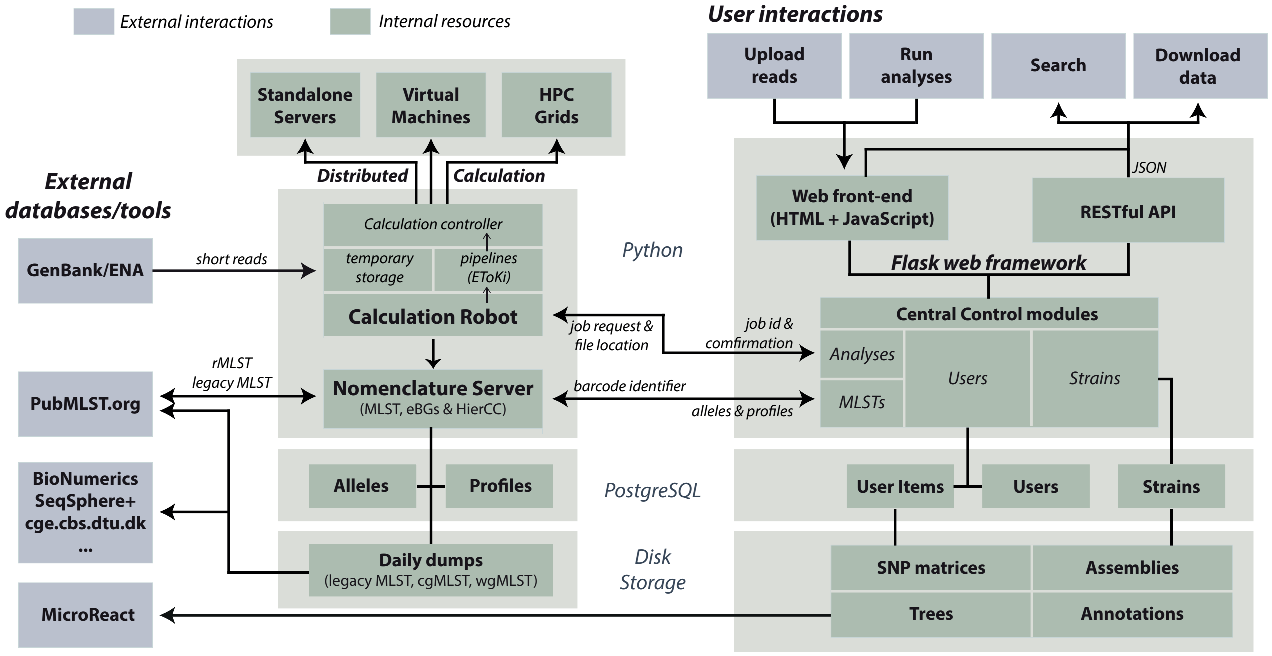
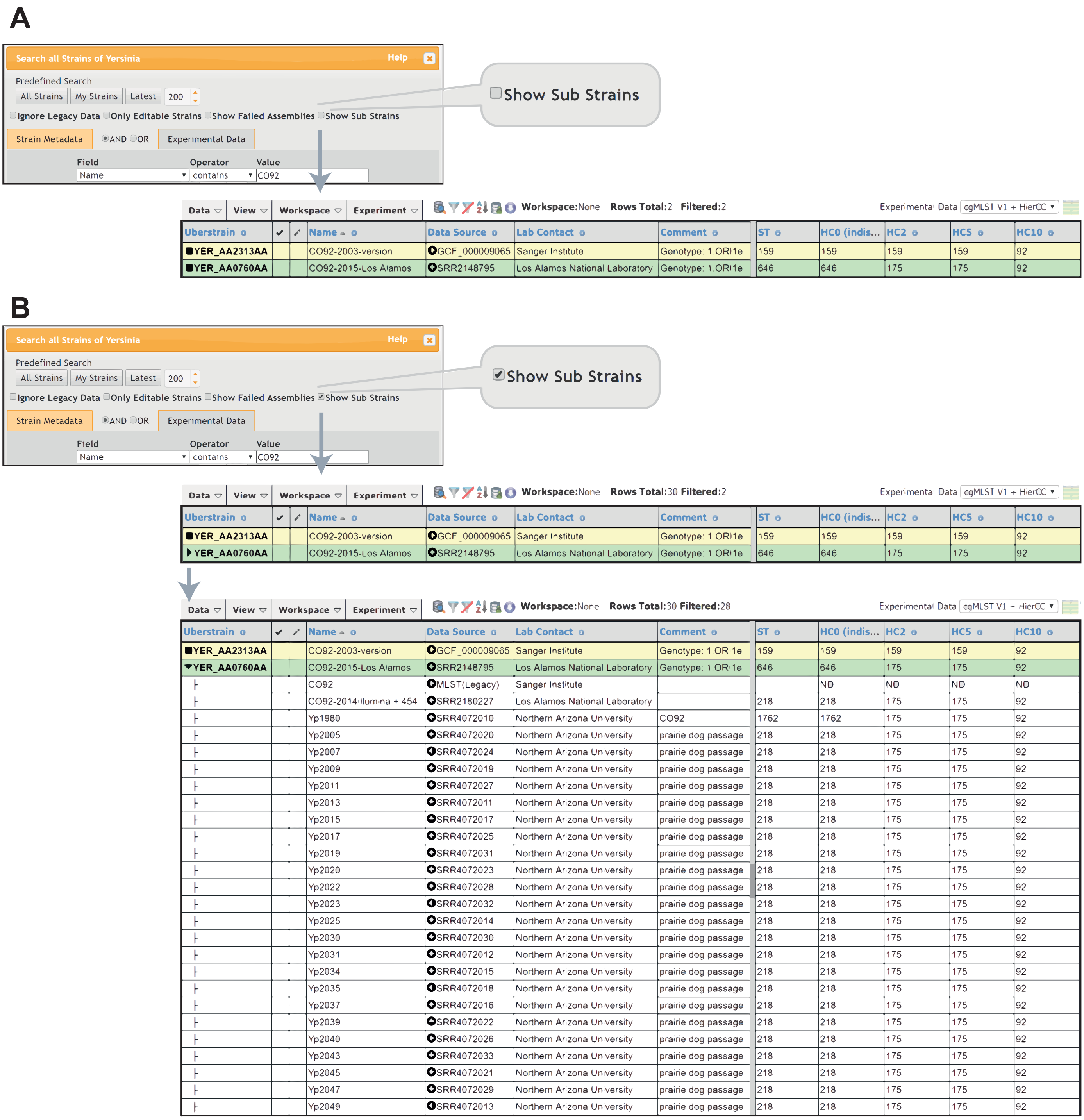


Supplementary Figure 1. Overview of workflow within EnteroBase. EnteroBase offers its users interactions via a web browser front end or a RESTful API (Application Programming Interface), both of which connect to a Python Flask web framework environment that handles all further interactions with structured 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.



A

Search all Strains of Yersinia

Predefined Search: All Strains | My Strains | Latest | 200

□ Ignore Legacy Data □ Only Editable Strains □ Show Failed Assemblies □ Show Sub Strains

Strain Metadata ● AND ○ OR Experimental Data

Field: Name | Operator: contains | Value: CO92

□ Show Sub Strains

Data | View | Workspace | Experiment | Workspace: None | Rows Total: 2 | Filtered: 2 | Experimental Data: cgMLST V1 + HierCC

Uberstrain	✓		Name	Data Source	Lab Contact	Comment	ST	HC0 (indis...	HC2	HC5	HC10
■YER_AA2313AA			CO92-2003-version	GCF_000009065	Sanger Institute	Genotype: 1.ORI1e	159	159	159	159	92
■YER_AA0760AA			CO92-2015-Los Alamos	SRR2148795	Los Alamos National Laboratory	Genotype: 1.ORI1e	646	646	175	175	92

B

Search all Strains of Yersinia

Predefined Search: All Strains | My Strains | Latest | 200

□ Ignore Legacy Data □ Only Editable Strains □ Show Failed Assemblies ☑ Show Sub Strains

Strain Metadata ● AND ○ OR Experimental Data

Field: Name | Operator: contains | Value: CO92

☑ Show Sub Strains

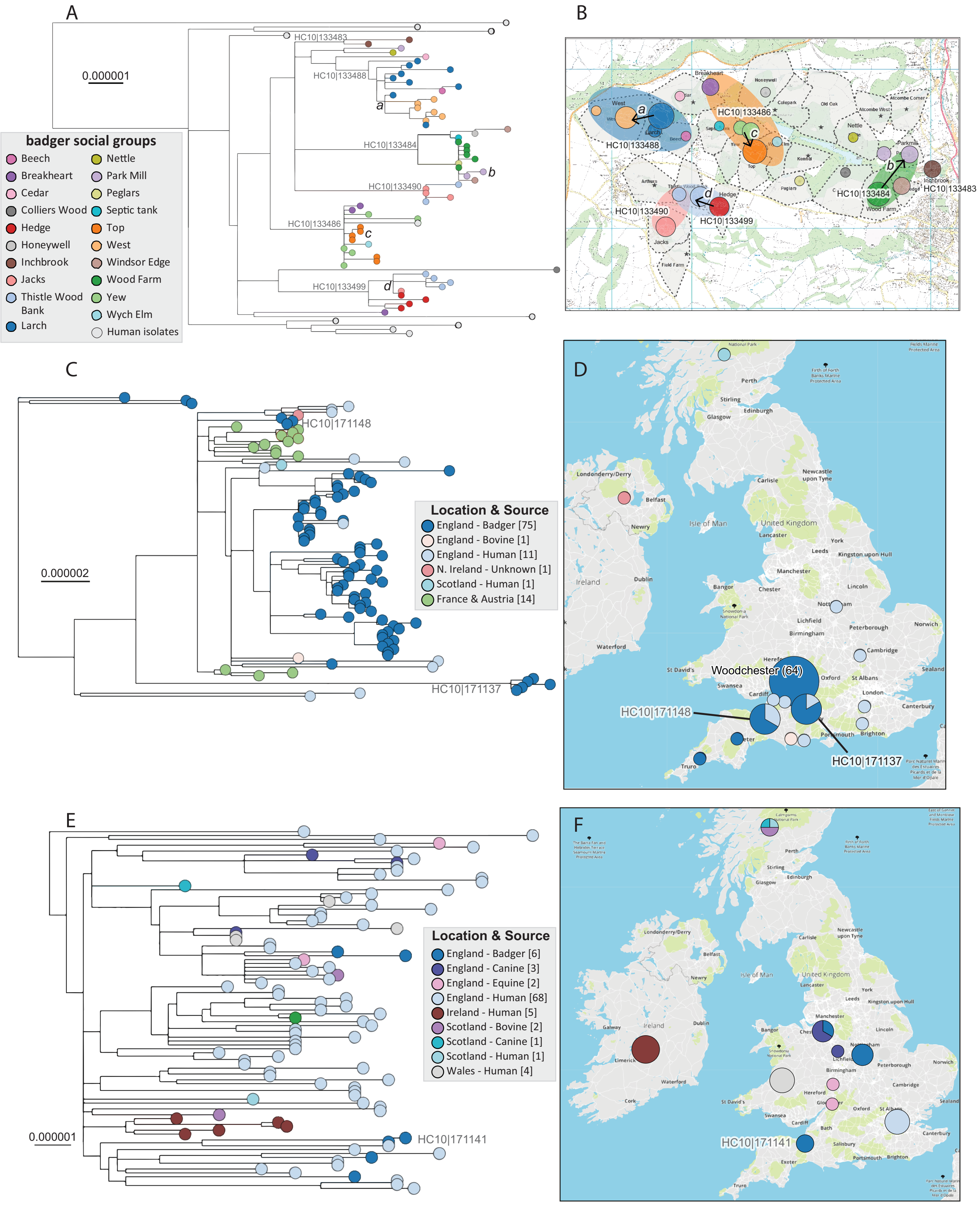
Data | View | Workspace | Experiment | Workspace: None | Rows Total: 30 | Filtered: 2 | Experimental Data: cgMLST V1 + HierCC

Uberstrain	✓		Name	Data Source	Lab Contact	Comment	ST	HC0 (indis...	HC2	HC5	HC10
■YER_AA2313AA			CO92-2003-version	GCF_000009065	Sanger Institute	Genotype: 1.ORI1e	159	159	159	159	92
▶YER_AA0760AA			CO92-2015-Los Alamos	SRR2148795	Los Alamos National Laboratory	Genotype: 1.ORI1e	646	646	175	175	92

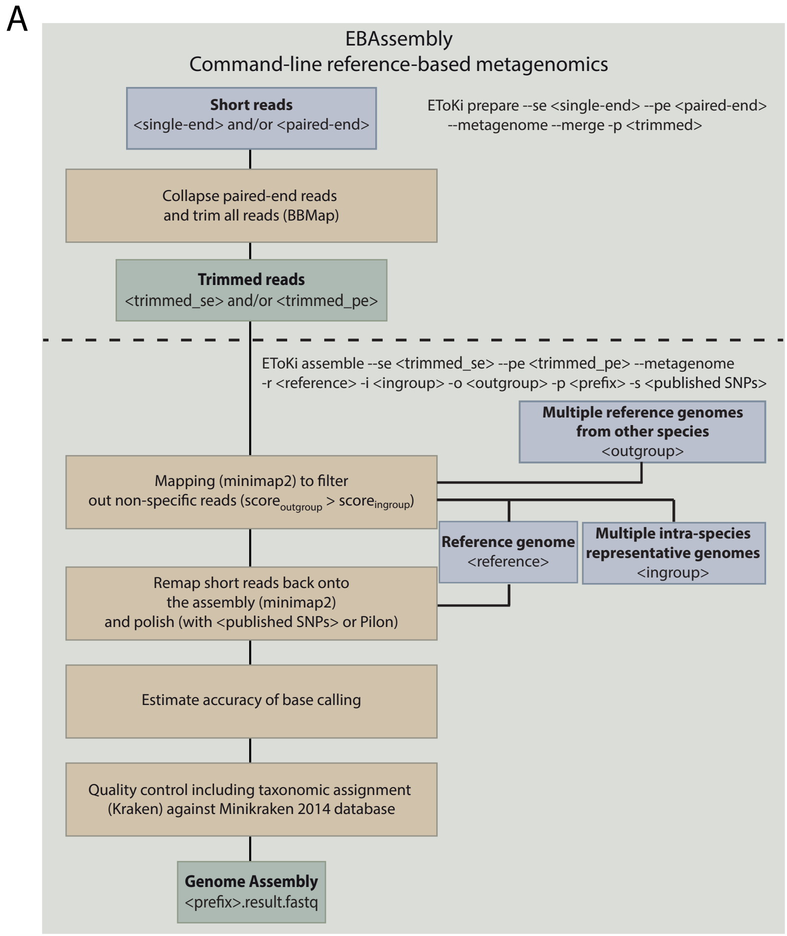
Data | View | Workspace | Experiment | Workspace: None | Rows Total: 30 | Filtered: 28 | Experimental Data: cgMLST V1 + HierCC

Uberstrain	✓		Name	Data Source	Lab Contact	Comment	ST	HC0 (indis...	HC2	HC5	HC10
■YER_AA2313AA			CO92-2003-version	GCF_000009065	Sanger Institute	Genotype: 1.ORI1e	159	159	159	159	92
▼YER_AA0760AA			CO92-2015-Los Alamos	SRR2148795	Los Alamos National Laboratory	Genotype: 1.ORI1e	646	646	175	175	92
├			CO92	MLST(Legacy)	Sanger Institute			ND	ND	ND	ND
├			CO92-2014Illumina + 454	SRR2180227	Los Alamos National Laboratory		218	218	175	175	92
├			Yp1980	SRR4072010	Northern Arizona University	CO92	1762	1762	175	175	92
├			Yp2005	SRR4072020	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2007	SRR4072024	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2009	SRR4072019	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2011	SRR4072027	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2013	SRR4072011	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2015	SRR4072017	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2017	SRR4072025	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2019	SRR4072031	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2020	SRR4072023	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2022	SRR4072028	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2023	SRR4072032	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2025	SRR4072014	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2030	SRR4072030	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2031	SRR4072012	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2034	SRR4072015	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2035	SRR4072018	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2037	SRR4072016	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2039	SRR4072022	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2040	SRR4072026	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2043	SRR4072033	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2045	SRR4072021	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2047	SRR4072029	Northern Arizona University	prairie dog passage	218	218	175	175	92
├			Yp2049	SRR4072013	Northern Arizona University	prairie dog passage	218	218	175	175	92

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.



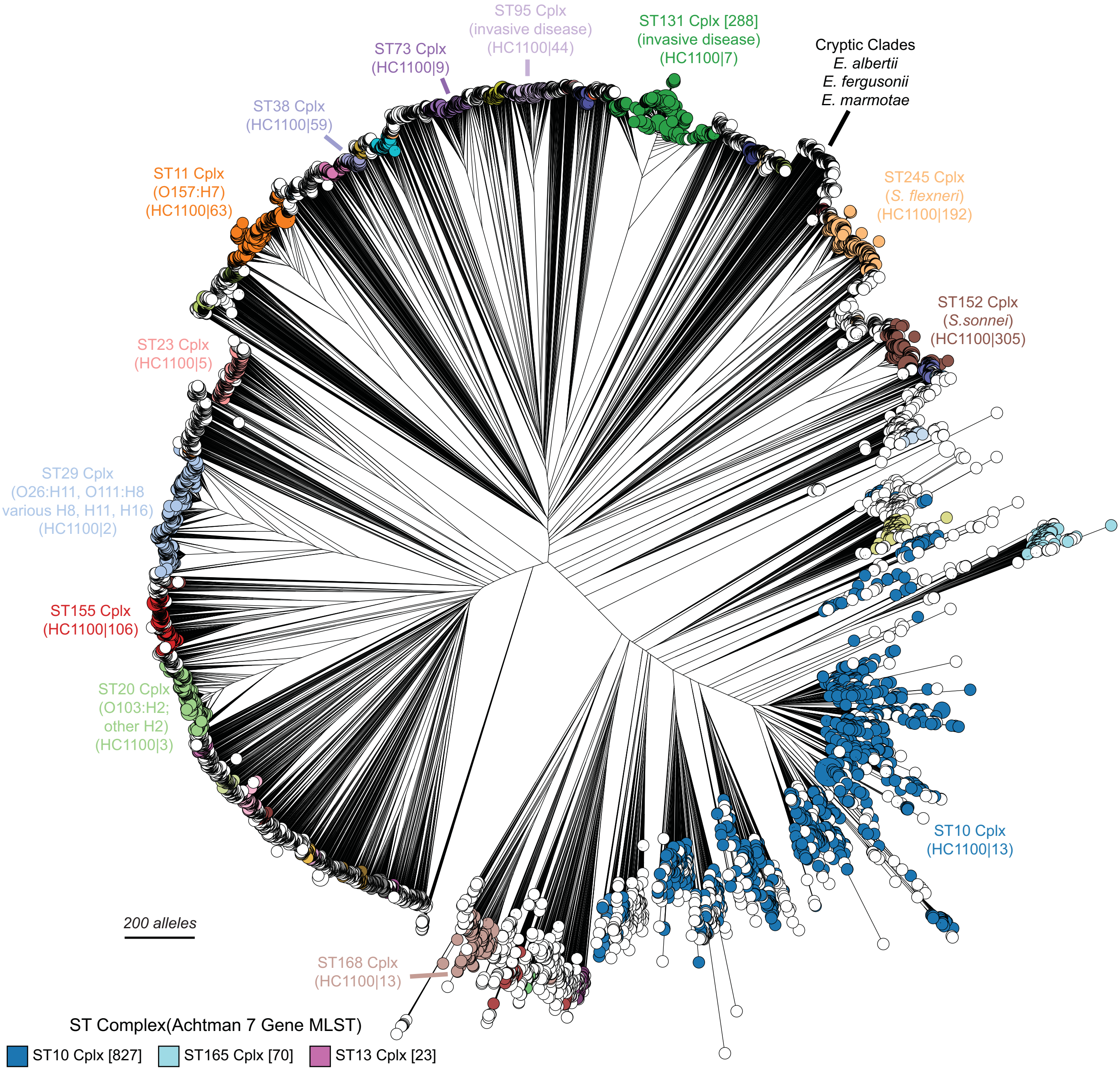
Supplementary Figure 3. Phylodynamics of isolates from badgers. A GrapeTree of Agama isolates was used to transfer individual subtrees plus their GPS coordinates and metadata to MicroReact (Argimon *et al.* 2016) as described in the GrapeTree Reference Manual (http://tinyurl.com/GrapeTreeRefManual). 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.


A
EBAssembly
Command-line reference-based metagenomics
Short reads
<single-end> and/or <paired-end>
Collapse paired-end reads and trim all reads (BBMap)
Trimmed reads
<trimmed_se> and/or <trimmed_pe>
Multiple reference genomes from other species
<outgroup>
Mapping (minimap2) to filter out non-specific reads (score$_{outgroup}$ > score$_{ingroup}$)
Reference genome
<reference>
Multiple intra-species representative genomes
<ingroup>
Remap short reads back onto the assembly (minimap2) and polish (with <published SNPs> or Pilon)
Estimate accuracy of base calling
Quality control including taxonomic assignment (Kraken) against Minikraken 2014 database
Genome Assembly
<prefix>.result.fastq

```
EToKi prepare --se <single-end> --pe <paired-end>
    --metagenome --merge -p <trimmed>
```

```
EToKi assemble --se <trimmed_se> --pe <trimmed_pe> --metagenome
-r <reference> -i <ingroup> -o <outgroup> -p <prefix> -s <published SNPs>
```

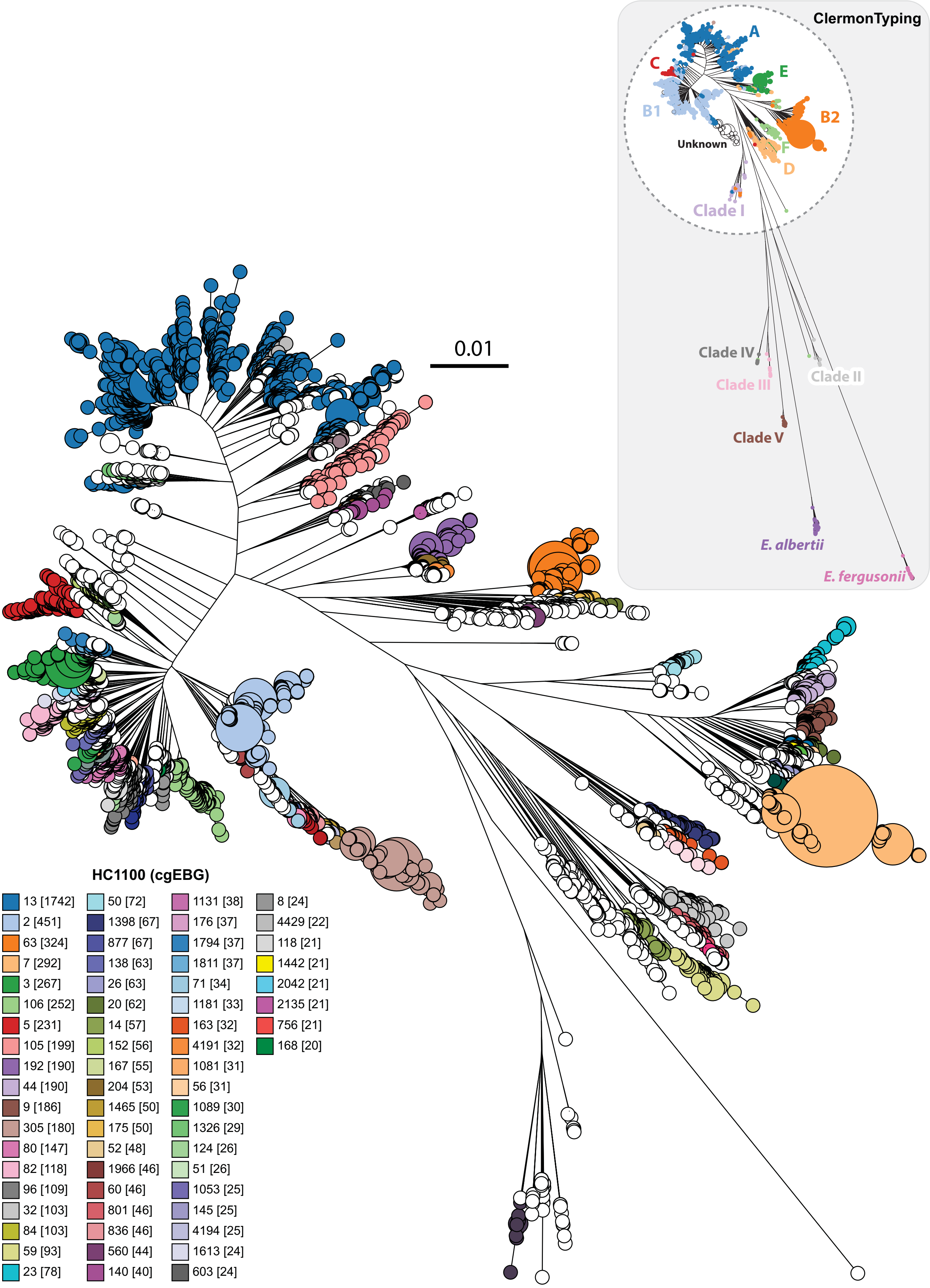
Supplementary Figure 4. Extracting aDNA assemblies from metagenomic sequences with the EBAssembly module of EToKi. EBAssembly includes functions for extracting genome-specific reads from metagenomic sequences which are only accessible in the stand-alone, command-line version of EToKi. The EToKi prepare 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.



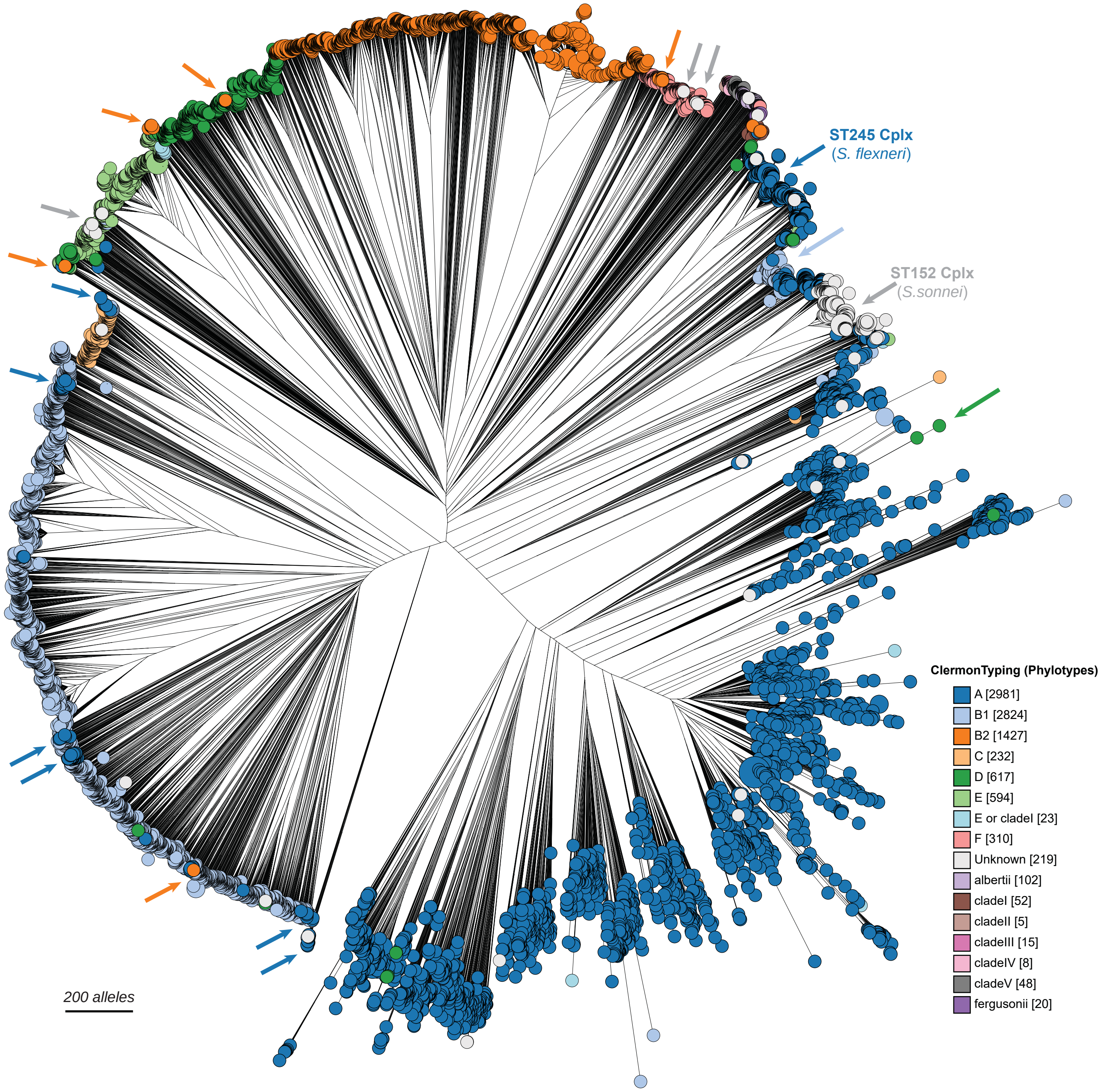
ST Complex(Achtman 7 Gene MLST)

- ■ ST10 Cplx [827]
- ■ ST29 Cplx [417]
- ■ ST11 Cplx [326]
- ■ ST245 Cplx [289]
- ■ ST131 Cplx [288]
- ■ ST20 Cplx [262]
- ■ ST155 Cplx [225]
- ■ ST23 Cplx [194]
- ■ ST73 Cplx [194]
- ■ ST95 Cplx [187]
- ■ ST152 Cplx [170]
- ■ ST168 Cplx [127]
- ■ ST38 Cplx [101]
- ■ ST86 Cplx [87]
- ■ ST101 Cplx [85]
- ■ ST278 Cplx [80]
- ■ ST12 Cplx [77]
- ■ ST206 Cplx [76]
- ■ ST69 Cplx [75]
- ■ ST165 Cplx [70]
- ■ ST648 Cplx [68]
- ■ ST14 Cplx [64]
- ■ ST243 Cplx [60]
- ■ ST405 Cplx [57]
- ■ ST350 Cplx [51]
- ■ ST59 Cplx [48]
- ■ ST32 Cplx [46]
- ■ ST469 Cplx [44]
- ■ ST205 Cplx [43]
- ■ ST31 Cplx [39]
- ■ ST156 Cplx [36]
- ■ ST354 Cplx [32]
- ■ ST448 Cplx [30]
- ■ ST46 Cplx [30]
- ■ ST147 Cplx [29]
- ■ ST28 Cplx [28]
- ■ ST398 Cplx [27]
- ■ ST446 Cplx [24]
- ■ ST13 Cplx [23]
- ■ ST40 Cplx [22]
- ■ ST226 Cplx [21]
- ■ ST148 Cplx [20]
- ■ ST349 Cplx [18]
- ■ ST399 Cplx [17]
- ■ ST568 Cplx [17]
- ■ ST590 Cplx [17]
- ■ ST394 Cplx [16]
- ■ ST270 Cplx [14]
- ■ ST538 Cplx [13]
- ■ ST522 Cplx [7]
- ■ ST122 Cplx [6]
- ■ ST280 Cplx [6]
- ■ ST582 Cplx [6]
- ■ ST467 Cplx [4]
- ■ ST272 Cplx [3]

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.