Public Health England

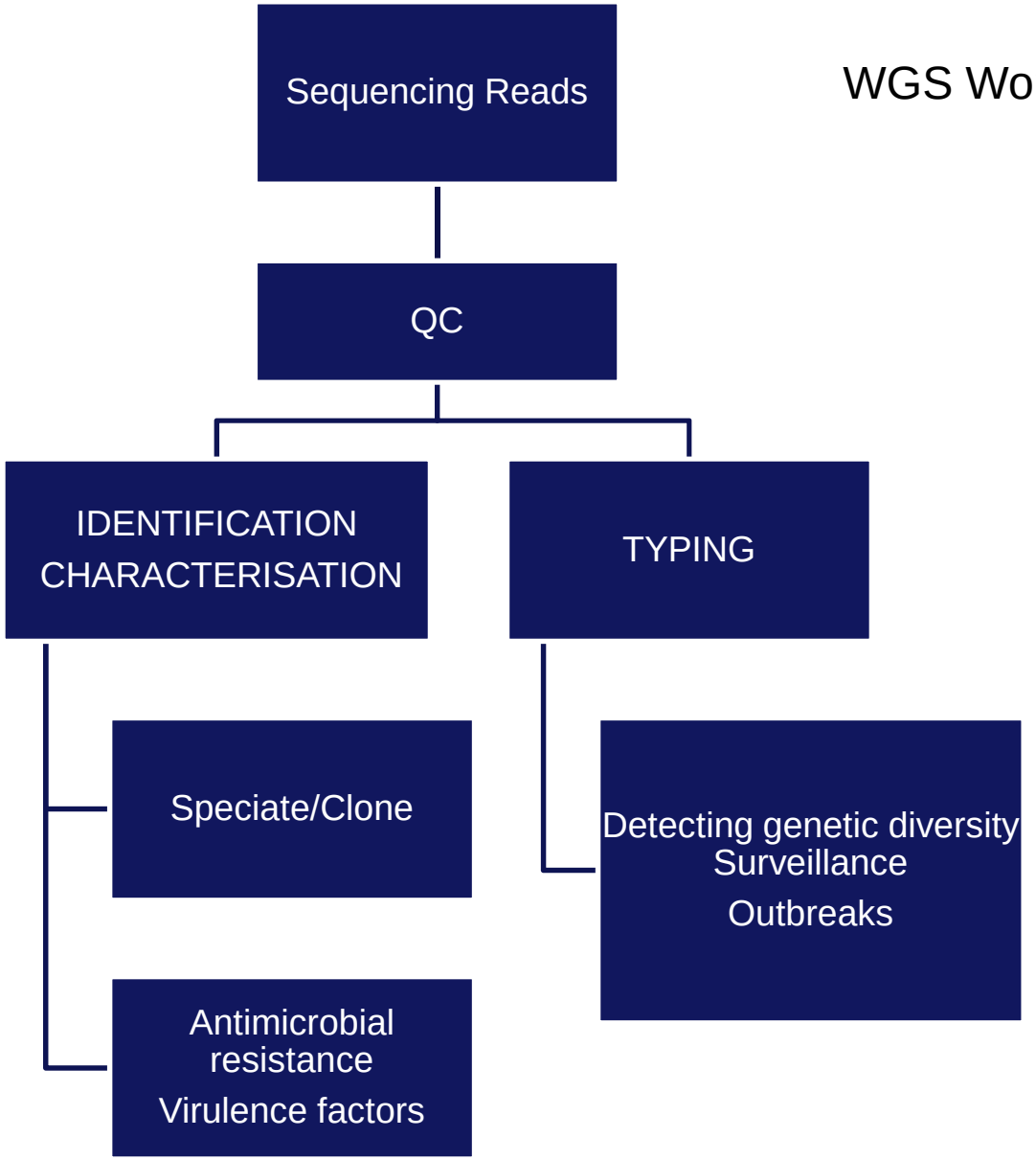
Routine WGS Analysis of GI Pathogens

Dr Tim Dallman
Gastrointestinal Bacteria Reference Unit

29th January 2015

Public Health England

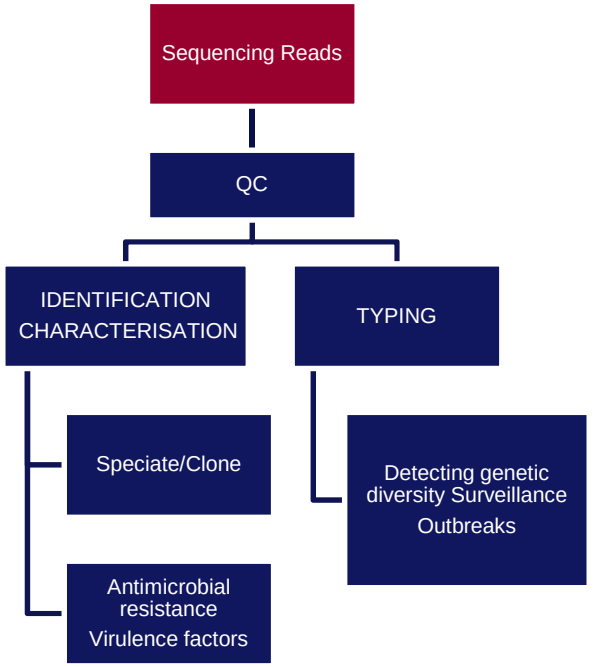
WGS Workflow



Public Health England

Sequencing Platform

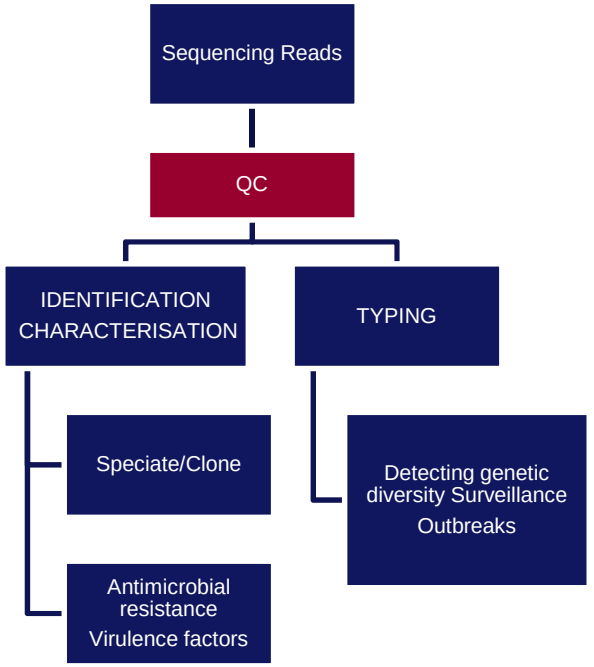
- Illumina HiSeq 2500
  - Rapid 27hr run
  - 2x100 bp read length
  - >150Mb of sequence



Public Health England

QC

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

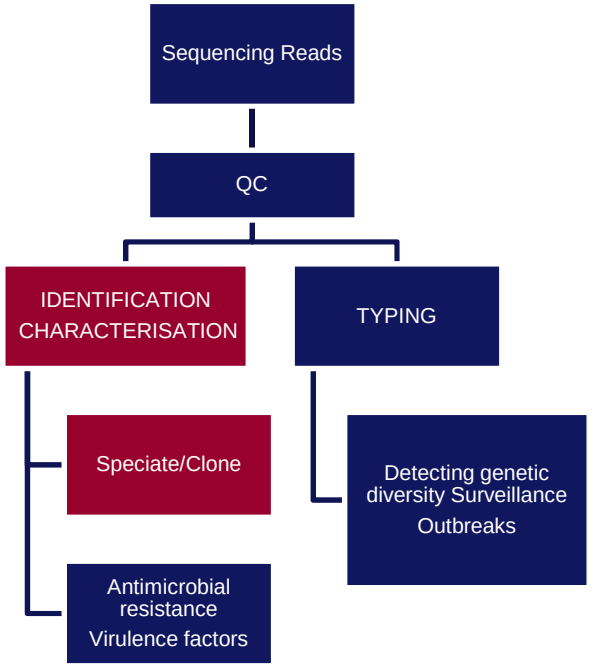


Public Health England

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.



Public Health England

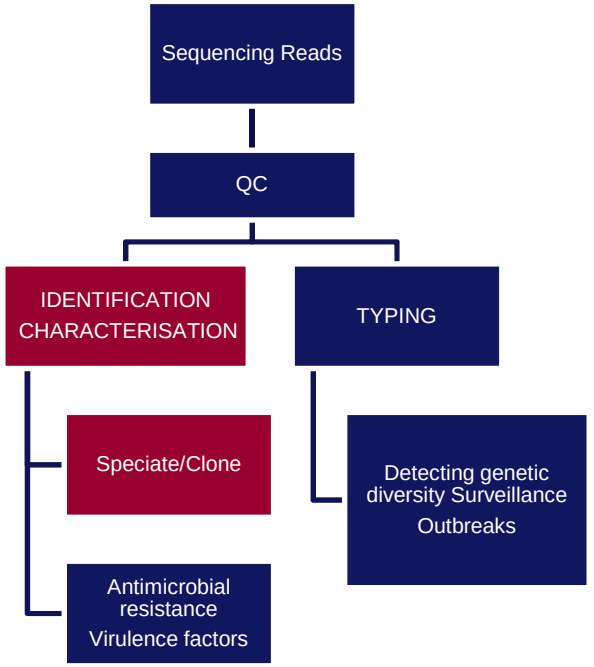
Identification / Mixed

- 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?

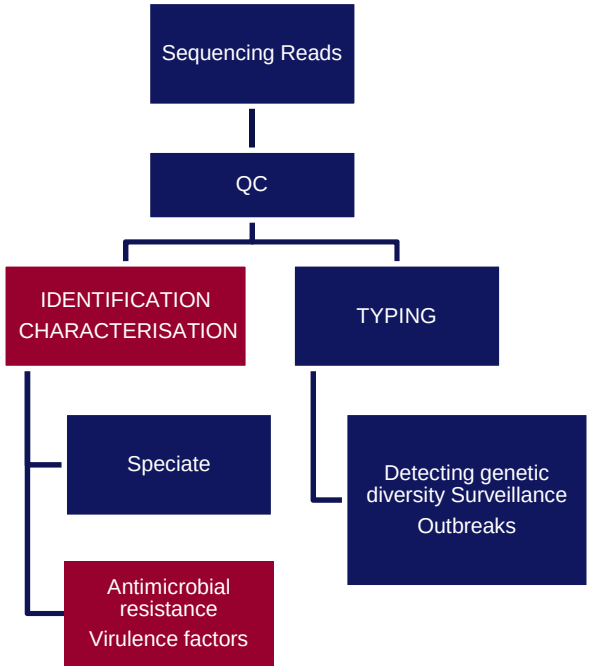
Can be used to identify cross species contamination



Public Health England

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*



AMS :Genotype *vs* phenotype

Public Health England

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	**1.40**
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	**1.25**
NAL	485	2	127	10	**1.87**
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	**0.95**

Michel Doumith/Martin Day

Major errors

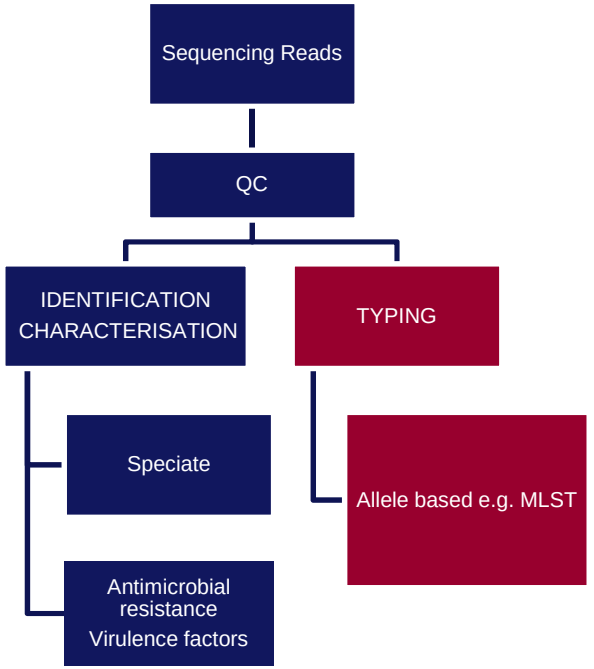
Very major errors

Public Health England

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

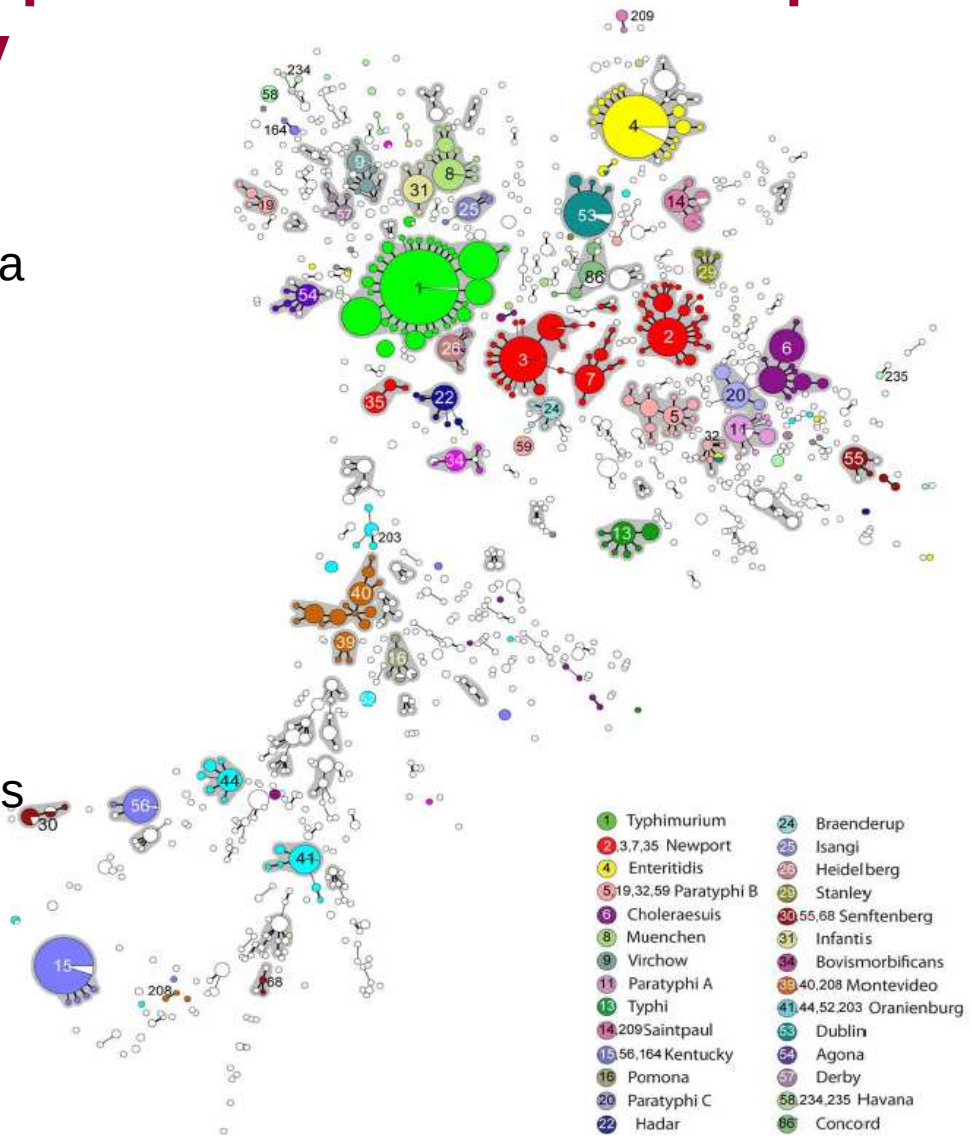


Public Health England

Salmonella population structure is complicated – 21st 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



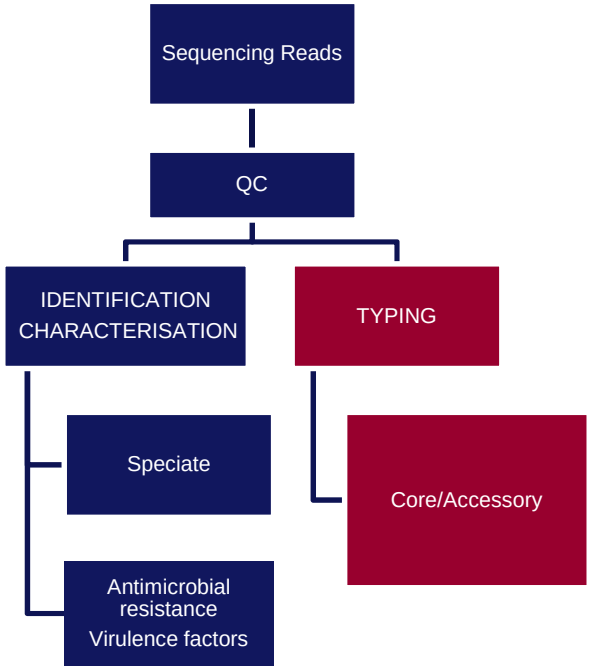
Achtman et al., 2012

Public Health England

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

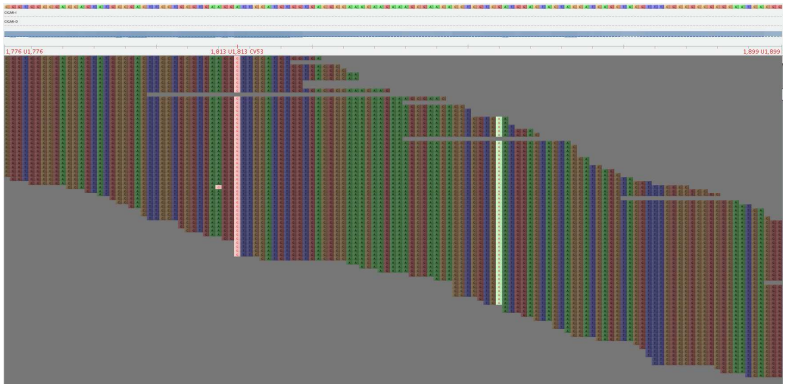


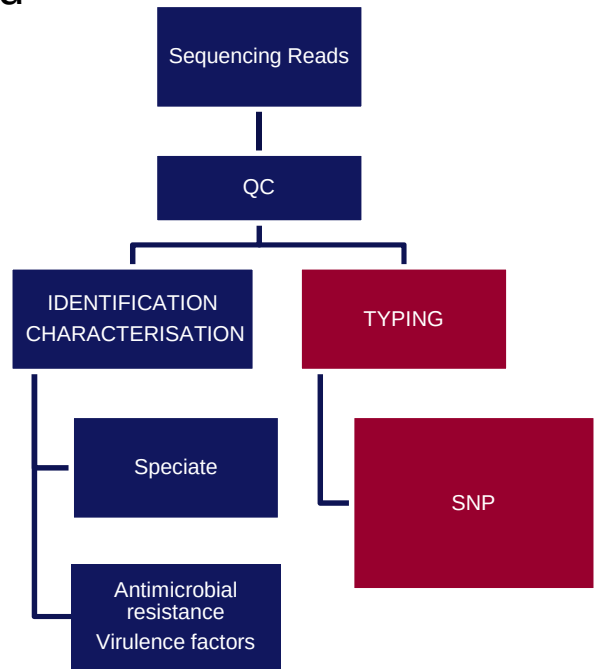
Public Health England

Typing – SNP

SnapperDb

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



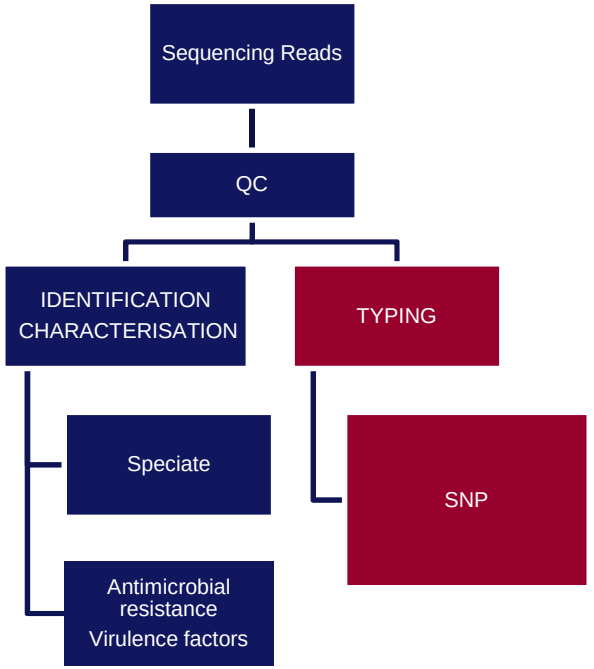


Public Health England

Typing – SNP

SnapperDb - Outputs

- SNP alignments
- Annotated variants

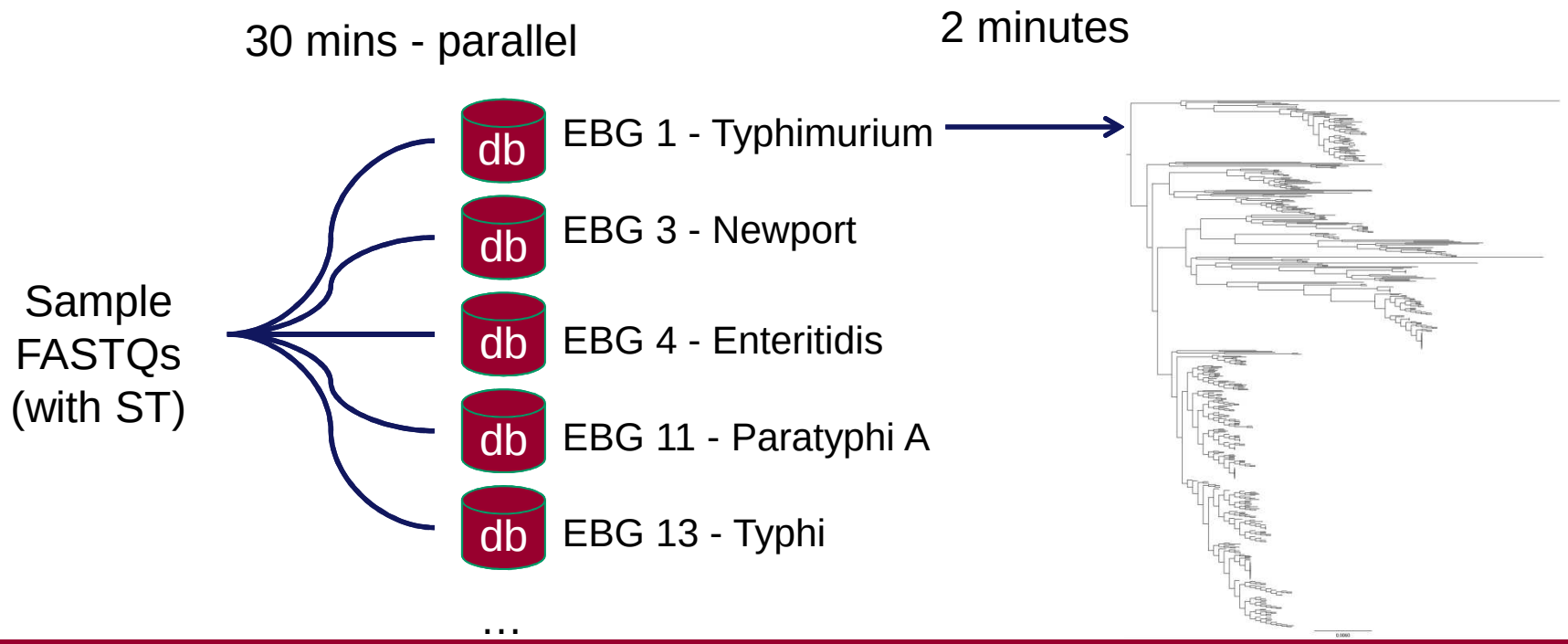


Public Health England

Salmonella SnapperDb

Challenges:

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

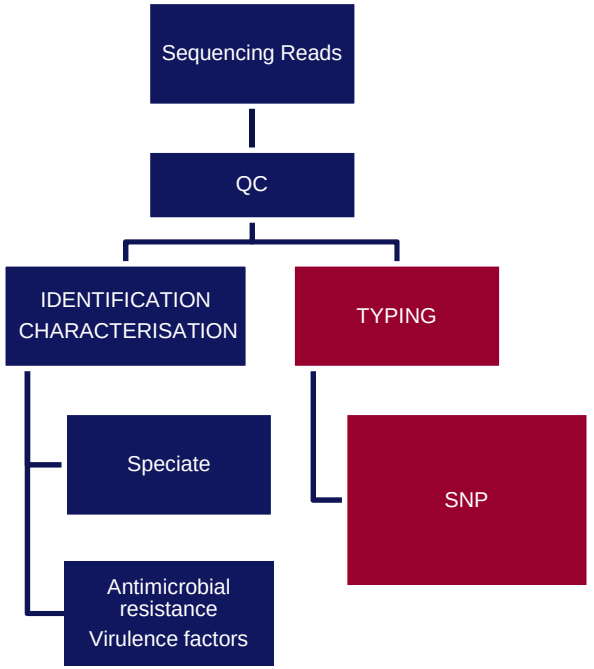


Public Health England

Typing – SNP

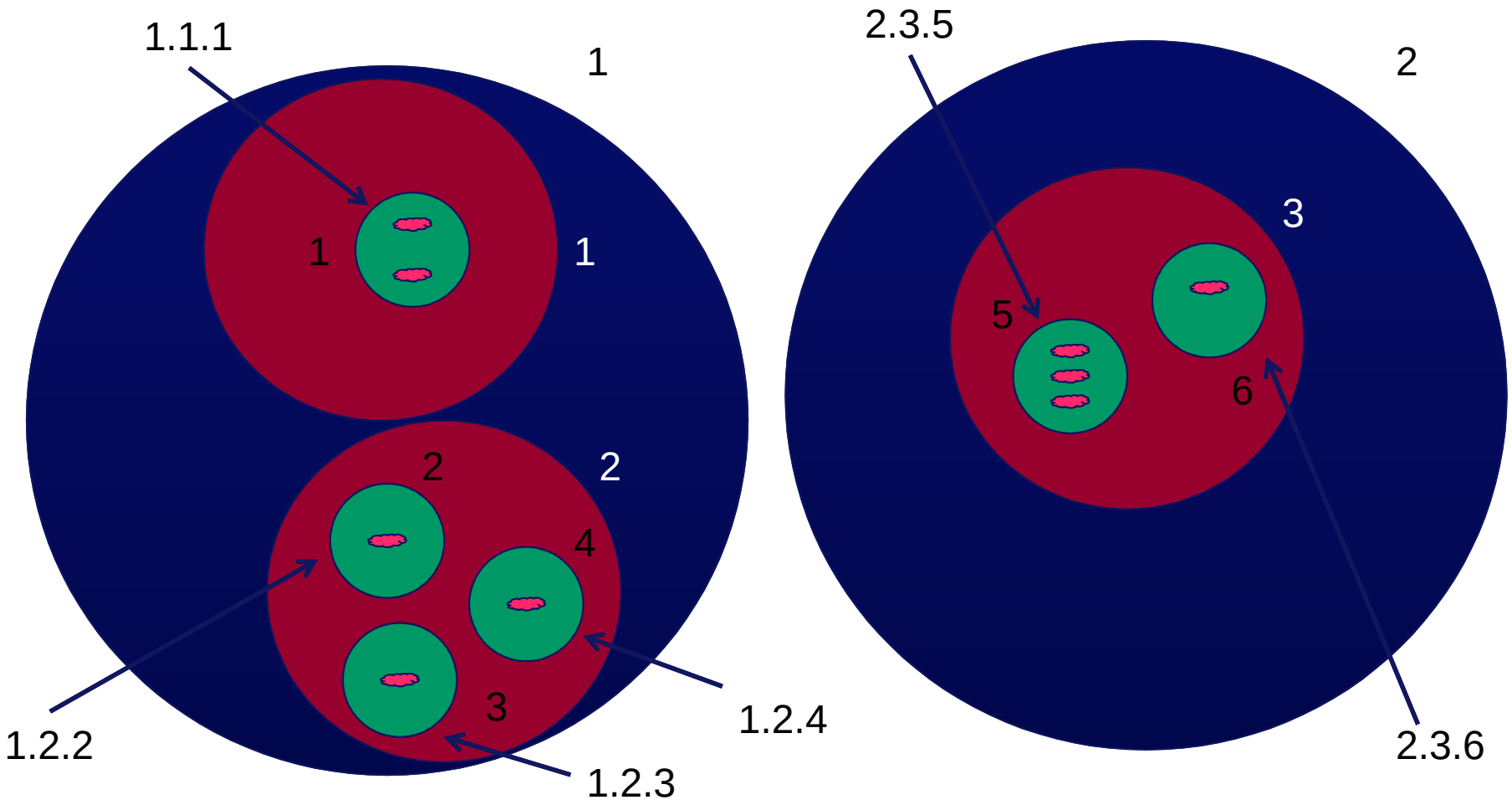
SnapperDb – Clustering

- Maintain a SNP distance matrix
- Hierarchical Clustering



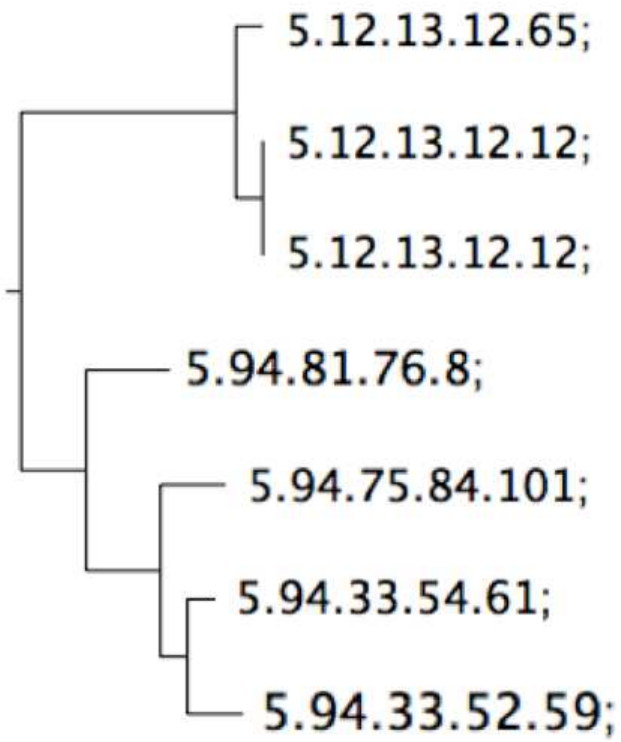
Public Health England

SNP Address



Public Health England

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
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)

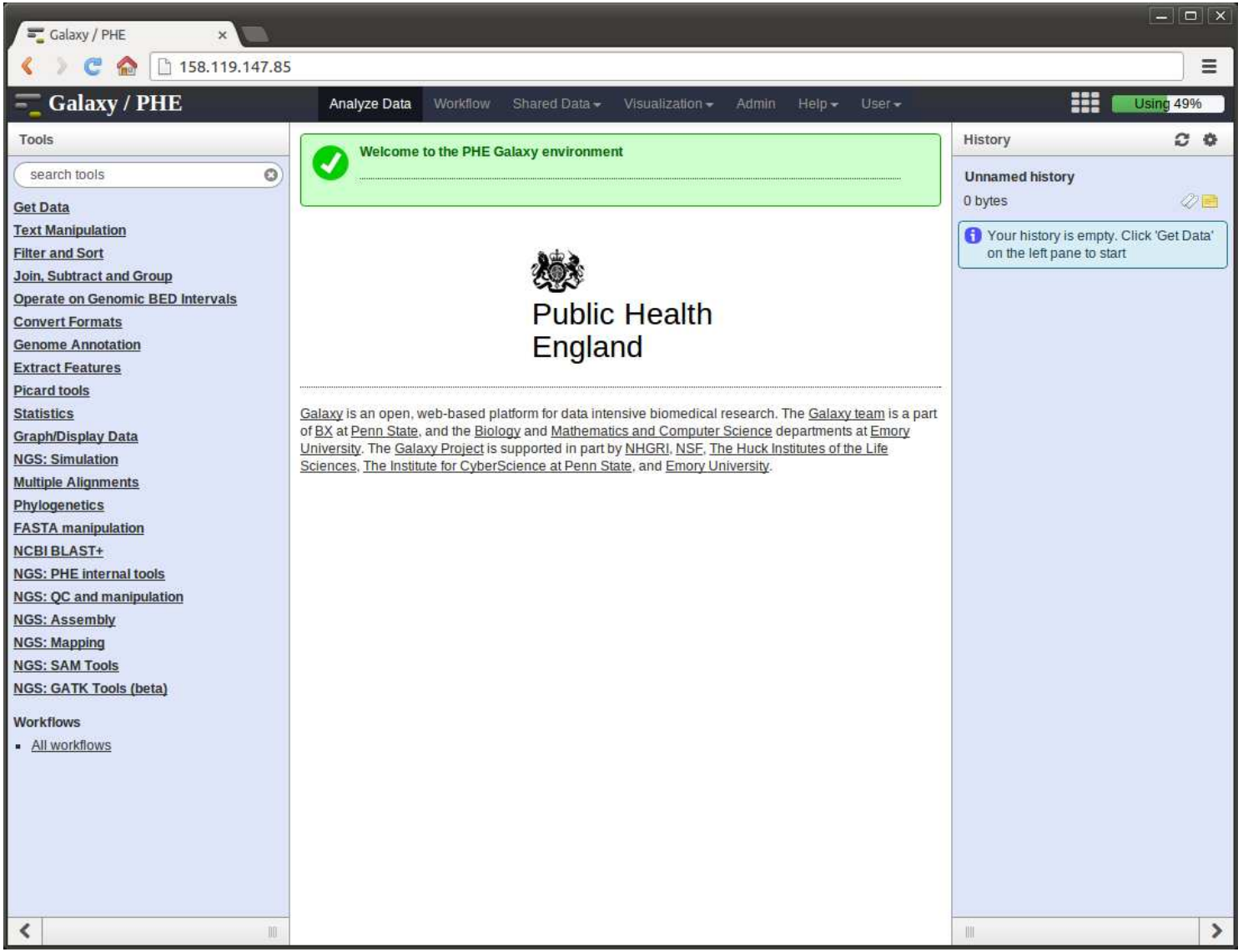
Submission:

Registration date: 19-May-2014

Public Health England

NCBI BioProject accession: PRJNA248064

Public Health England



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

Public Health England

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