Institution: University of Warwick

Unit of Assessment: UOA1 – Clinical medicine

Title of case study: EnteroBase, a platform for analysis of bacterial genomes to trace foodborne bacterial disease outbreaks

Period when the underpinning research was undertaken: 2014- 2020

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
Professor Mark Achtman	Professor of Bacterial Population Genetics	May 2013- present

Period when the claimed impact occurred: 2017- 2020

Is this case study continued from a case study submitted in 2014? N

1. Summary of the impact (indicative maximum 100 words)

The detection, tracking and notification of food-borne bacterial disease outbreaks has been challenging for national public health reference laboratories due to within-species diversity, inconsistent nomenclature and reliance on phenotype rather than genotype. To address this, Professor Achtman Fellow of the Royal Society (FRS), developed and launched Enterobase in 2016, a world-class, genome database for bacterial pathogens *Salmonella*, *Escherichia*, *Helicobacter, Vibrio, Yersinia*, *Clostridioides* and *Streptococcus*. It provides a standardised and scalable typing method enabling the genomic investigation of food-borne disease outbreaks, and is used by national public health reference laboratories, including those in England, Scotland, Ireland, France, Denmark, Canada, China and South Africa. Its powerful bioinformatic tools empower public health scientists to promptly identify the causes of outbreaks, track multi-country outbreaks, inform prompt public health decision making and subsequent recall of contaminated products in increasingly global food chains. For example, use of Enterobase enabled the *Salmonella enterica* serotype Poona 2019 outbreak in infants in France to be linked to infant formula and led to contamination investigations and product recall.

2. Underpinning research (indicative maximum 500 words)

The huge within-species diversity of bacterial pathogens presents challenges for microbiologists and public health scientists in recognising and classifying strain types. The global introduction of high throughput genomic sequencing resulted in extensive valuable information but it could only be evaluated by expert bioinformaticians.

Professor Mark Achtman FRS, a leading bacterial population geneticist, developed Multilocus Sequence Typing (MLST) of *Salmonella* and *Escherichia,* and established that phenotypic methods such as serotyping should be replaced by MLST. He also proposed that MLST should be conducted at the genomic level, but appreciated that many microbiologists would struggle with such analyses. From 2014 and funded by the BBSRC, Achtman, now at the University of Warwick, set out together with post-doctoral fellows Dr Martin Sergeant, Dr Nabil-Fareed Alikhan and Dr Zhemin Zhou,to create a powerful but user-friendly, online website and database to overcome this challenge.

Designed to support analyses ranging from 7-gene MLST to whole genomes, Enterobase was developed as a unique web-based platform for the automated assembly of short read sequences (Illumina) into genomic contigs which were made available for public access and analysis together with their metadata and genotyping data. The database contains more than 450,000 genomes for the genera *Salmonella*, *Escherichia*, *Yersinia*, *Clostridiodes*, *Helicobacter*, *Vibrio*,

and *Moraxella*. This genotyping now includes assigning allelic designations to all 1500-3000 core genes (cgMLST) and hierarchical clustering (HierCC) of their genomic similarities at multiple levels of resolution. High resolution HierCC can be used for identifying food-borne disease outbreaks. Intermediate resolution HierCC reliably identifies natural populations, whereas low resolution HierCC can replace traditional taxonomic measures. Further work at Warwick has enabled *EnteroBase* to combine genomes from modern cultivated bacteria with genomes calculated from metagenomic analyses, including from ancient DNA. EnteroBase enables epidemiology and population genetics investigations of isolates from distinct geographical sources and over extended time scales, and has been used to describe the population genomic structure of *Salmonella* [3.1] and reconstruct the long-term evolutionary history of bacterial pathogens [3.2, 3.3].

Novel graphical tools support facile interrogation of genomic relationships among bacterial pathogens. Achtman, in collaboration with Zhou and Sergeant, created GrapeTree, overcoming the difficulties of using phylograms in the visual presentation of large numbers of genotypes. The novel minimum spanning tree algorithm can depict genetic relationships between 100,000 genomes or more, accommodating high levels of missing data [3.4].

EnteroBase's hierarchical clustering can be used to identify populations of bacterial pathogens in multiple bacterial genera at all epidemiological levels, including to inform understanding of the transmission of *Clostridioides difficile*, the primary infectious cause of antibiotic-associated diarrhoea [3.5]., Achtman, Alikhan, and Zhou have documented the use of EnteroBase to reveal broad population structures with HierCC, as well as describe a historical collection of 10,000 new genomes isolated between 1891-2010 in 73 different countries. The Genomic DNA was sequenced and analysed using EnteroBase. The analyses demonstrate that discrete clusters with geographical specificity can be reliably recognized by hierarchical clustering approaches while confirming the polyphyletic nature of multiple serovars [3.6].

Key University of Warwick staff: Professor Mark Achtman (May 2013- present); Dr Martin Sergeant (2014- 2018); Dr Nabil-Fareed Alikhan (Senior Research Fellow 2014- 2018); Dr Zhemin Zhou (Senior Research Fellow, May 2013- present)

3. References to the research (indicative maximum of six references)

[3.1] **Alikhan, NF, Zhou, Z, Sergeant, MJ,** and **Achtman, M** (2018) A genomic overview of the population structure of *Salmonella*. PLoS Genetics, 14 (4). e1007261. doi:10.1371/journal.pgen.1007261

[3.2] **Zhou, Z,** Lundstrøm, I, Tran-Dien, A, Duchêne, S, **Alikhan, NF, Sergeant, MJ**, Langridge, G, Fotakis, AK, Nair, S, Stenøien, HK, Hamre, SS, Casjens, S, Christophersen, A, **Quince**, **C**, Thomson, NR, Weill, FX, Ho, SYW, Gilbert, MTP, and **Achtman, M** (2018) Pan-genome analysis of ancient and modern *Salmonella enterica* demonstrates genomic stability of the invasive Para C lineage for millennia. Current Biology, 28 (15). pp. 2420-2428. doi:10.1016/j.cub.2018.05.058

[3.3] **Zhou, Z, Alikhan, NF., Mohamed, K, Fan, Y,** Agama Study Group, & **Achtman, M** (2020). The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. Genome research, *30* (1), pp. 138–152. doi:10.1101/gr.251678.119

[3.4] **Zhou, Z, Alikhan, NF, Sergeant, MJ, Luhmann, N**, Vaz, C, Francisco, AP, Carriço, JA and **Achtman, M** (2018) GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Research, 28 (9). pp. 1395-1404. doi:10.1101/gr.232397.117

[3.5] Frentrup, M, **Zhou, Z**, Steglich, M, Meier-Kolthoff, JP, Göker, Markus, R, Thomas, Bunk, B, Spröer, C, Overmann, J, Blaschitz, M, Indra, A, von Müller, L, Kohl, TA., Niemann, S, Seyboldt, C, Klawonn, F, Kumar, N, Lawley, TD., García-Fernández, S, Cantón, R, del Campo, R, Zimmermann, O, Groß, U, **Achtman, M** and Nübel, U. (2020) A publicly accessible database for Clostridioides difficile genome sequences supports tracing of

transmission chains and epidemics. Microbial Genomics 6(8) doi:10.1099/mgen.0.000410

[3.6] **Achtman, M., Zhou, Z., Alikhan, NF., Tyne, W**., Parkhill, J., Cormican, M., Chiou, CS., Torpdahl, M., Litrup, E., Prendergast, DM., Moore, JE., Strain, S., Kornschober, C., Meinersmann, R., Uesbeck, A., Weill, FX., Coffey, A., Andrews-Polymenis, H., Curtiss R., Fanning, S. (2020) Genomic diversity of Salmonella enterica -The UoWUCC 10K genomes project, Wellcome Open Res 2020, 5:223, doi: 10.12688/wellcomeopenres.16291.1

Key grants

(PI) Mark Achtman; EnteroBase: A Powerful, User-Friendly Online Resource for Analyzing and Visualizing Genomic Variation within Escherichia coli and Salmonella enterica; BBSRC BB/L020319/1; 2014 –2019; £1,005,462

(PI) Mark Achtman; Deep evolutionary history of bacterial pathogens; Wellcome Trust 202792/Z/16/Z 2016-2021, £1,944,236

4. Details of the impact (indicative maximum 750 words)

Enteric pathogens such as *Escherichia coli* and *Salmonella enterica* are important causes of human and animal infections and are a significant public health challenge. Salmonella is a key global cause of diarrhoeal diseases, and estimated to cause 93,000,000 enteric infections annually (Majowicz et al, 2010). National Reference Laboratories are responsible for investigating food-borne bacterial disease outbreaks to inform control and prevention measures and inform national surveillance. Public health scientists typically used methods such as serotyping using pulsed-field gel electrophoresis to identify strains. The methods were often time-consuming with additional challenges in the reliability of subtyping and the use of inconsistent nomenclature between regions and countries. Outbreaks are often multi-country and therefore require high-resolution, accessible, and replicable isolate typing schemes. Adopted by National Reference Laboratories, Achtman's Enterobase has improved and accelerated the identification and tracking of outbreaks.

EnteroBase was first made publicly accessible in 2016 as a genotyping website for selected enteric pathogens. It enables rapid analysis of bacterial genomics using a common and stable nomenclature. Microbiologists with limited bioinformatic skills can use its user-friendly software platform to upload short reads, assemble and genotype genomes, and immediately investigate their genomic relationships and population structures. It now has more than 3,500 users in public health organisations and industrial and allied research laboratories globally. On average, 400 of them visit its web pages per month. Achtman has supported the adoption of EnteroBase by the provision of training and advice in its use.

In 2017, PulseNet International, the global network for laboratory-based surveillance for food-borne disease outbreaks confirmed the adoption of whole genome multilocus sequence typing (wgMLST) to standardise subtyping thanks to its higher resolution information. The scheme of genes selected for whole genome MLST in EnteroBase was selected as the global master scheme for *Escherichia coli* and *Salmonella enterica* [5.1]. The MLST schemes maintained by EnteroBase serve as the reference schemes for the public databases for molecular typing and microbial genome diversity (PubMLST), and used within the MLST module in CLC Genomics Workbench (Qiagen, Germany) and SeqSphere (Ridom, Germany)

The European Centre for Disease Prevention and Control's (ECDC's) 2018 external quality assessments for Salmonella typing (2018) and Shiga toxin-producing Escherichia coli (2019) state: "*For inter-laboratory comparability and communication about cluster definitions, cgMLST using a standard scheme (e.g. Enterobase) gives a very high degree of homogeneity in the results, whereas the use of non-standardised SNP analysis may be more challenging for comparison and communication between laboratories*", with participating labs using Enterobase for allele analysis [5.2]. Enterobase has been used globally by national reference laboratories including Public Health England; the Scottish Reference Microbiology Laboratories; the Irish

National Reference Laboratory Services for Salmonella, Shigella, *Listeria monocytogenes* and Carbapenemase Producing Enterobacteriaceae; the French National Reference Centre for *Escherichia coli-Shigella-Salmonella,* Institut Pasteur; Statens Serum Institut, Denmark; the National Microbiology Laboratory, Canada; the China Center for Disease Control and Prevention and the Centre for Enteric Diseases, National Institute for Communicable Diseases South Africa [5.3-5.10].

Enterobase's hierarchical clustering of core genome MLST sequence types facilitates immediate isolate characterisation and identification of close relatives of bacteria within those genera. Consistent nomenclature saves National Reference Laboratories valuable resources and time when liaising with the Epidemic Intelligence Information System, hosted by the European Centre for Disease Prevention and Control and improves communication between public health scientists, epidemiologists, medical and food safety officers, inspectors, and healthcare and environmental health professionals. The French National Reference Centre for *Escherichia coli-Shigella-Salmonella* (FNRC-ESSS) at Institut Pasteur has used EnteroBase to analyse more than 20,000 *Escherichia coli*, *Shigella* and *Salmonella* isolates since April 2017. The Head of the French National Reference Centre confirms "*the cgMLST hierarchical clustering tool has in particular revolutionized the typing of these pathogens and has now become the key information used when dealing with the epidemiologists from Public Health France and beyond"* [5.5]. The Statens Serum Institut, Denmark has used EnteroBase since 2016 for all *Salmonella*, *E. coli* (STEC), and *Clostridium* isolates received and shares sequence types with their local hospitals: *"the hierarchical clustering within EnteroBase is important to us because it gives us a means of fast communication which is crucial in outbreak investigations and further using cluster names makes our lives easier as no further analysis is required"* [5.6].

EnteroBase's tools such as GrapeTree enable users to help identify a possible source of the outbreak and perform genomic comparisons. Genomic comparisons are particularly helpful in the investigation of multi-national outbreaks and enable rapid exchange of harmonised data and reduce the hazardous exchange of isolates for comparison. Enterobase was used to confirm the close genetic relationship in the 2014-2018 multi-country *S. Agona* outbreak isolates, which along with seasonal peak timings indicated an intermittent common source (European Centre for Disease Prevention and Control/European Food Safety Authority. Multi-country outbreak of Salmonella Agona possibly linked to ready-to-eat food – 26 July 2018. Stockholm and Parma: ECDC/EFSA; 2018). The Director of the Irish National Reference Laboratory Services for Salmonella, Shigella, *Listeria monocytogenes* and Carbapenemase Producing Enterobacteriaceae (CPE) confirmed that in "*one recent outbreak of Salmonella enterica we were able to identify 3 closely related sequences in another jurisdiction which was relevant to the outbreak investigation. In the absence of EnteroBase we would not have been able to make that connection in such a rapid and convenient manner*" [5.4]. The Head of the French National Reference Centre affirms that "*During several investigations here in France, EnteroBase has allowed us to identify within minutes several outbreak-related cases in other European countries*" [5.5].

Genotyping and comparisons to earlier outbreak surveillance data increases resolution for epidemiological tracking of bacterial disease outbreaks and control. The Scottish Reference Microbiology Laboratories have used EnteroBase since 2017 for *Salmonella* and *Shigella,* using its hierarchical clustering function and SNP trees. The Principal Clinical Scientist confirms in 2019 they "*sequenced more than 1,400 isolates of Salmonella and Shigella from human and veterinary cases. Almost half of these cases were linked to new or previously identified clusters on the basis of cgMLST data provided by Enterobase. This constitutes a significant step forward in our ability to detect linked cases compared with traditional methods. The resolution obtained, particularly for the more common Salmonella enteritidis infections which have until now been extremely difficult to subtype, has greatly increased the statistical power essential to effective epidemiological investigation, by being able to forensically define an isolate as being part of, or being unrelated, to an outbreak. This allows more focus on confirmed cases, maximising the efficient use of investigative resources"* [5.3].

The efficiency of EnteroBase is highlighted in the 2018-2019 investigation of a *Salmonella enterica* serotype Poona outbreak in infants in France. All isolates from children under 3 years of age underwent serotyping and whole genome sequencing by the French National Reference Centre for *Escherichia coli, Shigella* and *Salmonella* to rapidly identify the *S. Poona* genome cluster affiliation and inform the epidemiological investigation interviews. Initial epidemiological investigations identified consumption of the same brand of rice-based powdered infant formula and triggered a recall of the formula manufactured at the identified facility on 24 January 2019. An urgent enquiry in the European Centre for Disease Prevention and Control Epidemic Intelligence Information System for Food and Waterborne Diseases and Zoonoses identified two further confirmed cases in Belgium and Luxemburg. The rapid response was attributed to the routine whole genome sequencing of human isolates received at the French National Reference Centre. The genomic analysis using EnteroBase provided further advantages through the linking of isolates to an outbreak in in 2010–11 also connected to the identified infant formula facility. The 2010–11 and 2019 genetical relationship (EnteroBase cgMLST profile HC20-44730) has led to further investigations to identify the persistent source of contamination. [5.6, 5.10].

5. Sources to corroborate the impact (indicative maximum of 10 references)

[5.1] PulseNet International recommendation of EnteroBase as the core genome database for Escherichia. coli and Salmonella enterica: Nadon C, Van Walle I, Gerner-Smidt P, et al. (2017) PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro Surveill., 22 (23).
[5.2] European Centre for Disease Prevention and Control. Eighth external quality assessment scheme for Salmonella typing. Stockholm: ECDC; 2018; and External quality assessment scheme for typing of Shiga toxin-producing Escherichia coli. Stockholm: ECDC; 2019.
[5.3] Factual Statement from the Principal Clinical Scientist, Scottish Reference Microbiology Laboratories
[5.4] Factual Statement from the Director of the Irish National Reference Laboratory Services for Salmonella, Shigella, *Listeria monocytogenes* and Carbapenemase Producing Enterobacteriaceae;
[5.5] Factual Statement from the Head of the French National Reference Centre for *Escherichia coli-Shigella-Salmonella,* Institut Pasteur
[5.6] Factual Statement from the Senior Researcher, Statens Serum Institut, Denmark
[5.7] Factual Statement from the Chief, Innovation and Application Development Section, Division of Enteric Diseases, National Microbiology Laboratory, Canada
[5.8] Factual Statement from the Deputy Director, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention
[5.9] Factual Statement from the Principal Medical Scientist, Centre for Enteric Diseases, National Institute for Communicable Diseases South Africa
[5.10] Jones G, Pardos de la Gandara M, Herrera-Leon L, 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):1900161. doi:10.2807/1560-7917.ES.2019.24.13.1900161