EnteroBase: Hierarchical clustering of 100,000s of bacterial genomes into species/sub-species and populations

Mark Achtman*, Zhemin Zhou*[,a], Jane Charlesworth and Laura Baxter

University of Warwick, Coventry CV4 7AL, UK

[a]Current address: Pasteurien College, Soochow University, No.199 Ren'ai Road, Suzhou, Jiangsu 215123, China.

***Corresponding authors: MA, m.achtman@warwick.ac.uk; ZZ, zheminzhou@qq.com

MA, ORCID 0000-0001-6815-0070; ZZ, ORCID 0000-0001-9783-0366; JC, ORCID 0000-0001-9759-9838; LB, ORCID 0000-0003-3287-7547

Word count: 7559

Word Count incl. 103 Citations 10846

Plus:

4 Tables

5 Figures

Abstract

The definition of bacterial species is traditionally a taxonomic issue while defining bacterial populations is done with population genetics. These assignments are species specific, and depend on the practitioner. Legacy multilocus sequence typing is commonly used to identify sequence types (STs) and clusters (ST Complexes). However, these approaches are not adequate for the millions of genomic sequences from bacterial pathogens that have been generated since 2012. EnteroBase (http://enterobase.warwick.ac.uk) automatically clusters core genome MLST alleles into hierarchical clusters (HierCC) after assembling annotated draft genomes from short read sequences. HierCC clusters span core sequence diversity from the species level down to individual transmission chains. Here we evaluate HierCC's ability to correctly assign 100,000s of genomes to the species/subspecies and population levels for *Salmonella, Clostridoides, Yersinia, Vibrio* and *Streptococcus*. HierCC assignments were more consistent with maximum-likelihood super-trees of core SNPs or presence/absence of accessory genes than classical taxonomic assignments or 95% ANI. However, neither HierCC nor ANI were uniformly consistent with classical taxonomy of *Streptococcus.* HierCC was also consistent with legacy eBGs/ST Complexes in *Salmonella* or *Escherichia* and revealed differences in vertical inheritance of O serogroups. Thus, EnteroBase HierCC supports the automated identification of and assignment to species/subspecies and populations for multiple genera.

1. Introduction

"Microbiologists need large databases to identify and communicate about clusters of related bacteria, … Such databases should contain the reconstructed genomes of bacterial isolates,… together with metadata describing their sources and phenotypic properties… and we are currently developing EnteroBase, a genome-based successor to the *E. coli* and *S. enterica* MLST databases." Excerpted from Achtman and Zhou, 2014 [1].

The Linnean system of genus and species designations was applied to bacterial nomenclature in the late 19th century, soon after their ability to cause infectious diseases had been recognised. These taxonomic labels initially reflected disease specificity and phenotypic similarities. Over the following decades, phenotypic criteria for distinguishing bacterial taxa were extended to include serologically distinct groupings and even differential sensitivity to bacteriophages. A primary goal for taxonomic designations for bacterial pathogens was that they be useful for epidemiology and clinical diagnoses. As a result, *Salmonella enterica* serovars, which differ by dominant epitopes on lipopolysaccharide and flagella, were each assigned a distinct species designation, often referring to the disease syndrome and host e.g. *Salmonella typhimurium* [2]. Similarly, even though *Yersinia pestis* is a clone of *Yersinia pseudotuberculosis* [3,4], it was designated as a distinct species because *Y. pestis* causes plague whereas *Y. pseudotuberculosis* causes gastroenteritis. The use of species designations for *Salmonella* serovars is now disparaged, although it is still in common use [5], but *Y. pestis* has retained its species designation.

Percentage DNA-DNA hybridization levels reflect genomic relationships and became the "Gold Standard" metric for new taxonomic designations after 1987 [6],

but distinctive phenotypic differences have remained a requirement for the definition of a novel species until the present. As a result, the international committee which controls taxonomic designations continues to reject the validity of taxonomic designations based solely on DNA similarities [7,8].

a) ANI and MLST

An alternative genomic approach for taxonomic designations was proposed in 2005 by Konstantinidis [9], namely the definition of species on the basis of average nucleotide identity (ANI). Pairs of genomes with 95% ANI usually belong to the same species whereas pairs with lower ANI values belong to different species. Fast computational methods based on k-mer searches, such as FastANI [10], can perform ANI-like calculations on large numbers of genomes, and these calculations are now routinely used by bioinformaticians for taxonomic assignments. An approach based on ANI is very enticing because it could provide defined criteria for the definition of a species across all Bacteria, and allow species assignment based exclusively on genome sequences. However, the use of ANI for species assignments has been criticized because it does not completely correlate with DNA-DNA hybridization [11,12]. Furthermore, multiple taxa that are designated as single species each contain multiple 95%ANI groups [13,14]. For example, *Streptococcus mitis* is designated a single species because those strains colonise the human oropharyngeal tract and have similar phenotypes. However, *S. mitis* encompasses myriad, genetically distinct strains [15-17], and encompasses at least 44 distinct 95% ANI clusters [17]. Other problematic genera include *Pseudomonas* [12,18] and *Aeromonas* [19]. Furthermore, large databases including 100,000s of genomes per genus would struggle to implement methods such as pairwise clustering by ANI because each single new entry

would require testing against all genomes. There is also no consensus on ANI criteria for recognizing lower taxonomic entities, such as sub-species and populations. Indeed, there is not even a consensus on the definitive properties of what constitutes a bacterial species [20].

Bottom-up population genetic approaches such as multilocus sequence typing (MLST) can provide an alternative to top-down taxonomy, and deal efficiently with large numbers of bacterial strains. Legacy MLST based on the sequence differences of several housekeeping genes was introduced in 1998 [21], and has now been applied to more than 100 bacterial species [22]. MLST defines sequence types (STs), consisting of unique integer designations for each unique sequence (allele) of each of the MLST loci. Some STs mark individual clones with special pathogenic properties and which seem to have arisen fairly recently, such as *Escherichia coli* ST131 which is a globally prominent cause of urinary tract infection (UTI) and invasive disease [23]. Similarly, *Salmonella enterica* subsp. *enterica* serovar Typhimurium ST313 is a common cause of extra-intestinal, invasive salmonellosis in Africa [24]. Clusters of related STs are also well known. Such clusters can be recognized by eBurst analyses [25], and are referred to as ST Complexes in *E. coli* [26] and eBGs (eBurst groups) in *S. enterica* [27]. ST Complexes and eBGs seem to reflect natural populations, but their broad properties are still not well understood, in part because the resolution of legacy MLST is fairly low.

The principles of legacy MLST were extended to rMLST, which uses sequences of 53 universal bacterial genes encoding ribosomal proteins [28] and provides a universal MLST scheme for all Bacteria [28]. For both *Escherichia* and *Salmonella,* rMLST offers a slight improvement in resolution over legacy MLST, and the

identification of *Salmonella* eBGs is reasonably consistent across both approaches [29]. rMLST has been used to identify sets of representative genomes from large collections for calculating phylogenetic trees [28] and pan-genomes [17].

MLST has also been extended to cgMLST, which encompasses all the genes in a soft core genome [29-31], and cgMLST nomenclatures have been implemented for multiple bacterial genera [28]. The large number of loci encompassed by cgMLST schemes results in enormous numbers of cgMLST STs (cgSTs), but these can be clustered into groups of bacterial genomes at multiple levels of genomic diversity (hierarchical clustering) [32]. cgMLST provides considerably higher resolution than legacy MLST or rMLST, and initial analyses indicated that it is ideal for investigating transmission chains within single source outbreaks or for identifying population structures up to the genus level [28]. Here we focus on the automated assignments of genome assemblies from six genera of important bacterial pathogens (Table 1) to taxonomic and population structures by hierarchical clustering of cgSTs within EnteroBase with the HierCC pipeline [32]. For five of those genera, cgMLST is a full solution for taxonomic designations of species, and incorrect taxonomic designations have now been unilaterally changed within EnteroBase.

b) EnteroBase and HierCC

EnteroBase (http://enterobase.warwick.ac.uk) automatically downloads short read sequences and their accompanying metadata from the European Nucleotide Archive (ENA). Metadata are then converted to consistent categories by a dedicated EnteroBase parser with the EToKi pipeline [28]. Draft genomes are assembled from the short reads, and annotated according to a genus-wide pan-genome scheme created with PEPPAN [17] (Table 2). In September 2021, EnteroBase contained over 600,000

draft genomes from the six genera in Table 1, and likely provides a nearly comprehensive overview of their global diversity. Many of these samples reflect a focus on food-borne disease in the U.S. and United Kingdom, but this bias is increasingly being reduced by the global sources of many genomes (Supplemental Text, Sample Bias).

One of the primary goals for EnteroBase was a hierarchical overview of the population structure of the genera in Table 1. We therefore developed HierCC [32] based on cgMLST assignments to support the rapid recognition and deeper investigation of differing levels of population structure, and EnteroBase provides cluster assignments and designations for 10-13 levels of allelic differences for all six genera [28] (Table 1). Some single source outbreaks of food-borne disease are caused by uniform clades [33], and HC5 – HC10 HierCC have been used to identify tight cgMLST clusters of *S. enterica* and *E. coli/Shigella* from single source outbreaks across Europe [34-37]. Pathogen species can also include more distantly related bacterial populations that are endemic in a single location or that have spread to multiple countries [38-40]. Higher HierCC levels can be used to track such populations, both locally and globally [28,41-43]. EnteroBase has even been used to classify all *Shigella* [44], which correspond to discrete lineages of *E. coli* [26,45], and to define novel sub-species of *S. enterica* [46].

The highest level of HierCC in EnteroBase can assign genomes to species/sub-species for *Salmonella*, *Clostridioides*, *Yersinia* and *Vibrio* (Table 1). Here we compare those assignments with assignments by classical taxonomic approaches and by 95% ANI. The results illustrate problems with both classical approaches and 95% ANI, and we conclude that HierCC is preferable for automated assignment of genomes to

species/subspecies for those genera. However, neither ANI nor HierCC is universally satisfactory for species/sub-species assignments within *Streptococcus*. This manuscript also provides an initial overview of the abilities of HierCC to assign genomes to populations and Lineages [41,47] within *Salmonella* and *Escherichia*/*Shigella*, and compare those assignments with the distributions of O antigens within lipopolysaccharide. Detailed analysis of *Escherichia* will be presented elsewhere.

Results

a. Species and sub-species

Classical taxonomic approaches for assigning individual strains and genomes to species depend on human expertise, and are not suitable for automated pipelines in real-time databases such as EnteroBase. This problem is acute because many short read sequences in ENA have incorrect taxonomic assignments, or none at all, and phenotypic distinctions are inappropriate for databases of 100,000s of genomes. We therefore annotated the branch topologies within super-trees calculated by a divide and conquer maximum likelihood approach (henceforth ML super-trees) from core genome SNPs of representative genomes of *Salmonella*, *Clostridioides*, *Yersinia*, *Vibrio* and *Streptococcus* with the groupings according to a 95% ANI cutoff calculated with FastANI [10]. We also identified HierCC levels that corresponded closely to peak values in normalized mutual information and silhouette plots, and which identified well-defined species/sub-species in each genus [32]. For example, for *Salmonella*, this optimal HC level was HC2850, which corresponds to a maximum internal branch length of 2850 different alleles among the 3,002 loci used for cgMLST within a minimal spanning tree of pair-wise allelic distances. The optimal HC levels for the other five genera ranged from HC363 – HC2500 (Table 1).

***Salmonella*.** ML super-trees based on 1,410,331 core SNPs from 10,002 representative genomes showed a strong concordance between traditional taxonomic assignments and clustering according to either HC2850 (Fig. 1A) or 95% ANI (Fig. 1B). High concordance was also obtained with ML super-trees based on the presence or absence of accessory genes (https://enterobase.warwick.ac.uk/ms_tree/53258). Both HC2850 and ANI distinguished *S. enterica* from *S. bongori*, and confirmed the

recent conclusion by Pearce *et al*. [48] based on 569 genomes that former subspecies IIIa corresponds to a separate species, *S. arizonae.* HC2850 also identified a new *Salmonella* species, cluster HC2850_215890, and that conclusion was confirmed by 95% ANI. HC2850_215890 consists of five strains that have been isolated from humans in the U.K. since 2018, and a gene for gene comparison indicated that roughly half of the core genes were more similar to *S. enterica* and the other half to *S. bongori*.

95% ANI did not detect any subspecies structure within *S. enterica* whereas HC2850 differentiated a total of 12 clusters that corresponded to distinct phylogenetic branches in the ML super-tree (Fig. 1). Most of these have been designated as subspecies by traditional taxonomy [46,48], but one of them, a singleton genome in HC2850_222931, likely represents a new subspecies. Despite the correct identification of so many sub-species, HC2850 did not support Pearce's definition of subspecies XI, and did not distinguish those genomes from subsp. II. Similarly, HC2850 did not differentiate Pearce's subsp. VIII from subsp. IV. The prior definition of these subspecies was largely dependent on branch topologies within phylogenetic trees based on rMLST loci, and was also contradicted by population statistics [48]. HierCC did differentiate HC2850_7171, which consists of a short phylogenetic branch within subsp. I, and which would not have been assigned to a subspecies according to its ML topology. Genomic comparisons of multiple single genes from genomes within HC2850_7171 revealed that some gene sequences were more similar to subsp. I than to other subspecies while others were most similar to subsp. II. These observations suggest that HC2850_7171 may be a hybrid between those two subspecies, but do not address the question whether HC2850_7171 is a novel subspecies or not. With this sole possible exception, HierCC seems to be

suitable for the automated detection of new *Salmonella* species and subspecies, and for routinely assigning novel genomes to the appropriate taxa.

Clostridioides. Clustering according to 95% ANI differentiated *C. difficile* from *C. mangenotii*, and also identified five other clusters that seem to be novel species on the basis of the phylogeny of an ML SNP super-tree (Fig. 2). HC2500 identified the same clusters, and also separated out two additional distinct clusters which resemble novel sub-species (arrows: HC2500_15334, HC2500_15408). A subset of the genomes in these clusters (Fig. 2) were assigned to cryptic clades C-I, CII and C-III by a recent publication [49], which also concluded that they represent novel species. Thus, HierCC is suitable for the automated detection of new *Clostridioides* species and subspecies, and for routinely assigning novel genomes to the appropriate taxa.

Similar comparisons were also performed with *Escherichia*. These will be presented in detail elsewhere.

Yersinia* and *Vibrio. Detailed results with *Yersinia* and *Vibrio* can be found in Supplemental Text. The general take-home lessons from those analyses resemble those with *Salmonella* and *Clostridioides*: 95% ANI clustering is largely congruent with existing taxonomic designations, and HierCC identifies both species and sub-species with better resolution than 95% ANI. However, multiple, unanticipated taxonomic problems were evident in both genera. Firstly, ENA uses the strain designation as a species designation when a species is not specified during the submission process. The resulting public metadata is bloated with inaccurate species names. Secondly, multiple distinct phylogenetic branches were found without any species identifier. Thirdly, in multiple cases the species identifier did not correspond with ANI and/or HierCC, and comparisons with both the SNP and presence/absence

ML topologies confirmed that those species designations in ENA were incorrect. We addressed these problems within EnteroBase by performing radical manual curation to the metadata to ensure that the taxonomic species designations correspond to the phylogenies and population structures.

95% ANI supports the taxonomic convention that *Y. enterocolitica* represents a single species, and all genomes designated as *Y. enterocolitica* were in a single 95% ANI cluster. However, in accord with the conclusions by Reuter *et al.* [50], HierCC clustering assigned the non-pathogenic biotype 1A genomes to HC1490_73 and HC1490_764, the highly pathogenic biotype 1B to HC1490_2, and pathogenic biotypes 2-5 to HC1490_10. We therefore renamed these four groups by adding the HierCC cluster to the species name, e.g. *Y. enterocolitica* HC1490_2.

In contrast to *Y. enterocolitica*, the *Yersinia pseudotuberculosis* Complex corresponds to a single phylogenetic cluster according to both HC1490 and 95% ANI (Supplemental Fig. S1). Taxonomists have split these bacteria into *Y. pseudotuberculosis*, *Y. pestis*, *Y. similis* and *Y. wautersii* [51,52]. DNA-DNA hybridization [53] and gene sequences of several housekeeping genes [4] previously demonstrated that *Y. pestis* is a clade of *Y. pseudotuberculosis*, and the new observations demonstrate that a distinct species status is also inappropriate for *Y. similis* and *Y. wautersii*. We have therefore downgraded all three taxa within EnteroBase to the category of subspecies of *Y. pseudotuberculosis*, and extended the designation of *Y. pseudotuberculosis* to *Y. pseudotuberculosis sensu stricto*. These assignments are not reflected by distinct HC1490 clusters, and HC1490 clustering can only be used to automatically assign new genomes to the *Y. pseudotuberculosis* Complex.

We also defined eight other novel species/sub-species designations within *Yersinia* by unique designations based on HierCC clusters: e.g. *Yersinia* HC1490_419. These clusters had previously been incorrectly designated with the names of other species which formed distinct ANI and HC1490 clusters, or had not been named at all (SupplementalText, Fig. S1).

General concordance was also found for *Vibrio* between phylogenetic trees based on either core SNPs or presence/absence of accessory genes, and also with the clustering according to 95% ANI and HC1090 (SupplementalText). *Vibrio* encompassed 152 HC1090 clusters (Supplemental Fig. S2, SupplementalText), which corresponds to much greater taxonomic diversity than for the other genera dealt with above. The concordance between HC1090 and 95% ANI clusters was absolute for most clusters, including a large numbers of genomes from *V. cholerae*, *V. parahaemolyticus*, *V. vulnificans,* and *V. anguillarum* (Supplemental Table S1), and we did not find any definitive evidence indicating the existence of sub-species. However, eleven HC1090 clusters each encompassed between 2 and 5 ANI clusters (Supplemental Table S2) and three ANI clusters each encompassed 2-3 HC1090 clusters (Supplemental Table S3). In order to support the automated assignment of genomes to named species, we implemented taxonomic assignments according to HC1090, and renamed the species of all genomes that were contradictory to this principle. The resulting dataset contains 109 HC1090 clusters with a unique, classical species designation as well as 43 other species level clusters designated as *Vibrio* HC1090_xxxx. Seventeen species names were eliminated because they were contradictory to the phylogenetic topologies or were incoherently applied (Supplemental Text, Supplemental Table 4). Our radical approach now permits future automated assignment of novel genomes to species

designations, and the recognition of novel species, and has provided a clean and consistent basic taxonomy that can be progressively expanded.

Prior work has assigned many *Vibrio* species into so-called higher order clades of species on the basis of MLSA (multilocus sequence analysis) [54]. These clades are also apparent in the ML super-tree of core SNPs, and Fig. S2 indicates the three largest (Cholerae, Harveyii, Splendidus).

***Streptococcus*.** In all five genera summarized above, clustering according to HierCC was largely concordant with the ML super-trees based on core SNPs or presence/absence of accessory genes. 95% ANI and taxonomic designations were also largely concordant with the phylogenetic trees. Similar concordance with the trees based on core SNPs (Fig. 3) and presence/absence of accessory genes (Supplemental Fig. S3) was also found for most of the named species within *Streptococcus*. 102 HC363 clusters were concordant with 95% ANI clusters, and each of those HierCC clusters was specific for a single species after eliminating *S. milleri, S. periodonticum* and *S. ursoris* (Supplemental Table S5). These results once again demonstrate strong agreement between the two methods with classical taxonomy. However, genomes from a large number of species within *Streptococcus* were not clustered satisfactorily by either method.

Prior results had demonstrated that *Streptococcus mitis* and *oralis* each represents multiple related sequence clusters rather than two single species [15,16,55,56], and similar results were obtained with 95% ANI clustering [17]. The current analyses confirmed these observations. 23 of the 102 concordant HC363 and 95% ANI clusters from the species *S. cristatus, S. infantis, S. mitis, S. oralis, S. sanguinis,* and *S. suis* corresponded to only a sub-set of the named species. Furthermore, HC363 identified

single clusters for 18 species or groups of species which were subdivided extensively by 95% ANI for a total of 155 discrepancies (Supplemental Table S6). These discrepant clusters were within the species *S. suis, S. criceti, S. sanguinis, S. suis, S. oralis, S. hyointestinalis, S. parasanguinis, S. australis, S. infantis,* and *S. cristatus.* 95% ANI was not able to differentiate *S. pneumoniae* from *S. pseudopneumoniae, S. pyogenes* from *S. dysgalactiae, S. salivarius* from *S. vestibularis,* or between *S. equinus, S. infantarius,* and *S. lutentiensis.* Contrariwise, three other 95% ANI clusters each encompassed two HC363 clusters (Supplemental Table S7). HC363 could not distinguish *S. pneumoniae* or *S. pseudopneumoniae* from fifty-seven 95% ANI clusters within *S. mitis.* In addition to these problems, eight additional type strains should not have been given a species name because they clearly belonged to one of these chaotic clusters. We conclude that numerous species that have been defined on taxonomic grounds cannot be correctly identified by either 95% ANI or HierCC, and that multiple discrepancies exist between the two approaches within *Streptococcus*.

b. Populations, eBGs and Lineages

For each genus in which it has been implemented, HierCC supports the identification of populations at multiple hierarchical levels. The highest HC level was aimed at species/sub-species and the lowest levels at deducing transmission chains. However, an important measure of the success of HierCC would be its reliable detection of the natural populations previously recognized by legacy MLST. These were initially designated as clonal complexes [25], but are now referred to as ST Complexes in *Escherichia* [26] and eBGs (eBurst Groups) in *Salmonella* [27]. These populations should correspond to intermediate HierCC clustering levels, lower than for sub-

species and higher than for transmission chains. Here we focus on examples of their from *Salmonella*, *E.coli/Shigella,* and *S. pneumoniae* where they have been examined in greatest detail.

***Salmonella*.** Phylogenetic trees and population genetic analyses of core genome SNP diversity identified a limited number of deep subdivisions within *S. enterica* subsp. *enterica*, which were variably designated as lineages [57], Clades [58-61] or Branches [62]. Their phylogenetic branching topology may be unreliable because of extensive inter-branch exchange of nucleotide diversity by homologous recombination [47,62]. We also failed to find any HierCC clustering level in *Salmonella* that was consistently concordant with these branch structures [32]. Instead, populations were most clearly associated with HC2000 and HC900 clusters, which represent later sub-divisions within the phylogenetic trees. At those levels, recombination is much less prominent, or even largely lacking [47].

Traditional nomenclature of *Salmonella* subgroupings below the sub-species level has been predominantly based on assigning serovar designations by serotyping. Serovar designations consist of common names that have been assigned to a total of more than 2500 unique antigenic formulas based on epitopes within lipopolysaccharide (O antigen) and two alternately expressed flagellar subunits (H1, H2). These antigenic formulas are written in the form O epitopes:H1 epitopes:H2 epitopes in the revised Kauffmann-White scheme [63,64]. Detailed metabolic maps have been elucidated for the synthesis of the LPS from 47 O serogroups which result in the expression of the O epitopes [65,66]. These O groups have distinct numeric designations, e.g. O:4 for the O group of serovar Typhimurium. Serogroups can be determined for natural isolates using specific antisera for serological tests or *in silico* from genomic sequences with

the programs SeqSero [67] or SISTR [68]. In practice. each of these methods has an error rate of a few percent [27,69].

Several hundred legacy eBGs within subsp. *enterica* were relatively uniform for serovar [27], with multiple exceptions, and many serovars were associated with multiple, apparently unrelated eBGs, likely due to the extensive exchange of genes encoding LPS and flagellar epitopes between eBGs early in their evolutionary history. We proposed that legacy MLST was an even better metric for identifying the natural populations and could replace serotyping completely [27]. Our new data indicate that HC900 clusters are even more reliable than legacy eBGs for recognizing natural populations within *Salmonella*.

Several phylogenetic comparisons indicated that HC900 clusters based on cgMLST are concordant with eBGs in *Salmonella* [32,41]. Here we provide independent analyses confirming these observations. There is a very strong quantitative concordance between eBG and HC900 clusters according to the Adjusted Mutual Information Score [AMI] (Table 4), a metric that is suitable for samples with a heterogeneous size distribution [70]. Comparisons of eBG or HC900 with HC2000 clustering yielded slightly lower AMI scores (Table 4). We also tested whether HC900 clusters are uniform for serovar. First, we corrected false serological data in the metadata Serovar field in EnteroBase by manual curation of 790 HC900 clusters that contained at least 5 entries in late 2019. After manual curation, 97% (767/790) of HC900 clusters with at least five entries were uniform (≥95%) for serovar (Supplemental Table S8), as were most HC2000 clusters (342/409 clusters; 83.4%). Changes in the O group between HC900 clusters within an HC2000 cluster was even rarer: in 95.3% of 404 HC2000 clusters within *S. enterica* subsp. *enterica*, the O

group was uniform across all internal HC900 clusters and the same was the case for 80.5% of 77 HC2000 clusters in other *Salmonella* species and sub-species (Fig. 4; Supplemental Table S9). These observations confirm and extend the prior conclusions about concordance between natural populations and serogroup within *Salmonella.*

Escherichia coli. Patterns of multilocus isoenzyme electrophoresis were used in the 1980s to provide an overview of the genetic diversity of *E. coli* (see overview by Chaudhuri and Henderson [71]) . Those analyses yielded a representative collection of 72 isolates [72], the EcoR collection, whose deep phylogenetic branches were designated haplogroups A, B1, B2, C, D and E [73]. Several haplogroups have since been added [74,75]. The presence or absence of several accessory genes can be used for the assignment of genomes to haplogroups with the Clermont scheme [76,77], and the haplogroup can also be predicted *in silico* from genomic assemblies with ClermontTyping [78] or EZClermont [79], both of which are implemented within EnteroBase. However, the Clermont scheme ignores *Shigella*, which consists of *E. coli* clades despite its differing genus designation [26,44,45], and makes multiple discrepant assignments according to phylogenetic trees [28]. The Clermont scheme also does not properly handle the entire diversity of species and environmental clades/subspecies within the genus *Escherichia* [28,80-82].

EnteroBase HierCC automatically assigns genomes within the *Escherichia/Shigella* database to the cgMLST equivalents of Lineages (HC2000 clusters) and ST Complexes (HC1100 clusters) (Table 1). It also perpetuates the ST Complexes that were initially defined for legacy MLST by Wirth *et al.* [26]. However, the numbers and composition of legacy ST Complexes have not been updated since 2009 because additional sequencing data defined intermediate, recombinant genotypes that were

equidistant and similar to more than one ST Complex, and threatened to merge existing ST Complexes. In contrast, HC1100 clusters do not merge: Intermediate genotypes are rare because cgMLST involves 2512 loci while legacy MLST is based on only seven. Secondly, new genotypes which are similar to and equidistant to multiple clusters would not trigger merging because HierCC arbitrarily assigns them to the oldest of the existing alternatives. Finally, unlike legacy MLST which has not been actively pursued in *Escherichia* since 2006, HierCC clustering is an ongoing, automated process for all new genotypes.

Frequent recombination in *E. coli* results in poor bootstrap support for the deep branches in phylogenetic trees of concatenated genes [26], and we were unable to identify a unique HierCC level which was largely concordant with the Clermont scheme. However, similar to *Salmonella,* HC1100 clusters identify clear population groups, which are highly concordant with legacy ST Complexes (Table 4). Unlike *Salmonella*, HC2000 clusters did not in general mark recognizable additional population structure beyond that which was provided by ST Complexes, and HC2000 clusters are only moderately concordant with either legacy ST Complexes or their cgMLST equivalent, HC1100 clusters (Table 4). Exceptionally, legacy ST Complexes which originally grouped multiple *Shigella* species [26] equate to HC2000 clusters rather than HC1100 clusters according to phylogenetic trees. HC2000_305 replaces ST152 Cplx, and encompasses all *S. sonnei*. HC2000_192 replaces ST245 Cplx, and contains many *S. flexneri*. However, *S. flexneri* O:group F6 is in an HC1100 cluster together with *S. boydii* B2 and B4 within HC2000_1465 (ST243 Cplx). HC2000_1465 also include a second HC1100 cluster with *S. boydii* O groups B1 and B18, and a third with *S. dysenteriae* D3, D9 and D13 (Fig. 5). HC2000_4118 replaces

the combination of ST250 Complex and ST149 Complex, which have merged, and contains both *S. dysenteriae* as well as *S. boydii*.

Unlike *Salmonella*, O serogroups are remarkably variable within HC1100 clusters of *E. coli* (Fig. 5). HC1100_63 (ST11 Complex) combines classical EHEC strains of serovar O157:H7 and atypical EPEC isolates of serovar O55:H7, which are thought to be ancestral [83]. Multiple other HC1100 clusters that contain EHEC strains also encompass multiple O groups, including HC1100_2 (ST29 Cplx; O26, O103, O111) [84] and HC1100_3 (ST20 Cplx; O45, O103, O111, O128). Similarly, multiple HC1100 clusters of *E. coli* that cause extra-intestinal diseases also contain a variety of O groups, including HC1100_7 (ST131 Cplx; O16, O25) [23,85,86] and HC1100_44 (K1-encapsulated bacteria of the ST95 Cplx; O1, O2, O18, O25, O45) [26,87,88]. Indeed, the primary impression from Fig. 5 is that O group variation within an HC1100 cluster is nearly universal throughout *E. coli*, confirming prior conclusions about the high frequency of homologous recombination in this species [26,74].

Other genera. Correlations have been described in *Clostridioides difficile* [42] between intra-specific HierCC clusters and population groupings at multiple levels, including ST Complexes and ribotypes. It remains to be proven whether HierCC can identify populations within *Yersinia* because bacterial populations in *Y. pseudotuberculosis* are largely obscured by recombination [89]. Similarly, we are not aware of extensive efforts to identify bacterial populations within *Vibrio* which could be compared with HierCC. However, recent work has described the assignment of large numbers of genomes of *Streptococcus pneumoniae* to legacy Clonal Complexes as well as GSPC (Global Pneumococcal Sequence Clusters) [90], initially with PopPunk [91] and more recently with Mandrake [92]. Most of the Illumina short reads

for those genomes had been uploaded to ENA, and were automatically imported into EnteroBase, which then assembled genomic sequences and assigned them to HierCC clusters. We therefore compared the HierCC and GSPC assignments for 18,147 genomes, and also compared them to Clonal Complexes (CCs) from legacy MLST (13,396 genomes). HC160 clusters yielded the greatest AMI and ARI scores with HC160 clusters while CC was most concordant with HC100 clusters (Table 4). Visual examination of of nodes color-coded by the different assignments within a Neighbor-Joining tree based on cgMLST distances supported these conclusions (Supplemental Fig. S4).

Discussion

We initially anticipated that EnteroBase might contain as many as 10,000 genomes of *Salmonella* and *Escherichia*. Instead, EnteroBase currently contains >600,000 assembled genomes of *Salmonella, Escherichia/Shigella, Streptococcus, Vibrio, Clostridioides* and *Yersinia*, and has become one of the primary global sources of their genomic data. An analyses of their genomic properties is facilitated by software tools (Table 2) that support indexing of genetic diversity, and searching as well as investigating stable population structures. Here we focused on the automated identification and designation of species/subspecies and populations on the basis of HierCC of core genome MLST genotypes. The data demonstrate that HierCC can completely replace classical taxonomic methods for genomes of *Salmonella, Escherichia/Shigella, Vibrio, Clostridioides* and *Yersinia*, with only few minor exceptions. HierCC is also a complete replacement for legacy MLST regarding the assignments to populations within *Salmonella, Escherichia/Shigella* and *Streptococcus pneumoniae.*

Taxonomic assignments. Disjointed tree merging (DTM) within a divide-and-conquer approach allows large phylogenetic trees to be calculated in a reasonable time [93]. We developed cgMLSA, a novel DTM approach that enabled analyses of the core and accessory genome diversity of up to 10,000 representative genomes from each of six bacterial genera. ML super-trees were calculated for each genus based on core genome SNPs and from a presence/absence matrix for accessory genes. The paired trees were generally topologically concordant. Therefore, species/subspecies are well defined and reproducible in these genera. We compared current taxonomic species designations with their topologies in the paired trees, and against clusters

based on ANI and HierCC patterns. HierCC was at least as compatible with current taxonomy as 95% ANI for *Salmonella*, *Escherichia/Shigella*, *Vibrio*, *Clostridioides* and *Yersinia*, and was much better for detecting sub-species. Classical taxonomical designations had multiple, glaring problems in each genus and we introduced an automated HierCC-based species/subspecies taxonomy in EnteroBase for problematical branches. We also defined novel species/sub-species and gave them HierCC designations. Why did we find so many obvious discrepancies? This may reflect the additional novel possibilities introduced by EnteroBase and HierCC, and the previously unprecedented analysis of so many representative genomes.

Our taxonomic changes in *Yersinia* are likely to be highly controversial. According to both ML trees and HierCC, the current designation "*Y. enterocolitica*" encompasses four, previously unnamed species/sub-species. These taxa are not distinguished by 95% ANI and had not previously been assigned distinct names. We maintain continuity with the traditional designation of *Y. enterocolitica*, and simply add HierCC affiliations to the species name, e.g. *Y. enterocolitica* HC1490_2. Alternatively, they could be considered to represent sub-species and their name modified slightly, e.g. *Y. enterocolitica* subspecies HC1490_2, similar to our downgrading of the named species *Y. pestis*, *Y. similis* [52] and *Y. wautersii* [51] on phylogenetic grounds to #to subspecies of the *Y. pseudotuberculosis* Complex (Fig. S1). Finally, we defined eight new species/sub-species within *Yersinia* that were not previously differentiated, and named them after their HC cluster, e.g. *Yersinia HC1490_4399*, without submitting traditional evidence for simple phenotypic differences. We also followed comparable strategies for the other genera analyzed here.

In *Salmonella*, HierCC and 95% ANI identified a new species, *Salmonella HC2850_215890*, the first since decades (Fig. 1). HierCC also identified a new subspecies, *S. enterica* subsp. HC2850_222931. Five distinct new species were identified within *C. difficile* by ANI as well as by HierCC, and HierCC identified two additional new sub-species. *Escherichia* taxonomy has already been dealt with previously [28], and an updated, more comprehensive analysis including HierCC assignments will be published elsewhere.

The ML trees presented here (Fig. S2) supported prior definitions of high order clades containing multiple species in *Vibrio* on the basis of MLSA [54]. We also observed a general concordance between ANI and HierCC for 109 named *Vibrio* species, However, 43 HierCC clusters of species rank had not previously been properly classified with species designations and 17 other species names were superfluous. These changes are fully described in Supplemental text, allowing closer scrutiny by the *Vibrio* taxonomic community for consistency with other properties.

HierCC was so effective in clarifying the taxonomies of these five species that we had hoped that it could also facilitate the taxonomic classification of *Streptococcus*. 102 HC363 clusters were indeed concordant with 95% ANI and taxonomic designations, with only minor exceptions. However, extensive discrepancies existed between 95% ANI and HierCC for numerous other clusters, and both approaches showed multiple discrepancies with classical taxonomic designations. The existence of multiple 95% ANI clusters within *S. mitis* and *S. oralis* and other species of *Streptococcus* has previously been commented on by Kilian and his colleagues [15,16,55,56]. Our observations extend these taxonomic problems to multiple additional species in which HC363 does not distinguish between pairs of named species (Fig. 3). Neither 95%

ANI nor cgMLST HierCC seems to provide a suitable general strategy for elucidating the taxonomy of all of *Streptococcus*, and we failed to find a general solution to this problem.

HierCC *versus* classical taxonomy. Classical microbiological species taxonomy involves identifying a type strain whose phenotype can be distinguished from all other type strains, identifying additional isolates with similar phenotypes, demonstrating distinct clustering from other known species in phylogenetic trees based on DNA or amino acid sequences, and publishing a report in one of a very limited number of acceptable journals. Such species definitions are then considered tentative until an international committee has approved them. Other forms of identifying species that include sole reliance on DNA sequence differences are not acceptable [7,8]. The metagenomics community and scientists working with uncultivated organisms from the environment have largely liberated themselves from such intellectually obsolete regulations, and tend to use operational taxonomic units (OTUs) as taxonomic entities. However, the taxonomic species structure of many microbes from environmental sources remains fuzzy, or does not clearly correspond to classical taxonomy (e.g. *Prochlorococcus* [94] or *Synechococcus* [95]). Furthermore, we are not aware of a prior method that has been adequate to the tasks of assigning 1000s of genomes per day to existing taxa and also identifying new taxa automatically as they appear. HierCC performs this task with bravura for the genera in Table 1, with the notable exception of *Streptococcus*.

We have taken the liberty of dropping all attempts to reconcile HierCC species/subspecies clusters with classical prescriptions for how to define a species. Instead we have adopted the practice of using HierCC cluster designations within

EnteroBase for the nomenclatures for species/subspecies groupings which had not yet been identified by others. We split *Y. enterocolitica* into four related taxa and changed the nomenclatures of species within the *Y. pseudotuberculosis* Complex to that of subspecies. We also eliminated from EnteroBase multiple species designations of type strains in *Vibrio* and *Streptococcus* where they did not match the ML tree topologies and HierCC clusters. These actions provided a uniform base for the future additions of additional species and ensured that future genomes can be correctly assigned to uniform clusters of related strains. Scientists wishing to use these schemes to identify the species of their bacterial isolates can upload their sequenced genomes to EnteroBase. The HierCC assignments will be available within hours. We also welcome additional curators of these databases who are willing to test and improve the current taxonomic assignments we have implemented. But we reject the concept of designating our assignments as tentative until they are confirmed in several years by an international committee.

HierCC and Populations. The first bacterial taxonomic designations were assigned over 100 years ago. Bacterial population genetics is much younger, and has many fewer practitioners. Initial population genetic analyses in the early 1980s subdivided multiple bacterial species into intra-specific lineages based on multilocus enzyme electrophoresis [72]. This methodology was replaced in the late 1990s by legacy MLST based on several housekeeping genes [21], which is currently being replaced by cgMLST based on all common genes in the soft core genome [22,30]. EnteroBase calculates genotypes for both types of MLST, as well as for rMLST [28]. Legacy STs differ from cgSTs due to different levels of resolution. However, the boundaries of ST Complexes/eBGs are highly concordant independent of the genes used for MLST (Table 4). eBGs in *S. enterica* correspond to HC900 clusters [27,46], ST Complexes

in *E. coli* correspond to HC1100 clusters and Clonal Complexes (CCs) in *S. pneumoniae* to HC100 clusters (Table 4). Similarly, serovars are largely uniform within HC900 clusters/eBGs in *S. enterica* [27] as are serotypes across CCs in *S. pneumoniae* [90].

We interpret these consistencies between legacy, rMLST and cgMLST as reflecting the existence of natural populations. We previously claimed that legacy ST Complexes in *E. coli* were unstable due to frequent recombination [26], and were therefore not surprised at their tendency to merge as additional isolates were sequenced. However, a major problem with legacy MLST is that it is based on only seven genes. The high resolution of cgMLST and the decision to assign new genotypes that are equidistant from multiple HierCC clusters to the older cluster have negated these problems, and *E. coli* HC1100 clusters correspond to bacterial groupings that have been independently identified by multiple phenotypic patterns. In the early 1980s, MA began his academic research on bacterial pathogens with *E. coli* that expressed the K1 polysaccharide capsule [87]. He observed unusually uniform patterns of electrophoretic mobility of major outer membrane proteins across multiple isolates, and interpreted those bacteria as representing “clones”. These “clones” correspond to HC1100 clusters, including HC1100_44 (ST95 Complex) which is still a major cause of extra-invasive disease in humans and animals around the globe [88]. The 1980s analyses showed that these K1 bacteria were variably O1, O2, or O18, and that the serotype variants differed both in their invasiveness and in the hosts which they infected [96]. As indicated in Fig. 5, HC1100_44 includes O groups O25 and O45 in addition to O1, O2 and O18. Similarly, HC1100_7 (ST131) represents another major cause of extraintestinal disease in humans [23,85,86]. EHEC bacteria are notorious for their association with hemolytic uremic syndrome (HUS). Many of them

belong to distinctive HC1100 clusters, including O157:H7/O55:H7 (HC1100_63, ST11 Cplx [83]) and O26, O103 and O111 EHEC bacteria (HC1100_2, ST29 Cplx [84]) (Fig. 5).

In contrast to such correlations with epidemiological groupings, we were unable to identify any HierCC levels that were consistently concordant with other intra-specific subdivisions based on phylogenetic SNP trees, including haplogroups [73] or Clermont typing [76,77] in *E. coli* [28] and Clades [58-61], lineages [57] or Branches [62] in *S. enterica.* Those intra-specific phylogenetic sub-divisions may not necessarily be informative across the entire species.

Lineages. Instead of phylogenetic sub-divisions, we introduce Lineages, a subset of population structures that are of higher order than ST Complexes/eBGs and lower than sub-species. In *E. coli*, HC2000 Lineages were particularly appropriate for defining the clustering level of seven groups of *Shigella* genomes (Fig. 5). In *S. enterica* subspecies *enterica*, several HC2000 Lineages combined multiple serovars. These included the Para C Lineage (HC2000_1272; serovars Paratyphi C, Typhisuis, Choleraesuis and Lomita [47]) which emerged about 3500 years ago (Supplemental Text) [47,62]. The observation that the Para C Lineage is uniformly serogroup O:7 is not particularly surprising because all its serovars have identical antigenic formulas, and are only distinguished by biotype. The Enteritidis Lineage (HC2000_12) consists of serovars Enteritidis (1,9,12:g,m:-), Gallinarum and its variant Pullorum (1,9,12:-:-), and Dublin (1,9,12:g,p:-:-) [41] (Fig. 4), and it is also not surprising that it is uniform for O:9. However, we were not prepared for the Typhimurium Lineage (HC2000_2), which includes serovars Typhimurium, Heidelberg, Reading, Saintpaul, Haifa, Stanleyville and others; [41]) All these serovars are O:4, and this Lineage is therefore

also uniform for O group. The Mbandaka Lineage (HC2000_4) is uniformly serogroup O:7 because serovar Mbandaka is 6,7,14: z10:e,n,z15 and serovar Lubbock is 6,7:g,m,s:e,n,z15. However, three Lineages did contain more than one O serogroup. HC2000_299 is a mixture of O:4 (HC900_299; serovar Agama [28]) and O:7 (HC900_127074; serovar Nigeria). HC900 clusters within HC2000_54 are O:4 (HC900_54: Bredeney, Schwarzengrund; HC24937: Kimuenza) or O:9,46 (HC900_57: Give). Similarly, HC900 groups within HC2000_106 are O:3,10; O:9; O:9,46; O:4: or O:8. These exceptional HC2000 Lineages might be worth exploring in greater detail to reconstruct the recombination that has resulted in such differences in LPS epitopes.

HC160 clusters are the equivalent of Lineages in *S. pneumonia*, and HC100 the equivalent of Clonal Complexes. HC160 clusters were strongly concordant with PopPunk GSPC clustering. Otherwise, little is yet known about the general properties of Lineages, except that they may vary across multiple species or genera. For example, whereas *S. enterica* HC2000 Lineages were uniform for O group, O serogroups were heterogeneous within HC1100 clusters/ST Complexes within *E. coli* (Fig. 5).

Future prospects. EnteroBase was conceived to satisfy a need that MA and ZZ perceived in 2014 [1]. We contend that it is now fit for purpose for microbial epidemiology and population genetics of multiple bacterial genera for scientists ranging from beginners through to experts in those areas. Its original creators have now all left this project, but EnteroBase is being maintained as a service for the global community by the University of Warwick. Maintenance of its technical functions and databases are thereby assured for the near future. Further functional developments

will, however, depend on increased participation and ownership by its users. We perceive a general trend to focus on insular solutions that can satisfy the demands of individual bioinformaticians and regional diagnostic laboratories. Such approaches can yield relatively rapid progress in solving short-term needs. However, a global overview of genomic diversity needs central databases to ensure definitive terminology. Pooling efforts on a central endeavor at the scale represented by EnteroBase would ensure that it continues to function over decades, is representative over all continents, and serves the global community even better. We therefore welcome additional curators and scientific experts as well as bioinformatics collaborations to improve EnteroBase even more.

Methods (see Supplemental Text)

Funding

MA was supported by grant 202792/Z/16/Z from the Wellcome Trust (https://wellcome.org/).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Code.

cgMLSA is available for downloading at https://github.com/zheminzhou/cgMLSA. All trees in figures and genomic data are freely available within EnteroBase for interactive examination. Other code is described in Supplemental Text

Acknowledgements

We are very grateful to François-Xavier Weill for permission to include Table S8 for public access in Supplemental Material. We are very grateful to François-Xavier Weill and Tandy Warnow for their comments on an earlier draft of this manuscript.

References

1. Achtman, M. and Zhou, Z. 2014 Distinct genealogies for plasmids and chromosome. *PLoS Genet* **10**, e1004874. (doi:10.1371/journal.pgen.1004874 [doi];PGENETICS-D-14-02780 [pii])

2. Kauffmann, F. 1961 *Die Bakteriologie der Salmonella-Species*. Copenhagen, Denmark: Munksgaard.

3. Morelli, G. *et al.*. 2010 *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nature Genet* **42**, 1140-1143.

4. Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A., and Carniel, E. 1999 *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci USA* **96**, 14043-14048.

5. Pulford, C. V. *et al.*. 2021 Stepwise evolution of *Salmonella* Typhimurium ST313 causing bloodstream infection in Africa. *Nat Microbiol* **6**, 327-338. (doi:10.1038/s41564-020-00836-1 [doi];10.1038/s41564-020-00836-1 [pii])

6. Wayne, L. G. *et al.*. 1987 Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *In J Syst Bacteriol* **37**, 463-464.

7. Murray, A. E. *et al.*. 2020 Roadmap for naming uncultivated Archaea and Bacteria. *Nat Microbiol* **5**, 987-994. (doi:10.1038/s41564-020-0733-x [doi];10.1038/s41564-020-0733-x [pii])

8. Sutcliffe, I. C., Dijkshoorn, L., Whitman, W. B., and Executive Board. 2020 Minutes of the International Committee on Systematics of Prokaryotes online discussion on the proposed use of gene sequences as type for naming of prokaryotes, and outcome of vote. *Int J Syst Evol Microbiol* **70**, 4416-4417. (doi:10.1099/ijsem.0.004303 [doi])

9. Konstantinidis, K. T. and Tiedje, J. M. 2005 Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* **102**, 2567-2572.

10. Jain, C., Rodriguez, R., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. 2018 High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114. (doi:10.1038/s41467-018-07641-9 [doi];10.1038/s41467-018-07641-9 [pii])

11. Auch, A. F., von, J. M., Klenk, H. P., and Goker, M. 2010 Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand Genomic Sci* **2**, 117-134. (doi:10.4056/sigs.531120 [doi])

12. Garrido-Sanz, D., Meier-Kolthoff, J. P., Goker, M., Martin, M., Rivilla, R., and Redondo-Nieto, M. 2016 Genomic and Genetic Diversity within the *Pseudomonas fluorescens* Complex. *PLoS One* **11**, e0150183. (doi:10.1371/journal.pone.0150183 [doi];PONE-D-15-47408 [pii])

13. Richter, M. and Rosselló-Mora, R. 2009 Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A* **106**, 19126-19131. (doi:0906412106 [pii];10.1073/pnas.0906412106 [doi])

14. Maderankova, D., Jugas, R., Sedlar, K., Vitek, M., and Skutkova, H. 2019 Rapid bacterial species delineation based on parameters derived from genome numerical representations. *Comput Struct Biotechnol J* **17**, 118-126. (doi:10.1016/j.csbj.2018.12.006 [doi];S2001-0370(18)30122-3 [pii])

15. Jensen, A., Scholz, C. F., and Kilian, M. 2016 Re-evaluation of the taxonomy of the Mitis group of the genus *Streptococcus* based on whole genome phylogenetic analyses, and proposed reclassification of *Streptococcus dentisani* as *Streptococcus oralis* subsp. *dentisani* comb. nov., *Streptococcus tigurinus* as *Streptococcus oralis* subsp. *tigurinus* comb. nov., and *Streptococcus oligofermentans* as a later synonym of *Streptococcus cristatus*. *Int J Syst Evol Microbiol* **66**, 4803-4820. (doi:10.1099/ijsem.0.001433 [doi])

16. Kilian, M., Poulsen, K., Blomqvist, T., Havarstein, L. S., Bek-Thomsen, M., Tettelin, H., and Sorensen, U. B. 2008 Evolution of *Streptococcus pneumoniae* and its close commensal relatives. *PLoS ONE* **3**, e2683.

17. Zhou, Z., Charlesworth, J., and Achtman, M. 2020 Accurate reconstruction of bacterial pan- and core genomes with PEPPAN. *Genome Res* **30**, 1667-1679. (doi:DOI 10.1011/gr.260828.120)

18. Gomila, M., Pena, A., Mulet, M., Lalucat, J., and Garcia-Valdes, E. 2015 Phylogenomics and systematics in *Pseudomonas*. *Front Microbiol* **6**, 214. (doi:10.3389/fmicb.2015.00214 [doi])

19. Beaz-Hidalgo, R., Hossain, M. J., Liles, M. R., and Figueras, M. J. 2015 Strategies to avoid wrongly labelled genomes using as example the detected wrong taxonomic affiliation for *Aeromonas* genomes in the GenBank database. *PLoS ONE* **10**, e0115813. (doi:10.1371/journal.pone.0115813 [doi];PONE-D-14-36626 [pii])

20. Achtman, M. and Wagner, M. 2008 Microbial diversity and the genetic nature of microbial species. *Nat Rev Microbiol* **6**, 431-440.

21. Maiden, M. C. J. *et al.*. 1998 Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA* **95**, 3140-3145.

22. Jolley, K. A., Bray, J. E., and Maiden, M. C. 2018 Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* **3**, 124. (doi:doi: 10.12688/wellcomeopenres.14826.1)

23. Stoesser, N. *et al.*. 2016 Evolutionary history of the global emergence of the *Escherichia coli* epidemic clone ST131. *MBio* **7**, e02162. (doi:mBio.02162-15 [pii];10.1128/mBio.02162-15 [doi])

24. Kingsley, R. A. *et al.*. 2009 Epidemic multiple drug resistant *Salmonella* Typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype. *Genome Res* **19**, 2279-2287.

25. Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P., and Spratt, B. G. 2004 eBURST: Inferring patterns of evolutionary descent among clusters of related bacterial genotypes from Multilocus Sequence Typing data. *J Bacteriol* **186**, 1518-1530.

26. Wirth, T. *et al.*. 2006 Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* **60**, 1136-1151.

27. Achtman, M. *et al.*. 2012 Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLoS Pathog* **8**, e1002776.

28. Zhou, Z., Alikhan, N. F., Mohamed, K., Fan, Y., and Achtman, M. 2020 The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Res* **30**, 138-152. (doi:gr.251678.119 [pii];10.1101/gr.251678.119 [doi])

29. Mellmann, A. *et al.*. 2011 Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS ONE* **6**, e22751.

30. Maiden, M. C., van Rensburg, M. J., Bray, J. E., Earle, S. G., Ford, S. A., Jolley, K. A., and McCarthy, N. D. 2013 MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Microbiol* **11**, 728-736.

31. Moura, A. *et al.*. 2016 Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. *Nat Microbiol* **2**, 16185. (doi:nmicrobiol2016185 [pii];10.1038/nmicrobiol.2016.185 [doi])

32. Zhou, Z., Charlesworth, J., and Achtman, M. 2021 HierCC: A multi-level clustering scheme for population assignments based on core genome MLST. *Bioinformatics* **37**, 3645-3646. (doi:10.1093/bioinformatics/btab234 [doi])

33. Bielaszewska, M., Mellmann, A., Zhang, W., Kock, R., Fruth, A., Bauwens, A., Peters, G., and Karch, H. 2011 Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study. *Lancet Infect Dis* **11**, 671-676.

34. Jones, G. *et al.*. 2019 Outbreak of *Salmonella enterica* serotype Poona in infants linked to persistent *Salmonella* contamination in an infant formula manufacturing facility, France, August 2018 to February 2019. *Euro Surveill* **24**, 13. (doi:doi.org/10.2807/1560-7917.ES.2019.24.13.1900161)

35. Jones, G. *et al.*. 2019 Outbreak of *Shiga* toxin-producing *Escherichia coli* (STEC) O26 paediatric haemolytic uraemic syndrome (HUS) cases associated with the consumption of soft raw cow's milk cheeses, France, March to May 2019. *Euro Surveill* **24**, 1900305. (doi:10.2807/1560-7917.ES.2019.24.22.1900305)

36. Van den Bossche, A., Ceyssens, P. J., Denayer, S., Hammami, N., van den Beld, M., Dallman, T. J., and Mattheus, W. 2021 Outbreak of Central American born *Shigella sonnei* in two youth camps in Belgium in the summer of 2019. *Eur J Clin Microbiol Infect Dis* **40**, 1573-1577. (doi:10.1007/s10096-021-04164-y [doi];10.1007/s10096-021-04164-y [pii])

37. European Centre for Disease Prevention and Control, E. F. S. A. 2020 Multi-country outbreak of *Salmonella* Typhimurium and *S.* Anatum infections linked to Brazil nuts - 21 October 2020. *EFSA Supporting Publications* **17**, 1944E. (doi:https://doi.org/10.2903/sp.efsa.2020.EN-1944)

38. Hawkey, J. *et al.*. 2021 Global population structure and genotyping framework for genomic surveillance of the major dysentery pathogen, *Shigella sonnei*. *Nat Commun* **12**, 2684. (doi:10.1038/s41467-021-22700-4 [doi];10.1038/s41467-021-22700-4 [pii])

39. Park, S. E. *et al.*. 2018 The phylogeography and incidence of multi-drug resistant typhoid fever in sub-Saharan Africa. *Nat Commun* **9**, 5094. (doi:10.1038/s41467-018-07370-z [doi];10.1038/s41467-018-07370-z [pii])

40. Weill, F. X. *et al.*. 2017 Genomic history of the seventh pandemic of cholera in Africa. *Science* **358**, 785-789. (doi:358/6364/785 [pii];10.1126/science.aad5901 [doi])

41. Achtman, M. *et al.*. 2020 Genomic diversity of *Salmonella enterica* -The UoWUCC 10K genomes project. *Wellcome Open Res* **5**, 223. (doi:10.12688/wellcomeopenres.16291.2 [doi])

42. Frentrup, M. *et al.*. 2020 A publicly accessible database for *Clostridioides difficile* genome sequences supports tracing of transmission chains and epidemics. *Microbial Genomics*, mgen.0.000410. (doi:10.1099/mgen.0.000410 [doi])

43. Achtman, M., van den Broeck, F., Cooper, K. K., Lemey, P., Parker, C. T., Zhou, Z., and the ATCC14028s Study Group. 2021 Genomic population structure associated with repeated escape of *Salmonella enterica* ATCC14028s from the laboratory into nature. *PLoS Genet* **17**, e1009820. (doi:DOI: 10.1371/journal.pgen.1009820)

44. Yassine, I. *et al.*. 2021 Population structure analysis and laboratory monitoring of *Shigella* with a standardised core-genome multilocus sequence typing scheme: a validation study. *BioRxiv*, 2021.

45. Pupo, G. M., Lan, R., and Reeves, P. R. 2000 Multiple independent origins of Shigella clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci USA* **97**, 10567-10572.

46. Alikhan, N.-F., Zhou, Z., Sergeant, M. J., and Achtman, M. 2018 A genomic overview of the population structure of *Salmonella*. *PLoS Genet* **14**,

e1007261. (doi:10.1371/journal.pgen.1007261 [doi];PGENETICS-D-18-00122 [pii])

47. Zhou, Z. *et al.*. 2018 Pan-genome analysis of ancient and modern *Salmonella enterica* demonstrates genomic stability of the invasive Para C Lineage for millennia. *Curr Biol* **28**, 2420-2428. (doi:https://doi.org/10.1016/j.cub.2018.05.058)

48. Pearce, M. E., Langridge, G. C., Lauer, A. C., Grant, K., Maiden, M. C. J., and Chattaway, M. A. 2021 An evaluation of the species and subspecies of the genus *Salmonella* with whole genome sequence data: Proposal of type strains and epithets for novel *S. enterica* subspecies VII, VIII, IX, X and XI. *Genomics* **113**, 3152-3162. (doi:S0888-7543(21)00281-0 [pii];10.1016/j.ygeno.2021.07.003 [doi])

49. Knight, D. R., Imwattana, K., Kullin, B., Guerrero-Araya, E., Paredes-Sabja, D., Didelot, X., Dingle, K. E., Eyre, D. W., Rodríguez, C., and Riley, T. V. 2021 Major genetic discontinuity and novel toxigenic species in *Clostridioides difficile* taxonomy. *eLife* **10**, e64325.

50. Reuter, S. *et al.*. 2014 Parallel independent evolution of pathogenicity within the genus *Yersinia*. *Proc Natl Acad Sci U S A* **111**, 6768-6773. (doi:1317161111 [pii];10.1073/pnas.1317161111 [doi])

51. Savin, C. *et al.*. 2014 The *Yersinia pseudotuberculosis* complex: characterization and delineation of a new species, *Yersinia wautersii*. *Int J Med Microbiol* **304**, 452-463. (doi:S1438-4221(14)00012-5 [pii];10.1016/j.ijmm.2014.02.002 [doi])

52. Sprague, L. D., Scholz, H. C., Amann, S., Busse, H. J., and Neubauer, H. 2008 *Yersinia similis* sp. nov. *Int J Syst Evol Microbiol* **58**, 952-958.

53. Bercovier, H., Mollaret, H. H., Alonso, J. M., Brault, J., Fanning, G. R., Steigerwalt, A. G., and Brenner, D. J. 1980 Intra- and interspecies relatedness

of *Yersinia pestis* by DNA hybridization and its relationship to *Yersinia pseudotuberculosis*. *Curr Microbiol* **4**, 225-229.

54. Gomez-Gil, B., Thompson, C. C., Matsumura, Y., Sawabe, T., Iida, T., Christen, R., Thompson, F., and Sawabe, T. 2014 The family *Vibrionaceae*. In *The prokaryotes - Gammaproteobacteria* (eds. E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, and F. Thompson), pp. 660-747. Berlin Heidelberg.

55. Kilian, M., Riley, D. R., Jensen, A., Bruggemann, H., and Tettelin, H. 2014 Parallel evolution of *Streptococcus pneumoniae* and *Streptococcus mitis* to pathogenic and mutualistic lifestyles. *MBio* **5**, e01490-14. (doi:mBio.01490-14 [pii];10.1128/mBio.01490-14 [doi])

56. Bishop, C. J., Aanensen, D. M., Jordan, G. E., Kilian, M., Hanage, W. P., and Spratt, B. G. 2009 Assigning strains to bacterial species via the internet. *BMC Biol* **7**, 3.

57. Didelot, X. *et al.*. 2011 Recombination and population structure in *Salmonella enterica*. *PLoS Pathog* **7**, e1002191.

58. den Bakker, H. C. *et al.*. 2011 Genome sequencing reveals diversification of virulence factor content and possible host adaptation in distinct subpopulations of *Salmonella enterica*. *BMC Genomics* **12**, 425.

59. Timme, R. E., Pettengill, J. B., Allard, M. W., Strain, E., Barrangou, R., Wehnes, C., Van Kessel, J. S., Karns, J. S., Musser, S. M., and Brown, E. W. 2013 Phylogenetic diversity of the enteric pathogen *Salmonella enterica* subsp. *enterica* inferred from genome-wide reference-free SNP characters. *Genome Biol Evol* **5**, 2109-2123. (doi:evt159 [pii];10.1093/gbe/evt159 [doi])

60. Parsons, S. K., Bull, C. M., and Gordon, D. M. 2011 Substructure within *Salmonella enterica* subspecies *enterica* isolated from Australian wildlife. *Appl Environ Microbiol* **77**, 3151-3153.

61. Worley, J., Meng, J., Allard, M. W., Brown, E. W., and Timme, R. E. 2018 *Salmonella enterica* phylogeny based on whole-genome sequencing reveals two new clades and novel patterns of horizontally acquired genetic elements. *MBio* **9**. (doi:mBio.02303-18 [pii];10.1128/mBio.02303-18 [doi])

62. Key, F. M. *et al.*. 2020 Emergence of human-specific *Salmonella enterica* is linked to the Neolithization process. *Nature Ecology & Evolution* **4**, 324-333. (doi:https://doi.org/10.1038/s41559-020-1106-9)

63. Guibourdenche, M., Roggentin, P., Mikoleit, M., Fields, P. I., Bockemuhl, J., Grimont, P. A., and Weill, F.-X. 2010 Supplement 2003-2007 (No. 47) to the White-Kauffmann-Le Minor scheme. *Res Microbiol* **161**, 26-29.

64. Issenhuth-Jeanjean, S., Roggentin, P., Mikoleit, M., Guibourdenche, M., De, P. E., Nair, S., Fields, P. I., and Weill, F.-X. 2014 Supplement 2008-2010 (no. 48) to the White-Kauffmann-Le Minor scheme. *Res Microbiol* **165**, 526-530. (doi:S0923-2508(14)00106-5 [pii];10.1016/j.resmic.2014.07.004 [doi])

65. Grimont, P. A. and Weill, F.-X. 2007 *Antigenic formulae of the Salmonella serovars*--9th edition edition. Paris, France: WHO Collaborating Centre for Reference and Research on *Salmonella*.

66. Seif, Y., Monk, J. M., Machado, H., Kavvas, E., and Palsson, B. O. 2019 Systems biology and pangenome of *Salmonella* O-Antigens. *MBio* **10**. (doi:mBio.01247-19 [pii];10.1128/mBio.01247-19 [doi])

67. Zhang, S., den Bakker, H. C., Li, S., Chen, J., Dinsmore, B. A., Lane, C., Lauer, A. C., Fields, P. I., and Deng, X. 2019 SeqSero2: rapid and improved *Salmonella* serotype determination using whole genome sequencing data. *Appl Environ Microbiol*. (doi:AEM.01746-19 [pii];10.1128/AEM.01746-19 [doi])

68. Robertson, J., Yoshida, C., Kruczkiewicz, P., Nadon, C., Nichani, A., Taboada, E. N., and Nash, J. H. E. 2018 Comprehensive assessment of the quality of

Salmonella whole genome sequence data available in public sequence databases using the *Salmonella* in silico Typing Resource (SISTR). *Microb Genom*, 1-11. (doi:10.1099/mgen.0.000151 [doi])

69. Uelze, L., Borowiak, M., Deneke, C., Szabo, I., Fischer, J., Tausch, S. H., and Malorny, B. 2020 Performance and accuracy of four open-source tools for *in silico* serotyping of *Salmonella* spp. based on whole-genome short-read sequencing data. *Appl Environ Microbiol* **86**. (doi:AEM.02265-19 [pii];10.1128/AEM.02265-19 [doi])

70. Romano, S., Vinh, N. X., Bailey, J., and Verspoor, K. 2016 Adjusting for chance clustering comparison measures. *Journal of Machine Learning Research* **17**, 1-32.

71. Chaudhuri, R. R. and Henderson, I. R. 2012 The evolution of the *Escherichia coli* phylogeny. *Infect Genet Evol* **12**, 214-226.

72. Ochman, H. and Selander, R. K. 1984 Standard reference strains of *Escherichia coli* from natural populations. *J Bacteriol* **157**, 690-693.

73. Selander, R. K., Caugant, D. A., and Whittam, T. S. 1987 Genetic structure and variation in natural populations of *Escherichia coli*. In *Escherichia coli and Salmonella typhimurium cellular and molecular biology Vol.II* (eds. F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger), pp. 1625-1648. Washington, D.C.: American Society for Microbiology.

74. Touchon, M., Perrin, A., de Sousa, J. A. M., Vangchhia, B., Burn, S., O'Brien, C. L., Denamur, E., Gordon, D., and Rocha, E. P. 2020 Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. *PLoS Genet* **16**, e1008866. (doi:10.1371/journal.pgen.1008866 [doi];PGENETICS-D-20-00247 [pii])

75. Denamur, E., Clermont, O., Bonacorsi, S., and Gordon, D. 2021 The population genetics of pathogenic Escherichia coli. *Nat Rev Microbiol* **19**, 37-54. (doi:10.1038/s41579-020-0416-x [doi];10.1038/s41579-020-0416-x [pii])

76. Gordon, D. M., Clermont, O., Tolley, H., and Denamur, E. 2008 Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. *Environ Microbiol* **10**, 2484-2496.

77. Clermont, O., Christenson, J. K., Denamur, E., and Gordon, D. M. 2013 The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ Microbiol Rep* **5**, 58-65. (doi:10.1111/1758-2229.12019 [doi])

78. Beghain, J., Bridier-Nahmias, A., Le, N. H., Denamur, E., and Clermont, O. 2018 ClermonTyping: an easy-to-use and accurate *in silico* method for *Escherichia* genus strain phylotyping. *Microb Genom* **4**. (doi:10.1099/mgen.0.000192 [doi])

79. Waters, N. R., Abram, F., Brennan, F., Holmes, A., and Pritchard, L. 2020 Easy phylotyping of *Escherichia coli* via the EzClermont web app and command-line tool. In (Anon.), pp. 1-5.

80. Walk, S. T. 2015 The "Cryptic" Escherichia. *EcoSal Plus* **6**. (doi:10.1128/ecosalplus.ESP-0002-2015 [doi])

81. Shen, Z., Koh, X. P., Yu, Y., Woo, C. F., Tong, Y., and Lau, S. C. K. 2019 Draft genome sequences of 16 strains of *Escherichia* cryptic clade II isolated from intertidal sediment in Hong Kong. *Microbiol Resour Announc* **8**. (doi:8/29/e00416-19 [pii];10.1128/MRA.00416-19 [doi])

82. van der Putten, B. C. L., Matamoros, S., COMBAT consortium, and Schultsz, C. 2021 *Escherichia ruysiae* sp. *nov.*, a novel Gram-stain-negative bacterium, isolated from a faecal sample of an international traveller. *Int J Syst Evol Microbiol* **71**, 004609. (doi:DOI 10.1099/ijsem.0.004609)

83. Wick, L. M., Qi, W., Lacher, D. W., and Whittam, T. S. 2005 Evolution of genomic content in the stepwise emergence of *Escherichia coli* O157:H7. *J Bacteriol* **187**, 1783-1791.

84. Eichhorn, I., Semmler, T., Mellmann, A., Pickard, D., Anjum, M. F., Fruth, A., Karch, H., and Wieler, L. H. 2018 Microevolution of epidemiological highly relevant non-O157 enterohemorrhagic *Escherichia coli* of serogroups O26 and O111. *Int J Med Microbiol* **308**, 1085-1095. (doi:S1438-4221(18)30241-8 [pii];10.1016/j.ijmm.2018.08.003 [doi])

85. Liu, C. M. *et al.*. 2018 *Escherichia coli* ST131-H22 as a foodborne uropathogen. *MBio* **9**. (doi:mBio.00470-18 [pii];10.1128/mBio.00470-18 [doi])

86. Price, L. B. *et al.*. 2013 The epidemic of extended-spectrum-beta-lactamase-producing *Escherichia coli* ST131 is driven by a single highly pathogenic subclone, H30-Rx. *MBio* **4**, e00377-13. (doi:mBio.00377-13 [pii];10.1128/mBio.00377-13 [doi])

87. Achtman, M., Mercer, A., Kusecek, B., Pohl, A., Heuzenroeder, M., Aaronson, W., Sutton, A., and Silver, R. P. 1983 Six widespread bacterial clones among *Escherichia coli* K1 isolates. *Infect Immun* **39**, 315-335.

88. Gordon, D. M. *et al.*. 2017 Fine-scale structure analysis shows epidemic patterns of Clonal Complex 95, a cosmopolitan *Escherichia coli* lineage responsible for extraintestinal infection. *mSphere* **2**. (doi:10.1128/mSphere.00168-17 [doi];mSphere00168-17 [pii])

89. Laukkanen-Ninios, R. *et al.*. 2011 Population structure of the *Yersinia pseudotuberculosis* complex according to multilocus sequence typing. *Environ Microbiol* **13**, 3114-3127.

90. Gladstone, R. A. *et al.*. 2019 International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. *EBioMedicine* **43**, 338-346. (doi:doi: 10.1016/j.ebiom.2019.04.021)

91. Lees, J. A., Harris, S. R., Tonkin-Hill, G., Gladstone, R. A., Lo, S. W., Weiser, J. N., Corander, J., Bentley, S. D., and Croucher, N. J. 2019 Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res* **29**, 304-316. (doi:gr.241455.118 [pii];10.1101/gr.241455.118 [doi])

92. Lees, J. A., Tonkin-Hill, G., Yang, Z., and Corander, J. 2021 Mandrake: visualising microbial population structure by embedding millions of genomes into a low-dimensional representation. *BioRxiv*, 2021. (doi:doi: https://doi.org/10.1101/2021.10.28.466232)

93. Zaharias, P. and Warnow, T. 2022 Recent progress on methods for estimating and updating large phylogenies. In (Anon.).

94. Tschoeke, D., Salazar, V. W., Vidal, L., Campeao, M., Swings, J., Thompson, F., and Thompson, C. 2020 Unlocking the genomic taxonomy of the *Prochlorococcus* collective. *Microb Ecol* **80**, 546-558. (doi:10.1007/s00248-020-01526-5 [doi];10.1007/s00248-020-01526-5 [pii])

95. Salazar, V. W., Tschoeke, D. A., Swings, J., Cosenza, C. A., Mattoso, M., Thompson, C. C., and Thompson, F. L. 2020 A new genomic taxonomy system for the *Synechococcus* collective. *Environ Microbiol* **22**, 4557-4570. (doi:10.1111/1462-2920.15173 [doi])

96. Achtman, M. and Pluschke, G. 1986 Clonal analysis of descent and virulence among selected *Escherichia coli*. *Annu Rev Microbiol* **40**, 185-210.

97. Griffiths, D. *et al.*. 2010 Multilocus sequence typing of *Clostridium difficile*. *J Clin Microbiol* **48**, 770-778.

98. Zhou, Z., Alikhan, N.-F., Sergeant, M. J., Luhmann, N., Vaz, C., Francisco, A. P., Carrico, J. A., and Achtman, M. 2018 GrapeTree: Visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res* **28**, 1395-1404. (doi:DOI: 10.1101/gr.232397.117)

99. Achtman, M. and Zhou, Z. 2020 Metagenomics of the modern and historical human oral microbiome with phylogenetic studies on *Streptococcus mutans* and *Streptococcus sobrinus*. *Phil Trans R Soc B* **375**, 20190573. (doi:http://dx.doi.org/10.1098/rstb.2019.0573)

100. Zhou, Z., Luhmann, N., Alikhan, N.-F., Quince, C., and Achtman, M. 2018 Accurate reconstruction of microbial strains from metagenomic sequencing using representative reference genomes. In *RECOMB 2018*, pp. 225-240: Springer, Cham.

101. Luhmann, N., Holley, G., and Achtman, M. 2021 BlastFrost: fast querying of 100,000s of bacterial genomes in Bifrost graphs. *Genome Biol* **22**, 30. (doi:10.1186/s13059-020-02237-3 [doi];10.1186/s13059-020-02237-3 [pii])

102. Pedregosa, F. *et al.*. 2011 Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* **12**, 2825-2830.

103. Bostock, M., Ogievetsky, V., and Heer, J. 2011 D^3: Data-Driven Documents. *IEEE Trans Vis Comput Graph* **17**, 2301-2309. (doi:10.1109/TVCG.2011.185 [doi])

Table 1. Legacy strains, newly assembled genomes and hierarchical cgST clusters in EnteroBase (09/2021).

Genus	Legacy MLST	cgMLST	# Genomic Loci		# HierCC clusters (HierCC level)		
	(# strains)	(# genomes)	wgMLST	cgMLST	ST Complexes	Lineages	Species/ Sub-species
Salmonella	4,930	312,196	21,065	3,002	3648 (900)	2185 (2000)	15 (2850)
Escherichia/ Shigella	9,525	175,256	25,002	2,512	1379 (1100)	343 (2000)	16 (2350)
Streptococcus		76,718	33,887	372	3681 (100)		132 (363)
Clostridioides		23,197	11,490	2,556	299 (950)		12 (2500)
Vibrio		12,267	152,249	1,128	2000 (800)		155 (1090)
Yersinia	4,286	4,023	19,591	1,553	451 (600)		34 (1490)

NOTE: # Genomic Loci: numbers of CDS that are automatically called for the wgMLST and cgMLST schemes. cut-off: maximum numbers of allelic differences between cgSTs that define HC clusters. EnteroBase automatically calculates Legacy MLST for all *Clostridioides* genomes according to the Griffiths scheme [97], but does not maintain a database of legacy data from ABI sequencing. Additional public databases are presented by EnteroBase for *Helicobacter* (>3500 genomes) and *Moraxella* (>2350 genomes), but these lack a cgMLST scheme. EnteroBase also has a database for >73,000 *Mycobacterium* genomes, but this is not yet publicly available, and still lacks a cgMLST scheme.

Table 2. EnteroBase-related software tools that support the analysis of large numbers of bacterial genomes

Name	Purpose	Citation	URL for Stand-alone version
GrapeTree	GUI for depicting and analysing minimal spanning and NJ trees of character data	[98]	https://github.com/achtman-lab/GrapeTree
PEPPAN	Pan-genome calculation including pseudo-genes from numerous representatives of an entire genus	[17]	https://github.com/zheminzhou/PEPPAN
SPARSE	Assignment of metagenomic reads to individual taxa	[99,100]	https://github.com/zheminzhou/SPARSE
EToKi	EnteroBase toolkit of useful functions and pipelines	[28]	https://github.com/zheminzhou/EToKi
BlastFrost	Efficient k-mer based search for and sequence extraction from genes in large datasets	[101]	https://github.com/nluhmann/BlastFrost
HierCC	Automated hierarchical clustering of cgSTs to existing or novel clusters	[32]	https://github.com/zheminzhou/pHierCC

NOTE: PEPPAN is a stand-alone programme that was used to generate the wgMLST schemes used in EnteroBase. SPARSE is a second stand-alone programme that has been used to extract taxon-specific reads from metagenomes of ancient DNA that were then used with EToKi to define pseudo-MAGs that were uploaded to EnteroBase for phylogenetic comparisons. HierCC was used as a stand-alone programme in development mode to define an initial set of clusters from a representative set of genomes. Subsequent assignments to existing clusters or to novel clusters were performed automatically within EnteroBase in production mode. BlastFrost is used by EnteroBase to determine the presence/absence of toxins and other pathovar characteristics of gastrointestinal pathogens within *E. coli*, and to assign gastrointestinal pathovar designations.

Table 3. Parameters of datasets used for calculating super-trees.

Genus	# Type strains	Total # Genomes	# Strict core genes	# Accessory genes	# Soft core genes (cgMLST)	# SNPs
Salmonella	7	10,002	1409	19,486	3002	1,410,331
Streptococcus	84	5937	90	31,630	372	263,080
Clostridioides	1	6725	515	10,975	2556	725,240
Vibrio	119	5032	124	146,041	1128	776,439
Yersinia	23	1847	744	18,799	1553	656,143

Table 4. Quantitative concordance between eBG and HierCC clustering

Salmonella			
Test statistic	eBG-HC900	eBG-HC2000	HC900-HC2000
Adjusted Mutual Information Score (AMI)	0.985	0.948	0.942
Adjusted Rand Index (ARI)	0.982	0.856	0.866
Escherichia/Shigella			
Test statistic	ST Cplx-HC1100	ST Cplx-HC2000	HC1100-HC2000
Adjusted Mutual Information Score (AMI)	0.940	0.688	0.692
Adjusted Rand Index (ARI)	0.871	0.299	0.337
Streptococcus pneumoniae			
Test statistic	CC-HC100	GSPC-HC100	GSPC-HC160
Adjusted Mutual Information Score (AMI)	0.959	0.956	0.980
Adjusted Rand Index (ARI)	0.987	0.906	0.949

NOTE: *Salmonella*: Calculations were performed on 319,490 entries in EnteroBase which had been assigned to eBGs, HC900 and HC2000 clusters by December, 2021. New eBGs have not been created in recent years, and the dataset only represented 411 eBGs, 690 HC900 clusters and 312 HC2000 clusters.

Escherichia/Shigella: HC1100 and HC2000 assignments were from 143,520 genomes which had been assigned to ST Complexes as well as to HC1100 and HC2350 clusters in December, 2021. At that time 186,852 genomes had been assigned to HC1100 and HC2000 clusters, with AMI of 0.59 and ARI of 0.2.

Streptococcus pneumoniae consisted of GSPC assignments for 18147 genomes from Supplemental Table 2 and CC assignments for 13396 genomes from Supplemental Table 1 to Gladstone et al. [90] which were imported into User defined fields in EnteroBase as described [28].

ARI and AMI were calculated with the functions adjusted_rand_score() and adjusted_mutual_info_score() in the Python 3 library sklearn.metrics V1.0.1 [102].

Figure Legends

Figure 1. Species and sub-species assignments within *Salmonella* with HierCC and ANI. ML super-tree of 1,410,331 SNPs among 3002 core genes from 10,002 representative genomes of *Salmonella*. Former sub-species IIIa is designated *S. arizonae* in accordance with Pearce *et al.*, 2021 [48]. A. Partitions coloured by HC2850 clusters (legend). Arrows indicate a new species, *HC2850_215890* (five strains from the U.K., 2018-2020), and a new sub-species, HC2850_222931 (one strain from France, 2018). All other HC2850 clusters correspond to species (*S. bongori* and *S. arizonae*) or subspecies, except for HC2850_7171 (starred), which is subsp. *enterica* (I) according to the ML tree. B. Partitions differentiated by ANI 95% clusters (legend) correspond to species (*S. enterica*, *S. bongori*, *S. arizonae*, *S. HC2850_215890*), as indicated by arrows, and sub-species are not differentiated. An interactive version of this GrapeTree rendition can be found at https://enterobase.warwick.ac.uk/ms_tree?tree_id=53257. The corresponding presence/absence tree can be found at https://enterobase.warwick.ac.uk/ms_tree?tree_id=53258.

Figure 2. Species and sub-species assignments within *Clostridioides* according to ANI (A) and HierCC (B). ML super-tree of 725,240 SNPs among 2556 core genes from 6,724 representative genomes of *Clostridioides difficile* and one genome of *Clostridioides mangenotii*. A. ANI 95% clusters (legend) differentiate *C. mangenotii* (cluster 1010) and five other clusters (clusters 255, 1370, 373, 2147, 1011) from *C. difficile* (cluster 0). As indicated, four of these correspond to cryptic clades C-I [clusters 255, 373], C-II [cluster 1370] and C-III [cluster 2147]) according to Knight *et al.* [49]. Arrows indicate two additional clusters that were distinguished by HierCC

in part B. B. Partitions coloured by HierCC assigns the same genomes as ANI to the corresponding HC2500 clusters, except that it defines HC2500_15334 and HC2500_15408 to one genome each, which might represent sub-species. An interactive version of this GrapeTree rendition can be found at https://enterobase.warwick.ac.uk/ms_tree?tree_id=53253 and a presence/absence tree can be found at https://enterobase.warwick.ac.uk/ms_tree?tree_id=53254.

Figure 3. Comparison of HC363 clusters with taxonomic designations in *Streptococcus*. ML super-tree of 263,080 SNPs among 372 core genes from 5937 representative *Streptococcus* genomes. Species names are indicated next to the phylogenetic clusters where genomes mapped for type strains, and according to published metadata. Nodes are colored by HC363 clusters, and exceptional assignments are indicated by asterisks next to *S. pneumoniae* and *S. pseudopneumoniae*, which were both HC363_99; multiple phylogenetic and HierCC clusters grouped within *S. suis*; *S. salivarius* and *S. vestibularis*, which were both HC353_202; *S. lutetiensis* and *S. equinus*, which were both HC363_181; and *S. dysgalactiae* and *S. pyogenes*, which were both HC363_139. An interactive version of the GrapeTree rendition of the SNP tree can be found at https://enterobase.warwick.ac.uk/ms_tree?tree_id=53261 and the presence/absence tree at https://enterobase.warwick.ac.uk/ms_tree?tree_id=53262.

Figure 4. Hierarchical population structure of O serogroups in *Salmonella*.

Hierarchical bubble plot of species/subspecies (HC2850; light grey circles), Lineages (HC2000; dark grey circles) and eBurst groups (HC900; other O:group specific colours) of 274,824 *Salmonella* genomes in 790 HC900 clusters for which a consensus O serogroup could be deduced by metadata, or bioinformatic analyses with SeqSero V2 [67] or SISTR 1.1.1 [68]. Additional information is indicated by yellow

text for selected HC2000 and HC900 circles which are specifically mentioned in the Discussion. The diameters of HC900 circles are proportional to the numbers of genomes. An interactive version of this figure can be found at https://observablehq.com/@laurabaxter/salmonella-serovar-piechart from which the representation, raw data and d3 Java code [103] for generating the plot can be downloaded.

Figure 5. Hierarchical population structure of O serogroups in *Escherichia coli/Shigella*. Hierarchical bubble plot for 167,312 genomes of *Escherichia coli* or *Shigella* in HC2350_1 (large grey circle) that were available in EnteroBase in April, 2021. Seven HC2000 Lineages encompassing 15 HC1100/ST Complexes of *Shigella* are arbitrarily indicated at the right. HC2350_1 also includes 15 other *E. coli* HC2000 Lineages that encompass 144 other HC1100 clusters that each contains at least 50 *E. coli* genomes. The figure shows only the HC1100 clusters and not the corresponding HC1100 clusters. Numbers of genomes assigned to individual O serogroups (legend) are indicated by pie chart wedges within the HC1100 circles. Selected HC1100 clusters are drawn on larger scale with indications of phenotype and nomenclature outside the main circle, connected to the original circles by lines. An interactive version of this figure can be found at https://observablehq.com/@laurabaxter/escherichia-serovar-piechart, from which the representation, raw data and d3 Java code [103] for generating the plot can be downloaded.

Supplemental Material

Supplemental Text:

Correspondence between species designations and ANI or HierCC

Yersinia

Vibrio

Historical reconstructions

Sample bias

Methods

Collections of representative genome sequences

ML super-trees

ASTRID guide tree

Frequencies of O groups in *E. coli* and *Salmonella*

Supplemental Table S1. Consistency and inconsistency between HC1090 and ANI95% in *Vibrio*.

Supplemental Table S2. Instances within *Vibrio* of one HC1090 cluster with multiple 95% ANI clusters.

Supplemental Table S3. Instances within *Vibrio* of one 95% ANI cluster vs multiple HC1090 clusters.

Supplemental Table S5. Consistent clustering in *Streptococcus* for 102 clusters according to HC363 and ANI95%.

Supplemental Table S6. One HC363 cluster vs multiple ANI95% clusters in *Streptococcus*.

Supplemental Table S7. One 95% ANI cluster vs two HC363 clusters in *Streptococcus*.

Supplemental Table S8. Lookup table for *Salmonella* from HC2000 through HC900 to Serovar.

Supplemental Table S9. Detailed analysis of O groups in 790 HC900 clusters of *Salmonella* with at least 5 entries.

Supplemental Table S10 (Supplemental Text). Lookup table from Serovar name to antigenic epitopes and O:groups (Institut Pasteur, Paris, with permission from F.X. Weill).

Supplemental Figure S1. Species and sub-species assignments to 1847 representative genomes of *Yersinia* with 95% ANI (A,C) and HC1490 (B,D). Maximum Likelihood ML presence/absence trees of 18,799 accessory genes (A,B) or 656,143 SNPs within 1553 soft core cgMLST genes (C,D). The most common color codes are indicated in legends next to each plot. Within each tree, curated species names in EnteroBase are indicated next to the appropriate clusters. Note that *Y. enterocolitica* is split into four species by HC1490 cluster as confirmed by 95% ANI clustering. Similarly, asterisks in A and C indicate sub-species *similis*, *wautersii* and *pestis* of the *Y. pseudotuberculosis* Complex that were previously assigned to the invalid former species *Y. similis*, *Y. wautersii* and *Y. pestis.* The novel *Yersinia* species *HC1490_1007*, *HC1490_440*, *HC1490_435*, and *HC1490_449* are also included in the figure. Further details can be found in Supplemental Text and interactive versions of the GrapeTree renditions can be found at

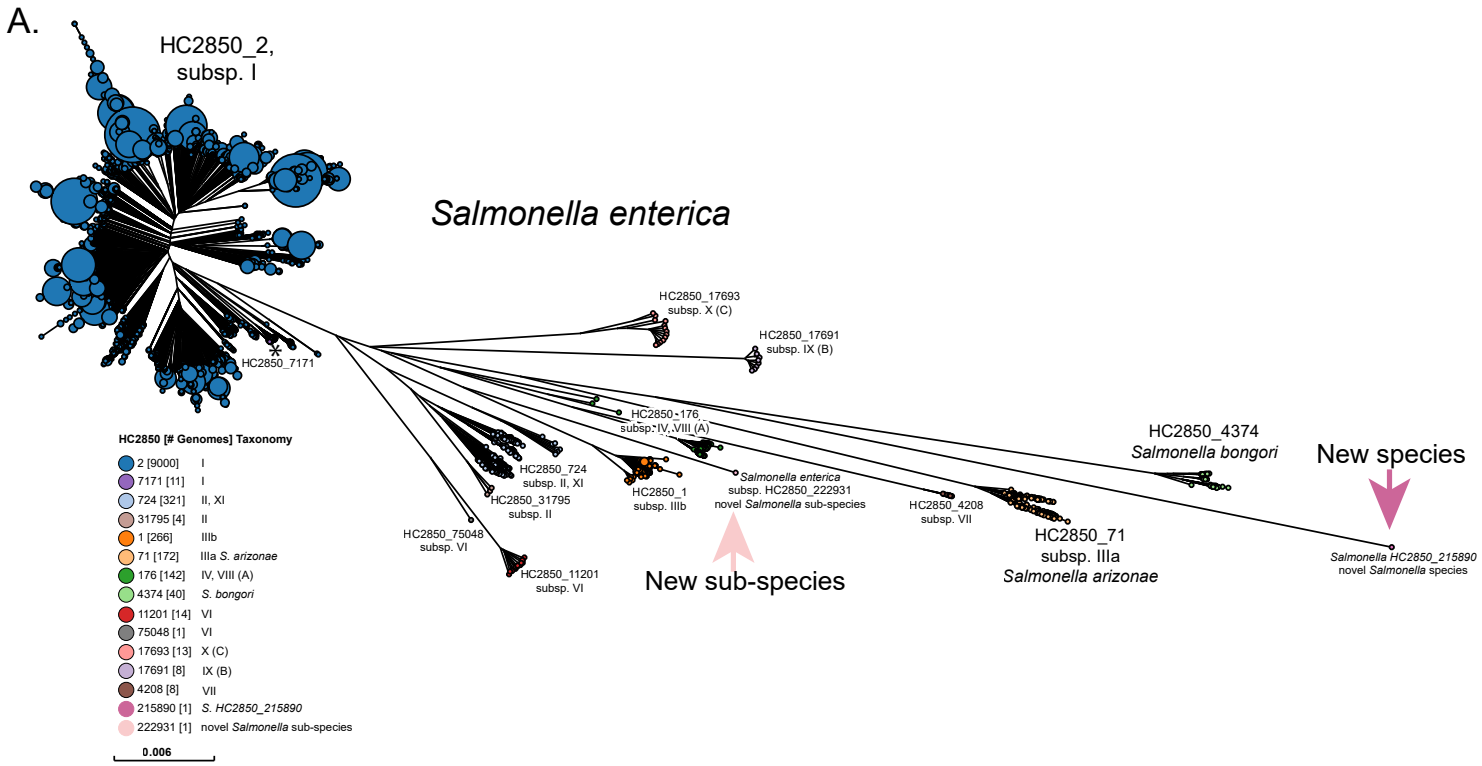
https://enterobase.warwick.ac.uk/ms_tree?tree_id=53270 (A,B) and https://enterobase.warwick.ac.uk/ms_tree?tree_id=53269 (C,D).

Supplemental Figure S2. Curated species, clade, HC1090 and ANI95% assignments to 5032 representative genomes of *Vibrio*. ML SNP super-tree of 776, 439 SNPs among 1128 soft core genes, colour-coded by HC1090 cluster The most common colour codes are indicated in the legend at the bottom. The scope of three higher order groups of species, Cholerae, Harveyii and Splendidus, that were defined by MLSA by Gomez-Gil *et al.*, 2014 are indicated by arcs at the periphery of the plot. Further details can be found in Supplemental Text and interactive version of the GrapeTree renditions can be found at https://enterobase.warwick.ac.uk/ms_tree?tree_id=53265 (SNP tree) and https://enterobase.warwick.ac.uk/ms_tree?tree_id=53266 (presence/absence tree).

Supplemental Figure S3. Comparison of HC363 clusters with taxonomic designations in *Streptococcus* in an ML super-tree of 31,630 accessory genes from 5937 representative *Streptococcus* genomes. Species names are indicated next to the phylogenetic clusters where genomes mapped together with type strains, and according to published metadata. Nodes are colored by HC363 clusters, and exceptional assignments are indicated by asterisks next to *S. mitis*, which could not be distinguished by HC363 clustering from *S. pneumoniae* or *S. pseudopneumoniae*, multiple phylogenetic and HierCC clusters within *S. suis*; *S. salivarius* and *S. vestibularis*, which were both HC353_202; *S. lutetiensis* and *S. equinus*, which were both HC363_181; and *S. dysgalactiae* and *S. pyogenes*, which were both HC363_139. An interactive version of this GrapeTree rendition can be found at https://enterobase.warwick.ac.uk/ms_tree?tree_id=53262.

Supplemental Figure S4. GrapeTree NJ plot of cgMLST allelic distances between 18,147 genomes of *Streptococcus pneumoniae* which had been assigned to GSPC clusters [90]. An interactive version of this figure may be examined at https://enterobase.warwick.ac.uk/ms_tree?tree_id=67948 and color-coded by HC100 cluster level, legacy Clonal Complex (CC), GPSC designation of HC100 cluster number or by other metadata stored for those genomes within EnteroBase. Key Legend: HC160 cluster numbers and their color-codes. Legend: Measure of 40 allelic distances.

Supplemental Figure S5. Cartoon of the workflow of cgMLSA (https://github.com/zheminzhou/cgMLSA). (A) Construct gene trees for each core gene with RaXML (B) Build an initial guide super-tree with ASTRID and calculate local branch supports with ASTRAL. (C) Split the initial guide tree into pairs of disjoint subsets at each highly supported branch, and append an artificial branch (dotted black line) to both members of the subset to allow later reconstruction of this branch. (D) Split all gene trees into the same sets of genomes for each pair of disjoint subsets. (E) Calculate a constraint super-tree for each subset with ASTRAL. (F) Construct the final super-tree by merging constraint trees via the connections marked in C.



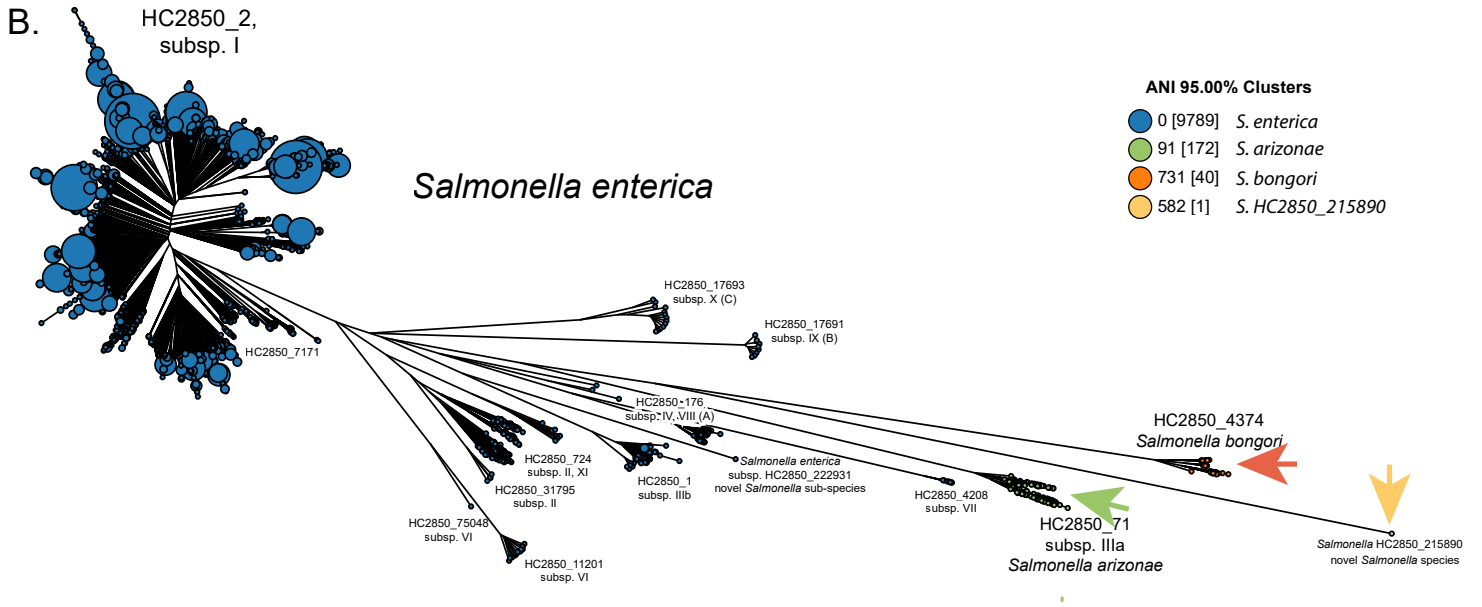


Figure 1. Species and sub-species assignments within *Salmonella* with HierCC and ANI. ML super-tree of 1,410,331 SNPs among 3002 core genes from 10,002 representative genomes of *Salmonella*. Former sub-species IIIa is designated *S. arizonae* in accordance with Pearce et al., 2021 [48]. A. Partitions coloured by HC2850 clusters (legend). Arrows indicate a new species, HC2850_215890 (five strains from the U.K., 2018-2020), and a new sub-species, HC2850_222931 (one strain from France, 2018). All other HC2850 clusters correspond to species (*S. bongori* and *S. arizonae*) or subspecies, except for HC2850_7171 (starred), which is subsp. *enterica* (I) according to the ML tree. B. Partitions differentiated by ANI 95% clusters (legend) correspond to species (*S. enterica, S. bongori, S. arizonae, S.* HC2850_215890), as indicated by arrows, and sub-species are not differentiated. An interactive version of this Grape-Tree rendition can be found at https://enterobase.warwick.ac.uk/ms_tree?tree_id=53257. The corresponding presence/absence tree can be found at https://enterobase.warwick.ac.uk/ms_tree?tree_id=53258.

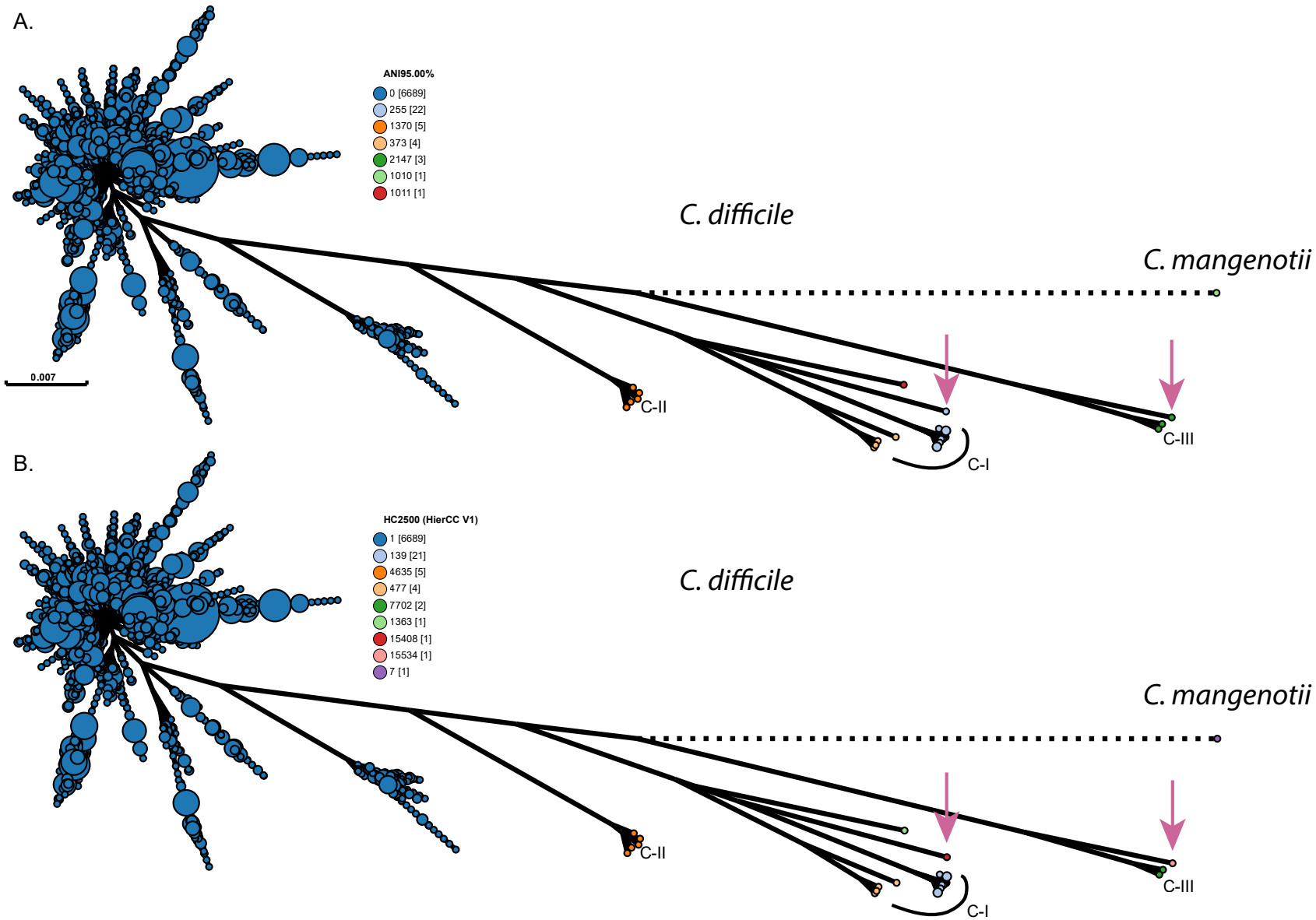


Figure 2. Species and sub-species assignments within *Clostridioides* according to ANI (A) and HierCC (B). ML super-tree of 725,240 SNPs among 2556 core genes from 6,724 representative genomes of *Clostridioides difficile* and one genome of *Clostridioides mangenotii*. A. ANI 95% clusters (legend) differentiate *C. mangenotii* (cluster 1010) and five other clusters (clusters 255, 1370, 373, 2147, 1011) from *C. difficile* (cluster 0). As indicated, four of these correspond to cryptic clades C-I [clusters 255, 373], C-II [cluster 1370] and C-III [cluster 2147]) according to Knight et al. [49]. Arrows indicate two additional clusters that were distinguished by HierCC in part B. B. Partitions coloured by HierCC assigns the same genomes as ANI to the corresponding HC2500 clusters, except that it defines HC2500_15334 and HC2500_15408 to one genome each, which might represent sub-species. An interactive version of this GrapeTree rendition can be found at https://enterobase.warwick.ac.uk/ms_tree?tree_id=53253 and a presence/absence tree can be found at https://enterobase.warwick.ac.uk/ms_tree?tree_id=53254.

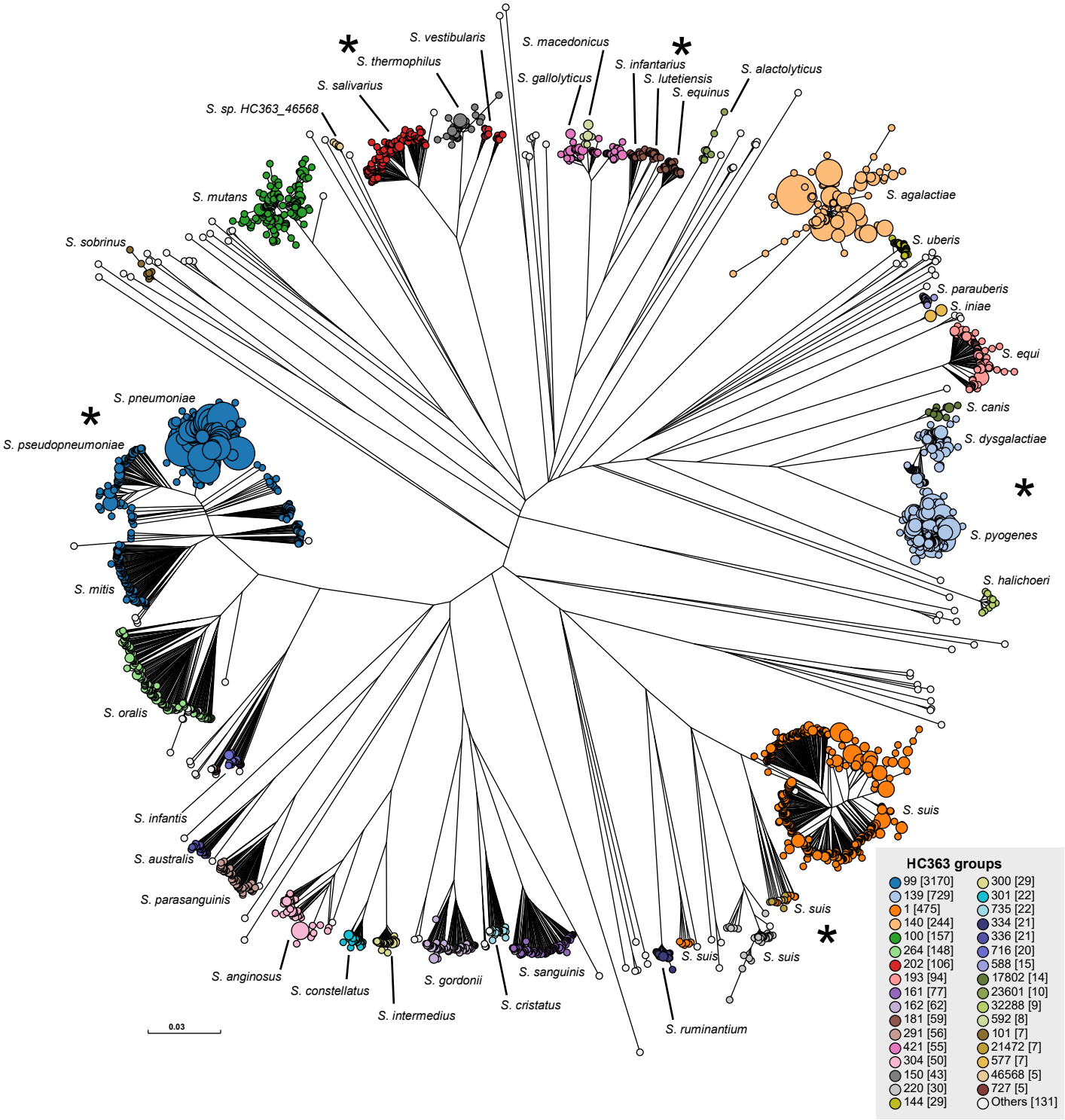
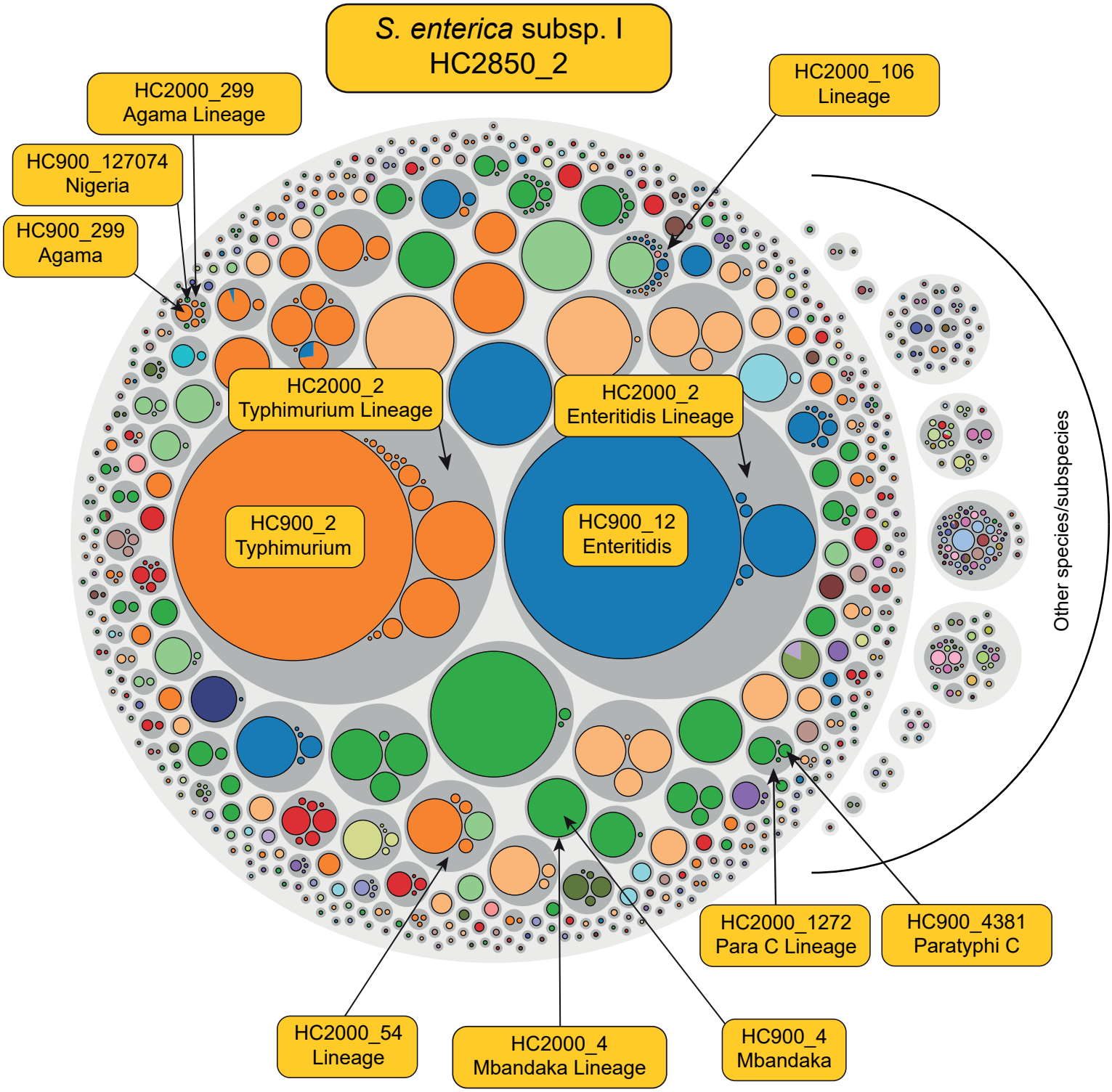


Figure 3. Comparison of HC363 clusters with taxonomic designations in *Streptococcus*. ML super-tree of 263,080 SNPs among 372 core genes from 5937 representative *Streptococcus* genomes. Species names are indicated next to the phylogenetic clusters where genomes mapped for type strains, and according to published metadata. Nodes are colored by HC363 clusters, and exceptional assignments are indicated by asterisks next to *S. pneumoniae* and *S. pseudopneumoniae,* which were both HC363_99; multiple phylogenetic and HierCC clusters grouped within *S. suis*; *S. salivarius* and *S. vestibularis*, which were both HC353_202; *S. lutetiensis and S. equinus*, which were both HC363_181; and *S. dysgalactiae and S. pyogenes*, which were both HC363_139. An interactive version of the GrapeTree rendition of the SNP tree can be found at https://enterobase.warwick.ac.uk/ms_tree?tree_id=53261 and the presence/absence tree at https://enterobase.warwick.ac.uk/ms_tree?tree_id=53262.



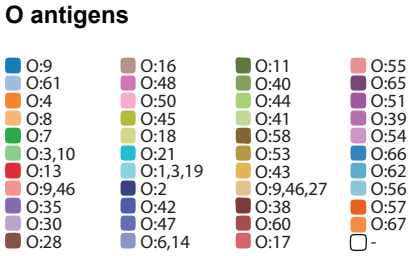


Figure 4. Hierarchical population structure of O serogroups in *Salmonella*. Hierarchical bubble plot of species/subspecies (HC2850; light grey circles), Lineages (HC2000; dark grey circles) and eBurst groups (HC900; other O:group specific colours) of 274,824 *Salmonella* genomes in 790 HC900 clusters for which a consensus O serogroup could be deduced by metadata, or bioinformatic analyses with SeqSero V2 [67] or SISTR 1.1.1 [68]. Additional information is indicated by yellow text for selected HC2000 and HC900 circles which are specifically mentioned in the Discussion. The diameters of HC900 circles are proportional to the numbers of genomes. An interactive version of this figure can be found at https://observablehq.com/@laurabaxter/salmonella-serovar-piechart from which the representation, raw data and d3 Java code [103] for generating the plot can be downloaded.

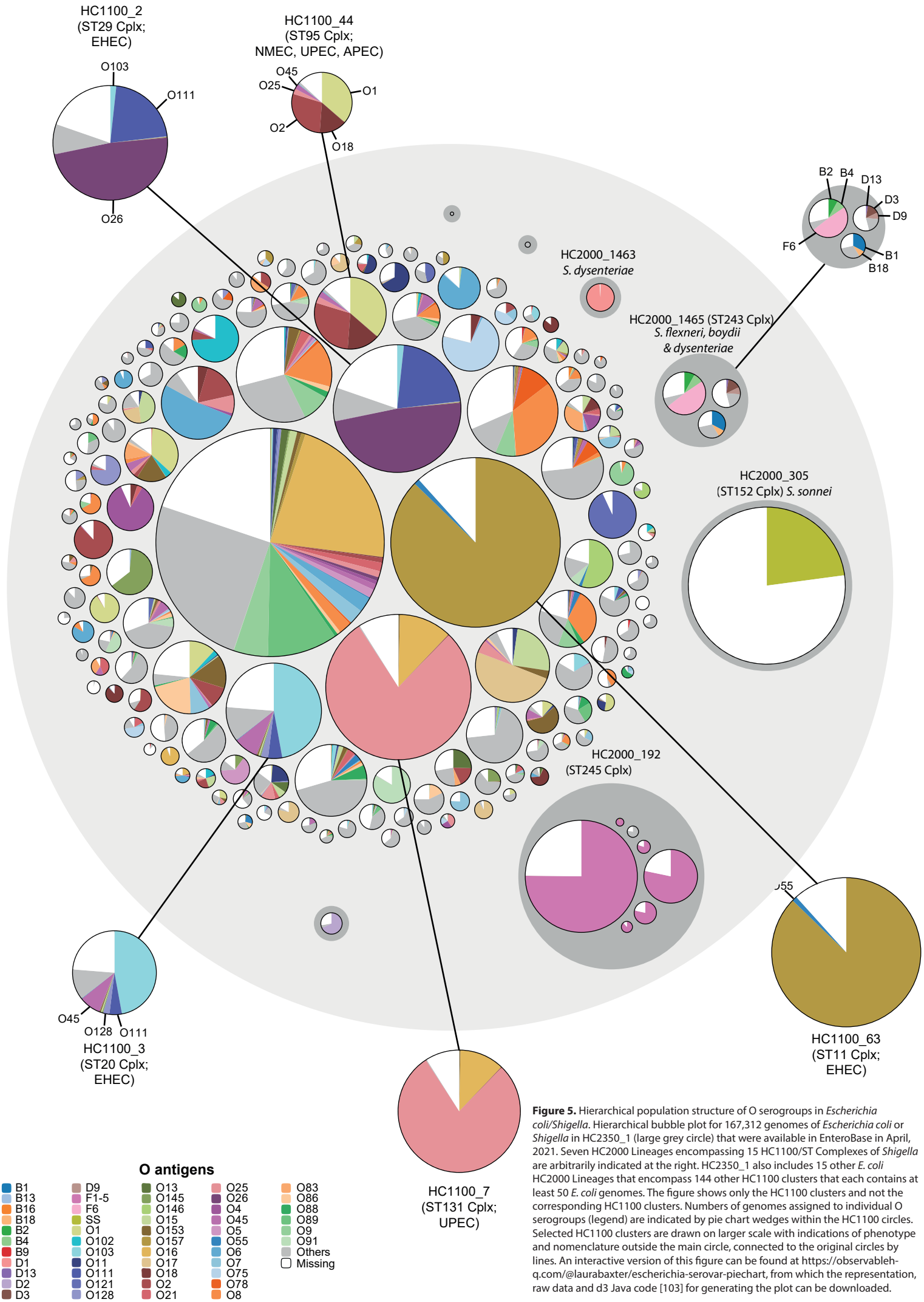


Figure 5. Hierarchical population structure of O serogroups in *Escherichia coli/Shigella*. Hierarchical bubble plot for 167,312 genomes of *Escherichia coli* or *Shigella* in HC2350_1 (large grey circle) that were available in EnteroBase in April, 2021. Seven HC2000 Lineages encompassing 15 HC1100/ST Complexes of *Shigella* are arbitrarily indicated at the right. HC2350_1 also includes 15 other *E. coli* HC2000 Lineages that encompass 144 other HC1100 clusters that each contains at least 50 *E. coli* genomes. The figure shows only the HC1100 clusters and not the corresponding HC1100 clusters. Numbers of genomes assigned to individual O serogroups (legend) are indicated by pie chart wedges within the HC1100 circles. Selected HC1100 clusters are drawn on larger scale with indications of phenotype and nomenclature outside the main circle, connected to the original circles by lines. An interactive version of this figure can be found at https://observableh-q.com/@laurabaxter/escherichia-serovar-piechart, from which the representation, raw data and d3 Java code [103] for generating the plot can be downloaded.