PeerJ

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

Philip M. Ashton[1], Satheesh Nair[1], Tansy M. Peters[1], Janet A. Bale[1], David G. Powell[1], Anaïs Painset[1], Rediat Tewolde[2], Ulf Schaefer[2], Claire Jenkins[1], Timothy J. Dallman[1], Elizabeth M. de Pinna[1], Kathie A. Grant[1,3] and Salmonella Whole Genome Sequencing Implementation Group

[1] Gastrointestinal Bacterial Reference Unit, Public Health England, London, United Kingdom
[2] Applied Laboratory and Bio-Informatics Unit, Public Health England, London, United Kingdom
[3] Gastrointestinal Infections, NIHR Health Protection Research Unit in Gastrointestinal Infections, London, United Kingdom

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 6,887 isolates of *S. enterica* subspecies I, and of these, 6,616 (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.

Subjects Bioinformatics, Genomics, Microbiology, Public Health
Keywords Whole genome sequencing, *Salmonella*, Bioinformatics, Multi-locus sequence typing, Public health

Submitted 9 October 2015
Accepted 12 February 2016
Published 5 April 2016

Corresponding authors
Claire Jenkins,
claire.jenkins@phe.gov.uk
Timothy J. Dallman,
tim.dallman@phe.gov.uk

Academic editor
Jose Izarzugaza

Additional Information and Declarations can be found on page 14

DOI 10.7717/peerj.1752

INTRODUCTION

The *Salmonella*e are major human pathogens and represent a significant global public health issue causing morbidity and mortality resulting in a high social and economic burden worldwide (Majowicz et al., 2010). The genus consists of two species; *Salmonella enterica* and *S. bongori*. There are six subspecies of *S. enterica* differentiated by biochemical variations, namely subspecies *enterica* (I), *salamae* (II), *arizonae* (IIIa), *diarizonae* (IIIb), *houtenae* (IV) and *indica* (VI) (*Threlfall, Ward & Old, 1999*). Subspecies I, *S. enterica* subsp. *enterica* cause 99% of human and animal infections. The two main pathologies associated with *S. enterica* are gastroenteritis and typhoidal disease. The typhoidal *Salmonellae* include *S.* Typhi and *S.* Paratyphi A, B and C. They are host restricted, monophyletic, rarely undergo recombination events and exhibit convergent evolution driven by genome degradation (*Wain et al., 2015*). The majority of gastroenteritis in the UK is caused by the host generalist serovars, such as *S.* Typhimurium and *S.* Enteritidis, and host adapted serovars that are adapted to a specific animal reservoir but can infect man and include *S.* Dublin, *S.* Gallinarum *S.* Choleraesuis, and *S.* Bovismorbificans (*Langridge et al., 2015*).

Approximately 8,000 isolates are referred to the *Salmonella* Reference Service (SRS) at Public Health England (PHE) each year from local and regional hospital laboratories. In April 2015, PHE implemented whole genome sequencing (WGS) as the routine typing tool for public health surveillance of *Salmonella* infections. Prior to April 2015, presumptive *Salmonella* isolates referred to SRS were speciated and sub-speciated using PCR (*Hopkins et al., 2009*; *Hopkins et al., 2011*) and grouped into serovars as described in the White-Kauffman-Le Minor scheme (*Grimont & Weill, 2007*; *Guibourdenche et al., 2010*; *Issenhuth-Jeanjean et al., 2014*). This methodology is based on reactions of rabbit antisera to the lipopolysaccharide (O antigen encoded by *rfb* genes) and flagellar antigens (phases 1 and 2 of H antigen encoded by *fli*C and *flj*B). The scheme utilises this phenotypic variation, expressed as an antigenic formulae, to divide *Salmonella* into more than 2,600 serovars. Epidemiological investigations of *Salmonella* infecting humans and animals have relied on serotyping for over 70 years; national and international governmental agencies base guidelines and regulations on the serotyping method and the use of this nomenclature is a globally recognised form of communication (*Swaminathan et al., 2006*; *EFSA, 2010*). Furthermore, serovars have often been shown to correlate with host range and disease sequelae (*Gordon, 2011*; *Wain et al., 2015*; *Langridge et al., 2015*).

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. (2012)* 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

Bacterial strains

All isolates ($n = 7{,}465$) of *Salmonella* from human cases of gastrointestinal disease submitted to SRS from local and regional hospital laboratories in England & Wales between 1st April 2014 and 31st March 2015 were sequenced in parallel with phenotypic serotyping (Table S1). Of these, 7,338 were identified as subspecies I and included 263 different serovars. The ten most common serovars in this dataset were Enteriditis (2,310), Typhimurium (1,407), Infantis (184), Typhi (184), Newport (173), Virchow (162), Kentucky (160), Stanley (146), Paratyphi A (135) and Java (99). One hundred and twenty seven isolates were identified as subspecies II–IV (*S. enterica* subspecies *salamae* $n = 28$; *S. enterica* subspecies *arizonae* $n = 25$; *S. enterica* subspecies *diarizonae* $n = 49$; *S. enterica* subspecies *houtenae* $n = 20$) and there was one isolate of *S. bongori*. No isolates belonging to subspecies VI (*S. enterica* subspecies *indica*) were submitted to SRS during the study period.

DNA extraction for WGS

DNA extraction of *Salmonella* isolates was carried out using a modified protocol of the Qiasymphony DSP DNA midi kit (Qiagen). In brief, 0.7 ml of overnight *Salmonella* culture in a 96 deep well plate was harvested. Bacterial cells were pre-lysed in 220 μl of ATL buffer (Qiagen) and 20 μl Proteinase K (Qiagen), and incubated shaking for 30 min at 56 °C. Four μl of RNase at 100 mg/ml (Qiagen) was added to the lysed cells and re-incubated for a further 15 min at 37 °C. This step increases the purity of the DNA for further downstream sequencing. Extraction of DNA from the treated cells was performed on the Qiasymphony SP platform (Qiagen) and eluted in 100 μl of water. DNA concentration using the GloMax system (Promega) was determined for the following sequencing steps.

DNA sequencing

Extracted DNA was then processed using the NexteraXT sample preparation method and sequenced with a standard 2x101 base protocol on a HiSeq 2500 Instrument in fast mode (Illumina, San Diego, CA, USA).

Bioinformatics workflow

FASTQ reads were quality trimmed using Trimomatic (*Bolger, Lohse & Usadel, 2015*) with bases removed from the trailing end that fell below a PHRED score of 30. If the read length post trimming was less than 50 bp the read and its pair were discarded. The PHE KmerID pipeline (https://github.com/phe-bioinformatics/kmerid) was used to compare the sequenced reads with 1,769 published genomes to 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 1,769 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 min 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 125 Gb 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 7,000 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, Banks & Poplin, 2011*) in unified genotyper mode. Core genome positions that had a high quality SNP (>90% consensus, minimum depth 10×, GQ $\geq$ 30, MQ $\geq$ 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

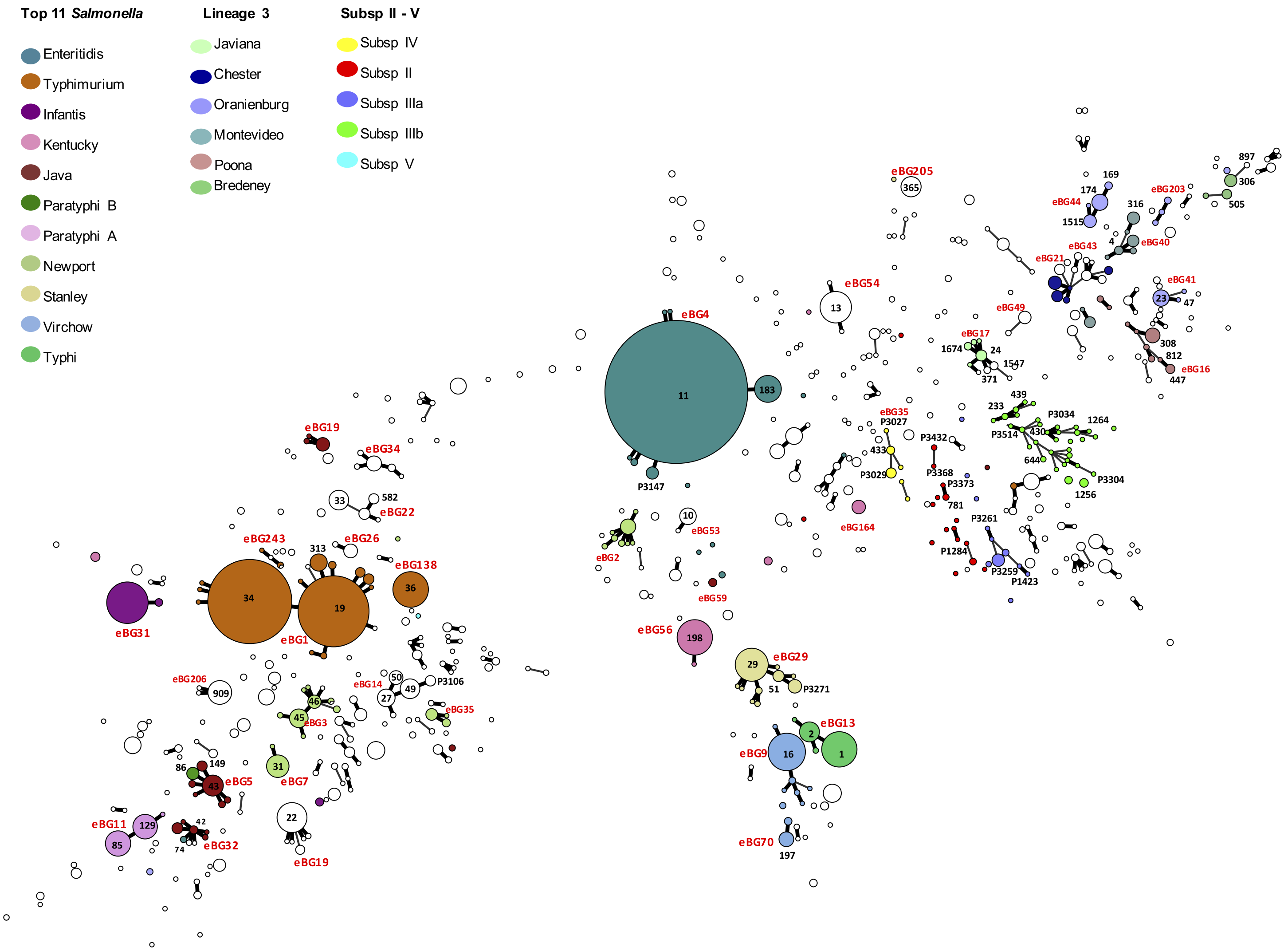

Figure 1 Population structure of all *Salmonella enterica* isolates submitted to PHE from local and regional hospital laboratories in England and Wales between April 2014 and March 2015 (see Table S1 for details).

to PHE between April 2014 and March 2015 is illustrated by the minimum spanning tree in Fig. 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 6,887 (94%) of 7,338 isolates of subspecies 1, and of these, 6,616 (96%) had the same result by both methods (Table S1). 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 7,338 strains tested, 70 (1%) had no serotype designation in the MLST database, of which 28 (40%) were serotyped phenotypically (Table S1).

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 (Table S1). These 654 isolates belonged to a total of 326 novel STs, designated PST; the modal number of isolates identified per PST was one (Fig. 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) (Fig. 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).

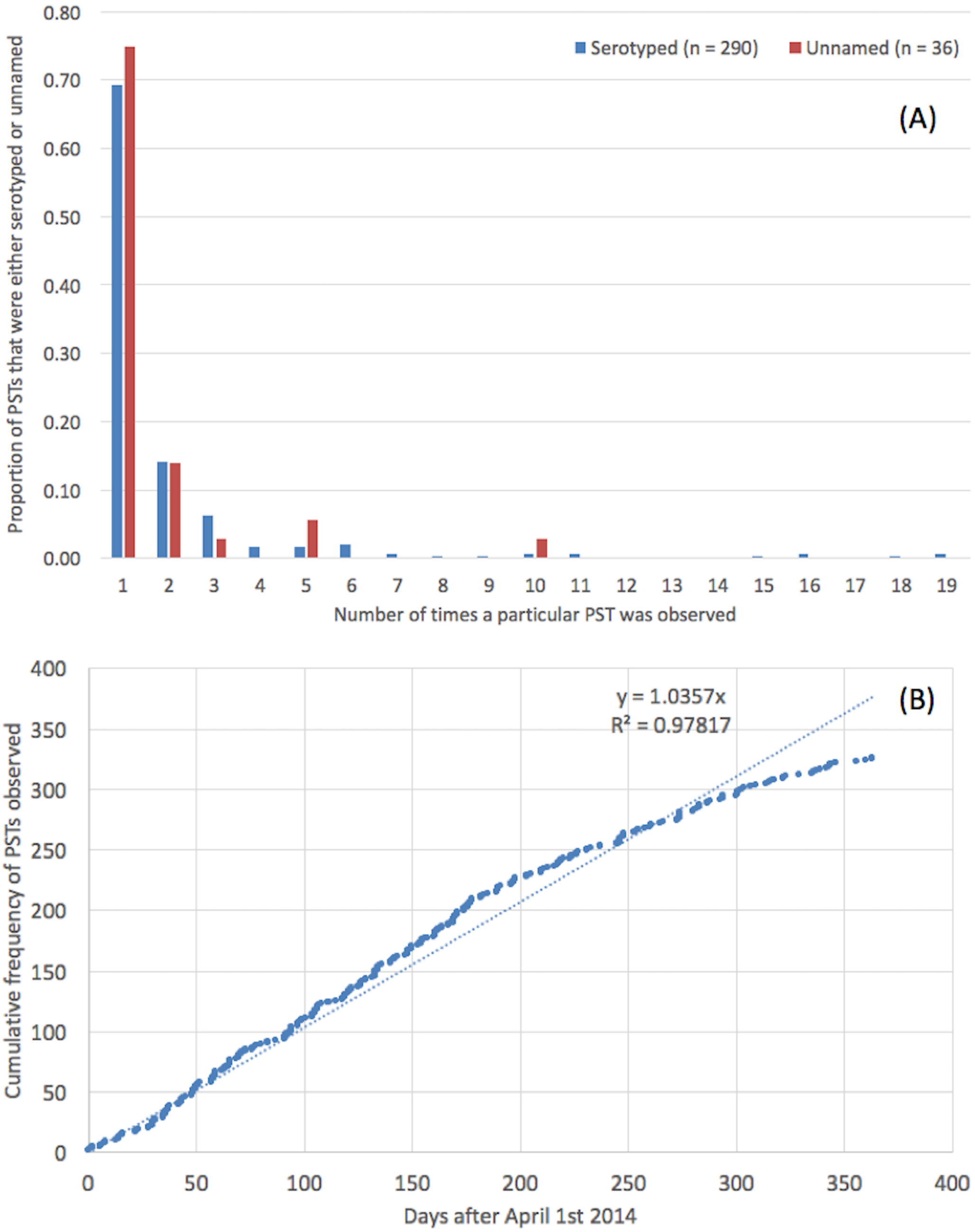

Figure 2 Trends in preliminary sequence types. (A) Novel, preliminary STs (PST) and the modal number of isolates identified per PST. (B) The rate at which PSTs were identified throughout the time frame of the study.

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

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

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 2,308 isolates of *S*. Enteritidis identified by both phenotypic serotyping and WGS MLST derived serotyping, 2,296 belonged to eBG4, including 2,200 ST11 and 76 ST183 (Fig. 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 2,308 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 1,392 isolates of *S*. Typhimurium and monophasic *S*. Typhimurium (rough and non-motile variants) (*Hopkins, De Pinna & Wain, 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 (Fig. 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

Table 2 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

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, Bunge & Helmuth, 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 Fig. 1). Two of these 99 isolates 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) (Fig. 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 (Fig. 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 (Table S1).

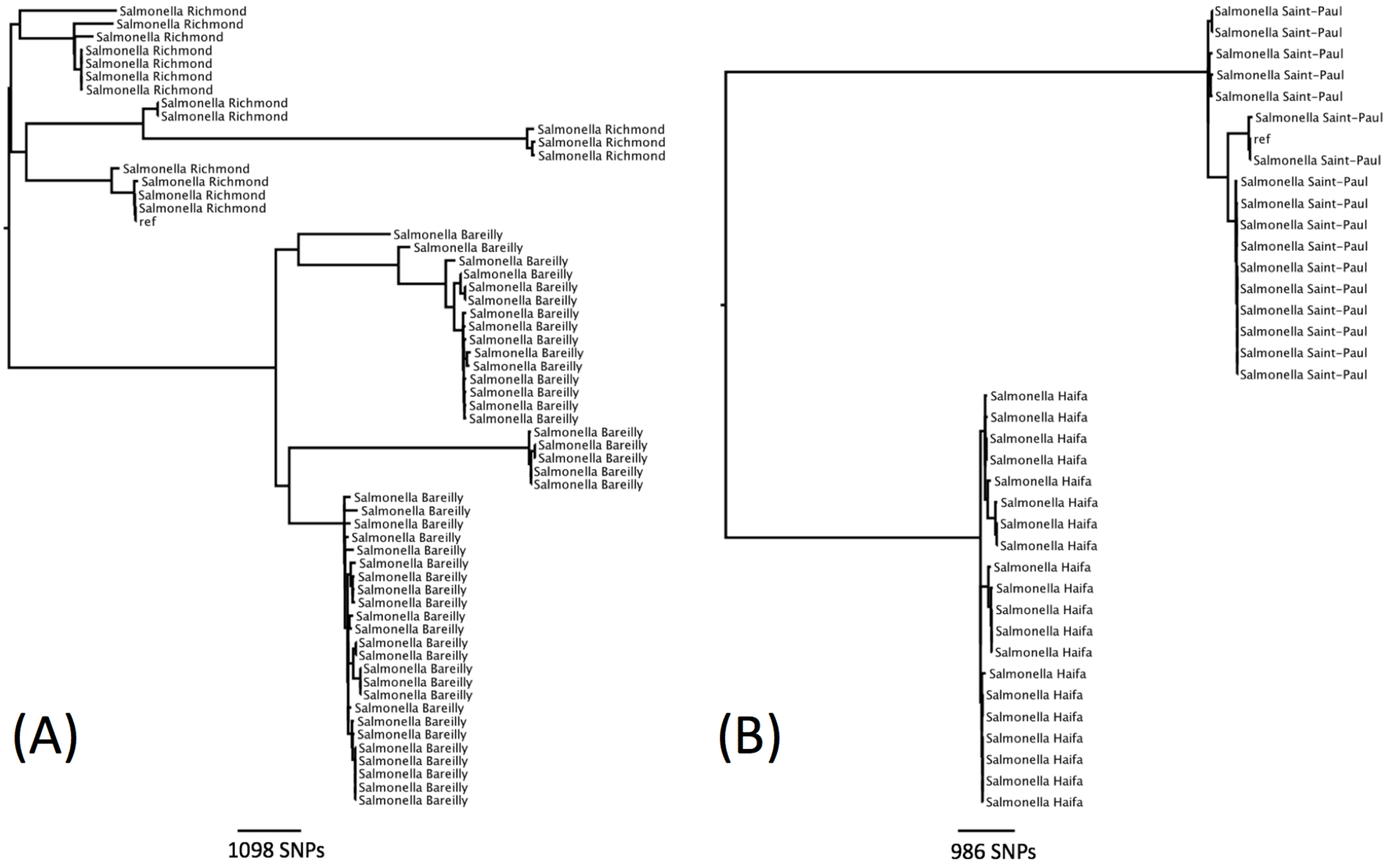

Figure 3 Phylogenetic relationship within two sequence types containing multiple serotypes. (A) Phylogenetic relationship of *S.* Richmond and *S.* Bareilly (ST909). (B) Phylogenetic relationship of *S.* Saintpaul and *S.* Haifa (ST49).

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 (Figs. 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 (Fig. 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

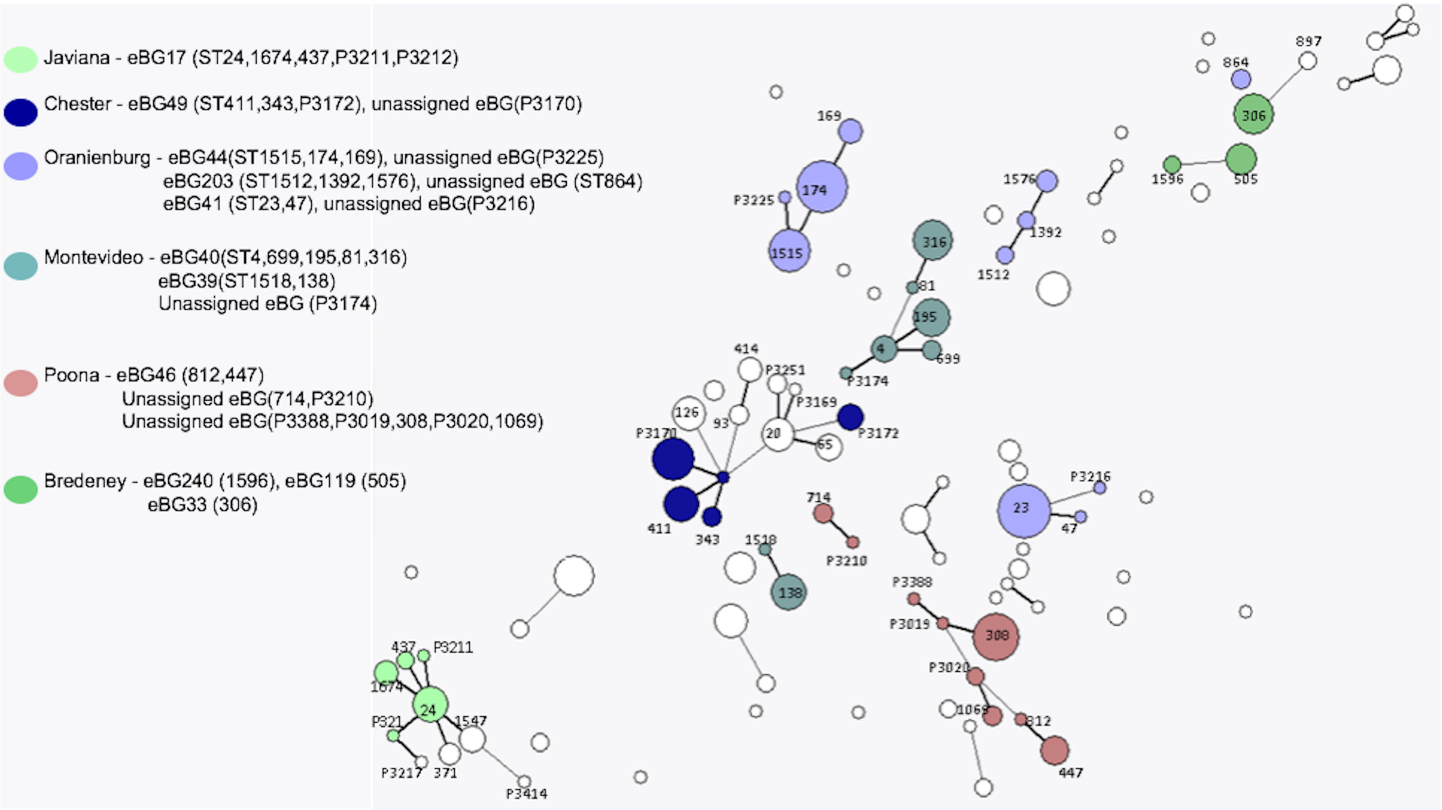
Javiana - eBG17 (ST24,1674,437,P3211,P3212)
Chester - eBG49 (ST411,343,P3172), unassigned eBG(P3170)
Oranienburg - eBG44(ST1515,174,169), unassigned eBG(P3225)
eBG203 (ST1512,1392,1576), unassigned eBG (ST864)
eBG41 (ST23,47), unassigned eBG(P3216)
Montevideo - eBG40(ST4,699,195,81,316)
eBG39(ST1518,138)
Unassigned eBG (P3174)
Poona - eBG46 (812,447)
Unassigned eBG(714,P3210)
Unassigned eBG(P3388,P3019,308,P3020,1069)
Bredeney - eBG240 (1596), eBG119 (505)
eBG33 (306)
897
864
306
169
174
P3225
1576
1596
505
1515
316
1392
1512
81
195
4
699
414
P3251
P3174
P3169
126
93
20
P3172
65
P3170
411
343
1518
714
P3210
P3216
23
47
138
P3388
P3019
308
P3020
1069
812
447
437
P3211
1674
24
P321
1547
P3217
371
P3414

Figure 4 ***Salmonella* lineage 3 population structure.** Serovars in lineage 3 mainly consist of multiple eBGs and are polyphyletic by nature.

connected network (Fig. 4). The five most common examples of this in the current study, were *S.* Oranienburg, *S.* Montevideo, *S.* Chester, *S.* Poona and *S.* Bredeney (Fig. 4 and Table S1). 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 (*2012*) argued convincingly for replacing serotyping with a MLST approach based on genetic population groupings for typing *S. enterica*. 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 6,887 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.

Isolates exhibiting monophasic properties that could not be fully serotyped phenotypically because they had an incomplete antigenic structure were matched to a ST derived serotype. The monophasic variants in this study mainly belonged to eBG1, eBG138 and eBG243 and previous studies have also shown that monophasic variants of *S*. Typhimurium have emerged as a result of multiple independent genetic events (*Soyer et al., 2009*; *Switt et al., 2009*; *Tennant et al., 2010*). Strains with monophasic properties are reportable to European Centre for Disease Prevention and Control (ECDC) but cannot be determined using the ST approach. Alternative strategies for determining monophasic characteristics by PCR are available (*Prendergast et al., 2013*) and methods for extracting this information from the genome sequencing data have been developed at PHE (P Ashton & A Lewis, pers. comm., 2015).

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

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

Two main issues arose during the evaluation of the MLST approach: (i) a number of STs did not have a serovar designation in the MLST database (including subspecies II–IV) and (ii) the unexpectedly large number of novel STs identified. Traditional phenotypic serotyping was required to type these isolates and the MLST database was modified and up-dated to incorporate the new data. Clearly, as we move forward the PHE MLST database will be constantly evolving and this data will be shared with colleagues in the field via existing MLST databases and their WGS compliant successors e.g., EnteroBase & BIGSdb. While it is difficult to draw conclusions based on our small sample size, MLST may not currently be an appropriate tool for the classification of *Salmonella* sub-species II–IV, due to the lack of a discrete population structure of EBGs. However, non-subspecies I isolates which are mainly adapted to cold blooded animals and/or reptiles contributed to less than 1.7% of the workload during the time frame of the study. Although MLST approach is generally more discriminatory than serotyping, it does not always provide the fine resolution required for public health surveillance. Further analysis based on single nucleotide polymorphisms in the core genome compared to a type strain representing the most common eBGs is performed for outbreak detection and investigation (*Ashton et al., 2015*).

In conclusion, serotyping inferred from 7-gene MLST results derived from WGS data is an accurate, robust, reliable, high throughput typing method that is well suited to routine public health surveillance of *Salmonella*. This approach supports the maintenance of traditional serovar nomenclature and provides further insight on the true evolutionary relationship between isolates, as well as a framework for fine level typing within eBGs for surveillance, outbreak detection and source attribution.

ACKNOWLEDGEMENTS

We would like to thank all the members of the *Salmonella Whole Genome Sequencing Implementation Group* including Steve Connell, Anna Lewis, Andy Levy, Clare Maguire, Clare Wend-Hansen, Martin Day, James Rogers, Siham Ibrahim, Arlene Barcenilla, Vineet Patel, Kiran Jayan, Anthony Underwood, Catherine Arnold and Ian Harrison.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The research was partially funded by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Gastrointestinal Infections at the University of Liverpool in partnership with Public Health England (PHE), University of East Anglia, University of Oxford and the Institute of Food Research. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Institute for Health Research Health Protection Research Unit (NIHR HPRU).
Public Health England (PHE).
University of East Anglia.
University of Oxford.
Institute of Food Research.

Competing Interests

The authors declare there are no competing interests

Author Contributions

- Philip M. Ashton and Satheesh Nair conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Tansy M. Peters conceived and designed the experiments, performed the experiments, reviewed drafts of the paper.
- Janet A. Bale conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- David G. Powell analyzed the data, contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- Anaïs Painset analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, reviewed drafts of the paper.
- Rediat Tewolde conceived and designed the experiments, performed the experiments, contributed reagents/materials/analysis tools, reviewed drafts of the paper.

- Ulf Schaefer contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- Claire Jenkins analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Timothy J. Dallman conceived and designed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, reviewed drafts of the paper.
- Elizabeth M. de Pinna and Kathie A. Grant conceived and designed the experiments, reviewed drafts of the paper.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

All data from the Salmonella surveillance project are deposited in the BioProject of the SRA: PRJNA248792. Per sample accessions are available in Table S1.

Data Availability

The following information was supplied regarding data availability:

MOST code for MLST and KmerID code are available here: https://github.com/phe-bioinformatics.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.1752#supplemental-information.

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 DOI 10.1371/journal.ppat.1002776.

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 Currents* Epub ahead of print Feb 10 2015 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)**:305–312 DOI 10.1093/cid/civ318.

DePristo MA, Banks E, Poplin RE, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43(5)**:491–498 DOI 10.1038/ng.806.

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 DOI 10.1371/journal.ppat.1002191.

EFSA Panel on Biological Hazards (BIOHAZ). 2010. Scientific Opinion on monitoring and assessment of the public health risk of "*Salmonella* Typhimurium-like" strains. *EFSA Journal* **8(10)**:1826 DOI 10.2903/j.efsa.2010.1826./efsajournal.htm.

Gordon MA. 2011. Invasive nontyphoidal *Salmonella* disease: epidemiology, pathogenesis and diagnosis. *Current Opinion Infectious Diseases* **24**:484–489 DOI 10.1097/QCO.0b013e32834a9980.

Grimont PAD, Weill FX. 2007. *Antigenic Formulae of the Salmonella Serovars (ninth ed.)*. Paris: WHO Collaborating Center for Reference and Research on Salmonella, Institut Pasteur.

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 DOI 10.1016/j.resmic.2009.10.002.

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)**: 20275.

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)**:278–280 DOI 10.1016/j.diagmicrobio.2011.01.015.

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)**:452–454 DOI 10.1016/j.diagmicrobio.2009.03.022.

Inouye M, Conway TC, Zobel J, Holt KE. 2012. Short read sequence typing (SRST): multi-locus sequence types from short reads. *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)**:526–530 DOI 10.1016/j.resmic.2014.07.004.

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 DOI 10.1371/journal.ppat.1002824.

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 of the United States of America* **112(3)**:863–868 DOI 10.1073/pnas.1416707112.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25(14)**:1754–1760 DOI 10.1093/bioinformatics/btp324.

Majowicz SE, Musto J, Scallan E, Angulo FJ, Kirk M, O'Brien SJ, Jones TF, Fazil A, Hoekstra RM. 2010. International collaboration on enteric disease 'burden of illness' studies. *Clinical Infectious Diseases* **6**:882–889.

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)**:4292–4297 DOI 10.1128/JCM.41.9.4292-4297.2003.

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)**:1786–1797 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)**:6465–6476 DOI 10.1128/JB.00969-10.

Shi C, Singh P, Ranieri ML, Wiedmann M, Moreno Switt AI. 2015. Molecular methods for serovar determination of *Salmonella*. *Critical Reviews Microbiology* **41(3)**:309–325 DOI 10.3109/1040841X.2013.837862.

Soyer Y, Moreno SA, Davis MA, Maurer J, McDonough PL, Schoonmaker-Bopp DJ, Dumas NB, Root T, Warnick LD, Grohn YT, Wiedmann M. 2009. *Salmonella enterica* serotype 4,5,12:i:-, an emerging *Salmonella* serotype that represents multiple distinct clones. *Journal of Clinical Microbiology* **47**:3546–3556 DOI 10.1128/JCM.00546-09.

Swaminathan B, Gerner-Smidt P, Ng LK, Lukinmaa S, Kam KM, Rolando S, Gutierrez EP, Binsztein N. 2006. Building PulseNet International: an interconnected system of laboratory networks to facilitate timely public health recognition and response to foodborne disease outbreaks and emerging foodborne diseases. *Foodborne Pathogens and Disease* **3**:36–50 DOI 10.1089/fpd.2006.3.36.

Switt AI, Soyer Y, Warnick LD, Wiedmann M. 2009. Emergence, distribution, and molecular and phenotypic characteristics of *Salmonella enterica* serotype 4,5,12:i:-. *Foodborne Pathogens and Disease* **6(4)**:407–415 DOI 10.1089/fpd.2008.0213.

Tennant SM, Diallo S, Levy H, Livio S, Sow SO, Tapia M, Fields PI, Mikoleit M, Tamboura B, Kotloff KL, Nataro JP, Galen JE, Levine MM. 2010. Identification by PCR of non-typhoidal *Salmonella enterica* serovars associated with invasive infections among febrile patients in Mali. *PLoS Neglected Tropical Diseases* **4**:e621 DOI 10.1371/journal.pntd.0000621.

Thomson NR, Clayton DJ, Windhorst D, Vernikos G, Davidson S, Churcher C, Quail MA, Stevens M, Jones MA, Watson M, Barron A, Layton A, Pickard D, Kingsley RA, Bignell A, Clark L, Harris B, Ormond D, Abdellah Z, Brooks K, Cherevach I, Chillingworth T, Woodward J, Norberczak H, Lord A, Arrowsmith C, Jagels K, Moule S, Mungall K, Sanders M, Whitehead S, Chabalgoity JA, Maskell D, Humphrey T, Roberts M, Barrow PA, Dougan G, Parkhill J. 2008. Comparative genome analysis of *Salmonella* Enteritidis PT4 and *Salmonella* Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. *Genome Research* **18(10)**:1624–1637 DOI 10.1101/gr.077404.108.

Threlfall J, Ward L, Old D. 1999. Changing the nomenclature of *Salmonella*. *Communicable Diseases Public Health* **2(3)**:156–157.

Wain J, Hendriksen RS, Mikoleit ML, Keddy KH, Ochiai RL. 2015. Typhoid fever. *Lancet* **385(9973)**:1136–1145 DOI 10.1016/S0140-6736(13)62708-7.

Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, Fitzgerald C, Fields PI, Deng X. 2015. Salmonella serotype determination utilizing high-throughput genome sequencing data. *Journal of Clinical Microbiology* **53(5)**:1685–1692 DOI 10.1128/JCM.00323-15.