AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE

What's in a Name? Species-Wide Whole-Genome Sequencing Resolves Invasive and Noninvasive Lineages of *Salmonella enterica* Serotype Paratyphi B

Thomas R. Connor,[a,b] Sian V. Owen,[a,c] Gemma Langridge,[c] Steve Connell,[d] Satheesh Nair,[d] Sandra Reuter,[b] Timothy J. Dallman,[e] Jukka Corander,[f,p] Kristine C. Tabing,[g] Simon Le Hello,[h] Maria Fookes,[b] Benoît Doublet,[i,j] Zhemin Zhou,[k] Theresa Feltwell,[b] Matthew J. Ellington,[l*] Silvia Herrera,[m] Matthew Gilmour,[g] Axel Cloeckaert,[i,j] Mark Achtman,[k] Julian Parkhill,[b] John Wain,[n] Elizabeth De Pinna,[d] François-Xavier Weill,[h] Tansy Peters,[d] Nick Thomson[b,o]

Cardiff University School of Biosciences, Cardiff University, Cardiff, United Kingdom[a]; Wellcome Trust Sanger Institute, Hinxton, United Kingdom[b]; Institute of Integrative Biology, University of Liverpool, Liverpool, United Kingdom[c]; Gastrointestinal Bacteria Reference Unit, Public Health England, London, United Kingdom[d]; Public Health England, London, United Kingdom[e]; Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland[f]; Bacteriology and Enteric Disease Program, National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, Canada[g]; Institut Pasteur, Unité des Bactéries Pathogènes Entériques, Paris, France[h]; INRA, UMR1282 Infectiologie et Santé Publique, Nouzilly, France[i]; Université François Rabelais de Tours, UMR1282 Infectiologie et Santé Publique, Tours, France[j]; Warwick Medical School, University of Warwick, Coventry, United Kingdom[k]; Public Health England, Addenbrooke's Hospital, Cambridge, United Kingdom[l]; Instituto de Salud Carlos III, Centro Nacional de Microbiología, Majadahonda, Madrid, Spain[m]; Norwich Medical School, UEA, Norwich, United Kingdom[n]; London School of Hygiene and Tropical Medicine, London, United Kingdom[o]; Department of Biostatistics, University of Oslo, Oslo, Norway[p]

* Present address: Matthew J. Ellington, Public Health England, Antimicrobial Resistance and Healthcare Associated Infections, London, United Kingdom.

ABSTRACT **For 100 years, it has been obvious that *Salmonella enterica* strains sharing the serotype with the formula 1,4,[5],12:b:1,2—now known as Paratyphi B—can cause diseases ranging from serious systemic infections to self-limiting gastroenteritis. Despite considerable predicted diversity between strains carrying the common Paratyphi B serotype, there remain few methods that subdivide the group into groups that are congruent with their disease phenotypes. Paratyphi B therefore represents one of the canonical examples in *Salmonella* where serotyping combined with classical microbiological tests fails to provide clinically informative information. Here, we use genomics to provide the first high-resolution view of this serotype, placing it into a wider genomic context of the *Salmonella enterica* species. These analyses reveal why it has been impossible to subdivide this serotype based upon phenotypic and limited molecular approaches. By examining the genomic data in detail, we are able to identify common features that correlate with strains of clinical importance. The results presented here provide new diagnostic targets, as well as posing important new questions about the basis for the invasive disease phenotype observed in a subset of strains.**

IMPORTANCE ***Salmonella enterica* strains carrying the serotype Paratyphi B have long been known to possess Jekyll and Hyde characteristics; some cause gastroenteritis, while others cause serious invasive disease. Understanding what makes up the population of strains carrying this serotype, as well as the source of their invasive disease, is a 100-year-old puzzle that we address here using genomics. Our analysis provides the first high-resolution view of this serotype, placing strains carrying serotype Paratyphi B into the wider genomic context of the *Salmonella enterica* species. This work reveals a history of disease dating back to the middle ages, caused by a group of distinct lineages with various abilities to cause invasive disease. By quantifying the key genomic differences between the invasive and noninvasive populations, we are able to identify key virulence-related targets that can form the basis of simple, rapid, point-of-care tests.**

Received 23 March 2016 **Accepted** 12 July 2016 **Published** 23 August 2016

Citation Connor TR, Owen SV, Langridge G, Connell S, Nair S, Reuter S, Dallman TJ, Corander J, Tabing KC, Le Hello S, Fookes M, Doublet B, Zhou Z, Feltwell T, Ellington MJ, Herrera S, Gilmour M, Cloeckaert A, Achtman M, Parkhill J, Wain J, De Pinna E, Weill F-X, Peters T, Thomson N. 2016. What's in a name? Species-wide whole-genome sequencing resolves invasive and noninvasive lineages of *Salmonella enterica* serotype Paratyphi B. mBio 7(4):e00527-16. doi:10.1128/mBio.00527-16.

Editor Tom Chiller, CDC

Address correspondence to Thomas R. Connor, connortr@cardiff.ac.uk.

Salmonella paratyphosus B, *Salmonella schottmuelleri*, *Salmonella* Java: the serotype of *Salmonella enterica* subspecies *enterica* with the formula 1,4,[5],12:b:1,2 that is now known as *S. enterica* serotype Paratyphi B has not always been named thus. In the last 100 years, it has been known by various names, several of which are still commonly used today. The reason for this multiplicity of names is because isolates possessing the serotype have long been observed to cause either invasive disease (characterized by life-threatening paratyphoid fever) or gastroenteritis. It was clear to many microbiologists in the late 19th and early 20th centuries that, despite the shared serotype, there were differences between strains that related to the different disease outcomes. However, categorizing the differences in the form of classical, reproducible biochemical tests has proven to be a nontrivial problem, as the differences are more subtle than the disease phenotypes might suggest. Muller and, later, Kauffmann ultimately subdi-

vided the serotype into two biovars on the basis of an ability to form a slime wall and to ferment dextrorotatory tartrate (*d*Ta) (1, 2). In his classification of *Salmonella* isolates, Kauffmann named those isolates that formed a slime wall, were unable to ferment *d*Ta (*d*Ta$^-$), and caused paratyphoid fever in humans *S.* Paratyphi B. Those isolates that did not form a slime wall and were able to ferment *d*Ta (*d*Ta$^+$), he named *S.* Java (3). The already unclear delineation was further confused when Le Minor (4) ultimately rejected this nomenclature to redefine *S.* Java as a biovar of *S.* Paratyphi B. In practice, the result of the *d*-tartrate test remains the principal method for distinguishing *S.* Paratyphi B isolates causing invasive disease from those causing gastroenteritis (5). The implications of an isolate being classified as *S.* Paratyphi B or *S.* Java are significant for patients, reference laboratories, and public health authorities. Although in the first instance, treatment will depend on presentation, as a member of the so-called typhoidal *Salmonella* group, cases of disease where *S.* Paratyphi B is detected generally necessitate household follow-up and contact tracing. This is not required for *S.* Java infections, even when these cause systemic infections. In laboratory research, there are also significantly different handling requirements depending on the result—*d*Ta$^-$ strains are treated as biological safety level 3 (BSL3) organisms, while *d*Ta$^+$ organisms are handled at BSL2.

The molecular explanation for the differences in the ability to ferment *d*-tartrate is known: one single-nucleotide polymorphism (SNP) in the first codon of the gene upstream from *ttdA* and *ttdB*, the genes responsible for *d*-tartrate metabolism, ablates their expression (5). Based on IS*200* profiling (6), multilocus sequence typing (MLST) (7), and latterly, clustered regularly interspaced short palindromic repeat (CRISPR) typing (8), it is clear that the serotype falls into a number of discrete groups (7) and that possession of the *d*Ta$^-$ SNP is characteristic of groups of strains that are predominantly associated with invasive disease. However, the relationship between groups of isolates carrying the common serotype remains unresolved (9).

Salmonella strains possessing this serotype remain a common cause of gastroenteritis, being responsible for recent outbreaks in the United Kingdom (10), Belgium (11), Scandinavia (12), Canada (13, 14), and the United States (15), as well as a cause of invasive disease around the world in travelers (16). Moreover, since the late 1990s, two different clones of *S.* Paratyphi B *d*Ta$^+$ with resistance to multiple antibiotics have become increasingly associated with human infections (17), poultry, and poultry products (11, 18). These clones carry two different multidrug resistance-encoding integrons, *Salmonella* genomic island 1 (SGI-1) (19) and a chromosomally located class 2 integron carrying the *dfrA1-sat2-aadA1* (Tn*7*) array of gene cassettes (20), which confer resistance to trimethoprim, streptothricin, and aminoglycosides, respectively. Because of an inability to unambiguously subdivide this grouping in a phylogenetically meaningful way and an almost complete lack of knowledge about the genomic content and phylogenetic relationships of these strains, the significance of these observations is difficult to quantify. The work presented here defines the population structure of strains carrying the Paratyphi B serotype and reveals how strains carrying this serotype can be subdivided and how they vary in gene content. By gaining a better understanding of the population structure of strains carrying this Paratyphi B serotype, we are able to define what separates invasive from noninvasive strains, opening up new opportunities to better diagnose and track this organism.

RESULTS AND DISCUSSION

Whole-genome sequencing reveals the extent of divergence between isolates sharing the Paratyphi B serotype. Based upon its O antigen (formula 1,4,[5],12), *S.* Paratyphi B is a member of the group O:4 (formally group B) salmonellae. Forty-six different O serogroups have been identified within *Salmonella* (21), and these serogroups provide a structure to group together the >2,500 serotypes of the species (22). To quantify the population structure of *S.* Paratyphi B, we assembled a collection of 191 strains collected over 120 years that possess the serotype with the formula 1,4,[5],12:b:[1,2], encompassing both diphasic (b:1,2-type flagella present) and monophasic (b-type flagella only) Paratyphi B/Java isolates. To place these samples into a wider context, we also included a selection of 25 other salmonellae, including 10 other representatives of group O:4, as well as published reference genomes for 6 other *Salmonella* serotypes associated with invasive disease in humans and animals (see Table S1 in the supplemental material). We began our analysis by identifying the core and accessory genomes across the isolate collection. The diversity of isolates carrying this serotype is evidenced in the fact that the pangenome of the serotype itself is open (see Fig. S1) (23) and the core genome size of isolates sharing the Paratyphi B serotype (2,949 genes) is smaller by almost 1,000 genes than those reported for other serotypes, such as *S. enterica* serotype Typhimurium (3,846 genes) (24). Relative to the core genome across the sample data set, approximately 53% of the genome of the *S.* Paratyphi B reference strain is core to the serotype, in comparison to 43% that is core to the entire sample data set. Following the removal of putative recombinant regions using Gubbins (see Fig. S1) (25), a phylogenetic analysis based on the remaining SNPs found at positions that are shared by all isolates visualizes the extreme level of diversity between isolates carrying the Paratyphi B serotype (Fig. 1). However, it is also clear that there is not a continuum of diversity present within this group, but rather, strains carrying the common serotype fall into discrete clusters of strains. Recognizing that classical subdivisions have introduced a confused and inconsistent nomenclature to describe this group, we used a population genetic statistical framework, Bayesian Analysis of Population Structure (BAPS) (26), to perform an unsupervised subdivision of the Paratyphi B complex into a set of phylogroups (PGs). Iteratively clustering across the population, BAPS identified a set of 10 Paratyphi B clusters that we define as PG1 to PG10. Only isolates from PG1 possessed the SNP that is diagnostic for *d*Ta$^-$ strains. This represents the first occasion, to our knowledge, where an automated approach has unambiguously separated out *d*Ta$^-$ strains into a cluster that is distinct from other closely related strains without the need for a marker such as the *d*Ta$^-$ SNP to be provided as a basis for subdivision. Our results show that the *d*Ta$^-$ isolates fall within a group of 5 PGs (PG1 to -5) that, although separated by between 1,000 and 10,000 SNPs, are more closely related to one another than they are to other PGs or serotypes. Intriguingly, this difference is of a scale similar to the ~6,000 SNPs that separate the closely related host generalist and invasive serotypes *S. enterica* serotypes Enteritidis and Gallinarum/Pullorum (see Fig. S2) (27). Moving beyond PG1 to -5, it is clear that the serotype 1,4,[5],12:b:[1,2] is found in genomic backgrounds across the *S. enterica* species tree (Fig. 1) and that all isolates sharing this serotype clearly do not share a recent common ancestor. The extent of the separation between PGs is marked: isolates from *S. enterica* sero-

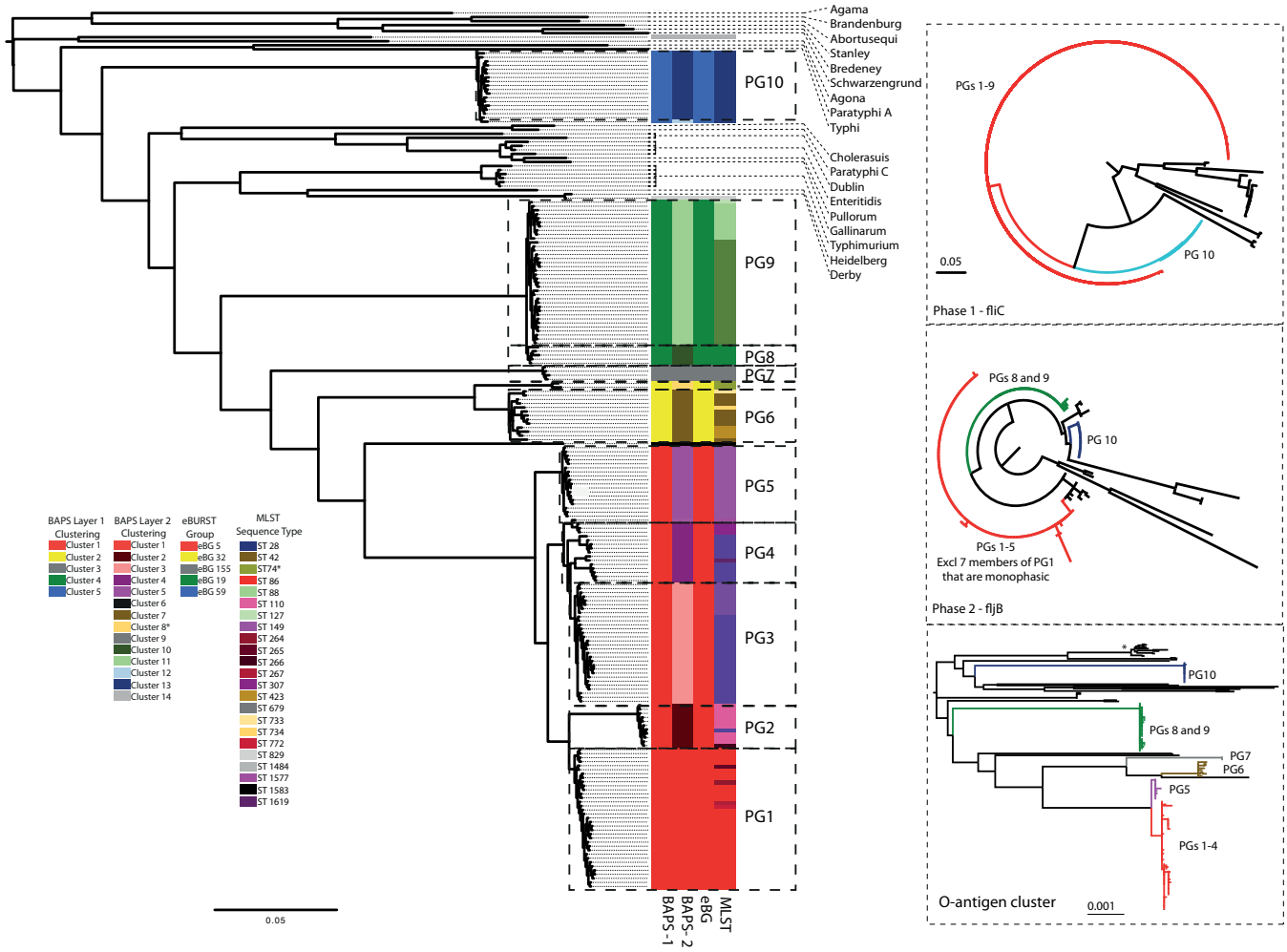

FIG 1 Whole-genome and LPS gene maximum-likelihood phylogenies of strains carrying serotype Paratyphi B and reference strains from other serotypes. Main figure: whole-genome phylogeny of the Paratyphi B strains, with representatives from other serogroup B and invasive serotypes shown. The tree was constructed based on the core genome for the isolates and drawn using RAxML. Next to the tree, colored blocks indicate the BAPS cluster that an isolate has been assigned to and the MLST and eBURST group that an isolate belongs to. Right, top to bottom: maximum-likelihood trees based on the nucleotide sequences of the phase 1 antigen *fliC* and the nucleotide sequences of *fljB*, drawn using PhyML, and maximum-likelihood tree based on the nucleotide sequences for the O-antigen cluster extracted from Paratyphi B and reference strains used to generate the whole-genome phylogeny, drawn using RAxML. In all cases, the phylogenetic trees were generated using a general time-reversible (GTR) model with gamma correction for between-site heterogeneity.

types Heidelberg, Derby, Paratyphi C, Choleraesuis, Gallinarum var. Gallinarum, Gallinarum var. Pullorum, Enteritidis, and Dublin all share more SNPs with PG1 to -9 than the isolates of PG10 do.

The common serotype is a result of recombination events at the flagellum loci. Kauffmann suggested that the evident variability in the genetic background of Paratyphi B strains may be due to recombination (28). The Paratyphi B serotype is defined based on its O antigen (responsible for the first part of the serotype formula, 4,5,12) and phase 1 and phase 2 flagella (the H antigens). The phase 1 flagella are of type b, and either the phase 2 flagella are of type 1,2 or, in the case of monophasic isolates, no phase 2 is present. Extracting the complete O-antigen gene cluster, as well as the *fliC* (phase 1) and *fljB* (phase 2) genes, and generating a phylogeny of these components individually revealed a startling similarity in the topology of the O-antigen cluster compared with the core genome tree (Fig. 1, inset). In contrast, the phylogenies for *fliC* (Fig. 1, inset) and *fljB* (Fig. 1, inset) individually revealed evolutionary histories that were markedly different from those of both the core genome and the O-antigen gene cluster. In the case of *fljB*, nonmonophasic Paratyphi B isolates all possessed identical or nearly identical gene sequences (Fig. 1, inset). Looking at the 20 kb of DNA around these two genes reveals a topology that is markedly different from the core genome tree (see Fig. S3 in the supplemental material), implying that while the common flagella with limited diversity could be due to selection, the more parsimonious explanation for the lack of diversity within *fljB* and *fliC* is that of homologous recombination. These results suggest that, in a number of cases, the import of flagellum genes has led to the establishment of clear lineages that have been able to cause widespread disease in humans and animals. It is, however, also interesting to note that when collecting isolates sharing the serotype 1,4,[5],12:b:[1,2] from clinical settings, we also collected a small number of strains that are singletons (Fig. 1, highlighted in grey).

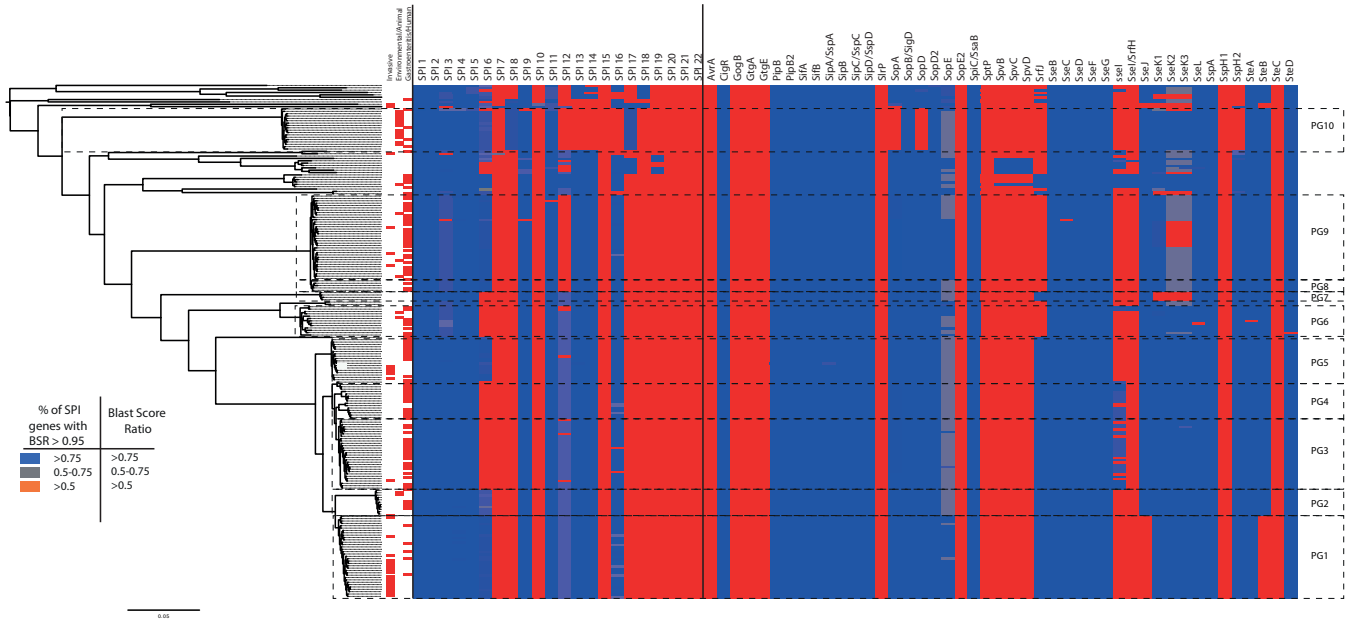

FIG 2 Distribution of SPIs and effectors, with disease phenotypes of isolates indicated immediately to the right of the phylogenetic tree. Color coding for the SPIs is based on the percentage of genes on an SPI that are present in a given isolate (defined as genes with a Blast score ratio to the expected gene of >0.95). Color coding for effectors corresponds to the Blast score ratio recorded for each genome when screened with the reference effectors. Disease phenotypes have been classified into three categories where metadata are available; invasive, gastroenteritis, or environmental. For a full outline of sources, see Table S1 in the supplemental material.

Of these cases, one strain appears to be novel, with an MLST profile that was uncharacterized previously. We observe that this strain is most closely related (1,191 SNPs; see Fig. S2) to a sample that has been serotyped as *S.* Derby and so may represent a serotype switch from *S.* Derby (antigenic formula 1,4,[5],12:f,g:[1,2]), based on the acquisition of the b-type phase 1 flagella. In the case of the other singleton, no close relatives were identified, although other isolates sharing the same sequence type are recorded in the MLST database with a Paratyphi B serotype, suggesting that this observation is not simply the result of a sequencing or serotyping error.

***d*Ta$^-$ strains have limited differences in their virulence repertoire compared with close relatives.** The origin of the confusion around the subdivision of the Paratyphi B complex is not simply that its constituent members possess a shared serotype, it is that a subset of the strains with this serotype has frequently been found to cause invasive disease. Thus, the significance of serotype Paratyphi B is built not only upon the phylogenetic distribution of the *fliC*/*fljB* genes but, also, the distribution of disease phenotypes and virulence genotypes among strains sharing this serotype. To examine the congruence of disease types with phylogenetic positions and known virulence determinants, we examined our sample collection in light of clinical metadata and the known virulence repertoire of these organisms (Fig. 2).

Within PG1, we found that 20/34 of the isolates for which we had clinical metadata were associated with cases of invasive disease. However, while all of the *d*Ta$^-$ samples in our data set are found in PG1, it was apparent that there were also cases of invasive disease caused by isolates from outside this PG. Most notably, 5/17 isolates from PG5 (Fig. 2) were also associated with invasive disease. However, performing a pairwise comparison of the rates of invasive disease between PGs using the χ^2 test reveals that after correcting for multiple sampling, only PG1 shows a significant difference in the rate of invasive disease-causing isolates relative to those of the other PGs (*P* value of 3.9E^{-6} for PG1 versus PG3, 1.2E^{-5} versus PG4, and 0.002 versus PG5 following a Holm-Bonferroni correction). While it is unknown whether the genes affected by the *d*Ta$^-$ SNP are causative or merely indicative in terms of the disease phenotype observed, what is clear is that the only Paratyphi B lineage that is strongly associated with invasive disease is PG1. Moving beyond the classical single-site test for subdividing these lineages, we examined the core and pangenomes of the PGs to look for variation in gene contents, specifically in the virulence repertoire of the Paratyphi B complex. We identified the distribution of known *Salmonella* virulence factors associated with invasive disease, including *Salmonella* pathogenicity islands (SPIs) and their associated type three secretion system (TTSS) effector proteins, as well as previously characterized fimbriae (Fig. 2; see also Fig. S4 in the supplemental material).

Our analysis indicates that across all of the PGs, the complement of virulence-related factors is mostly consistent, with all PGs sharing SPI-1 to -5, -9, and -11. All except PG3, -4, and -6 appeared to possess SPI-6, including the type VI secretion system. All of the lineages other than the poultry-associated PG10 possess SPI-12, -13, -14, and -16, while PG10 alone possesses SPI-8 and -17. Across the population, in terms of the SPI complement, there is therefore little to distinguish the apparently more invasive members of PG1 from most of the PGs of the Paratyphi B complex. One possible explanation could be found in the effector and fimbriae proteins.

While the fimbrial content is broadly similar across the sample set (see Fig. S4 in the supplemental material), there is some variation evident when examining TTSS effectors (Fig. 2). Isolates belonging to PG1 to -5 and -7 possess the effector gene *srfJ*,

whereas members of other PGs lack it, while PG1 has lost the effectors *sseJ* and *steB*, mirroring losses that have also occurred in *S. enterica* serotype Typhi. This is particularly of note since complementation of *S.* Typhi with a functional *sseJ* decreases cytotoxicity (29), a capability that is thought to aid *S.* Typhi in entering the blood stream. *steB* was recently discovered in *S.* Typhimurium (30), and its role in virulence is yet to be elucidated, but its absence in *S.* Paratyphi B and *S.* Typhi is suggestive of a role that could limit the invasiveness of these lineages were it present. Contrary to previous work (31), we found that *sopE* was not present in all of the isolates of PG1. We did observe that 2 of the members of PG1 had a gene homologous to the *sopE* used in the 2003 paper (31); however, this hit was also found in other isolates across our data set, demonstrating that it is not a suitable marker for identification of invasive strains of Paratyphi B. We also investigated the pangenome to identify any genes that were lost by all of PG1 but present in their close relatives and any genes that have been gained by the PG1 ancestor and retained by all samples. This revealed a limited number of gains and losses within the whole group, with 31 genes being apparently lost across PG1 that are present in every other isolate found with the same serotype. Interestingly, eight of these genes are hydrogenases, and one is in the cellulose biosynthesis pathway (see Table S2)—losses that are also found in other invasive lineages, such as *S.* Typhi and S. Gallinarum (27), as well as in host-adapted strains of other enterobacteriaceae, such as *Yersinia pestis* and *Yersinia enterocolitica* (32). These genes have been previously associated with adaptation to the inflamed mammalian gut (32–34), and so these losses would be consistent with an organism that has adapted or is adapting into an invasive niche.

Paratyphi B *d*Ta$^-$ strains share an ancestor predicted to have existed in the 12th century. Collectively, the analyses of the virulence-related characteristics of the Paratyphi B PGs revealed a set of lineages that are relatively consistent in terms of their core gene content, with core genomes for PGs with >2 isolates ranging from 3,951 to 4,511 genes. Members of PG1 share 4,236 genes, a level of gene conservation that suggests that the genomes of the Paratyphi B PGs are stable, with a limited amount of gene gain and loss. To place this into a temporal context and to better understand the natural history of the PGs, we used Bayesian Evolutionary Analysis by Sampling Trees (BEAST) to perform a set of population genomic analyses within the PGs where we had sufficient dating information/coverage to produce robust estimates; these were PG1, -2, -5, -8/-9, and -10 (see Fig. S5 in the supplemental material). PG3, -4, and -6 were too diverse to examine using BEAST, and PG7 had too few isolates. These analyses revealed that the invasive lineage, PG1, is more ancient than may have been predicted based upon the low frequency with which its members cause disease today. The median date predicted by BEAST (35) for the most recent common ancestor (MRCA) of this group is 1188 AD (95% confidence interval [CI], 1799 AD to 469 BC), implying that PG1 is older than the Paratyphi A serotype, whose common ancestor is dated to ~450 years ago (36). This is an interesting finding for two reasons. First, this suggests that the core genome of 4,236 genes has been conserved within PG1 for over 750 years, implying that the genome is very stable. Second, this finding is notable given that Paratyphi A strains are now more frequently isolated than Paratyphi B strains but Paratyphi B appears to have emerged first. In comparison, the other main PGs of clinical significance appear to have emerged more recently. PG8/-9 have an MRCA in 1726 (95% CI, 1880 to 1448), while the poultry-associated PG10 has an MRCA that dates to 1977 (95% CI, 2001 to 1859), pointing toward its recent emergence as a pathogen associated with intensive farming of poultry. Finally, the most recent strains in PG5 have an MRCA dated to the beginning of the 1980s (95% CI, 2008 to 1738). This observation suggests that the clone has recently expanded, a surprising observation given the lack of antimicrobial resistance found in this group. This observation is true of most of the PGs examined. We see generally low levels of inter-PG recombination (see Fig. S6) and very limited evidence of the acquisition of antimicrobial resistance (see Fig. S7). Our analysis reveals that the acquisitions of resistance elements have been single, local events that occurred within PG3 (SGI-1) and PG10 (a chromosomal class 2 integron along with extended-spectrum β-lactamase [ESBL]-encoding plasmids). Subsequently, we only observe evidence for vertical inheritance, with no evidence of the spread of these elements to other lineages.

Conclusion. Using next-generation sequencing, we have been able to uncover much of the genomic basis for the confusion around this serotype. It is clear, when examining the data on a genomic scale, that strains possessing a serotype with the antigenic formula 1,4,[5],12:b:[1,2] can be subdivided into at least 10 groups. Unlike MLST, which could only ever indicate that these groupings were diverse, whole-genome sequencing provides us with the capacity to quantify the divergence between groups and place the resulting data in the context of other inter- and intra-*Salmonella* serotype variation. The analysis presented here demonstrates the advantages of using whole-genome approaches over eBURST groups (eBGs) for subdividing *S.* Paratyphi B. eBGs have previously been suggested as a basis for diagnostics for *S. enterica* (7), but the results presented here clearly show that eBGs do not distinguish *d*Ta$^-$ strains from *d*Ta$^+$ strains, making these unsuitable for rapid diagnostics from either traditional MLST or genome sequence data. This analysis also reveals clearly that, over the last 1,000 years, there have been a set of independent clonal expansions of lineages that share the common serotype with the formula 1,4,5,12:b:1,2, and based upon an examination of the flagellum genes responsible for the phase 1 and phase 2 components of the serotype, these expansions have been predated by recombination events importing one or other of these genes into different chromosomal backgrounds. The number of lineages carrying serotype Paratyphi B is suggestive of the fact that this serotype is successful in a number of niches. Within the population of strains carrying serotype Paratyphi B, there are strains that have been isolated from humans, where they caused invasive disease or gastroenteritis, from poultry, and from aquatic organisms and/or the aquatic environment. This large number of lineages carrying the same serotype is suggestive that serotype switching is more frequent than in other successful lineages, such as *S.* Typhimurium, which have remained discrete lineages.

Conversely, we find that both the core genome more generally and lipopolysaccharide (LPS) genes specifically are remarkably stable, with the genes within the O-antigen cluster producing a phylogeny mirroring that produced by the core genome. Of the 10 PGs that we identified, one (PG1) comprised the canonical Paratyphi B group. This group is relatively closely related to PG2 to -5, producing a larger cluster of lineages where the intergroup divergence is comparable to that found in other closely related host generalist/host-specialized lineages, such as the *S.* Enteritidis/Gallinarum group. Examining phylogenetic clustering in the context of disease type, it is clear, however, that invasive disease is not

limited to PG1 to -3; PG5, -6, -9, and -10 can all cause invasive disease, albeit to various extents. However, based upon our sample data set, only PG1 is significantly associated with invasive disease. While isolates that are *d*Ta$^-$ have historically been classified as "invasive," isolates from PG5 may represent a lineage, similar to the invasive sequence type 313 (ST313) of *S.* Typhimurium (37), that has a higher propensity to cause invasive disease given particular host-associated factors (advanced age, suppressed immune system, etc.) but which is not an "invasive" lineage *per se*. It is our hope that by defining the population structure of this serotype, it will be possible to investigate this question more precisely in the future.

Although there are fewer clues as to the genomic basis for the difference in disease type than may be expected, given the difference in symptoms and infection sites, there are several meaningful signals—most notably, the variations in the complements of effectors and the presence of SPIs—that point toward the genomic differences that underpin the observed disease types. The variations in SPI and effector presence in particular are significant for two reasons. First, the degradation of these elements in lineages may be mechanistic, and since they are in virulence-related factors, this degradation may relate directly to the invasiveness of the isolates. Second, because SPIs/effectors are differentially present within lineages within the Paratyphi B complex, they may provide possible diagnostic testing targets that are linked to the machinery actually used to cause disease, rather than characteristics which are probably related to disease but not causal, such as the tartrate test. In particular, the presence/absence of effectors *sseJ* and *steB* would appear to provide a simple mechanism for identifying PG1 (and *S.* Paratyphi A and *S.* Typhi) using PCR, potentially reducing the time taken to detect BSL3 *d*Ta$^-$ *S.* Paratyphi B from 13 days to a few hours in virtually any laboratory in the world.

This work underlines the difficulty posed when genomic approaches are not used to subdivide lineages and exemplifies the challenges that face classical typing, reinforcing the need for unambiguous molecular methods for characterizing members of the Paratyphi B complex. This work also demonstrates both the limitations of genomics alone to unpick the complex biological processes that translate genotype to phenotype and a methodological framework that can be used to explore other polyphyletic *S. enterica* serotypes where variations in gene content have been observed, such as 4,[5]12:b:−, using PCR-based approaches (38). The lack of consistent genomic differences between PG1 to -5 suggests that the invasiveness of PG1 may be due to hitherto undiscovered virulence determinants or to other factors, such as transcriptional control. The loss of hydrogenases and elements of the cellulose biosynthesis pathway—established components of the blueprint for invasive salmonellae—is also suggestive that metabolic changes have also occurred within PG1 as these organisms adapt to an invasive niche but that Paratyphi B may represent a sort of evolutionary halfway house, sitting somewhere between a host-adapted and invasive serotype. This work therefore provides a basis for reinterpreting what we already know about invasive salmonellae, providing a simple practical basis for distinguishing the invasive PG1 strains from other, noninvasive strains, while building a foundation for future work to better understand what makes *d*Ta$^-$/PG1 strains so invasive compared to their close *d*Ta$^+$ relatives in PG2 to -5.

MATERIALS AND METHODS

Samples. In order to explore the population structure of *Salmonella* Paratyphi B, as defined by isolates sharing the antigenic formula 1,4,[5],12:b:[1,2], we sequenced the genomes of a collection of 191 isolates collected from the United Kingdom, France, Spain, Ireland, and Canada that encompasses all currently known MLST eBURST groups that are labeled as being serotype Paratyphi B, including both monophasic and diphasic strains. In addition to the deliberate selection of a range of historical strains that covers the full population of this serotype ($n = 81$), we also collected samples from clinical episodes of disease in the United Kingdom ($n = 63$) and Spain ($n = 33$), as well as isolates from poultry-associated disease reported previously ($n = 14$) (see Table S1 in the supplemental material for a full list of the sources and strains used in this study) (11). All of our strains were classically serotyped prior to sequencing by the respective originating reference laboratory: the Spanish National Reference Laboratory, the Institut Pasteur, the *Salmonella* Reference Laboratory of Health Protection England, or the National Microbiology Laboratory, Public Health Agency of Canada. This classical serotyping was confirmed using genome sequencing. To provide a wider phylogenetic context, we also include within our study a further 27 *Salmonella enterica* isolates, representing isolates carrying 18 different serotypes. Of these, 21 are previously published strains and 6 represent new sequences that were generated as part of this study. Additionally, we include a further 2 isolates that are of serotype Dublin/Enteritidis but are grouped by MLST with the predominantly Paratyphi B eBG 32 (see Table S1 for a full list of strains, with accession numbers, serotype information, and other relevant metadata).

Genome sequencing. The genomes were sequenced using the Illumina sequencing platform, with Genome Analyzer IIx (GAIIx), HiSeq, and MiSeq instruments being used to sequence isolates to approximately 200× coverage, as described previously (39). The samples were generated with a mean insert size of between 200 and 300 bp, and depending upon the instrument used, underwent 2 × 50 bp paired-end sequencing (GAIIx), 2 × 100 bp paired-end sequencing (HiSeq), or 2 × 250 bp paired-end sequencing (MiSeq). The data were assembled *de novo* using Velvet (40), with assemblies improved using the Velvet Columbus module and the software package iCORN (41).

Phylogenetic analysis and variant calling. Using the tool Snippy running on the Cloud Infrastructure for Microbial Bioinformatics (46), we performed mapping against the reference Paratyphi B strain SPB7, identifying 147,963 positions that are present in all samples in our collection but vary in at least one isolate. Extracting these variable sites, we removed putative regions of recombination, using the tool Gubbins (25), to produce a set of core positions that are free from recombination. These were then used to generate a phylogenetic tree using RAxML, which was generated using a general time-reversible model with gamma correction (GTR-gamma) for between-site heterogeneity. Additionally, in order to better quantify the core/accessory genomes for the isolate collection, we made use of the large-scale Blast score ratio (LS-BSR) tool (42), which was run against the assembled genomes of the complete data set. The matrix generated by LS-BSR was then processed using the PanGP tool to visualize the size of the core genome across the data set.

Comparative genomics. We identified genes of interest through literature searches and examination of annotated *Salmonella* genomes, producing a set of query sequences to explore the SPI complement and to allow us to find and extract fimbriae and effectors from our *de novo* assemblies. To examine the SPI contents, we identified the genes present within the pangenome that are carried on SPI-1 to -22, and using these, calculated the number of genes associated with each SPI that were represented in each sample. To examine fimbria, effector, and flagellum genes and the O-antigen gene cluster, we located and screened genes or regions of interest using the LS-BSR tool (42). We then visualized the results using a simple script that converts comma-separated-value (CSV) tables into vector graphics, developed in house (fimbriae, SPIs, and effectors—available at https://github.com/tomrconnor/Basicscripts). To

extract genes, we performed a Blast search across the assembled genomes for targets of interest using an approach described previously (32), extracting sequences, aligning them with MUSCLE (43), and generating trees using a GTR-gamma model under PhyML (44). To confirm the presence of *sopE* in Paratyphi B samples, we used SRST2 in addition to the approach described above. SRST2 (45) uses a mapping-based approach to map sequence reads against a set of target sequences, identifying whether there is evidence for those sequences within a file of sequence reads from a genomic sample. This complements the assembly-based approach and compensates for potential limitations around the detection of genes from assemblies. The results generated from this analysis are reproduced in Table S3 in the supplemental material.

Population genetic analyses. To subdivide the population, we made use of the software package Bayesian Analysis of Population Structure (BAPS) (26). We provided the software with the mapping-based SNP alignment for the data set prior to the removal of recombinations and performed a hierarchical BAPS (26) run to 2 levels with a maximum number of 50 populations, using the second level of clustering to define phylogroups (PGs) from the population. We imposed an artificial limitation that a phylogroup must contain a minimum of 2 isolates. The analysis was run three times to confirm the clustering results. As well as identifying a set of PGs containing 2 isolates or more, the algorithm also identified a number of other isolates that may or may not constitute new PGs, and we anticipate that these candidate PGs will be confirmed (or not) over time. To estimate a dated phylogeny, we made use of BEAST (35) 1.8 and performed the analysis on an SNP alignment for the isolates that we had good dating information on. Performing BEAST on each PG individually, we used three chains with a total length of 100,000,000 states each and with trees sampled every 10,000 states for each data set. To identify the best combination of models to use, for each sample set, we performed this analysis for constant and lognormal clock models and for constant, logistics, expansion, exponential, and skyline demographic models. For each data set, the runs that converged and generated effective sample size (ESS) values of >200 were compared, and the best model was selected based on the AICM (Akaike's information criterion through Markov chain Monte Carlo), calculated using Tracer. The best models determined on this basis for each run were as follows: a lognormal clock and skyline model for PG1, a lognormal clock and expansion model for PG2, a lognormal and skyline model for PG5, a lognormal and skyline model for PG8/-9, and a lognormal and expansion model for PG10. In the case of PG3 and -4, BEAST did produce results, but the predicted MRCA for the best model was over 300 years in the past, despite the fact that we only had samples going back 15 to 20 years. In both of these cases, we concluded that our sample was too diverse to derive accurate BEAST results. In the case of PG6, we did not have enough samples with good date information for BEAST to produce usable results. For the final selected BEAST runs, the ESS values were >>200 in all cases. The tree files were combined using LogCombiner, processed using TreeAnnotator, and visualized with FigTree, all tools that are part of or published with the BEAST package and are freely available from http://beast.bio.ed.ac.uk.

Accession numbers. The sequence data for this project have been deposited in the European Nucleotide Archive. Please see Table S1 for accession numbers and metadata for the individual strains examined.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00527-16/-/DCSupplemental.

Figure S1, PDF file, 0.2 MB.
Figure S2, PDF file, 0.03 MB.
Figure S3, PDF file, 0.05 MB.
Figure S4, PDF file, 0.1 MB.
Figure S5, PDF file, 0.2 MB.
Figure S6, PDF file, 0.9 MB.
Figure S7, PDF file, 0.1 MB.
Table S1, XLSX file, 0.1 MB.
Table S2, XLSX file, 0.02 MB.
Table S3, XLSX file, 0.04 MB.

ACKNOWLEDGMENTS

We thank David Harris and the sequencing teams at the Sanger Institute for sequencing the samples and Rebecca Connor for providing assistance with proofreading.

T.R.C., N.T., G.L., T.F., and M.F. all received Wellcome Trust funding as part of this project (grant number 098051). T.R.C.: The Bioinformatics analysis used resources funded by the MRC (grant number MR/L015080/1) and Cardiff University (data storage was funded by the Cardiff University Research Infrastructure Fund).

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

FUNDING INFORMATION

This work, including the efforts of Thomas R. Connor, Gemma Langridge, Theresa Feltwell, Julian Parkhill, and Nicholas Thomson, was funded by Wellcome Trust (098051). This work, including the efforts of Thomas R. Connor, was funded by Medical Research Council (MRC) (MR/L015080/1).

F.-X.W. and S.L. are funded by the Institut Pasteur, the Institut de Veille Sanitaire, and the French Government's Investissement d'Avenir program Integrative Biology of Emerging Infectious Diseases Laboratory of Excellence (grant number ANR-10-LABX-62-IBEID).

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication

REFERENCES

1. **Kauffmann F.** 1941. Die bakteriologie der Salmonella-gruppe. E Munksgaard, Copenhagen, Denmark.
2. **Kauffmann F.** 1954. Enterobacteriaceae, 2nd ed. E Munksgaard Publishers, Copenhagen, Denmark.
3. **Kauffmann F.** 1955. Zur Differentialdiagnose und Pathogenität von Salmonella java und Salmonella paratyphi B. Zeitschr Hygiene **141:**546–550. http://dx.doi.org/10.1007/BF02156850.
4. **Le Minor L.** 1988. Typing of Salmonella species. Eur J Clin Microbiol Infect Dis **7:**214–218. http://dx.doi.org/10.1007/BF01963091.
5. **Malorny B, Bunge C, Helmuth R.** 2003. Discrimination of d-tartrate-fermenting and -nonfermenting Salmonella enterica subsp. enterica isolates by genotypic and phenotypic methods. J Clin Microbiol **41:** 4292–4297. http://dx.doi.org/10.1128/JCM.41.9.4292-4297.2003.
6. **Ezquerra E, Burnens A, Jones C, Stanley J.** 1993. Genotypic typing and phylogenetic analysis of Salmonella paratyphi B and S. java with IS200. J Gen Microbiol **139:**2409–2414. http://dx.doi.org/10.1099/00221287-139-10-2409.
7. **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 Pathog **8:**e1002776. http://dx.doi.org/10.1371/journal.ppat.1002776.
8. **Fabre L, Zhang J, Guigon G, Le Hello S, Guibert V, Accou-Demartin M, de Romans S, Lim C, Roux C, Passet V, Diancourt L, Guibourdenche M, Issenhuth-Jeanjean S, Achtman M, Brisse S, Sola C, Weill FX.** 2012. CRISPR typing and subtyping for improved laboratory surveillance of Salmonella infections. PLoS One **7:**e36995. http://dx.doi.org/10.1371/journal.pone.0036995.
9. **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. J Bacteriol **192:**6465–6476. http://dx.doi.org/10.1128/JB.00969-10.
10. **Francis S, Rowland J, Rattenbury K, Powell D, Rogers WN, Ward L, Palmer SR.** 1989. An outbreak of paratyphoid fever in the UK associated with a fish-and-chip shop. Epidemiol Infect **103:**445–448. http://dx.doi.org/10.1017/S0950268800030843.
11. **Doublet B, Praud K, Nguyen-Ho-Bao T, Argudín MA, Bertrand S, Butaye P, Cloeckaert A.** 2014. Extended-spectrum beta-lactamase- and

AmpC beta-lactamase-producing D-tartrate-positive Salmonella enterica serovar Paratyphi B from broilers and human patients in Belgium, 2008–10. J Antimicrob Chemother **69:**1257–1264. http://dx.doi.org/10.1093/jac/dkt504.

12. **Denny J, Threlfall J, Takkinen J, Lofdahl S, Westrell T, Varela C, Adak B, Boxall N, Ethelberg S, Torpdahl M, Straetemans M, van Pelt W.** 2007. Multinational Salmonella Paratyphi B variant Java (Salmonella Java) outbreak, August–December 2007. Euro Surveill **12:**E071220–E071222. http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=3332.
13. **Gaulin C, Vincent C, Alain L, Ismail J.** 2002. Outbreak of Salmonella paratyphi B linked to aquariums in the province of Quebec, 2000. Can Commun Dis Rep **28:**89–93.
14. **Stratton J, Stefaniw L, Grimsrud K, Werker DH, Ellis A, Ashton E, Chui L, Blewett E, Ahmed R, Clark C, Rodgers F, Trottier L, Jensen B.** 2001. Outbreak of Salmonella paratyphi B var java due to contaminated alfalfa sprouts in Alberta, British Columbia and Saskatchewan. Can Commun Dis Rep **27:**133–137.
15. **Centers for Disease Control and Prevention.** 2008. Multistate outbreak of human Salmonella infections associated with exposure to turtles—United States, 2007-2008. MMWR Morb Mortal Wkly Rep **57:**69–72. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5703a3.htm.
16. **Kendall ME, Crim S, Fullerton K, Han PV, Cronquist AB, Shiferaw B, Ingram LA, Rounds J, Mintz ED, Mahon BE.** 2012. Travel-associated enteric infections diagnosed after return to the United States, Foodborne Diseases Active Surveillance Network (FoodNet), 2004–2009. Clin Infect Dis **54**(Suppl 5):S480–S487. http://dx.doi.org/10.1093/cid/cis052.
17. **Gobin M, Launders N, Lane C, Kafatos G, Adak B.** 2011. National outbreak of Salmonella Java phage type 3b variant 9 infection using parallel case-control and case-case study designs, United Kingdom, July to October 2010. Euro Surveill **16:**20023. http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20023.
18. **Davies R, Deuchande R, Larki L, Collins R, Irvine RM.** 2013. Multidrug resistant salmonella Java found in British broiler flocks. Vet Rec **172:**617–618. http://dx.doi.org/10.1136/vr.f3664.
19. **Weill FX, Fabre L, Grandry B, Grimont PA, Casin I.** 2005. Multiple-antibiotic resistance in Salmonella enterica serotype Paratyphi B isolates collected in France between 2000 and 2003 is due mainly to strains harboring Salmonella genomic islands 1, 1-B, and 1-C. Antimicrob Agents Chemother **49:**2793–2801. http://dx.doi.org/10.1128/AAC.49.7.2793-2801.2005.
20. **Miko A, Pries K, Schroeter A, Helmuth R.** 2003. Multiple-drug resistance in D-tartrate-positive Salmonella enterica serovar paratyphi B isolates from poultry is mediated by class 2 integrons inserted into the bacterial chromosome. Antimicrob Agents Chemother **47:**3640–3643. http://dx.doi.org/10.1128/AAC.47.11.3640-3643.2003.
21. **Brenner FW, Villar RG, Angulo FJ, Tauxe R, Swaminathan B.** 2000. Salmonella nomenclature. J Clin Microbiol **38:**2465–2467.
22. **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. Res Microbiol **165:**526–530. http://dx.doi.org/10.1016/j.resmic.2014.07.004.
23. **Pallen MJ, Wren BW.** 2007. Bacterial pathogenomics. Nature **449:**835–842. http://dx.doi.org/10.1038/nature06248.
24. **Fu S, Octavia S, Tanaka MM, Sintchenko V, Lan R.** 2015. Defining the core genome of Salmonella enterica serovar Typhimurium for genomic surveillance and epidemiological typing. J Clin Microbiol **53:**2530–2538. http://dx.doi.org/10.1128/JCM.03407-14.
25. **Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR.** 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res **43:** http://dx.doi.org/10.1093/nar/gku1196.
26. **Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J.** 2013. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Mol Biol Evol **30:**1224–1228. http://dx.doi.org/10.1093/molbev/mst028.
27. **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 Res **18:**1624–1637. http://dx.doi.org/10.1101/gr.077404.108.
28. **Kauffmann F.** 1953. On the transduction of serological properties in the Salmonella group. Acta Pathol Microbiol Scand **33:**409–420. http://dx.doi.org/10.1111/j.1699-0463.1953.tb01537.x.
29. **Trombert AN, Berrocal L, Fuentes JA, Mora GC.** 2010. S. Typhimurium sseJ gene decreases the S. Typhi cytotoxicity toward cultured epithelial cells. BMC Microbiol **10:**312. http://dx.doi.org/10.1186/1471-2180-10-312.
30. **Geddes K, Worley M, Niemann G, Heffron F.** 2005. Identification of new secreted effectors in Salmonella enterica serovar Typhimurium. Infect Immun **73:**6260–6271. http://dx.doi.org/10.1128/IAI.73.10.6260-6271.2005.
31. **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. J Clin Microbiol **41:**4270–4278. http://dx.doi.org/10.1128/JCM.41.9.4270-4278.2003.
32. **Reuter S, Connor TR, Barquist L, Walker D, Feltwell T, Harris SR, Fookes M, Hall ME, Petty NK, Fuchs TM, Corander J, Dufour M, Ringwood T, Savin C, Bouchier C, Martin L, Miettinen M, Shubin M, Riehm JM, Laukkanen-Ninios R, Sihvonen LM, Siitonen A, Skurnik M, Falcao JP, Fukushima H, Scholz HC, Prentice MB, Wren BW, Parkhill J, Carniel E, Achtman M, McNally A, Thomson NR.** 2014. Parallel independent evolution of pathogenicity within the genus Yersinia. Proc Natl Acad Sci U S A **111:**6768–6773. http://dx.doi.org/10.1073/pnas.1317161111.
33. **Barrett EL, Clark MA.** 1987. Tetrathionate reduction and production of hydrogen sulfide from thiosulfate. Microbiol Res **51:**192–205.
34. **Winter SE, Thiennimitr P, Winter MG, Butler BP, Huseby DL, Crawford RW, Russell JM, Bevins CL, Adams LG, Tsolis RM, Roth JR, Bäumler AJ.** 2010. Gut inflammation provides a respiratory electron acceptor for salmonella. Nature **467:**426–429. http://dx.doi.org/10.1038/nature09415.
35. **Drummond AJ, Suchard MA, Xie D, Rambaut A.** 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol **29:**1969–1973. http://dx.doi.org/10.1093/molbev/mss075.
36. **Zhou Z, McCann A, Weill FX, Blin C, Nair S, Wain J, Dougan G, Achtman M.** 2014. Transient Darwinian selection in Salmonella enterica serovar Paratyphi A during 450 years of global spread of enteric fever. Proc Natl Acad Sci U S A **111:**12199–12204. http://dx.doi.org/10.1073/pnas.1411012111.
37. **Okoro CK, Kingsley RA, Connor TR, Harris SR, Parry CM, Al-Mashhadani MN, Kariuki S, Msefula CL, Gordon MA, de Pinna E, Wain J, Heyderman RS, Obaro S, Alonso PL, Mandomando I, MacLennan CA, Tapia MD, Levine MM, Tennant SM, Parkhill J, Dougan G.** 2012. Intracontinental spread of human invasive Salmonella Typhimurium pathovariants in sub-Saharan Africa. Nat Genet **44:**1215–1221. http://dx.doi.org/10.1038/ng.2423.
38. **Toboldt A, Tietze E, Helmuth R, Junker E, Fruth A, Malorny B.** 2013. Population structure of Salmonella enterica serovar 4,[5],12:b:– strains and likely sources of human infection. Appl Environ Microbiol **79:**5121–5129. http://dx.doi.org/10.1128/AEM.01735-13.
39. **Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y.** 2012. A tale of three next generation sequencing platforms: comparison of Ion torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics **13:**341. http://dx.doi.org/10.1186/1471-2164-13-341.
40. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **18:**821–829. http://dx.doi.org/10.1101/gr.074492.107.
41. **Otto TD, Sanders M, Berriman M, Newbold C.** 2010. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. BioInformatics **26:**1704–1707. http://dx.doi.org/10.1093/bioinformatics/btq269.
42. **Sahl JW, Caporaso JG, Rasko DA, Keim P.** 2014. The large-scale Blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. PeerJ **2:**e332. http://dx.doi.org/10.7717/peerj.332.
43. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:**1792–1797. http://dx.doi.org/10.1093/nar/gkh340.
44. **Guindon S, Delsuc F, Dufayard JF, Gascuel O.** 2009. Estimating maxi-

mum likelihood phylogenies with PhyML. Methods Mol Biol **537:** 113–137. http://dx.doi.org/10.1007/978-1-59745-251-9_6.

45. **Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, Zobel J, Holt KE.** 2014. SRST2: rapid genomic surveillance for public health and hospital microbiology labs. Genome Med **6:**90. http://dx.doi.org/10.1186/s13073-014-0090-6.

46. **Connor TR, Loman NJ, Thompson S, Smith A, Southgate J, Poplawski R, Bull MJ, Richardson E, Ismail M, Elwood-Thompson S, Kitchen C, Guest M, Bakke M, Sheppard SK, Pallen MJ.** CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. http://biorxiv.org/content/early/2016/07/19/064451.