The user's guide to comparative genomics with EnteroBase. Three case studies: micro-clades within *Salmonella enterica* serovar Agama, ancient and modern populations of *Yersinia pestis*, and core genomic diversity of all *Escherichia.*

Authors: Zhemin Zhou[1¶], Nabil-Fareed Alikhan[1¶], Khaled Mohamed[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, (APHA - Animal and Plant Health Agency, Addlestone, UK); Andy Robertson (Environment & Sustainability Institute, University of Exeter, Penryn, UK); William Tyne and Yulei Fan (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 including pathogens. It currently contains more than 300,000 genomes that have been assembled from Illumina short reads from the genera *Salmonella*, *Escherichia*, *Yersinia*, *Clostridiodes, Helicobacter, Vibrio*, and *Moraxella*. With the recent introduction of hierarchical clustering of core genome MLST sequence types, EnteroBase now facilitates the identification of close relatives of bacteria within those genera inside of a few hours of uploading their short reads. It also supports private collaborations between groups of users, and the comparison of genomic data that were assembled from short reads with SNP calls that were extracted from metagenomic sequences. Here we provide an overview for its users on how EnteroBase works, what it can do, and its future prospects. This user's guide is illustrated by three case studies ranging in scale from the miniscule (local transmission of *Salmonella* between neighboring social groups of badgers) through pandemic transmission of plague and microevolution of *Yersinia pestis* over the last 5,000 years to a novel, global overview of the population structure of all of *Escherichia*.

Introduction

Multiple publications have described comparative genomics and epidemiological transmission chains of *Salmonella, Escherichia* or *Yersinia* (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). These studies were based on single nucleotide polymorphisms (SNPs) from hundreds or even thousands of core genomes. However, that scale pales in comparison to the numbers of publicly available archives of short read sequences (SRAs) of bacterial pathogens which have been deposited since the recent drop in price of high throughput sequencing (Wetterstrand 2019). In January 2019, the SRA at NCBI contained genomic sequence reads from 319,359 *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 several recent publications providing an overview of the population structure of *Salmonella* (Alikhan et al. 2018), 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 largely been promulgated by word of mouth. Here we provide a high-level overview of the functionality of EnteroBase, followed by an illustrated user's guide based on 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 (Fig. 2), many of which are also available as part of a publicly available repository of useful tools called EToKi (EnteroBase ToolKit) (https://github.com/zheminzhou/EToKi) (Fig. 3). EnteroBase performs daily scans of the GenBank SRA via its Entrez APIs (Clark et al. 2016) for all the new Illumina short reads that have been deposited within the last 30 days. It uploads novel Illumina short read sequences, and assembles them (EBAssembly, Fig. 3A) into annotated draft genomes, which must pass quality control (Table S1; Fig. 3A). For each Enterobase database, it also fetches the metadata associated with the records, and transcribes it into Enterobase metadata format. 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 into pre-defined categories using a pre-trained Native Bayesian classifier implemented in Python NLTK library (Loper and Bird 2002).

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. 3B) and their Sequence Types (STs) are assigned to population groupings. For *Salmonella*, serotypes are predicted based on MLST and/or by SISTR (Yoshida *et al.* 2016; Robertson *et al.* 2018) according to genomic sequence plus legacy MLST. We have also developed a comparable program, EB*Eis*, which performs the same tasks for *Escherichia,* and which is also included in EToKi. By default, assemblies and strain metadata are immediately publicly available, but users can specify a delayed release date of up to 12 months.

In April 2019, EnteroBase provided access to 318,123 genomes (Table 1). In order to allow comparisons with historical data, EnteroBase also maintains additional legacy 7-gene MLST results plus their metadata from 16,466 strains that were based on classical Sanger sequencing.

Ownership, access and privacy. All users can upload new entries to EnteroBase, consisting of paired-end Illumina short reads plus their metadata. EnteroBase stores those short reads locally, and does not allow anybody to download them. Local storage would allow automated, brokered uploads of reads and metadata for the user to European Nucleotide Archive in the future. However, such brokered uploads have only first been initiated very recently.

All assembled genomes and metadata within EnteroBase are available for comparative genomic analyses. All users can find records using Enterobase tools and include these genomes in their analyses within EnteroBase. However, users own their uploaded data, and can embargo public downloading for up to

12 months after the reads and metadata were uploaded. Prior to that release date, only the owner and their buddies, admins or curators can download the genomic assemblies, their metadata or MLST genotypes, or trees containing nodes. Procedures have already been implemented to strip pre-release metadata, genotypes and genomic assemblies from downloaded data when users lack ownership privileges. GrapeTree and Dendrogram are currently being modified to strip pre-release nodes from trees, figures and SNP matrices prior to downloading those items, but those modifications are not yet complete.

Uploaded metadata belong to the owner indefinitely, and can only be edited by the owner or buddies to whom they have assigned editing rights, or by admins and curators. In general, metadata that were imported from a SRA are not editable, except by admins and curators. However, it is possible to assign editing rights to users with claims to ownership or who possess special insights.

MLST Population structures. Unique sequence variants of each of the genes in an MLST scheme are assigned unique numerical designations. 7-gene MLST STs consist of seven integers for the alleles in 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). 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 *S. enterica* (Achtman et al. 2012). EnteroBase has also implemented similar population groups (reBGs) for rMLST in *Salmonella*, which are largely consistent with eBGs or their sub-populations (Alikhan *et al.* 2018). The Nomenclature Server module in EnteroBase (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, such growing chains are terminated

when they reach proximity to other populations (2 alleles different 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. As a result, almost every genome results in a unique cgST number, resulting in 100,000s of cgST numbers in EnteroBase, many of which only differ from other cgSTs by missing data. Since late 2018, EnteroBase has also implemented HierCC (Hierarchical Clustering of cgMLST STs), a novel approach which supports analyses of population structures at multiple levels of resolution. It was not immediately obvious which cut-off values in stepwise allelic distances would reliably resolve natural populations on the basis of HierCC of cgMLST distances. We therefore calculated a matrix of allelic distances (excluding missing data in pairwise fashion) 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. These CPU-intensive calculations only needed to be performed once because subsequent assignments of new cgSTs to HierCC clusters only require the expansion of existing matrices for each new cgST. EnteroBase only displays a small subset of the HierCC levels, as illustrated here, and leaves the choice of desired level of resolution up to the user. For *Salmonella*, in which the cgMLST scheme consists of 3,002 loci*,* thirteen HierCC levels are reported, ranging from HC0 (indistinguishable except for missing data) to HC2850 (Fig. 4). Currently, 10 levels are reported for the other genera, ranging from HC0 to HC1100 for *Escherichia* or *Clostridioides* and from HC0 to HC600 for *Yersinia.* Our experience with *Salmonella* indicates that HC2000 corresponds to the super-lineages described elsewhere (Zhou *et al.* 2018c). In *Salmonella*, HC900 corresponds to eBGs; long-term endemic persistence seems to be associated with HC100 or HC200; and epidemic outbreaks with HC2, HC5 or HC10. Similarly, HC1100 seems to correspond to ST Complexes in *Escherichia* whose

cgMLST scheme consists of 2,512 loci. 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, some entries have two or more sets of genotyping data. For example, EnteroBase includes some 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 because it automatically imports all new SRA records. In still 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 provides a mechanism for dealing with such duplications. It implements the concept of an Uberstrain, which can be a parent to one or more sub-strains. Sub-strains remain invisible unless they are specified in the search dialog (Fig. S2). For example, the EnteroBase *Yersinia* database includes two distinct genomes for *Y. pestis* CO92, the genome originally sequenced in 2001 (Parkhill *et al.* 2001) and a subsequent, corrected genome from 2015 (Johnson *et al.* 2015). The more recent genome sequence is the Uberstrain for genomes from multiple bacterial colonies that were sequenced after infecting prairie dogs in the laboratory with CO92 (project PRJNA340278). The older genome is maintained in EnteroBase as a separate Uberstrain because it was previously used as a reference genome for SNP calling (Morelli et al. 2010; Bos et al. 2011), and continues to be used as a reference for *Y. pestis* ancient DNA (aDNA) projects based on archaeological samples (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).

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

S. enterica subsp. *enterica* encompasses more than 1,547 defined serovars (Guibourdenche et al. 2010). 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). We were therefore interested to learn that the University of Liverpool possessed a historical collection of *Salmonella* isolates that had been isolated in 2006-2007 from European badgers (*Meles meles*) in Woodchester Park, Gloucestershire, England. Badgers commonly excrete rare serovars of *S. enterica* (Wray *et al.* 1977; Wilson *et al.* 2003), and many of the isolates in the University of Liverpool collection had been assigned to serovar Agama (antigenic formula: 4,12:i:1,6). These isolates thus provided an opportunity to analyze the genetic diversity of a rare serovar within a single host species over a limited geographical area, and to compare that diversity and population structure with genomes of the same serovar from other hosts and multiple geographical sources.

When this project began in late 2018, only 0.09% (134/156,347) of the genome assemblies in EnteroBase contained “Agama” in the serovar metadata field. The existing metadata indicated that all were from humans in the British Isles even though EnteroBase contains many *Salmonella* genomes from other geographical sources, from humans as well as the environment. We sequenced the genomes of 72 supposed Agama isolates from badgers within the University of Liverpool collection, and uploaded the short reads and strain metadata into EnteroBase.

Search Strains. The browser interface to EnteroBase is based on a workspace concept (https://tinyurl.com/EnteroBase-WS), implemented as a spreadsheet-like window that can page through 1,000s of entries. However visual scanning of 1,000s of entries is inefficient. EnteroBase therefore offers powerful search functions for identifying sets of isolates that share common phenotypes (metadata) and/or genotypes (experimental data).

Metadata consist of information on the bacterial strain, including isolation year, host and geographical location as well as unique identifier barcodes, strain name, and accession codes that act as hyperlinks to the original data sources. For *Salmonella,* metadata also includes serovars specified in the SRAs or uploaded by our users. Experimental data includes assembly statistics, and MLST ST (Sequence Type) assignments plus population groupings for all the MLST schemes in Table 1. For *Salmonella,* it also includes predictions of serovar based on SISTR and/or MLST STs.

Many entries lack metadata information on serovar, or contain erroneous assignments, and the software predictions are also not fail-proof. We therefore used the Search Strains dialog box to search for entries where the serovar field in metadata contains “Agama” or the software predictions of serovar made by SISTR contain “Agama” (Fig. 5A). The search results are displayed in two co-ordinated windows, a metadata window at the left and an experimental data window at the right (Fig. 5B). Fig. 5B shows experimental data for cgMLST + HierCC, but other types of experimental data can be chosen from a drop-down list.

International participation in a collaborative network. Our initial phylogenetic analyses indicated that Agama consisted of multiple micro-clusters, one of which included all Agama isolates from badgers. However, these results were possibly distorted by a highly skewed geographical sampling bias because almost all isolates were from England. We therefore formed the Agama Study Group, consisting of our 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) in order to provide shared access to the Workspaces and phylogenetic trees in the ‘Zhou et al. Agama’ folder (Fig. 6 left). That folder has now been made publicly available and can be found at http://enterobase.warwick.ac.uk/species/index/senterica.

Analyses can benefit from aggregating fields from multiple sources of experimental data and/or User-defined Fields for text annotations in a user-defined Custom View. A new Custom View was established for the Agama Study Group (Fig. 6 right), saved in the Agama folder, and thereby also shared with the entire Study Group. This view thereby remained private during the initial analyses, and became public when the Agama folder became publicly visible.

Members of the Agama Study Group were requested to sequence genomes from all the Agama strains in their collections, and to upload those short reads to EnteroBase, or alternatively to send the bacterial strains or their DNAs to University of Warwick for sequencing and uploading. The new entries were added to the 'Zhou et al., All Agama Strains' workspace (Fig. 6), which now contains 344 genomes. The isolates were from Europe, Africa and Australia, with collection years ranging from 1956 to 2018 (Table S2).

Global population Structure of Agama. The genetic relationships within serovar Agama were investigated by a neighbor joining GrapeTree (Zhou *et al.* 2018a) of cgMLST allelic distances between all genomes in the workspace (Fig. 7). Color coding the nodes of the tree by SISTR serovar predictions confirmed that most isolates were Agama (Fig. 7A). However, there were some interesting exceptions. One micro-cluster (light orange shade at 11:00) consisted of seven monophasic isolates with a defective or partial *fljB* (H2) CDS. As a result, SISTR scored these genomes as monophasic Typhimurium. Monophasic Typhimurium and monophasic Agama share identical antigenic formulas, and monophasic Typhimurium isolates are more common. SISTR classified six other inter-dispersed isolates as unknown serovars (arrows in Fig. 7) because it could not predict their O antigens. For all 13 isolates, their genomes were interspersed among true Agama genomes to which they are genetically closely related.

The tree also included 16 isolates on long branches, which represented false calls. They had only been designated Agama by the accompanying metadata, and SISTR assigned then to a variety of other serovars, as did their eBG

associations (Fig. 6 right, Fig. 7 grey shading). Coloring the nodes in the tree by HierCC (Fig. 7B) confirmed that these 16 isolates had been falsely serotyped as Agama, or represented contamination with other *Salmonella*. All nodes in HC2000 cluster number 299 (HC2000|299) clustered together, and all genomes called Agama by SISTR were HC2000|299 (Fig. 7B). The 16 assemblies in the grey zone belong to other, unrelated HC2000 clusters, which are associated with their predicted serovars.

These results shows 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 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 the higher resolution HierCC HC400|299 cluster, as were multiple other isolates from humans and other animals (Fig. 7B, dashed box). This cluster of genomes was analyzed in greater depth using the dendrogram GUI within EnteroBase to calculate and visualize Maximum-Likelihood trees (RAxML) based on core, non-repetitive SNPs called against a reference draft genome. We calculated one tree (Fig. 8A) prior to establishing the Agama Study group and a second tree (Fig. 8B) after additional Agama genomes had been sequenced. A comparison of the two trees is highly instructive on the effects of sample bias.

The original HC400|299 genomes were obtained from 149 isolates from the British Isles plus one isolate from Germany (Fig. 8A). Those genomes fell into four micro-clades (Fig. 8A) which corresponded to HC100|299, HC100|2433, HC100|28865 and HC100|67355. HC100|2433 contained all badger isolates plus a few human isolates. At this time point, all isolates from badgers had been collected at Woodchester Park in 2006-2007 within the context of a long-term

live capture-mark-recapture study (McDonald *et al.* 2018). Agama isolates from those badgers were more distal than the isolates from humans in HC100|2433, suggesting that the most recent common ancestor of all badger isolates had been transmitted from humans. One human isolate was from Bath/North East Somerset, which is compatible with a back-transmission from badgers to humans because that genome was within the cluster of badger isolates, and that location is not far from Woodchester Park.

The badgers in Woodchester Park occupy adjacent social group territories which contain several setts (burrows) where the trapping took place. To investigate the transmissions between setts and social groups, GrapeTree was used to transmit a Newick sub-tree of HC100|2433 plus metadata to an external program, MicroReact (Argimon *et al.* 2016), which is specialized in analyzing geographical associations. That sub-tree contained multiple distinct internal micro-clusters of Agama from badgers, which corresponded to distinct HC10 clusters (Fig. S3A). Four HC10 clusters each contained isolates from two social groups in close proximity (Fig. S3B), in accordance with observations of occasional movements of badgers between social groups in adjacent territories (Rogers *et al.* 1998).

Long-term dispersals and inter-host transmissions. The additional isolates that were subsequently sequenced by the Agama Study Group provided insights on the dissemination of Agama over a much longer time frame. The additional genomes included 17 isolates from English badgers from multiple locations in south-west England (Fig. S3C,D) that had been isolated between 1998 and 2016, and stored at APHA. Eleven of those isolates were in HC100|2433, the same micro-clade as all badger isolates described above (Fig. 8B). However, they were not interspersed among the initial genomes. Instead they defined novel micro-clusters, including HC10|11137 and HC10|171148, which were the most basal clades in HC100|2433. The other six badger isolates were from different geographical areas than HC100|2433 (Fig. S3B,C) and were interspersed among human isolates in HC100|299 (Fig. 8B), which had previously not contained any badger isolates at all. 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.

Still further insights on Agama were provided by other additional Agama genomes, which included isolates from humans in Ireland, France or Austria, as well as multiple isolates from animals and food (Fig. 8B). HC100|299 contained human isolates from Ireland as well as isolates from dogs, cows and horses. All isolates in HC100|67355 were from Ireland, 12 from humans and one from mussels. HC100|2433 now contained not only isolates from the British Isles, but also isolates from France and Austria. We were particularly struck by four genomes in that clade (Fig. 8B, green arrow at 04:00) which were in HC10|140035, and had all been isolated in 2018. Three of these isolates were from Austria, two from frozen chives and one from a blood culture from a case of human septicemia. They were assigned to HC5|162029, and differ from each other in pair-wise comparisons by 2-5 non-repetitive core SNPs and 2-4 cgMLST loci. The fourth isolate was from a human in France. It differed by 5 SNPs and 5 cgMLST loci from the three Austrian isolates and by 8-35 SNPs and 6-23 cgMLST loci from other Agama in France. We do not know of any epidemiological data that support food-borne transmission of these organisms between France and Austria, or that indicate that frozen chives are a vehicle for food-borne invasive Agama disease.

These observations confirm that Agama is rare. They also demonstrate that 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 2: Combining modern *Y. pestis* genomes with ancient metagenomes.

EnteroBase automatically scours short read archives for Illumina short reads from cultivated isolates, assembles their genomes and publishes draft assemblies that pass quality control. In March 2019, EnteroBase had assembled >1,250 genomes of *Y. pestis* from short reads, 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 it would be difficult to implement adequate, automated quality control measures. However, EnteroBase administrators can upload such genomes after *ad hoc* assessment of sequence quality, and EnteroBase already contains standard complete genomes such as CO92 (Parkhill *et al.* 2001) and the other genomes used to derive the *Y. pestis* phylogeny (Morelli *et al.* 2010).

Ancient *Y. pestis*. 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. Indeed, EnteroBase already contains several reconstructed *S. enterica* genomes from ancient archaeological material in order to facilitate analyses of bacterial evolution (Zhou *et al.* 2018c). 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). 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 publicly available, command-line version of the EnteroBase EToKi calculation package can now calculate a modified genome assembly by modifying a reference genome by SNP calls from metagenomic data (Fig. S4). This module was used to generate such modified genome assemblies from all published metagenomes of ancient material containing *Y. pestis* after specifying an in-group of 4 *Y. pestis* genomes, an out-group of 4 genomes from the related taxa and CO92 as the reference *Y. pestis* genome to be used for SNP mapping, and the resulting assemblies and metadata were uploaded to EnteroBase. EnteroBase users can now include them in a workspace of their choice together with other reconstructed genomes and modern genomic assemblies, and use the EnteroBase SNP dendrogram module to calculate and visualize a Maximum Likelihood tree based on a SNP matrix relative to the reference genome of choice. Users are currently limited to calculating SNP dendrograms for no more than 200 genomes due to the computing demands imposed by generating large ML trees.

Fig. 9 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 (Morelli *et al.* 2010; Cui *et al.* 2013) as extended in a recent review article (Achtman 2016) 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, 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 three other species: *E. fergusonii, E. albertii* and *E. marmotae*. Furthermore, *E. coli* itself includes the genus *Shigella* (Pupo *et al.* 2000), which was assigned a distinctive genus name because it causes dysentery, a distinctive form of gastroenteritis. Initial analyses of the phylogenetic structure of *E. coli* identified multiple deep branches, called haplogroups (Selander *et al.* 1987). Analyses of environmental isolates from lakes revealed the existence of multiple “cryptic clades” which are distinct from the main *E. coli* haplogroups and the other *Escherichia* species (Walk et al. 2009; Luo et al. 2011). Bacterial isolates are routinely assigned to haplogroups by PCR tests for the presence of variably present genes from the accessory genome (Clermont *et al.* 2013) or by software programs that identify the presence of those genes from genomic sequences (Beghain *et al.* 2018; Waters *et al.* 2018).

In 2006, legacy MLST was introduced for *E. coli,* including the assignment of STs to ST Complexes (Wirth *et al.* 2006). EnteroBase continues to assign novel STs to the original ST Complexes, and 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 order to render this sample amenable to calculating an ML tree of core SNPs, we took advantage of low rates of neutral diversity which are characteristic of ribosomal proteins, and reduced the sample size by selecting one genome from each of the *Escherichia* rSTs in EnteroBase. In homage to the EcoR collection (Ochman and Selander 1984), a classical group of 72 bacterial strains that represented the genetic diversity found with multilocus enzyme electrophoresis, we designate this collection of 9,479 genomes as the EcoRPlus collection.

Core genome genetic diversity within *Escherichia.* Homologous recombination is widespread within *E. coli* (Wirth *et al.* 2006). We had 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 differences in numbers of different cgMLST alleles between the strains of EcoRPlus (Fig. 10). The most predominant, discrete sets of clustered nodes were also largely congruent with cgMLST HC1100 hierarchical clustering (Fig. 10). Furthermore, with one exception (ST168 Complex), assignments to common ST Complexes based on legacy 7-gene MLST were also predominantly congruent with the HC1100 clustering (compare Fig. S5 with Fig. 10).

Fig. 10 may represent the first detailed overview of the entire genetic diversity of the core genome of *Escherichia*. It is readily possible to examine its features in real time (http://tinyurl.com/ECOR-RNJ) due to the ability of the GrapeTree algorithm to deal with large numbers of cgSTs (Zhou *et al.* 2018a). Nodes can be readily colored by any available metadata or experimental data (Fig. S5, S6, S7), 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 used FastTree (Price *et al.* 2010; Liu *et al.* 2011) to calculate a Maximum-Likelihood (ML) tree of the 1,230,995 core SNPs within all 9,479 genomes (Fig. S6). 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 V formed distinct long branches of comparable lengths (Fig. S6 inset).

EnteroBase also provides phylotype predictions for *Escherichia* based on Clermont typing according to two distinct algorithms (Beghain *et al.* 2018; Waters *et al.* 2018), as well as *fimH* allelic designations (Roer *et al.* 2017). The assignments by both Clermont typing algorithms are largely congruent. And those assignments are also largely congruent with the general structure of the

ML tree (Fig. S6 inset). However, Clermont typing is based on the presence/absence of accessory genes, which are more variable than are sequences of the core genome. As a result, close examination of the ML tree shows multiple exceptions, as indicated by flashes of discrepant colors in Fig. S7. Clermont typing never included *S. flexneri* (HC1100|192) and scores it incorrectly as haplogroup A, or *S. sonnei,* which it scores as Unknown (Fig. S7).

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). The underlying infrastructure of EnteroBase is sufficiently generic that it was readily extended to *Clostridioides, Helicobacter* and *Vibrio*, and could in principle be extended to other taxa.

EnteroBase was developed 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 also 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.

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 also be suitable for the annotation of clusters within phylogenetic trees (Case Study 2). 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 currently under development which will support similar approaches for ancient and modern genomes of *S. enterica* and for 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). The user's guide presented here provides initial insights into these various capacities, as illustrated by the three case studies. More details can be found in the online documentation (https://enterobase.readthedocs.io/en/latest/).

EnteroBase is already being used to identify genetically related groups of isolates (Johnson *et al.* 2019; Haley *et al.* 2019; Diemert and Yan 2019; Numberger *et al.* 2019), and an international outbreak of *S. enterica* serovar Poona in baby food was concluded to be associated with HierCC HC5|164707 (Jones *et al.* 2019). Case Study 1 on Agama illustrates how to explore HierCC genomic relationships at multiple levels, ranging from HC2000 (super-lineages) for inter-continental dispersion down to HC5-HC10 for detecting local transmission chains. This case study also illustrated the sharing of private data and analyses with colleagues, including the use of Custom Views and User-defined Fields, and how to work to a common goal by using EnteroBase to assemble and genotype the genomic sequences produced in multiple laboratories.

Case Study 1 may be also become a paradigm for how to examine chains of transmission between humans or between humans and their companion or domesticated animals because of four Agama isolates in HC10|140035. Their assignment to this HierCC cluster indicates that the most distinct of the pairs of

genomes differed at no more than 10 of the 3,002 cgMLST loci. At the core genome level, these isolates differed by 5 to 10 non-repetitive SNPs. It seems likely that this observation reflects hitherto undiscovered long-distance transmissions *via* the food chain because three of these isolates were from Austria, two from frozen chives and the third from blood culture from a human patient. The fourth isolate was a human isolate from France. We anticipate that large numbers of such previously silent transmission chains will be revealed as EnteroBase is used more extensively.

The second case study uses *Y. pestis,* the cause of plague, to illustrate how genomic data in an EnteroBase database can facilitate the reconstruction and phylogenetic analyses of genomes from metagenomic sequences. The resulting phylogenetic tree (Fig. 9) presents a unique overview of the core genomic diversity over 5,000 years of evolution and pandemic spread. EnteroBase will continue to automatically assemble modern genomes of *Yersinia,* including all *Y. pestis*, and this tree will be updated at regular intervals, or upon request, as additional genomes or metagenomes become available. The *Yersinia* database includes cgMLST plus HierCC, and it appears that HC10 designations are roughly comparable to the manual annotations shown in Fig. 9. However, the single linkage clustering used for hierarchical clustering requires a reasonably large dataset before it is reliable, and it is uncertain whether the current HierCC clusters would be stable with time because they were based on only 1,250 *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.

This case study illustrates how to combine genomes from single colony isolates with genomes reconstructed from metagenomic sequences. 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 results in the form of a GrapeTree can now be evaluated and used by non-bioinformaticians (Green 2018). In principle, this

approach could be extended to metagenomes from both modern and ancient samples for any taxon where a reference genome already exists, thus providing a general tool for comparing genomes from cultivated bacteria with metagenomic sequences from modern and/or ancient material.

EnteroBase currently limits users to a maximum of 200 genomes for ML trees and only admins can import external data from command-line procedures. However, we hope to be able to increase the capacities for tree calculation in the future and would welcome collaborations with scientists performing metagenomic analyses of ancient DNA to ensure continuing updates of this *Y. pestis* tree and of comparable trees for other pathogens.

The third case study addresses how to choose a fully representative set of genomes for a nearly definitive overview of core-genomic diversity within a genus. This question was previously addressed for *Salmonella* (Alikhan et al. 2018), by choosing 926 representative genomes to represent known populations (subspecies and reBGs). However, the populations within *Escherichia* were not as well established as for *Salmonella.* We therefore assembled the EcoRPlus Collection of 9,479 genomes by choosing one genome per rST among all 52,876 genomes available in EnteroBase in 2018. EcoRPlus is a genome-based successor of EcoR (Ochman and Selander 1984), which consisted of 72 representatives from a collection of 2,600 strains of *E. coli* that had been tested by multilocus enzyme electrophoresis. The genomic assemblies and known metadata of EcoRPlus are now publicly available (http://tinyurl.com/ECOR-Plus), and can serve as a reference set of genomes for future analyses with other methods.

Several discrete *E. coli* populations that have each been the topics of multiple publications were prominent upon visual examination of an NJ tree of cgMLST allelic diversity color-coded by HierCC HC1100 (Fig. 10). These included a primary cause of hemolytic uremic syndrome (O157:H7), a common cause of invasive disease in the elderly (ST131 and other members of the ST131

Complex), as well as various groups of *Shigella* that cause dysentery. However, despite years of association with *E. coli* by the authors, they did not recognize multiple other discrete populations which were also prominent in this tree, and are apparently also common causes of global disease in humans and animals. The annotation of this tree would be a laudable task for the entire scientific community interested in *Escherichia.* We also note that HierCC is apparently a 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. Fig. S3A 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 this 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. S3B and S3C and 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. 7 are publicly available in EnteroBase in the workspace 'Zhou et al. All Agama strains' (https://tinyurl.com/AgamaWS). The genomic properties, metadata and accession codes for the 329 genomic assemblies in HC2000|299 are summarized in Table S2 and S3.

Integration of ancient *Yersinia pestis* genomes in EnteroBase. Metagenomic reads from ancient samples might contain sequence reads from both the species of interest and also from other environment taxa that are genetically similar. In order to remove such non-specific reads after extraction with the EToKi prepare module, we used EToKi assemble module to align the extracted reads in comparison with an ingroup of related genomes and an outgroup of genomes from other species. The ingroup consisted of *Y. pestis* genomes CO92 (2001), Pestoides F, KIM10+ and 91001. 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 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 14 genomes, the remaining 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 27 pseudo-genomes were stored in EnteroBase together with their associated metadata.

Data access

Illumina reads for 344 genomes of *S. enterica* serovar Agama are available in the NCBI short read archive with accessions shown in Supplementary Table 3. The corresponding metadata, genomic assemblies and annotations are available from the publicly available workspaces entitled "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 7 is available at https://tinyurl.com/AgamaFig7. Trees presented in Figure 8 are available separately at (A) https://tinyurl.com/AgamaFig8A and (B) https://tinyurl.com/AgamaF8B; An interactive version of Figure 9 is available at http://tinyurl.com/Ypestis-ancSNP; The tree present in Figure 10, Figure S5 and S7 is available at http://tinyurl.com/ECOR-RNJ; The MicroReact projects of Figure S3 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;

Acknowledgements

EnteroBase development was funded by the BBSRC (BB/L020319/1) and the Wellcome Trust (202792/Z/16/Z).

Author Contributions.

Figure Legends

Figure 1. Overview of EnteroBase Features. A) Data uploads come from public databases, user uploads and existing legacy and rMLST databases at PubMLST (https://pubmlst.org/). B) EnteroBase offers a spreadsheet-like interface to 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 and tab-delimited text-files can be used to import (upload) metadata for entries to which the user has editing rights. C) The search strains dialog box supports flexible combinations (and/or) of metadata and experimental data to load matching entries into the spreadsheet. Find ST(s) retrieves STs that user-defined, minimal number of identical alleles with a specified ST. Locus search identifies the locus designation in any MLST scheme in EnteroBase according to searches for a given sequence with BLASTn (Altschul *et al.* 1990) and UBlastP in USEARCH (Edgar 2010). D). UserSpace OS. This interface is used for file-explorer like manipulations of workspaces, trees, SNP projects and custom views. By default, all of these objects are private to the user who has created them, but they can be shared with buddies or made globally accessible. E) Processes and analyses which are implemented within EnteroBase, including the external programs Prokka (Seemann 2014), SPADES (Bankevich *et al.* 2012), and Pilon (Walker *et al.* 2014). F) Visualization tools, including transferring data to external websites such as MicroReact (Argimon *et al.* 2016) from GrapeTree (Zhou *et al.* 2018a).

Figure 2. Algorithmic data flow within EnteroBase. EnteroBase produces a large variety of data and graphical outputs that result from multiple calculation and graphic modules. All modules are explained in greater detail in the EnteroBase documentation (https://enterobase.readthedocs.io/en/latest//about.html). Those that are most frequently used (EBAssembly, MLSType, SNP Project and EB*Eis*) have been combined into EToKi (EnteroBase ToolKit), a standalone package for external command line usage (Fig. 3). EnteroBase also incorporates a variety of stand-alone programs, including Prokka (Seemann 2014), SISTR (Robertson *et al.* 2018), EzClermont (Waters *et al.* 2018), ClermonTyping (Beghain *et al.* 2018), FimTyper (Roer *et al.* 2017), GrapeTree (Zhou *et al.* 2018a) and JBrowse (Buels *et al.* 2016).

Figure 3. The EToKi toolkit and its primary modules. A) EBAssembly is responsible for conducting assemblies, including trimming, automatic down-sampling if necessary, remapping and polishing. It estimates the accuracy of base calls and the probability of assignment to the correct taxon. These summary statistics are presented in the Experimental Data field “Assembly stats”. The command line standalone version of EToKi offers the separate EBAssembly modules ‘prepare’ and ‘assemble’ which can be used for extracting a genomic assembly from metagenomic data (see Fig. S4). B) MLSType is responsible for MLST allele calling. EnteroBase maintains “Exemplar Alleles”, a list of exemplar allele sequences for each of the genes in the wgMLST scheme and in cgMLST, which is a subset of the wgMLST scheme. Exemplar Alleles are used to identify the ends of the corresponding loci after alignment of each genome, which are in turn used to assign existing or novel allele numbers and STs for each of the MLST schemes. C) EnteroBase offers the opportunity of creating a SNP project which consists of the SNPs within multiple assemblies that are calculated relative to the reference genome chosen by the user. This module excludes repetitive DNA regions prior to calculating the SNP matrix. The SNP matrix can be downloaded by the user and/or used to calculate an ML tree with RAxML V8 for up to 200 genomes (Stamatakis 2014), or to calculate NJ or minimal spanning trees. D) EB*Eis* is responsible for predicting *Escherichia* O serotypes based on the marker genes *wzx, wzy, wzt,* and *wzm* (Joensen *et al.* 2015; DebRoy *et al.* 2016) and H serotypes based on the flagellar *fliC* antigen (Joensen *et al.* 2015). The standalone version of EToKi is available at GitHub (https://github.com/zheminzhou/EToKi). It includes the following external programs: BBMap in BBTools (Bushnell 2016), SPADES (Bankevich *et al.* 2012), BWA (Li and Durbin 2010), Pilon (Walker *et al.* 2014), Kraken and MiniKraken database (Wood and Salzberg 2014), BLASTn (Altschul *et al.* 1990), USEARCH module UBlastP (Edgar 2010), trf (Benson 1999), pilercr (Edgar 2007), LASTAL in Last (Kielbasa *et al.* 2011), RaXML(Stamatakis 2014).

Figure 4. The hierarchical cgMLST clustering (HierCC) scheme in Enterobase. A) A screenshot of *Salmonella* cgMLST V2 plus HierCC data for five randomly selected genomes. The numbers in the columns are the cluster numbers. Cluster numbers are the smallest cgMLST ST number in single-linkage clusters of pairs of STs that are joined by up to a 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 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 an 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 5. Searching strains. A) Search strains dialog box. This dialog box enables searching for all entries that match flexible combinations of metadata (left) and experimental data (right). The dialog box lists the available fields in drop-down lists and supports the operators “contains” (text), “in” (opens a list box for pasting comma- or CRLF-delimited sets of data), “equals”, “>”, “<” and their negations. In this case, the search was designed to identify all *Salmonella* genomes where the Serovar metadata contained “Agama” or the serovar Agama had been predicted from the genomic assembly by SISTR (Robertson *et al.* 2018). B) Results of search. All entries that matched the search are shown in a spread-sheet like interface consisting of a metadata window (left|) and an experimental data window (right), which was sorted in ascending order by clicking on the column header for HC2000 at the right. The drop-down box at the upper right for *Salmonella* allows a choice between experimental data: 7-gene MLST, Assembly statistics, Annotation downloads (General Feature Format [GFF] or GenBank format [GBK]), wgMLST (Alleles and ST numbers), SISTR serotype prediction, rMLST (Alleles, ST numbers, reBG, serovar prediction), cgMLST V2 and HierCC (Alleles, ST numbers and HC clusters based on cgMLST at 13 levels ranging from HC0 to HC2850). Experimental Data also includes Custom Views and User Defined Fields. Similar experimental data are available for each of the taxa covered by EnteroBase. If the experimental data include ST designations, clicking on the GrapeTree symbol in the menu will create a GrapeTree (Zhou *et al.* 2018a) of allelic differences. The red box highlights that multiple Agama entries all share the HC2000|299 HC cluster. Right clicking on any cell within the right window gives access to a List Box offering Select All (entries in browser; indicated in second column from left), Unselect All, Find STs (with up to a user-defined number of different alleles). Get at this level (find all Strains with the same HC cluster designation or eBG/ST Complex), Download Allelic Profile (all entries in the spread-sheet) and Download Allelic Profile (Selected Only).

Figure 6. UserSpace OS. The menu item labelled Workspace (red arrow) allows access to multiple tools implemented within the UserSpace OS for working with workspaces, including Empty (contents). Save (current changes), Summary (Description and external hyperlinks), Save As, Load, Copy Selected (entries), Paste Strains (which were copied in a separate workspace), Add Selected to Workspace (selected in current workspace to a different workspace), Delete Selected (from current workspace). “Load” opens the dialog box shown on the left, which provides an overview of the user’s personal area (Mine), of workspaces shared by other users (Shared) and of publicly available workspaces (Public). The experimental data tab at the upper right shows that the Custom View “Agama View” has been chosen whereas the dialog box shows how this view was created within Custom View\New/Edit. Custom views can consist of any combination of Experimental Data fields and User Defined Fields (text). They can be saved in the user’s WorkSpace folder, and shared with other users in My Buddies (left menu). Further details can be found in EnteroBase Documentation at https://tinyurl.com/WorkSpaceToolInfo.

Figure 7. Diversity among putative Agama isolates. The results of the Search Dialog in Fig. 5 were used to create a GrapeTree depiction of a RapidNJ tree (Simonsen *et al.* 2011) of cgMLST allelic distances between putative Agama isolates. Those strains are included in a *Salmonella* workspace called 21320 - Zhou et al. All Agama Strains that was shared among the Agama Study Group. That workspace is now publicly available at Workspace\Load\Public\Zhou et al. Agama. The GrapeTree is also present in that folder as 24006 – Fig7AB. Agama and mistakes (RapidNJ) or directly at https://tinyurl.com/AgamaFig7. A) All entries color coded by Predicted Serovar (SISTR). Arrows indicate isolates whose serovar could not be predicted by SISTR. Orange shading emphasizes isolates that were wrongly predicted to be monophasic Typhimurium. Gray shading indicates isolates from serovars other than Agama, including true monophasic Typhimurium. B) All entries color-coded by HC2000 cluster. Note that all Agama entries are HC2000|299, as were the genetically related entries marked with arrows or emphasized by orange shading. In contrast, entries from other serovars that are emphasized by gray shading were in diverse HC2000 clusters whose most common serovars were the same as those predicted by SISTR. 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. 8. Scale bar: number of cgMLST allelic differences.

Figure 8. Effects of sample bias on inferred transmission chains within HC400|299 Agama isolates. A) Left: map of hosts and countries from which genomic assemblies from 149 Agama isolates were in EnteroBase by August, 2018. Right: Maximum-likelihood radial dendrogram (https://tinyurl.com/AgamaFig8A) based on RAxML (Stamatakis 2014) of 8,791 non-repetitive core SNPs as calculated by EnteroBase against reference genome 283179 (EnteroBase barcode SAL_QA0456AA). This reference genome was selected due to its high coverage, low number of contigs, and a reasonably high N50. Color-coding is according to a User-defined Field (Location & Source). Three HC100 micro-clades are visible in the tree. HC100|2433 contained all Agama from badgers. Almost all basal isolates in that micro-clade were from humans in England, suggesting that badgers were infected by Agama after a sole transmission from humans or their waste products. B) Left: map of hosts and countries from which the genomes from 64 additional Agama isolates had been uploaded by the Agama Study Group by March 2019. Right: Maximum-likelihood radial dendrogram (https://tinyurl.com/AgamaF8B) based on 9,701 SNPs from 213 isolates. Micro-clade HC100|2433 now contained multiple isolates of Agama from humans and food in France and Austria. Agama were also found in multiple livestock, companion animals and mussels in micro-clades HC100|299 and HC100|67355, which indicates that the almost exclusive association of Agama with humans and badgers in A was due to sample bias. The stars indicate the roots of multiple transmission chains of Agama into English badgers while the pink arrow indicates a potential back-transmission from badgers to a human in Bath/North East Somerset. The green arrow at 04:00 indicates four closely related Agama isolates in HC10|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. S4.

Figure 9. Maximum-Likelihood tree of modern and ancient genomes of *Y. pestis.* EnteroBase contained 1,270 ancient and modern *Yersinia pestis* genomes in February 2019, of which several hundred genomes that had been isolated in Madagascar and Brazil over a short time period 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 countries, leaving a total of 624 modern *Y. pestis* genomes. 27 ancient genomes of *Y. pestis* from existing publications (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; Damgaard *et al.* 2018) were assembled with EToKi (see Methods), resulting in a total of 651 *Y. pestis* genomes plus *Yersinia pseudotuberculosis* IP32953, as an outgroup (http://tinyurl.com/Ypestis-ancWS). The EnteroBase pipelines (Fig. 3C) were used to create a SNP project in which all genomes were aligned against CO92 (2011) using LASTAL. The SNP project identified 25,795 non-repetitive SNPs plus 8,137 short inserts/deletions over 3.9 Mbps of core genomic sites which had been called in ≥95% of the genomes. The resulting ML tree (RAxML) of the core SNPs plus all metadata is publicly available at http://tinyurl.com/Ypestis-ancSNP. 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). Note that one of the two genomes associated with the Justinianic plague (DA101) was originally described by Damgaard *et al*., 2018 as 0.PE5, whereas we assign it to 0.ANT4 due to the prior usage of this designation for *Y. pestis* from the Justinianic plague by Wagner *et al*. 2014.

Figure 10. Neighbour-Joining tree of core-genomic distances in the EcoRPlus Collection. EcoRPlus consists of 9,479 genomes, the draft genome with the greatest N50 value for each rST among the 52,876 genomes of *Escherichia* that were available within EnteroBase in August, 2018 (see http://tinyurl.com/ECOR-Plus). The population structure of these representative genomes was investigated by a neighbor-joining (RapidNJ) phylogenetic tree based on pair-wise differences in their cgMLST alleles. The nodes in this tree are color-coded by HC1100 clusters, as indicated in the Key at the bottom left. The locations of common ST Complexes is indicated at the circumference of the tree, and color-coding by ST Complex can be seen in Fig. S5. HC1100 clustering is largely congruent with assignments to common ST Complexes, except that HC1100|13 includes ST10 Complex and ST168 Complex. Still other discrepancies exist among the smaller, unlabeled populations. Color-coding by Clermont typing can be seen in Fig. S7 which highlights multiple discrepancies between taxonomical designations and population groupings. 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 showing estimates of branch lengths can be found in Fig. S6.

Supplementary Figure 1. Overview of workflow within EnteroBase. EnteroBase offers its users interactions *via* a web browser front end or a RESTful API (Application Programming Interface), both of which connect to a Python Flask web framework environment that handles all further interactions with structured data (PostgreSQL databases) and stored files in disk storage *via* Central Control modules. These central control modules are also responsible for interacting *via* JSON strings with the Calculation Robot and Nomenclature Server, which in turn also interact with external database environments. Further details on the API commands can be found at http://tinyurl.com/EnteroBaseAPIDocs.

Supplementary Figure 2. Uberstrains and Sub-strains. A) The Search Dialog in default mode only retrieves Uberstrains, and does not retrieve any associated sub-strains. Uberstrains are indicated by a black square to the left of the Uberstrain barcode designation at the left. The example shows two distinct variants of *Y. pestis* CO92, one which was sequenced in 2001 and a second sequenced in 2015, in which 13 erroneous SNP calls have been corrected. B) Showing Sub-strains. Even when the workspace contains sub-strains, it does not show them as a default behaviour. When the checkbox Show Sub Strains is clicked in the search dialog, the browser shows a triangle at the far left of Uberstrains that contain one or more Sub-Strains. Clicking on the triangle opens a previously hidden tree-like hierarchy containing all its sub-strains. View\Show All Sub-strains in the top browser Menu opens such hierarchies for all Uberstrains in the browser window, and clicking on View\Close all Sub-strains reverts to showing only Uberstrains.

Supplementary Figure 3. Phylodynamics of isolates from badgers. A GrapeTree of Agama isolates was used to transfer individual subtrees plus their GPS coordinates and metadata to MicroReact (Argimon *et al.* 2016) as described in the GrapeTree Reference Manual (http://tinyurl.com/GrapeTreeRefManual). The trees at the left side are phylogenetic trees drawn by MicroReact and the maps at the right side are geographic locations within MicroReact, except that in part B they were overlaid by the idealized spatial distributions of badger social groups and setts as elucidated by (McDonald *et al.* 2018). (A, B) Phylogenetic tree and geographical map of 64 Agama isolates from badgers in Woodchester Park that were collected in 2006-2007. Four HierCC HC10 clusters (colored ovals in part B) of genetically related genomes which differ by <11 cgMLST alleles were isolated from neighbouring social groups, and are inferred to have moved by local transmission chain *a, b, c, d*. An interactive version of the MicroReact project can also be found at https://microreact.org/project/t7qlSSslh/3e634888. (C, D) Phylogenetic tree and geographical map of 75 Agama isolates from badgers in Woodchester Park and elsewhere in England in HC100|2433 that were collected between 1998 and 2010. An interactive version of the MicroReact project can also be found at https://microreact.org/project/9XUC7i-Fm/fed65ff5. (E, F) Phylogenetic tree and geographical map of 6 Agama isolates from badgers in England in HC100|299 that were collected between 2009 and 2016. An interactive version of the MicroReact project can also be found at https://microreact.org/project/XaJm1cNjY/69748fe3.

Supplementary Figure 4. Extracting aDNA assemblies from metagenomic sequences with the EBAssembly module of EToKi. EBAssembly includes functions for extracting genome-specific reads from metagenomic sequences which are only accessible in the stand-alone, command-line version of EToKi. The EToKi prepare module can collapse paired-end reads and trim both paired-end and single-end reads without down-sampling. As described in the documentation (https://github.com/zheminzhou/EToKi), the EToKi assemble module incorporates elements from SPARSE (Zhou *et al.* 2018b) to identify genome-specific short reads within metagenomic sequences after specifying a reference genome sequence, an in-group of related genomes and a related but distinct out-group of genomes . The module replaces nucleotides in the reference genome by their calculated SNVs after checking that they are supported by at least 3 metagenomic reads, and the supporting read frequencies occur with at least one-third of the average read depth. It also allows constraining SNP calls to (published) SNPS within a text file, and save the modified sequence of the reference genome in a form which can be uploaded to EnteroBase by admins and curators.

Supplementary Figure 5. A Neighbour-Joining tree of cgMLST distances in the EcoRPlus Collection color-coded by ST Complex. This tree is identical to the tree in Fig. 10 and Fig. S7 except that the nodes are colored according to the ST Complex for legacy 7-gene MLST (Wirth *et al.* 2006). The correspondence between ST Complex and HC1199 clustering (which is based on much higher resolution cgMLST) is striking for the most common ST Complexes, with the exception of ST168 Complex (07:00) which is assigned to HC1100|13 by Hierarchical Clustering. The discrete nature of multiple ST Complexes is also noteworthy, and prominent examples of such discrete Complexes are indicated by text. Recent publications have provided interesting details on the ST131 Complex (usually erroneously referred to as ST131) (Johnson *et al.* 2016; Stoesser *et al.* 2016; Ben Zakour *et al.* 2016; Liu *et al.* 2018), the ST95 Complex (Achtman *et al.* 1983; Wirth *et al.* 2006; Gordon *et al.* 2017) and ST11 Complex/O157:H7 (Leopold *et al.* 2009; Dallman *et al.* 2015). However, little attention has been directed at the others although they are a very common cause of disease in humans and animals according to the frequencies of their genomes in EnteroBase.

Supplementary Figure 6. A Maximum-Likelihood (ML) tree of the EcoRPlus Collection. 1,230,995 core SNPs were extracted from the concatenated sequences (2.33 Mbps) of the 2,513 core gene alignments (MAFFT (Katoh and Standley 2013)) from 9,479 core genomes. A maximum likelihood tree was calculated using FASTTREE 2 (Price *et al.* 2010). Inset) Tree of all genomes color-coded by ClermonTyping, including the Cryptic Clades I-V and the *Escherichia* species *albertii*, *fergusonii*, and *marmotae*. Note that the three genomes of *E. marmotae* are on a deep branch within Clade V. The white circle encloses genetically-related populations within *E. coli*, including Clade I, whereas the other Clades and species are on external branches within the gray rectangle. These topological relationships are similar to those described earlier on smaller datasets (Luo *et al.* 2011) except that *E. marmotae* was not included. Main figure) Closeup of genomes on branches within the inner circle of the inset, color-coded by HC1100 HierCC cluster. This ML clustering of individual genomes is concordant with the clustering according to the neighbour-joining algorithm in Fig. 10, while providing much more accurate branch lengths. However, this tree also took several weeks to complete.

Supplementary Figure 7. A Neighbour-Joining tree of cgMLST distances in the EcoRPlus Collection color-coded by Clermont types. This tree is identical to the tree in Fig. 10 and Fig. S5 except that the nodes show the Clermont Types predicted by the program ClermonTyping (Beghain *et al.* 2018), which has been implemented within EnteroBase. Large parts of the tree are relatively homogeneous, indicating that Clermont Typing often correlates well with HC2000 clustering. However, arrows indicate multiple nodes which differ in Clermont Type from their close neighbors, illustrating that the presence/absence of genes from the accessory genome which is used for the Clermont scheme does not correlate completely with the phylogenetic relationships revealed by cgMLST. As a result, nodes assigned to Clermont Types A and B2 are found at multiple positions within the tree, far from most other strains of those Clermont types. In addition, two groups of *Shigella* are inaccurately labelled by Clermont Types. ST245 Complex largely corresponds to *Shigella flexneri* (Wirth *et al.* 2006), but is inappropriately assigned to Clermont Type A. Similarly, ST152 Complex largely corresponds to *Shigella sonnei* but is not recognized by Clermont Typing.

Table 1. Basic statistics on EnteroBase (https://enterobase.warwick.ac.uk) (April, 2019)

Genus	Legacy MLST	Assembled genomes (user uploads)	wgMLST (Loci)	cgMLST (Loci)	rMLST (Loci)	MLST (Loci)	HierCC (cgMLST)
Salmonella	5,555	198,183 (19,638)	21,065	3,002	51	7	√
Escherichia/ Shigella	9,614	93,747 (7,155)	25,002	2,512	51	7	√
Clostrioides		13,721 (1,381)	11,490	2,556	53	7	√
Vibrio		6,760 (79)			51		
Yersinia	881	2,649 (949)	19,531	1,553	51	7	√
Helicobacter		2,427 (831)			53		
Moraxella	416	636 (283)			52	8	
Total	16,466	318,123 (30,319)					

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

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

Reference List

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

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

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

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

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

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

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

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

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**: 455-477.

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

Ben Zakour NL, Alsheikh-Hussain AS, Ashcroft MM, Khanh Nhu NT, Roberts LW, Stanton-Cook M, Schembri MA, Beatson SA. 2016. Sequential acquisition of virulence and fluoroquinolone resistance has shaped the evolution of *Escherichia coli* ST131. *MBio* **7**: e00347-16.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573-580.

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

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

Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elsik CG, Lewis SE, Stein L, et al. 2016. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol* **17**: 66.

Bushnell B. BBMap short read aligner. 2016. *URL*

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

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

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

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

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

Dallman TJ, Ashton PM, Byrne L, Perry NT, Petrovska L, Ellis R, Allison L, Hanson M, Holmes A, Gunn GJ, et al. 2015. Applying phylogenomics to understand the emergence of Shiga-toxin-producing *Escherichia coli* O157:H7 strains causing severe human disease in the UK. *Microb Genom* **1**: e000029.

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

DebRoy C, Fratamico PM, Yan X, Baranzoni G, Liu Y, Needleman DS, Tebbs R, O'Connell CD, Allred A, Swimley M, et al. 2016. Correction: Comparison of O-Antigen Gene Clusters of All O-Serogroups of Escherichia coli and Proposal for Adopting a New Nomenclature for O-Typing. *PLoS ONE* **11**: e0154551.

Diemert S and Yan T. 2019. Comparative genomic analyses of municipal wastewater *Salmonella* isolates detects a clinically unreported salmonellosis outbreak. *Appl Environ Microbiol*.

Edgar RC. 2007. PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* **8**: 18.

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

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

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

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

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

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

Gordon DM, Geyik S, Clermont O, O'Brien CL, Huang S, Abayasekara C, Rajesh A, Kennedy K, Collignon P, Pavli P, 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**.

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

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

Guibourdenche M, Roggentin P, Mikoleit M, Fields PI, Bockemuhl J, Grimont PA, Weill FX. 2010. Supplement 2003-2007 (No. 47) to the White-Kauffmann-Le Minor scheme. *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.

Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. 2015. Rapid and easy *in silico* serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. *J Clin Microbiol* **53**: 2410-2426.

Johnson JR, Davis G, Clabots C, Johnston BD, Porter S, DebRoy C, Pomputius W, Ender PT, Cooperstock M, Slater BS, et al. 2016. Household clustering of *Escherichia coli* sequence type 131 clinical and fecal isolates according to whole genome sequence analysis. *Open Forum Infect Dis* **3**: ofw129.

Johnson SL, Daligault HE, Davenport KW, Jaissle J, Frey KG, Ladner JT, Broomall SM, Bishop-Lilly KA, Bruce DC, Coyne SR, et al. 2015. Thirty-two complete genome assemblies of nine *Yersinia* species, Including *Y. pestis*, *Y. pseudotuberculosis*, and *Y. enterocolitica*. *Genome Announc* **3**.

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, 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. 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.

Katoh K and Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772-780.

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.

Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Res* **21**: 487-493.

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.

Leopold SR, Magrini V, Holt NJ, Shaikh N, Mardis ER, Cagno J, Ogura Y, Iguchi A, Hayashi T, Mellmann A, et al. 2009. A precise reconstruction of the emergence and constrained radiations of *Escherichia coli* O157 portrayed by backbone concatenomic analysis. *Proc Natl Acad Sci USA* **106**: 8713-8718.

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

Li H and Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589-595.

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 K, Linder CR, Warnow T. 2011. RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. *PLoS ONE* **6**: e27731.

Loper E, Bird S. NLTK: the Natural Language Toolkit. 2002. Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. **1:** 63-70. Philadelphia, Pennsylvania, Association for Computational Linguistics. ETMTNLP '02.

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.

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.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**: e9490.

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.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**: 2068-2069.

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, 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.

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

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

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

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

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

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

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

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

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

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

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

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

Wood DE and Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15**: R46.

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

Wray C, Baker K, Gallagher J, Naylor P. 1977. *Salmonella* infection in badgers in the South West of 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.

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

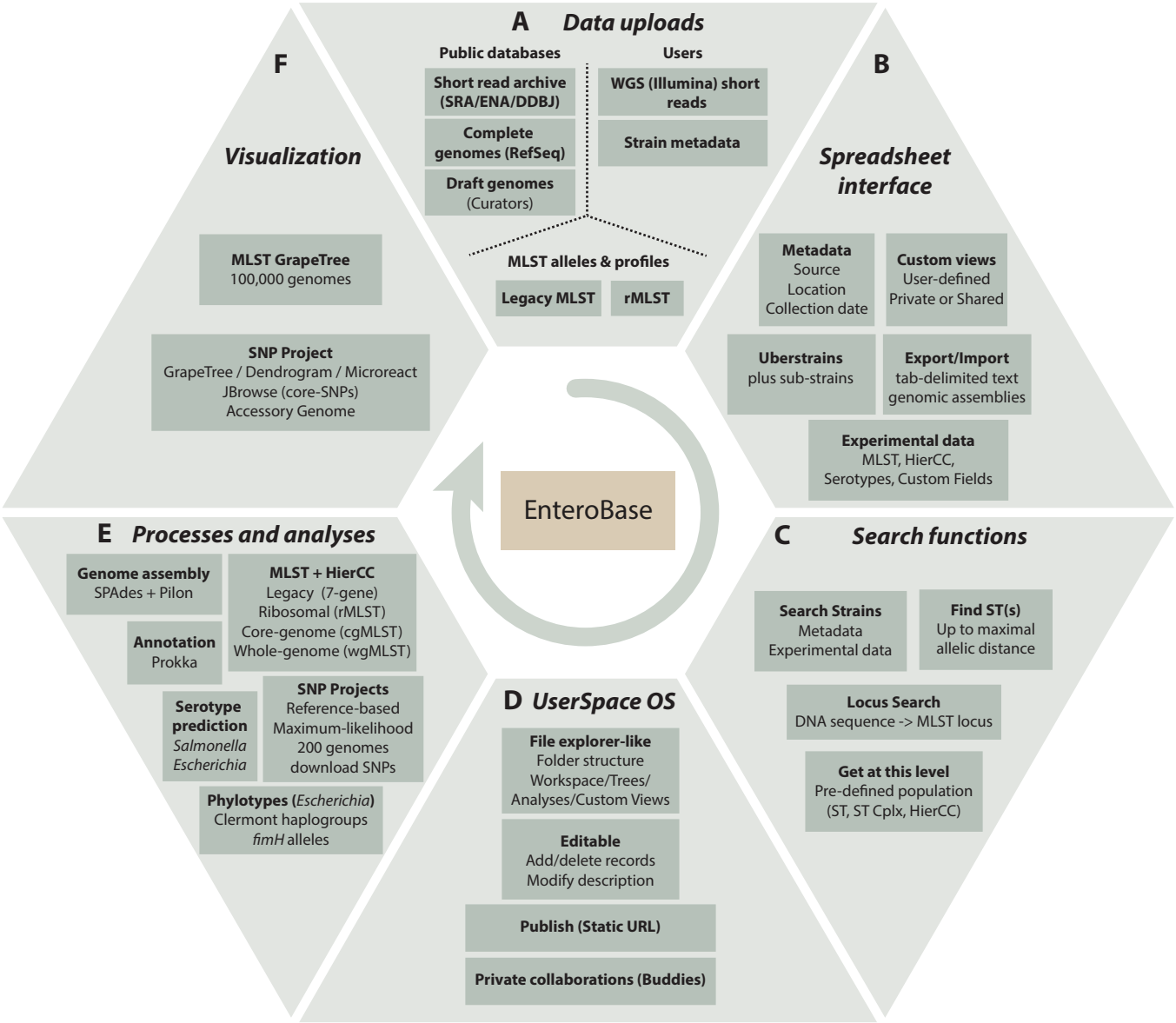


Figure 1. Overview of EnteroBase Features. A) Data uploads come from public databases, user uploads and existing legacy and rMLST databases at PubMLST (https://pubmlst.org/). B) EnteroBase offers a spreadsheet-like interface to sets of strains (one Uberstrain plus any number of sub-strains) each containing metadata, and their associated experimental data and custom views. Post-release data can be exported (downloaded) as genome assemblies or tab-delimited text files and tab-delimited text-files can be used to import (upload) metadata for entries to which the user has editing rights. C) The search strains dialog box supports flexible combinations (and/or) of metadata and experimental data to load matching entries into the spreadsheet. Find ST(s) retrieves STs that user-defined, minimal number of identical alleles with a specified ST. Locus search identifies the locus designation in any MLST scheme in EnteroBase according to searches for a given sequence with BLASTn (Altschul *et al*. 1990) and UBlastP in USEARCH (Edgar 2010). D). UserSpace OS. This interface is used for file-explorer like manipulations of workspaces, trees, SNP projects and custom views. By default, all of these objects are private to the user who has created them, but they can be shared with buddies or made globally accessible. E) Processes and analyses which are implemented within EnteroBase, including the external programs Prokka (Seemann 2014), SPADES (Bankevich *et al*. 2012), and Pilon (Walker *et al.* 2014). F) Visualization tools, including transferring data to external websites such as MicroReact (Argimon *et al.* 2016) from GrapeTree (Zhou *et al.* 2018a).

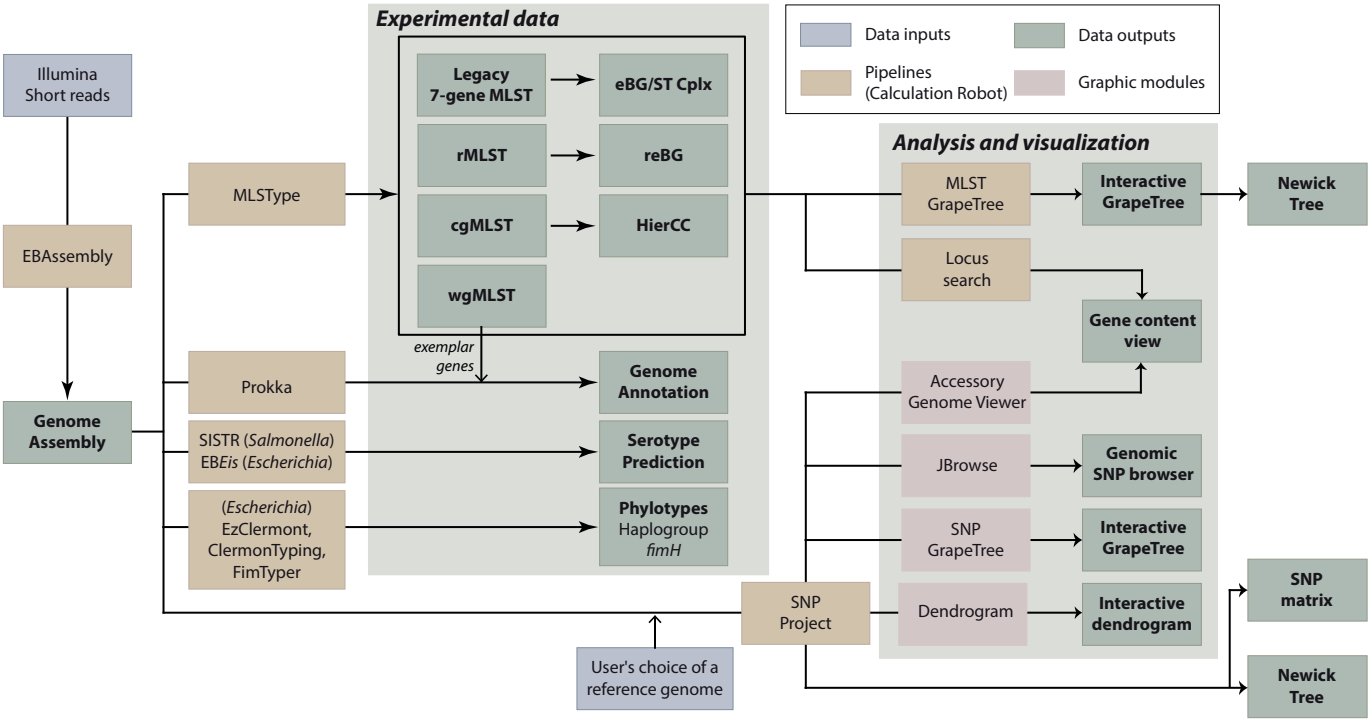


Figure 2. Algorithmic data flow within EnteroBase. EnteroBase produces a large variety of data and graphical outputs that result from multiple calculation and graphic modules. All modules are explained in greater detail in the EnteroBase documentation (https://enterobase.readthedocs.io/en/latest//about.html). Those that are most frequently used (EBAssembly, MLSType, SNP Project and EB*Eis*) have been combined into EToKi (EnteroBase ToolKit), a stand-alone package for external command line usage (Fig. 3). EnteroBase also incorporates a variety of stand-alone programs, including Prokka (Seemann 2014), SISTR (Robertson *et al.* 2018), EzClermont (Waters *et al.* 2018), ClermonTyping (Beghain *et al. 2018*), FimTyper (Roer *et al.* 2017), GrapeTree (Zhou et al. 2018a) and JBrowse (Buels *et al.* 2016).

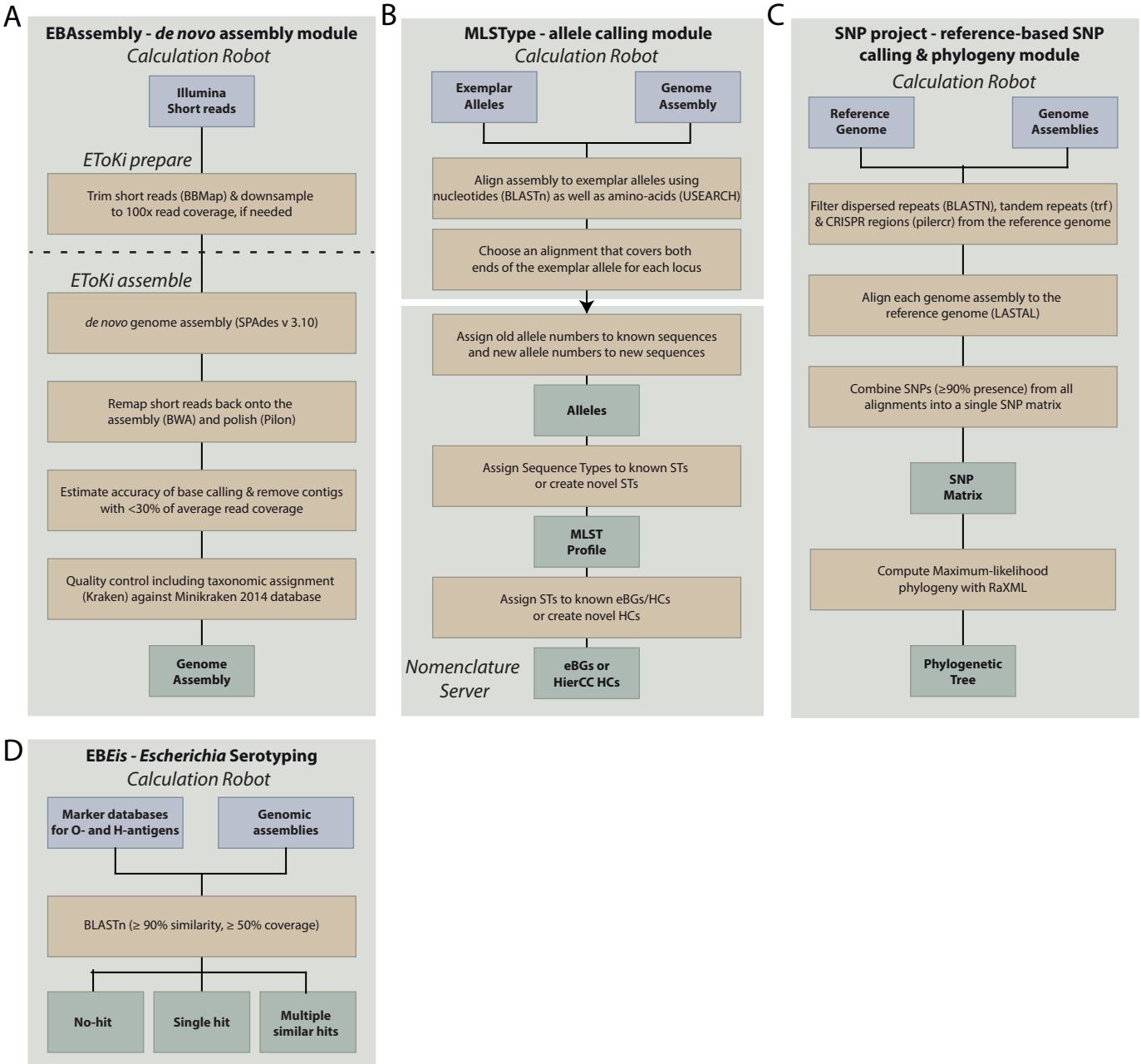
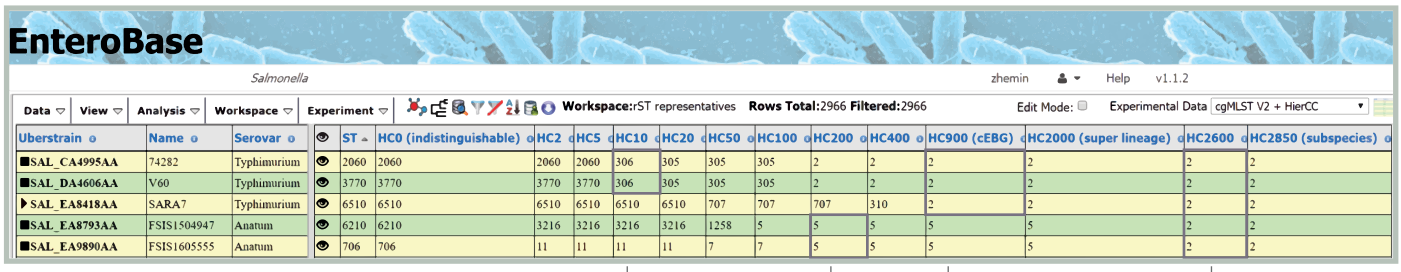


Figure 3. The EToKi toolkit and its primary modules. A) EBAssembly is responsible for conducting assemblies, including trimming, automatic down-sampling if necessary, remapping and polishing. It estimates the accuracy of base calls and the probability of assignment to the correct taxon. These summary statistics are presented in the Experimental Data field "Assembly stats". The command line standalone version of EToKi offers the separate EBAssembly modules 'prepare' and 'assemble' which can be used for extracting a genomic assembly from metagenomic data (see Fig. S4). B) MLSType is responsible for MLST allele calling. EnteroBase maintains "Exemplar Alleles", a list of exemplar allele sequences for each of the genes in the wgMLST scheme and in cgMLST, which is a subset of the wgMLST scheme. Exemplar Alleles are used to identify the ends of the corresponding loci after alignment of each genome, which are in turn used to assign existing or novel allele numbers and STs for each of the MLST schemes. C) EnteroBase offers the opportunity of creating a SNP project which consists of the SNPs within multiple assemblies that are calculated relative to the reference genome chosen by the user. This module excludes repetitive DNA regions prior to calculating the SNP matrix. The SNP matrix can be downloaded by the user and/or used to calculate an ML tree with RAxML V8 for up to 200 genomes (Stamatakis 2014), or to calculate NJ or minimal spanning trees. D) EB*Eis* is responsible for predicting *Escherichia* O serotypes based on the marker genes *wzx, wzy, wzt* and *wzm* (Joensen *et al*. 2015; DebRoy *et al.* 2016) and H serotypes based on the flagellar *fliC* antigen (Joensen et al. 2015). The standalone version of EToKi is available at GitHub (https://github.com/zheminzhou/EToKi). It includes the following external programs: BBMap in BBTools (Bushnell 2016), SPADES (Bankevich *et al.* 2012), BWA (Li and Durbin 2010), Pilon (Walker *et al.* 2014), Kraken and MiniKraken database (Wood and Salzberg 2014), BLASTn (Altschul *et al.* 1990), USEARCH module UBlastP (Edgar 2010), trf (Benson 1999), pilercr (Edgar 2007), LASTAL in Last (Kielbasa *et al.* 2011), RaXML(Stamatakis 2014).

A



Uberstrain	Name	Serovar	ST	HC0 (indistinguishable)	HC2	HC5	HC10	HC20	HC50	HC100	HC200	HC400	HC900 (cEBG)	HC2000 (super lineage)	HC2600	HC2850 (subspecies)
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	1258	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

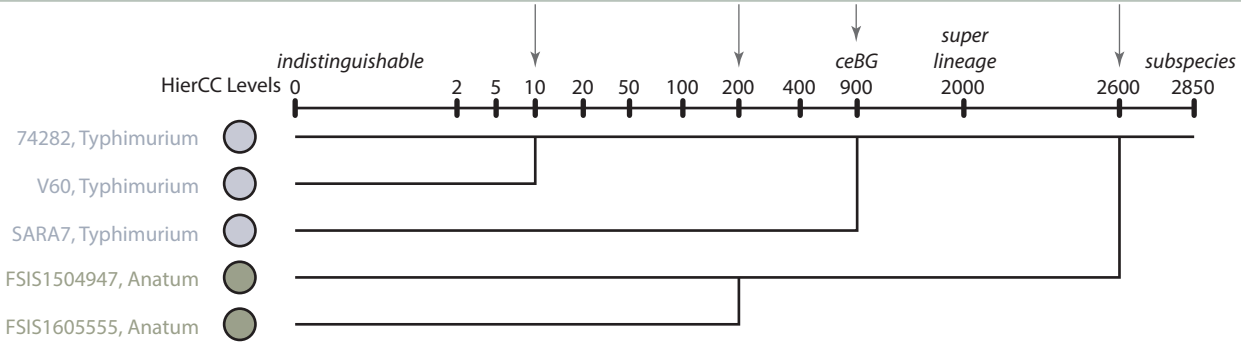


Figure 4. The hierarchical cgMLST clustering (HierCC) scheme in Enterobase. A) A screenshot of *Salmonella* cgMLST V2 plus HierCC data for five randomly selected genomes. The numbers in the columns are the cluster numbers. Cluster numbers are the smallest cgMLST ST number in single-linkage clusters of pairs of STs that are joined by up to a 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 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 an 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.

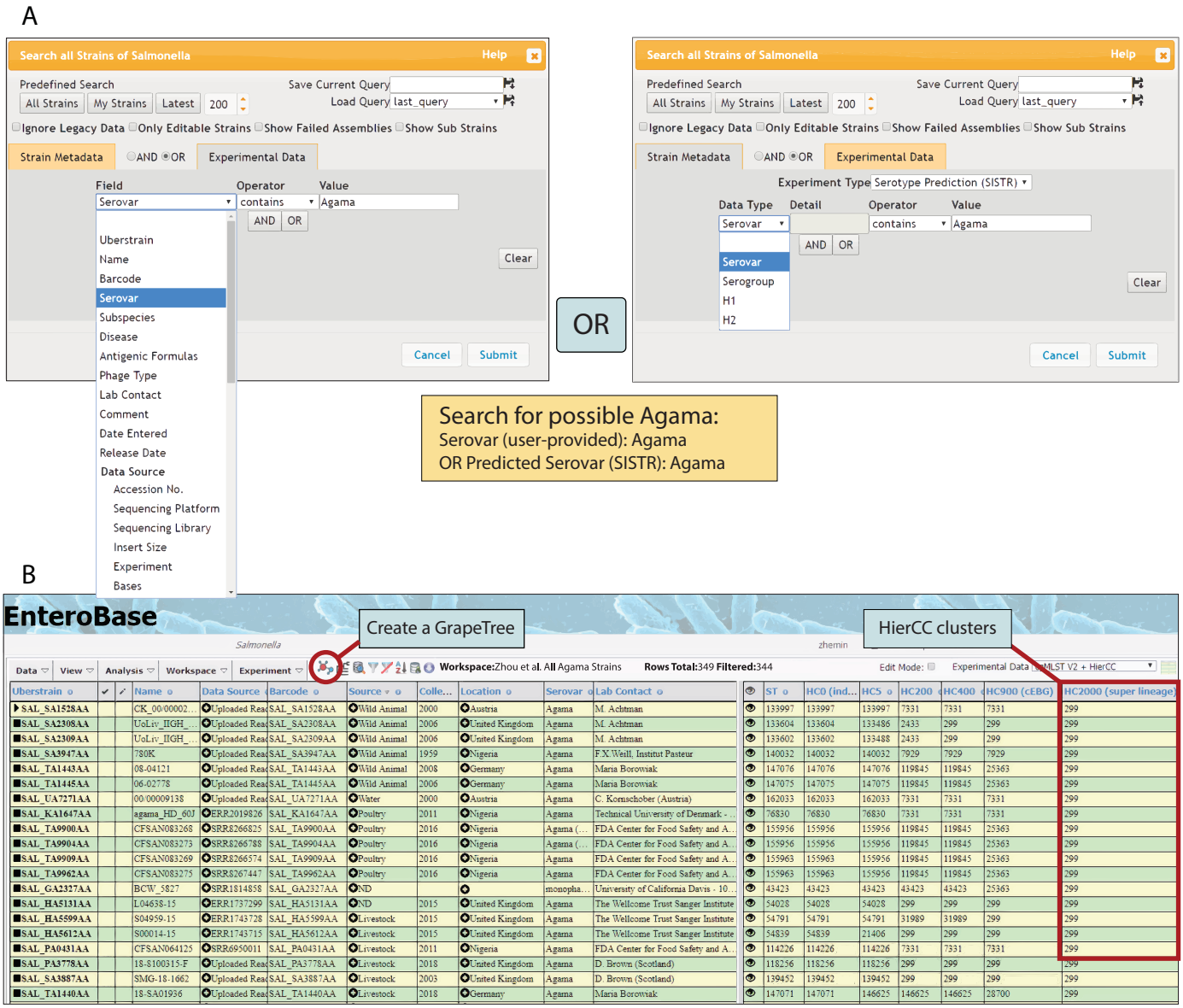


Figure 5. Searching strains. A) Search strains dialog box. This dialog box enables searching for all entries that match flexible combinations of metadata (left) and experimental data (right). The dialog box lists the available fields in drop-down lists and supports the operators “contains” (text), “in” (opens a list box for pasting comma- or CRLF-delimited sets of data), “equals”, “>”, “<” and their negations. In this case, the search was designed to identify all Salmonella genomes where the Serovar metadata contained “Agama” or the serovar Agama had been predicted from the genomic assembly by SISTR (Robertson *et al.* 2018). B) Results of search. All entries that matched the search are shown in a spread-sheet like interface consisting of a metadata window (left|) and an experimental data window (right), which was sorted in ascending order by clicking on the column header for HC2000 at the right. The drop-down box at the upper right for Salmonella allows a choice between experimental data: 7-gene MLST, Assembly statistics, Annotation downloads (General Feature Format [GFF] or GenBank format [GBK]), wgMLST (Alleles and ST numbers), SISTR serotype prediction, rMLST (Alleles, ST numbers, reBG, serovar prediction), cgMLST V2 and HierCC (Alleles, ST numbers and HC clusters based on cgMLST at 13 levels ranging from HC0 to HC2850). Experimental Data also includes Custom Views and User Defined Fields. Similar experimental data are available for each of the taxa covered by EnteroBase. If the experimental data include ST designations, clicking on the GrapeTree symbol in the menu will create a GrapeTree (Zhou *et al.* 2018a) of allelic differences. The red box highlights that multiple Agama entries all share the HC2000|299 HC cluster. Right clicking on any cell within the right window gives access to a List Box offering Select All (entries in browser; indicated in second column from left), Unselect All, Find STs (with up to a user-defined number of different alleles). Get at this level (find all Strains with the same HC cluster designation or eBG/ST Complex), Download Allelic Profile (all entries in the spread-sheet) and Download Allelic Profile (Selected Only).

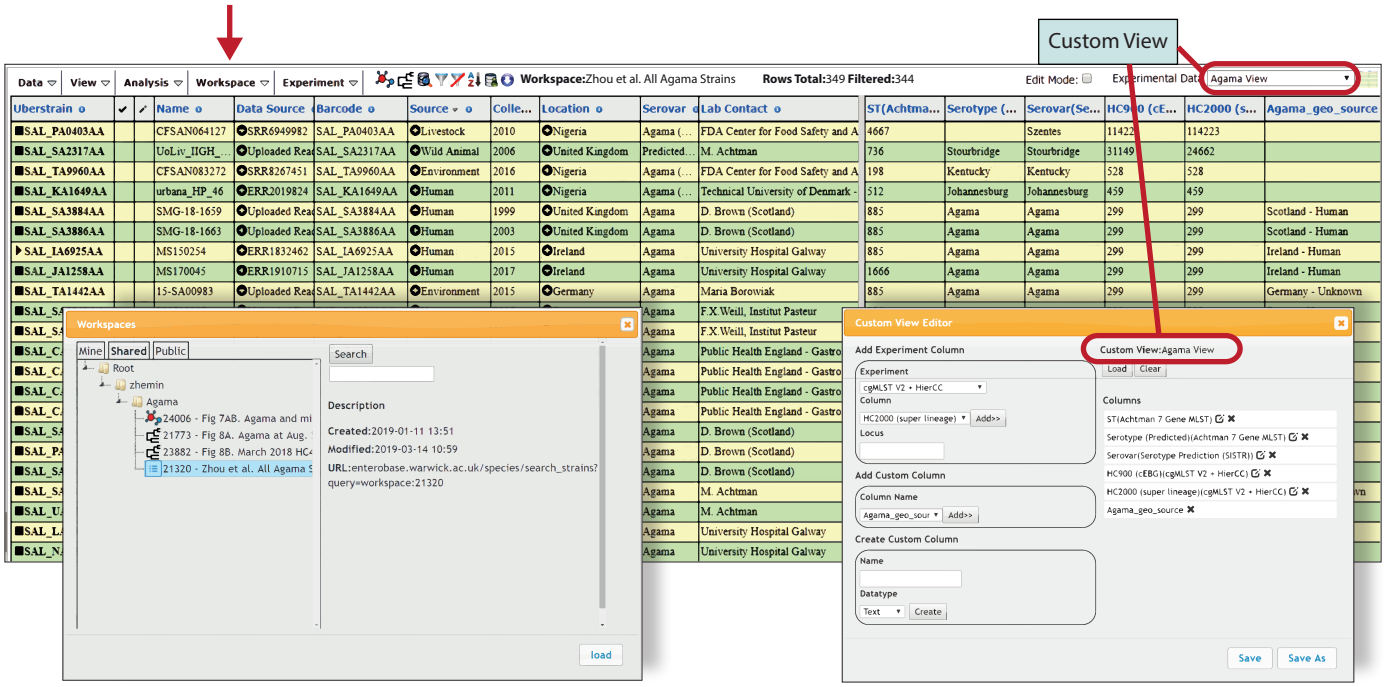


Figure 6. UserSpace OS. The menu item labelled Workspace (red arrow) allows access to multiple tools implemented within the UserSpace OS for working with workspaces, including Empty (contents). Save (current changes), Summary (Description and external hyperlinks), Save As, Load, Copy Selected (entries), Paste Strains (which were copied in a separate workspace), Add Selected to Workspace (selected in current workspace to a different workspace), Delete Selected (from current workspace). “Load” opens the dialog box shown on the left, which provides an overview of the user’s personal area (Mine), of workspaces shared by other users (Shared) and of publicly available workspaces (Public). The experimental data tab at the upper right shows that the Custom View “Agama View” has been chosen whereas the dialog box shows how this view was created within Custom View\New/Edit. Custom views can consist of any combination of Experimental Data fields and User Defined Fields (text). They can be saved in the user’s WorkSpace folder, and shared with other users in My Buddies (left menu). Further details can be found in EnteroBase Documentation at https://tinyurl.com/WorkSpaceToolInfo

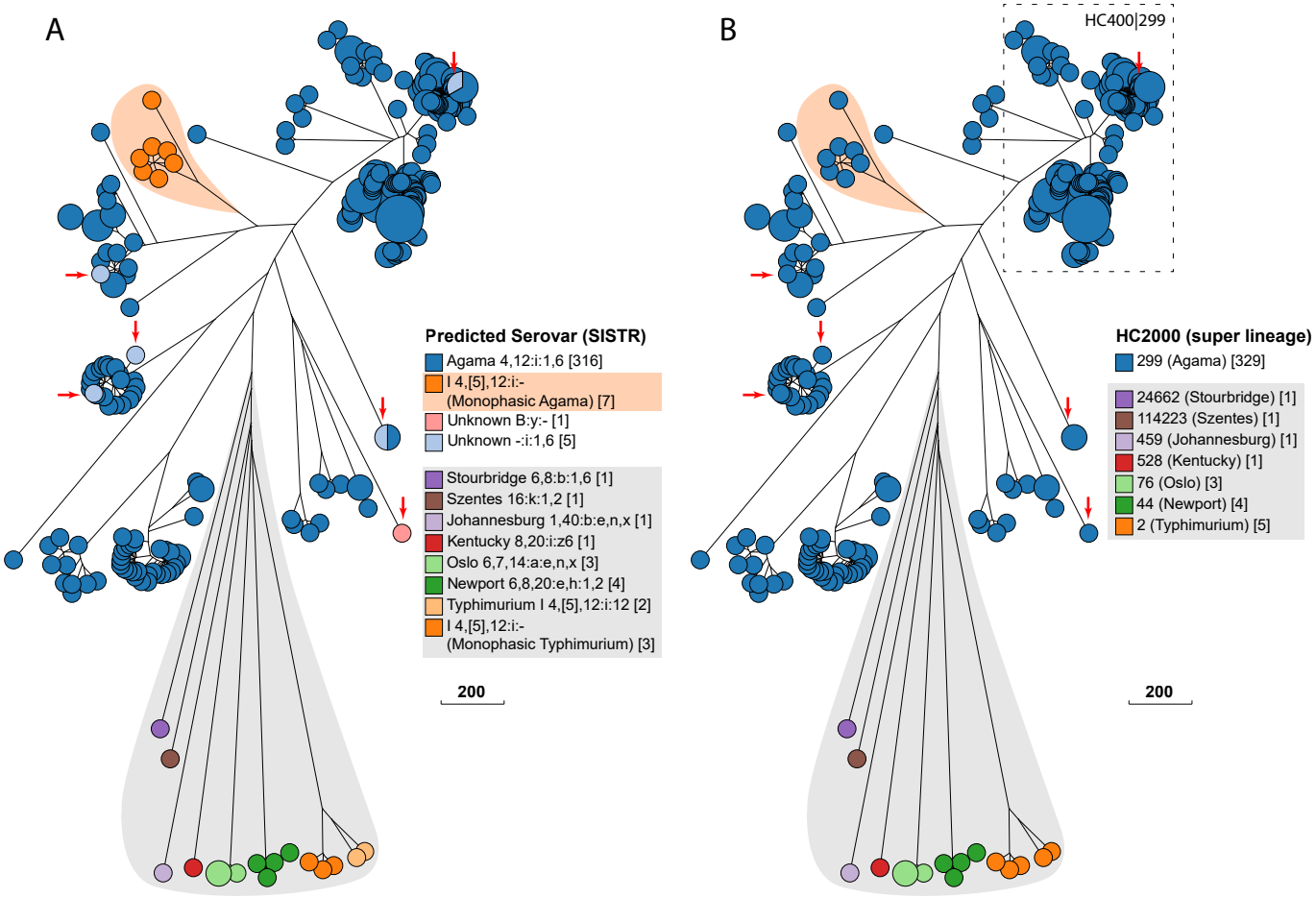


Figure 7. Diversity among putative Agama isolates. The results of the Search Dialog in Fig. 5 were used to create a GrapeTree depiction of a RapidNJ tree (Simonsen *et al.* 2011) of cgMLST allelic distances between putative Agama isolates. Those strains are included in a *Salmonella* workspace called 21320 - Zhou et al. All Agama Strains that was shared among the Agama Study Group. That workspace is now publicly available at Workspace\Load\Public\Zhou et al. Agama. The GrapeTree is also present in that folder as 24006 – Fig7AB. Agama and mistakes (RapidNJ) or directly at https://tinyurl.com/AgamaFig7. A) All entries color coded by Predicted Serovar (SISTR). Arrows indicate isolates whose serovar could not be predicted by SISTR. Orange shading emphasizes isolates that were wrongly predicted to be monophasic Typhimurium. Gray shading indicates isolates from serovars other than Agama, including true monophasic Typhimurium. B) All entries color-coded by HC2000 cluster. Note that all Agama entries are HC2000|299, as were the genetically related entries marked with arrows or emphasized by orange shading. In contrast, entries from other serovars that are emphasized by gray shading were in diverse HC2000 clusters whose most common serovars were the same as those predicted by SISTR. 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. 8. Scale bar: number of cgMLST allelic differences.

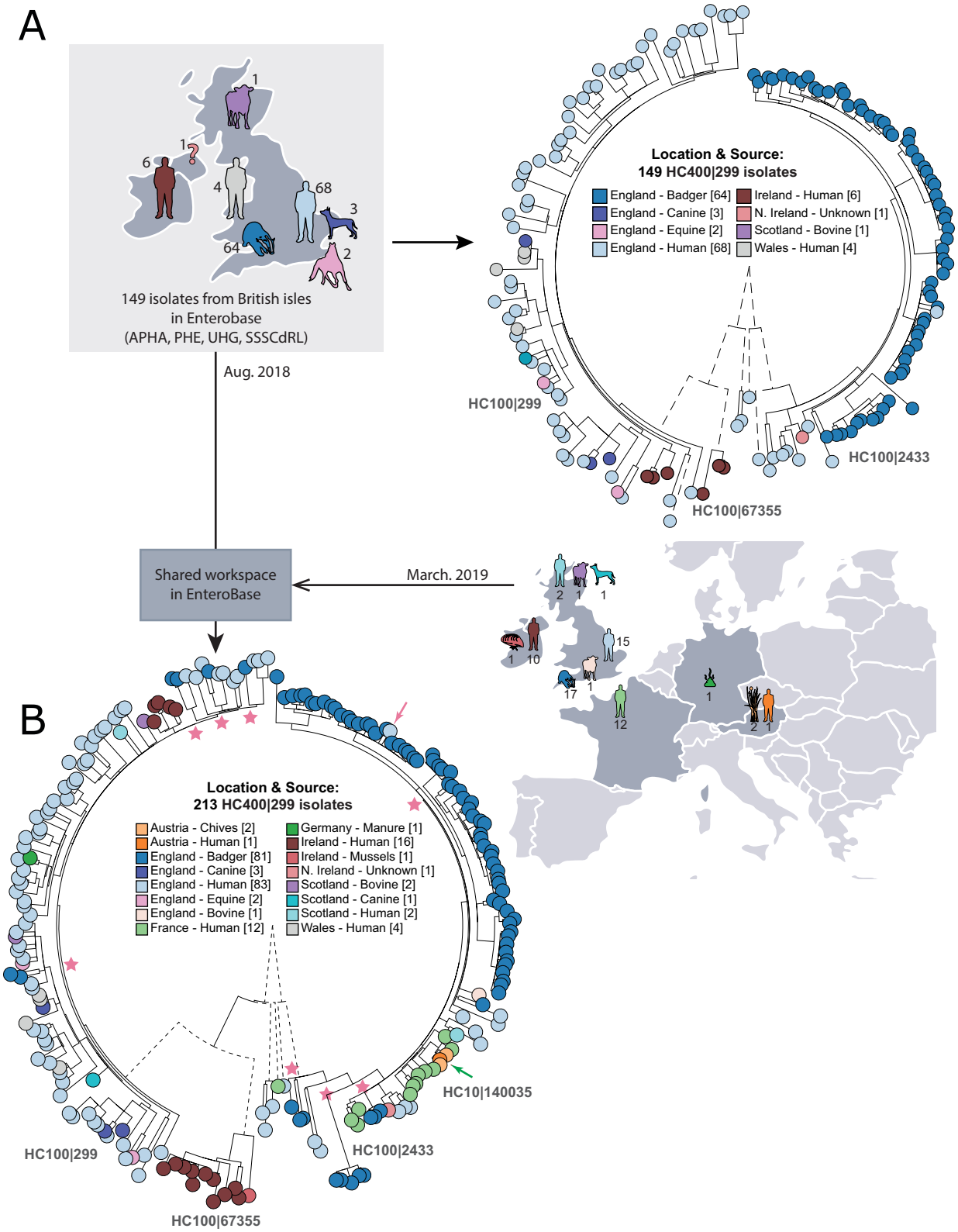


Figure 8. Effects of sample bias on inferred transmission chains within HC400|299 Agama isolates. A) Left: map of hosts and countries from which genomic assemblies from 149 Agama isolates were in EnteroBase by August, 2018. Right: Maximum-likelihood radial dendrogram (https://tinyurl.com/AgamaFig8A) based on RAxML (Stamatakis 2014) of 8,791 non-repetitive core SNPs as calculated by EnteroBase against reference genome 283179 (EnteroBase barcode SAL_QA0456AA). This reference genome was selected due to its high coverage, low number of contigs, and a reasonably high N50. Color-coding is according to a User-defined Field (Location & Source). Three HC100 micro-clades are visible in the tree. HC100|2433 contained all Agama from badgers. Almost all basal isolates in that micro-clade were from humans in England, suggesting that badgers were infected by Agama after a sole transmission from humans or their waste products. B) Left: map of hosts and countries from which the genomes from 64 additional Agama isolates had been uploaded by the Agama Study Group by March 2019. Right: Maximum-likelihood radial dendrogram (https://tinyurl.com/AgamaF8B) based on 9,701 SNPs from 213 isolates. Micro-clade HC100|2433 now contained multiple isolates of Agama from humans and food in France and Austria. Agama were also found in multiple livestock, companion animals and mussels in micro-clades HC100|299 and HC100|67355, which indicates that the almost exclusive association of Agama with humans and badgers in A was due to sample bias. The stars indicate the roots of multiple transmission chains of Agama into English badgers while the pink arrow indicates a potential back-transmission from badgers to a human in Bath/North East Somerset. The green arrow at 04:00 indicates four closely related Agama isolates in HC10|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. S4.

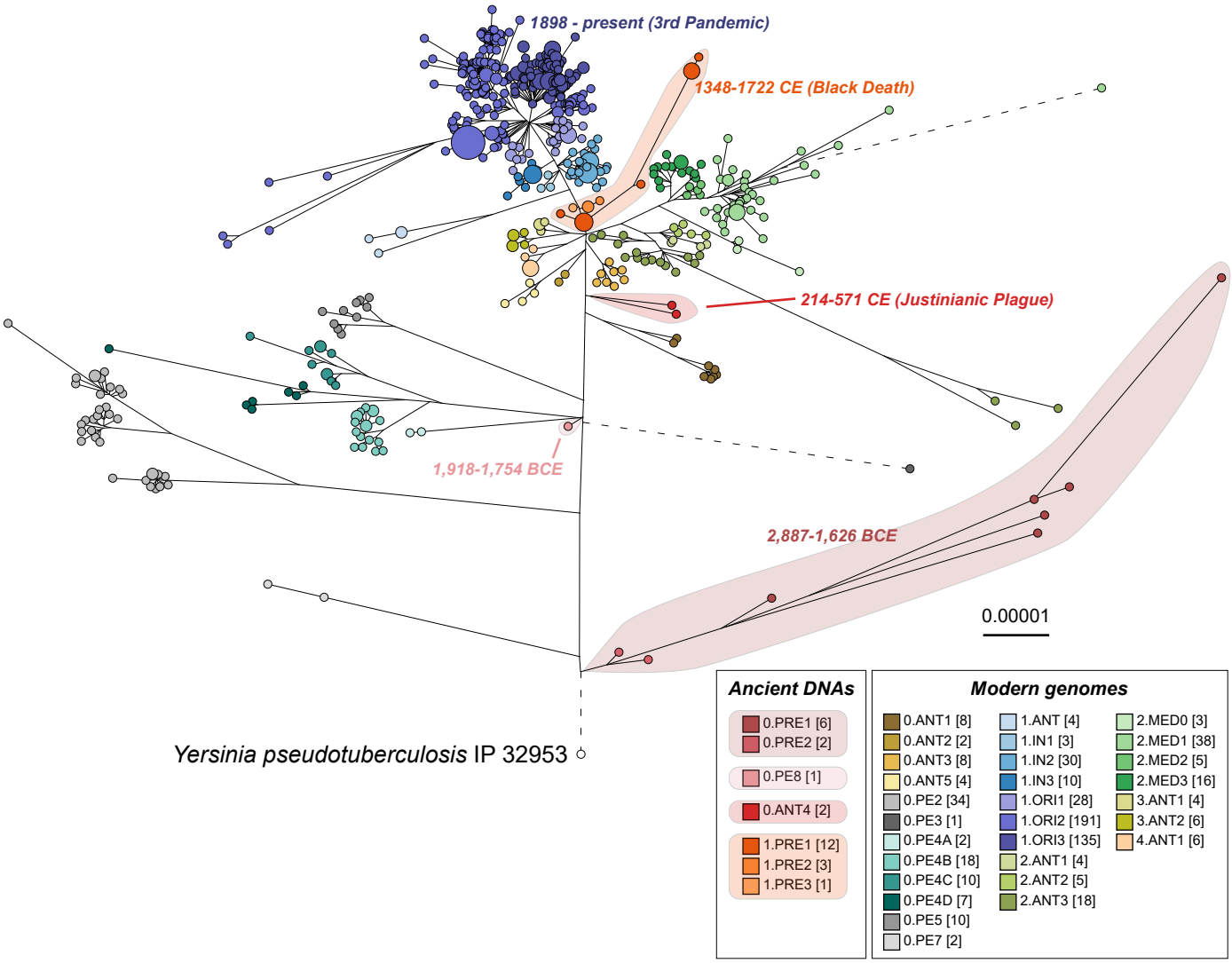


Figure 9. Maximum-Likelihood tree of modern and ancient genomes of *Y. pestis*. EnteroBase contained 1,270 ancient and modern *Yersinia pestis* genomes in February 2019, of which several hundred genomes that had been isolated in Madagascar and Brazil over a short time period demonstrated very low levels 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 countries, leaving a total of 624 modern *Y. pestis* genomes. 27 ancient genomes of *Y. pestis* from existing publications (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; Damgaard *et al.* 2018) were assembled with EToKi (see Methods), resulting in a total of 651 *Y. pestis* genomes plus *Yersinia pseudotuberculosis* IP32953, as an outgroup (http://tinyurl.com/Ypestis-ancWS). The EnteroBase pipelines (Fig. 3C) were used to create a SNP project in which all genomes were aligned against CO92 (2011) using LASTAL. The SNP project identified 25,795 non-repetitive SNPs plus 8,137 short inserts/deletions over 3.9 Mbps of core genomic sites which had been called in ≥95% of the genomes. The resulting ML tree (RAxML) of the core SNPs plus all metadata is publicly available at http://tinyurl.com/Ypestis-ancSNP. 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). Note that one of the two genomes associated with the Justinianic plague (DA101) was originally described by Damgaard et al., 2018 as 0.PE5, whereas we assign it to 0.ANT4 due to the prior usage of this designation for *Y. pestis* from the Justinianic plague by Wagner *et al.* 2014.

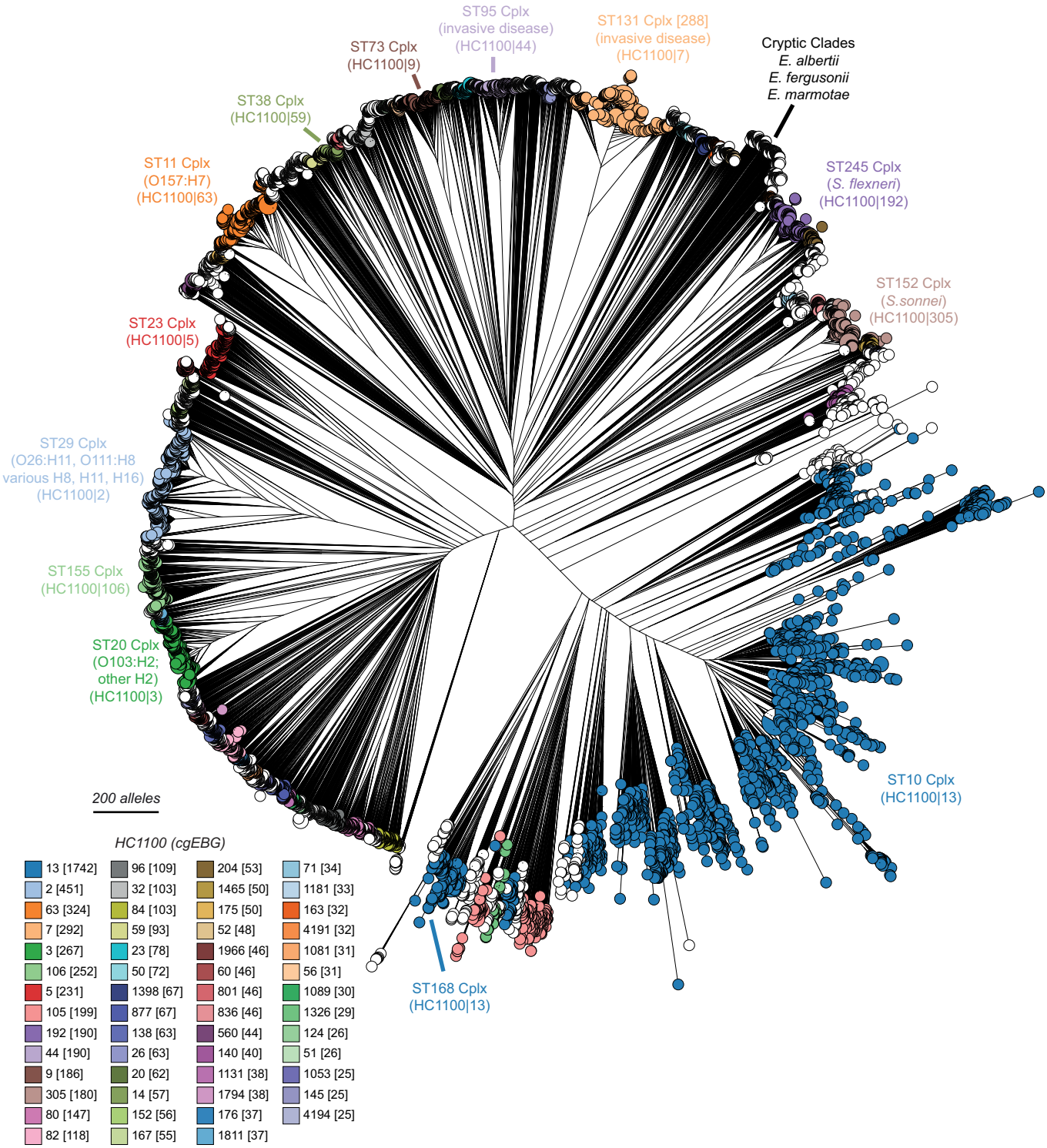


Figure 10. Neighbour-Joining tree of core-genomic distances in the EcoRPlus Collection. EcoRPlus consists of 9,479 genomes, the draft genome with the greatest N50 value for each rST among the 52,876 genomes of *Escherichia* that were available within EnteroBase in August, 2018 (see http://tinyurl.com/ECOR-Plus). The population structure of these representative genomes was investigated by a neighbor-joining (RapidNJ) phylogenetic tree based on pair-wise differences in their cgMLST alleles. The nodes in this tree are color-coded by HC1100 clusters, as indicated in the Key at the bottom left. The locations of common ST Complexes is indicated at the circumference of the tree, and color-coding by ST Complex can be seen in Fig. S5. HC1100 clustering is largely congruent with assignments to common ST Complexes, except that HC1100|13 includes ST10 Complex and ST168 Complex. Still other discrepancies exist among the smaller, unlabeled populations. Color-coding by Clermont typing can be seen in Fig. S7 which highlights multiple discrepancies between taxonomical designations and population groupings. 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 showing estimates of branch lengths can be found in Fig. S6.