#165 Understanding genomic landscapes in EnteroBase with cgMLST & GrapeTree.

Nabil-Fareed Alikhan[1], Zhemin Zhou[1], Nina Luhmann[1], Catia Vaz[2,3], Alexandre P. Francisco[2,4], João André Carriço[5], Mark Achtman[1]

[1]Warwick Medical School, University of Warwick, Coventry, United Kingdom, [2]Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento (INESC-ID), Lisboa, Portugal, [3]ADEETC, Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Lisboa, Portugal, [4]Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal, [5]Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal

Motivation

High throughput sequencing continues to produce vasts amount of data for bacterial pathogens (Fig 1). Majority of which is provided by public health institutions for routine surveillance as short reads (Fig 2). Analysing such data is difficult without training in bioinformatics.

Results

EnteroBase [1] retrieves short reads from public databases, processes and provides access to a range of routine analyses for *Salmonella*, *Escherichia*, *Yersinia* & *Clostridioides* (Fig 3). EnteroBase includes consistent high-resolution genotyping by core genome multi-locus sequence typing (cgMLST) schemes and their intuitive visualization by GrapeTree [2]. Phylogenetic analyses via single nucleotide polymorphisms is also available on-demand. EnteroBase makes it easy to find strains of interest and construct datasets for analysis. The recent development of hierarchical clustering provides a method for finding genetically related clusters. We are also expanding EnteroBase to include genomes assembled from ancient metagenomes (Fig 4).

Conclusion

EnteroBase (http://enterobase.warwick.ac.uk) now provides a one-stop solution to analyses at all scales and provides users with the opportunity to quickly identify the genetic relatives of the bacteria whose genomes they have recently sequenced within the framework of the currently sequenced global diversity at previously unknown scales (>150,000 *Salmonella* genomes; >70,000 *Escherichia* genomes). EnteroBase supports teamwork through the sharing of work spaces and trees with other selected users or publicly.

Fig 1: Exponential increase of sequenced reads available for the genera in EnteroBase

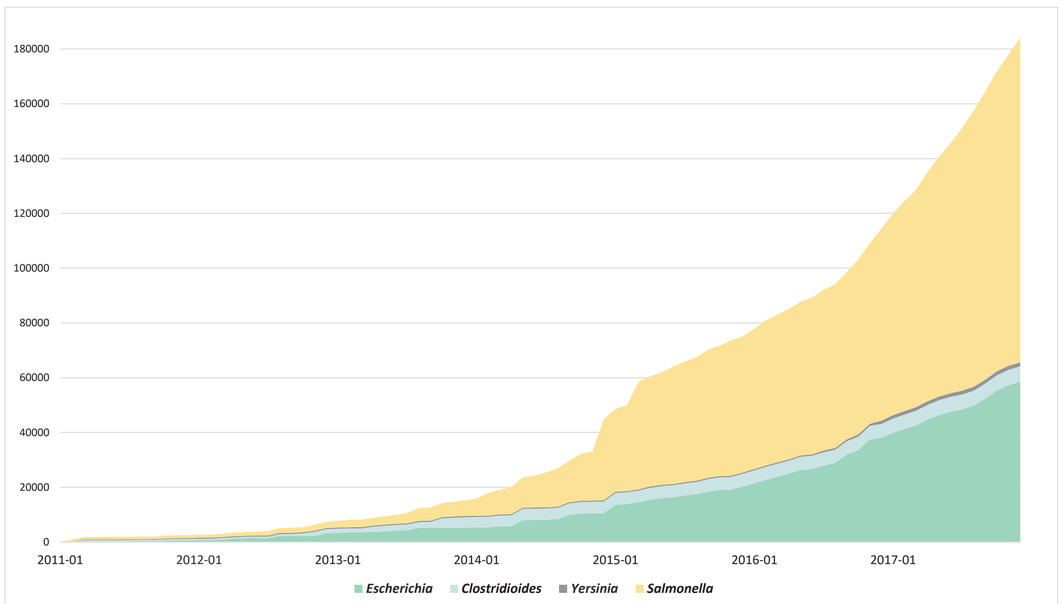


Fig 2: Top 5 Sequencing contributors for the genera in EnteroBase

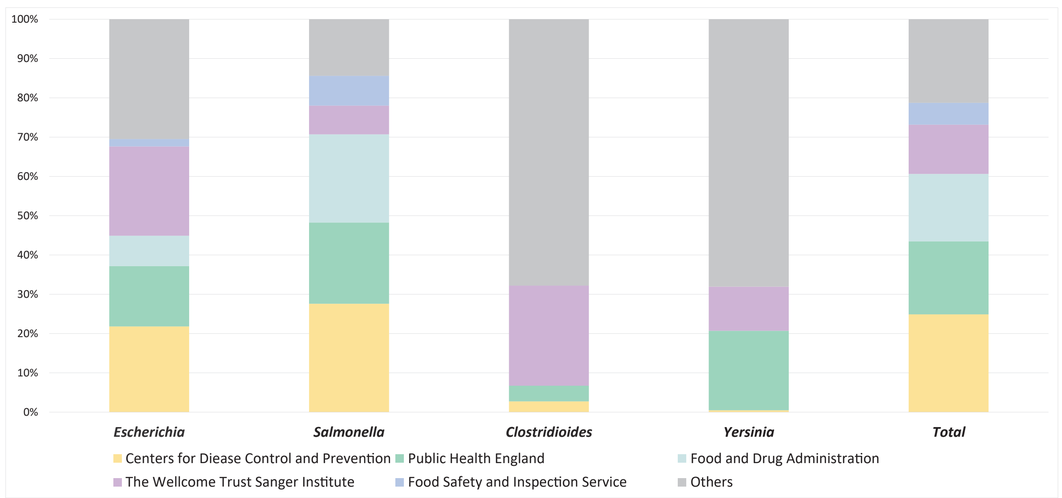
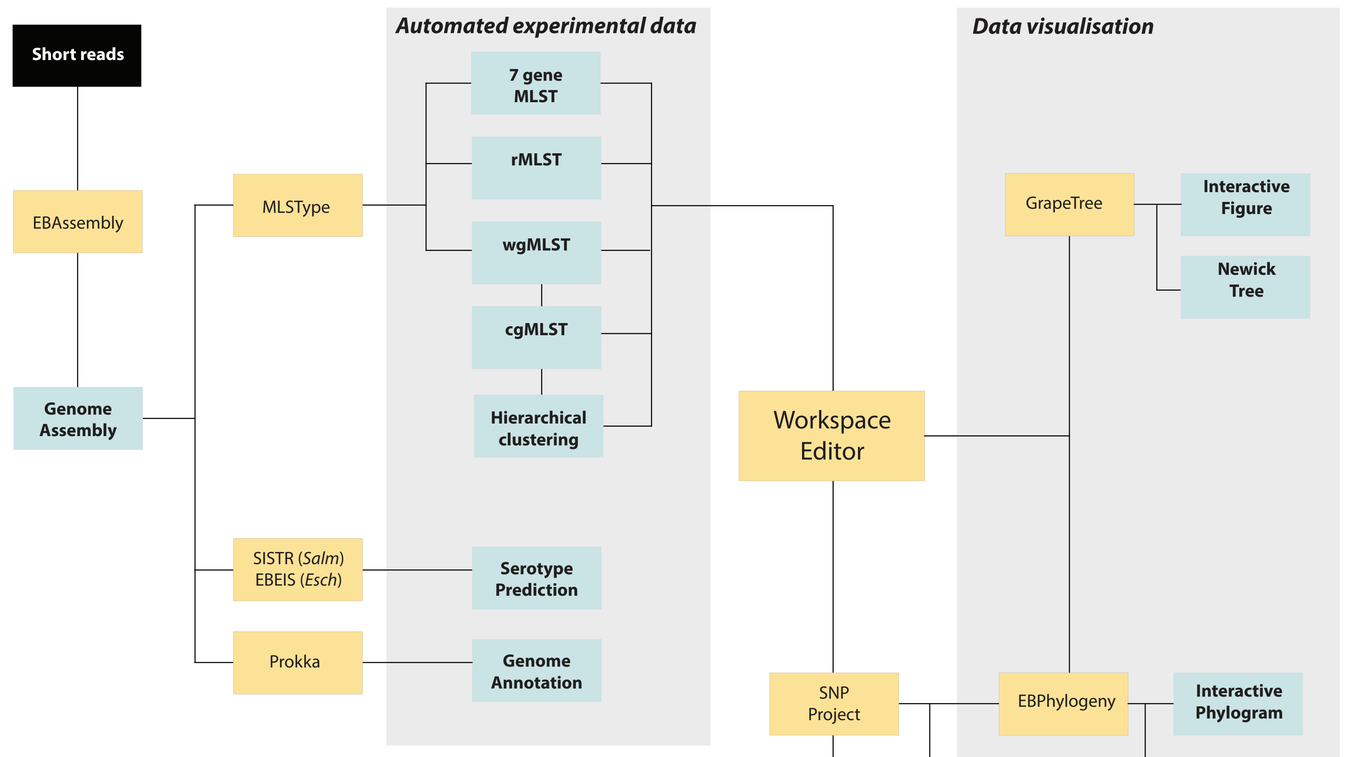


Fig 3: Overview of data analytic tools in EnteroBase and their outputs.

Short reads are fetched automatically from public databases or uploaded by users, which are then assembled by a standard pipeline (EBAssembly). A number of automated analyses are run based on the assembly. An overview of MLST methods can be found in the table. All data outputs (blue boxes) are available for users to download. EnteroBase records can be viewed through a workspace (see Box 1) and visualised in GrapeTree or as a phylogram. SISTR: Salmonella In silico Typing Resource [3]. EBEIS: EnteroBase Escherichia In silico Serotyping. Genome annotation with Prokka [4].

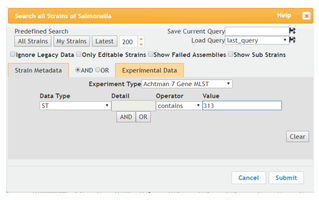


Summary of MLST schemes in EnteroBase (*Salmonella*)

Legacy MLST [5]	Ribosomal rMLST [6]	Core Genome cgMLST	Whole Genome wgMLST
7 Loci	51 Loci	3,002 Loci	21,065 Loci
Conserved Housekeeping genes	Ribosomal proteins	Core genes	Any coding sequence
Highly conserved; Low resolution	Highly conserved; Medium resolution	Variable; High resolution	Highly variable; Extreme resolution
Different scheme for each species/genus	Single scheme across tree of life	Different scheme for each species/genus	Different scheme for each species/genus

Box 1: How to navigate thousands of genomes in EnteroBase

1. Construct your own workspace of genomes based on a range of categories



Search EnteroBase with a variety of fields:

- Genotyping (e.g. MLST)
- Serotyping (e.g. Typhimurium)
- Metadata (Host, Geography, Isolation date)
- Accession codes (e.g. Biosample, Bioproject)

2. Expand your workspace with Hierarchical clustering

Hierarchical clustering applies single linkage clustering on cgMLST, providing discrete cluster groups at a range of thresholds. In *Salmonella,* clusters of strains with up to 20 alleles difference could indicate possible outbreaks, whereas *Salmonella* subspecies seem defined by clusters allowing up to 2100 alleles differece.

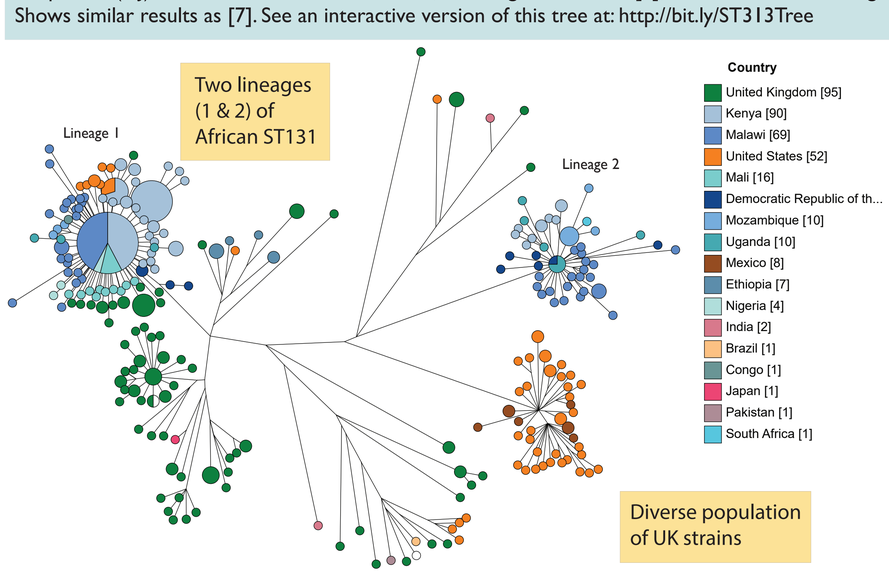
Data | View | Analysis | Workspace — Workspace:st131_example Rows Total:1956 Filtered:1150

Uberstrain	Name	Data Source	Barcode	Source	Collection ...	Location	Serovar	<1	<5	<20	<50	<100	<400	Clonal complex	Subsp.
SAL_DA4...	M2907772	ERR024420	SAL_DA484...	Human	2003	Mozambique	Typhimurium	17833	17833	728	728	728	2	2	2
SAL_DA4...	DT2B	ERR024398	SAL_DA486...	Human	2009	United Kingdo	Typhimurium	17817	17817	1611	1611	728	2	2	2
SAL_DA4...	A38596	ERR023792	SAL_DA487...	Human	2005	Malawi	Typhimurium	17816	17816	728	728	728	2	2	2
SAL_DA4...	J3	ERR023793	SAL_DA487...	Human	2005	Mali	Typhimurium	17815	17815	728	728	728	2	2	2
SAL_DA4...	J17	ERR023794	SAL_DA487...	Human	2002	Mali	Typhimurium	17814	729	728	728	728	2	2	2
SAL_DA4...	J27	ERR023795	SAL_DA487...	Human	2002	Mali	Typhimurium	17813	17813	728	728	728	2	2	2
SAL_DA4...	C13184	ERR023796	SAL_DA487...	Human	2002	Malawi	Typhimurium	17812	729	728	728	728	2	2	2
SAL_DA4...	A39051	ERR023797	SAL_DA487...	Human	2002	Malawi	Typhimurium	17811	729	728	728	728			
SAL_DA4...	A39129	ERR023798	SAL_DA487...	Human	2006	Malawi	Typhimurium	17810	17810	728	728	728			
SAL_DA4...	A39155	ERR023799	SAL_DA487...	Human	2006	Malawi	Typhimurium	17809	17809	728	728	728			
SAL_DA4...	P78	ERR023800	SAL_DA487...	Human	2005	Mali	Typhimurium	17808	17808	728	728	728			
SAL_DA4...	P62	ERR023801	SAL_DA487...	Human	2005	Mali	Typhimurium	17807	17807	728	728	728			
SAL_DA4...	P73	ERR023802	SAL_DA488...	Human	2005	Mali	Typhimurium	17806	10743	728	728	728			
SAL_DA4...	P51	ERR023803	SAL_DA488...	Human	2006	Mali	Typhimurium	17805	17805	728	728	728			
SAL_DA4...	A018	ERR023766	SAL_DA489...	Human	1997	Malawi	Typhimurium	17793	17793	14945	14851	14851	2	2	2
SAL_DA4...	C2167	ERR023767	SAL_DA489...	Human	2000	Malawi	Typhimurium	17792	17792	14945	14851	14851	2	2	2

Closely related strains (< 20 cgMLST alleles different)

3. Launch analyses (e.g. GrapeTree) for genomes in your workspace.

GrapeTree (NJ) of 380 S. enterica ST313 strains, including strains from [7], based on EnteroBase cgMLST. Shows similar results as [7]. See an interactive version of this tree at: http://bit.ly/ST313Tree



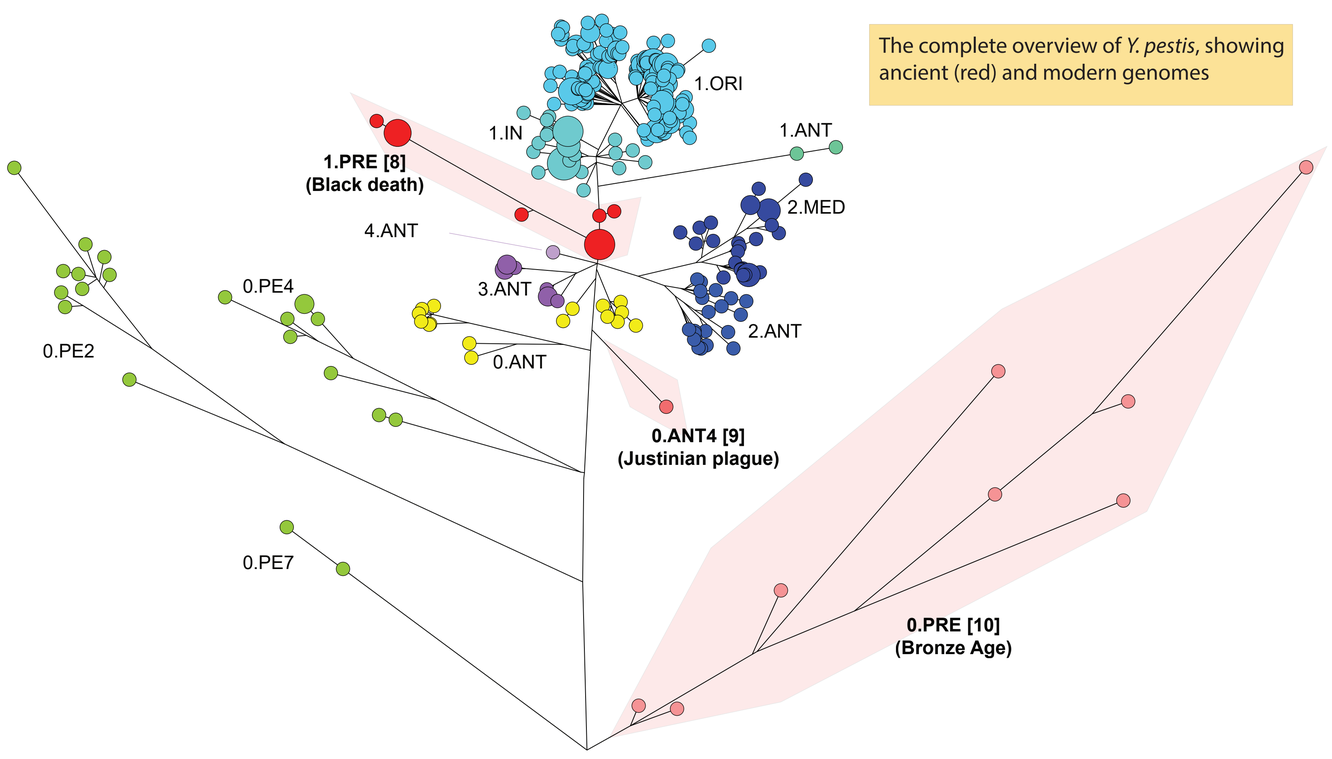
4. Share your workspace privately with buddies or make it public

See an example workspace at: http://bit.ly/st313example. Workspaces can be:

- Edited, saved and later reloaded;
- Made public so you can link it in a publication

Fig 4: Modern genomes and genomes from ancient metagenomes of *Yersinia pestis* in EnteroBase

A maximum-likelihood phylogeny of 6,811 SNPs in the core genomic regions of 714 modern *Yersinia pestis* genomes and 22 ancient strains [8-10]. The circles present tip nodes in the tree, and are connected by branches as in the phylogeny. These nodes are scaled by the numbers of genomes assigned to them and color-coded by the corresponding lineages. The names of the lineages are also shown nearby the nodes. The numbers of ancient strains in each of the three lineages (pink polygon) are indicated in brackets after the names of lineages (bold). See an interactive version of this tree at: http://bit.ly/PestisTree



The complete overview of *Y. pestis*, showing ancient (red) and modern genomes

References

[1] Alikhan et al. 2018. PLoS Genetics 14 (4): e1007261. & http://enterobase.warwick.ac.uk
[2] Zhou et al. 2018. Genome Research 28:1395-1404. & https://achtman-lab.github.io/GrapeTree/
[3] Yoshida et al. 2016. PLoS One 11:1.
[4] Seemann 2014. Bioinformatics 30:14
[5] Achtman et al. 2012 PLoS Pathogens 8:6
[6] Jolley et al. 2012 Microbiology 158:4
[7] Ashton et al. 2017. Genome Medicine 9:92.
[8] Bos et al. 2011. Nature 478:506-510 & Spyrou et al. 2016 Cell Host & Microbe 19:6
[9] Feldman et al. 2016. Molecular Biology and Evolution 33:11 & Andrades et al. 2017 Current Biology 27:23
[10] Rasmussen et al. 2015 Cell 163:571-582

EnteroBase is funded by the BBSRC (BB/L020319/1).
Additional grant support was from the Wellcome Trust (202792/Z/16/Z)

Warwick Medical School
UNIVERSIDADE DE LISBOA