The user's guide to comparative genomics with EnteroBase, including case studies on transmissions of micro-clades of *Salmonella*, the phylogeny of ancient and modern *Yersinia pestis* genomes, and the core genomic diversity of all *Escherichia.*

Authors: Zhemin Zhou[1¶], Nabil-Fareed Alikhan[1¶], Khaled Mohamed[1], Yulei Fan[1], the Agama Study Group[§], and Mark Achtman[1*]

Affiliations:

[1]Warwick Medical School, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, United Kingdom

[¶]Co-equal first author. [*]Corresponding author: M.A.: m.achtman@warwick.ac.uk

[§]The co-authors included in the Agama Study Group consist of: Derek Brown (Scottish Salmonella Reference Laboratory, Glasgow, UK); Marie Chattaway and Tim Dallman (PHE - Public Health England, Colindale, UK); Richard Delahay (National Wildlife Management Centre, APHA, Sand Hutton, York, UK); Christian Kornschober and Ariane Pietzka (AGES - Austrian Agency for Health and Food Safety, Institute for Medical Microbiology and Hygiene Graz, Austria); Burkhard Malorny (German Federal Institute for Risk Assessement, Berlin, Germany [Study Centre for Genome Sequencing and Analysis]); Liljana Petrovska and Rob Davies, (APHA - Animal and Plant Health Agency, Addlestone, UK); Andy Robertson (Environment & Sustainability Institute, University of Exeter, Penryn, UK); William Tyne (Warwick Medical School, University of Warwick, Coventry, UK); François-Xavier Weill and Marie Accou-Demartin (Institut Pasteur, Paris, France); Nicola Williams (Department of Epidemiology and Population Health, Institute of Infection and Global Health, University of Liverpool).

Abstract

EnteroBase is an integrated software environment which supports the identification of global population structures within several bacterial genera that include pathogens. Here we provide an overview on how EnteroBase works, what it can do, and its future prospects. EnteroBase has currently assembled more than 300,000 genomes from Illumina short reads from *Salmonella*, *Escherichia*, *Yersinia*, *Clostridiodes, Helicobacter, Vibrio*, and *Moraxella*, and genotyped those assemblies by core genome Multilocus Sequence Typing (cgMLST). Hierarchical clustering of cgMLST sequence types allows mapping, a new bacterial strain to predefined population structures at multiple levels of resolution within a few hours after uploading its short reads. Case study 1 illustrates this process for local transmissions of *Salmonella enterica* serovar Agama between neighboring social groups of badgers and humans. EnteroBase also supports SNP calls from both genomic assemblies and after extraction from metagenomic sequences, as illustrated by case study 2 which summarizes the microevolution of *Yersinia pestis* over the last 5,000 years of pandemic plague. EnteroBase can also provide a global overview of the genomic diversity within an entire genus, as illustrated by case study 3 which presents a novel, global overview of the population structure of all of the species, subspecies and clades within *Escherichia*.

Introduction

Epidemiological transmission chains of *Salmonella, Escherichia* or *Yersinia* have been reconstructed with the help of single nucleotide polymorphisms (SNPs) from hundreds or even thousands of core genomes (Zhou *et al.* 2013; Zhou *et al.* 2014; Langridge *et al.* 2015; Connor *et al.* 2016; Dallman *et al.* 2016; Wong *et al.* 2016; Ashton *et al.* 2017; Waldram *et al.* 2017; Worley *et al.* 2018; Alikhan *et al.* 2018; Zhou *et al.* 2018c; Johnson *et al.* 2019). However, the scale of these studies pales in comparison to the numbers of publicly available archives (SRAs) of short read sequences of bacterial pathogens which have been deposited since the recent drop in price of high throughput sequencing

(Wetterstrand 2019). In October 2019, the SRA at NCBI contained genomic sequence reads from 430,417 *Salmonella*, *Escherichia/Shigella*, *Clostridiodes*, *Vibrio* and *Yersinia*. However, until very recently (Sanaa *et al.* 2019), only relatively few draft genomic assemblies were publicly available, and even the current comparative genomic analyses in GenomeTrakr (https://www.ncbi.nlm.nih.gov/pathogens/) are restricted to relatively closely related genetic clusters. Since 2014, EnteroBase (https://enterobase.warwick.ac.uk) has attempted to address this gap for selected genera that include bacterial pathogens (Table 1). EnteroBase provides an integrated software platform (Fig. 1) that can be used by microbiologists with limited bioinformatic skills to upload short reads, assemble and genotype genomes, and immediately investigate their genomic relationships to all natural populations within those genera. These aspects have been illustrated by recent publications providing overviews of the population structures of *Salmonella* (Alikhan et al. 2018) and *Clostridioides* (Frentrup *et al*. 2019), a description of the GrapeTree GUI (Zhou *et al.* 2018a) and a reconstruction of the genomic history of the *S. enterica* Para C Lineage (Zhou *et al.* 2018c). However, EnteroBase also provides multiple additional features, which have hitherto largely been promulgated by word of mouth. Here we provide a high-level overview of the functionality of EnteroBase, followed by exemplary case studies of *Salmonella enterica* serovar Agama, *Yersinia pestis* and all of *Escherichia*.

Results

Overview of EnteroBase

The Enterobase back-end consists of multiple, cascading automated pipelines (Supplemental Fig. S1) which implement the multiple functions that it provides (Supplemental Fig. S2A). Many of these EnteroBase pipelines are also available within EToKi (EnteroBase ToolKit), a publicly available repository (https://github.com/zheminzhou/EToKi) of useful modules (Fig. S2B-D) that facilitate genomic assemblies, MLST typing, calling non-repetitive SNPs against a reference genome, or predicting serotypes of *E. coli* from genome assemblies (EB*Eis*).

EnteroBase performs daily scans of the GenBank SRA via its Entrez APIs (Clark et al. 2016) for novel Illumina short read sequences for each of the bacterial genera that it supports. It uploads the new reads, and assembles them (EBAssembly, Fig. S2B) into annotated draft genomes, which are published if they pass quality control (Supplemental Table S1). EnteroBase fetches the metadata associated with the records, and attempts to transcribe it automatically into Enterobase metadata format (Supplemental Tables S2, S3). During the conversion, geographic metadata are translated into structured format using the Nominatim engine offered by OpenStreetMap (OpenStreetMap contributors 2017) and the host/source metadata are assigned to pre-defined categories using a pre-trained Native Bayesian classifier implemented in the NLTK Natural Language Toolkit for Python (Bird *et al.* 2009) (Supplemental Material, Supplemental Fig. S3; estimated accuracy of 60%). Registered users can upload their own Illumina short reads and metadata into EnteroBase; these are then processed with the same pipelines.

The annotated genomes are used to call alleles for Multilocus Sequence Typing (MLST) (MLSType; Fig. S2B) and their Sequence Types (STs) are assigned to population groupings as described below. *Salmonella* serovars are predicted from the legacy MLST eBurstGroups (eBGs), which are strongly associated with individual serovars (Achtman *et al.* 2012), or by two external programs (SISTR1 (Yoshida *et al.* 2016; Robertson *et al.* 2018); SeqSero2 (Zhang *et al.* 2019)) which evaluate genomic sequences. *Escherichia* serotypes are predicted from the genome assemblies by the EnteroBase module EB*Eis.* Clermont haplogroups are predicted for *Escherichia* by two external programs (ClermonTyping (Beghain *et al.* 2018); EZClermont (Waters *et al.* 2018)) and *fimH* type by a third (FimTyper (Roer *et al.* 2017)). By default, public free access to strain metadata and the genome assemblies, predicted genotypes and predicted phenotypes is immediate, but a delay in the release date of up to 12 months can be imposed when uploading short read sequences.

In September 2019, EnteroBase provided access to 364,690 genomes and their associated metadata and predictions (Table 1). In order to allow comparisons

with historical data, EnteroBase also maintains additional legacy 7-gene MLST assignments (and metadata) that were obtained by classical Sanger sequencing from 18,478 strains,

Ownership, permanence, access and privacy. EnteroBase users can upload new entries, consisting of paired-end Illumina short reads plus their metadata. Short reads are deleted after genome assembly, or after automated, brokered uploading of the reads and metadata to the European Nucleotide Archive upon user request.

The search and graphical tools within EnteroBase include all assembled genomes and their metadata, even if they are pre-release. However, ownership of uploaded data remains with the user, and extends to all calculations performed by EnteroBase. Only owners and their buddies, admins or curators can edit the metadata. And only those individuals can download any data or calculations prior to their release date. In order to facilitate downloading of post-release data by the general community, downloads containing metadata and genotypes or genomic assemblies are automatically stripped of pre-release data for users who lack ownership privileges. Similarly, pre-release nodes within trees in the GrapeTree and Dendrogram graphical modules must be hidden before users without ownership privileges can download those trees.

In general, metadata that were imported from an SRA are not editable, except by admins and curators. However, the admins can assign editing rights to users with claims to ownership or who possess special insights.

MLST Population structures. Each unique sequence variant of a gene in an MLST scheme is assigned a unique numerical designation. 7-gene MLST STs consist of seven integers for the alleles of seven housekeeping gene fragments (Maiden *et al.* 1998). rSTs consist of 51-53 integers for ribosomal protein gene alleles (Jolley *et al.* 2012). cgMLST STs consist of 1,553 – 3,002 integers for the number of genes in the soft core genome for that genus (Table 1), which were chosen as described elsewhere (Frentrup *et al.* 2019). However, STs are

arbitrary constructs, and natural populations can each encompass multiple, related ST variants. Therefore, 7-gene STs are grouped into ST Complexes in *Escherichia*/*Shigella* (Wirth et al. 2006) by an eBurst approach (Feil *et al.* 2004), and into their equivalent eBurst groups (eBGs) in *Salmonella* (Achtman et al. 2012). EnteroBase implements similar population groups (reBGs) for rMLST in *Salmonella*, which are largely consistent with eBGs or their sub-populations (Alikhan *et al.* 2018). The EnteroBase Nomenclature Server (Fig. S1) calculates these population assignments automatically for each novel ST on the basis of single linkage clustering chains with maximal pairwise differences of one allele for 7-gene MLST and two alleles for rMLST. In order to prevent overlaps between ST Complexes, growing chains are terminated when they extend too closely to other existing populations (2 alleles difference in 7-gene MLST and 5 in rMLST).

cgMLST has introduced additional complexities over MLST and rMLST. Visual comparisons of cgSTs are tedious, and rarely productive, because each consists of up to 3,002 integers. Furthermore, almost all cgSTs contain some missing data because they are called from draft genomes consisting of multiple contigs. EnteroBase contains 100,000s of cgST numbers because almost every genome results in a unique cgST number, even though many cgSTs only differ from others by missing data. EnteroBase supports working with so many cgSTs through HierCC (Hierarchical Clustering), a novel approach which supports analyses of population structures based on cgMLST at multiple levels of resolution. In order to identify the cut-off values in stepwise cgMLST allelic distances which would reliably resolve natural populations, we first calculated a matrix of pair-wise allelic distances (excluding pairwise missing data) for all existing pairs of cgSTs, and one matrix for the HierCC clusters at each level of allelic distance, i.e. one matrix for HC0, HC1, HC2…HC3,001. A genus-specific subset of the most reliable HierCC clusters is reported by EnteroBase.

For *Salmonella*, thirteen HierCC levels are reported, ranging from HC0 (indistinguishable except for missing data) to HC2850 (Fig. 2). Our experience

with *Salmonella* indicates that HC2850 corresponds to subspecies, HC2000 to super-lineages (Zhou *et al.* 2018c) and HC900 to cgMLST versions of eBGs. Long-term endemic persistence seems to be associated with HC100 or HC200; and epidemic outbreaks with HC2, HC5 or HC10. Eleven levels are reported for the other genera, ranging from HC0 up to HC2350 for *Escherichia*, HC2500 for *Clostridioides* and HC1450 for *Yersinia. Escherichia* HC1100 corresponds to ST Complexes (man. In prep.) and the correspondences to population groupings in *Clostridioides* are described elsewhere (Frentrup et al. 2019). Further information on HierCC can be found in the EnteroBase documentation (https://tinyurl.com/HierCC-doc).

Uber- and Sub-strains. Most bacterial isolates/strains in EnteroBase are linked to one set of metadata and one set of genotyping data. However, EnteroBase includes strains for which legacy MLST data from classical Sanger sequencing exists in addition to MLST genotypes from genomic assemblies. Similarly, some users have uploaded the same reads to both EnteroBase and SRAs, and both sets of data are present in EnteroBase. In other cases, genomes of the same strain have been sequenced by independent laboratories, or multiple laboratory variants have been sequenced that are essentially indistinguishable (e.g. *S. enterica* LT2 or *E. coli* K-12).

EnteroBase deals with such duplications by implementing the concept of an Uberstrain, which can be a parent to one or more identical sub-strains. Sub-strains remain invisible unless they are specified in the search dialog (Supplemental Fig. S4), in which case they are shown with a triangle in the Uberstrain column (Fig. 3A). Examples of the usage of this approach can be found in Supplemental Material.

Examples of the utility of EnteroBase.

Often the utility of a tool first becomes clear through examples of its use. Here we present three case studies that exemplify different aspects of EnteroBase. Case study 1 demonstrates how geographically separated laboratories can collaborate in private on an EnteroBase project until its completion, upon which EnteroBase

publishes the results. This example focuses on geographical micro-variation and transmission chains between various host species of a rare serovar of *S. enterica*. Case study 2 demonstrates how to combine modern genomes of *Yersinia pestis* with partially reconstructed genomes from ancient skeletons of plague victims. It also demonstrates how EToKI can extract SNPs from metagenomic sequence reads. Case study 3 provides a quantitative overview of the genomic diversity of an entire genus, thereby defining the EcoRPlus set of representative genomes of all *Escherichia*.

Case Study 1: A group collaboration on *S. enterica* serovar Agama

S. enterica subsp. *enterica* encompasses more than 1,586 defined serovars (Guibourdenche *et al.* 2010; Issenhuth-Jeanjean *et al.* 2014). These differ in the antigenic formulas of their lipopolysaccharide (O antigen) and/or two alternative flagellar antigens (H1, H2), which are abbreviated as O:H1:H2. Some serovars are commonly isolated from infections and the environment, and have been extensively studied. Others are rare, poorly understood and often polyphyletic (Achtman *et al.* 2012), including *Salmonella* that colonize badgers (Wray *et al.* 1977; Wilson *et al.* 2003).

In late 2018, serovar Agama (antigenic formula: 4,12:i:1,6) was specified in the Serovar metadata field for only 134/156,347 (0.09%) genome assemblies in EnteroBase, and all 134 isolates were from humans. We were therefore interested to learn that the University of Liverpool possessed serovar Agama isolates that had been isolated in 2006-2007 from European badgers (*Meles meles*) in Woodchester Park, Gloucestershire, England. We sequenced the genomes of 72 such isolates, and uploaded the short reads and strain metadata into EnteroBase. This data was used to analyze the population structure of a rare serovar within a single host species over a limited geographical area, and to compare Agama genomes from multiple hosts and geographical sources.

Search Strains. The browser interface to EnteroBase is implemented as a spreadsheet-like window called a “Workspace” that can page through 1,000s of entries, showing metadata at the left and experimental data at the right

(https://tinyurl.com/EnteroBase-WS). However visual scanning of 1,000s of entries is inefficient. EnteroBase therefore offers powerful search functions (https://tinyurl.com/EnteroBase-search) for identifying isolates that share common phenotypes (metadata) and/or genotypes (experimental data).

EnteroBase also predicts serovars from assembled *Salmonella* genomes and from MLST data. However, the software predictions are not fail-proof, and many entries lack metadata information, or the metadata is erroneous. We therefore used the Search Strains dialog box to find entries containing “Agama” in the metadata field or by the predictions from SISTR1. Phylogenetic analyses of the cgMLST data from those entries indicated that Agama consisted of multiple micro-clusters.

International participation in a collaborative network. Almost all Agama isolates in EnteroBase were from England, which represents a highly skewed geographical sampling bias that might lead to phylogenetic distortions. We therefore formed the Agama Study Group, consisting of colleagues at national microbiological reference laboratories in England, Scotland, Ireland, France, Germany and Austria. The participants were declared as ‘buddies’ within EnteroBase (https://tinyurl.com/EnteroBase-buddies) with explicit rights to access the Workspaces and phylogenetic trees in the Workspace\Load\Shared\Zhemin\Agama folder. After completion of this manuscript, that folder was made publicly available.

We facilitated the analysis of the Agama data by creating a new user-defined Custom View (https://tinyurl.com/EnteroBase-customview), which can aggregate various sources of experimental data as well as User-defined Fields. The Custom View was saved in the Agama folder, and thereby shared with the Study Group. It too was initially private but became public together with the other workspaces and trees when the folder was made public.

Members of the Agama Study Group were requested to sequence genomes from all Agama strains in their collections, and to upload those short reads to

EnteroBase, or to send their DNAs to University of Warwick for sequencing and uploading. The new entries were added to the 'All Agama Strains' workspace. The final set of 345 isolates had been isolated in Europe, Africa and Australia, with collection years ranging from 1956 to 2018 (Supplemental Table S3).

Global population Structure of Agama. We created a neighbor joining GrapeTree (Zhou *et al.* 2018a) of cgMLST data to reveal the genetic relationships within serovar Agama. Color coding the nodes of the tree by SISTR1 serovar predictions confirmed that most isolates were Agama (Fig. 3A). However, one micro-cluster (shaded in light orange) consisted of seven monophasic Agama isolates with a defective or partial *fljB* (H2) CDS, which prevented a serovar prediction. SISTR1 also could not predict the O antigens of three other related isolates (arrows in Fig. 3). Sixteen other isolates on long branches were assigned to other serovars by SISTR1 (Fig. 3A, grey shading). Comparable results were obtained with SeqSero2 or eBG serovar associations, and these sixteen isolates represent erroneous Serovar assignments within the metadata. Interestingly, three of these erroneous Agama had the same predicted antigenic formula (1,4,[5],12:i:-) as the monophasic Agama isolates (orange shading), but these represent monophasic Typhimurium.

In contrast to serovar, coloring the tree nodes by HC2000 clusters (Fig. 3B) immediately revealed that all genomes that were called Agama by SISTR1 belonged to HC2000 cluster number 299 (HC2000_299), and all HC2000_299 were genetically related and clustered together in the tree (Fig. 3B). In contrast, the 16 other isolates on long branches (gray shading) belonged to other HC2000 clusters.

These results show that Agama belongs to one super-lineage, HC2000_299, which has been isolated globally from humans, badgers, companion animals and the environment since at least 1956. The genetic relationships would not have been obvious with lower resolution MLST: some Agama isolates belong to

eBG167, others to eBG336 and thirteen Agama MLST STs do not belong to any eBG.

Transmission patterns at different levels of HierCC resolution. All isolates from badgers were in HierCC cluster HC400_299 (Fig. 3B, dashed box), which also included other isolates from humans and other animals. HC400_299 was investigated by Maximum-Likelihood trees of core, non-repetitive SNPs called against a reference draft genome with the help of the EnteroBase Dendrogram GUI. One tree encompassed 149 isolates from the British Isles which were in EnteroBase prior to establishing the Agama Study group. A second tree (Fig. 4B) contained the final data set of 213 genomes, including isolates from additional badgers and multiple countries. A comparison of the two trees is highly instructive on the effects of sample bias.

Almost all of the initial HC400_299 genomes fell into three micro-clades designated HC100_299, HC100_2433 and HC100_67355. All badger isolates were from Woodchester Park (2006-2007) within the context of a long-term live capture-mark-recapture study (McDonald *et al.* 2018). The Agama isolates from those badgers formed a monophyletic clade within HC100_2433, whose basal nodes represented human isolates. This branch topology suggested that a single recent common ancestor of all badger isolates which had been transmitted from humans or their waste products.

The badgers in Woodchester Park occupy adjacent social group territories which each contain several setts (burrows). HC100_2433 contains multiple HC10 clusters of Agama from badgers (Supplemental Fig. S5A). To investigate whether these micro-clusters might mark transmission chains between setts and social groups, a Newick sub-tree of HC100_2433 plus geographical co-ordinates was transmitted from GrapeTree to MicroReact (Argimon *et al.* 2016), an external program which is specialized in depicting geographical associations. Badgers occasionally move between neighboring social groups (Rogers *et al.* 1998). Transmissions associated with such moves are supported by the

observation that five distinct HC10 clusters each contained isolates from two social groups in close proximity (Fig. S5B).

Long-term dispersals and inter-host transmissions. The 63 additional HC400_299 Agama genomes that were sequenced by the Agama Study Group provided important insights on the dissemination of Agama over a longer time frame, and demonstrated the dramatic effects of sample bias. Seventeen Agama strains had been isolated from English badgers at multiple locations in south-west England between 1998 and 2016 (Fig. S5B), and stored at APHA. Eleven of them were in HC100_2433. However, rather than being interspersed among the initial genomes from badgers, they defined novel micro-clusters, including HC10_171137 and HC10_171148, which were the most basal clades in HC100_2433 (Fig. 4B). The other six badger isolates from additional geographical sources were interspersed among human isolates in HC100_299 (Fig. 4B), which had previously not included any badger isolates (Fig. S5F). These results show that the diversity of Agama from English badgers is comparable to their diversity within English humans, and that it would be difficult to reliably infer the original host of these clades or the directionality of inter-host transmissions. Further observations on micro-epidemiology of Agama transmissions between hosts and countries are presented in Supplemental Material.

Case Study 2: Combining modern *Y. pestis* genomes with ancient metagenomes.

EnteroBase automatically scours sequence read archives for Illumina short reads from cultivated isolates, assembles their genomes and publishes draft assemblies that pass quality control. In October 2019, EnteroBase had assembled >1,300 genomes of *Y. pestis*, including genomes that had already been assigned to population groups (Cui *et al.* 2013), other recently sequenced genomes from central Asia (Eroshenko *et al.* 2017; Kutyrev *et al.* 2018) and numerous unpublished genomes from Madagascar and Brazil. EnteroBase does not upload assembled genomes, for which adequate, automated quality control

measures would be difficult to implement. However, EnteroBase administrators can upload such genomes after *ad hoc* assessment of sequence quality, and EnteroBase contains standard complete genomes such as CO92 (Parkhill *et al.* 2001) and other genomes used to derive the *Y. pestis* phylogeny (Morelli *et al.* 2010).

Enterobase also does not automatically assemble genomes from metagenomes containing mixed reads from multiple taxa, but similar to complete genomes, administrators can upload reconstructed ancient genomes derived from SNP calls against a reference genome.

Ancient *Y. pestis*. The number of publications describing ancient *Y. pestis* genomes has increased dramatically over the last few years as ancient plague has been progressively deciphered (Bos *et al.* 2011; Wagner *et al.* 2014; Rasmussen *et al.* 2015; Bos *et al.* 2016; Feldman *et al.* 2016; Spyrou *et al.* 2016; Spyrou *et al.* 2018; Margaryan *et al.* 2018; Namouchi *et al.* 2018; Keller *et al.* 2019; Spyrou *et al.* 2019). The metagenomic short reads used to reconstruct these genomes are routinely deposited in the public domain but the reconstructed ancient genomes are not. This practice has made it difficult for non-bioinformaticians to evaluate the relationships between ancient and modern genomes from *Y. pestis*. However, EnteroBase now provides a solution to this problem.

The EnteroBase EToKi calculation package can reconstruct an ancient genome assembly by unmasking individual nucleotides in a fully masked reference genome based on reliable SNP calls from metagenomic data (Supplemental Fig. S6). We ran EToKi on 56 published ancient metagenomes containing *Y. pestis* and the resulting assemblies and metadata were uploaded to EnteroBase. EnteroBase users can now include those ancient genomes together with other reconstructed genomes and modern genomic assemblies in a workspace of their choice, and use the EnteroBase SNP dendrogram module to calculate and

visualize a Maximum Likelihood tree (of up to a current maximum of 200 genomes).

Fig. 5 presents a detailed overview of the genomic relationships of all known *Y. pestis* populations from pandemic plague over the last 5,500 years, including 100s of unpublished modern genomes. This tree was manually annotated using a User-defined Field and Custom View with population designations from the literature on modern isolates to include reconstructed ancient genomes. These population designations have now been updated for additional modern genomes from central Asia and elsewhere. An interactive version of this tree and all related metadata in EnteroBase is publicly available (https://tinyurl.com/YpestisSNP), thus enabling its detailed interrogation by a broad audience from multiple disciplines (Green 2018), and providing a common language for scientific discourse.

Case Study 3: Thinking big – an overview of the core genomic diversity of *Escherichia/Shigella.*

Escherichia coli has long been one of the primary work-horses of molecular biology. Most studies of *Escherichia* have concentrated on a few well-characterized strains of *E. coli*, but the genus *Escherichia* includes other species: *E. fergusonii, E. albertii, E. marmotae* (Liu *et al.* 2015) and *E. ruysiae* (van der Putten *et al.* 2019). *E. coli* itself includes the genus *Shigella* (Pupo *et al.* 2000), which was assigned a distinctive genus name because it causes dysentery. Initial analyses of the phylogenetic structure of *E. coli* identified multiple deep branches, called haplogroups (Selander *et al.* 1987), and defined the EcoR collection (Ochman and Selander 1984), a classical group of 72 bacterial strains that represented the genetic diversity found with multilocus enzyme electrophoresis. The later isolation of environmental isolates from lakes revealed the existence of “cryptic clades” I-VI which were distinct from the main *E. coli* haplogroups and the other *Escherichia* species (Walk et al. 2009; Luo et al. 2011). Currently, bacterial isolates are routinely assigned to haplogroups or clades by PCR tests for the presence of variably present genes from the

accessory genome (Clermont *et al.* 2013) or by programs that identify the presence of those genes in genomic sequences (Beghain *et al.* 2018; Waters *et al.* 2018).

An alternative scheme for subdividing *Escherichia* was introduced in 2006, legacy MLST which includes the assignment of STs to ST Complexes (Wirth *et al.* 2006). Several ST Complexes are common causes of invasive disease in humans and animals, such as ST131 (Stoesser *et al.* 2016; Liu *et al.* 2018), ST95 Complex (Wirth *et al.* 2006; Gordon *et al.* 2017) and ST11 Complex (O157:H7) (Eppinger *et al.* 2011a; Eppinger *et al.* 2011b; Newell and La Ragione 2018). The large number of *Escherichia* genomes in EnteroBase (Table 1) now provides an opportunity to re-investigate the population structure of *Escherichia* on the basis of the greater resolution provided by cgMLST, and within the context of a much larger and more comprehensive sample. in 2018 EnteroBase contained 52,876 genomes. In order to render this sample amenable to calculating an ML tree of core SNPs, we selected a representative sample consisting of one genome from each of the 9,479 *Escherichia* rSTs. In homage to the EcoR collection, we designate this as the EcoRPlus collection.

Core genome genetic diversity within *Escherichia.* Homologous recombination is widespread within *E. coli* (Wirth *et al.* 2006). We therefore anticipated that a phylogenetic tree of core genomic differences in EcoRPlus would be 'fuzzy', and that ST Complexes and other genetic populations would be only poorly delineated. Instead, considerable core genome population structure is visually apparent in a RapidNJ tree based on pairwise differences at cgMLST alleles between the EcoRPlus genomes (Fig. 6). The most predominant, discrete sets of node clusters were also largely uniform according to cgMLST HC1100 hierarchical clustering. Furthermore, with occasional exceptions, assignments to HC1100 clustering were also largely congruent with ST Complexes based on legacy 7-gene MLST (Supplemental Fig. S7) and with Clermont typing (Supplemental Fig. S8; Supplemental material).

Fig. 6 may represent the first detailed overview of the entire genetic diversity of the core genome of *Escherichia*. Real time examination of its features (http://tinyurl.com/ECOR-RNJ) is feasible because the GrapeTree algorithm can handle large numbers of cgSTs (Zhou *et al.* 2018a). Nodes can be readily colored by metadata or experimental data (Supplemental Figs. S7-S9), and GrapeTree also readily supports analyses of sub-trees in greater detail. However, although cgMLST allelic distances are reliable indicators of population structures, SNPs are preferable for examining genetic distances. We therefore calculated a Maximum-Likelihood (ML) tree of the 1,230,995 core SNPs within all 9,479 genomes (Supplemental Fig. S9). This tree confirmed the clustering of the members of HC1100 groups within *E. coli,* and also showed that the other *Escherichia* species and cryptic clades II to VIII formed distinct long branches of comparable lengths (Fig. S9 inset).

Discussion

EnteroBase was originally developed as a genome-based successor to the legacy MLST websites for *Escherichia* (Wirth et al. 2006), *Salmonella* (Achtman et al. 2012), *Yersinia pseudotuberculosis* (Laukkanen-Ninios et al. 2011) and *Moraxella catarrhalis* (Wirth et al. 2007). Its underlying infrastructure is sufficiently generic that EnteroBase was readily extended to *Clostridioides, Helicobacter* and *Vibrio*, and could in principle be extended to other taxa.

EnteroBase was intended to provide a uniform and reliable pipeline that can assemble consistent draft genomes from the numerous short read sequences in public databases (Achtman and Zhou 2014), and to link those assemblies with metadata and genotype predictions. It was designed to provide access to an unprecedentedly large global set of draft genomes to users at both extremes of the spectrum of informatics skills. A further goal was to provide analytical tools, such as GrapeTree (Zhou *et al.* 2018a), that could adequately deal with cgMLST from >100,000 genomes, and Dendrogram, which generates phylograms from non-repetitive core SNPs called against a reference genome. Still another important goal was to support private analyses by groups of colleagues, with the

option of subsequently making those analyses publicly available. Case Study 1 illustrates how EnteroBase can be used for all of these tasks and more.

EnteroBase has expanded beyond its original goals, and is morphing in novel directions. It has implemented HierCC for cgMLST, which supports the automated recognition of population structures at multiple levels of resolution (Case Study 1), and may help with the annotation of clusters within phylogenetic trees (Case Study 2; see below). EnteroBase has also been extended to support analyses of metagenomic data from ancient genomes (Zhou *et al.* 2018c) by implementing a subset of the functionality of SPARSE (Zhou *et al.* 2018b) within the stand-alone EToKi package. Case Study 2 illustrates this capability for *Y. pestis.* Additional EnteroBase databases are under development for ancient and modern genomes of *S. enterica* and biofilms within dental calculus. EnteroBase has also demonstrated its capacities for providing overviews of the core genome diversity of entire genera, with currently extant examples consisting of *Salmonella* (Alikhan et al. 2018) and *Escherichia* (Case Study 3).

EnteroBase is already being used by the community to identify genetically related groups of isolates (Johnson *et al.* 2019; Haley *et al.* 2019; Numberger *et al.* 2019; Diemert and Yan 2019), and HierCC has been used to mark international outbreaks of *S. enterica* serovar Poona (Jones *et al.* 2019b) and *E. coli* O26 (Jones *et al.* 2019a). Case Study 1 illustrates how to explore HierCC genomic relationships at multiple levels, ranging from HC2000 (super-lineages) for inter-continental dispersion down to HC5-10 for detecting local transmission chains.

Case Study 1 confirms that although *S. enterica* serovar Agama is rare, it has been isolated from multiple hosts and countries, and is clearly not harmless for humans. The results also document that an enormous sample bias exists in current genomic databases because they largely represent isolates that are relevant to human disease from a limited number of geographic locations.

Case Study 1 may also become a paradigm for identifying long-distance chains of transmission between humans or between humans and their companion or domesticated animals: Four Agama isolates in the HC5_140035 cluster from France (human) and Austria (frozen chives and a human blood culture) differed by no more than 5 of the 3,002 cgMLST loci. These isolates also differed by no more than 5 non-repetitive core SNPs. We anticipate that large numbers of such previously silent transmission chains will be revealed as EnteroBase is used more extensively.

Case study 2 illustrates how EnteroBase can facilitate combining reconstructed genomes from metagenomic sequences with draft genomes from cultured strains. In this case, the metagenomes were from ancient tooth pulp which had been enriched for *Y. pestis*, and the bacterial isolates were modern *Y. pestis* from a variety of global sources since 1898. The resulting phylogenetic tree (Fig. 5) presents a unique overview of the core genomic diversity over 5,000 years of evolution and pandemic spread of plague, which can now be evaluated and used by a broad audience. This tree will be updated at regular intervals as additional genomes or metagenomes become available.

The manual population designations in Fig. 5 are largely reflected by HC10 clusters. However, it is uncertain whether the current HierCC clusters would be stable with time because they were based on only 1,300 *Y. pestis* genomes. EnteroBase will therefore maintain manual annotations in parallel with automated HierCC assignments until a future date when a qualified choice is possible.

Case study 3 defines the EcoRPlus Collection of 9,479 genomes which represents the genetic diversity of 52,876 genomes. It is a worthy successor of EcoR (Ochman and Selander 1984), which contained 72 representatives of 2,600 *E. coli* strains that had been tested by multilocus enzyme electrophoresis in the early 1980s. The genomic assemblies and known metadata of EcoRPlus

are publicly available (http://tinyurl.com/ECOR-Plus), and can serve as a reference set of genomes for future analyses with other methods.

Visual examination of an NJ tree of cgMLST allelic diversity color-coded by HierCC HC1100 immediately revealed several discrete *E. coli* populations that have each been the topics of multiple publications (Fig. 6). These included a primary cause of hemolytic uremic syndrome (O157:H7), a common cause of invasive disease in the elderly (the ST131 Complex), as well as multiple distinct clusters of *Shigella* that cause dysentery. However, it also contains multiple other discrete clusters of *E. coli* that are apparently also common causes of global disease in humans and animals but which have not yet received comparable attention. The annotation of this tree would therefore be a laudable task for the entire scientific community interested in *Escherichia.* We also note that HierCC is apparently a one stop, complete replacement for haplogroups, Clermont Typing and ST Complexes, some of whose deficiencies are also illustrated here.

This user's guide provides an overview of what EnteroBase can do now. With time, we hope to include additional, currently missing features, such as community annotation of the properties of bacterial populations, predicting antimicrobial resistance/sensitivity, and distributing core pipelines to multiple mirror sites. However, EnteroBase is already able to help a broad community of users with a multitude of tasks for the selected genera it supports. More detailed instructions are available in the online documentation (https://enterobase.readthedocs.io/en/latest/) and questions can be addressed to the support team (enterobase@warwick.ac.uk).

Methods

Isolation of serovar Agama from badgers. Supplemental Fig. S5B provides a geographical overview of the area in Woodchester, Gloucestershire in which badger setts and social groups were investigated in 2006-2007. This area has been subject to a multi-decade investigation of badger mobility and patterns of

infection with *Mycobacterium bovis* (McDonald *et al.* 2018). According to the standard protocol for that study, badgers were subjected to routine capture using steel mesh box traps baited with peanuts, examination under anesthesia and subsequent release. Fecal samples were cultivated at University of Liverpool after selective enrichment (Rappaport-Vassiliadis broth and semi-solid agar), followed by cultivation on MacConkey agar. Lactose-negative colonies that swarmed on Rappaport-Vassiliadis agar but not on nutrient agar, and were catalase-positive and oxidase-negative, were serotyped by slide agglutination tests according to the Kaufmann and White scheme. Additional isolates from badgers from the geographical areas in England that are indicated in Fig. S5D and S5F were collected during routine investigations of animal disease at the APHA.

Laboratory manipulations and genomic sequencing. At University of Warwick, *Salmonella* were cultivated and DNA was purified by automated procedures as described (O'Farrell et al. 2012). Paired-end 150 bp genomic sequencing was performed in multiplexes of 96-192 samples on an Illumina NextSeq 500 using the High Output Kit v2.5 (FC-404-2002) according to the manufacturer's instructions. Other institutions used their own standard procedures. Metadata and features of all 344 genomes in Fig. 4 are publicly available in EnteroBase in the workspace 'Zhou et al. All Agama strains' (https://tinyurl.com/AgamaWS).

Integration of ancient *Yersinia pestis* genomes in EnteroBase. Metagenomic reads from ancient samples may contain a mixture of sequence reads from the species of interest as well as from genetically similar taxa that represent environmental contamination. In order to deal with this issue and remove such non-specific reads after extraction with the EToKi prepare module, the EToKi assemble module can be used to align the extracted reads after comparisons with an ingroup of genomes related to the species of interest and with an outgroup of genomes from other species. In the case of Fig. 5, the ingroup consisted of *Y. pestis* genomes CO92 (2001), Pestoides F, KIM10+ and 91001 and the outgroup consisted of

genomes *Y. pseudotuberculosis* IP32953 and IP31758, *Y. similis* 228 and *Y. enterocolitica* 8081. Reads were excluded which had higher alignment scores to the outgroup genomes than to the ingroup genomes. Prior to mapping reads to the *Y. pestis* reference genome (CO92 (2001)), a pseudo-genome was created in which all nucleotides were masked in order to ensure that only nucleotides supported by metagenomic reads would be used for phylogenetic analysis. For the 13 ancient genomes whose publications included complete SNP lists, we unmasked the sites in the pseudo-genomes which were included in the published SNP lists. For the other 43 genomes, the filtered metagenomic reads were mapped onto the pseudo-genome with minimap2 (Li 2018), and evaluated with Pilon (Walker *et al.* 2014), and sites in the pseudo-genome were unmasked which were covered by ≥3 reads and had a consensus base that was supported by ≥80% of the mapped reads. All 56 pseudo-genomes were stored in EnteroBase together with their associated metadata.

Data Access

The Illumina sequence reads for 161 new genomes of *S. enterica* serovar Agama generated in this study have been submitted to the European Nucleotide Archive database (ENA; https://www.ebi.ac.uk/ena) under study accession numbers ERP114376, ERP114456, ERP114871 and ERP115055. The genomic properties, metadata and accession codes for the 329 genomic assemblies in HC2000_299 are summarized in Supplemental Table S3 and in Online Table 1 (https://wrap.warwick.ac.uk/128112). The metadata, genomic assemblies and annotations are also available from the publicly available workspace “Zhou et al. All Agama Strains” (https://tinyurl.com/AgamaWS). The EToKi package and its documentation are accessible at https://github.com/zheminzhou/EToKi. EnteroBase documentation is accessible at https://enterobase.readthedocs.io/en/latest/. An interactive version of Figure 3 is available at https://tinyurl.com/AgamaFig3. Trees presented in Fig. 4 are available separately at (A) https://tinyurl.com/AgamaFig4A and (B) https://tinyurl.com/AgamaF4B. An interactive version of Fig. 5 is available at https://tinyurl.com/YpestisSNP. The MicroReact projects of Figure S5 are available at (A,B) https://microreact.org/project/t7qlSSslh/3e634888; (C,D) https://microreact.org/project/9XUC7i-Fm/fed65ff5 and (E,F)

https://microreact.org/project/XaJm1cNjY/69748fe3. The tree shown in Figure 6 as well as Supplemental Figs. S7-S8 are available at http://tinyurl.com/ECOR-RNJ;

Acknowledgements

EnteroBase development was funded by the BBSRC (BB/L020319/1) and the Wellcome Trust (202792/Z/16/Z). We gratefully acknowledge sharing of strains and data by Niall Delappe and Martin Cormican, Salmonella Reference Laboratory, Galway, Ireland, and critical comments on the text by Nina Luhmann and Jane Charlesworth.

Author Contributions.

MA wrote the manuscript with the help of all other authors. ZZ, N-FA, KM and YF were responsible for the development of EnteroBase under the guidance of MA. N-FA and KM were responsible for the online manual. Analyses were performed and figures were drawn by ZZ and N-FA under the guidance of MA. The Agama Study Group provided information, bacterial strains, DNAs and genomic sequences from Agama isolates from all over Europe, and was involved in writing the manuscript and evaluating the conclusions.

730 Table 1. Basic statistics on EnteroBase (https://enterobase.warwick.ac.uk) (19.09.2019)

Genus	Legacy MLST	Assembled genomes (user uploads)	wgMLST (Loci)	cgMLST (Loci)	rMLST (Loci)	MLST (Loci)	HierCC (cgMLST)
Salmonella	6,480	225,026 (30,636)	21,065	3,002	51	7	√
Escherichia/ Shigella	10,155	110,302 (12,584)	25,002	2,512	51	7	√
Clostrioides		14,592 (1,422)	11,490	2,556	53	7	√
Vibrio		7,010 (128)			51		
Yersinia	1,054	3,412 (1,066)	19,531	1,553	51	7	√
Helicobacter		2,458 (846)			53		
Moraxella	789	1,890 (349)			52	8	
Total	18,478	364,690 (47,031)					

731 NOTE: The numbers of assemblies refers to the number of Uberstrain/substrain sets, and
732 ignores known duplicates. Legacy MLST refers to strain metadata and sequences from ABI
733 sequencing of 7 loci for the genera *Salmonella* (Kidgell *et al.* 2002; Achtman *et al.* 2012),
734 *Escherichia/Shigella* (Wirth *et al.* 2006), *Yersinia* (Laukkanen-Ninios *et al.* 2011; Hall *et al.*
735 2015) and *Moraxella* (Wirth *et al.* 2007) that are maintained at EnteroBase as a legacy of
736 data originally provided at http://MLST.warwick.ac.uk. The 7 gene MLST scheme for
737 *Clostrioides difficile* (Griffiths *et al.* 2010) and all rMLST schemes (Jolley *et al.* 2012) are
738 coordinated on a daily basis with the schemes that are maintained at PubMLST
739 (https://pubmlst.org/).

740 Abbreviations: wgMLST: whole genome MultiLocus Sequence Typing (Maiden *et al.* 2013);
741 cgMLST: core genome MultiLocus Sequence Typing (Mellmann *et al.* 2011); rMLST:
742 ribosomal MultiLocus Sequence Typing (Jolley *et al.* 2012).

743

744 Reference List

745

746 Achtman M. 2016. How old are bacterial pathogens? *Proc Biol Sci* **283**: 1836.

747 Achtman M, Wain J, Weill F-X, Nair S, Zhou Z, Sangal V, Krauland MG, Hale JL, Harbottle H,
748 Uesbeck A, et al. 2012. Multilocus sequence typing as a replacement for serotyping in
749 *Salmonella enterica*. *PLoS Pathog* **8**: e1002776.

750 Achtman M and Zhou Z. 2014. Distinct genealogies for plasmids and chromosome. *PLoS Genet* **10**:
751 e1004874.

752 Alikhan N-F, Zhou Z, Sergeant MJ, Achtman M. 2018. A genomic overview of the population structure
753 of *Salmonella*. *PLoS Genet* **14**: e1007261.

754 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol*
755 *Biol* **215**: 403-410.

756 Argimon S, Abudahab K, Goater RJ, Fedosejev A, Bhai J, Glasner C, Feil EJ, Holden MT, Yeats CA,
757 Grundmann H, et al. 2016. Microreact: visualizing and sharing data for genomic epidemiology
758 and phylogeography. *Microb Genom* **2**: e000093.

759 Ashton PM, Owen S, Kaindama L, Rowe WPM, Lane C, Larkin L, Nair S, Jenkins C, de Pinna E,
760 Feasey N, et al. 2017. *Salmonella enterica* serovar Typhimurium ST313 responsible for
761 gastroenteritis in the UK are genetically distinct from isolates causing bloodstream infections
762 in Africa. *BioRxiv*.

763 Beghain J, Bridier-Nahmias A, Le NH, Denamur E, Clermont O. 2018. ClermonTyping: an easy-to-use
764 and accurate *in silico* method for *Escherichia* genus strain phylotyping. *Microb Genom* **4**.

765 Bird S, Klein E, Loper E. 2009. *Natural Language Processing with Python: Analyzing Text with the*
766 *Natural Language Toolkit*, 1 edition. O'Reilly Media, Sebastopol, CA.

767 Bos KI, Herbig A, Sahl J, Waglechner N, Fourment M, Forrest SA, Klunk J, Schuenemann VJ, Poinar
768 D, Kuch M, et al. 2016. Eighteenth century *Yersinia pestis* genomes reveal the long-term
769 persistence of an historical plague focus. *Elife* **5**.

770 Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N, Coombes BK, McPhee JB,
771 Dewitte SN, Meyer M, Schmedes S, et al. 2011. A draft genome of *Yersinia pestis* from
772 victims of the Black Death. *Nature* **478**: 506-510.

773 Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2016. GenBank. *Nucleic Acids Res* **44**:
774 D67-D72.

775 Clermont O, Christenson JK, Denamur E, Gordon DM. 2013. The Clermont *Escherichia coli* phylo-
776 typing method revisited: improvement of specificity and detection of new phylo-groups.
777 *Environ Microbiol Rep* **5**: 58-65.

778 Connor TR, Owen SV, Langridge G, Connell S, Nair S, Reuter S, Dallman TJ, Corander J, Tabing KC,
779 Le HS, et al. 2016. What's in a name? Species-wide whole-genome sequencing resolves
780 invasive and noninvasive lineages of *Salmonella enterica* serotype Paratyphi B. *MBio* **7**:
781 e00527-16.

782 Cui Y, Yu C, Yan Y, Li D, Li Y, Jombart T, Weinert LA, Wang Z, Guo Z, Xu L, et al. 2013. Historical
783 variations in mutation rate in an epidemic pathogen, *Yersinia pestis*. *Proc Natl Acad Sci USA*
784 **110**: 577-582.

785 Dallman T, Inns T, Jombart T, Ashton P, Loman N, Chatt C, Messelhaeusser U, Rabsch W, Simon S,
786 Nikisins S, et al. 2016. Phylogenetic structure of European *Salmonella* Enteritidis outbreak
787 correlates with national and international egg distribution network. *Microb Genom* **2**: e000070.

788 Damgaard PB, Marchi N, Rasmussen S, Peyrot M, Renaud G, Korneliussen T, Moreno-Mayar JV,
789 Pedersen MW, Goldberg A, Usmanova E, et al. 2018. 137 ancient human genomes from
790 across the Eurasian steppes. *Nature* **557**: 369-374.

791 Diemert S and Yan T. 2019. Clinically unreported salmonellosis outbreak detected via comparative
792 genomic analysis of municipal wastewater *Salmonella* isolates. *Appl Environ Microbiol* **85**: 10.

793 Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**:
794 2460-2461.

795 Eppinger M, Mammel MK, LeClerc JE, Ravel J, Cebula TA. 2011a. Genome signatures of *Escherichia*
796 *coli* O157:H7 from the bovine host reservoir. *Appl Environ Microbiol* **77**: 2916-2925.

797 Eppinger M, Mammel MK, LeClerc JE, Ravel J, Cebula TA. 2011b. Genomic anatomy of *Escherichia*
798 *coli* O157:H7 outbreaks. *Proc Natl Acad Sci USA* **108**: 20142-20147.

799 Eroshenko GA, Nosov NY, Krasnov YM, Oglodin YG, Kukleva LM, Guseva NP, Kuznetsov AA,
800 Abdikarimov ST, Dzhaparova AK, Kutyrev VV. 2017. *Yersinia pestis* strains of ancient
801 phylogenetic branch 0.ANT are widely spread in the high-mountain plague foci of Kyrgyzstan.
802 *PLoS ONE* **12**: e0187230.

803 Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG. 2004. eBURST: Inferring patterns of
804 evolutionary descent among clusters of related bacterial genotypes from Multilocus Sequence
805 Typing data. *J Bacteriol* **186**: 1518-1530.

806 Feldman M, Harbeck M, Keller M, Spyrou MA, Rott A, Trautmann B, Scholz HC, Paffgen B, Peters J,
807 McCormick M, et al. 2016. A high-coverage *Yersinia pestis* genome from a sixth-century
808 Justinianic plague victim. *Mol Biol Evol* **33**: 2911-2923.

809 Frentrup M, Zhou Z, Steglich M, Meier-Kolthoff JP, Göker M, Riedel T, Bunk B, Spröer C, Overmann
810 J, Blaschitz M, et al. 2019. Global genomic population structure of *Clostridioides difficile*.
811 *BioRxiv* 727230.

812 Gordon DM, Geyik S, Clermont O, O'Brien CL, Huang S, Abayasekara C, Rajesh A, Kennedy K,
813 Collignon P, Pavli P, et al. 2017. Fine-scale structure analysis shows epidemic patterns of
814 Clonal Complex 95, a cosmopolitan *Escherichia coli* lineage responsible for extraintestinal
815 infection. *mSphere* **2**.

816 Green MH. 2018. Putting Africa on the Black Death map: Narratives from genetics and history.
817 *Afriques [Online]* **9**.

818 Griffiths D, Fawley W, Kachrimanidou M, Bowden R, Crook DW, Fung R, Golubchik T, Harding RM,
819 Jeffery KJ, Jolley KA, et al. 2010. Multilocus sequence typing of *Clostridium difficile*. *J Clin*
820 *Microbiol* **48**: 770-778.

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

Haley BJ, Kim SW, Haendiges J, Keller E, Torpey D, Kim A, Crocker K, Myers RA, Van Kessel JAS. 2019. *Salmonella enterica* serovar Kentucky recovered from human clinical cases in Maryland, USA (2011-2015). *Zoonoses Public Health*.

Hall M, Chattaway MA, Reuter S, Savin C, Strauch E, Carniel E, Connor T, Van D, I, Rajakaruna L, Rajendram D, et al. 2015. Use of whole-genus genome sequence data to develop a multilocus sequence typing tool that accurately identifies *Yersinia* isolates to the species and subspecies levels. *J Clin Microbiol* **53**: 35-42.

Issenhuth-Jeanjean S, Roggentin P, Mikoleit M, Guibourdenche M, De PE, Nair S, Fields PI, Weill F-X. 2014. Supplement 2008-2010 (no. 48) to the White-Kauffmann-Le Minor scheme. *Res Microbiol* **165**: 526-530.

Johnson TJ, Elnekave E, Miller EA, Munoz-Aguayo J, Flores FC, Johnston B, Nielson DW, Logue CM, Johnson JR. 2019. Phylogenomic analysis of extraintestinal pathogenic *Escherichia coli* Sequence Type 1193, an emerging multidrug-resistant clonal group. *Antimicrob Agents Chemother* **63**.

Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna H, Harrison OB, Sheppard SK, Cody AJ, et al. 2012. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* **158**: 1005-1015.

Jones G, Lefevre S, Donguy MP, Nisavanh A, Terpant G, Fougere E, Vaissiere E, Guinard A, Mailles A, De VH, et al. 2019a. 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**.

Jones G, Pardos de la Gandaro M, Herrera-Leon L, Herrera-Leon S, Varela Martinez C, Hureaux-Roy R, Abdallah Y, Nisavanh A, Fabre L, Renaudat C, et al. 2019b. 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.

Keller M, Spyrou MA, Scheib CL, Neumann GU, Kropelin A, Haas-Gebhard B, Paffgen B, Haberstroh J, Ribera IL, Raynaud C, et al. 2019. Ancient *Yersinia pestis* genomes from across Western Europe reveal early diversification during the First Pandemic (541-750). *Proc Natl Acad Sci U S A* **116**: 12363-12372.

Kidgell C, Reichard U, Wain J, Linz B, Torpdahl M, Dougan G, Achtman M. 2002. *Salmonella typhi*, the causative agent of typhoid fever, is approximately 50,000 years old. *Infect Genet Evol* **2**: 39-45.

Kutyrev VV, Eroshenko GA, Motin VL, Nosov NY, Krasnov JM, Kukleva LM, Nikiforov KA, Al'khova ZV, Oglodin EG, Guseva NP. 2018. Phylogeny and classification of *Yersinia pestis* through the lens of strains From the plague foci of Commonwealth of Independent States. *Frontiers in Microbiology* **9**: 1106.

Langridge GC, Fookes M, Connor TR, Feltwell T, Feasey N, Parsons BN, Seth-Smith HM, Barquist L, Stedman A, Humphrey T, et al. 2015. Patterns of genome evolution that have accompanied host adaptation in *Salmonella*. *Proc Natl Acad Sci U S A* **112**: 863-868.

Laukkanen-Ninios R, Didelot X, Jolley KA, Morelli G, Sangal V, Kristo P, Brehony C, Imori PF, Fukushima H, Siitonen A, et al. 2011. Population structure of the *Yersinia pseudotuberculosis* complex according to multilocus sequence typing. *Environ Microbiol* **13**: 3114-3127.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-3100.

Liu CM, Stegger M, Aziz M, Johnson TJ, Waits K, Nordstrom L, Gauld L, Weaver B, Rolland D, Statham S, et al. 2018. *Escherichia coli* ST131-H22 as a foodborne uropathogen. *MBio* **9**.

Liu S, Jin D, Lan R, Wang Y, Meng Q, Dai H, Lu S, Hu S, Xu J. 2015. *Escherichia marmotae sp. nov.*, isolated from faeces of *Marmota himalayana*. *Int J Syst Evol Microbiol* **65**: 2130-2134.

Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT. 2011. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proc Natl Acad Sci USA* **108**: 7200-7205.

Maiden MC, van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Microbiol* **11**: 728-736.

Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, 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.

Margaryan A, Hansen HB, Rasmussen S, Sikora M, Moiseyev V, Khoklov A, Epimakhov A, Yepiskoposyan L, Kriiska A, Varul L, et al. 2018. Ancient pathogen DNA in human teeth and petrous bones. *Ecol Evol* **8**: 3534-3542.

McDonald JL, Robertson A, Silk MJ. 2018. Wildlife disease ecology from the individual to the population: Insights from a long-term study of a naturally infected European badger population. *J Anim Ecol* **87**: 101-112.

Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, 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.

Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, Feldkamp M, Kusecek B, Vogler AJ, Li Y, et al. 2010. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nature Genet* **42**: 1140-1143.

Namouchi A, Guellil M, Kersten O, Hänsch S, Ottoni C, Schmid BV, Pacciani E, Quaglia L, Vermunt M, Bauer EL, et al. 2018. Integrative approach using *Yersinia pestis* genomes to revisit the historical landscape of plague during the Medieval Period. *Proc Natl Acad Sci U S A*.

Newell DG and La Ragione RM. 2018. Enterohaemorrhagic and other Shiga toxin-producing *Escherichia coli* (STEC): Where are we now regarding diagnostics and control strategies? *Transbound Emerg Dis* **65 Suppl 1**: 49-71.

Numberger D, Riedel T, McEwen G, Nubel U, Frentrup M, Schober I, Bunk B, Sproer C, Overmann J, Grossart HP, et al. 2019. Genomic analysis of three *Clostridioides difficile* isolates from urban water sources. *Anaerobe* **56**: 22-26.

O'Farrell B, Haase JK, Velayudhan V, Murphy RA, Achtman M. 2012. Transforming microbial genotyping: A robotic pipeline for genotyping bacterial strains. *PLoS ONE* **7**: e48022.

Ochman H and Selander RK. 1984. Standard reference strains of *Escherichia coli* from natural populations. *J Bacteriol* **157**: 690-693.

OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org. 2017.

Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, Prentice MB, Sebaihia M, James KD, Churcher C, Mungall KL, et al. 2001. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413**: 523-527.

Pupo GM, Lan R, Reeves PR. 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.

Rasmussen S, Allentoft ME, Nielsen K, Orlando L, Sikora M, Sjogren KG, Pedersen AG, Schubert M, Van DA, Kapel CM, et al. 2015. Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. *Cell* **163**: 571-582.

Robertson J, Yoshida C, Kruczkiewicz P, Nadon C, Nichani A, Taboada EN, Nash JHE. 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.

Roer L, Tchesnokova V, Allesoe R, Muradova M, Chattopadhyay S, Ahrenfeldt J, Thomsen MCF, Lund O, Hansen F, Hammerum AM, et al. 2017. Development of a web tool for *Escherichia coli* subtyping based on *fimH* alleles. *J Clin Microbiol* **55**: 2538-2543.

Rogers LM, Delahay R, Cheeseman CL, Langton S, Smith GC, Clifton-Hadley RS. 1998. Movement of badgers (*Meles meles*) in a high-density population: individual, population and disease effects. *Proc Biol Sci* **265**: 1269-1276.

Sanaa M, Pouillot R, Vega FG, Strain E, Van Doren JM. 2019. GenomeGraphR: A user-friendly open-source web application for foodborne pathogen whole genome sequencing data integration, analysis, and visualization. *PLoS ONE* **14**: e0213039.

Selander RK, Caugant DA, Whittam TS. 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, H.E. Umbarger), pp. 1625-1648. American Society for Microbiology, Washington, D.C.

Simonsen M, Mailund T, Pedersen CNS. Inference of large phylogenies using Neighbour-Joining. 2011. Biomedical Engineering Systems and Technologies: 3rd International Joint Conference, BIOSTEC 2010. *Communications in Computer and Information Science* , 334-344. Springer Verlag.

Spyrou MA, Keller M, Tukhbatova RI, Scheib CL, Nelson EA, Andrades VA, Neumann GU, Walker D, Alterauge A, Carty N, et al. 2019. Phylogeography of the second plague pandemic revealed through analysis of historical *Yersinia pestis* genomes. *Nat Commun* **10**: 4470.

Spyrou MA, Tukhbatova RI, Feldman M, Drath J, Kacki S, Beltran de HJ, Arnold S, Sitdikov AG, Castex D, Wahl J, et al. 2016. Historical *Y. pestis* genomes reveal the European black death as the source of ancient and modern plague pandemics. *Cell Host Microbe* **19**: 874-881.

Spyrou MA, Tukhbatova RI, Wang CC, Valtuena AA, Lankapalli AK, Kondrashin VV, Tsybin VA, Khokhlov A, Kuhnert D, Herbig A, et al. 2018. Analysis of 3800-year-old *Yersinia pestis* genomes suggests Bronze Age origin for bubonic plague. *Nat Commun* **9**: 2234.

946 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
947 phylogenies. *Bioinformatics* **30**: 1312-1313.

948 Stoesser N, Sheppard AE, Pankhurst L, De MN, Moore CE, Sebra R, Turner P, Anson LW, Kasarskis
949 A, Batty EM, et al. 2016. Evolutionary history of the global emergence of the *Escherichia coli*
950 epidemic clone ST131. *MBio* **7**: e02162.

951 van der Putten BCL, Matamoros S, COMBAT consortium, Schultsz C. 2019. Genomic evidence for
952 revising the *Escherichia* genus and description of *Escherichia ruysiae sp. nov. BioRxiv*
953 781724.

954 Wagner DM, Klunk J, Harbeck M, Devault A, Waglechner N, Sahl JW, Enk J, Birdsell DN, Kuch M,
955 Lumibao C, et al. 2014. *Yersinia pestis* and the plague of Justinian 541-543 AD: a genomic
956 analysis. *Lancet Infect Dis* **14**: 319-326.

957 Waldram A, Dolan G, Ashton P, Jenkins C, Dallman T. 2017. Whole genome sequencing reveals an
958 outbreak of *Salmonella* Enteritidis associated with reptile feeder mice in the United Kingdom,
959 2012-2015. *Food Microbiology*.

960 Walk ST, Alm EW, Gordon DM, Ram JL, Toranzos GA, Tiedje JM, Whittam TS. 2009. Cryptic
961 lineages of the genus *Escherichia*. *Appl Environ Microbiol* **75**: 6534-6544.

962 Walker BJ, Abeel T, Shea T, Priest M, Abouelleil A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J,
963 Young SK, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection
964 and genome assembly improvement. *PLoS ONE* **9**: e112963.

965 Waters NR, Abram F, Brennan F, Holmes A, Pritchard L. 2018. Easily phylotyping *E. coli* via the
966 EzClermont web app and command-line tool. *BioRxiv* 317610.

967 Wetterstrand, KA. 2019. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing
968 Program (GSP). https://www.genome.gov/sequencingcostsdata/.

969 Wilson JS, Hazel SM, Williams NJ, Phiri A, French NP, Hart CA. 2003. Nontyphoidal salmonellae in
970 United Kingdom badgers: Prevalence and spatial distribution. *Appl Environ Microbiol* **69**:
971 4312-4315.

972 Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman
973 H, et al. 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol*
974 *Microbiol* **60**: 1136-1151.

975 Wirth T, Morelli G, Kusecek B, Van Belkum A, van der Schee C, Meyer A, Achtman M. 2007. The rise
976 and spread of a new pathogen: seroresistant *Moraxella catarrhalis*. *Genome Res* **17**: 1647-
977 1656.

978 Wong VK, Baker S, Connor TR, Pickard D, Page AJ, Dave J, Murphy N, Holliman R, Sefton A, Millar
979 M, et al. 2016. An extended genotyping framework for *Salmonella enterica* serovar Typhi, the
980 cause of human typhoid. *Nat Commun* **7**: 12827.

981 Worley J, Meng J, Allard MW, Brown EW, Timme RE. 2018. *Salmonella enterica* phylogeny based on
982 whole-genome sequencing reveals two new clades and novel patterns of horizontally
983 acquired genetic elements. *MBio* **9**.

984 Wray C, Baker K, Gallagher J, Naylor P. 1977. *Salmonella* infection in badgers in the South West of
985 England. *Br Vet J* **133**: 526-529.

Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VP, Nash JH, Taboada EN. 2016. The *Salmonella In Silico* Typing Resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft *Salmonella* genome assemblies. *PLoS One* **11**: e0147101.

Zhang S, Den-Bakker HC, Li S, Chen J, Dinsmore BA, Lane C, Lauer AC, Fields PI, Deng X. 2019. SeqSero2: rapid and improved *Salmonella* serotype determination using whole genome sequencing data. *Appl Environ Microbiol*.

Zhou Z, Alikhan N-F, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, Carrico JA, Achtman M. 2018a. GrapeTree: Visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res* **28**: 1395-1404.

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

Zhou Z, Lundstrøm I, Tran-Dien A, Duchêne S, Alikhan N-F, Sergeant MJ, Langridge G, Fokatis AK, Nair S, Stenøien HK, et al. 2018c. 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.

Zhou Z, McCann A, Litrup E, Murphy R, Cormican M, Fanning S, Brown D, Guttman DS, Brisse S, Achtman M. 2013. Neutral genomic microevolution of a recently emerged pathogen, *Salmonella enterica* serovar Agona. *PLoS Genet* **9**: e1003471.

Zhou Z, McCann A, Weill F-X, Blin C, Nair S, Wain J, Dougan G, Achtman M. 2014. Transient Darwinian selection in *Salmonella enterica* serovar Paratyphi A during 450 years of global spread of enteric fever. *Proc Natl Acad Sci U S A* **111**: 12199-12204.

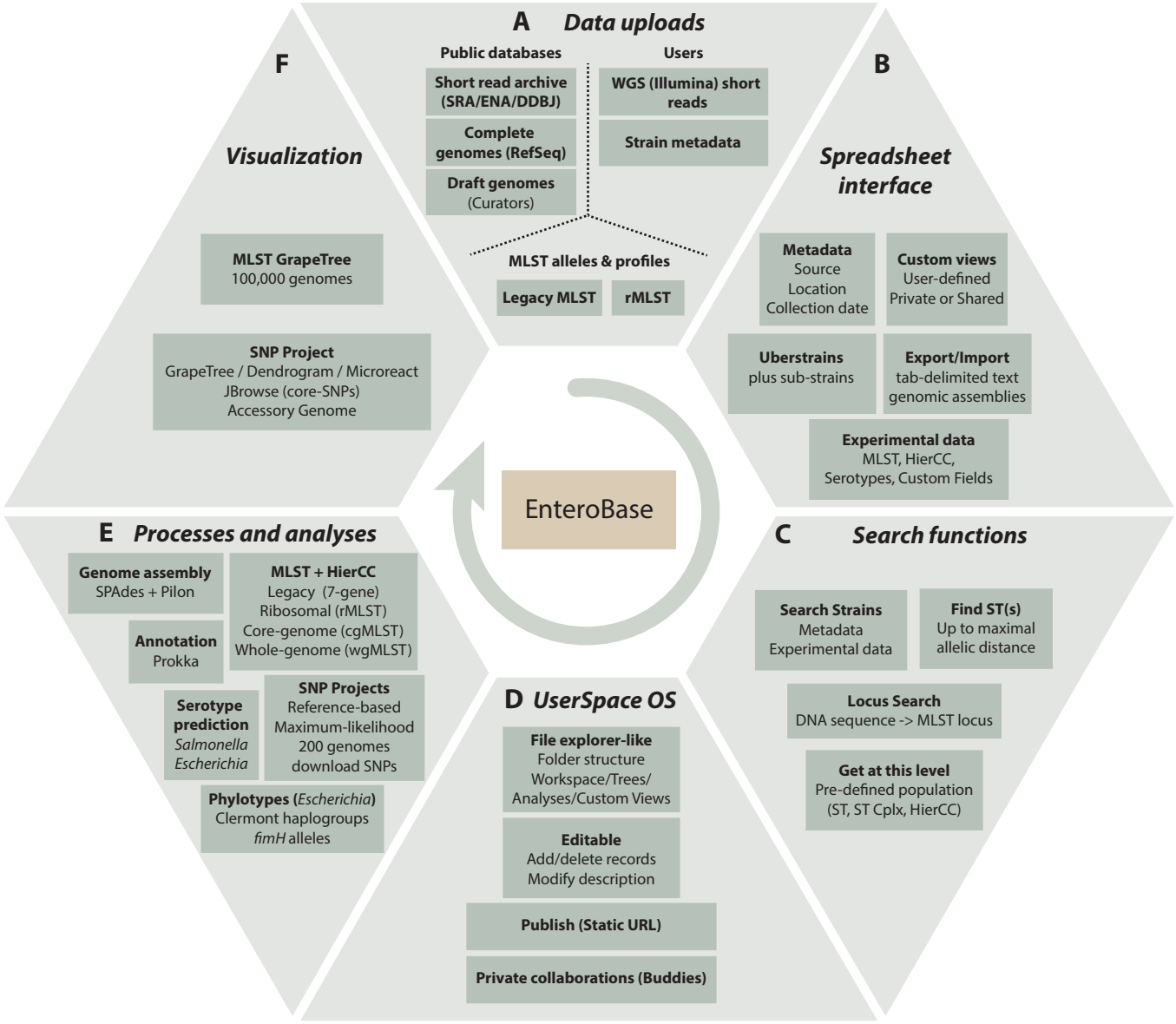
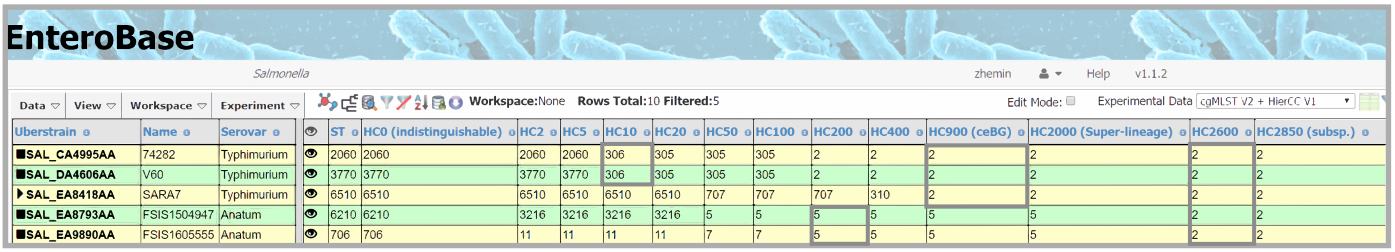


Figure 1. Overview of EnteroBase Features. A) Data uploads. Data are imported from public databases, user uploads and existing legacy MLST and rMLST databases at PubMLST (https://pubmlst.org/). B) Spreadsheet Interface. The browser-based interface visualizes sets of strains (one Uberstrain plus any number of sub-strains) each containing metadata, and their associated experimental data and custom views. Post-release data can be exported (downloaded) as genome assemblies or tab-delimited text files containing metadata and experimental data. Metadata can be imported to entries for which the user has editing rights by uploading tab-delimited text-files. C) Search Strains supports flexible (AND/OR) combinations of metadata and experimental data for identifying entries to load into the spreadsheet. Find ST(s) retrieves STs that differ from a given ST by no more than a maximal number of differing alleles. Locus Search uses BLASTn (Altschul et al. 1990) and UBlastP in USEARCH (Edgar 2010) to identify the MLST locus designations corresponding to an input sequence. Get at this level: menu item after right clicking on experimental MLST ST or cluster numbers. D). UserSpace OS. A file-explorer like interface for manipulations of workspaces, trees, SNP projects and custom views. These objects are initially private to their creator, but can be shared with buddies or rendered globally accessible. E) Processes and analyses, EnteroBase uses EToKi and external programs as described in Supplemental Fig. S1. F) Visualization. MLST trees are visualized with the EnteroBase tools GrapeTree (Zhou et al. 2018a) and Dendrogram, which in turn can transfer data to external websites such as MicroReact (Argimon et al. 2016).

A



Uberstrain	Name	Serovar	ST	HC0 (indistinguishable)	HC2	HC5	HC10	HC20	HC50	HC100	HC200	HC400	HC900 (ceBG)	HC2000 (Super-lineage)	HC2600	HC2850 (subsp.)
SAL_CA4995AA	74282	Typhimurium	2060	2060	2060	2060	306	305	305	305	2	2	2	2	2	2
SAL_DA4606AA	V60	Typhimurium	3770	3770	3770	3770	306	305	305	305	2	2	2	2	2	2
SAL_EA8418AA	SARA7	Typhimurium	6510	6510	6510	6510	6510	6510	707	707	707	310	2	2	2	2
SAL_EA8793AA	FSIS1504947	Anatum	6210	6210	3216	3216	3216	3216	5	5	5	5	5	5	2	2
SAL_EA9890AA	FSIS1605555	Anatum	706	706	11	11	11	11	7	7	5	5	5	5	2	2

B

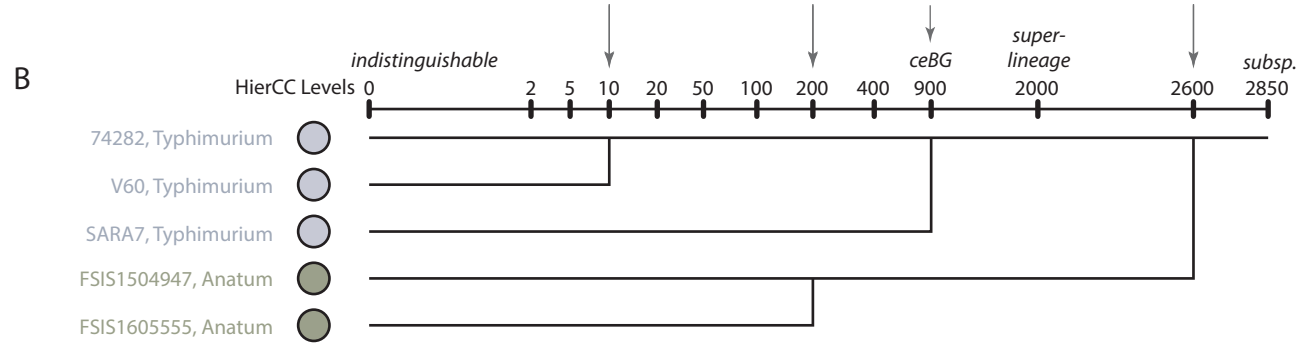


Figure 2. The hierarchical cgMLST clustering (HierCC) scheme in Enterobase. A) A screenshot of *Salmonella* cgMLST V2 plus HierCC V1 data for five randomly selected genomes. The numbers in the columns are the HierCC cluster numbers. Cluster numbers are the smallest cgMLST ST number in single-linkage clusters of pairs of STs that are joined by up to the specified maximum number of allelic differences. These maximum differences are indicated by the suffix of each HC column, starting with HC0 for 0 cgMLST allelic differences through to HC2850 for 2850 allelic differences. The cluster assignments are greedy because individual nodes which are equidistant from multiple clusters are assigned to the cluster with the smallest cluster number. B) Interpretation of HierCC numbers. The assignments of genomic cgMLST STs to HC levels can be used to assess their genomic relatedness. HC0 indicates identity except for missing data. The top two genomes are both assigned to HC10_306, which indicates a very close relationship, and may represent a transmission chain. The top three genomes are all assigned to HC900_2, which corresponds to a legacy MLST eBG. HC2000 marks super-lineages (Zhou *et al*. 2018c) and HC2850 marks subspecies. This figure illustrates these interpretations in the form of a cladogram drawn by hand.

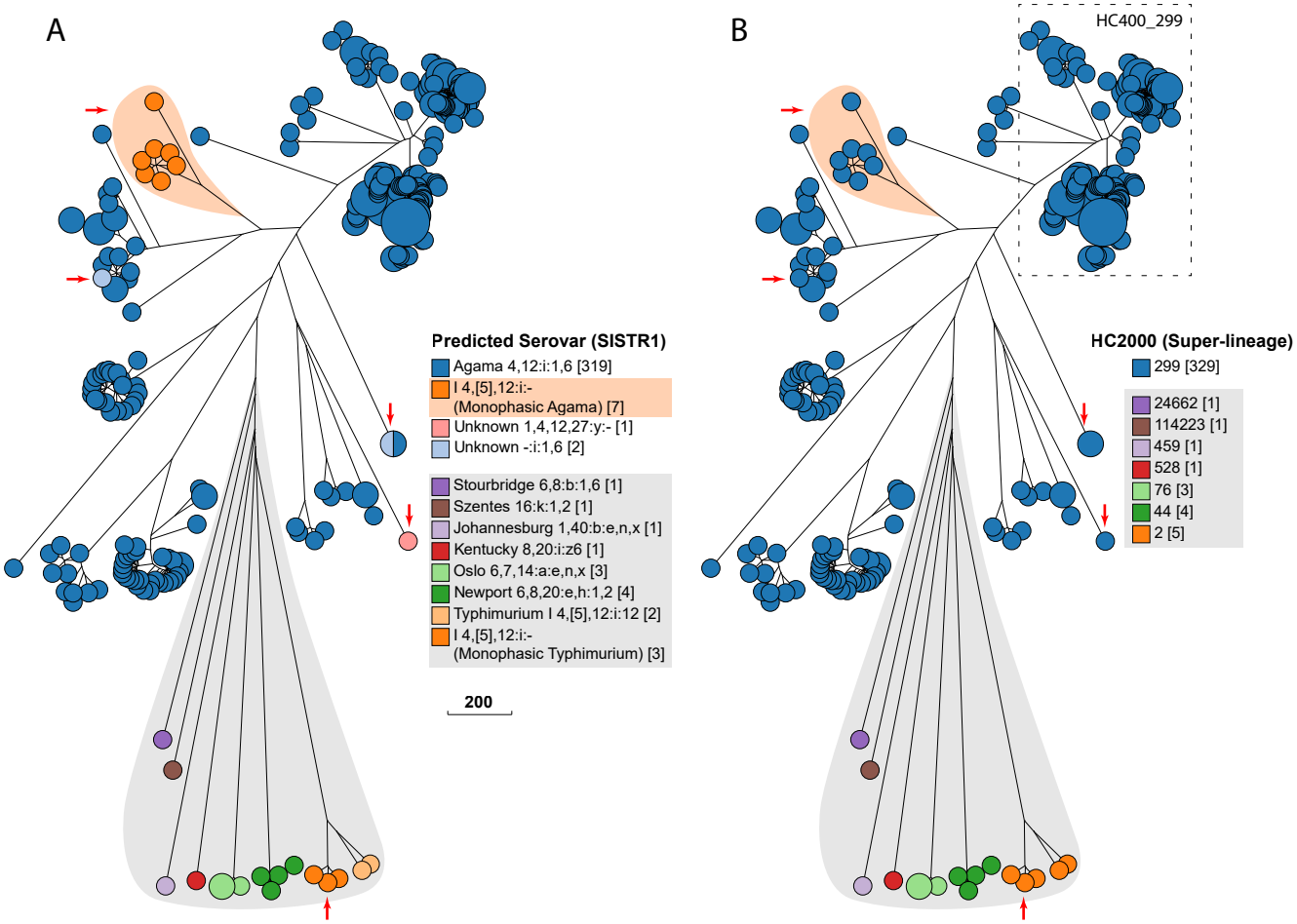


Figure 3. Serovar *versus* HierCC clustering in serovar Agama. GrapeTree (Zhou et al. 2018a) depiction of a Rapid-NJ tree (Simonsen *et al*. 2011) of cgMLST allelic distances between genomic entries whose metadata Serovar field contained Agama or SISTR1 (Robertson *et al*. 2018) Serovar predictions contained Agama. A) Color coding by Predicted Serovar (SISTR1). Arrows indicate isolates whose serovar was not predicted. Orange shading emphasizes 1,4,[5],12:i:- isolates that were monophasic Agama. Gray shading indicates isolates with incorrect Serovar metadata, including 1,4,[5],12:i:- isolates that were monophasic Typhimurium (arrow). B) Color-coding by HC2000 cluster. All Agama entries are HC2000_299, as were the genetically related entries marked with arrows or emphasized by orange shading. Entries from other serovars (gray shading) were in diverse other HC2000 clusters. The dashed box indicates a subset of Agama strains within HC400_299, including all isolates from badgers, which were chosen for deeper analyses in Fig. 4. Scale bar: number of cgMLST allelic differences.

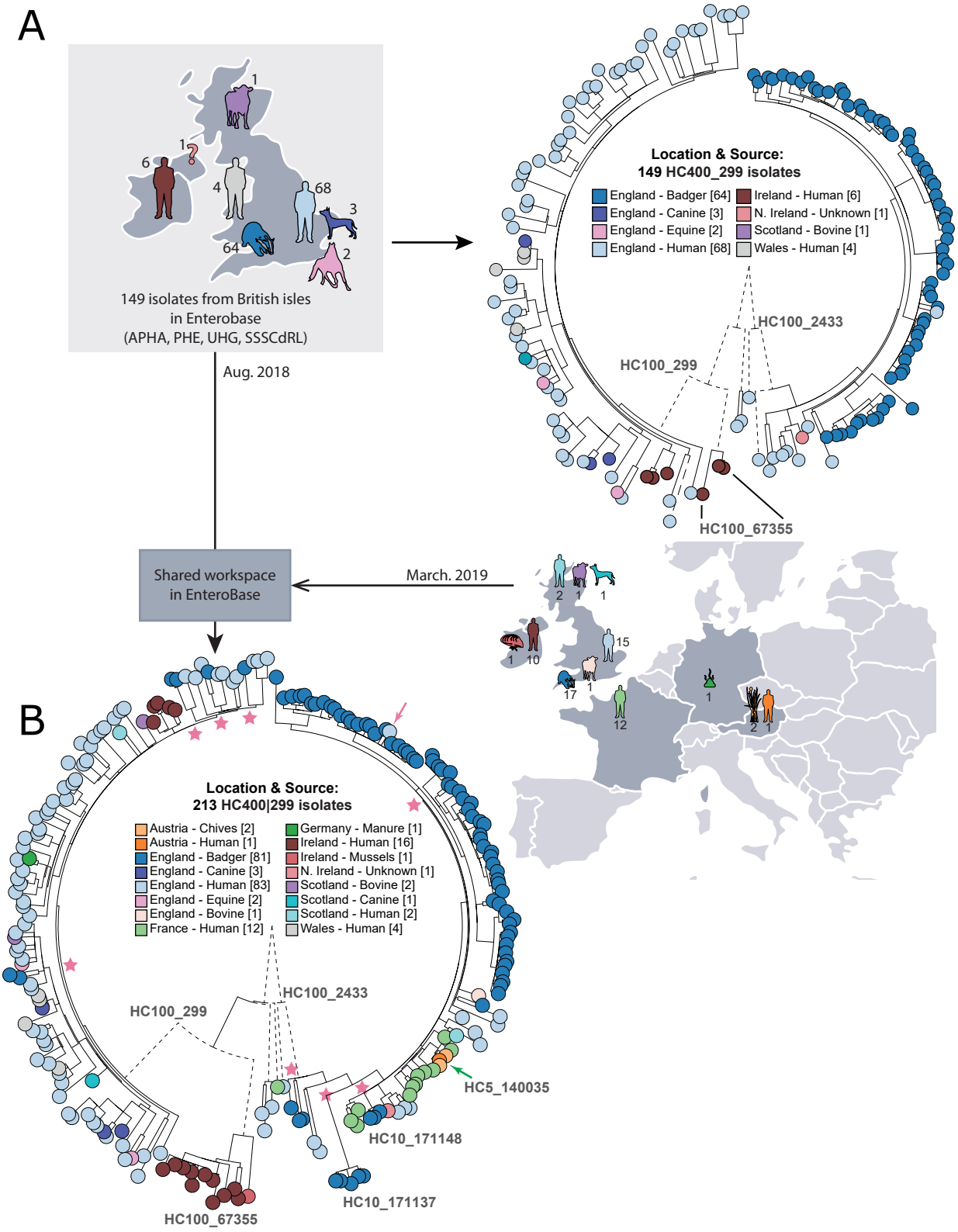


Figure 4. Effects of sample bias on inferred transmission chains within HC400_299 Agama isolates. A) Left: map of hosts in the British Isles of 149 Agama isolates in EnteroBase in August, 2018. Right: Maximum-likelihood radial phylogeny (https://tinyurl.com/AgamaFig4A) based on RAxML (Stamatakis 2014) of 8,791 non-repetitive core SNPs as calculated by EnteroBase Dendrogram against reference genome 283179. Color-coding is according to a User-defined Field (Location & Source). HC100 cluster designations for three micro-clades are indicated. HC100_2433 contained all Agama from badgers. B) Right: summary of hosts and countries from which 64 additional Agama isolates had been sequenced by March 2019. Left: Maximum-likelihood radial dendrogram (https://tinyurl.com/AgamaF4B) based on 9,701 SNPs from 213 isolates. Multiple isolates of Agama in HC100_2433 were now from humans and food in France and Austria. HC100_299 and HC100_67355 now contained multiple isolates from badgers, livestock, companion animals and mussels, demonstrating that the prior strong association of Agama with humans and badgers in part A reflected sample bias. Stars indicate multiple MRCAs of Agama in English badgers while the pink arrow indicates a potential transmission from badgers to a human in Bath/North East Somerset, which is close to Woodchester Park. The green arrow indicates a potential food-borne transmission chain consisting of four closely related Agama isolates in HC5_140035 from Austria (chives x 2; human blood culture x 1) and France (human x 1) that were isolated in 2018. The geographical locations of the badger isolates are shown in Fig. S5.

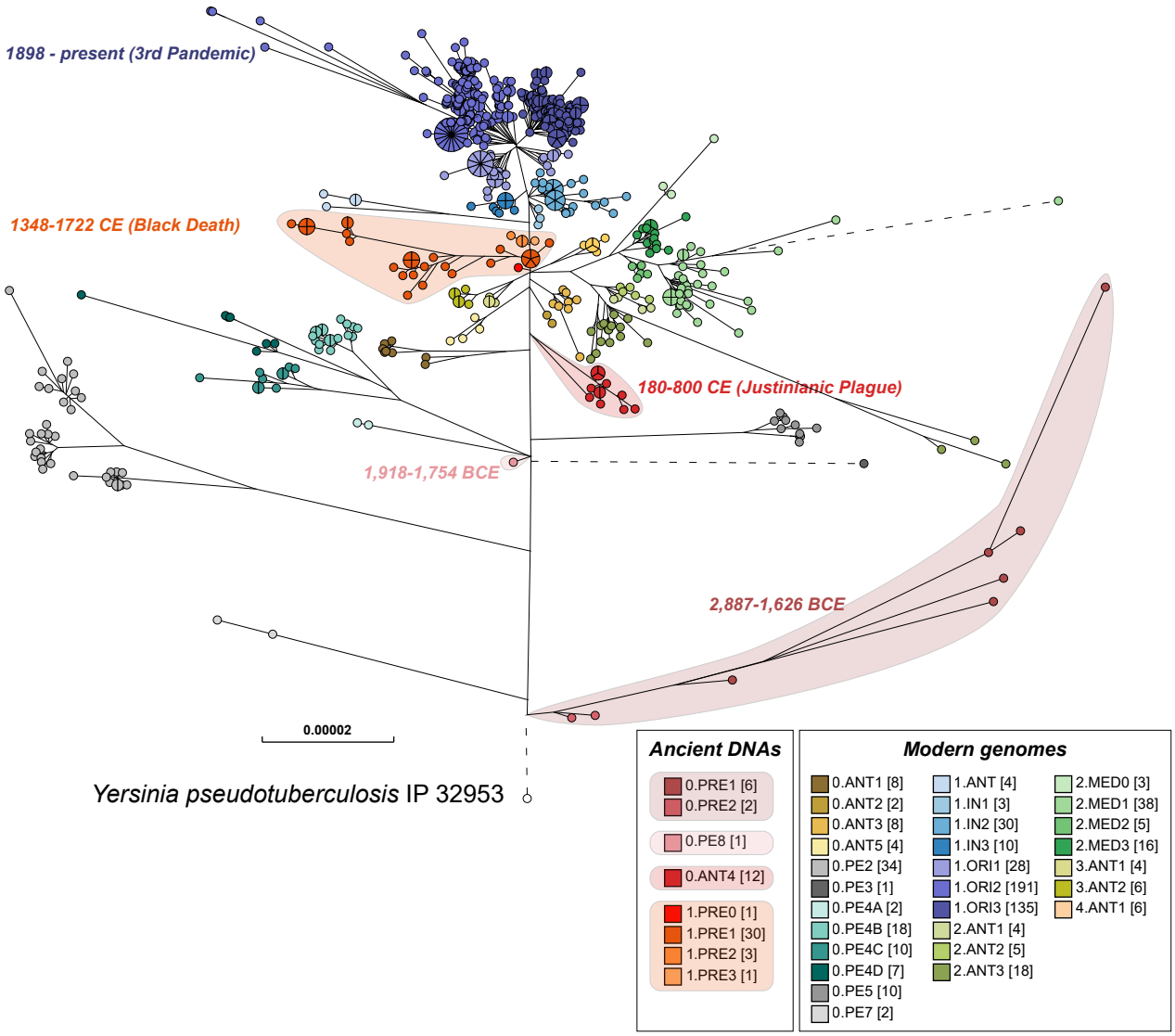


Figure 5. Maximum-Likelihood tree of modern and ancient genomes of *Y. pestis*. EnteroBase contained 1,368 ancient and modern *Yersinia pestis* genomes in October 2019, of which several hundred genomes that had been isolated in Madagascar and Brazil over short time periods demonstrated very low levels of genomic diversity. In order to reduce this sample bias, the dataset used for analysis included only one random representative from each HC0 group from those two countries, leaving a total of 622 modern *Y. pestis* genomes. 56 ancient genomes of *Y. pestis* from existing publications were assembled with EToKi (see Methods), resulting in a total of 678 *Y. pestis* genomes plus *Yersinia pseudotuberculosis* IP32953 as an outgroup (https://tinyurl.com/YpestisWS). The EnteroBase pipelines (Fig. S2D) were used to create a SNP project in which all genomes were aligned against CO92 (2001) using LASTAL. The SNP project identified 23,134 non-repetitive SNPs plus 7,534 short inserts/deletions over 3.8 Mbps of core genomic sites which had been called in ≥95% of the genomes. In this figure, nodes are color-coded by population designations for *Y. pestis* according to published sources (Morelli *et al*. 2010; Cui *et al*. 2013; Achtman 2016), except for 0.PE8 which was assigned to a genome from 1,918-1754 BCE (Spyrou *et al*. 2018). The designation 0.ANT was applied to *Y. pestis* from the Justinianic plague by Wagner et al. 2014, and that designation was also used for a genome associated with the Justinianic plague (DA101) that was later described by Damgaard *et al*., 2018 as 0.PE5.

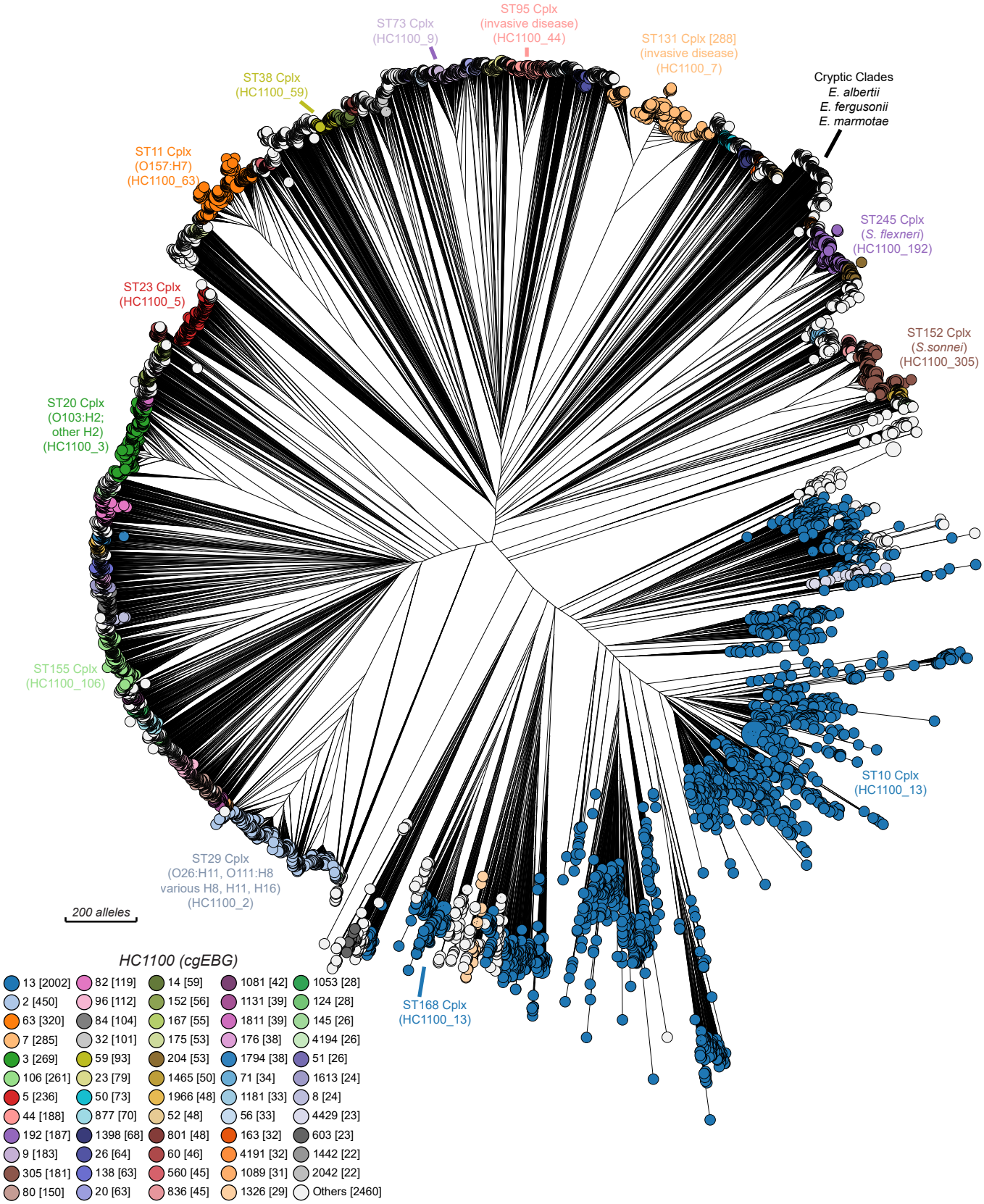


Figure 6. Neighbour-Joining (RapidNJ) tree of core-genome allelic distances in the EcoRPlus Collection of 9,479 genomes. EcoRPlus includes the draft genome with the greatest N50 value from each of the 9,479 rSTs among the 52,876 genomes of *Escherichia* within EnteroBase (August, 2018) (http://tinyurl.com/ECOR-Plus). The nodes in this tree are color-coded by HC1100 clusters, as indicated in the Key at the bottom left. Common HC1100 clusters (plus corresponding ST Complexes) are indicated at the circumference of the tree. These are largely congruent, except that HC1100_13 corresponds to ST10 Complex plus ST168 Complex, and other discrepancies exist among the smaller, unlabeled populations. See Figs S7 and S8, respectively, for color-coding by ST Complex and Clermont typing. An interactive version in which the nodes can be freely color-coded by all available metadata is available at http://tinyurl.com/ECOR-RNJ. A Maximum-Likelihood tree based on SNP differences can be found in Fig. S9.