# Identification of *Salmonella* for Public Health Surveillance using Whole Genome Sequencing

Philip M Ashton, Satheesh Nair, Tansy M Peters, Janet Bale, David G Powell, Anaïs Painset, Rediat Tewolde, Ulf Schaefer, Claire Jenkins, Timothy J Dallman, Elizabeth M de Pinna, Kathie A Grant

In April 2015, Public Health England implemented whole genome sequencing (WGS) as a routine typing tool for public health surveillance of *Salmonella*, adopting a multilocus sequence typing (MLST) approach as a replacement for traditional serotyping. The WGS derived sequence type (ST) was compared to the phenotypic serotype for 6887 isolates of *S. enterica* subspecies I, and of these, 6616 (96%) were concordant. Of the 4% (n=271) of isolates of subspecies I exhibiting a mismatch, 119 were due to a process error in the laboratory, 26 were likely caused by the serotype designation in the MLST database being incorrect and 126 occurred when two different serovars belonged to the same ST. The population structure of *S. enterica* subspecies II-IV differs markedly from that of subspecies I and, based on current data, defining the serovar from the clonal complex may be less appropriate for the classification of this group. Novel sequence types that were not present in the MLST database were identified in 8.6% of the total number of samples tested (including *S. enterica* subspecies I-IV and *S. bongori*) and these 654 isolates belonged to 326 novel STs. For *S. enterica* subspecies I, WGS MLST derived serotyping is a high throughput, accurate, robust, reliable typing method, well suited to routine public health surveillance. The combined output of ST and serovar supports the maintenance of traditional serovar nomenclature while providing additional insight on the true phylogenetic relationship between isolates.

# Identification of *Salmonella* for public health surveillance using whole genome sequencing

Philip M. Ashton, Satheesh Nair, Tansy M. Peters, Janet Bale, Ulf Schaefer, David G. Powell, Anais Painset, Rediat Tewolde, Claire Jenkins, Timothy J. Dallman*, Elizabeth M. de Pinna, Kathie A. Grant and the *Salmonella* Whole Genome Sequencing Implementation Group

Gastrointestinal Bacteria Reference Unit, Public Health England, 61 Colindale Ave, London, NW9 5HT

Corresponding author
Timothy J. Dallman, Gastrointestinal Bacteria Reference Unit, Public Health England, 61 Colindale Ave, London, NW9 5HT
Email: tim.dallman@phe.gov.uk

Running title: WGS derived MLST serotyping for *Salmonella*

**Abstract**

In April 2015, Public Health England implemented whole genome sequencing (WGS) as a routine typing tool for public health surveillance of *Salmonella*, adopting a multilocus sequence typing (MLST) approach as a replacement for traditional serotyping. The WGS derived sequence type (ST) was compared to the phenotypic serotype for 6887 isolates of *S. enterica* subspecies I, and of these, 6616 (96%) were concordant. Of the 4% (n=271) of isolates of subspecies I exhibiting a mismatch, 119 were due to a process error in the laboratory, 26 were likely caused by the serotype designation in the MLST database being incorrect and 126 occurred when two different serovars belonged to the same ST. The population structure of *S. enterica* subspecies II-IV differs markedly from that of subspecies I and, based on current data, defining the serovar from the clonal complex may be less appropriate for the classification of this group. Novel sequence types that were not present in the MLST database were identified in 8.6% of the total number of samples tested (including *S. enterica* subspecies I-IV and *S. bongori*) and these 654 isolates belonged to 326 novel STs. For *S. enterica* subspecies I, WGS MLST derived serotyping is a high throughput, accurate, robust, reliable typing method, well suited to routine public health surveillance. The combined output of ST and serovar supports the maintenance of traditional serovar nomenclature while providing additional insight on the true phylogenetic relationship between isolates.

43 **Introduction**

44 The *Salmonella*e are major human pathogens and represent a significant global public health issue
45 causing morbidity and mortality resulting in a high social and economic burden worldwide (*Majowicz et*
46 *al., 2010*). The genus consists of 2 species; *Salmonella enterica* and *S. bongori*. There are six subspecies
47 of *S. enterica* differentiated by biochemical variations, namely subspecies *enterica* (I), *salamae* (II),
48 *arizonae* (IIIa), *diarizonae* (IIIb), *houtenae* (IV) and *indica* (VI) (*Threlfall et al. 1999*). Subspecies I, *S.*
49 *enterica* subsp. *enterica* cause 99% of human and animal infections. The two main pathologies
50 associated with *S. enterica* are gastroenteritis and typhoidal disease. The typhoidal *Salmonella*e include
51 *S*. Typhi and *S*. Paratyphi A, B and C. They are host restricted, monophyletic, rarely undergo
52 recombination events and exhibit convergent evolution driven by genome degradation (*Wain et al.*
53 *2015*). The majority of gastroenteritis in the UK is caused by the host generalist serovars, such as *S*.
54 Typhimurium and *S*. Enteritidis, and host adapted serovars that are adapted to a specific animal
55 reservoir but can infect man and include *S*. Dublin, *S*. Gallinarum *S*. Choleraesuis, and *S*. Bovismorbificans
56 (*Langridge et al. 2015*).
57
58 Approximately 8,000 isolates are referred to the *Salmonella* Reference Service (SRS) at Public Health
59 England (PHE) each year from local and regional hospital laboratories. In April 2015, PHE implemented
60 whole genome sequencing (WGS) as the routine typing tool for public health surveillance of *Salmonella*
61 infections. Prior to April 2015, presumptive *Salmonella* isolates referred to SRS were speciated and sub-
62 speciated using PCR *(Hopkins et al. 2009, 2011*) and grouped into serovars as described in the White-
63 Kauffman-Le Minor scheme (*Grimont & Weill 2007, Guibourdenche et al. 2010, Issenhuth-Jeanjean et al.*
64 *2014*). This methodology is based on reactions of rabbit antisera to the lipopolysaccharide (O antigen
65 encoded by *rfb* genes) and flagellar antigens (phases 1 and 2 of H antigen encoded by *fli*C and *flj*B). The
66 scheme utilises this phenotypic variation, expressed as an antigenic formulae, to divide *Salmonella* into
67 more than 2600 serovars. Epidemiological investigations of *Salmonella* infecting humans and animals
68 have relied on serotyping for over 70 years; national and international governmental agencies base
69 guidelines and regulations on the serotyping method and the use of this nomenclature is a globally
70 recognised form of communication (*Swaminathan et al. 2009, EFSA 2010*). Furthermore, serovars have
71 often been shown to correlate with host range and disease sequelae (*Gordon et al. 2011, Wain et al.*
72 *2015, Langridge et al. 2015*).
73

There are, however, a number of issues with the serotyping approach; specifically, the expense and expertise required to produce the antisera and , furthermore, serotyping does not reflect the genetic relatedness between serovars, nor does it provide an evolutionary perspective. Alternative molecular serotyping methods have been described previously including Pulsed-field gel electrophoresis, ribotyping, repetitive extragenic palindromic sequence-based PCR (rep-PCR) and combined PCR- and sequencing-based approach that directly targets O- and H-antigen-encoding genes (*Ranieri et al. 2013, Shi et al. 2015*). In 2012, Achtman *et al.* proposed a sequenced based approach, multilocus sequence typing (MLST), based on the sequences of multiple house-keeping genes. Isolates that possess identical alleles for the seven gene fragments analysed are assigned a common sequence type (ST) and related STs from clonal complexes are termed e-Burst Groups (eBGs). They showed that ST and eBGs strongly correlated with serovar and so utilising this approach would facilitate backward compatibility with historical data, minimise disruption for reference laboratory service users and facilitate data exchange with other colleagues in the field.

Advances in whole genome sequencing (WGS) methodologies have resulted in the ability to perform high throughput sequencing of bacterial genomes at low cost making WGS an economically viable alternative to traditional typing methods for public health surveillance and outbreak detection (*Koser et al. 2012*). Whilst WGS provides the opportunity to resolve bacterial strains to the single nucleotide resolution needed for identifying cases linked to a common source of infection (*Dallman et al. 2015*), grouping isolates into higher taxonomical clones (e.g. those defined by serotyping) is an important step. The decision to adopt WGS as a routine typing method at PHE provided the opportunity to review our approach to typing *Salmonella* and to implement the MLST approach in parallel with WGS.

The aim of this study was to evaluate MLST, as derived from WGS data, as a replacement for conventional serotyping of *Salmonella* for routine public health surveillance and to provide insight into the genetic population structure of all *Salmonella* species in England and Wales during a 12 month period.

**Methods**

106 **Bacterial strains**

107 All isolates (n=7465) of *Salmonella* from human cases of gastrointestinal disease submitted to SRS from
108 local and regional hospital laboratories in England & Wales between 1st April 2014 and 31st March 2015
109 were sequenced in parallel with phenotypic serotyping (Supplementary Table). Of these, 7338 were
110 identified as subspecies I and included 263 different serovars. The ten most common serovars in this
111 dataset were Enteriditis (2310), Typhimurium (1407), Infantis (184), Typhi (184), Newport (173), Virchow
112 (162), Kentucky (160), Stanley (146), Paratyphi A (135) and Java (99). One hundred and twenty seven
113 isolates were identified as subspecies II-IV (*S. enterica* subspecies *salamae* n=28; *S. enterica* subspecies
114 *arizonae* n=25; *S. enterica* subspecies *diarizonae* n=49; *S. enterica* subspecies *houtenae* n=20) and there
115 was one isolate of *S. bongori*. No isolates belonging to subspecies VI (*S. enterica* subspecies *indica*) were
116 submitted to SRS during the study period.
117
118 **DNA extraction for WGS**
119 DNA extraction of *Salmonella* isolates was carried out using a modified protocol of the Qiasymphony
120 DSP DNA midi kit (Qiagen). In brief, 0.7 ml of overnight *Salmonella* culture in a 96 deep well plate was
121 harvested. Bacterial cells were pre-lysed in 220 µl of ATL buffer (Qiagen) and 20 µl Proteinase K
122 (Qiagen), and incubated shaking for 30 mins at 56°C. Four µl of RNase at 100 mg/ml (Qiagen) was added
123 to the lysed cells and re-incubated for a further 15 mins at 37°C. This step increases the purity of the
124 DNA for further downstream sequencing. Extraction of DNA from the treated cells was performed on
125 the Qiasymphony SP platform (Qiagen) and eluted in 100 µl of water. DNA concentration using the
126 GloMax system (Promega) was determined for the following sequencing steps.
127
128 **DNA sequencing**
129 Extracted DNA was then processed using the NexteraXT sample preparation method and sequenced
130 with a standard 2x101 base protocol on a HiSeq 2500 Instrument in fast mode (Illumina, San Diego).
131
132 **Bioinformatics Workflow**
133 FASTQ reads were quality trimmed using Trimomatic (*Bolger et al. 2014*) with bases removed from the
134 trailing end that fell below a PHRED score of 30. If the read length post trimming was less than 50bp the
135 read and its pair were discarded. The PHE KmerID pipeline (https://github.com/phe-
136 bioinformatics/kmerid) was used to compare the sequenced reads with 1769 published genomes to
137 identify the bacterial species (and *Salmonella* subspecies) and to detect cultures submitted by the local

and regional hospital laboratories that contained more than one bacterial species (mixed cultures). KmerID determines a similarity index between the FASTQ reads and each of the 1769 published reference genomes by calculating the percentage of 18-mers in the reference that are also present in the FASTQs. Only 18-mers that occur at least twice in the FASTQ are considered present. Mixed cultures are detected by comparing the list of similarities between the sample and the references with the similarities of the references to each other, and filtering this comparison for inconsistencies. ST assignment was performed using the Metric Orientated Sequence Typer (MOST), a modified version of SRST (Inouye et al., 2012), available from https://github.com/phe-bioinformatics/MOST. The primary difference between SRST and MOST is in the metrics provided around the result, while SRST gives a single score, MOST provides a larger array of metrics to give users more details on the read level associated with their result. Preliminary analysis was undertaken using the MLST database described in Achtman *et al.* (2012). It takes approximately 10-15 minutes to run MOST using a single core on the PHE infrastructure which consists of Intel Xeon CPU E5-2680 0 @ 2.70GHz, 16 cores sharing 125Gb Memory.

For isolates that had novel STs, or a ST but no associated serovar in the Achtman MLST database, the serovar was determined by phenotypic serotyping at PHE. STs and corresponding serovars of isolates serotyped and sequenced during this study were added to a modified version of the Achtman MLST database, held and curated at PHE. These novel STs were assigned a preliminary ST (PST) and an inferred serovar was determined. The PHE MLST database currently holds 7000 strains and 1,200 serovars and is up-dated every three months.

For some STs that contained two serotypes, whole genome SNP phylogenetic analysis was carried out by mapping the strains of interest against a reference genome from within the same sequence type (for ST909 H145100685 was used; for ST49, H143720759 was used), using BWA mem (Li & Durbin, 2009). SNPs were called using GATK2 (DePristo et al., 2011) in unified genotyper mode. Core genome positions that had a high quality SNP (>90% consensus, minimum depth 10x, GQ >= 30, MQ >=30) in at least one strain were extracted and RAxML v8.1.17 phylogenies determined with the gamma model of rate heterogeneity and 100 bootstraps undertaken.

**Results**

Achtman *et al.* (2012) described the population structure of *Salmonella enterica* as monophyletic lineages of STs that have evolved from a single founder node and termed these discrete clusters eBGs. The population structure of all the *Salmonella* species submitted to PHE between April 2014 and March 2015 is illustrated by the minimum spanning tree in Figure 1.

*Salmonella subspecies I*

The ST and corresponding serovar designation obtained from the MLST database were used to compare the WGS derived ST to the phenotypic serotype for 6887 (94%) of 7338 isolates of subspecies 1, and of these, 6616 (96%) had the same result by both methods (Supplementary Table). It was not possible to compare phenotypic serotyping with MLST-based serotyping for 451 (6%) subspecies I isolates because either the phenotypic serotype could not be determined due to an incomplete antigenic structure (*S*. Unnamed) (n=423) or the serovar could not be determined because the ST did not have a designated serotype in the MLST database (n=70). Forty-two isolates were both *S*. Unnamed and had no MLST designated serotype.

For the 423 (5.8%) subspecies I isolates reported as *S*. Unnamed, 318 (90%) were designated a serotype from the WGS derived MLST data. The most common serovars typed in this way included *S*. Typhimurium (118), *S*. Virchow (30), S. Stanley (17), *S*. Enteritidis (16), *S*. Infantis (14) and *S*. Thompson (13). Of the 7338 strains tested, 70 (1%) had no serotype designation in the MLST database, of which 28 (40%) were serotyped phenotypically (Supplementary Table).

*Subspecies I novel sequence types*

Novel sequence types that were not present in the MLST database were identified in 8.6% (n=654) of the strains (Supplementary Table). These 654 isolates belonged to a total of 326 novel STs, designated PST; the modal number of isolates identified per PST was one (Figure 2a). There was no difference in the distribution of number of isolates per PST depending on whether the PST had a known serovar or belonged to an unnamed or ambiguous serotype. The rate at which PSTs were received throughout the year was plotted and revealed a linear relationship ($R^2 = 0.98$, $y = 1.04 * x$, where x = number of days since April 1st 2014) (Figure 2b).

The serovars with the highest number of new PSTs were *S*. Typhimurium (n=9), *S*. Stanley (n=9), *S*. Enteritidis (n=9) and *S*. Newport (n=8), although the majority of these PSTs were single locus variants

(SLVs) of established STs, belonging to these serovars (*S*. Typhimurium 8/9, *S*. Stanley 7/9 and *S*. Newport 7/8). There were also serovars for which a large number of PSTs were identified that were not SLVs of established STs (S. Agama 5/5, *S*. Agbeni 5/5, *S*. Saint-Paul 5/5, *S*. Enteritidis 4/9) which may represent new eBGs that share these serotypes.

*Subspecies I mismatches*

Four percent (n= 271) of the isolates tested exhibited a mismatch between the WGS MLST derived serovar and the phenotypic serotyping results. Of the 271 mismatches, 119 were due to a process error in the laboratory either in the phenotypic serotyping or the DNA extraction part of the WGS pipeline. With respect to the phenotypic serotyping, common errors included mislabelling samples and misinterpreting or incorrectly transcribing the antigenic structure, especially when the antigenic structures were similar. For example, *S*. Agona (I 4, 12:f, g, s:-) and *S.* Derby (I 4, 12:f, g:-). DNA extraction errors were associated with mislabelled samples.

Twenty-six mismatches were potentially caused by the predicted serotype designation in the Achtman MLST database being incorrect which may be attributed to single entries that had been misidentified at the laboratory from which the MLST data was submitted. For example, in the original database ST1499 is represented by one entry identified by the submitter as *S*. Litchfield. Subsequently, phenotypic serotyping at PHE identified this ST as *S.* Bovis-morbificans in more than five isolates. ST1499 belongs to eBG34 which comprises two other STs both associated with S. Bovis-morbificans, indicating that the original entry in the MLST database is likely to be incorrect.

The most common reason for mismatches occurring between the WGS MLST derived serotype and the phenotypic serotype (n=126) occurred when two different serovars belonged to the same eBG and the same ST (see Table 1 and discussed in more detail below).

For 157 of the 271 mismatches, assembly based ST calls were available from Enterobase (http://enterobase.warwick.ac.uk/). We compared the STs called by our mapping pipeline and the assembly based Enterobase pipeline and the same ST was called in 100% of samples.

*Serovars Enteritidis and Dublin*

Of the 2308 isolates of *S.* Enteritidis identified by both phenotypic serotyping and WGS MLST derived serotyping, 2296 belonged to eBG4, including 2200 ST11 and 76 ST183 (Figure 1). There were five additional SLVs of ST11, four of which were novel types. *S*. Gallinarum and *S*. Pulloram can be difficult to distinguish from *S*. Enteritidis (*Thomson et al. 2008*) but neither of these serovars were identified in this study. Serologically, *S*. Dublin ([1],9,12:g,p:-) has a similar antigenic structure to *S.* Enteritidis ([1],9,12:g,m:-), and in Achtman *et al.* (2012), eBG32 (ST74) contained both *S.* Enteritidis and *S*. Dublin. However, in this study both isolates belonging to ST74 eBG32 typed as *S.* Enteriditis. Of the 2308 isolates, 26 belonged to nine new PSTs. The most common was P3147, a previously undescribed SLV of ST11, comprising 16 cases including 10 known to have travelled to Malaysia or Singapore.

*Serovar Typhimurium*

In this study, eBG1 contained 1392 isolates of *S.* Typhimurium and monophasic *S.* Typhimurium (rough and non-motile variants) (*Hopkins et al. 2012*). The monophasic variants also belong to eBG138 (primarily ST 36) and eBG243. In contrast to eBG1 described in Achtman *et al.* (2012), which was represented by a large central ST19 node with at least 27 SLV STs comprising much smaller numbers of strains, eBG1 in the PHE dataset shows a predominance of both ST19 and ST34 and less allelic variation. Only nine SLVs to ST19 were identified including three undesignated STs (Figure 1).

*Serovars Java/Paratyphi B data*

Despite the different disease outcomes associated with *S.* Paratyphi B (most commonly associated with invasive disease and paratyphoid fever) and *S.* Java (most commonly associated with gastroenteritis) it is not possible to differentiate the two serotypes by serotyping alone. *S.* Java and *S.* Paratyphi B are therefore differentiated in the laboratory by their ability to ferment dextrorotatory tartrate (*S.* Java *d*Ta+ and *S.* Paratyphi B *d*Ta-) (*Malorny et al. 2003*).

The 99 isolates identified by both phenotypic serotyping and WGS MLST derived serotyping as *S.* Java, belonged to a diverse range of eBGs, STs and PSTs (Table 2 and Figure 1). Two of these 99 isolates (marked with * in Table 2) belonged to ST86 and the predicted serotype from the MLST database was *S.* Paratyphi B. One of these isolates was from a blood culture (associated with invasive disease) and, therefore, likely to have been misidentified phenotypically. All 12 isolates identified as *S.* Paratyphi B phenotypically, were identified as *S.* Paratyphi B ST86 by WGS MLST.

*Subspecies II-IV and S. bongori*

Isolates from subspecies II, III, IV, VI and *S. bongori* were not well represented in Achtman's MLST database and thus the majority of isolates from these sub-species sequenced in this study did not belong to a previously designated eBG or ST. The population structure of the 127 non-subspecies I isolates differs markedly from that of *Salmonella enterica* (subspecies 1) (Figure 1) and shows some similarity to the population structure of lineage 3 in being a connected network of STs.

Sixteen of the 28 isolates belonging to subspecies II were previously designated *S.* Unnamed and the 28 strains belonged to 20 different STs. There were 25 isolates classed as subspecies IIIa (belonging to 10 different STs) and 49 in subspecies IIIb (belonging to 27 different STs). Of the 20 isolates identified as subspecies IV, 10 were designated *S.* Wassenaar (P3029) by phenotypic serotyping and the 20 isolates belonged to five different STs. All isolates of subspecies II-IV and *S. bongori* were correctly speciated using the k-mer ID approach.

*Population structure*

As highlighted by Achtman *et al.* (2012), the majority of isolates in the dataset belong to eBGs that have a one-to-one relationship with a specific serovar including *S*. Typhi, *S*. Paratyphi A and *S*. Heidelberg. In this study, of the serovars comprising more than 25 isolates, there were 17 serovar specific eBGs and 10 examples of a single serovar being associated with multiple eBGs (Figure 1). There were at least six examples of more than one serovar belonging to the same eBG but different STs, for example *S.* Hadar (ST33) and *S.* Kottbus (ST582) both belong to eBG22 and *S.* Bredeney (ST306) and *S.* Schwarzengrund (ST96) both belong to eBG33 (Supplementary Table).

There were seven examples where two serovars belonged to the same eBG and the same ST (Table 1). In all of these examples, the antigenic structures of the two serovars were similar with only one antigen differentiating the two serovars. Further analysis was carried out on two examples to determine whether this difference in antigenic structure represented a true difference in strain relatedness or a random change that is not reflected in phylogeny (for example, the insertion of phage encoded antigen). The analysis showed that the change in antigenic structure in *S.* Richmond (I 6,7:y:1,2) and *S.* Bareilly (I 6, 7:y:1,5), both ST 909, and in *S.* Saintpaul (I 4,5,12: e,h: 1,2) and *S.* Haifa (I 4,5,12: z,10: 1,2), both ST49, reflected a true phylogenetic difference (Figures 3a and 3b).

The same higher strata population structure referred to as lineage 3 for *S. enterica* subspecies I, as described by Achtman *et al.* (2012), was observed in this dataset (Figure 4). Genomes of these *Salmonella*e are in constant flux and homologous recombination among unrelated eBGs is frequent (*Achtman et al. 2012, Didelot et al. 2011*). Serovars in this lineage mainly consists of multiple eBGs and are polyphyletic by nature. Achtman *et al.* (2012) suggested that the population structure of lineage 3 does not comprise of independent startbursts, as observed with other serovars of subspecies I, but rather a connected network (Figure 4). The five most common examples of this in the current study, were *S.* Oranienburg, *S.* Montevideo, *S.* Chester, *S.* Poona and *S.* Bredeney (Figure 4 and Supplementary Table). These five serovars are not represented in the top 10 serovars submitted to SRS during this surveillance period.

*K-mer identification*

There were 249 cultures submitted to SRS by the local hospital and regional laboratories for *Salmonella* typing that were a mix of *Salmonella* and non-*Salmonella* species. These were identified by the k-mer identification step and included 138 *Escherichia coli*, 40 *Morganella morganii*, 11 *Citrobacter species* and four *Escherichia albertii*.

**Discussion**

In their seminal 2012 paper Achtman and colleagues argued convincingly for replacing serotyping with a MLST approach based on genetic population groupings for typing *S. enterica* (Achtman *et al. 2012*). The key aspects of this approach that led PHE to adopt this strategy were (i) the robustness of the population structure as defined by the natural eBG clusters (ii) the fact that eBG designation provides an accurate representation of strain relatedness and (iii) that this approach lends itself to automation. At the same time, it was necessary for PHE to maintain serovar nomenclature in order to facilitate data exchange with other colleagues in the field and maintain backward compatibility with historical data. It was suggested that by using the MLST approach to infer serovar, and by reporting both inferred serovar and ST, it would be possible to utilise the advantages of both methods and implement a state-of-the-art typing system while keeping disruption for reference laboratory service users to a minimum.

The PHE dataset of 6887 subspecies I isolates that were serotyped using both traditional phenotypic methods and a derived serotype based on MLST data extracted from the genome during a 12 month time frame, provided further evidence of the robustness of the ST/eBG approach to typing. The 96%

concordance between the two techniques in a reference laboratory setting is evidence of the validity and suitability of this approach. There were 451 isolates that had to be excluded from the comparison because both types of data (phenotypic and genotypic) were not available. Of these, for 94% of the isolates, it was the phenotypic serotype that could not be determined indicating that WGS MLST derived serotyping is more robust.

The PHE dataset included single serovars associated with multiple eBG, for example *S.* Typhimurium and *S.* Newport (*Sangal et al. 2010, Achtman et al. 2012*) and multiple serovars belonging to the same eBG but with different STs, for example *S.* Java (ST43) and *S.* Paratyphi (ST86) both belong to eBG5 (*Achtman et al. 2012*). In both these scenarios, the correct serovar was determined from the MLST WGS data and the combination of serovar and ST/eBG provided insight into the true phylogenetic relationship between isolates. This data clearly supports Achtman and colleagues argument that eBG and ST designation provides a more accurate representation of strain relatedness than the traditional serovar designation. The phenomenon of multiple serovars belonging to the same ST (for example *S.* Richmond/*S.* Bareilly and *S.* Haifa/*S.* Saintpaul) was a rare but important example of serotyping providing a higher level of strain discrimination within a ST. These strains could be differentiated *in silico* using a tool to infer serovar from the genes that determine antigenic struture, such as seqsero (*Zhang et al. 2015*).

Despite the implementation of WGS, a limited phenotypic serotyping facility continues to be maintained at PHE in order to serotype isolates that cannot be matched to a serovar; either because the ST in the MLST database has no serovar designation or the ST is a novel type. Additionally, it ensures that we maintain the ability to perform the standard reference method for serotyping *Salmonella*. The PHE MLST database is regularly up-dated to include STs recently matched to a serotype by linking the ST to PHE phenotypic serotyping data and novel PSTs. This approach was adopted because at the time of analysis, the Achtman MLST database was not accepting submissions generated by WGS. There was no decrease in the rate at which PSTs were observed during the 12 month study period and the majority of PSTs were only sampled once in that time frame. Many PSTs were SLVs of known STs, indicating that we have not yet sampled the full diversity of known eBGs. New PSTs, not part of any previously identified eBG, were also observed and further diversity was found within *S. enterica* subspecies II-IV and the lineage 3 population. This suggests that there is a large amount of previously unidentified diversity within the species *Salmonellae* associated with both domestically acquired and travel related gastrointestinal disease in human cases resident in England and Wales.

362

363 Isolates exhibiting monophasic properties that could not be fully serotyped phenotypically because they
364 had an incomplete antigenic structure were matched to a ST derived serotype. The monophasic variants
365 in this study mainly belonged to eBG1, eBG138 and eBG243 and previous studies have also shown that
366 monophasic variants of *S.* Typhimurium have emerged as a result of multiple independent genetic
367 events (*Soyer et al. 2009, Switt et al. 2009, Tennant et al. 2010*). Strains with monophasic properties are
368 reportable to European Centre for Disease Prevention and Control (ECDC) but cannot be determined
369 using the ST approach. Alternative strategies for determining monophasic characteristics by PCR are
370 available (*Prendergast et al. 2013*) and methods for extracting this information from the genome
371 sequencing data have been developed at PHE (Personal communication: Philip Ashton & Anna Lewis,
372 publication in preparation).

373

374 In contrast to *S.* Typhurmurium, where ST could not be used to determine monophasic characteristics, in
375 this study ST was able to differentiate the complex relationship between *S.* Java (Hazard Group (HG) 2
376 organism) and *S.* Paratyphi B (HG3) with the latter belonging to either ST42 or ST86. If this ST
377 designation proves to be robust, MLST will facilitate the diagnosis of invasive disease and life
378 threatening paratyphoid fever.

379

380 The MLST derived serovar correlated well with the traditional serovar designation and demonstrated
381 many advantages over traditional phenotypic serotyping. Monophasic strains with incomplete antigenic
382 structures were accurately assigned to serotypes. Phenotypic serotyping errors, such as misinterpreting
383 or incorrectly transcribing the antigenic structure, were avoided. Novel types were identified, confirmed
384 and given a PST designation. Finally, this approach lends itself to automation and rapid, high-throughput
385 processing.

386

387 Two main issues arose during the evaluation of the MLST approach: (i) a number of STs did not have a
388 serovar designation in the MLST database (including subspecies II to IV) and (ii) the unexpectedly large
389 number of novel STs identified. Traditional phenotypic serotyping was required to type these isolates
390 and the MLST database was modified and up-dated to incorporate the new data. Clearly, as we move
391 forward the PHE MLST database will be constantly evolving and this data will be shared with colleagues
392 in the field via existing MLST databases and their WGS compliant successors e.g. EnteroBase & BIGSdb.
393 While it is difficult to draw conclusions based on our small sample size, MLST may not currently be an

394 appropriate tool for the classification of *Salmonella* sub-species II-IV, due to the lack of a discrete
395 population structure of EBGs. However, non-subspecies I isolates which are mainly adapted to cold
396 blooded animals and/or reptiles contributed to less than 1.7% of the workload during the time frame of
397 the study. Although MLST approach is generally more discriminatory than serotyping, it does not always
398 provide the fine resolution required for public health surveillance. Further analysis based on single
399 nucleotide polymorphisms in the core genome compared to a type strain representing the most
400 common eBGs is performed for outbreak detection and investigation (*Ashton et al. 2014*).
401
402 In conclusion, serotyping inferred from 7-gene MLST results derived from WGS data is an accurate,
403 robust, reliable, high throughput typing method that is well suited to routine public health surveillance
404 of *Salmonella*. This approach supports the maintenance of traditional serovar nomenclature and
405 provides further insight on the true evolutionary relationship between isolates, as well as a framework
406 for fine level typing within eBGs for surveillance, outbreak detection and source attribution.
407
408
409 **Acknowledgements**
410 We would like to thank all the members of the *Salmonella Whole Genome Sequencing Implementation*
411 *Group* including Steve Connell, Anna Lewis, Andy Levy, Clare Maguire, Clare Wend-Hansen, Martin Day,
412 James Rogers, Siham Ibrahim, Arlene Barcenilla, Vineet Patel, Kiran Jayan, Anthony Underwood,
413 Catherine Arnold and Ian Harrison.
414
415 **Data Deposition**
416 All data from the *Salmonella* surveillance project are deposited in the BioProject of the SRA
417 PRJNA248792. A per sample listing of the Sequence Read Archive accessions is available in
418 Supplementary table 1.

## References

**Achtman M, Wain J, Weill FX, Nair S, Zhou Z, Sangal V, Krauland MG, Hale JL, Harbottle H, Uesbeck A, Dougan G, Harrison LH, Brisse S, S. Enterica MLST Study Group. 2012.** Multilocus sequence typing as a replacement for serotyping in *Salmonella* enterica. *PLoS Pathogens* 8(6):e1002776.

**Ashton PM, Peters T, Ameh L, McAleer R, Petrie S, Nair S, Muscat I, de Pinna E, Dallman T. 2015.** Whole Genome Sequencing for the Retrospective Investigation of an Outbreak of Salmonella Typhimurium DT 8. *PLoS Curr* 10:7. doi: 10.1371/currents.outbreaks.2c05a47d292f376afc5a6fcdd8a7a3b6.

**Bolger AM, Lohse M, Usadel B. 2015**. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014:1-7. doi:10.1093/bioinformatics/btu170.

**Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, Petrovska L, Ellis RJ, Elson R, Underwood A, Green J, Hanage WP, Jenkins C, Grant K, Wain J. 2015**. Whole-genome sequencing for national surveillance of Shiga Toxin-producing *Escherichia coli* O157. *Clinical Infectious Diseases* 61(3):305-12.

**DePristo MA, Banks E, Poplin RE, et al. 2011.** A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5):491-498.

**Didelot X, Bowden R, Street T, Golubchik T, Spencer C, McVean G, Sangal V, Anjum MF, Achtman M, Falush D, Donnelly P. 2011.** Recombination and population structure in *Salmonella enterica*. *PLoS Pathogens* 7:e1002191

**European Food Safety Authority. 2010.** Scientific opinion on monitoring and assessment of public health risk of "*Salmonella* Typhimurium" strains. *EFSA Journal* 8.

**Grimont PAD, Weill FX.** 2007. Antigenic Formulae of the *Salmonella* Serovars (ninth ed.)WHO Collaborating Center for Reference and Research on Salmonella, Institut Pasteur, Paris (2007) http://www.pasteur.fr/sante/clre/cadrecnr/salmoms/WKLM_En.pdf

**Gordon MA. 2011.** Invasive nontyphoidal *Salmonella* disease: epidemiology, pathogenesis and diagnosis. *Current Opinion Infectious Diseases* 24:484-489.

**Guibourdenche M, Roggentin P, Mikoleit M, Fields PI, Bockemuhl J, Grimont PA, Weill FX. 2010.** Supplement 2003-2007 (No. 47) to the White-Kauffmann-Le Minor scheme. *Research in Microbiology* 161:26-29.

**Hopkins KL, Lawson AJ, Connell S, Peters TM, de Pinna E. 2011.** A novel real-time polymerase chain reaction for identification of *Salmonella enterica* subspecies enterica. *Diagnostic Microbiology & Infectious Disease* 70(2):278-80.

**Hopkins KL, Peters TM, Lawson AJ, Owen RJ. 2009.** Rapid identification of *Salmonella enterica* subsp. arizonae and S. enterica subsp. diarizonae by real-time polymerase chain reaction. *Diagnostic Microbiology & Infectious Disease* 64(4):452-4.

**Hopkins KL, de Pinna E, Wain J. 2012.** Prevalence of *Salmonella enterica* serovar 4,[5],12:i:- in England and Wales, 2010. *Euro Surveillance* 17(37). pii: 20275

**Inouye M, Conway TC, Zobel J, Holt KE. Short read sequence typing (SRST): multi-locus sequence types from short reads. 2012.** BMC Genomics;13(1):338. doi:10.1186/1471-2164-13-338.

**Issenhuth-Jeanjean S, Roggentin P, Mikoleit M, Guibourdenche M, de Pinna E, Nair S, Fields PI, Weill FX. 2014.** Supplement 2008-2010 (no. 48) to the White-Kauffmann-Le Minor scheme. *Research in Microbiology* 165(7):526-30.

**Koser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, Holden MT, Dougan G, Bentley SD, Parkhill J, Peacock SJ. 2012.** Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathogens* 8:e1002824.

**Langridge GC, Fookes M, Connor TR, Feltwell T, Feasey N, Parsons BN, Seth-Smith HM, Barquist L, Stedman A, Humphrey T, Wigley P, Peters SE, Maskell DJ, Corander J, Chabalgoity JA, Barrow P, Parkhill J, Dougan G, Thomson NR. 2015.** Patterns of genome evolution that have accompanied host adaptation in *Salmonella*. *Proceedings of the National Academy of Sciences USA* 112(3):863-8.

**Li H, Durbin R. 2009.** Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25(14):1754–60.

**Majowicz SE, Musto J, Scallan E, Angulo FJ, Kirk M, O'Brien SJ, Jones TF, Fazil A, Hoekstra RM; International Collaboration on Enteric Disease 'Burden of Illness' Studies. 2010.** The global burden of nontyphoidal *Salmonella* gastroenteritis. *Clinical Infectious Diseases* (6):882-9.

**Malorny B, Bunge C, Helmuth R. 2003.** Discrimination of d-tartrate-fermenting and -nonfermenting *Salmonella enterica* subsp. enterica isolates by genotypic and phenotypic methods. *Journal of Clinical Microbiology* 41(9):4292-7.

**Prager R, Rabsch W, Streckel W, Voigt W, Tietze E, Tschäpe H. 2003.** Molecular properties of *Salmonella enterica* serotype paratyphi B distinguish between its systemic and its enteric pathovars. *Journal of Clinical Microbiology* 41(9):4270-8.

**Prendergast DM, Hand D, Ní Ghallchóir E, McCabe E, Fanning S, Griffin M, Egan J, Gutierrez M. 2013.** A multiplex real-time PCR assay for the identification and differentiation of *Salmonella enterica* serovar Typhimurium and monophasic serovar 4,[5],12:i:-. *International Journal Food Microbiology* 166(1):48-53. doi: 10.1016/j.ijfoodmicro.2013.05.031.

**Ranieri ML, Shi C, Moreno Switt AI, den Bakker HC, Wiedmann M. 2013.** Comparison of typing methods with a new procedure based on sequence characterization for *Salmonella* serovar prediction. *Journal of Clinical Microbiology* 51(6):1786-97. doi: 10.1128/JCM.03201-12.

**Sangal V, Harbottle H, Mazzoni CJ, Helmuth R, Guerra B, Didelot X, Paglietti B, Rabsch W, Brisse S, Weill FX, Roumagnac P, Achtman M. 2010.** Evolution and population structure of *Salmonella enterica* serovar Newport. *Journal of Bacteriology* 192(24):6465-76.

514 **Shi C, Singh P, Ranieri ML, Wiedmann M, Moreno Switt AI. 2015.** Molecular methods for serovar
515 determination of *Salmonella*. *Critical Reviews Microbiology* 41(3):309-25. doi:
516 10.3109/1040841X.2013.837862.
517
518 **Soyer Y, Moreno SA, Davis MA, Maurer J, McDonough PL, Schoonmaker-Bopp DJ, Dumas NB, Root T,**
519 **Warnick LD, Grohn YT, Wiedmann M. 2009.** *Salmonella* enterica serotype 4,5,12:i:-, an emerging
520 *Salmonella* serotype that represents multiple distinct clones. *Journal of Clinical Microbiology* 47: 3546–
521 3556.
522
523 **Swaminathan B, Gerner-Smidt P, Ng LK, Lukinmaa S, Kam KM, Rolando S, Gutierrez EP, Binsztein N.**
524 **2006.** Building PulseNet International: an interconnected system of laboratory networks to facilitate
525 timely public health recognition and response to foodborne disease outbreaks and emerging foodborne
526 diseases. *Foodborne Pathogens and Disease* 3:36-50
527
528 **Switt AI, Soyer Y, Warnick LD, Wiedmann M. 2009.** Emergence, distribution, and molecular and
529 phenotypic characteristics of ***Salmonella* enterica** serotype 4,5,12:i:-. *Foodborne Pathogens and Disease*
530 6(4):407-15. doi: 10.1089/fpd.2008.0213.
531
532 **Tennant SM, Diallo S, Levy H, Livio S, Sow SO, Tapia M, Fields PI, Mikoleit M, Tamboura B, Kotloff KL,**
533 **Nataro JP, Galen JE, Levine MM. 2010.** Identification by PCR of non-typhoidal *Salmonella enterica*
534 serovars associated with invasive infections among febrile patients in Mali. *PLOS Neglected Tropical*
535 *Diseases* 4: e621.
536
537 **Thomson NR, Clayton DJ, Windhorst D, Vernikos G, Davidson S, Churcher C, Quail MA, Stevens M,**
538 **Jones MA, Watson M, Barron A, Layton A, Pickard D, Kingsley RA, Bignell A, Clark L, Harris B, Ormond**
539 **D, Abdellah Z, Brooks K, Cherevach I, Chillingworth T, Woodward J, Norberczak H, Lord A, Arrowsmith**
540 **C, Jagels K, Moule S, Mungall K, Sanders M, Whitehead S, Chabalgoity JA, Maskell D, Humphrey T,**
541 **Roberts M, Barrow PA, Dougan G, Parkhill J. 2008.** Comparative genome analysis of *Salmonella*
542 Enteritidis PT4 and *Salmonella* Gallinarum 287/91 provides insights into evolutionary and host
543 adaptation pathways. *Genome Research* 18(10):1624-37.
544
545 **Threlfall J, Ward L, Old D. 1999.** Changing the nomenclature of *Salmonella*. *Communicable Diseases*
546 *Public Health* 2(3):156-7.
547
548 **Wain J, Hendriksen RS, Mikoleit ML, Keddy KH, Ochiai RL. 2015.** Typhoid fever. *Lancet* 385(9973):1136-
549 45.
550
551 **Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, Fitzgerald C, Fields PI, Deng X.**
552 **2015.** Salmonella serotype determination utilizing high-throughput genome sequencing data. *J Clin*
553 *Microbiol* 53(5):1685-92. doi: 10.1128/JCM.00323-15.

**Tables**

Table 1. Examples where two serovars belonged to the same eBG and the same ST

Table 2. *S.* Java isolates in this study belonged to a diverse range of eBGs and STs associated with *S.* Java whereas *S.* Paratyphi B belonged to ST86 only

559 **Figures**

560 Figure 1. Population structure of all *Salmonella enterica* isolates submitted to PHE from local and
561 regional hospital laboratories in England and Wales between April 2014 and March 2015 (see
562 Supplementary Table for details)

563

564 Figure 2a. Novel, preliminary STs (PST) and the modal number of isolates identified per PST

565

566 Figure 2b. The rate at which PSTs were identified throughout the time frame of the study

567

568 Figure 3a. Phylogenetic relationship of *S.* Richmond and *S.* Bareilly (ST909) (Figure 3a) and *S.* Saintpaul
569 and *S.* Haifa (ST49) (Figure 3b)

570

571 Figure 4. Serovars in lineage 3 mainly consist of multiple eBGs and are polyphyletic by nature

572

## Table 1(on next page)

Examples where two serovars belonged to the same eBG and the same ST

| Serotype | Antigenic structure | ST |
|---|---|---|
| Bareilly<br>Richmond | I 6,7: y: 1,2<br>I 6,7: y: 1,5 | 909 |
| Saintpaul<br>Haifa | I 4,5,12: e,h: 1,2<br>I 4,5,12: z,10: 1,2 | 49 |
| Sandiego<br>Brandenburg | I 4,12: I,v: e,n,z15<br>I 4,12: e,h: e,n,z15 | 20 |
| Uganda<br>Sinstorf | I 3, 10: I,z13: 1,5<br>I 3, 10: I,v: 1,5 | 684 |
| Agona<br>Essen | I 4,12: f,g,s:-<br>I 4,12: f,g,m:- | 13 |
| Napoli<br>Zaiman | I 1,9,12: I,z13: enx<br>I 1,9,12: I,v: enx | P3141 |

1

## Table 2(on next page)

Differences in ST between S. Java and S. Paratyphi B

*S.* Java isolates in this study belonged to a diverse range of eBGs and STs associated with *S.* Java whereas *S.* Paratyphi B belonged to ST86 only

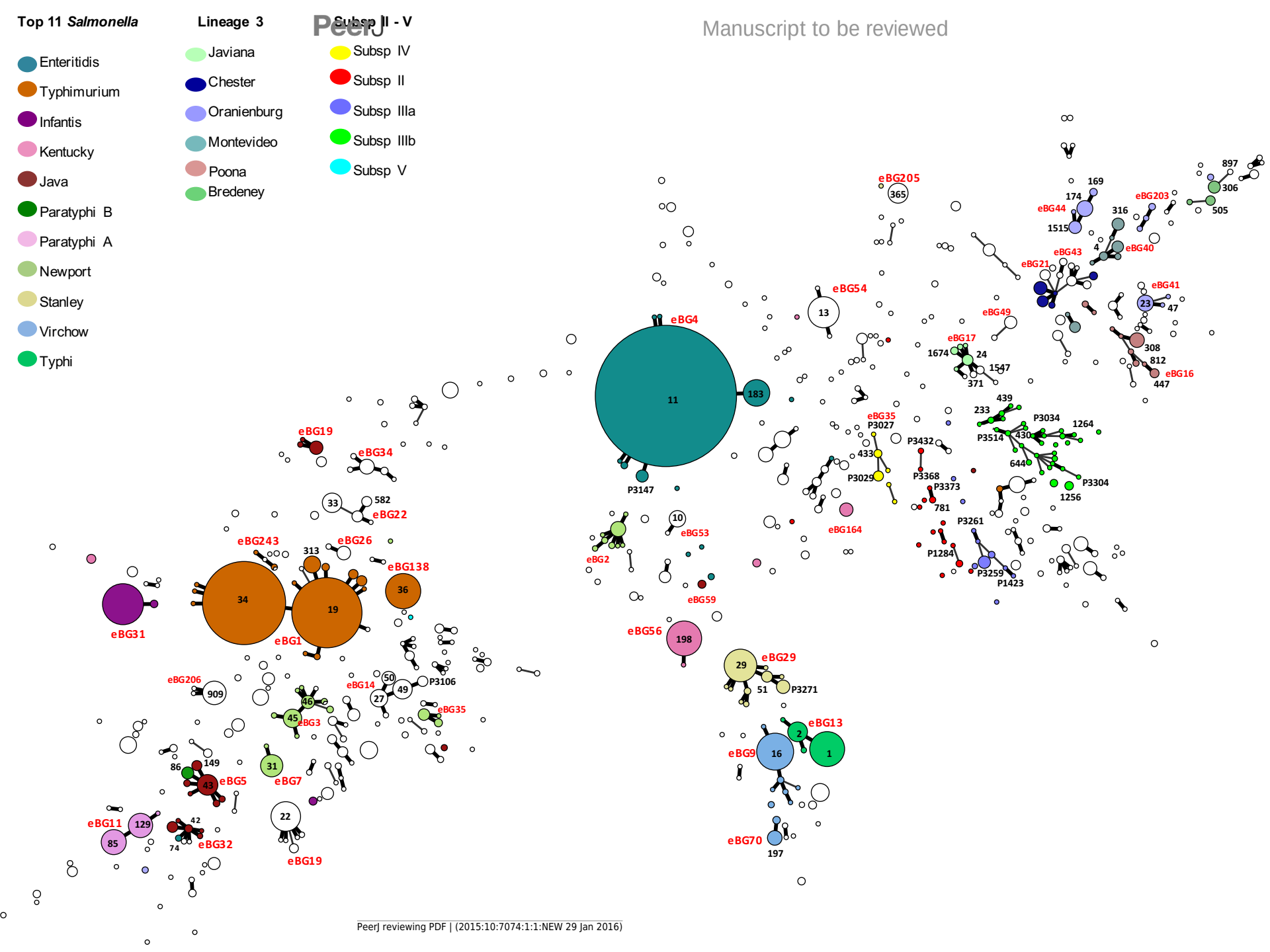
| | eBG5 | | | | | eBG 9 | eBG59 | eBG32 | | eBG95 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Phenotypic serovar** | ST43 | ST149 | ST307 | ST1577 | ST86 | ST88/127 | ST28 | ST423 | ST682/1588 | 1583 |
| Java | 45 | 7 | 4 | 3 | 2* | 18 | 6 | 5 | 2 | 1 |
| Paratyphi B | 0 | 0 | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 0 |

1

## Figure 1(on next page)

Population Structure of *Salmonella enterica* submitted to PHE

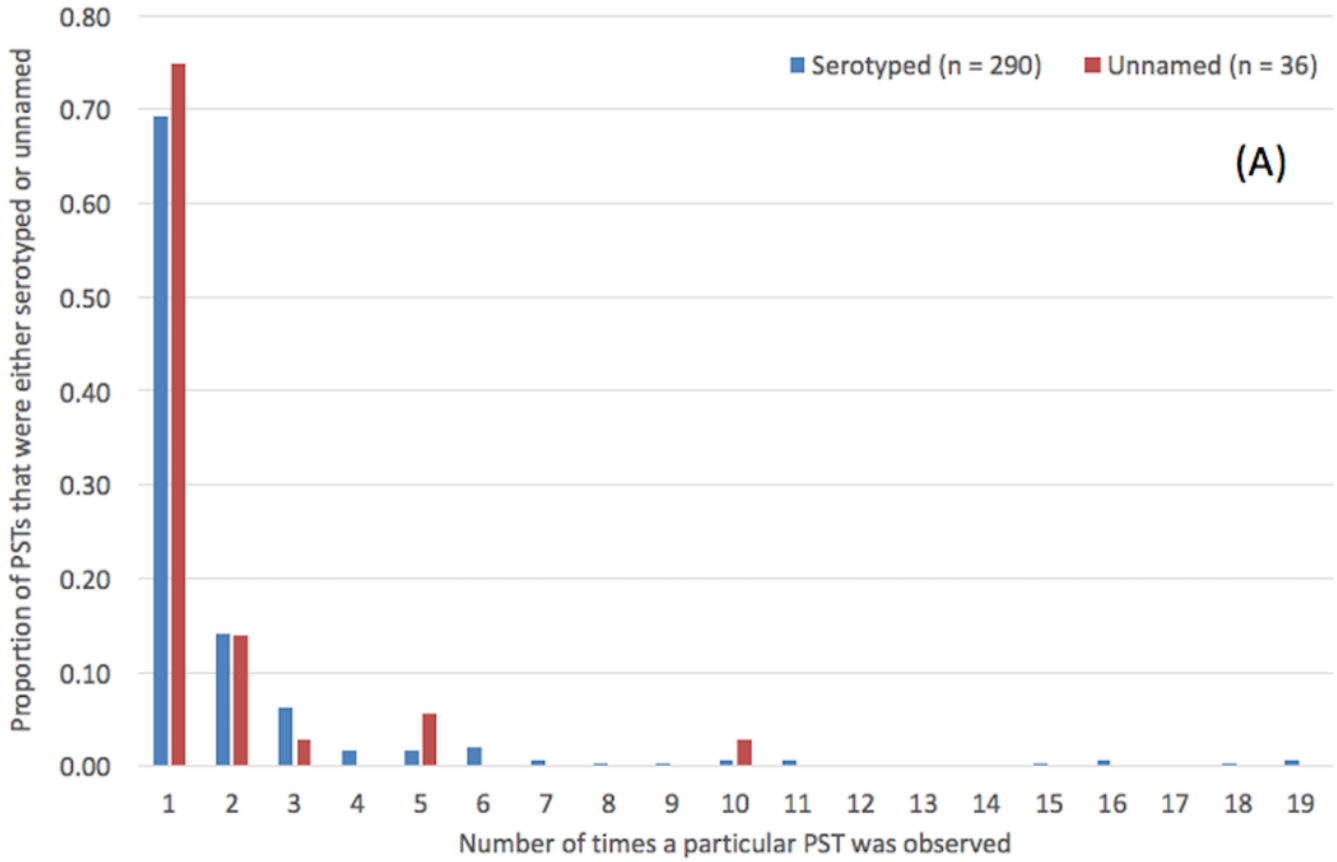
Population structure of all *Salmonella enterica* isolates submitted to PHE from local and regional hospital laboratories in England and Wales between April 2014 and March 2015 (see Supplementary Table for details)

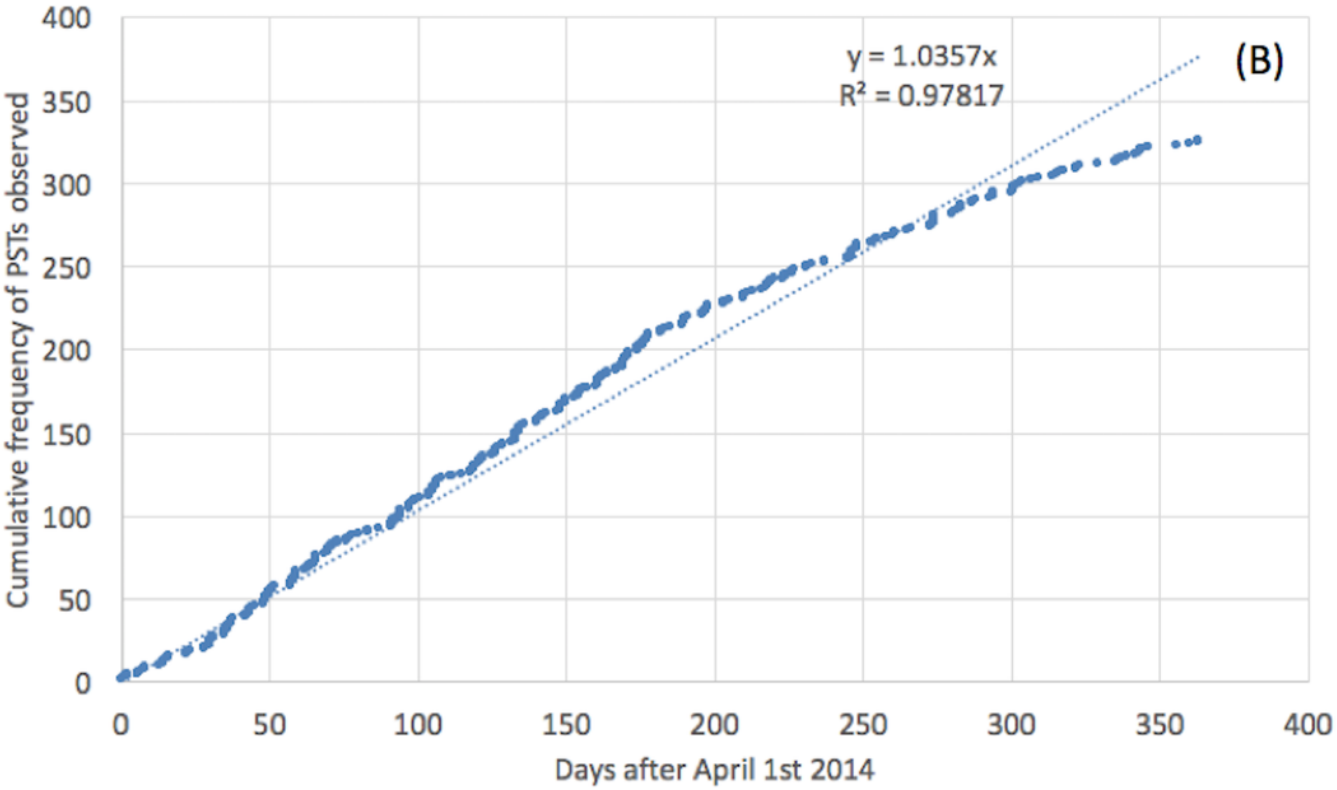

Top 11 Salmonella
Enteritidis
Typhimurium
Infantis
Kentucky
Java
Paratyphi B
Paratyphi A
Newport
Stanley
Virchow
Typhi
Lineage 3
Javiana
Chester
Oranienburg
Montevideo
Poona
Bredeney
Subsp II - V
Subsp IV
Subsp II
Subsp IIIa
Subsp IIIb
Subsp V
eBG4
eBG205
eBG54
eBG19
eBG34
eBG22
eBG243
eBG26
eBG138
eBG31
eBG1
eBG206
eBG14
eBG3
eBG35
eBG5
eBG7
eBG11
eBG32
eBG19
eBG53
eBG2
eBG59
eBG56
eBG29
eBG13
eBG9
eBG70
eBG44
eBG203
eBG40
eBG43
eBG21
eBG41
eBG49
eBG17
eBG16
eBG164
P3147
P3027
P3029
P3106
P3271
P3034
P3514
P3432
P3368
P3373
P3304
P3261
P1284
P3259
P1423

# 2

Trends in Preliminary Sequence Types

(A) Novel, preliminary STs (PST) and the modal number of isolates identified per PST (B) The rate at which PSTs were identified throughout the time frame of the study.

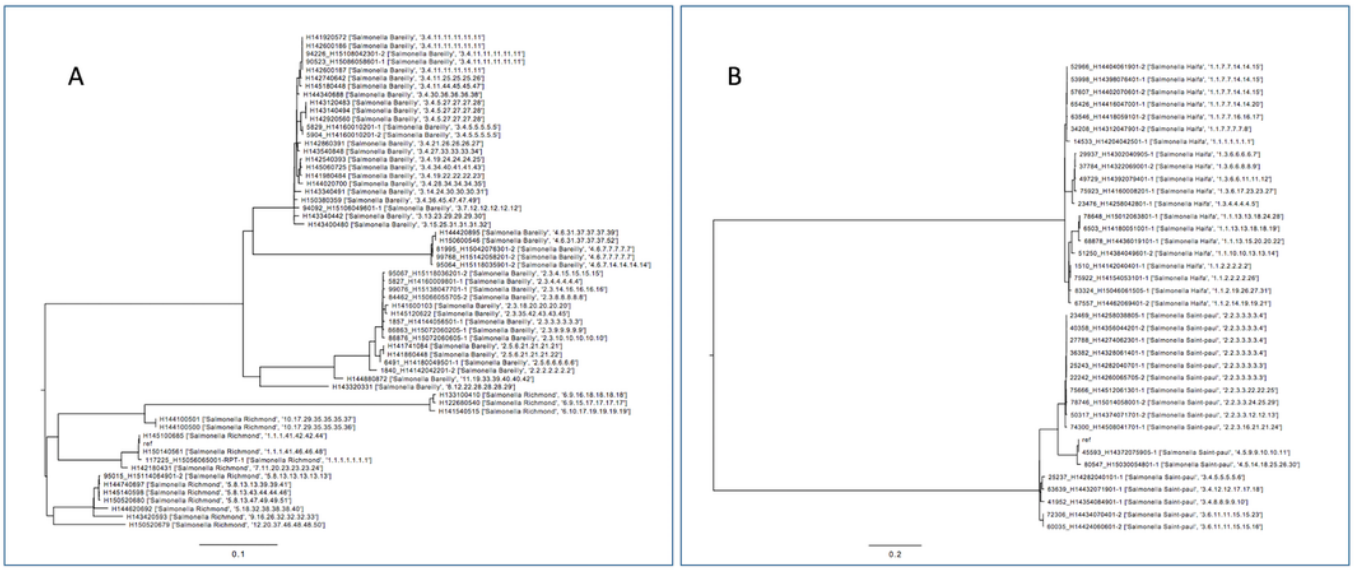

Serotyped (n = 290)
Unnamed (n = 36)
(A)
Proportion of PSTs that were either serotyped or unnamed
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Number of times a particular PST was observed


y = 1.0357x
R² = 0.97817
(B)
Cumulative frequency of PSTs observed
400
350
300
250
200
150
100
50
0
0 50 100 150 200 250 300 350 400
Days after April 1st 2014

# 3

Phylogenetic relationship within two sequence types containing multiple serotypes

(A) Phylogenetic relationship of S. Richmond and S. Bareilly (ST909) (B) Phylogenetic relationship of S. Saintpaul and S. Haifa (ST49)



# 4

*Salmonella* lineage 3 population structure

Serovars in lineage 3 mainly consist of multiple eBGs and are polyphyletic by nature