Genome Biology

METHOD

Open Access

An accessible, efficient and global approach for the large-scale sequencing of bacterial genomes

Blanca M. Perez-Sepulveda[1*], Darren Heavens[2†], Caisey V. Pulford[1†], Alexander V. Predeus[1], Ross Low[2], Hermione Webster[1], Gregory F. Dykes[1], Christian Schudoma[2], Will Rowe[1,3], James Lipscombe[2], Chris Watkins[2], Benjamin Kumwenda[4], Neil Shearer[2], Karl Costigan[1], Kate S. Baker[1], Nicholas A. Feasey[5,6], Jay C. D. Hinton[1*], Neil Hall[2,7,8*] and The 10KSG consortium

* Correspondence: blanca.m.perez@gmail.com; jay.hinton@liverpool.ac.uk; neil.hall@earlham.ac.uk
†Darren Heavens and Caisey V. Pulford contributed equally to this work.
[1]Institute of Infection, Veterinary & Ecological Sciences, University of Liverpool, Liverpool, UK
[2]Earlham Institute, Norwich Research Park, Norwich, UK
Full list of author information is available at the end of the article

Abstract

We have developed an efficient and inexpensive pipeline for streamlining large-scale collection and genome sequencing of bacterial isolates. Evaluation of this method involved a worldwide research collaboration focused on the model organism *Salmonella enterica*, the 10KSG consortium. Following the optimization of a logistics pipeline that involved shipping isolates as thermolysates in ambient conditions, the project assembled a diverse collection of 10,419 isolates from low- and middle-income countries. The genomes were sequenced using the LITE pipeline for library construction, with a total reagent cost of less than USD$10 per genome. Our method can be applied to other large bacterial collections to underpin global collaborations.

Keywords: Thermolysates, *Salmonella*, Whole-genome sequencing, iNTS

Background

Whole-genome sequencing (WGS) is an important tool that has revolutionized our understanding of bacterial disease over the past decade [1–4]. Recognizing the immense advantages that WGS data provides for surveillance, functional genomics, and population dynamics, both public health and research communities have adopted genome-based approaches.

Until recently, large-scale bacterial genome projects could only be performed in a handful of sequencing centers around the world. Here, we aimed to make this technology accessible to bacterial laboratories worldwide. The high demand for sequencing human genomes has driven down the costs of sequencing reagents to below USD$1000 per sample [5–7]. However, the genome sequencing of thousands of microorganisms has remained expensive, largely due to costs associated with sample transportation and library construction.

The number of projects focused on sequencing the genomes of collections of key pathogens has increased markedly over recent years. While the first *Vibrio cholerae* next-generation WGS study was based on 23 genomes [8], a recent study involved 1070 isolates from 45 African countries [9] and identified the origin of the most recent cholera pandemic. *Mycobacterium tuberculosis*, another major human pathogen, was originally sequenced on the 100-isolate scale in 2010 [10], while recent publications used 3651 [11] or 10,209 [12] genomes to evaluate the accuracy of antibiotic resistance prediction. Other successful large-scale next-generation WGS projects for pathogens include *Salmonella*, *Shigella, Staphylococcus*, and pneumococcus (*Streptococcus pneumoniae*) [13–16]. Indeed, a recent study by Achtman et al. [17] sequenced 9591 *Salmonella* genomes isolated mainly from water and animal sources in the USA, Europe, and Taiwan. This collaboration between the University of Warwick (UK) and the University College Cork (Ireland) focused on the analysis of three serovars, mostly obtained from the environment, animals, and human feces, adding an important level of diversity to publicly available genomes for the *Salmonella* community.

One of the most significant challenges facing scientific researchers in low- and middle-income (LMI) countries is the streamlining of surveillance with scientific collaborations. For a combination of reasons, the regions associated with the greatest burden of severe bacterial disease have inadequate access to WGS technology and have had to rely on expensive and bureaucratic processes for sample transport and sequencing. This has prevented the adoption of large-scale genome sequencing and analysis of bacterial pathogens for public health and surveillance in LMI countries [18]. Here, we have established an efficient and relatively inexpensive pipeline for the worldwide collection and sequencing of bacterial genomes. To evaluate our pipeline, we used the model organism *Salmonella enterica*, a pathogen with a global significance [19].

Non-typhoidal *Salmonella* (NTS) are widely associated with enterocolitis in humans, a zoonotic disease that is linked to the industrialization of food production. Because of the scale of human cases of enterocolitis and concerns related to food safety, more genome sequences have been generated for *Salmonella* than for any other genus. The number of publicly available sequenced *Salmonella* genomes reached 350,000 in 2021 [20] and are available from several public repositories such as the EMBL European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena), the Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra), and Enterobase (https://enterobase.warwick.ac.uk/species/index/senterica). However, limited genome-based surveillance of foodborne infections has been done in LMI countries, and the genomic dataset did not accurately represent the *Salmonella* pathogens that are currently causing disease across the world.

In recent years, new lineages of NTS serovars Typhimurium and Enteritidis have been recognized as common causes of invasive bloodstream infections (iNTS disease), responsible for about 77,000 deaths per year worldwide [21]. Approximately 80% of deaths due to iNTS disease occurs in sub-Saharan Africa, where iNTS disease has become endemic [22]. The new *Salmonella* lineages responsible for bloodstream infections of immunocompromised individuals are characterized by genomic degradation, altered prophage repertoires, and novel multidrug-resistant plasmids [23, 24].

We saw a need to simplify and expand genome-based surveillance of salmonellae from Africa and other parts of the world, involving isolates associated with invasive disease and gastroenteritis in humans, and extended to bacteria derived from animals and the environment. We optimized a pipeline for streamlining the large-scale collection and sequencing of samples from LMI countries with the aim of facilitating access to WGS and worldwide collaboration. Our pipeline represents a relatively inexpensive and robust tool for the generation of bacterial genomic data from LMI countries, allowing investigation of the epidemiology, drug resistance, and virulence factors of isolates.

Results

Development of an optimized logistics pipeline

The "10,000 *Salmonella* genomes project" (10KSG; https://10k-salmonella-genomes.com/) is a global consortium that includes collaborators from 25 institutions and a variety of settings, including research and reference laboratories across 16 countries. Limited funding resources prompted us to design an approach that ensured accurate sample tracking and captured comprehensive metadata for individual bacterial isolates while minimizing costs for the consortium. A key driver was to assemble a set of genomic data that would be as informative and robust as possible.

Members of the 10KSG consortium provided access to 10,419 bacterial isolates from collections that spanned 53 LMI countries and regions (such as Reunion Island, an overseas department and region of the French Republic). The samples covered seven bacterial genera: *Acinetobacter*, *Enterobacter*, *Klebsiella*, *Pseudomonas*, *Shigella*, and *Staphylococcus.* We optimized the logistics of specimen collection and the transport of materials to the sequencing center in the UK. The standardized protocols for metadata and sample submission were coordinated in three different languages (English, French, and Spanish), which facilitated collaboration across several countries (Fig. 1).

A crucial criterion for inclusion of *Salmonella* isolates in this study was the availability of detailed metadata and phenotypic information, to maximize the insights that could be generated from bacterial genomics. We created a standardized metadata table for input of relevant parameters. Table 1 summarizes some of the metadata collected. The metadata associated with unique sample identifiers included date of isolation, geographical location, source niche (human, animal or environmental isolate), and type

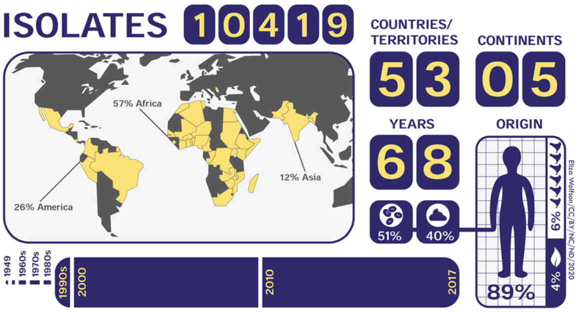

Fig. 1 Summary of the geographical origin, timeline, and body site source of 10,419 bacterial isolates. The 10,419 isolates were collected from 53 countries/territories spanning 5 continents (America, Africa, Asia, Europe, and Oceania), with most isolates originating from Africa (56%) and America (26%). The samples were mostly of human origin (86%), of which 52% were blood isolates, 41% were stool isolates, and 7% from other body compartments. About 5% samples originated from environmental sources, 6% were of animal origin, and 3% unknown. The bacterial pathogens were isolated over a 68-year time period, from 1949 to 2017. The majority of samples were isolated after 1990

Table 1 Sources of isolates collected by the 10KSG consortium

	Africa	America	Other[a]	Total
Animal	**294**	**253**	**94**	**641**
Farm animal	166	51	12	229
Other	128	202	82	412
Environmental	**6**	**357**	**80**	**443**
Food	0	56	12	68
Other	6	301	68	375
Human	**5612**	**2107**	**1567**	**9286**
Blood	3461	1238	12	4711
Stool	1591	772	1334	3697
Other	560	97	221	878
Total	**5912**	**2717**	**1741**	**10,370**[b]

[a] Asia, Europe, and Oceania

[b] The source location (continent) was unknown for 49 of the total 10,419 isolates

(body compartment). We also collected data regarding the antimicrobial susceptibility of isolates (Table 2) and captured additional information related to individual studies. We created a unified metadata master form (Additional file 2: Table S1) by manual concatenation and curation of individual metadata forms.

Development of thermolysates and sample collection

The main challenges for the global collection of bacterial samples are temperature-control and biological safety during transport. As refrigerated logistic chains are expensive, shipments should be at ambient temperature to minimize costs. To ensure biosafety, it was important to avoid the accidental transport of hazard group 3 (HG3) isolates (e.g., *S.* Typhi and *S.* Paratyphi A) [25]. Accordingly, we optimized a protocol for production of "thermolysates" that inactivated bacterial cells and permitted ambient temperature transport and adherence to containment level 2 (CL2) laboratory regulations, coupled with effective genomic DNA extraction for WGS (Additional file 2: Table S2 and Additional file 1: Fig. S3). Production of thermolysates also permitted simplification of material transfer agreements (MTAs) as there was no distribution of live cultures. Inactivation of *Salmonella* can be achieved at temperatures between 55 and 70 °C for as little as 15 s at high temperature (≥ 95 °C) [26]. We optimized the method for generation of "thermolysates" by inactivating bacterial cultures at high temperature (95 °C for 20 min). The optimization

Table 2 AMR phenotypes of 3463 isolates collected by the 10KSG consortium

Antimicrobial resistance[b]	Africa	America	Other[a]	Total
MDR[c]	1271	122	21	1414
1–2 agents[d]	581	69	56	706
Susceptible[e]	1139	142	62	1343
Total	**2991**	**333**	**139**	**3463**

[a] Asia, Europe, and Oceania

[b] Antimicrobial resistance profile performed by Kirby-Bauer technique. The antimicrobials used for profiling varied depending on the study, and it included ampicillin, chloramphenicol, streptomycin, tetracycline, gentamicin, kanamycin, nalidixic acid, trimethoprim, ciprofloxacin, ceftriaxone, and cotrimoxazole

[c] Resistance to ampicillin, cotrimoxazole, and chloramphenicol

[d] Antimicrobial resistance to 1–2 tested agents

[e] Susceptible to tested agents

involved testing under three different temperatures (90, 95, or 100 °C) and different incubation times (10 and 20 min). We also tested the effective inactivation of other non-*Salmonella* Gram-positive (*Staphylococcus aureus*) and Gram-negative (*Escherichia coli*) organisms (Additional file 2: Table S2).

Temperature is a key factor in the transportation of samples, especially in some LMI countries where dry ice is expensive and difficult to source, and access to international courier companies is limited or very costly. To allow transport without refrigeration, we tested the stability of the resulting thermolysates at room temperature for more than 7 days by assessing the quality of extracted DNA (Additional file 2: Table S2). Minimizing the steps required for sample collection allowed us to reach collaborators with limited access to facilities and personnel.

We collected samples using screwed-cap barcoded tubes (FluidX tri-coded jacket 0.7 mL, Brooks Life Sciences, 68-0702-11) costing USD$0.23 each, which we distributed from the UK to collaborators worldwide. Individually barcoded tubes were organized in FluidX plates in a 96-well format, each with their own barcode. Both QR codes and human-readable barcodes were included on each tube to ensure that the correct samples were always sequenced, and to permit the replacement of individual tubes when required.

The combination of method optimization, development, distribution of easy-to-follow protocols in English and Spanish (French was used only for communication), and collection of the bacterial isolates was completed within 1 year. Barcoded tubes were distributed to collaborators, including an extra ~ 20% to permit replacements as required. In total, 11,823 tubes were used in the study, of which 10,419 were returned to the sequencing center containing bacterial thermolysates for DNA extraction and genome sequencing.

To validate this approach for bacteria other than *Salmonella*, ~ 25% (2573, 24.7%) of the samples were isolates from a variety of genera, including Gram-negatives such as *Shigella* and *Klebsiella*, and Gram-positives such as *Staphylococcus.*

DNA extraction, library construction, quality control, and genome sequencing

Our high-throughput DNA extraction and library construction pipeline was designed to be versatile, scalable, and robust, capable of processing thousands of samples in a time and cost-efficient manner. The procedure included DNA extraction, quality control (QC), normalization, sequencing library construction, pooling, size selection, and sequencing. The time taken for each step, and the associated consumable cost, is shown in Table 3. All the parts of the pipeline are scalable and can be run simultaneously with robots, allowing hundreds of samples to be processed each day, in a 96-well format. With dedicated pre- and post-PCR robots, up to 768 bacterial samples were processed each day. The total consumable cost for extraction of DNA and genome sequence generation was less than USD$10 per sample (excluding staff time). Given the high-throughput nature of this project, and the difficulty in optimizing the processes to account for every possible variation in DNA/library quality and quantity, this cost includes a 20% contingency.

In designing the DNA extraction pipeline, we anticipated that samples would contain a wide range of DNA concentrations due to the different approaches by collaborators,

Table 3 Processing time and consumable costs for DNA extraction and sequencing

Activity	Processing time (h)[a]	Hands-on time (h)[a]	Consumable cost (USD$)[a,b]
DNA extraction	1	0.5	93.88
DNA QC and normalization	1	0.5	136.44
Library Construction, QC, pooling and size selection	6	1	277.86
Sequencing[c]	85	1	459.35
Total	**93 h**	**3 h**	**USD$ 967.53**

[a] Per 96-well plate
[b] Converted from GBP (1 GBP = 1.25 USD)
[c] Based on Illumina HiSeq4000 runs

some of whom sent thermolysates and others extracted DNA. The DNA was isolated in a volume of 20 μL, and the total yield ranged from 0 to 2170 ng (average of 272 ng). Less than 6% samples contained less than 2.5 ng (Additional file 1: Fig. S1).

To facilitate large-scale low-cost whole-genome sequencing, we developed the LITE (Low Input, Transposase Enabled; Fig. 2) pipeline, a low-cost high-throughput library construction protocol based on the Nextera kits (Illumina). Prior to LITE library construction, all DNA samples were normalized to 0.25 ng/μL unless the concentration was below that limit, in which case samples remained undiluted. We calculated that given a bacterial genome size of 4.5 Mbp, 1 ng of DNA equated to over 200,000 bacterial genome copies. Hence, the LITE pipeline was optimized to work with inputs ranging from 0.25 to 2 ng DNA. As the ratio of DNA to transposase enzyme determines the

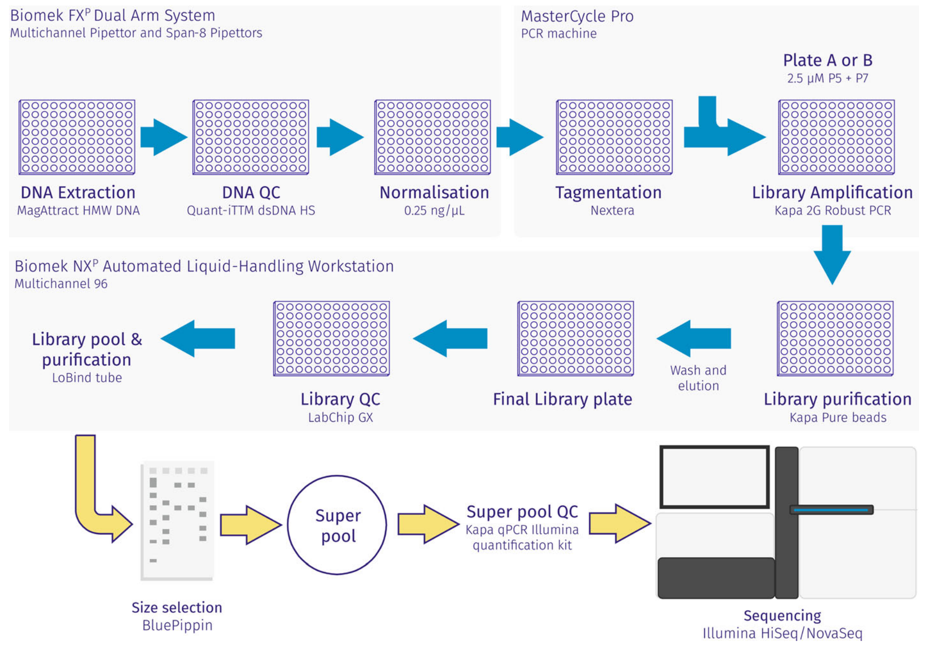

Fig. 2 LITE (Low Input, Transposase Enabled) pipeline for library construction. The DNA was extracted using a protocol based on the MagAttract HMW DNA isolation kit (Qiagen). Library construction was performed by tagmentation using Nextera tagmentation kit, size selected on a BluePippin, and quantified using a High Sensitivity BioAnalyzer kit (Agilent) and Qubit dsDNA HS Assay (ThermoFisher). Genome sequencing of "super pools" was performed in a HiSeq^{TM} 4000 (Illumina) system, and re-sequencing in NovaSeq^{TM} 6000 (Illumina) when needed, both with a 2 × 150 bp paired ends read metric

insert size of the libraries being constructed, this input amount allowed us to minimize reagent use and reaction volumes. The LITE pipeline permitted the construction of over 1000 Illumina-compatible libraries from the 24-reaction Illumina kits, Tagment DNA Enzyme (Illumina FC 15027865), and Illumina Tagment DNA Buffer (Illumina FC 15027866).

To maximize the multiplexing capability for the LITE pipeline, we designed 438 bespoke 9-bp barcodes (Additional file 2: Table S4), each with a hamming distance of 4 bp, giving the option to pool over 190,000 samples or uniquely dual-index more than 200 samples. The 438 barcodes allowed multiplexing capability to be maximized, and a further reduction in costs as sequencer throughputs increase in the future.

For this study, we used 9-bp barcoded P7 PCR primers (Illumina) and employed twelve 6-bp barcoded P5 PCR primers (Illumina) when multiplexing 12 × 96-well plates on a HiSeq 4000 system (Illumina) and targeted a median 30× genome coverage. Using an input of only 0.5 ng DNA and 14 PCR cycles consistently provided detectable amounts of library across the majority of samples.

Quality control (QC) of the resulting LITE libraries involved a Perkin Elmer LabChip® GX Nucleic Acid Analyzer. The LITE libraries typically gave three different GX electropherogram profiles depending upon whether the DNA was high molecular weight, partially degraded, or completely degraded (Additional file 1: Fig. S2). A wide range of electropherogram profiles and the resultant molarity of library molecules was expected at this point, due to the varied approaches used by collaborators to produce and transport samples.

Up to 12 of the 96 pooled and size-selected libraries were then combined and run on a single HiSeq 4000 system lane, with a 2 × 150 bp paired end read metric. After the initial screen was completed, samples that failed to produce 30× genome coverage were re-sequenced on a NovaSeq 6000 system, also with a 2 × 150 bp read metric. In total, 1525 (15.2%) of the 9976 samples required re-sequencing, a proportion that was within the 20% contingency included in our unit cost.

Bioinformatic analysis and data provision

To complete our WGS approach, we developed and implemented a bespoke sequence analysis bioinformatic pipeline for the *Salmonella* samples included in the study. The full pipeline is available from https://github.com/apredeus/10k_genomes including versions of all packages used. Because the estimation of sequence identity and assembly quality is relatively species-independent, and annotation is strongly species-specific, the pipeline can be easily adapted to other bacterial species by changing quality control criteria and specifying relevant databases of known proteins.

Following DNA extraction, sequencing, and re-sequencing, we generated sequence reads for 9976 (96.0%) samples, of which 7236 were bioinformatically classified as *Salmonella enterica* using Kraken2 and Bracken [27, 28]. A small proportion of the samples (209 out of 9976; 2.1%) had been y mis-identified as *Salmonella* prior to sequencing. The remaining samples corresponded to 1157 Gram-positive and Gram-negative bacterial isolates that were included to validate the study. The 443 (4.3%, out of the 10,419 bacterial isolates received) samples that did not generate sequence reads reflected poor-quality DNA extraction, due to either low biomass input or partial cell lysis. A comprehensive list of the quality control of individual isolates is in Additional file 2: Table S3. Overall, the generation of sequence data from the vast majority of samples demonstrated the robustness of the use of thermolysates coupled with the high-throughput LITE pipeline for processing thousands of samples from a variety of different collaborating organizations.

To assess the quality of sequence data, we focused on the 7236 (69.5%) genomes identified as *Salmonella enterica* (Fig. 3). To allow the bioinformatic analysis to be customizable for other datasets, we developed a robust quality control (QC) pipeline to do simple uniform processing of all samples, and to yield the maximum amount of reliable genomic information. Well-established software tools were used to assess species-level identity from raw reads, trim the reads, assess coverage and duplication rate, assemble genomes, and make preliminary evaluation of antibiotic resistance and virulence potential.

Trimming abundant adapters from the reads produced by the LITE pipeline was critical for optimal genome assembly. Using Quast [29] and simple assembly metrics, we evaluated the performance of Trimmomatic [30] in palindrome mode with and without retention of singleton reads, compared with BBDuk (https://jgi.doe.gov/data-and-tools/bbtools) in paired end mode. BBDuk was selected for our analysis because this tool generated genomes with a higher N50, and a comparable number of mis-assemblies.

Genome assembly was performed using SPAdes [31] via Unicycler v.0.4.7 [32] in short-read mode. SPAdes is an established and widely used tool for bacterial genome assembly, while Unicycler optimizes SPAdes parameters and performs assembly polishing by mapping reads back to the assembled genomes. Genome assembly QC was done using the criteria established by the genome database EnteroBase [33]. Specifically, these “stringent criteria” required (1) total assembly length between 4 and 5.8 Mb, (2)

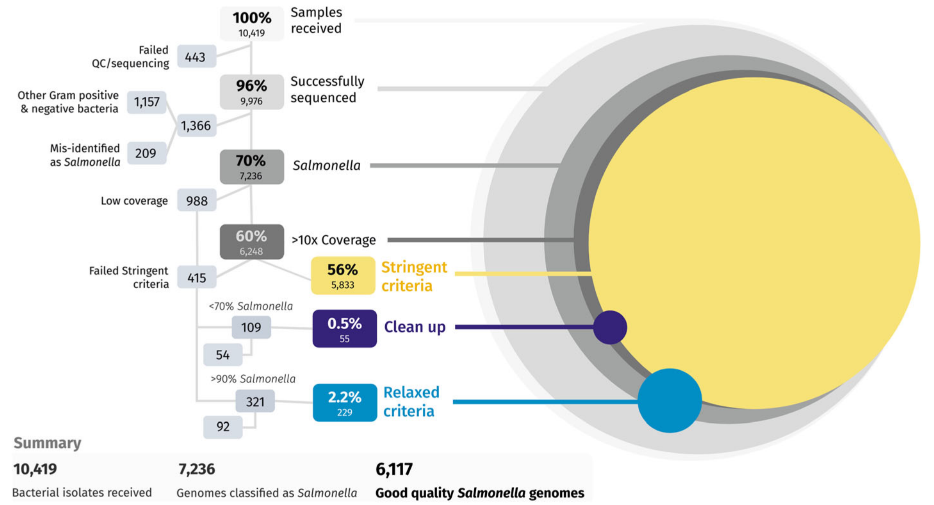

Fig. 3 The sequential quality control process used to select whole-genome sequences for detailed analysis. Of the 10,419 isolates, 443 failed the DNA extraction or quality control prior to genome sequencing. We produced sequencing libraries of 9975 samples, of which 1366 were not bioinformatically identified as *Salmonella enterica*. These 1366 corresponded to 1157 which were part of the 25% non-*Salmonella* component of the project, plus 209 isolates that had been mis-identified as *Salmonella* before sequencing. Of the 7236 *Salmonella* genomes, 6248 had sequence coverage over 10×, of which 5833 passed the “stringent criteria.” Of the 415 samples that failed the “stringent criteria,” 284 samples were rescued based on a “clean up” (55) or a “relaxed criteria” (229). Overall, we generated 6117 high-quality *Salmonella* genomes

N50 of 20 kb or more, (3) fewer than 600 contigs, and (4) more than 70% sequence reads assigned to the correct species. Using this approach for *S. enterica*, 5833 of the *Salmonella* genomes (80.6%) passed QC (Fig. 3).

We found that the sequencing depth before trimming was too low (≤ 10×) for 988 or 13.7% *S. enterica* samples, which were not analyzed further. To "rescue" all possible *S. enterica* in the remaining assemblies with coverage greater than 10× that failed the stringent QC, two approaches were used: "relaxed criteria" and "clean up".

To "rescue" all possible *S. enterica* in the remaining assemblies with coverage greater than 10× that failed the stringent QC, two approaches were used: "relaxed criteria" and "clean up". The "relaxed criteria" accepted assemblies of 4 Mb to 5.8 Mb overall length, species purity of 90% or more, N50 > 10 kb, and fewer than 2000 contigs. In contrast, the "clean up" approach was used for assemblies that had < 70% *Salmonella* sequence reads using the "stringent criteria." The raw reads of these samples were "cleaned" using Kraken2 & Bracken, with the reads assigned to *Salmonella* being retained, and subjected to the "stringent criteria" for QC detailed above. The assemblies rescued by these two approaches accounted for a further 3.9% (284) assemblies from our initial *Salmonella* collection. In total, we generated 6117 high-quality *S. enterica* genomes, corresponding to 84.5% of the total *Salmonella* isolates successfully sequenced through the LITE pipeline (Figs. 3 and 4).

Genome sequence data were shared with collaborators via downloadable packages hosted by the Centre of Genomic Research, University of Liverpool (UK). These packages included sequencing statistics, raw (untrimmed) fastq files of sequence reads, and the individual genome assemblies. We included the genome-derived *Salmonella* serovar and sequence type of each isolate (Fig. 4).

Together with predicted sequence type and serovar, the genome-derived information was provided to permit local surveillance laboratories and infectious disease clinicians to derive important insights about the *Salmonella* variants circulating in their countries. The value of bacterial WGS data for generating epidemiological insights or understanding pathogen evolution has been summarized recently [20]. All the processed sequence reads and assemblies were deposited in the European Nucleotide Archive under the project accession number PRJEB35182 (ERP118197). Individual accession numbers are listed in Additional file 2: Table S5.

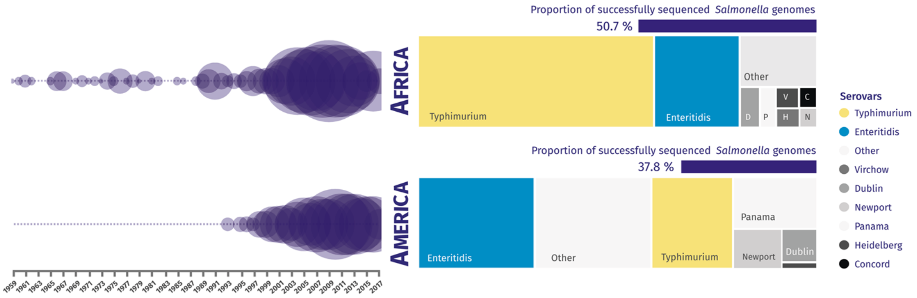

Fig. 4 Genome-based summary of *Salmonella enterica* from African and American datasets, organized by continent, year of isolation, and serovar. Of the 6117 *Salmonella enterica* genomes that were successfully sequenced and that passed QC, 3100 (50.7%) were from Africa and 2313 (37.8%) were from America. Bubble size represents the number of genomes isolated between 1959 and 2017. The graphs represent the proportion of the main *Salmonella* serovars predicted based on genome analysis: 1844 *S.* typhimurium and 657 *S.* enteritidis from Africa, and 474 *S.* typhimurium and 676 *S.* enteritidis from America

Discussion

We have optimized an efficient and relatively inexpensive method for large-scale collection and sequencing of bacterial genomes, by streamlining the collection of isolates and developing a logistics pipeline that permitted ambient shipment of thermolysates. The global focus of our study provided a diverse collection of 10,419 clinical and environmental bacterial isolates for a single sequencing study within 1 year.

Advantages and limitations of the study

The effectiveness and accessibility of our approach allowed all samples to be collected in a timely manner, and generated genomic data for LMI countries that lacked easy access to sequencing technology. A key aspect of our methodology was the involvement of researchers fluent in multiple languages in corresponding with collaborators, to maximize clear and continuous communication by email. Our inclusive approach was intended to provide access to researchers from countries across the world who might otherwise have been excluded, permitting the development of an international consortium of 26 institutions in 16 countries. This global approach allowed genome sequencing data from bacterial isolates from five continents (53 countries/territories, Fig. 1) to be incorporated into a single study.

The optimized DNA extraction and sequencing LITE pipeline generated an individual bacterial genome at a consumables cost of USD$10 per sample (the full economic cost cannot be calculated because collaborator staff time was an in-kind contribution). An advantage of the LITE pipeline is the low DNA requirement. For *Salmonella*, 1 ng equates to > 150,000 copies of the genome, making it a robust pipeline for the capture of all the genetic material.

A key innovative aspect of the project involved the use of thermolysates, both to reduce shipment costs and to increase throughput by facilitating automated DNA extraction in a central location. Thermolysates allowed collaborators to provide samples for genome sequencing without needing to extract genomic DNA.

The combination of the use of thermolysates with the optimized DNA extraction and sequencing LITE pipeline provides a robust approach for global collaboration on the genome-based mass surveillance of pathogens. Our method is suitable for other large collections of Gram-negative or Gram-positive bacteria.

However, our approach did pose manual and logistical challenges. The LITE pipeline represents a compromise in terms of data quality to maximize economic value. The biggest factor in sample failure was DNA degradation, which reduced the starting length of extracted DNA molecules and affected the final library size. Because sample failures increased the cost of subsequent bioinformatic QC steps, we identified examples of unsuccessful DNA extraction by assessment of DNA integrity (Additional file 1: Fig. S2), which resulted in less than 5% (443) samples failing to be sequenced (Fig. 3). Therefore, it is important that all QC steps, and the rigorous bioinformatic approach that we specify, are followed to produce a reliable dataset. Overall, we case generated 84.5% (6117) high-quality genomes of the 7236 *Salmonella* isolates that were successfully sequenced (Figs. 3 and 4).

The initial logistical optimization required a significant investment of time and manual curation to ensure that the project goals were met. To successfully coordinate this global collaboration, continuous electronic communication was needed to maintain mutual trust, to understand the requirements and concerns of all collaborators, and to maintain a focus on the various individual research questions. We suggest that future implementations of a similar approach for sequencing thousands of bacterial isolates begins with an early investment in the development of a shared, protected, and version-controlled database for the storage of epidemiological information. In addition, a streamlined system for the sending and receiving of samples and automated scripts to handle sequencing data are essential.

Conclusions

We have established an efficient and relatively inexpensive pipeline for the worldwide collection and sequencing of bacterial genomes. Our novel approach allows the transport and whole-genome sequencing of large collections of bacterial pathogens, by coupling the use of thermolysates with DNA extraction and sequencing using the innovative LITE pipeline for library construction.

We evaluated this method with the model organism *Salmonella enterica* through worldwide research collaboration, generating 6117 high-quality *Salmonella* genomes, which have already been used for a number of published studies [34–37]. In future, the method will facilitate rapid, low-cost, and collaborative genome sequencing of bacterial pathogens. Our concerted approach demonstrates the value of true global collaboration, and could contribute to the future investigation of international epidemics or pandemics.

Methods

Study design and optimization

We designed the project with the aim of validating an efficient method for large-scale assembly and sequencing of bacterial genomes. We selected *Salmonella* as a model organism due to its worldwide relevance and current burden of infection. We aimed to assemble a pool of bacterial samples that would represent the different scenarios, including a 25% of non-*Salmonella* isolates, to allow the method to be extrapolated to other bacterial datasets. The 25% of non-*Salmonella* organisms were selected to cover Gram-negative (*Shigella* and *Klebsiella*) and Gram-positive (*Staphylococcus*) bacteria. The targeted *Salmonella* isolates were predominantly *S.* Enteritidis and Typhimurium, and associated with human bloodstream infection. However, we expanded the sampling criteria to other serovars, body compartments, and source types to include some animal and environmental samples.

Method optimization focused on standardizing a safe protocol for sample transport and processing. The optimized method comprised bacterial isolates grown at 37 °C overnight directly in FluidX tubes (FluidX tri-coded jacket 0.7 mL, 68-0702-11, Brooks Life Sciences) that contained 100 μL rich media and were inoculated from a frozen stock prepared from one bacterial colony (one "scoop" or bead (Microbank^{TM}, Pro Lab Diagnostics Inc.)). The rich media used in the reference laboratories involved in this large collaborative study were either Lennox Broth or Buffered Peptone Water, depending on the protocols used in different laboratory settings. After overnight growth, the bacterial samples were inactivated by incubation at > 95 °C for 20 min, followed by storage at 4 °C until collection. Sample transportation was carried out at ambient temperature.

To ensure that the heat-inactivation step had the expected effect on organism viability, a controlled experiment was performed. Three bacterial strains were tested, namely *Salmonella enterica* serovar Typhimurium D23580, *Escherichia coli* K12, and *Staphylococcus aureus* Newman. Either a "scoop" with a 10-μL plastic loop taken from a bacterial glycerol (50% v/v) stock or 2 beads of bacteria stored at − 80 °C in Microbank tube™ cryotubes (Pro Lab Diagnostics) were used as inocula. The samples were grown at 37 °C and 220 rpm overnight in either 100 or 200 μL LB (1% tryptone, 0.5% yeast extract, 0.5% NaCl; pH 7.0) in FluidX tubes. The effect of three temperatures and two treatment times upon microbial viability was determined as follows: 100 μL of each sample was heated to either 90, 95, or 100 °C for 10 or 20 min, and then plated on nutrient agar (1.5% agar LB (w/v)) for CFU determination (Additional file 2: Table S2).

To test the effect of transport, the samples were subjected to genomic DNA extraction using a DNeasy Blood & Tissue Kit (Qiagen) after incubation at room temperature for more than 7 days. The quality of extracted DNA was assessed by 1% agarose gel electrophoresis, and fluorometric DNA quantification using Qubit™ dsDNA HS Assay Kit (Invitrogen™) (Additional file 2: Table S2 and Additional file 1: Fig. S3).

Detailed protocols were sent to collaborators, along with a metadata template and barcoded tubes. The design of the metadata template and protocol booklet was tested several times to maximize clarity and to obtain unified information that was interpreted in the same way by different users. The metadata template (Additional file 2: Table S1) was a Microsoft Excel spreadsheet divided in five main categories: (1) unique identifiers, with information about pre-read barcodes, including plate and tube barcode, tube location, and replacement barcode, (2) isolate details, encompassing information about strain name, bacterial species, and serovar (*Salmonella* only), sender, date and location of isolation, and type of sample submitted (DNA, thermolysates, or preserved culture), (3) sample type, with detailed information about source of isolation, such as human, animal, or environmental origin, and (4) antimicrobial resistance phenotype of tested antimicrobials (profile obtained by Kirby-Bauer technique). The metadata template also included an extra column for relevant information that could not be assigned to any other category, such as type of study and relevant citations.

The resulting metadata were stored per collaborator and then combined into a metadata master form for curation. Curation was done manually, standardizing each category by column and maintaining version control. The final metadata master form was cross-referenced with the list of sent barcodes to identify inconsistencies.

DNA extraction and normalization

DNA was extracted from bacterial thermolysates on a Biomek FX^P instrument using a reduced volume protocol of the MagAttract HMW DNA isolation kit (Qiagen). Incomplete barcoded 96-tube plates received were re-organized and FluidX barcodes re-read using the FluidX barcode reader and software prior to DNA extraction, to determine plate layouts. The tubes were de-capped using a manual eight-tube decapper and the cellular material was re-suspended using a multichannel pipette. Up to 100 μL of the suspension was transferred to a clean 96-well plate. The plate was spun at 4000 rpm in an Eppendorf 5810R centrifuge and visible pellets were observed in a majority of cases indicating the presence of cellular material. Plates were then upturned and the supernatant carefully discarded.

Cell pellets were re-suspended in a mixture of 12 μL of Qiagen ATL buffer and 2 μL Proteinase K, and incubated at 56 °C for 30 min in an Eppendorf Thermomixer C. The samples were cooled to room temperature, and 1 μL of MagAttract Suspension G was added. The samples were mixed, and 18.67 μL of Qiagen MB buffer was added, followed by mixing. The samples were incubated for 3 min and placed on a 96-well magnetic particle concentrator (MPC) to pellet the beads. The supernatant was discarded, while remaining on the MPC the beads were washed once with 45 μL Qiagen MW1 buffer and once with 45 μL Qiagen PE buffer. The recommended water washes were omitted to help increase yield.

The plate was then removed from the MPC and, using a new set of filter tips, 20 μL of Qiagen AE buffer was added and the samples mixed to re-suspend the beads. The samples were incubated at room temperature for 3 min to elute the DNA. The plate was placed back on the MPC and the DNA was transferred to a new 96-well plate.

The concentration of each sample was determined using the Quant-iT™ dsDNA Assay, high sensitivity kit (Thermo Fisher). A standard curve was generated by mixing 10 μL of the eight DNA standards provided (0 to 10 ng/μL) with 189 μL of 1× Quant-iTTM dsDNA HS buffer, 1 μL of Quant-iTTM dsDNA HS reagent, and 1 μL of DNA in a 96-well black Greiner plate. The fluorescence was detected on a Tecan Infinite F200 Pro plate reader (Tecan).

For samples received as DNA, 198 μL of 1× Quant-iTTM dsDNA HS buffer, 1 μL of Quant-iTTM dsDNA HS reagent, and 1 μL of DNA were combined in a 96-well black Greiner plate, and the fluorescence detected using the Tecan plate reader. Concentrations were calculated using the standard curve, and the DNA was normalized to 0.25 ng/μL in elution buffer using the Biomek FX^{P} instrument.

Library construction and sequencing

A master mix containing 0.9 μL of Nextera buffer, 0.1 μL Nextera enzyme, and 2 μL of DNAse-free water was combined with 2 μL of normalized DNA. This reaction was incubated at 56 °C for 10 min on an Eppendorf MasterCycle Pro PCR instrument. Then, 2 μL of an appropriately barcoded 2.5 μM P7 adapter was added, and then 18 μL of a master mix containing 2 μL of an appropriately barcoded 2.5 μM P5, 5 μL Kapa Robust 2G 5× reaction buffer, 0.5 μL 10 mM dNTPs, 0.1 μL Kapa Robust 2G polymerase, and 10.4 μL DNase-free water were added to the tube. This reaction was then subjected to PCR amplification as follows: 72 °C × 3 min, 98 °C for 2 min, then 14 cycles of 98 °C × 10 s, 62 °C × 30 s and 72 °C × 3 min, followed by a final incubation at 72 °C for 5 min on an Eppendorf MasterCycle Pro.

The amplified library was then subjected to a magnetic bead-based purification step on a Biomek NX^{P} instrument. Then, 25 μL of Kapa Pure beads (Roche, UK) was added to 25 μL of amplified library, and mixed. This library was incubated at room temperature for 5 min, briefly spun in an Eppendorf 5810R centrifuge and placed on a 96-well magnetic particle concentrator. Once the beads had pelleted, the supernatant was removed and discarded, and the beads washed twice with 40 μL of freshly prepared

70% ethanol. After the second ethanol wash, the beads were left to air dry for 5 min. The 96-well plate was removed from the MPC, and the beads were re-suspended in 25 μL of 10 mM TRIS-HCl, pH 8 (Elution Buffer). The DNA was eluted by incubating the beads for 5 min at room temperature. The plate was replaced on the MPC, the beads allowed to pellet, and the supernatant containing the DNA was transferred to a new 96-well plate.

To assess the concentrations of individual libraries, 20 μL of elution buffer was added to 2 μL of purified library, and run on a LabChip GX (Perking Elmer) using the High-throughput, High Sense reagent kit, and HT DNA Extended Range Chip according to the manufacturers' instructions. To determine the amount of material present in each library between 400 and 600 bp, a smear analysis was performed using the GX analysis software. The resulting value was used to calculate the amount of each library to pool. Pooling of each 96-libraries was performed using a Biomek Nx instrument. Then, 100 μL of the pooled libraries was added to 100 μL of Kapa Pure beads in a 1.5-mL LoBind tube. The sample was vortexed and incubated at room temperature for 5 min to precipitate the DNA onto the beads. The tube was then placed on an MPC to pellet the beads, the supernatant discarded, and the beads were washed twice with 200 μL of freshly prepared 70% ethanol. The beads were left to air dry for 5 min and then re-suspended in 30 μL Elution Buffer. The samples were incubated at room temperature for 5 min to elute the DNA. The plate was placed back on the MPC and the DNA was transferred to a new 1.5-mL tube.

The concentrated sample containing a pool of 96 libraries was subjected to size selection on a BluePippin (Sage Science, Beverly, USA). The 40 μL in each collection well of a 1.5% BluePippin cassette was replaced with fresh running buffer, and the separation and elution current checked prior to loading the sample. Then, 10 μL of R2 marker solution were added to 30 μL of the pooled library, and then the combined mixture was loaded into the appropriate well.

Using the smear analysis feature of Perkin Elmer GX software, we calculated the amount of material between 400 and 600 bp for each library. We targeted this region based on the electropherograms in Additional file 1: Fig. S2, to minimize the overlap between 150 bp paired end reads and maximize the number of libraries that would generate data. We determined the detection limit for the molarity within this size range to be 0.007 nM, meaning that libraries with lower concentrations were reported as 0.007 nM. The amount of library material between 400 and 600 bp ranged from 0.0 to 2.4 nM (average of 0.3 nM), with less than 6% having less than 0.007 nM (Additional file 1: Fig. S1).

Post size selection, the 40 μL from the collection well were recovered, and the library size was determined using a High Sensitivity BioAnalyzer kit (Agilent) and DNA concentration calculated using a Qubit dsDNA HS Assay (Thermo Fisher). "Super pools" were created by equimolar pooling of up to 12 size-selected 96-sample pools, each with a different P5 barcode. Using these molarity figures, 96 libraries were equimolarly pooled, concentrated, and then size-selected using a 1.5% cassette on the Sage Science Blue Pippin.

To determine the number of viable library molecules, the super pools were quantified using the Kapa qPCR Illumina quantification kit (Kapa Biosystems) prior to sequencing. For the initial screen, sequencing was performed on the HiSeq^{TM} 4000 (Illumina). For re-sequencing of samples, the sequencing was carried out in a lane of an S1 flowcell on the NovaSeq^{TM} 6000 (Illumina), both with a 2 × 150 bp read metric.

Bioinformatic analysis *and data distribution*

Raw sequencing reads (paired end, 2 × 150 bp) were examined using FastQC v0.11.8 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc), confirming 0–20% Nextera adapter sequence presence in all examined reads. Quick coverage estimation was done raw unaligned reads, assuming genome length of 4.8 Mb for *Salmonella enterica*. Taxonomic classification of raw reads was performed using Kraken v2.0.8-beta [27] with Minikraken 8GB 201904_UPDATE database, followed by species-level abundance estimation using Bracken v1.0.0 [28] with distribution for 150 bp k-mer. Sequence duplication level was estimated by alignment of reads using Bowtie v2.3.5 [38] to genome assembly of LT2 strain (NCBI accession number GCA_000006945.2), followed by MarkDuplicates utility from Picard tools v2.21.1(http://broadinstitute.github.io/picard).

Raw sequence reads were then trimmed and assembled using Unicycler v0.4. 7[32] in short-read mode. Several trimming strategies were tested including quality trimming with seqtk (https://github.com/lh3/seqtk) followed by Trimmomatic v0.3 9[39] in palindromic mode with and without retaining the single reads, and BBDuk v38.07 (https://jgi.doe.gov/data-and-tools/bbtools). We evaluated the resulting assemblies using overall length, N50, and number of contigs. Genome assembly quality was assessed using the criteria established on EnteroBase [33] (https://enterobase.readthedocs.io/en/latest) for *S. enterica*: (1) total assembly length between 4 and 5.8 Mb; (2) N50 of 20 kb or more; (3) fewer than 600 contigs; (4) more than 70% correct species assigned by Kraken (which we replaced with Kraken2 and Bracken assessment of the raw reads). Samples that failed the stringent criteria were divided into two groups. Group 1 were subjected to "relaxed criteria," which included assemblies of 4–5.8 Mb overall length, species purity of 90% or more, N50 > 10,000, and fewer than 2000 contigs. Group 2 included samples that had less than 70% *Salmonella* by original assessment, but produced assemblies passing the stringent criteria from "cleaned up" reads obtained by keeping only raw reads assigned *S. enterica* by Kraken2 + Bracken.

Assembled *Salmonella* genomes were annotated using Prokka v1.13.7 [40] using a custom protein database generated from *S. enterica* pan-genome analysis. Additionally, *Salmonella* assemblies were in silico serotyped using command line SISTR v1.0.2 [41] and assigned sequence type using mlst v2.11 [42] (https://github.com/tseemann/mlst). We have used cgMLST serovar assignment provided by SISTR for all further classification and comparison with metadata. Preliminary resistance and virulence gene profiling was done using Abricate v0.9.8 (https://github.com/tseemann/abricate). Our analysis of a sample of 680 *S.* typhimurium isolates showed that genome-based analysis accurately predicted the AMR phenotypes of 89.8% isolates, with 87–98% sensitivity (86.7% chloramphenicol, 98.0% ampicillin, 97.4% cotrimoxazole) and 77–96% specificity (83.6% chloramphenicol, 95.6% ampicillin, 77.1% cotrimoxazole) [34]. All processing scripts detailing command settings and custom datasets are available at https://github.com/apredeus/10k_genomes [43].

Data distribution was carried out by sharing packages through links created at the Centre for Genomic Research, University of Liverpool (UK). The packages contained sequencing stats, raw (untrimmed) fastq read files, assemblies, and a text files with information about serovar and sequence type details. All the processed reads and assemblies were deposited in the European Nucleotide Archive using the online portal Collaborative Open Plant Omics (COPO; https://copo-project.org/copo) under the project accession numbers PRJEB35182 [44] and PRJEB47910 [45]. COPO is an online portal for the description, storage, and submission of publication data. The COPO wizards allow users to describe their data, using ontologies to link and suggest metadata to include, based on past submissions and similar projects. This approach generates meaningful data descriptions and standardizes the format to facilitate the easy retrieval of information. The COPO strategy simplified the process of data submission. Individual accession numbers are listed in Additional file 2: Table S5.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02536-3.

Additional file 1: Includes supplementary figures S1, S2, and S3: **Fig. S1.** Average DNA concentration and molarity of libraries constructed using the LITE pipeline across individual 96-well plates. **Fig. S2:** Assessment of DNA integrity among libraries constructed using the LITE pipeline. **Fig. S3:** Integrity assessment of DNA extraction of 42 samples described in Table S2.

Additional file 2: Includes supplementary tables S1, S2, S3, S4, and S5. **Table S1.** Metadata template form. **Table S2.** Optimization of bacterial thermolysates generation and DNA extraction. **Table S3.** Metadata for sequenced isolates, including bioinformatic stats for Salmonella genomes. **Table S4.** Bespoke 9 bp barcodes for library construction using the LITE pipeline. **Table S5.** European Nucleotide Archive accession numbers.

Additional file 3. Review history.

Acknowledgements

Next-generation sequencing and library construction were delivered via the BBSRC National Capability in Genomics and Single Cell (BB/CCG1720/1) at Earlham Institute, by members of the Genomics Pipelines Group. Each of the following Earlham Institute staff and alumni made an enormous contribution to the project: Dr Helen Chapman, Mr. Jake Collins, and Miss Sophie Stephenson. We are grateful to the Centre for Genomic Research, University of Liverpool (UK) for computing support. The 10KSG consortium: Blanca M. Perez-Sepulveda, Darren Heavens, Caisey V. Pulford, María Teresa Acuña, Dragan Antic, Martin Antonio, Kate S. Baker, Johan Bernal, Hilda Bolaños, Marie Chattaway, John Cheesbrough, Angeziwa Chirambo, Karl Costigan, Saffiatou Darboe, Paula Díaz, Pilar Donado, Carolina Duarte, Francisco Duarte, Dean Everett, Séamus Fanning, Nicholas A. Feasey, Patrick Feglo, Adriano M. Ferreira, Rachel Floyd, Ronnie G. Gavilán, Melita A. Gordon, Neil Hall, Rodrigo T. Hernandes, Gabriela Hernández-Mora, Jay C. D. Hinton, Daniel Hurley, Irene N. Kasumba, Benjamin Kumwenda, Brenda Kwambana-Adams, James Lipscombe, Ross Low, Salim Mattar, Lucy Angeline Montaño, Cristiano Gallina Moreira, Jaime Moreno, Dechamma Mundanda Muthappa, Satheesh Nair, Chris M. Parry, Chikondi Peno, Jasnehta Permala-Booth, Jelena Petrović, Alexander V. Predeus, José Luis Puente, Getenet Rebrie, Martha Redway, Will Rowe, Terue Sadatsune, Christian Schudoma, Neil Shearer, Claudia Silva, Anthony M. Smith, Sharon Tennant, Alicia Tran-Dien, Chris Watkins, Hermione Webster, François-Xavier Weill, Magdalena Wiesner and Catherine Wilson.

Review history

The review history is available as Additional file 3.

Peer review information

Kevin Pang was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

NH and JCDH conceived the idea and received funding. BPS wrote the manuscript. JCDH, NH, CVP, AVP, DH, BK, KSB, WR, and NAF contributed to manuscript writing and editing. The 10KSG consortium reviewed the manuscript. JCDH, NH, NAF, KSB, CVP, and BPS designed the study. BPS and KC curated the metadata. BPS was the main point of contact for the 10KSG consortium, designed and prepared protocols and other material for collaborators, and distributed barcoded tubes. WR and BPS designed web page. RL uploaded generated data to ENA. CW and NS supervised logistics at the Earlham Institute. BPS, CVP, and HW optimized thermolysates generation. AVP, CVP, RL, and CS developed bioinformatic pipelines and analysis. DH and JL optimized LITE protocol. BPS, CVP, and the 10KSG consortium isolated and prepared bacterial samples. All authors read and approved the final manuscript.

Authors' information

Twitter handles: @BlancaPS (Blanca M. Perez-Sepulveda), @1caiseypulford (Caisey V. Pulford), @APredeus (Alexander V. Predeus), @joy_hermione (Hermione Webster), @wil_rowe (Will Rowe), @benjkum (Benjamin Kumwenda), @ksbakes (Kate S. Baker), @nfeasey1 (Nicholas A. Feasey), @jay_salsa (Jay C. D. Hinton), @neilhall_uk (Neil Hall).

Funding

The project was supported by both a Global Challenges Research Fund (GCRF) data & resources grant BBS/OS/GC/000009D and the BBSRC Core Capability Grant to the Earlham Institute BB/CCG1720/1 and Core Strategic Programme Grant BBS/E/T/000PR9817. This project was partly supported by the Wellcome Trust Senior Investigator Award (106914/Z/15/Z) to Jay C. D. Hinton.

Availability of data and materials

The dataset supporting the conclusions of this article is available in the EMBL European Nucleotide Archive (ENA) repository under the project accession numbers PRJEB35182 [44] and PRJEB47910 [45]. Individual accession numbers are listed in Additional file 2, Table S5. Our code is available as open source (GPL v3 license) at https://github.com/apredeus/10k_genomes [43].

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare no competing interests.

Author details

[1]Institute of Infection, Veterinary & Ecological Sciences, University of Liverpool, Liverpool, UK. [2]Earlham Institute, Norwich Research Park, Norwich, UK. [3]University of Birmingham, Birmingham, UK. [4]Kamuzu University of Health Sciences, Blantyre, Malawi. [5]Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, UK. [6]Malawi-Liverpool-Wellcome Programme, Blantyre, Malawi. [7]School of Biological Sciences, University of East Anglia, Norwich, UK. [8]Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia.

Received: 10 August 2020 Accepted: 21 October 2021
Published online: 21 December 2021

References

1. Loman NJ, Constantinidou C, Chan JZM, Halachev M, Sergeant M, Penn CW, et al. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol. Nature Publishing Group. 2012; 10(9):599–606. https://doi.org/10.1038/nrmicro2850.
2. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. Nature Publishing Group. 2016;17(6):333–51. https://doi.org/10.1038/nrg.2016.49.
3. Power RA, Parkhill J, De Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. Nat. Rev. Genet. Nature Publishing Group; 2016. p. 41–50.
4. Bentley SD, Parkhill J. Genomic perspectives on the evolution and spread of bacterial pathogens. Proc R Soc B Biol Sci. 2015;282(1821):20150488. https://doi.org/10.1098/rspb.2015.0488.
5. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 2012;30(5):434–9. https://doi.org/10.1038/nbt.2198.
6. Wetterstrand KA. DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP) [Internet]. [cited 2019 Nov 27]. Available from: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data
7. Quainoo S, Coolen JPM, van Hijum SAFT, Huynen MA, Melchers WJG, van Schaik W, et al. Whole-genome sequencing of bacterial pathogens: The future of nosocomial outbreak analysis. Clin. Microbiol. Rev. American Society for Microbiology. 2017;30:1015–63.
8. Chun J, Grim CJ, Hasan NA, Je HL, Seon YC, Haley BJ, et al. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic Vibrio cholerae. Proc Natl Acad Sci U S A. 2009;106(36):15442–7. https://doi.org/10.1073/pnas.0907787106.
9. Weill F-X, Domman D, Njamkepo E, Tarr C, Rauzier J, Fawal N, et al. Genomic history of the seventh pandemic of cholera in Africa. Science (80-). American Association for the Advancement of Science. 2017;358:785–9.
10. Schürch AC, Kremer K, Daviena O, Kiers A, Boeree MJ, Siezen RJ, et al. High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. J Clin Microbiol. 2010;48(9):3403–6. https://doi.org/10.1128/JCM.00370-10.
11. Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo EC, Bradley P, et al. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. Lancet Infect Dis. 2015; 15(10):1193–202. https://doi.org/10.1016/S1473-3099(15)00062-6.
12. Allix-Béguec C, Arandjelovic I, Bi L, Beckert P, Bonnet M, Bradley P, et al. Prediction of susceptibility to first-line tuberculosis drugs by DNA sequencing. N Engl J Med. 2018;379(15):1403–15. https://doi.org/10.1056/NEJMoa1800474.
13. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, et al. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics. 2015;15(2):141–61. https://doi.org/10.1007/s10142-015-0433-4.
14. Kwong JC, Mccallum N, Sintchenko V, Howden BP. Whole genome sequencing in clinical and public health microbiology. Pathology. 2015;47(3):199–210. https://doi.org/10.1097/PAT.0000000000000235.
15. Gladstone RA, Lo SW, Lees JA, Croucher NJ, van Tonder AJ, Corander J, et al. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. EBioMedicine. The Authors. 2019;43:338–46. https://doi.org/10.1016/j.ebiom.2019.04.021.
16. Bardsley M, Jenkins C, Mitchell HD, Mikhail AFW, Baker KS, Foster K, et al. Persistent Transmission of Shigellosis in England Is Associated with a Recently Emerged Multidrug-Resistant Strain of Shigella sonnei. J Clin Microbiol. American Society for Microbiology. 2020;58(4). https://doi.org/10.1128/JCM.01692-19.

17. Achtman M, Zhou Z, Alikhan N-F, Tyne W, Parkhill J, Cormican M, et al. Genomic diversity of *Salmonella enterica* -The UoWUCC 10 K genomes project [version 1; peer review: 2 approved]. Wellcome Open Res. 2020;5:223. https://doi.org/10.12688/wellcomeopenres.16291.1.
18. Black A, DR MC, Sibley TR, Bedford T. Ten recommendations for supporting open pathogen genomic analysis in public health. Nat Med. Springer US. 2020;26(6):832–41.
19. Kirk MD, Pires SM, Black RE, Caipo M, Crump JA, Devleesschauwer B, et al. World Health Organization Estimates of the Global and Regional Disease Burden of 22 Foodborne Bacterial, Protozoal, and Viral Diseases, 2010: A Data Synthesis. PLoS Med. 2015;12:1–21.
20. Moustafa AM, Lal A, Planet PJ. Comparative genomics in infectious disease. Curr Opin Microbiol. Elsevier Ltd. 2020;53:61–70. https://doi.org/10.1016/j.mib.2020.02.009.
21. Stanaway JD, Parisi A, Sarkar K, Blacker BF, Reiner RC, Hay SI, et al. The global burden of non-typhoidal *Salmonella* invasive disease: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Infect Dis. Lancet Publishing Group. 2019;19(12):1312–24. https://doi.org/10.1016/S1473-3099(19)30418-9.
22. Stanaway JD, Reiner RC, Blacker BF, Goldberg EM, Khalil IA, Troeger CE, et al. The global burden of typhoid and paratyphoid fevers: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Infect Dis. 2019;19(4):369–81. https://doi.org/10.1016/S1473-3099(18)30685-6.
23. Okoro CK, Kingsley RA, Quail MA, Kankwatira AM, Feasey NA, Parkhill J, et al. High-resolution single nucleotide polymorphism analysis distinguishes recrudescence and reinfection in recurrent invasive nontyphoidal salmonella typhimurium disease. Clin Infect Dis. 2012;54(7):955–63. https://doi.org/10.1093/cid/cir1032.
24. Feasey NA, Hadfield J, Keddy KH, Dallman TJ, Jacobs J, Deng X, et al. Distinct *Salmonella* Enteritidis lineages associated with enterocolitis in high-income settings and invasive disease in low-income settings. Nat Genet. 2016;48(10):1211–7. https://doi.org/10.1038/ng.3644.
25. Andrews JR, Ryan ET. Diagnostics for invasive *Salmonella* infections: Current challenges and future directions. Vaccine. Elsevier Ltd. 2015;33:C8–15. https://doi.org/10.1016/j.vaccine.2015.02.030.
26. Silva FVM, Gibbs PA. Thermal pasteurization requirements for the inactivation of Salmonella in foods. Food Res Int. 2012; 45(2):695–9. https://doi.org/10.1016/j.foodres.2011.06.018.
27. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. Genome Biology. 2019;20:1–13.
28. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: Estimating species abundance in metagenomics data. PeerJ Comput Sci. 2017;2017:1–17. https://doi.org/10.7717/peerj-cs.104.
29. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. England. 2013;29(8):1072–5. https://doi.org/10.1093/bioinformatics/btt086.
30. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. England. 2014;30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170.
31. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77. https://doi.org/10.1089/cmb.2012.0021.
32. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. United States. 2017;13:e1005595.
33. Alikhan N-F, Zhou Z, Sergeant MJ, Achtman M. A genomic overview of the population structure of Salmonella. PLoS Genet. Public Library of Science. 2018;14(4):e1007261. https://doi.org/10.1371/journal.pgen.1007261.
34. Pulford CV, Perez-Sepulveda BM, Canals R, Bevington JA, Bengtsson RJ, Wenner N, et al. Stepwise evolution of Salmonella Typhimurium ST313 causing bloodstream infection in Africa. Nat Microbiol. Nature Research. 2021;6(3):327–38. https://doi.org/10.1038/s41564-020-00836-1.
35. Kasumba IN, Pulford C V., Perez-Sepulveda BM, Sen S, Sayed N, Permala-Booth J, et al. Characteristics of Salmonella recovered from stools of children enrolled in the global enteric multicenter study. Clin Infect Dis. Oxford University Press (OUP); 2021;
36. Wilson CN, Pulford CV, Akoko J, Sepulveda BP, Predeus AV, Bevington J, et al. Salmonella identified in pigs in Kenya and Malawi reveals the potential for zoonotic transmission in emerging pork markets. PLoS Negl Trop Dis. Public Library of Science. 2020;14(11):e0008796. https://doi.org/10.1371/journal.pntd.0008796.
37. Darboe S, Bradbury R, Phelan J, Kanteh A, Muhammad A-K, Worwui A, et al. Genomic diversity and antimicrobial resistance among non-typhoidal *Salmonella* associated with human disease in The Gambia. bioRxiv. Cold Spring Harbor Laboratory. 2020;2020(10):30.316588.
38. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9. https://doi.org/10.1038/nmeth.1923.
39. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170.
40. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. England. 2014;30(14):2068–9. https://doi.org/10.1093/bioinformatics/btu153.
41. Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VPJ, Nash JHE, et al. The *Salmonella In Silico* Typing Resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft *Salmonella* genome assemblies. Hensel M, editor. PLoS One. 2016;11:e0147101.
42. Jolley KA, Maiden MCJ. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 2010;11(1). https://doi.org/10.1186/1471-2105-11-595.
43. Predeus A V. 10 k Salmonella genomes v1.0. https://github.com/apredeus/10k_genomes. https://doi.org/10.5281/zenodo.5512963 (2021).
44. Perez-Sepulveda BM, Heavens D, Pulford C V., Predeus A V., Low R, Webster HJ, et al. A worldwide effort to understand the epidemiology, transmission and virulence of invasive non-Typhoidal Salmonellosis. European Nucleotide Archives. www.ebi.ac.uk/ena/browser/view/PRJEB35182 (2021)
45. Perez-Sepulveda BM, Heavens D, Pulford C V., Predeus A V., Low R, Webster HJ, et al. A worldwide effort to understand the epidemiology, transmission and virulence of invasive non-Typhoidal Salmonellosis. European Nucleotide Archive. www.ebi.ac.uk/ena/browser/view/PRJEB47910 (2021)

Publisher's Note