Population structure analysis and laboratory monitoring of *Shigella* with a standardised core-genome multilocus sequence typing scheme: a validation study

Iman Yassine, MSc[1,2], Sophie Lefèvre, PharmD, PhD[1], Elisabeth E. Hansen, MSc[1‡], Corinne Ruckly[1], Isabelle Carle[1], Monique Lejay-Collin, BSc[1], Laëtitia Fabre, PhD[1], Rayane Rafei, PhD[2], Dominique Clermont, PhD[3], Maria Pardos de la Gandara MD, PhD[1], Fouad Dabboussi, PhD[2], Nicholas R. Thomson, PhD[4,5], François-Xavier Weill, MD, PhD[1]*

[1]Institut Pasteur, Unité des bactéries pathogènes entériques, Centre National de Référence des *Escherichia coli*, *Shigella* et *Salmonella*, Paris, 75015, France

[2]Laboratoire Microbiologie Santé et Environnement (LMSE), Doctoral School of Sciences and Technology, Faculty of Public Health, Lebanese University, Tripoli, Lebanon

[3]Institut Pasteur, Collection de l'Institut Pasteur, Paris, 75015, France

[4] Wellcome Sanger Institute, Cambridge, CB10 1SA, United Kingdom

[5]London School of Hygiene and Tropical Medicine, London, WC1E 7HT, United Kingdom

[‡]current affiliation: Harvard Medical School, Boston, 02115, United States

*corresponding author: francois-xavier.weill@pasteur.fr (F.- X. Weill).

Abstract

Background

The laboratory surveillance of bacillary dysentery is based on a *Shigella* typing scheme standardised in the late 1940s. This scheme classifies *Shigella* strains into four serogroups and more than 50 serotypes on the basis of biochemical tests and lipopolysaccharide O-antigen serotyping. Real-time genomic surveillance of *Shigella* infections has been implemented in several countries, but without the use of a standardised high-resolution typing scheme.

Methods

We studied over 4,000 reference strains and clinical isolates of *Shigella*, covering all serotypes, including provisional serotypes and atypical strains, with the current serotyping scheme. These strains and isolates were also subjected to whole-genome sequencing and analysis with the EnteroBase *Escherichia*/*Shigella* 2,513-locus core-genome multilocus sequence typing (cgMLST) scheme.

Findings

The *Shigella* genomes were grouped into eight phylogenetically distinct clusters, within the *E. coli* species. Three of these clusters contained strains from different serogroups and serotypes, the remaining five each consisting of a single serotype. The cgMLST hierarchical clustering (HC) analysis at different levels of resolution (HC2000 to HC400) recognised the natural groupings for *Shigella*. By contrast, the serotyping scheme was affected by horizontal gene transfer, leading to a conflation of genetically unrelated *Shigella* strains and a separation of some genetically related strains. We also curated the various provisional serotypes reported in the literature and described five new *Shigella* serotypes for addition to the typing scheme.

Interpretation

The EnteroBase *Escherichia*/*Shigella* cgMLST is a standardised, robust, portable, and high-resolution scheme that will enhance the laboratory surveillance of *Shigella* infections, particularly for *Shigella flexneri*. However, cgMLST data should be considered together with *in silico* serotyping data, to maintain backward compatibility with the current *Shigella* serotyping scheme.

Funding

Institut Pasteur; *Santé publique France*; Fellowship from Association AZM and Saadeh, and the Lebanese Association for Scientific Research (LASeR) for IY. NRT is funded by the Wellcome Trust (grant 206194). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Introduction

Shigella belongs to the *Enterobacteriaceae* family, and causes bacillary dysentery, a common cause of diarrhoea in low- and middle-income countries. It has been estimated that this intracellular human pathogen, which is transmitted via the faecal-oral route with very low infectious dose (10-100 cells), is responsible for over 210,000 deaths per year, mostly in children under the age of five years.[1–3] In high-income countries, *Shigella* infections also occur in travellers and in some high-risk groups, such as men who have sex with men (MSM) and Orthodox Jewish communities.[1,3] The morbidity of these infections is currently increasing due to growing resistance to antimicrobial drugs in these bacteria.[2,3]

Laboratory surveillance of *Shigella* infections was initiated several decades ago, and was facilitated by the adoption of a standardised *Shigella* typing scheme in the late 1940s.[4] This scheme, which is still in use today, is based on biochemical tests and serotyping (slide agglutination with typing sera directed against the different *Shigella* lipopolysaccharide O-antigens). It splits the *Shigella* genus into four serogroups (originally considered to be species): *Shigella dysenteriae*, *S. boydii*, *S. flexneri*, and *S. sonnei*; these four serogroups are then subdivided into more than 50 serotypes. However, modern population genetics methods, such as multilocus sequence typing (MLST) analysis, and, more recently, core-genome single-nucleotide variant (cgSNV) analysis, have shown that *Shigella* forms distinct lineages within the species *E. coli*, from which it emerged following the acquisition of a large virulence plasmid (VP) enabling the bacterium to invade intestinal cells.[5–8] In parallel, these host-restricted pathogens converged independently on the *Shigella* phenotype (non-motility, no decarboxylation of lysine, no use of citrate and malonate, and other characteristics, as reported by Pupo and colleagues[5]) through genome degradation. Furthermore, these recent methods have shown that the current typing scheme does not capture the natural groupings of this pathogen.[5] Some molecular data have been taken into account in an update of the *Shigella* serotyping scheme. *S. boydii* 13, for example, was withdrawn from the classification, because it was shown to belong to another species, *E. albertii*, and did not contain the VP.[9,10]

In an increasing number of countries, the laboratory surveillance of *Shigella* infections has now passed from conventional serotyping to real-time genomic surveillance.[7,11] The genomic methods used were developed recently, and most of their targets lie within the O-antigen gene cluster (*rfb*) or in the *S. flexneri* serotype-converting prophages, to ensure serotype specificity.[11,12] Several other genes in the accessory genome were recently targeted, resulting in the assignment of *Shigella* serotypes to eight clusters.[13] These methods undoubtedly facilitate backward compatibility between the genomic and serotyping data, but do not fully exploit the unprecedented resolution of genomics. An extension of the MLST method to cover a large number of core-genome genes has been developed. This high-resolution method, core-genome MLST (cgMLST), has been successfully used in the surveillance of many pathogens, including *Listeria monocytogenes*,[14] and *Salmonella enterica*.[15] Furthermore, cgMLST data are easy to interpret with clustering threshold methods, such as the hierarchical clustering (HierCC) implemented in EnteroBase.[15] However, cgMLST has never been used for the comprehensive description of *Shigella* populations, and the utility of this method for the genomic surveillance of *Shigella* infections has not previously been assessed.

In this study, by analysing over 4,000 genomes from phenotypically characterised *Shigella* strains representative of the global diversity of this pathovar of *E. coli*, we aimed: i) to resolve the population structure of *Shigella* by cgMLST, (ii) to create a dictionary of correspondence between cgMLST HC and serotyping data, and (iii) to update the *Shigella* serotyping scheme by describing new serotypes. We demonstrate that the combination of cgMLST HC with *rfb* gene cluster analysis would enhance the laboratory surveillance of *Shigella* infections, while maintaining backward compatibility with the current serotyping scheme.

Methods

Strains selection and typing

In total, 4,187 *Shigella* reference strains and clinical isolates from the French National Reference Centre for *E. coli*, *Shigella*, and *Salmonella* (FNRC-ESS), Institut Pasteur, Paris were studied (appendix 1). The collection consisted of two datasets. The first dataset – the reference dataset – consisted of 317 *Shigella* reference strains covering all the known serotypes – including provisional serotypes – of the four serogroups (at least one strain per serotype); most of the strains studied were historical strains from various geographic locations and time periods. The second dataset – the routine dataset – consisted of 3,870 clinical isolates (of the 3,942 isolates received) sequenced by the FNRC-ESS between 2017 and 2020 in the framework of the French national surveillance programme for *Shigella* infections. All these strains and isolates were thoroughly characterised with a panel of biochemical tests and serotyped by slide agglutination assays according to standard protocols, as previously described[16] (appendix 2 p 2).

DNA extraction and sequencing

The 4,187 strains and isolates were processed and sequenced with various Illumina platforms (appendix 2 p 2).

Other studied genomes

With the aim of capturing the broadest possible diversity of S*higella* populations, we searched the *E. coli*/*Shigella* database in EnteroBase,[15] and selected 81 additional *Shigella* genomes (reference+ dataset) not originating from the Institut Pasteur (appendix 2 p 2). We included 27 enteroinvasive *E. coli* (EIEC) and 68 *E. coli* strains from the ECOR collection (appendix 2 p 2), to place our *Shigella* genomes in the phylogenetic context of the broader diversity of *E. coli*. We also used the closed PacBio sequences available for all *Shigella* serotypes and described by Kim and colleagues,[17] to study the genetic organisation of the *rfb* gene cluster or various operons described in the "Gene analyses" section. However, these closed genomes were not included in the cgMLST analysis, as they were not edited with short reads and the numerous indels in the homopolymers therefore altered the allelic distances (appendix 2 p 7).

Characterisation of the O-antigen gene clusters

The *Shigella* O-antigen biosynthetic gene (*rfb*) cluster was analysed after extraction of the region between the housekeeping genes *galF* (encoding UTP-glucose-1-phosphate uridylyltransferase) and *gnd* (encoding 6-phosphogluconate dehydrogenase), which are known to flank the *rfb* cluster.[18] Newly identified *rfb* clusters were annotated based on a previously annotated closely matched *E. coli* cluster in the NCBI BLASTn nucleotide collection (nr/nt) database (100% coverage and at least 99% identity) or with ORFfinder (https://www.ncbi.nlm.nih.gov/orffinder/) when no matching cluster was found in the NCBI BLAST database (https://blast.ncbi.nlm.nih.gov/Blast.cgi). The GenBank accession codes of all the *Shigella rfb* clusters are listed in appendix 2 p 8. We also used three tools for *in silico* serotyping: SeroPred, the serotype prediction tool implemented in EnteroBase,[15] ShigaTyper,[11] and ShigEiFinder.[13] Short-read and SPAdes assemblies were used for ShigaTyper and ShigEiFinder, respectively.

Phylogenetic analyses

We used the *Escherichia*/*Shigella* cgMLST scheme (2,513 loci) implemented in EnteroBase to study our genomic datasets (appendix 2 p 2).[19] The cgMLST trees were inferred with the NINJA NJ algorithm, based on the "cgMLST V1 + HierCC" scheme. We visualised the cgMLST data with GrapeTree.[20] We also performed cgSNV analysis, to assess the phylogenetic relationships of 398 *Shigella* (317 from the reference dataset and 81 from the reference+ dataset) and 95 *E. coli* (68 ECOR and 27 EIEC) strains (appendix 2 p 2). A phylogenetic tree of *rfb* sequences was constructed with the sequences from 43 *Shigella* (appendix 2 p 2 and p 8) and 196 *E. coli* isolates from DebRoy and colleagues.[18]

Gene analyses

The presence of the *ipaH* gene, a multicopy gene unique to *Shigella* and EIEC,[21] the presence and structure of the mannitol (*mtl*),[22] raffinose,[23] and tryptophanase (*tna*) operons, [24] and the type of the O-antigen gene cluster (*rfb*) were determined on SPAdes assemblies using the NCBI BLASTn tool (https://blast.ncbi.nlm.nih.gov/Blast.cgi). The target sequences are described in appendix 2 p 9.

Data Availability Statement

Short-read sequence data were submitted to EnteroBase (https://enterobase.warwick.ac.uk/),[15] and to the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home) under study numbers PRJEB44801, PRJEB2846, and PRJEB2128. The GrapeTree of the reference and reference+ datasets is publicly available from EnteroBase (http://enterobase.warwick.ac.uk/ms_tree?tree_id=55118). All the GenBank and ENA accession numbers of the genomes used in this study are listed in appendix 1.

Results

We assembled and sequenced a collection of 317 *Shigella* strains chosen on the basis of their representativeness of the known diversity of *Shigella* populations (i.e., covering all serogroups and serotypes, and the different lineages or phylogroups of *S. sonnei* and *S. flexneri*). The genomic diversity of this reference dataset was increased further, by adding another 81 publicly available *Shigella* genomes. The 398 genomes studied were from strains belonging to the *S. flexneri* (*n* = 191), *S. dysenteriae* (*n* = 83), *S. boydii* (*n* = 80), and *S. sonnei* (*n* = 44) serogroups (appendix 2 pp 10-11). We determined the wider phylogenetic context of these *Shigella* genomes, by also analysing 95 *E. coli* genomes, including 27 EIEC from eight different EIEC genomic clusters and 68 (of the 72) strains from the ECOR collection, considered representative of the diversity of natural populations of *E. coli*.[25] These 493 genomes were studied by two different approaches: the EnteroBase *Escherichia*/*Shigella* cgMLST scheme and SNV-based clustering.

According to cgMLST, all these genomes belonged to the same hierarchical cluster, HC2350_1 (appendix 1). As expected, all the *Shigella* and EIEC genomes contained the pathogenicity gene *ipaH*, whereas the ECOR genomes did not (appendix 2 p 16). A NINJA neighbour-joining (NJ) tree of core genomic allelic distances was generated with the dataset for the 493 *Shigella* and *E. coli* genomes (Fig. 1). Visual examination of the colour-coded HC2000 tree revealed that the *Shigella* genomes were grouped into eight different HC2000 clusters (Fig. 1B). Seven of these HC2000 clusters contained exclusively *Shigella* genomes. The eighth, HC2000_2, contained *S. dysenteriae* type 8 and *E. coli* (EIEC and ECOR) genomes. Four HC2000 clusters contained *Shigella* genomes from a single serotype: HC2000_305 (*S. sonnei*), HC2000_1463 (*S. dysenteriae* type 1), HC2000_44944 (*S. dysenteriae* 10), and HC2000_45542 (*S. boydii* 12).

These clusters are referred to below as SON, SD1, SD10, and SB12, respectively. Three clusters, HC2000_1465, HC2000_4118, and HC2000_192, consisted of multiple serogroups and serotypes (Figs. 1-4). The first of these clusters, HC2000_1465, contained various serotypes of *S. dysenteriae* (3, 4-7, 9, 11-15, provisional (prov.) 93-119, prov. SH-103, prov. 97-10607, prov. SH-105, prov. 96-3162 and prov. 204/96), *S. boydii* (1-4, 6, 8, 10, 11, 14, 18-20, and prov. 07-6597), and *S. flexneri* type 6 (Fig. 2), consistent with Cluster 1 described by Pupo and colleagues in their MLST analysis of 46 diverse *Shigella* strains.[5] The HC2000_1465 cluster, named S1, can be divided into five HC1100 clusters (Fig. 2). Only the HC1100_36524 cluster (subcluster S1d) contained strains from a single serotype, *S. dysenteriae* 7. The HC1100_45518 cluster (S1e) contained only *S. flexneri* 6 strains, but most strains from this serotype were in another HC1100, HC1100_1465 (S1b), along with *S. dysenteriae* 3 and various serotypes of *S. boydii*. The HC1100_1466 cluster (S1c) contained *S. dysenteriae* 5 and various serotypes of *S. boydii*. Finally, the HC1100_4194 cluster (S1a) included only *S. dysenteriae* strains, but from diverse serotypes. *S. dysenteriae* 3 was found in two different S1 subclusters, S1a and S1b. At a higher level of resolution, four *Shigella* serotypes were grouped within specific HC400 clusters, whereas the other serotypes were split between two to six HC400 clusters (appendix 2 p 12).

The second cluster, HC2000_4118, comprised various serotypes of *S. dysenteriae* (2, prov. E670/74, prov. 96-265, and prov. BEDP 02-5104) and *S. boydii* (5, 7, 9, 11, 15-17) (Fig. 3). This cluster, consisting exclusively of indole-positive strains, corresponds to the Cluster 2 described by Pupo and colleagues.[5] The HC2000_4118 cluster, hereafter referred to as S2, could be divided into six distinct HC1100 clusters (Fig. 3). Five of these HC1100 clusters contained exclusively *S. boydii*; the sixth, HC1100_4191 (subcluster S2d), contained *S. boydii* 15 and all the *S. dysenteriae* serotypes found in S2. Three HC1100 clusters contained a single serotype:

HC1100_11401 (S2f) for *S. boydii* 7, HC1100_7057 (S2e) for *S. boydii* 9, and HC1100_11421 (S2c) for *S. boydii* 11. This last serotype was also found in the S1 cluster (S1b subcluster). At higher resolution, it was possible to assign some serotypes to a particular HC400 cluster. This was the case for *S. boydii* 16 (HC400_11449) and *S. boydii* 17 (HC400_11452). However, at this level of resolution, other serotypes were split between two to four clusters (appendix 2 p12).

The third cluster, HC2000_192, comprised *S. boydii* prov. E1621-54 and all serotypes and subserotypes of *S. flexneri*, with the exception of *S. flexneri* 6, which grouped in S1 (Fig. 4). This cluster seems to correspond to the Cluster 3 reported by Pupo and colleagues,[5] except that *S. boydii* 12 rather than *S. boydii* prov. E1621-54 was reported in Cluster 3 in this previous study (see discussion). This HC2000_192 cluster, hereafter referred to as S3, could be divided into seven distinct HC1100 clusters (Fig. 4A). One of these S3 subclusters, HC1100_11429, contained exclusively *S. boydii* prov. E1621-54. The other six HC1100 clusters contained two or more *S. flexneri* serotypes per cluster. Connor and colleagues previously subdivided >350 genomes of *S. flexneri* 1-5, X, Y into seven phylogenetic groups (PGs), based on a Bayesian analysis of population structure.[2] As 140 *S. flexneri* genomes from our study were included in the analysis by Connor and colleagues,[2] we compared the clustering by cgMLST HC1100 to that obtained by PG. HC1100_204, HC1100_543, HC1100_1468, HC1100_11594, HC1100_1530 corresponded to PG2, PG4, PG5, PG6 and PG7, respectively (Fig. 4). HC1100_192 encompassed PG1 and PG3, and the use of a higher HC resolution made it possible to link HC400_192 to PG3. However, PG1 did not correspond to a single HC400 cluster. Instead, it corresponded to two such clusters: HC400_237 and HC400_327.

We evaluated the accuracy of cgMLST HC for grouping *Shigella* genomes into different phylogenetic clusters by employing another approach: using the same dataset of 493 *E. coli* and *Shigella* genomes, we constructed a maximum-likelihood tree based on 8,003 SNV differences, and compared this SNV-based clustering (with strong bootstrap support) to the cgMLST HC data. There were no observable differences between the two approaches (Fig. 5).

To confirm the robustness of the population structure of *Shigella* established by cgMLST analysis of our reference datasets was robust, we also applied cgMLST to 3,870 clinical *Shigella* isolates received by the FNRC-ESS between 2017 and 2020, in the framework of the French national surveillance programme for *Shigella* infections. All these isolates were characterised phenotypically, on the basis of biochemical reactions and serotyping. They belonged to *S. dysenteriae* (*n* = 53), *S. boydii* (*n* = 101), *S. flexneri* (*n* = 1,555), and *S. sonnei* (*n* = 2,161). All but one of these 3,870 genomes were assigned to known serotype/HC2000/HC1100/HC400 combinations, without inconsistencies (appendix 1, appendix 2 p 17). The exception was an HC1100_204 (PG2) *S. flexneri* isolate, grouped into a new HC400 cluster, HC400_11853.

In silico serotyping tools have been developed by various groups, and can be used to maintain links with the current *Shigella* serotyping system. We assessed the performances of the three tools currently available: the EnteroBase "SeroPred" tool,[15] ShigaTyper,[11] and ShigEiFinder,[13] with our 317 genomes from well-characterised reference strains. ShigEiFinder (appendix 2 pp 13-14) gave the best serotype prediction results. However, 100% of the strains belonging to *S. boydii* 10 and to the new serotype *S. dysenteriae* 17, and 14-20% of the strains from *S. dysenteriae* 11, *S. dysenteriae* 14, and *S. boydii* 2 were not identified. All the strains from *S. dysenteriae* prov. BEDP 02-5104 were incorrectly predicted to be *S. dysenteriae* 2, whereas

83% of the strains from the new serotype *S. dysenteriae* 16 were incorrectly predicted to be *S. dysenteriae* prov. 96-265 and 13% were not assigned.

In recent decades, several provisional new serotypes of *S. dysenteriae* and *S. boydii* have been described by different groups across the world.[26,27] However, the phylogenetic relationships between these provisional serotypes and between these serotypes and other *Shigella* populations have not been investigated. We characterised these relationships in detail (appendix 2 pp 3-5). We found that all these provisional serotypes belonged to the three main *Shigella* clusters, S1 to S3 (Figs. 2-4), and that many of those reported under different names were actually identical (appendix 2 pp 3-5). We propose adding *S. dysenteriae* 16-18, and *S. boydii* 21 and 22 to the current serotyping scheme, retaining provisional status for *S. dysenteriae* prov. BEDP 02-5104. All the reference strains for these new serotypes are now available from the *Collection de l' Institut Pasteur* (CIP) or the National Collection of Type Cultures (NCTC) (appendix 2 pp 3-5).

Discussion

We present here a broad overview of the population of *Shigella*. The hierarchical clustering of cgMLST data and a cgSNV analysis showed that *Shigella* strains belong to eight phylogenetically distinct clusters, within the species *E. coli*. Our results are consistent with previous studies suggesting multiple origins of the *Shigella* phenotype.[5,28] However, the higher resolution of cgMLST, and comprehensive sampling from thousands of phenotypically characterised isolates and reference strains covering all serotypes, including provisional serotypes and atypical strains, made it possible to complete, and in some cases amend, the *Shigella* population structure obtained in previous MLST and genomic studies (appendix 2 pp 3-5).

The 70-year-old *Shigella* typing scheme, which is still in use today, was based on biochemical characteristics, antigenic relationships, and tradition.[10] We show here that, unlike cgMLST, this scheme does not always reveal natural groupings. In particular, the *Shigella* serogroups/species are artificial constructs developed from data for antigen and metabolic markers affected by Insertion Sequence (IS) element mobilisation and horizontal gene transfer (appendix 2 pp 3-4). The presence of large numbers of ISs and their expansions in *Shigella* genomes may alter the nature of both the O-antigen and the rare phenotypic markers identified in this bacterium with weak metabolic activity, by disrupting coding sequences or causing genome rearrangements and deletions.[6] For example, *S. boydii* 6 and 20 arose in subcluster 1c following the acquisition of a single IS within the *rfb* cluster of *S. boydii* 10 and 1, respectively. Serotype diversification, which is observed mostly in clusters S1 to S3, also occurs via horizontal gene transfer of the O antigen-encoding *rfb* cluster from *Escherichia* donors.[5,18] Horizontal gene transfer outside of the *rfb* cluster can also alter the serotype of a strain, as illustrated particularly clearly by the S3 cluster. All the *S. flexneri* strains in this cluster share the same O-antigen backbone structure and their serotypes are determined by glucosylation and/or O-acetylation modifications to the O-antigen tetrasaccharide repeat, conferred by prophage-encoded *gtr* and/or *oac* genes, respectively.[12] Plasmid-mediated serotype conversion by the O-antigen phosphoethanolamine transferase gene (*opt*) has also been reported in *S. flexneri.*[12] Each of the seven *S. flexneri* phylogenetic groups (PGs) described by Connor and colleagues,[2] based on a cgSNV analysis, contained two or more of these serotypes. As this serotyping method does not reflect the genetic relatedness between *Shigella* isolates, and has a number of other disadvantages, including being time-consuming, with intra- and interspecies cross-reactivity, and the impossibility of typing rough strains and new serotypes,[11,26] modern laboratory surveillance of *Shigella* infections should now be based on phylogenetically relevant methods rather than simply on molecular or *in silico* serotyping.[7,11–13] In our hands, the cgMLST HC analysis proved to be the method of choice for monitoring the trends in *Shigella* types. The different types of S*higella* can be identified with HC2000. Higher resolution, with HC1100 and, in certain cases, HC400, can reveal additional subclusters. This is particularly interesting for S3, which contains the *S. flexneri* 1-5, X, and Y serotypes generated via horizontal gene transfer rather than by vertical descent. We therefore recommend integrating the seven phylogenetic groups (PG1-PG7) described for *S. flexneri* into routine genomic surveillance for *S. flexneri*. These PGs can be easily inferred from cgMLST HC1000/HC400; it is even possible to obtain up to eight groups (after subdividing PG1 into PG1a and PG1b). The cgMLST HC analysis also provides, in a single step, a wide range of clustering levels, from HC0 (no allelic difference allowed) to HC2350 (maximum of 2,350 allelic differences), with a standard nomenclature. For the most frequent *Shigella* serotypes, such as *S. sonnei* and *S. flexneri* 2a, higher resolution levels, such as HC5 and HC10, can also help to identify a single-source outbreak or an epidemic strain, before confirmation by cgSNV analysis. The use of cgMLST HC data also makes it possible to query EnteroBase, which contains over 160,000 *E. coli*/S*higella* genomes, to identify strains with similar HC types. This can facilitate the investigation of unusual types of *Shigella* or outbreaks with an international dimension. HC10 was recently used to investigate the origins of an outbreak of *S. sonnei* infections in Belgium, and made it possible to link this outbreak to South America.[29]

However, the use of cgMLST HC data in surveillance should be paired with *in silico* serotyping, to achieve backward compatibility with the current serotyping scheme. This is a very important point for the maintenance of international surveillance with laboratories that cannot currently afford genomic surveillance and to prevent disjunction with the seven decades of serotyping data accumulated worldwide. For this purpose, we found that ShigEiFinder had the best performance of the three available tools.[13] However, it requires optimisation for certain serotypes. The complete set of *rfb* sequences provided by our study would be helpful for improving this tool.

In conclusion, by studying >4,000 serotyped reference strains and routine isolates covering the overall diversity of *Shigella*, we were able to demonstrate that cgMLST is a robust and portable genomic method revealing natural groupings for this pathovar of *E. coli*. The cgMLST method has strong added value in the framework of the laboratory monitoring of *Shigella*, as it prevents genetically unrelated strains being conflated, and genetically related strains being separated. However, we strongly recommend combining cgMLST with *in silico* serotyping to maintain backward compatibility with the current *Shigella* serotyping scheme.

ACKNOWLEDGMENTS

We thank all corresponding laboratories for sending isolates to the French National Reference Centre for *E. coli*, *Shigella*, and *Salmonella*. We also thank the sequencing teams at the Institut Pasteur (P2M-Plateforme de Microbiologie Mutualisée) and the Wellcome Sanger Institute for sequencing the samples.

AUTHOR CONTRIBUTIONS

FXW conceived and designed the study. IY, EH, LF and FXW did the genomic analyses. IY and FXW contributed to data interpretation and visualisation. CR, IC and MLC conducted the laboratory experiments. SL, CR, IC, MLC, MPG, DC, and FXW contributed to isolate acquisition and data collection. FXW, FD, RR and NRT were responsible for funding acquisition. FXW and IY drafted the article. LF, EH, RR, DC, MPG, SL, FD, and NRT critically reviewed the draft. All authors read and approved the final manuscript. IY and FXW accessed and verified the underlying data.

References

1 Khalil IA, Troeger C, Blacker BF, *et al.* Morbidity and mortality due to *Shigella* and enterotoxigenic *Escherichia coli* diarrhoea: the Global Burden of Disease Study 1990–2016. *Lancet Infect Dis* 2018; **18**: 1229–40.

2 Connor TR, Barker CR, Baker KS, *et al.* Species-wide whole genome sequencing reveals historical global spread and recent local persistence in *Shigella flexneri*. *Elife* 2015; **4**: e07335.

3 Hawkey J, Paranagama K, Baker KS, *et al.* Global population structure and genotyping framework for genomic surveillance of the major dysentery pathogen, *Shigella sonnei*. *Nat Commun* 2021; **12**: 2684.

4 Ewing WH. *Shigella* nomenclature. *J Bacteriol* 1949; **57**: 633–8.

5 Pupo GM, Lan R, Reeves PR. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *PNAS* 2000; **97**: 10567–72.

6 Hawkey J, Monk JM, Billman-Jacobe H, Palsson B, Holt KE. Impact of insertion sequences on convergent evolution of *Shigella* species. *PLoS Genet* 2020; **16**: e1008931.

7 Chattaway MA, Schaefer U, Tewolde R, Dallman TJ, Jenkins C. Identification of *Escherichia coli* and *Shigella* Species from Whole-Genome Sequences. *J Clin Microbiol* 2017; **55**: 616–23.

8 Pettengill EA, Pettengill JB, Binet R. Phylogenetic Analyses of *Shigella* and Enteroinvasive *Escherichia coli* for the Identification of Molecular Epidemiological Markers: Whole-Genome Comparative Analysis Does Not Support Distinct Genera Designation. *Front Microbiol* 2015; **6**: 1573.

9 Hyma KE, Lacher DW, Nelson AM, *et al.* Evolutionary genetics of a new pathogenic *Escherichia* species: *Escherichia albertii* and related *Shigella boydii* strains. *J Bacteriol* 2005; **187**: 619–28.

10 Strockbine NA, Maurelli AT. *Shigella*. In: Bergey's Manual of Systematics of Archaea and Bacteria. *American Cancer Society*, 2015: 1–26.

11 Wu Y, Lau HK, Lee T, Lau DK, Payne J. *In Silico* Serotyping Based on Whole-Genome Sequencing Improves the Accuracy of *Shigella* Identification. *Appl Environ Microbiol* 2019; **85**.

12 Brengi SP, Sun Q, Bolaños H, *et al.* PCR-based method for *Shigella flexneri* serotyping: International multicenter validation. *J Clin Microbiol* 2019; **57**.

13 Zhang X, Payne M, Nguyen T, Kaur S, Lan R. Cluster-specific gene markers enhance *Shigella* and Enteroinvasive *Escherichia coli in silico* serotyping. *bioRxiv* 2021; : 2021.01.30.428723.

14 Moura A, Criscuolo A, Pouseele H, *et al.* Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. *Nat Microbiol* 2016; **2**: 1–10.

15 Zhou Z, Alikhan N-F, Mohamed K, *et al.* The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Res* 2020; **30**: 138–52.

16 Langendorf C, Le Hello S, Moumouni A, *et al.* Enteric bacterial pathogens in children with diarrhea in Niger: diversity and antimicrobial resistance. *PLoS One* 2015; **10**: e0120275.

17 Kim J, Lindsey RL, Garcia-Toledo L, *et al.* High-Quality Whole-Genome Sequences for 59 Historical *Shigella* Strains Generated with Pacbio Sequencing. *Genome Announc* 2018; **6**.

18 DebRoy C, Fratamico PM, Yan X, *et al.* Comparison of O-Antigen Gene Clusters of All O-Serogroups of *Escherichia coli* and Proposal for Adopting a New Nomenclature for O-Typing. *PLoS One* 2016; **11**: e0147434.

19 Zhou Z, Charlesworth J, Achtman M. HierCC: A multi-level clustering scheme for population assignments based on core genome MLST. *Bioinformatics* 2021.

20 Zhou Z, Alikhan N-F, Sergeant MJ, *et al.* GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res* 2018; **28**: 1395–404.

21 Venkatesan MM, Buysse JM, Kopecko DJ. Use of *Shigella flexneri ipaC* and *ipaH* gene sequences for the general identification of *Shigella* spp. and enteroinvasive *Escherichia coli*. *J Clin Microbiol* 1989; **27**: 2687–91.

22 Figge RM, Ramseier TM, Saier MH. The mannitol repressor (MtlR) of *Escherichia coli*. *J Bacteriol* 1994; **176**: 840–7.

23 Aslanidis C, Schmitt R. Regulatory elements of the raffinose operon: nucleotide sequences of operator and repressor genes. *J Bacteriol* 1990; **172**: 2178–80.

24 Rezwan F, Lan R, Reeves PR. Molecular basis of the indole-negative reaction in *Shigella* strains: extensive damages to the *tna* operon by insertion sequences. *J Bacteriol* 2004; **186**: 7460–5.

25 Ochman H, Selander RK. Standard reference strains of *Escherichia coli* from natural populations. *J Bacteriol* 1984; **157**: 690–3.

26 Coimbra RS, Lenormand P, Grimont F, Bouvet P, Matsushita S, Grimont PAD. Molecular and Phenotypic Characterization of Potentially New *Shigella dysenteriae* Serotype. *J Clin Microbiol* 2001; **39**: 618–21.

27 Melito PL, Woodward DL, Munro J, *et al.* A Novel *Shigella dysenteriae* Serovar Isolated in Canada. *J Clin Microbiol* 2005; **43**: 740–4.

28 Yang J, Nie H, Chen L, *et al.* Revisiting the molecular evolutionary history of *Shigella* spp. *J Mol Evol* 2007; **64**: 71–9.

29 Van den Bossche A, Ceyssens P-J, Denayer S, *et al.* Outbreak of Central American born *Shigella sonnei* in two youth camps in Belgium in the summer of 2019. *Eur J Clin Microbiol Infect Dis* 2021.

Figure 1. A NINJA neighbour-joining GrapeTree showing the population structure of *Shigella* spp. based on the cgMLST allelic differences between 493 *Shigella* and *E. coli* reference genomes. (A) The tree nodes are colour-coded by *Shigella* serogroup and *E. coli* pathovar. (B) The tree nodes are colour-coded by HC2000 data. HC2000 clusters with fewer than two isolates are represented by white nodes. The different *Shigella* cgMLST clusters are labelled. For the SON cluster, the different genomic lineages of *S. sonnei* are indicated with Latin numerals. For the *S. flexneri* serotypes in cluster S3, the phylogenetic groups (PG1 to PG7) are also indicated. The interactive version of the tree is publicly available from http://enterobase.warwick.ac.uk/ms_tree?tree_id=55118

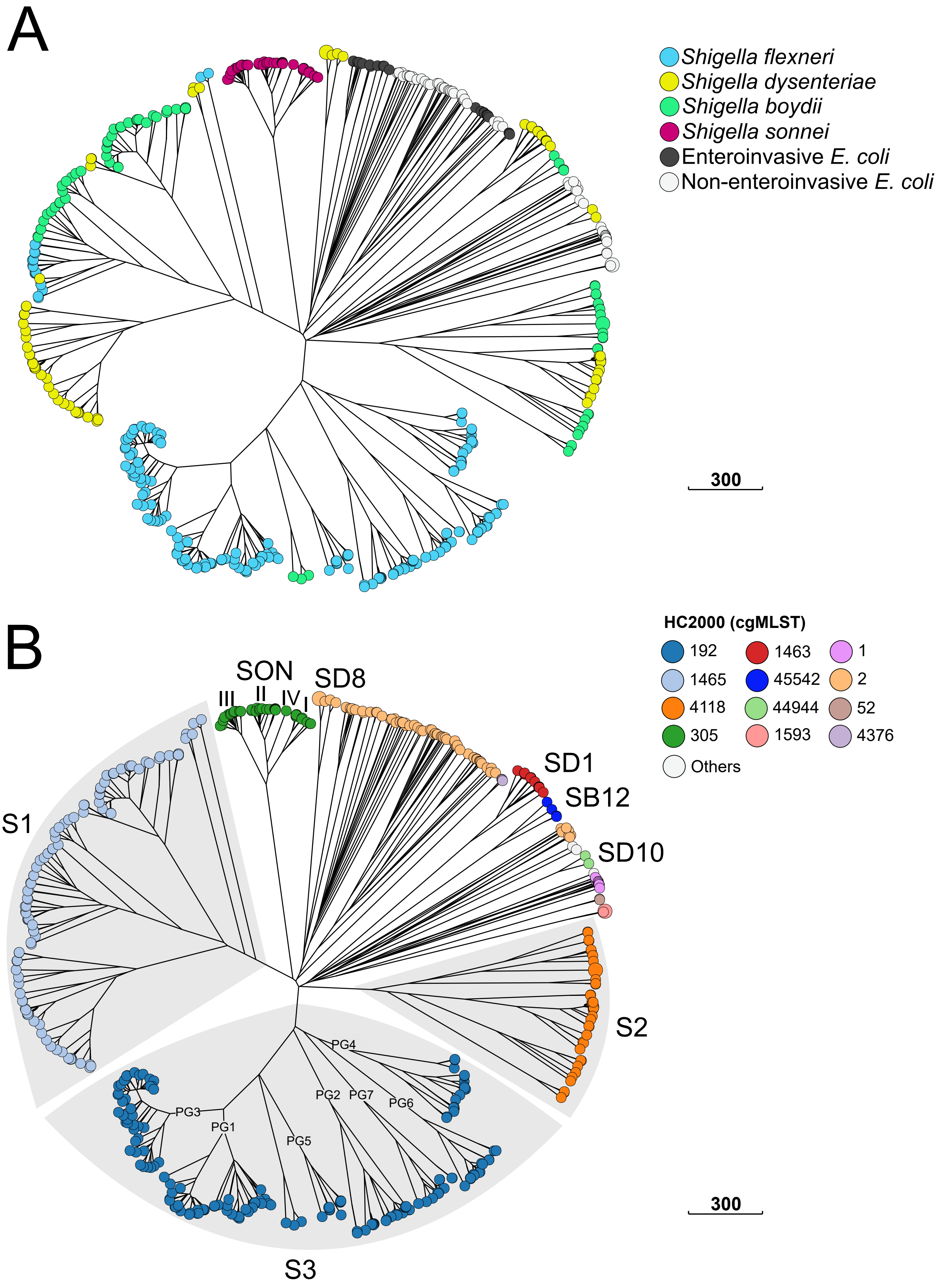
Figure 2. A NINJA neighbour-joining GrapeTree showing the population structure of the *Shigella* S1 cluster (HC2000_1465). This subtree is based on the tree shown in Figure 1. The tree nodes are colour-coded by serogroup. The numbers within nodes indicate the serotype. HC1100 designation is indicated next to each subcluster. Novel *Shigella* serotypes are also shown. NST, = non-serotypable.

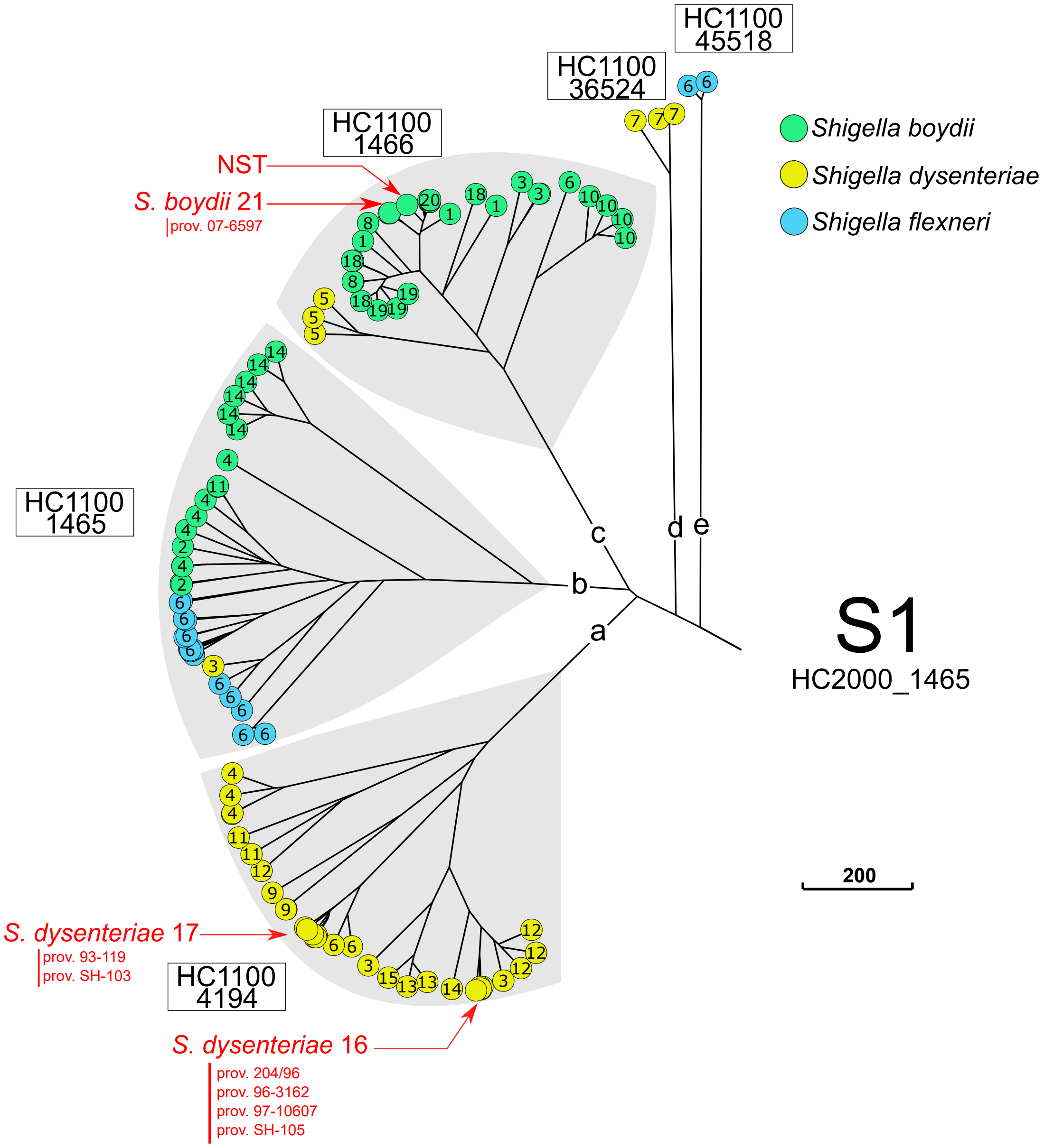
Figure 3. A NINJA neighbour-joining GrapeTree showing the population structure of the *Shigella* S2 cluster (HC2000_4118). This subtree is based on the tree shown in Figure 1. The tree nodes are colour-coded by serogroup. The numbers within nodes indicate the serotype. HC1100 designation is indicated next to each subcluster. Novel and provisional (prov.) *Shigella* serotypes are also shown.

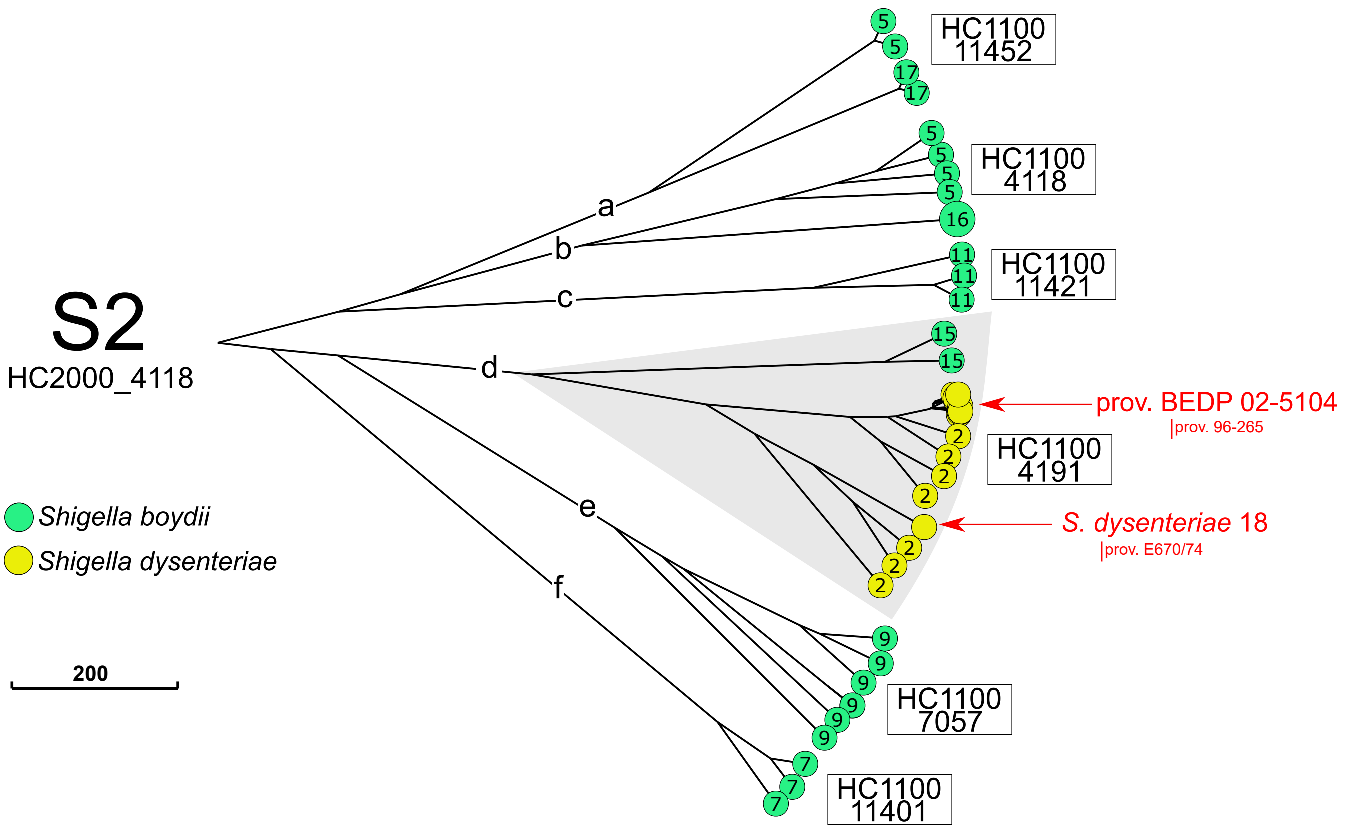
Figure 4. A NINJA neighbour-joining GrapeTree showing the population structure of the *Shigella* S3 cluster (HC2000_192). This subtree is based on the tree shown in Figure 1 (A). The tree nodes are colour-coded by HC1100 data. The *S. flexneri* phylogenetic groups (PG) identified by Connor and colleagues are indicated.[2] Some HC400 clusters are indicated to separate PG3 from PG1. *S. boydi* 21 (formerly prov. E1621-54) is shown. (B) The tree nodes are colour-coded by *S. flexneri* serotype.

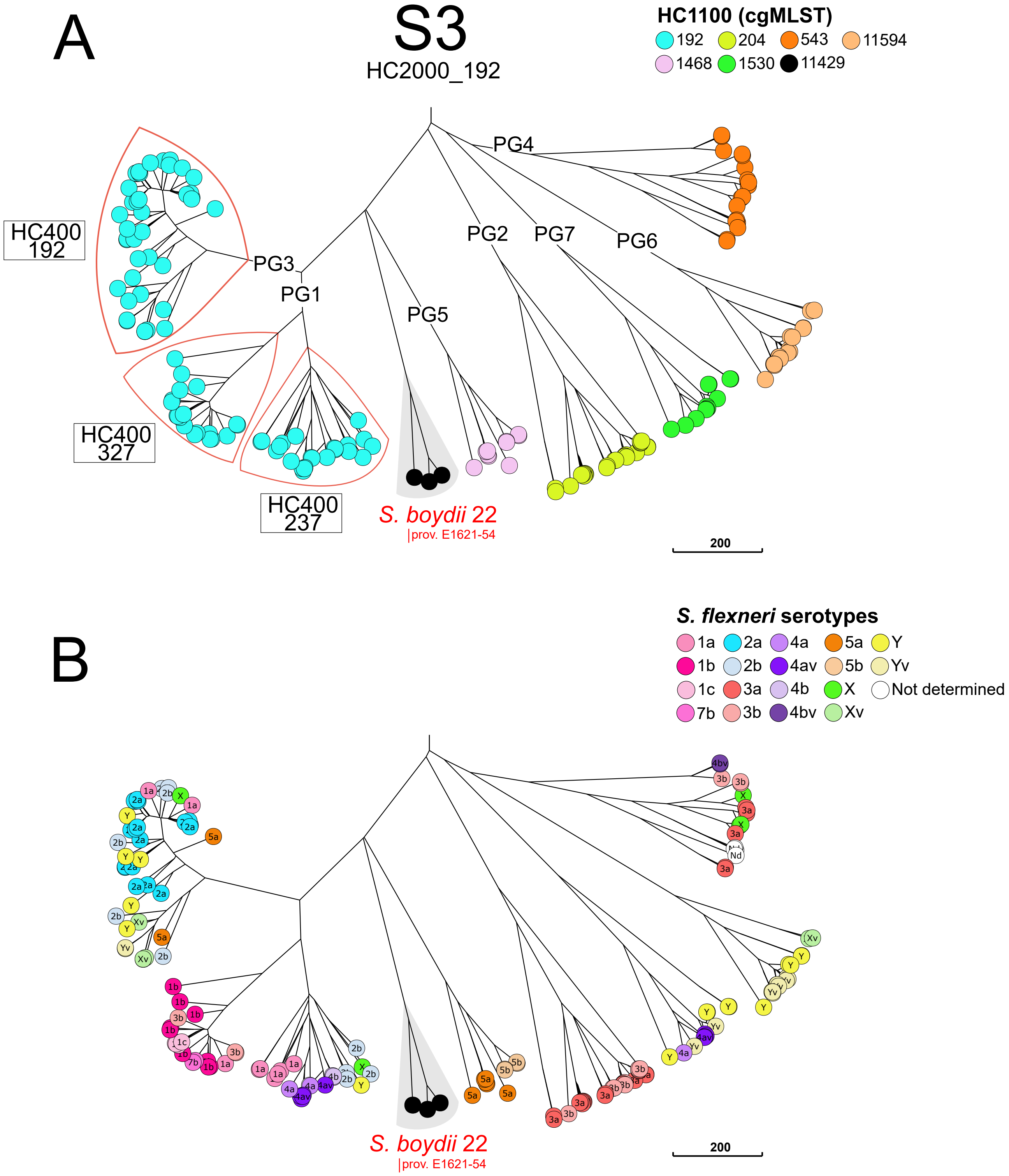
Figure 5. A maximum-likelihood phylogenetic tree showing the population structure of 493 *Shigella* and *E. coli* reference genomes based on 8,003 core-genome single-nucleotide variants

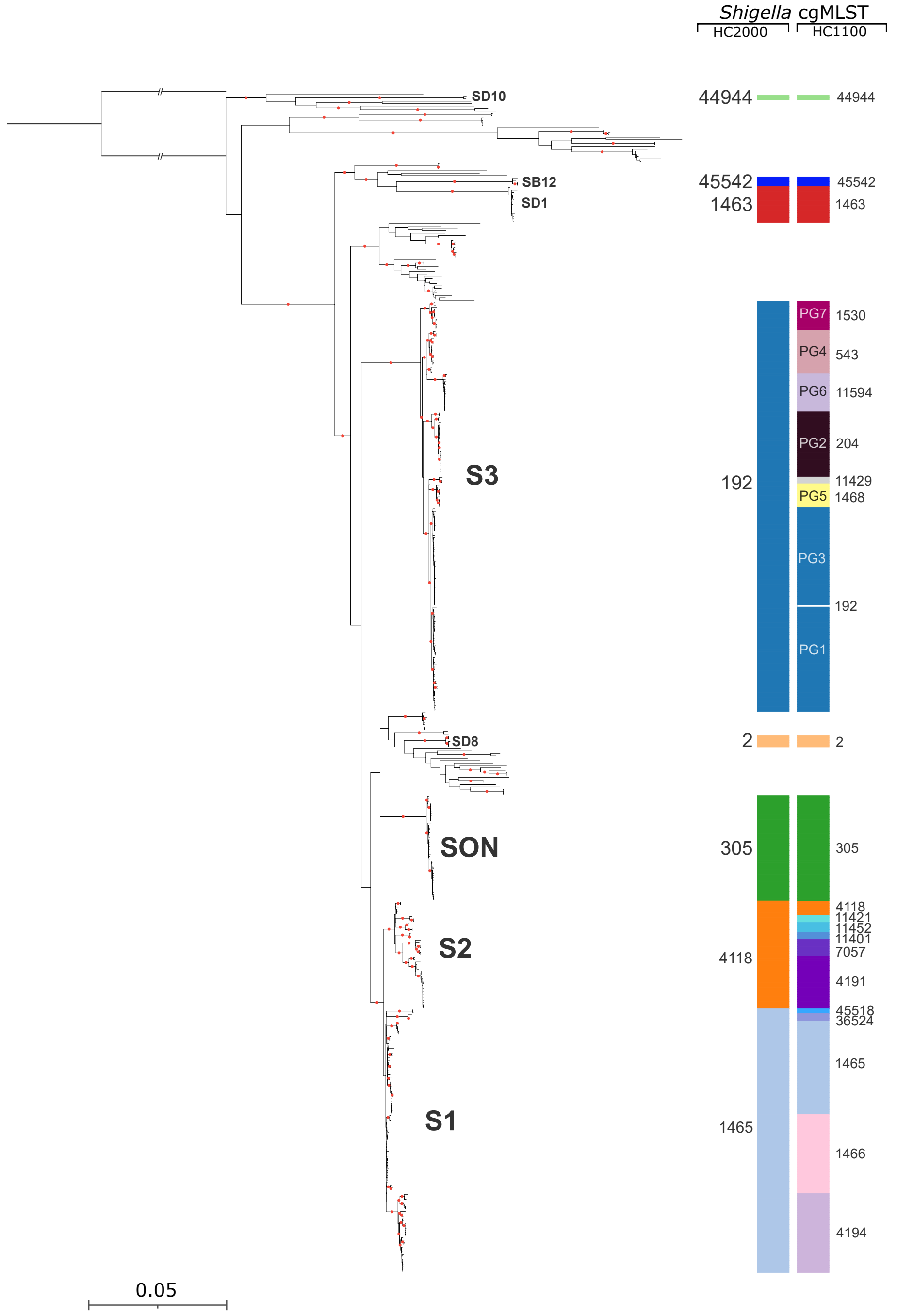
(SNVs). Nodes supported by bootstrap values ≥95% are indicated by red dots. Phylogenetic clades containing *Shigella* genomes are labelled with the same nomenclature as in Figure 1. All the *Shigella* genomes are also labelled on the right with cgMLST HC2000 and HC1000 data.


A
Shigella flexneri
Shigella dysenteriae
Shigella boydii
Shigella sonnei
Enteroinvasive E. coli
Non-enteroinvasive E. coli
300
B
HC2000 (cgMLST)
192
1465
4118
305
Others
1463
45542
44944
1593
1
2
52
4376
SON
III
II
IV
I
SD8
SD1
SB12
SD10
S1
S2
S3
PG1
PG2
PG3
PG4
PG5
PG6
PG7
300


HC1100
45518
HC1100
36524
HC1100
1466
NST
S. boydii 21
prov. 07-6597
Shigella boydii
Shigella dysenteriae
Shigella flexneri
HC1100
1465
a
b
c
d
e
S1
HC2000_1465
200
S. dysenteriae 17
prov. 93-119
prov. SH-103
HC1100
4194
S. dysenteriae 16
prov. 204/96
prov. 96-3162
prov. 97-10607
prov. SH-105


S2
HC2000_4118
HC1100 11452
HC1100 4118
HC1100 11421
HC1100 4191
HC1100 7057
HC1100 11401
a
b
c
d
e
f
5
5
17
17
5
5
5
5
16
11
11
11
15
15
2
2
2
2
2
2
2
9
9
9
9
9
9
7
7
7
prov. BEDP 02-5104
prov. 96-265
S. dysenteriae 18
prov. E670/74
Shigella boydii
Shigella dysenteriae
200


A
S3
HC2000_192
HC1100 (cgMLST)
192
204
543
11594
1468
1530
11429
PG4
PG3
PG1
PG5
PG2
PG7
PG6
HC400
192
HC400
327
HC400
237
S. boydii 22
|prov. E1621-54
200
B
S. flexneri serotypes
1a
2a
4a
5a
Y
1b
2b
4av
5b
Yv
1c
3a
4b
X
Not determined
7b
3b
4bv
Xv
S. boydii 22
|prov. E1621-54
200


Shigella cgMLST
HC2000
HC1100
SD10
SB12
SD1
S3
SD8
SON
S2
S1
0.05
44944
45542
1463
192
2
305
4118
1465
44944
45542
1463
PG7
1530
PG4
543
PG6
11594
PG2
204
11429
PG5
1468
PG3
192
PG1
2
305
4118
11421
11452
11401
7057
4191
45518
36524
1465
1466
4194