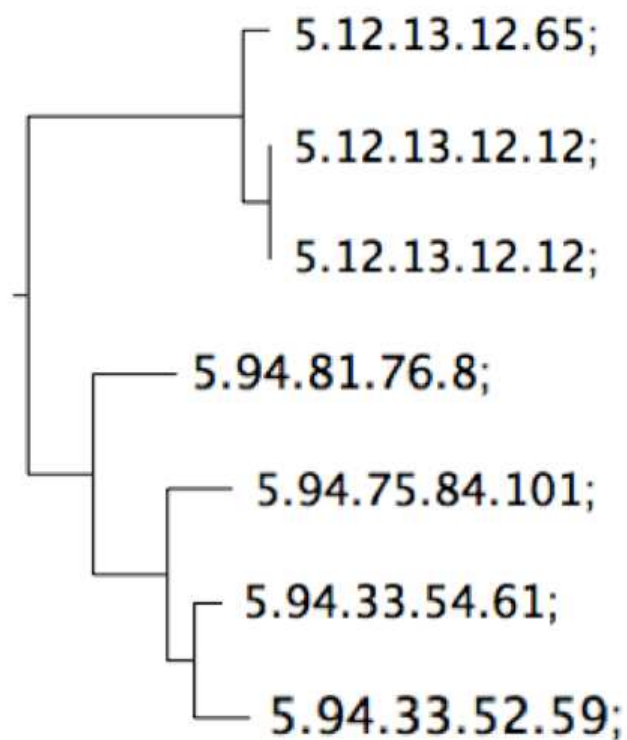




# SNP Address







# SNP address

- Hierarchical clustering based on full pairwise distance between two genomes
- Used to assign a SNP address to a strain based on specified index e.g. 100:50:25:10:5
- Can be used for surveillance purposes



Public Health  
England

# Uploading data into Short Read Archive

## Public Health England Pathogen Sequencing

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
<b>Genome sequencing</b> <i>Highest level of assembly :</i> 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



Public Health  
England

The screenshot shows the Galaxy / PHE web interface in a browser window. The address bar shows the URL 158.119.147.85. The interface has a dark blue header with the 'Galaxy / PHE' logo and navigation tabs: 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. A green status bar in the top right corner indicates 'Using 49%'. The left sidebar, titled 'Tools', contains a search bar and a list of tool categories: 'Get Data', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Operate on Genomic BED Intervals', 'Convert Formats', 'Genome Annotation', 'Extract Features', 'Picard tools', 'Statistics', 'Graph/Display Data', 'NGS: Simulation', 'Multiple Alignments', 'Phylogenetics', 'FASTA manipulation', 'NCBI BLAST+', 'NGS: PHE internal tools', 'NGS: QC and manipulation', 'NGS: Assembly', 'NGS: Mapping', 'NGS: SAM Tools', 'NGS: GATK Tools (beta)', and 'Workflows' (with a sub-link 'All workflows'). The main content area features a green welcome message 'Welcome to the PHE Galaxy environment' with a checkmark icon, followed by the Public Health England logo and name. Below this, a paragraph describes Galaxy as an open, web-based platform for data intensive biomedical research, mentioning its association with Penn State and Emory University. The right sidebar, titled 'History', shows 'Unnamed history' with '0 bytes' and a message: 'Your history is empty. Click 'Get Data' on the left pane to start'.



Public Health  
England

# Key Goals

- Pathogen Agnostic Analysis
- Pathogen Specific Interpretation
- Parallelisation & Scalability



Public Health  
England

# Key Challenges

- Enterprise level codebase
- Cross Compatibility
- Communication



Public Health  
England

# Key Opportunities

- Harmonisation
- Inter-operability
- Public Health Impact



# Acknowledgements

## GBRU

Claire Jenkins, Neil Perry, Elizabeth de Pinna, Tansy Peters, Satheesh Nair, Kathie Grant, Phil Ashton and other staff in the lab

## Genomic Services Unit

Cath Arnold and team

## Bioinformatics Unit

Jonathon Green, Anthony Underwood, Rediat Tewolde, Ulf Schaefer, Aleksey Jironkin, Michel Doumith