The EnteroBase user’s guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity

Zhemin Zhou,1 Nabil-Fareed Alikhan,1 Khaled Mohamed, Yulei Fan, the Agama Study Group,2 and Mark Achtman

Warwick Medical School, University of Warwick, Coventry CV4 7AL, United Kingdom

Enterobase is an integrated software environment that supports the identification of global population structures within several bacterial genera that include pathogens. Here, we provide an overview of how EnteroBase works, what it can do, and its future prospects. EnteroBase has currently assembled more than 300,000 genomes from Illumina short reads from Salmonella, Escherichia, Yersinia, Clostridioides, Helicobacter, Vibrio, and Moraxella and genotyped those assemblies by core genome multilocus sequence typing (cgMLST). Hierarchical clustering of cgMLST sequence types allows mapping a new bacterial strain to predefined population structures at multiple levels of resolution within a few hours after uploading its short reads. Case Study 1 illustrates this process for local transmissions of Salmonella enterica serovar Agama between neighboring social groups of badgers and humans. EnteroBase also supports single nucleotide polymorphism (SNP) calls from both genomic assemblies and after extraction from metagenomic sequences, as illustrated by Case Study 2 which summarizes the microevolution of Yersinia pestis over the last 5000 years of pandemic plague. EnteroBase can also provide a global overview of the genomic diversity within an entire genus, as illustrated by Case Study 3, which presents a novel, global overview of the population structure of all of the species, subspecies, and clades within Escherichia.

[Supplemental material is available for this article.]

Epidemiological transmission chains of Salmonella, Escherichia, or Yersinia have been reconstructed with the help of single-nucleotide polymorphisms (SNPs) from hundreds or even thousands of core genomes (Zhou et al. 2013, 2014, 2018c; Langridge et al. 2015; Connor et al. 2016; Dallman et al. 2016; Wong et al. 2016; Ashton et al. 2017; Alikhan et al. 2018; Waldram et al. 2018; Worley et al. 2018; Johnson et al. 2019). However, the scale of these studies pales in comparison to the numbers of publicly available archives (e.g., NCBI Sequence Read Archive [SRA]) of short-read sequences of bacterial pathogens that have been deposited since the recent drop in price of high-throughput sequencing (https://www.genome.gov/sequencingcostsdata/). In October 2019, SRA contained genomic sequence reads from 430,417 Salmonella, Escherichia/Shigella, Clostridioides, Vibrio, and Yersinia. However, until very recently (Sanaa et al. 2019), relatively few draft genomic assemblies were publicly available, and even the current comparative genomic analyses in NCBI Pathogen Detection (https://www.ncbi.nlm.nih.gov/pathogens/) are restricted to relatively closely related genetic clusters. Since 2014, EnteroBase (https://enterobase.warwick.ac.uk) has attempted to address this gap for selected genera that include bacterial pathogens (Table 1). EnteroBase provides an integrated software platform (Fig. 1) that can be used by microbiologists with limited bioinformatic skills to upload short reads, assemble and genotype genomes, and immediately investigate their genomic relationships to all natural populations within those genera. These aspects have been illustrated by recent publications providing overviews of the population structures of Salmonella (Alikhan et al. 2018) and Clostridioides (Frentrup et al. 2019), a description of the GrapeTree GUI (Zhou et al. 2018a), and a reconstruction of the genomic history of the Salmonella enterica Para C Lineage (Zhou et al. 2018c). However, EnteroBase also provides multiple additional features, which have hitherto largely been promulgated by word of mouth. Here, we provide a high-level overview of the functionality of EnteroBase, followed by exemplary case studies of S. enterica serovar Agama, Yersinia pestis, and all of Escherichia.

Results

Overview of EnteroBase

The EnteroBase back end consists of multiple, cascading automated pipelines (Supplemental Fig. S1) that implement the multiple functions that it provides (Supplemental Fig. S2A). Many of these EnteroBase pipelines are also available within EToKi (EnteroBase ToolKit) (Supplemental Code), a publicly available repository (https://github.com/zheminzhou/EToKi) of useful modules (Supplemental Fig. S2B–E) that facilitate genomic assemblies (EToKi modules prepare and assemble), MLST (MLSType), calling nonrepetitive SNPs against a reference genome (EToKi modules align and phylo), or predicting serotypes of Escherichia coli from genome assemblies (EBEs). EnteroBase performs daily scans of SRA via its Entrez APIs (Clark et al. 2016) for novel Illumina short-read sequences for

© 2020 Zhou et al. This article, published in Genome Research, is available under a Creative Commons License (Attribution 4.0 International), as described at http://creativecommons.org/licenses/by/4.0/.

1Coequal first author
2A complete list of the Agama Study Group coauthors appears at the end of this paper.
Corresponding author: m.achtman@warwick.ac.uk

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.251678.119. Freely available online through the Genome Research Open Access option.

138 Genome Research www.genome.org 30:138–152 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/20; www.genome.org
Escherichia serotypes are predicted from the genome assemblies by the EnteroBase module EB as described below. Their sequence types (STs) are assigned to population groupings (Supplemental Table S1). EnteroBase fetches the metadata associated with the records and attempts to transcribe it automatically into EnteroBase metadata format (Supplemental Table S2; Supplemental Fig. S3). During the conversion, geographic metadata are translated into structured format using the Nominatim engine offered by OpenStreetMap (OpenStreetMap contributors 2017; Planet dump retrieved from https://planet.osm.org) and the host/source metadata are assigned to predefined categories (Supplemental Table S4). Until recently, metadata was parsed using a pretrained Native Bayesian classifier implemented in the Natural Language Toolkit (NLTK) for Python (Bird et al. 2009) with an estimated accuracy of 60%. Since November 2019, a new metaparser is being used, with an estimated accuracy of 93% (Supplemental Material), and all old data will soon be re-parsed. 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 [Supplemental Fig. S2C]) and their sequence types (STs) are assigned to population groupings as described below. Salmonella serovars are predicted from the legacy MLST eBurstGroups (eBGs), which are strongly associated with individual serovars (Achtman et al. 2012), or by two external programs—SISTR (Yoshida et al. 2016; Robertson et al. 2018) and SeqSero2 (Zhang et al. 2019)—which evaluate genomic sequences. Escherichia serotypes are predicted from the genome assemblies by the EnteroBase module EBs (Supplemental Fig. S2E). Clermont haplogroups are predicted for Escherichia by two external programs—ClermontTyping (Begbain et al. 2018) and EZClermont (Waters et al. 2018)—and fimH type by a third (FimTyper) (Roer et al. 2017). By default, all registered users have full access to strain metadata and the genome assemblies, predicted genotypes, and predicted phenotypes, but a delay in the release date of up to 12 mo can be imposed by users when uploading short-read sequences.

In September 2019, EnteroBase provided access to 364,690 genomes and their associated metadata and predictions (Table 1). To allow comparisons with historical data, EnteroBase also maintains additional legacy seven-gene MLST assignments (and metadata) that were obtained by classical Sanger sequencing from 18,478 strains.

Each unique sequence variant of a gene in an MLST scheme is assigned a unique numerical designation. Seven-gene MLST STs consist of seven integers for the alleles of seven housekeeping gene fragments (Maiden et al. 1998). rSTs consist of 51–53 integers for ribosomal protein gene alleles (Jolley et al. 2012). cgMLST STs consist of 1553–3002 integers for the number of genes in the soft core genome for that genus (Table 1), which were chosen as described elsewhere (Frentrup et al. 2019). However, STs are arbitrary constructs, and natural populations can each encompass multiple, related ST variants. Therefore, seven-gene STs are grouped into ST complexes in Escherichia/Shigella (Wirth et al. 2006) by an eBurst

| Genus         | Legacy MLST | Assembled genomes (user uploads) | wgMLST loci | cgMLST loci | rMLST loci | MLST loci | HierCC |
|---------------|-------------|----------------------------------|------------|-------------|------------|----------|-------|
| Salmonella    | 6480        | 225,026 (30,636)                 | 21,065     | 3002        | 51         | 7        |       |
| Escherichia/Shigella | 10,155  | 110,302 (12,584)                 | 25,002     | 2512        | 51         | 7        |       |
| Clostridioides| 14,592      | 14,592 (1422)                    | 11,490     | 2556        | 53         | 7        |       |
| Vibrio        | 7010        | 7010 (128)                       | 51         |             |            |          |       |
| Yersinia      | 1054        | 3412 (1066)                      |            |             |            |          |       |
| Helicobacter  | 2458        | 2458 (846)                       |            |             |            |          |       |
| Moraxella     | 789         | 1890 (349)                       |            |             |            |          |       |
| Total         | 18,478      | 364,690 (47,031)                 |            |             |            |          |       |

Enterobase URL: https://enterobase.warwick.ac.uk (Date accessed: 09-19-2019).

Each unique sequence variant of a gene in an MLST scheme is assigned a unique numerical designation. Seven-gene MLST STs consist of seven integers for the alleles of seven housekeeping gene fragments (Maiden et al. 1998). rSTs consist of 51–53 integers for ribosomal protein gene alleles (Jolley et al. 2012). cgMLST STs consist of 1553–3002 integers for the number of genes in the soft core genome for that genus (Table 1), which were chosen as described elsewhere (Frentrup et al. 2019). However, STs are arbitrary constructs, and natural populations can each encompass multiple, related ST variants. Therefore, seven-gene STs are grouped into ST complexes in Escherichia/Shigella (Wirth et al. 2006) by an eBurst.

Ownership, permanence, access, and privacy

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

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

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

MLST population structures

Each unique sequence variant of a gene in an MLST scheme is assigned a unique numerical designation. Seven-gene MLST STs consist of seven integers for the alleles of seven housekeeping gene fragments (Maiden et al. 1998). rSTs consist of 51–53 integers for ribosomal protein gene alleles (Jolley et al. 2012). cgMLST STs consist of 1553–3002 integers for the number of genes in the soft core genome for that genus (Table 1), which were chosen as described elsewhere (Frentrup et al. 2019). However, STs are arbitrary constructs, and natural populations can each encompass multiple, related ST variants. Therefore, seven-gene STs are grouped into ST complexes in Escherichia/Shigella (Wirth et al. 2006) by an eBurst.
**Figure 1.** Overview of EnteroBase Features. (A) Data uploads. Data are imported from public databases, user uploads, and existing legacy MLST and rMLST databases at PubMLST (https://pubmlst.org/). (B) Spreadsheet Interface. The browser-based interface visualizes sets of strains (one Uberstrain plus any number of substrains) each containing metadata, and their associated experimental data and custom views. Post-release data can be exported (downloaded) as genome assemblies or tab-delimited text files containing metadata and experimental data. Metadata can be imported to entries for which the user has editing rights by uploading tab-delimited text files. (C) Search Strains supports flexible (AND/OR) combinations of metadata and experimental data for identifying entries to load into the spreadsheet. Find ST (s) retrieves STs that differ from a given ST by no more than a maximal number of differing alleles. Locus Search uses BLASTN (Altschul et al. 1990) and UBLastP in USEARCH (Edgar 2010) to identify the MLST locus designations corresponding to an input sequence. Get at this level: menu item after right clicking on experimental MLST ST or cluster numbers. (D) UserSpace OS. A file explorer-like interface for manipulations of workspaces, trees, SNP projects, and custom views. These objects are initially private to their creator but can be shared with buddies or rendered globally accessible. (E) Processes and analyses. EnteroBase uses EToKi and external programs as described in Supplemental Figure S1. (F) Visualization. MLST trees are visualized with the EnteroBase tools GrapeTree (Zhou et al. 2018a) and Dendrogram, which in turn can transfer data to external websites such as Microreact (Argümin et al. 2016).

**cgSTs differ from others only by missing data.** EnteroBase supports working with so many cgSTs through Hierarchical Clustering (HierCC), a novel approach which supports analyses of population structures based on cgMLST at multiple levels of resolution. To identify the cutoff values in stepwise cgMLST allelic distances which would reliably resolve natural populations, we first calculated a matrix of pairwise allelic distances (excluding pairwise missing data) for all existing pairs of cgSTs, and one matrix for the HierCC cluster numbers at each level of allelic distance, that is, one matrix for HC0, HC1, HC2, ..., HC3001. A genus-specific subset of the most reliable HierCC clusters is reported by EnteroBase.

For Salmonella, 13 HierCC levels are reported, ranging from HC0 (indistinguishable except for missing data) to HC2850 (Fig. 2). Our experience with Salmonella indicates that HC2850 corresponds to subspecies, HC2000 to superlineages (Zhou et al. 2018c), and HC900 to cgMLST versions of eBgs. Long-term endemic persistence seems to be associated with HC100 or HC200; and epidemic outbreaks with HC2, HC5, or HC10. Eleven levels are reported for the other genera, ranging from HC0 up to HC2350 for Escherichia, HC2500 for Clostridiodes, and HC1450 for Yersinia. Escherichia HC1100 corresponds to ST Complexes (see below) and the correspondences to population groupings in Clostridiodes are described elsewhere (Frentrup et al. 2019). Further information on HierCC can be found in the EnteroBase documentation (https://enterobase.readthedocs.io/en/latest/features/clustering.html).

**Ubberstrains and substrains**

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

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

---

**Zhou et al.**

Genome Research

www.genome.org
often polyphyletic (Achtman et al. 2012), including those that are rare, poorly understood, and monophyletic. 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 genomes of 

S. enterica

that colonize badgers (Wray et al. 1977; Wilson et al. 2003). Case Study 1 shows how to combine modern genomics and strain metadata into EnteroBase. This data was used to analyze the genomes of 72 such isolates and uploaded the short reads derived from those entries indicated that Agama consisted of multiple microclusters. The top two genomes are both assigned to HC10_306, which indicates a very close relationship, and may represent a transmission chain. The top three genomes are all assigned to HC900_2, which corresponds to a legacy MLST eBG. HC2000 marks superlineages (Zhou et al. 2018c), and HC2850 marks subspecies. This figure illustrates these interpretations in the form of a cladogram drawn by hand.

Examples of the utility of EnteroBase

Often the utility of a tool first becomes clear through examples of its use. Here, we present three case studies that exemplify different aspects of EnteroBase. Case Study 1 shows how geographically separated laboratories can collaborate in private on an EnteroBase project until its completion, upon which EnteroBase publishes the results. This example focuses on geographical microvariation and transmission chains between various host species of a rare serovar of 

S. enterica

Case Study 2 shows how to combine modern genomes of 

Yersinia pestis

with partially reconstructed genomes from ancient skeletons of plague victims. It also shows how EToKi can extract SNPs from metagenomic sequence reads. Case Study 3 provides a detailed overview of the genomic diversity of the genus Escherichia and defines the EcoRPlus set of representative genomes.

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

S. enterica subspecies enterica encompasses more than 1586 defined serovars (Guibourdenche et al. 2010; Issenhuth-Jeannet et al. 2014). These differ in the antigenic formulas of their lipopolysaccharide (O antigen) and/or two alternative flagellar antigens (H1, H2), which are abbreviated as O:H1:H2. Some serovars are commonly isolated from infections and the environment and have been extensively studied. Others are rare, poorly understood, and often polyphyletic (Achtman et al. 2012), including Salmonella that colonize badgers (Wray et al. 1977; Wilson et al. 2003). In late 2018, serovar Agama (antigenic formula: 4,12:i:1,6) was isolated from humans. We were therefore interested to learn that the University of Liverpool possessed serovar Agama isolates that had been isolated in 2006–2007 from European badgers (Meles meles) in Woodchester Park, Gloucestershire, England. We sequenced the genomes of 72 such isolates and uploaded the short reads and strain metadata into EnteroBase. This data was used to analyze the population structure of a rare serovar within a single host species over a limited geographical area and to compare Agama genomes from multiple hosts and geographical sources.

Search strains

The browser interface to EnteroBase is implemented as a spreadsheet-like window called a “Workspace” that can page through thousands of entries, showing metadata at the left and experimental data at the right (https://enterobase.readthedocs.io/en/latest/features/using-workspaces.html). However, visual scanning of so many entries is inefficient. EnteroBase therefore offers powerful search functions (https://enterobase.readthedocs.io/en/latest/enterobase-tutorials/search-agama.html) for identifying isolates that share common phenotypes (metadata) and/or genotypes (experimental data).

EnteroBase also predicts serovars from assembled Salmonella genomes and from MLST data. However, the software predictions are not failproof, and many entries lack metadata information or the metadata is erroneous. We therefore used the Search Strains dialog box to find entries containing “Agama” in the metadata Serovar field or by the Serovar predictions from SISTR1 (https://enterobase.readthedocs.io/en/latest/enterobase-tutorials/search-agama.html). Phylogenetic analyses of the cgMLST data from those entries indicated that Agama consisted of multiple microclusters.
International participation in a collaborative network

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

We facilitated the analysis of the Agama data by creating a new user-defined Custom View (https://enterobase.readthedocs.io/en/latest/features/user-defined-content.html), which can aggregate various sources of experimental data as well as user-defined fields. The Custom View was saved in the Agama folder, and thereby shared with the study group. It too was initially private but became public together with the other workspaces and trees when the folder was made public.

Members of the Agama Study Group were requested to sequence genomes from all Agama strains in their collections, and to upload those short reads to EnteroBase, or to send their DNAs to University of Warwick for sequencing and uploading. The new entries were added to the “All Agama Strains” workspace. The final set of 345 isolates had been isolated in Europe, Africa, and Australia, with collection years ranging from 1956 to 2018 (Supplemental Table S3).

Global population structure of Agama

We created a neighbor-joining GrapeTree (Zhou et al. 2018a) of cgMLST allelic differences to reveal the genetic relationships within serovar Agama. (A) Color coding by Predicted Serovar (SISTR1). Arrows indicate isolates whose serovar was not predicted. Orange shading emphasizes 1,4,[5],12:i:- isolates that were monophasic Agama. Gray shading indicates isolates with incorrect Serovar metadata, including 1,4,[5],12:i:- isolates that were monophasic Typhimurium (arrow). (B) Color coding by HC2000 cluster. All Agama entries are HC2000_299, as were the genetically related entries marked with arrows or emphasized by orange shading. Entries from other serovars (gray shading) were in other diverse HC2000 clusters. The dashed box indicates a subset of Agama strains within HC400_299, including all isolates from badgers, which were chosen for deeper analyses in Figure 4. (Scale bar) Number of cgMLST allelic differences.
In contrast to serovar, coloring the tree nodes by HC2000 clusters (Fig. 3B) immediately revealed that all genomes that were called Agama by SISTR1 belonged to HC2000 cluster number 299 (HC2000_299), and all HC2000_299 were genetically related and clustered together in the tree (Fig. 3B). In contrast, the 16 other isolates on long branches (gray shading) belonged to other HC2000 clusters.

These results show that Agama belongs to one superlineage, HC2000_299, which has been isolated globally from humans, badgers, companion animals, and the environment since at least 1956. The genetic relationships would not have been obvious with lower resolution MLST: Some Agama isolates belong to eBG167, others to eBG336, and 13 Agama MLST STs do not belong to any eBG.

**Transmission patterns at different levels of HierCC resolution**

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

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

The badgers in Woodchester Park occupy adjacent social group territories, which each contain several sets (burrows). HC100_2433 contains multiple HC10 clusters of Agama from badgers (Supplemental Fig. S5A). To investigate whether these microclusters might mark transmission chains between sets and social groups, a Newick subtree of HC100_2433 plus geographical coordinates was transmitted from GrapeTree to Microreact (Argimón et al. 2019), an external program which is specialized in depicting geographical associations. Badgers occasionally move between neighboring social groups (Rogers et al. 1998). Transmissions associated with such moves are supported by the observation that five distinct HC10 clusters each contained isolates from two social groups in close proximity (Supplemental Fig. S5B).

**Long-term dispersals and interhost transmissions**

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

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

EnteroBase automatically scour sequence read archives for Illumina short reads from cultivated isolates, assembles their genomes, and publishes draft assemblies that pass quality control. In October 2019, EnteroBase had assembled more than 1300 genomes of Y. pestis, including genomes that had already been assigned to population groups (Cui et al. 2013), other recently sequenced genomes from central Asia (Eroschenko et al. 2017; Kutyrev et al. 2018), and numerous unpublished genomes from Madagascar and Brazil. EnteroBase does not upload assembled genomes, for which adequate, automated quality control measures would be difficult to implement. However, EnteroBase administrators can upload such genomes after ad hoc assessment of sequence quality, and EnteroBase contains standard complete genomes such as CO92 (Parkhill et al. 2001) and other genomes used to derive the Y. pestis phylogeny (Morelli et al. 2010).

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

**Ancient Y. pestis**

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

The EnteroBase EToKi calculation package (Supplemental Code) can reconstruct an ancient genome assembly by unmasking individual nucleotides in a fully masked reference genome based on reliable SNP calls from metagenomic data (Supplemental Fig. S6). We ran EToKi on S6 published ancient metagenomes containing Y. pestis, and the resulting assemblies and metadata were uploaded to EnteroBase. EnteroBase users can now include those ancient genomes together with other reconstructed genomes and modern genomic assemblies in a workspace of their choice and use the EnteroBase SNP dendrogram module to calculate and visualize a maximum-likelihood tree (of up to a current maximum of 200 genomes).

Figure 5 presents a detailed overview of the genomic relationships of all known Y. pestis populations from pandemic plague
Figure 4. Effects of sample bias on inferred transmission chains within HC400_299 Agama isolates. (A, left) Map of hosts in the British Isles of 149 Agama isolates in EnteroBase in August, 2018. (Right) Maximum-likelihood radial phylogeny (http://enterobase.warwick.ac.uk/a/21773/d) based on RAxML (Stamatakis 2014) of 8791 nonrepetitive core SNPs as calculated by EnteroBase Dendrogram against reference genome 283179. Color coding is according to a user-defined field (Location & Source). HC100 cluster designations for three microclades are indicated. HC100_2433 contained all Agama from badgers. (B, right) Summary of hosts and countries from which 64 additional Agama isolates had been sequenced by March 2019. (Left) Maximum-likelihood radial dendrogram (http://enterobase.warwick.ac.uk/a/23882/d) based on 9701 SNPs from 213 isolates. Multiple isolates of Agama in HC100_2433 were now from humans and food in France and Austria. HC100_299 and HC100_67355 now contained multiple isolates from badgers, livestock, companion animals, and mussels, demonstrating that the prior strong association of Agama with humans and badgers in A reflected sample bias. Stars indicate multiple MRCAs of Agama in English badgers, whereas the pink arrow indicates a potential transmission from badgers to a human in Bath/North East Somerset, which is close to Woodchester Park. The green arrow indicates a potential food-borne transmission chain consisting of four closely related Agama isolates in HC5_140035 from Austria (chives × 2; human blood culture × 1) and France (human × 1) that were isolated in 2018. The geographical locations of the badger isolates are shown in Supplemental Figure S5.
over the last 5500 years, including hundreds of unpublished modern genomes. This tree was manually annotated using a User-defined Field and Custom View with population designations for reconstructed ancient genomes that are consistent with the literature on modern isolates. We also assigned consistent population designations to additional modern genomes from central Asia and elsewhere. An interactive version of this tree and all related metadata in EnteroBase is publicly available (http://enterobase.warwick.ac.uk/a/21977/g), 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 workhorses of molecular biology. Most studies of *Escherichia* have concentrated on a few well-characterized strains of *E. coli*, but the genus *Escherichia* includes other species: *E. fergusonii*, *E. albertii*, *E. marmotae* (Liu et al. 2015), and *E. ruysiae* (van der Putten et al. 2019). *E. coli* itself includes the genus *Shigella* (Pupo et al. 2000), which was assigned a distinctive genus name because it causes dysentery. Initial analyses of the phylogenetic structure of *E. coli* identified multiple deep branches, called haplogroups (Selander et al. 1987), and defined
the EcoR collection (Ochman and Selander 1984), a classical group of 72 bacterial strains that represented the genetic diversity found with multilocus enzyme electrophoresis. The later isolation of environmental isolates from lakes revealed the existence of “cryptic clades” I–V1 which were distinct from the main *E. coli* haplogroups and the other *Escherichia* species (Walk et al. 2009; Luo et al. 2011). Currently, bacterial isolates are routinely assigned to haplogroups or clades by PCR tests for the presence of variably present genes from the accessory genome (Clermont et al. 2013) or by programs that identify the presence of those genes in genomic sequences (Beghain et al. 2018; Waters et al. 2018).

Legacy MLST is an alternative scheme for subdividing *Escherichia*, which includes the assignment of STs to ST Complexes (Wirth et al. 2006). Several ST Complexes are common causes of invasive disease in humans and animals, such as ST131 (Stoesser et al. 2016; Liu et al. 2018), ST95 Complex (Wirth et al. 2006; Gordon et al. 2017), and ST11 Complex (O157:H7) (Epping et al. 2011a,b; Newell and La Ragione 2018). The large number of *Escherichia* genomes in EnteroBase (Table 1) offered the opportunity to reinvestigate the population structure of *Escherichia* on the basis of the greater resolution provided by cgMLST and within the context of a much larger and more comprehensive sample. In 2018, EnteroBase contained 52,876 genomes. To render this sample amenable to calculating a maximum-likelihood (ML) tree of core SNPs, we selected a representative sample consisting of one genome from each of the 9479 *Escherichia* sSTS. In homage to the EcoR collection, we designate this as the EcoRPlus Collection.

**Core genome genetic diversity within *Escherichia***

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

Figure 6 may represent the first detailed overview of the genetic diversity of the core genome of *Escherichia*. Real-time examination of its features (http://enterobase.warwick.ac.uk/a/15981) is feasible because the GrapeTree algorithm can handle large numbers of cgSTS (Zhou et al. 2018a). Nodes can be readily colored by metadata or experimental data (Supplemental Figs. S7, S8), and GrapeTree also readily supports analyses of subtrees in greater detail. However, although cgMLST allelic distances are reliable indicators of population structures, SNPs are preferable for examining genetic distances. We therefore calculated a ML tree of the 1,230,995 core SNPs within all 9479 genomes (Supplemental Fig. S9). This tree confirmed the clustering of the members of HC1100 groups within *E. coli*, and also showed that the other *Escherichia* species and cryptic clades II to VIII formed distinct long branches of comparable lengths (Supplemental Fig. S9 inset).

**Discussion**

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

EnteroBase was intended to provide a uniform and reliable pipeline that can assemble consistent draft genomes from the numerous short-read sequences in public databases (Achtman and Zhou 2014) and to link those assemblies with metadata and genotype predictions. It was designed to provide access to an unprecedentedly large global set of draft genomes to users at both extremes of the spectrum of informatics skills. A further goal was to provide analytical tools, such as GrapeTree (Zhou et al. 2018a), that could adequately deal with cgMLST from more than 100,000 genomes, and Dendrogram, which generates phylograms from nonrepetitive core SNPs called against a reference genome. Still another important goal was to support private analyses by groups of colleagues, with the option of subsequently making those analyses publicly available. Case Study 1 illustrates how EnteroBase can be used for all of these tasks, and more.

EnteroBase has expanded beyond its original goals and is morphing in novel directions. It has implemented HierCC for cgMLST, which supports the automated recognition of population structures at multiple levels of resolution (Case Study 1), and may help with the annotation of clusters within phylogenetic trees (Case Study 2; see below). EnteroBase has also been extended to support analyses of metagenomic data from ancient genomes (Zhou et al. 2018c; Achtman and Zhou 2019) 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 *F. pestis*. Additional EnteroBase databases are under development for ancient and modern genomes of *S. enterica* and biofilms within dental calculus. EnteroBase has also shown its capacities for providing overviews of the core genome diversity of entire genera, with currently extant examples consisting of *Salmonella* (AliKhan et al. 2018) and *Escherichia* (Case Study 3).

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

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

Case Study 1 may also become a paradigm for identifying long-distance chains of transmission between humans or between humans and their companion or domesticated animals: Four Agama isolates in the HCS-140035 cluster from France (human) and Austria (frozen chives and a human blood culture) differed by no more than five of the 3002 cgMLST loci. These isolates also differed by no more than five nonrepetitive core SNPs.
Similar discoveries of transmissions of *E. coli* between humans and wild birds are described below. We anticipate that large numbers of such previously silent transmission chains will be revealed as EnteroBase is used more extensively.

Case Study 2 illustrates how EnteroBase can facilitate combining reconstructed genomes from metagenomic sequences with draft genomes from cultured strains. In this case, the metagenomes were from ancient tooth pulp that had been enriched for *Y. pestis*.
and the bacterial isolates were modern Y. pestis from a variety of global sources since 1898. The resulting phylogenetic tree (Fig. 5) presents a unique overview of the core genomic diversity over 5000 years of evolution and pandemic spread of plague, which can now be evaluated and used by a broad audience. This tree will be updated at regular intervals as additional genomes or metagenomes become available.

The manual population designations in Figure 5 are largely reflected by HC10 clusters. However, it is uncertain whether the current HierCC clusters would be stable with time because they were based on only 1300 Y. pestis genomes. EnterorBase will therefore maintain these manual annotations in parallel with automated HierCC assignments until a future date when a qualified choice is possible.

Case Study 3 defines the EcoRPlus Collection of 9479 genomes, which represents the genetic diversity of 52,876 genomes. It is a worthy successor of EcoR (Ochman and Selander 1984), which contained 72 representatives of 2600 E. coli strains that had been tested by multilocus enzyme electrophoresis in the early 1980s. The genomic assemblies and known metadata of EcoRPlus are publicly available (http://enterobase.warwick.ac.uk/a/15931) and can serve as a reference set of genomes for future analyses with other methods.

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

This case study also opened up new perspectives during the review phase of this paper, such as how EnterorBase could be used for the analysis of interhost transmission of antimicrobial resistance (AMR). Seagulls often carry E. coli that are resistant to multiple antibiotics and can transmit those bacteria to other seagulls (Stedt et al. 2014; Ahlstrom et al. 2018, 2019b; Sandegren et al. 2018), including at multiple sites in a small area of Alaska between seagulls and other hosts. EnteroBase contained 406 E. coli genomes from seagulls, distributed over 322 HC5 clusters. Of those clusters, four contained E. coli strains isolated from other hosts (Supplemental Table SS5), including chickens, crows, swine, and humans. The dates of isolation of those isolates ranged over about 4 yr, and their geographical locations were separated by long distances: Alaska–New York; Alaska–Michigan; Tasmania–continental Australia. As indicated above, HC5 clusters in Salmonella are associated with recent transmission chains between badgers and across European borders. These additional observations suggest that E. coli from diverse ST Complexes which encode AMR have also been recently transmitted between humans and wild birds and domesticated animals.

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

Methods

Isolation of serovar Agama from badgers

Supplemental Figure SSB provides a geographical overview of the area in Woodchester, Gloucestershire, in which badger sets and social groups were investigated in 2006–2007. This area has been subject to a multidecade investigation of badger mobility and patterns of infection with Mycobacterium bovis (McDonald et al. 2018).

According to the standard protocol for that study, badgers were subjected to routine capture using steel mesh box traps baited with peanuts, examination under anesthesia, and subsequent release. Fecal samples were cultivated at University of Liverpool after selective enrichment (Rappaport–Vassiliadis broth and semisolid 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 (Issenhuth-Jeanjean et al. 2014). Additional isolates from badgers from the geographical areas in England that are indicated in Supplemental Figure S5D 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 Figure 4 are publicly available in EnterorBase in the workspace “Zhou et al. All Agama strains” (http://enterobase.warwick.ac.uk/a/21320).

Integration of ancient Yersinia pestis genomes in EnterorBase

Metagenomic reads from ancient samples may contain a mixture of sequence reads from the species of interest as well as from genetically similar taxa that represent environmental contamination. To deal with this issue and remove such nonspecific reads after extraction with the EToKi prepare module, the EToKi assembly module can be used to align the extracted reads after comparisons with an ingroup of genomes related to the species of interest and with an outgroup of genomes from other species. In the case of Figure 5, the ingroup consisted of Y. pestis genomes CO92 (2001), Pestoides F, KIM10+ and 91001, and the outgroup consisted of Y. pseudotuberculosis genomes IP32953 and IP31758, Y. similis 228, and Y. enterocolitica 8081. Reads were excluded which had high alignment scores to the outgroup genomes than to the ingroup genomes. Prior to mapping reads to the Y. pestis reference genome (CO92) (2001), a pseudogenome was created in which all nucleotides were masked to ensure that only nucleotides
supported by metagenomic reads would be used for phylogenetic analysis. For the 13 ancient genomes whose publications included complete SNP lists, we unmasked the sites in the pseudogenomes that were included in the published SNP lists. For the other 43 genomes, EToKi was used as in Supplemental Figure S6 to map the filtered metagenomic reads onto the pseudogenome with minimap2 (Li 2018), evaluate them with Pilon (Walker et al. 2014), and unmask sites in the pseudogenome that were covered by three or more reads and had a consensus base that was supported by ≥80% of the mapped reads. All 56 pseudogenomes were uploaded to EnteroBase together with their associated metadata.

Data access

The Illumina sequence reads for 161 new genomes of *S. enterica* serovar Agama generated in this study have been submitted to the European Nucleotide archive database (ENA; https://www.ebi.ac.uk/ena) under study accession numbers ERP114376, ERP114456, ERP114871, and ERP115055. The genomic properties, metadata, and accession codes for the 329 genomic assemblies in HC2000_299 are summarized in Supplemental Table S3 and in Online Table 1 (https://wrap.warwick.ac.uk/128112). The metadata, genomic assemblies, and annotations are also available from the publicly available workspace “Zhou et al. All Agama Strains” (http://enterobase.warwick.ac.uk/a/21320). The EToKi package and its documentation are accessible at https://github.com/zheminzhou/EToKi and as Supplemental Code. EnteroBase documentation is accessible at https://enterobase.readthedocs.io/en/latest/. An interactive version of Figure 3 is available at http://enterobase.warwick.ac.uk/a/24006. Trees presented in Figure 4A,B are available separately at http://enterobase.warwick.ac.uk/a/21773/d and http://enterobase.warwick.ac.uk/a/23882/d, respectively. An interactive version of Figure 5 is available at http://enterobase.warwick.ac.uk/a/21977/g. The MicroReact projects of Supplemental Figure SSA,B are available at https://microreact.org/project/t7qBSShh/36e534888; Supplemental Figure SSC,D at https://microreact.org/project/9XUC7fi-Fm/f665H5; and Supplemental Figure SSE,F at https://microreact.org/project/XajlmCJNy/69748fe3. The tree shown in Figure 6 as well as Supplemental Figures S7, S8 are available at http://enterobase.warwick.ac.uk/a/15981.

Agama Study Group

Derek Brown,3 Marie Chataway,4 Tim Dallman,4 Richard Delahay,5 Christian Kornschober,6 Ariane Pietzka,6 Burkhard Malorny,7 Liljana Petrovska,8 Rob Davies,8 Andy Robertson,9 William Tyne,10 François-Xavier Weill,11 Marie Accou-Demartin,11 and Nicola Williams12

References

Achtman M. 2016. How old are bacterial pathogens? *Proc Biol Sci* 283: 20160090. doi:10.1098/rspb.2016.0990

Achtman M, Zhou Z. 2014. Distinct genealogies for plasmids and chromosome. *PLoS Genet* 10: e1004874. doi:10.1371/journal.pgen.1004874

Achtman M, Zhou Z. 2019. Analysis of the human oral microbiome from modern and historical samples with SPARSE and EToKi. bioRxiv doi:10.1101/642542

Achtman M, Wain J, Wells FX, Nair S, Zhou Z, Sangal V, Krauland MG, Hale JL, Harbottle H, Uesbeck A, et al. 2012. Multicloud sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLoS Pathog* 8: e1002776. doi:10.1371/journal.ppat.1002776

Ahlstrom CA, Bonnedahl J, Woskpeh H, Hernandez J, Olsen B, Ramey AM. 2018. Acquisition and dissemination of cephalosporin-resistant *E. coli* in migratory birds sampled at an Alaska landfill as inferred through genomic analysis. *Sci Rep* 8: 7361. doi:10.1038/s41598-018-25474-w

Ahlstrom CA, Bonnedahl J, Woskpeh H, Hernandez J, Reed JA, Tibbitts L, Olsen B, Douglas DC, Ramey AM. 2019a. Satellite tracking of gulls and genomic characterization of facia bacteria reveals environmentally mediated acquisition and dispersal of antimicrobial-resistant *Escherichia coli* on the Kenai Peninsula, Alaska. *Mol Ecol* 28: 2531–2545. doi:10.1111/mec.15101

Ahlstrom CA, Ramey AM, Woskpeh H, Bonnedahl J. 2019b. Repeated detection of carbapenemase-producing *Escherichia coli* in gulls inhabiting Alaska. *Antimicrob Agents Chemother* 63: e00758-19. doi:10.1128/AAC.00758-19

Allikhan NF, Zhou Z, Sergeant MJ, Achtman M. 2018. A genomic overview of the population structure of *Salmonella*. *PLoS Genet* 14: e1007261. doi:10.1371/journal.pgen.1007261

Altshul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215: 403–410. doi:10.1016/S0022-2836(85)80360-2

Argimon S, Abudahab K, Goater RJ, Fedosejev A, Bhai J, Glasner C, Feil EJ, Nielsen JH, Harbottle H, Uesbeck A, et al. 2016. Microreact: visualizing and sharing data for genomic epidemiology and phyogeography. *Microb Genom* 2: e000093. doi:10.1099/mgen.0.000093

Ashton PM, Owen SV, Kaindama L, Rowe WPM, Lane CR, Larkin L, Nair S, Jenkins C, de Pinna EM, Feasey NA, et al. 2017. Public health surveillance in the UK revolutionises our understanding of the invasive *Salmonella Typhimurium* epidemic in Africa. *Genome Med* 9: 92. doi:10.1186/s12973-017-0480-7

Beghain J, Bridier-Nahmias A, Le NH, Denamur E, Clermont O. 2018. *Clermont*Typing: an easy-to-use and accurate in silico method for *Escherichia* genus strain phylotyping. *Microb Genom* 4: e000192. doi:10.1099/mgen.0.000192

Bird S, Klein E, Loper E. 2009. *Natural language processing with Python: analyzing text with the Natural Language Toolkit*, 1st ed. O’Reilly Media, Sebastopol, CA.

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. doi:10.1038/nature10549

Bos KI, Herbig A, Sahl J, Waglechner N, Fourment M, Forrest SA, Klunk J, Schuenemann VM, Poinar D, Kuch M, et al. 2016. Eighteenth century *Yersinia pestis* genomes reveal the long-term persistence of an historical plague focus. *eLife* 5: e12994. doi:10.7554/eLife.12994

Acknowledgments

EnteroBase development was funded by the Biotechnology and Biological Sciences Research Council (BB/L020319/1) and the Wellcome Trust (202792/Z/16/Z). We thank Niall Delappe and Martin Cormican, Salmonella Reference Laboratory, Galway, Ireland, for sharing strains and data, and Nina Luhmann and Jane Charlesworth for critical comments on the text.

Author contributions: M.A. wrote the manuscript with the help of all other authors. Z.Z., N.-F.A., K.M., and Y.F. were responsible for the development of EnteroBase under the guidance of M.A. N.-F.A. and K.M. were responsible for the online manual. Analyses were performed and figures were drawn by Z.Z. and N.-F.A. under the guidance of M.A. The Agama Study Group provided information, bacterial strains, DNAs, and genomic sequences from Agama isolates from all over Europe and was involved in writing the manuscript and evaluating the conclusions.
Spyrou MA, Tukhatova RI, Wang CC, Valuueva AA, Lankapalli AK, Kondrashin VV, Tsybin VA, Khokhlov A, Kühnert 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. doi:10.1038/s41467-018-04655-z

Spyrou MA, Keller M, Tukhatova RI, Scheib CL, Nelson EA, Andrades VA, Neumann GU, Walker D, Alterauge A, Carty N, et al. 2019. Phylogeography of the second plague pandemic revealed through analysis of historical Yersinia pestis genomes. Nat Commun 10: 4470. doi:10.1038/s41467-019-12154-0

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312–1313. doi:10.1093/bioinformatics/btu133

Stedl J, Bonnedahl J, Hernandez J, McMahon BJ, Hasen B, Olsen B, Drobni M, Waldenström J. 2014. Antibiotic resistance patterns in Escherichia coli from gulls in nine European countries. Infect Ecol Epidemiol 4: 21565. doi:10.3402/iee.v4i21565

Stoesser N, Sheppard AE, Pinkhurst L, De Maio N, Moore CE, Sobra 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: e01262. doi:10.1128/mbio.00126-15

van den Bogaart BCJ, Mailund T, Pedersen CNS. 2018. Enterohaemorrhagiae and other Shiga toxin-producing Escherichia coli (STEC): Where are we now regarding diagnostics and control strategies? Transbound Emerg Dis 65: 49–71. doi:10.1111/tbed.12789

Numberger D, Riedel T, McEwen G, Nübel U, Frentrup M, Schober I, Bunk B, Sporns O, Jost B, Grossart HP, et al. 2019. Genomic analysis of three Chlorotrichia difficile isolates from urban water sources. Anaerobe 56: 22–26. doi:10.1016/j.anaero.2019.01.002

Ochman H, Selander RK. 1984. Standard reference strains of Escherichia coli isolated naturally from enteric infections. J Bacteriol 157: 690–695.

O’Farrell B, Haase JK, Veludayuthan V, Murphy RA, Achtman M. 2012. Transforming microbial genotyping: a robotic pipeline for genotyping bacterial strains. PLoS One 7: e48022. doi:10.1371/journal.pone.0048022

Parkhill J, Ween BW, Thomson NR, Tishball RW, Holden MTJ, Sebaihia M, James KD, Churcher C, Mungall KL, et al. 2001. Genome sequencing of Yersinia pestis, the causative agent of plague. Nature 413: 523–529. doi:10.1038/35100531

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. doi:10.1073/pnas.97.18.10567

Rasmussen S, Allentoft ME, Nielsen K, Orlando L, Sikora M, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, Robertson J, Yoshida CE, Kruczkiewicz P, Nadon C, Nichani A, Taboada EN, Namouchi A, Guellil M, Kersten O, Hänsch S, Ottoni C, Schmid BV, Pacciani S, Morelli G, Zhou Z, McCann A, Litrup E, Murphy R, Cormican S, Brown D, Gutmann DS, Brisse S, Achtman M. 2013. Neutral genomic micro-evolution of a recently emerged pathogen, Salmonella enterica serovar Agona. PLoS Genet 9: e1003471. doi:10.1371/journal.pgen.1003471

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 tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9: e112963. doi:10.1371/journal.pone.0112963

Waters NR, Abram F, Brennan F, Holmes A, Pritchard L. 2018. Easily phyotyping E. coli via the EsClertmon web app and command-line tool. bioRxiv doi:10.1101/317610

Wilson JS, Hazel SM, Williams NJ, Pithi A, French NP, Hart CA. 2003. Nontyphoidal salmonellae in United Kingdom badgers: prevalence and spatial distribution. Appl Environ Microbiol 69: 4312–4315. doi:10.1128/AEM.69.7.4312-4315.2003

Wirth T, Faulhaber 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. doi:10.1111/j.1365-2958.2006.05170.x

Wirth T, Morelli G, Kuczek B, Van Belkum A, Van der Schee C, Meyer A, Achtman M. 2007. The rise and spread of a new pathogen: seroreistance Monash calves. Genome Res 17: 1647–1656. doi:10.1101/gr.66307

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. doi:10.1038/ncomms12827

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: e02530-18. doi:10.1128/mBio.02530-18

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. doi:10.1016/S0007-1935(17)39996-9

Yoshida CE, Kruczkiewicz P, Laing CR, Gannon VP, Nash JH, Frome DM, Koltenbah NK, et al. 2019. Emergence of a novel Salmonella enterica serovar Typhi, the cause of human typhoid. Nat Commun 10: 4470. doi:10.1038/s41467-019-12154-0

Zhang S, Den-Bakker HC, Li S, Chen J, Dinsmore BA, Lane C, Lauer AC, Fields PI, Deng X. 2019. SeqSero2: rapid and improved Salmonella serotype determination using whole genome sequencing data. Appl Environ Microbiol 85: e01746-19. doi:10.1128/AEM.01746-19

Zhao Z, McCann A, Litrup E, Murphy R, Cornican M, Fanning S, Brown D, Gutmann DS, Brisse S, Achtman M. 2013. Neutral genomic micro-evolution of a recently emerged pathogen, Salmonella enterica serovar Agona. PLoS Genet 9: e1003471. doi:10.1371/journal.pgen.1003471

Zhou Z, McCann A, Litrup E, Murphy R, Cornican M, Fanning S, Brown D, Gutmann DS, Brisse S, Achtman M. 2013. Neutral genomic micro-evolution of a recently emerged pathogen, Salmonella enterica serovar Agona. PLoS Genet 9: e1003471. doi:10.1371/journal.pgen.1003471
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 111: 12199–12204. doi:10.1073/pnas.1411012111

Zhou Z, Alikhan N-F, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, Carriço JA, Achtman M. 2018a. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Res 28: 1395–1404. doi:10.1101/gr.232397.117

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

Zhou Z, Lundstrøm I, Tran-Dien A, Duchêne S, Alikhan NF, 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.e10. doi:10.1016/j.cub.2018.05.058

Received April 20, 2019; accepted in revised form December 3, 2019.