Article
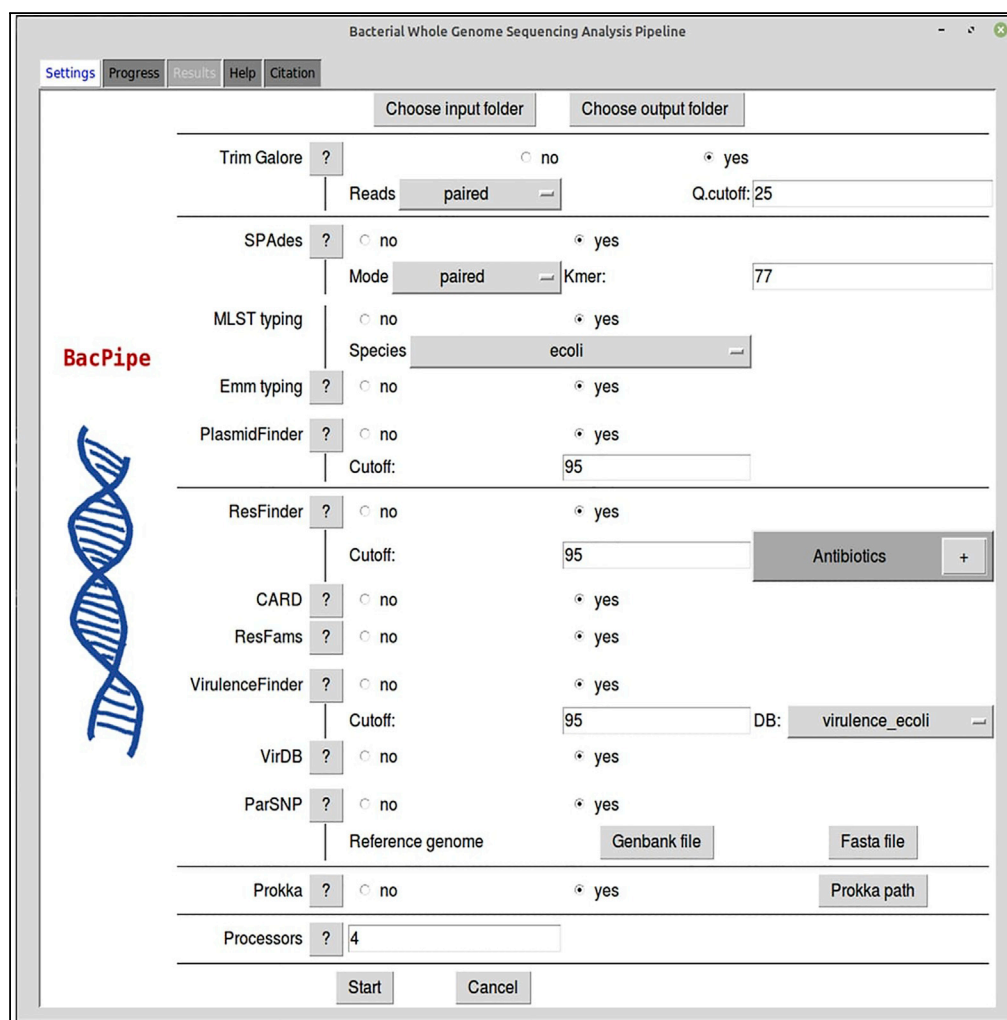
BacPipe: A Rapid, User-Friendly Whole-Genome Sequencing Pipeline for Clinical Diagnostic Bacteriology

Basil B. Xavier,
Mohamed Mysara,
Mattia Bolzan, ...,
João A. Carriço,
Guy Cochrane,
Surbhi Malhotra-
Kumar

surbhi.malhotra@uantwerpen.
be

HIGHLIGHTS

BacPipe is an automated whole genome sequencing pipeline

Interactive user-friendly GUI

BacPipe can process raw reads, contigs, or scaffolds

Time-to-analysis for a 5 Mb genome is ~30–40 min

iScience

Article

BacPipe: A Rapid, User-Friendly Whole-Genome Sequencing Pipeline for Clinical Diagnostic Bacteriology

Basil B. Xavier,[1,2,7] Mohamed Mysara,[1,2,6,7] Mattia Bolzan,[1] Bruno Ribeiro-Gonçalves,[5] Blaise T.F. Alako,[4] Peter Harrison,[4] Christine Lammens,[1,2] Samir Kumar-Singh,[1,3] Herman Goossens,[1,2] João A. Carriço,[5] Guy Cochrane,[4] and Surbhi Malhotra-Kumar[1,2,8,*]

SUMMARY

Despite rapid advances in whole genome sequencing (WGS) technologies, their integration into routine microbiological diagnostics has been hampered by the lack of standardized downstream bioinformatics analysis. We developed a comprehensive and computationally low-resource bioinformatics pipeline (BacPipe) enabling direct analyses of bacterial whole-genome sequences (raw reads or contigs) obtained from second- or third-generation sequencing technologies. A graphical user interface was developed to visualize real-time progression of the analysis. The scalability and speed of Bac-Pipe in handling large datasets was demonstrated using 4,139 Illumina paired-end sequence files of publicly available bacterial genomes (2.9–5.4 Mb) from the European Nucleotide Archive. BacPipe is integrated in EBI-SELECTA, a project-specific portal (H2020-COMPARE), and is available as an independent docker image that can be used across Windows- and Unix-based systems. BacPipe offers a fully automated "one-stop" bacterial WGS analysis pipeline to overcome the major hurdle of WGS data analysis in hospitals and public-health and for infection control monitoring.

INTRODUCTION

Next-generation sequencing (NGS) technologies hold the promise to revolutionize the public health sector especially clinical diagnostic microbiology, infection control, outbreak detection, and antibiotic stewardship in hospitals (Arnold, 2015; Kwong et al., 2015; Moran-Gilad, 2017). As costs of sequencing technologies are steadily decreasing and response times getting shorter, their utility as tools for tracking pathogens in real-time for routine hospital epidemiology or as an early warning system for outbreak detection and detecting multi-drug resistant (MDR) pathogens is steadily increasing (Punina et al., 2015). Currently, depending on the pathogen, the identification and characterization process may take one to seven days for culture, an additional one to two days for species identification and susceptibility testing, and one to several weeks for molecular typing. Whole genome sequencing (WGS) of bacterial isolates combines identification, molecular typing, and prediction of antimicrobial susceptibility and virulence, theoretically reducing the time-to-result for these procedures to a few days (Didelot et al., 2012; Joensen et al., 2014; Koser et al., 2012). However, despite rapid advances in WGS workflows and in NGS technologies, their integration into routine microbiological diagnostics and infection control has been hampered by the need for downstream bioinformatics analyses that is challenging and requires considerable expertise (Deurenberg et al., 2017; Muir et al., 2016). WGS analysis comprises different stages, and each stage is crucial for correct data interpretation. Although there are commercial softwares available such as CLC Genomics Workbench (Qiagen), DNA Star (DNASTAR Inc., USA), BioNumerics (Applied Maths), and SeqSphere[+] (Ridom GmbH, Münster, Germany), the current licensing costs are very high and cannot be sustained by small to medium laboratories. Furthermore, these tools handle the analysis as a black-box for the user and often lag when it comes to integrating state-of-the-art tools compared with publicly managed software/packages (Lüth et al., 2018). Thus, more extensive use of open-source software for whole-genome sequencing data analysis needs to be advocated (Deurenberg et al., 2017).

Several open access tools are available and are split into two categories, web-based analysis or locally downloadable tools. Few web-based open access pipelines such as Orione (http://orione.crs4.it) (Cuccuru et al., 2014) and the Bacterial analysis pipeline (https://cge.cbs.dtu.dk/services/cge/) (Thomsen et al., 2016) and the microbial genomics virtual laboratory (https://nectar.org.au/) are also available (Afgan et al., 2015).

[1]Laboratory of Medical Microbiology, Campus Drie Eiken, University of Antwerp, S6, Universiteitsplein 1, B-2610 Wilrijk, Belgium

[2]Vaccine & Infectious Disease Institute, University of Antwerp, Antwerp 2610, Belgium

[3]Molecular Pathology Group, Cell Biology and Histology, University of Antwerp, Antwerp 2610, Belgium

[4]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge CB10 1SD, UK

[5]Instituto de Microbiologia and Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Av. Professor Egaz Moniz, Lisboa 1649-028, Portugal

[6]Microbiology Unit, Belgian Nuclear Research Center (SCK•CEN), Mol 2400, Belgium

[7]These authors contributed equally

[8]Lead Contact

*Correspondence: surbhi.malhotra@uantwerpen.be

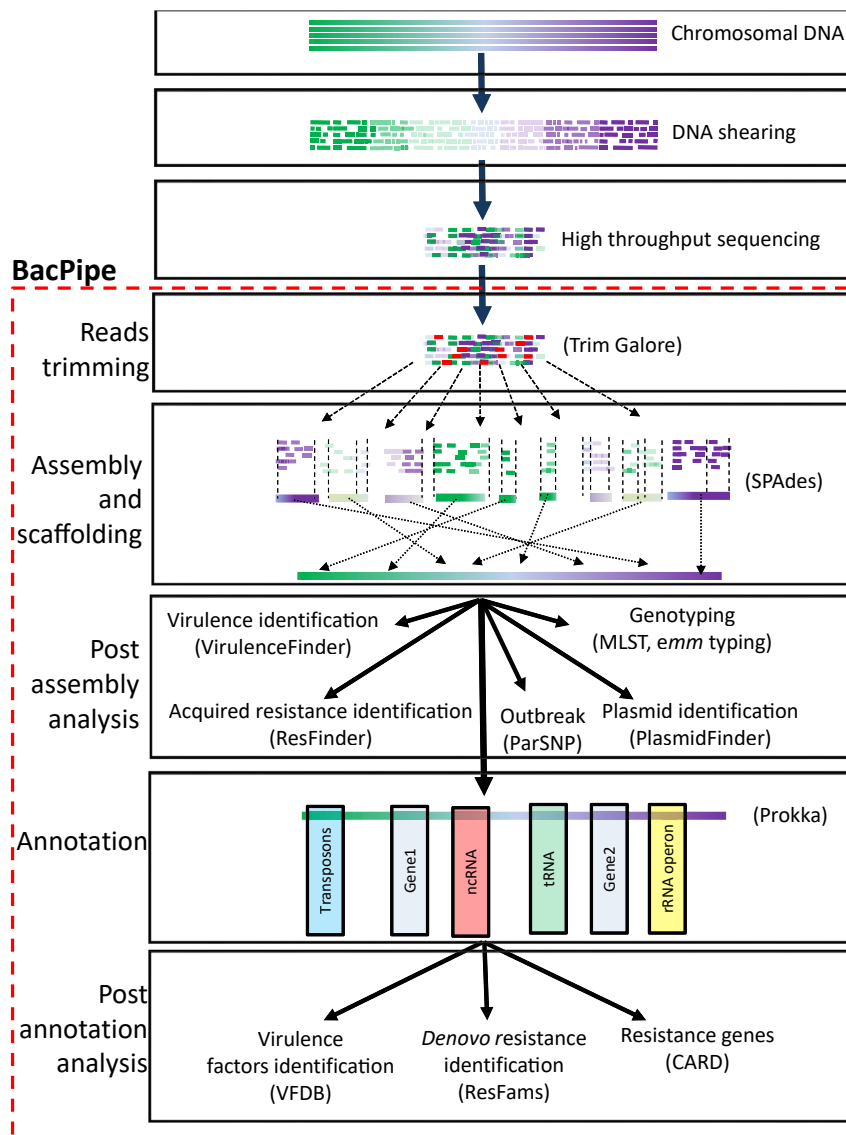https://doi.org/10.1016/j.isci.2019.100769

Figure 1. The Workflow of BacPipe
Complete overview of NGS workflow and analysis performed within BacPipe.

Orione is available in the Galaxy portal (https://usegalaxy.org/), and it offers WGS quality control, assembly and annotation, and variant calling (Cuccuru et al., 2014). The Bacterial analysis pipeline (https://cge.cbs. dtu.dk/services/cge/) offers molecular typing tools as well as resistance and virulence gene predictions and SNP-based phylogeny. However, the performance of web-server based analysis depends on the server load and requires a fast and consistent internet connection to upload large raw data files, which is unreliable when it comes to patient care. Moreover, due to the fact that the analysis is performed remotely, this forms a great barrier for hospital and data protection, which remains a sensitive matter with policies varying between countries (Akgün et al., 2015; Muir et al., 2016).

The second type of the open-source software that those locally installable tools developed specifically for running and managing microbial genomics pipelines includes IRIDA (irida.ca), Innuendo (http:// www.innuendoweb.org/project-definition), and nullarbor (https://github.com/tseemann/nullarbor). IRIDA provides a workflow for assembly (SPAdes), annotation (Prokka), SNP phylogeny (SNVPhyl), resistance (CARD), and virulence (Islandviewer) but not for plasmids and MLST typing. INNUENDO, with its INNUca workflow, provides quality control of reads, *de novo* assembly, and contigs quality assessment.

Figure 2. Snapshot of BacPipe

BacPipe graphical user interface (GUI). See also Figure S1.

Nullarbor supports Illumina paired-end sequencing data but not single-end reads from either Illumina or Ion Torrent.

To add in this list and overcome the various issues discussed above, we have developed a rapid, "one-stop" bacterial WGS analysis pipeline, BacPipe (Figures 1 and 2). This freely available pipeline offers a graphical user interface, parallel computing for fast execution and a containerized granting it standardization of the results across different hospitals. Its open-source software is capable of performing a plethora of analyses starting from raw data quality check, genome assembly, and annotation, resulting in bacterial typing, resistance, and virulence gene predictions, as well as single nucleotide polymorphisms (SNP)-based phylogeny. BacPipe has been successfully implemented to analyze sample from large-scale projects, such as EBI-SELECTA, a rule-based computational workflow engine developed as part of H2020 COMPARE project (https://www.compare-europe.eu/).

RESULTS AND DISCUSSION

BacPipe Implementation and Running Time on Small Number of Strains (as a Function of Genome Size and Sequencing Coverage)

To demonstrate the impact of bacterial genome sizes on the computational time required to obtain results with BacPipe, we used three pathogen genomes that vary considerably in size, *Streptococcus pyogenes*
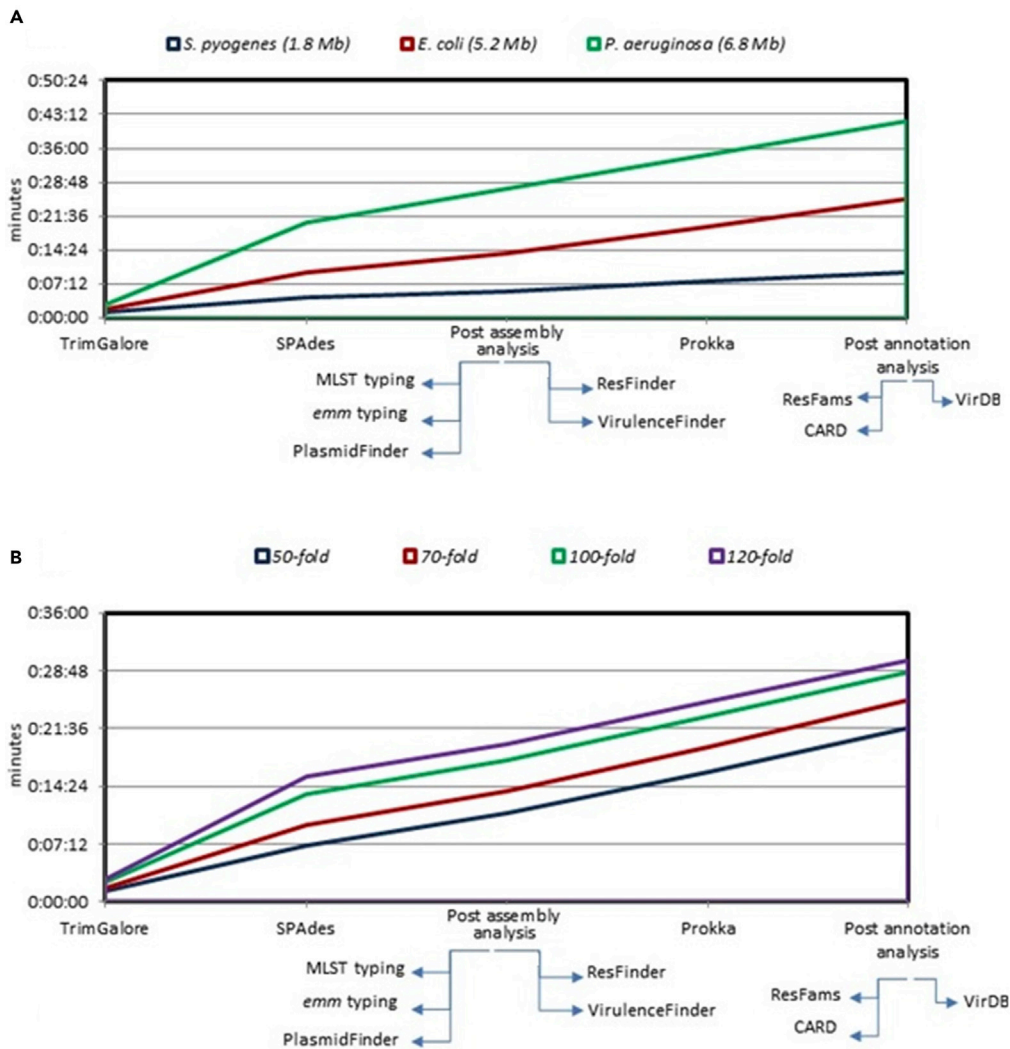
A

B

Figure 3. BacPipe Running Time

Impact of different genome sizes at equal sequencing coverage (70-fold) on the computational time taken for each analysis step in BacPipe (A). Impact of varying sequencing coverage of an *E. coli* genome on the computational time taken for each analysis step in BacPipe (B).

(~1.8 Mb), *Escherichia coli* (~5.2 Mb), and *Pseudomonas aeruginosa* (~6.8 Mb). Whole genome sequences of each pathogen were normalized to the same fold-coverage to demonstrate an increase in computational time as a function of genome size. These internal isolates were sequenced from our in-house MiSeq, and as the *P. aeruginosa* PAO1 had 70-fold coverage, we randomly selected reads from the other two strains resulting in the same coverage. Expectedly, computational time increased with increasing genome size totalling 9, 25, and 41 min for *S. pyogenes, E. coli,* and *P. aeruginosa*, respectively (Figure 3A). Among all the tools employed in the pipeline, as expected, genome assembly (SPAdes) was found to be the most computationally intensive, taking on average 36% of the total running time. Also, we assessed the added value of parallelizing the post-assembly tools (PlasmidFinder, ResFinder, VirulenceFinder, MLST, and *emm* typing) and post-annotation tools (ResFams, VirDB, and CARD search). Parallelizing these tools resulted in a reduction of time-to-result (computational time) by 56%, 29%, and 25% for the three pathogen sequences, respectively (data not shown). Additionally, to emphasize the increase in the computational time due to higher coverage, we subsampled the *E. coli* sequences at 50-, 70-, 100-, and 120-fold-coverage. The required computational time for 50-, 70-, 100-, 120-fold-coverages were 21, 25, 28 and 30 min, respectively (Figure 3B). For this benchmark, we used a MacBook Pro, 2.5 GHz, quad-core i7 with 16 GB RAMS (DDR3), 4 cores, and SSD hard drive.
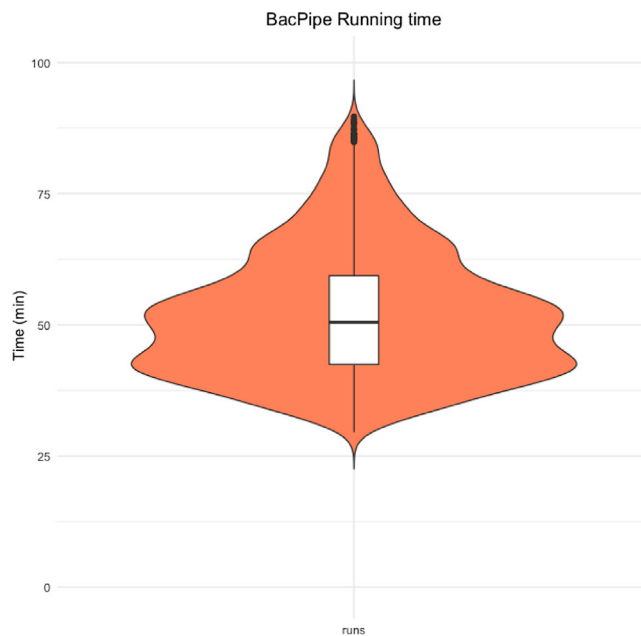
Figure 4. Large Scale Validation of BacPipe

BacPipe running time (on average 50 min/run) over 4,000 paired-end sequence reads of bacterial genomes. This process was performed on the EBI high-performance computing platform is an EBI shared facility made up of 130 nodes with 130Gb of RAM each and 2 core per node with 40 CPUs (See also Figures S2 and S3 and Table S1).

BacPipe Implementation and Running Time on Publicly Available Bacterial Genomes at Large Scale (EBI-SELECTA Framework)

Within the SELECTA framework, BacPipe was used to analyze 4,139 paired-end publicly available WGS sequence reads for the bacterial genomes listed in Table S1. An example of an analysis result can be found here (https://www.ebi.ac.uk/ena/data/view/ERZ799760). This implementation demonstrated the potential of BacPipe in processing a large number of runs on a short timescale (Figure 4).

Validation of BacPipe's Functionality Using Prior Published Data

We challenged BacPipe with various bacterial genomes including those with higher GC content and multiple repeat regions (*M. tuberculosis*). Mainly, five previously published and analyzed WGS datasets (raw reads or assembled contigs) from hospital outbreaks caused by MRSA and carbapenem-resistant *K. pneumoniae* (Snitkin et al., 2012), a 3-year long in-hospital transmission study of *C. difficile* (Jia et al., 2016), a community-based surveillance and transmission study of *M. tuberculosis* (Kohl et al., 2014), and finally a foodborne outbreak caused by *S. enterica* (Taylor et al., 2015) were utilized. We attempted to recreate the same analyses as reported in the respective publications to demonstrate the "one-stop" analysis with BacPipe.

Outbreak Dynamics of MRSA in an Academic Hospital of Paramaribo, Republic of Suriname

The recent work of Sabat et al reported an investigation of an MRSA outbreak at the Academic Hospital Paramaribo (AZP), Suriname from April to May 2013. The outbreak included 12 patients and one healthcare worker/nurse at the AZP totaling 24 isolates that were used to investigate phylogenetic relatedness and transmission (Sabat et al., 2017). In this study, isolates were sequenced on the MiSeq (V3 kit), and downstream analysis were done using commercial software SeqMan NGen and SeqMan Pro (DNASTAR Inc., USA). Annotation was done using NCBI prokaryotic genome annotation pipeline (PGAP) (Tatusova et al., 2016), and MLST, acquired resistance genes, and SNP analyses were performed using the CGE (http://genomicepidemiology.org/) tools. The data are available under the Bioproject accession number PRJNA312385.

We analyzed all raw reads belonging to 24 isolates and 63 plasmids from this study using BacPipe and produced same results. Firstly, we constructed an SNP-based phylogenetic tree similar to that of Sabat et al.
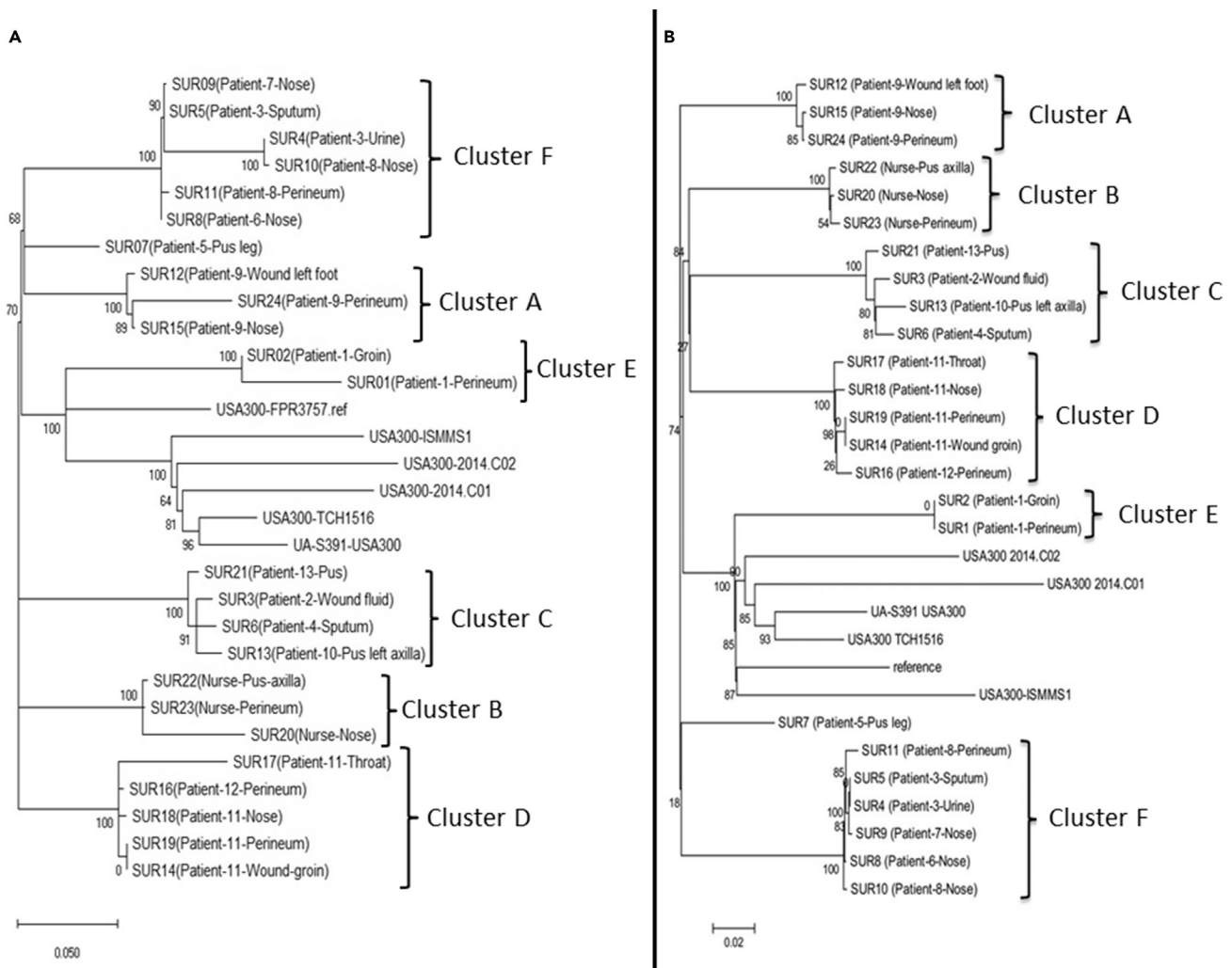
Figure 5. Comparison of Phylogenetic Analysis

Phylogenetic maximum likelihood tree generated from core-genome SNPs generated through BacPipe and visualized by TreeView tool (A) and from Sabat et al. (Sabat et al., 2017) (B). The scale bar indicates the evolutionary distance between the sequences determined by 0.1 substitutions per nucleotide at the variable positions. See also Data S1.

(Sabat et al., 2017), consisting of six distinct clusters (A–F) and one singleton (SUR7) (Figures 5A and 5B). Secondly, the pipeline assessed the MLST of all isolates as ST8, as reported, and confirmed the loss of *splD* and *splE* genes (representing important virulence factors) from Cluster F (Data S1). Similar to what was reported, antibiotic resistance patterns of all isolates showed the presence of *dfrG* trimethoprim resistance, with exception to ClusterE, whereas the *ermC* gene, conferring resistance to clindamycin, was identified in all isolates of ClusterF and two of ClusterA isolates. For the remainder, it was possible to confirm identical resistance profiles found in all isolates to the previously reported ones including *blaZ, mecA, ermC, aphA3, str, msrA,* and *mphC* genes. Thus, similar to the conclusions of Sabat et al. (Sabat et al., 2017), we also identified utilizing BacPipe, a heterogeneous population structure, during this outbreak driven by the different body sites of the same patient or existence of direct transmission between patients. Additionally, virulence factors *ssp, atl, efb,* and *esa* were also detected in all analyzed strains by the VFDB database in BacPipe (Chen et al., 2016) (Data S1).

Tracking a Hospital Outbreak of Carbapenem-Resistant *Klebsiella pneumoniae*

Snitkin et al described a carbapenem-resistant *K. pneumoniae* (CRE) outbreak in 2011 at the US National Institutes of Health Clinical Center that affected 18 patients of whom 11 died (Snitkin et al., 2012). The first
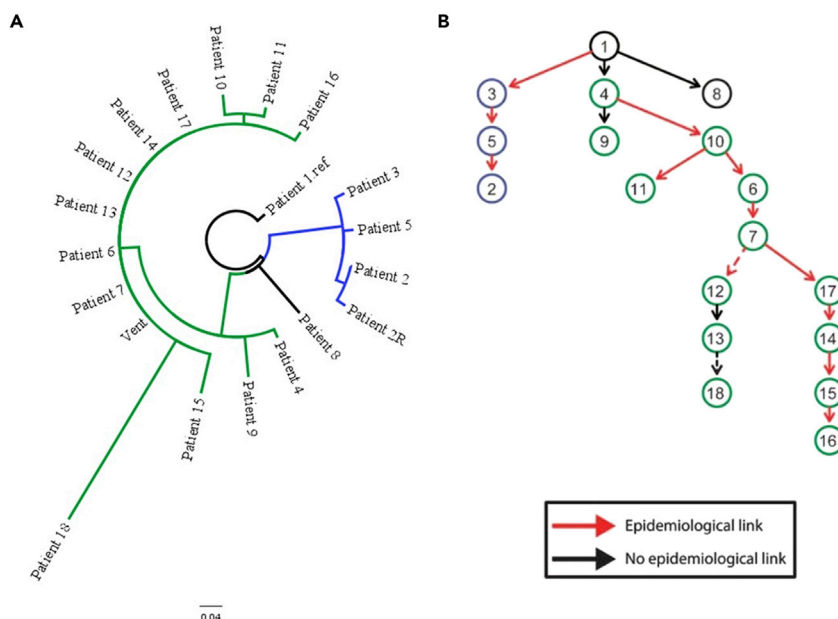
Figure 6. Comparison of Phylogenetic Analysis

Phylogenetic maximum likelihood tree generated through BacPipe and visualized by TreeView tool (A). Putative map of *K. pneumoniae* transmission during outbreak reproduced from Snitkin et al. (Snitkin et al., 2012). Nodes represent patients, and arrows indicate a transmission event directly or indirectly from one patient to another (B). See also Data S1.

patient colonized with CRE was placed under contact isolation and treated, yet after three weeks of her discharge, one new case of colonization or active infection was detected every week at the center totalling up to 17 patients. To answer the central question whether patient 1 had initiated the outbreak and if so, how was she linked to the other affected patients, CRE isolated from the 18 patients' samples were analyzed by WGS Roche/454 XLR instrument (Roche Life Sciences). Assembly and annotation was done using gsAssembler and NCBI PGAP, respectively (Snitkin et al., 2012). The data is available under Bioproject accession number PRJNA73841.

The 18 strain sequences were processed through BacPipe. As reported in the study, all 18 CRE belonged to the epidemic ST258 clone and harbored *bla*KPC-3. SNP-based phylogenetic construction showed two large clusters and a third cluster consisting only of patient 8 and demonstrated that patient 1 was not only linked to the outbreak but also that three independent transmissions of genetically distinct isolates occurred from patient 1 to other patients (Figures 6A and 6B).

Additional antibiotic resistance genes such as *bla*SHV, *blaTEM, blaOXA, fosA, mphA, catA, oqxA, oqxB, sul1, dfrA12,* and *aadA2* were also identified in the isolates as were plasmid types *IncFII(K), IncFIB (pQil), IncFIB (K),* and *ColRNAI* and a virulence gene, *cii* in a single process rather than multi-stage analysis as in publication (Data S1).

Tracing Nosocomial Transmission of *Clostridium difficile* Ribotype 027 in a Chinese Hospital, 2012–2014

In the study by Jia et al. 2016 (Jia et al., 2016), a rare case of *C. difficile* bloodstream infection (CDBI) was identified. Consequently, all cases or strains that had emerged from the same ward during the past three years were retrospectively analyzed by WGS. Of the 75 patients presenting with diarrhea, *C. difficile* was isolated from 20 patients, including the case with CDBI. Isolates were sequenced on the HiSeq platform, reads were mapped to R20291 (NAP1/BI/027, ST1) reference strain using the REALPHY tool (Bertels et al., 2014), and the phylogenetic tree was reconstructed by the BEAST tool, whereas the genomic SNP differences between strains were detected using SOAP2 (Li et al., 2009). The data are available under Bioproject accession number PRJNA271048.
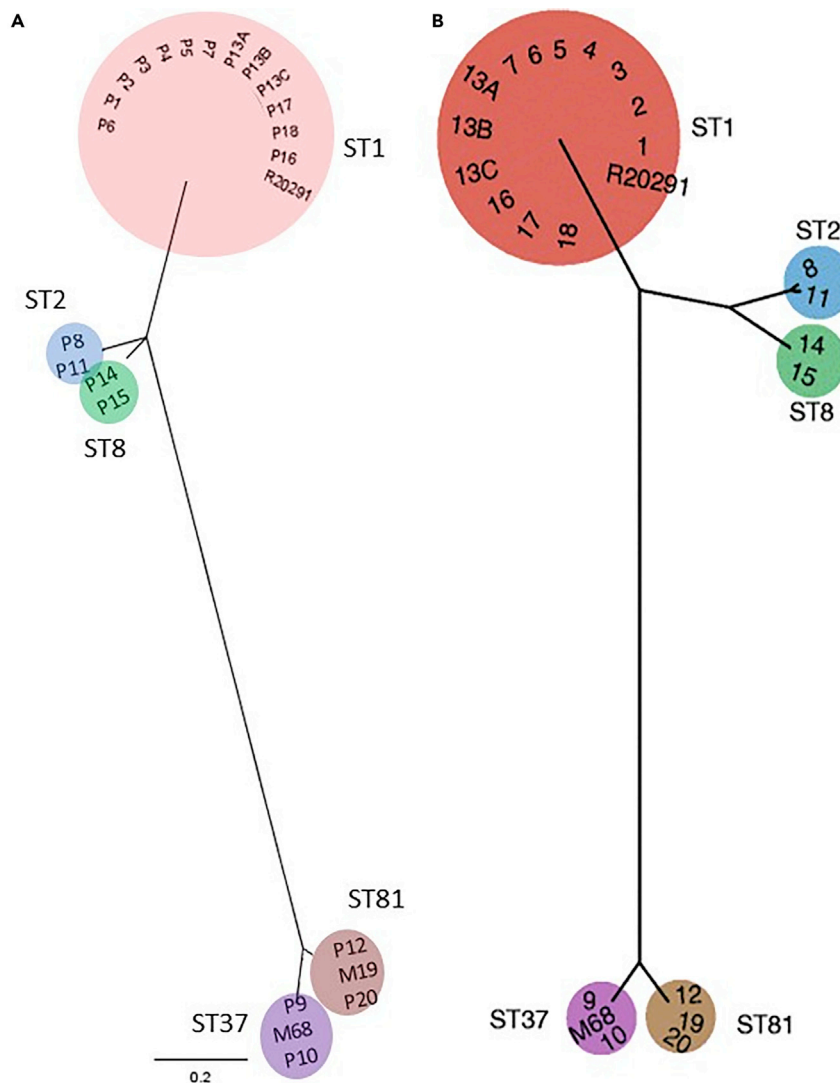
Figure 7. Comparison of Phylogenetic Analysis

Phylogenetic maximum likelihood tree of *C. difficile* generated through BacPipe and visualized by TreeView tool (A) and tree reconstructed from multimapping files via Bayesian evolutionary analysis by BEAST from Jia et al. (Jia et al., 2016) (B). See also Data S1.

BacPipe analysis was able to reproduce the MLST results, where the isolates were characterized into five STs: ST1 (11 patients), ST2 (2 patients), ST8 (2 patients), ST37 (2 patients), and, and ST81 (3 patients) (Data S1). From the SNP-based phylogenetic analysis, we confirmed the finding of Jia et al. of a clear separation between isolates of different STs and that all ST1 isolates were monoclonal (Figures 7A and 7B).

Additional data not reported in this study but generated through BacPipe were as follows: *aac(6′)-aph(2″)* gene conferring aminoglycoside resistance and *erm(B)* conferring macrolide resistance were identified in all isolates belonging to ST1, ST38, and ST81, whereas *tet(M)* conferring tetracycline resistance was identified in isolates belonging to ST37 and ST81. Additionally, for all isolates belonging to ST1, we were able to identify *rep1* plasmid, which was not detected in the other non-ST1 isolates (Data S1).

Whole-Genome-Based Surveillance of *Mycobacterium tuberculosis*

Kohl et al. trace an *M. tuberculosis* complex (MTBC) longitudinal outbreak compromising 26 isolates (between 2001 and 2010) showing identical *IS 6110* DNA fingerprint and spoligotype patterns. These
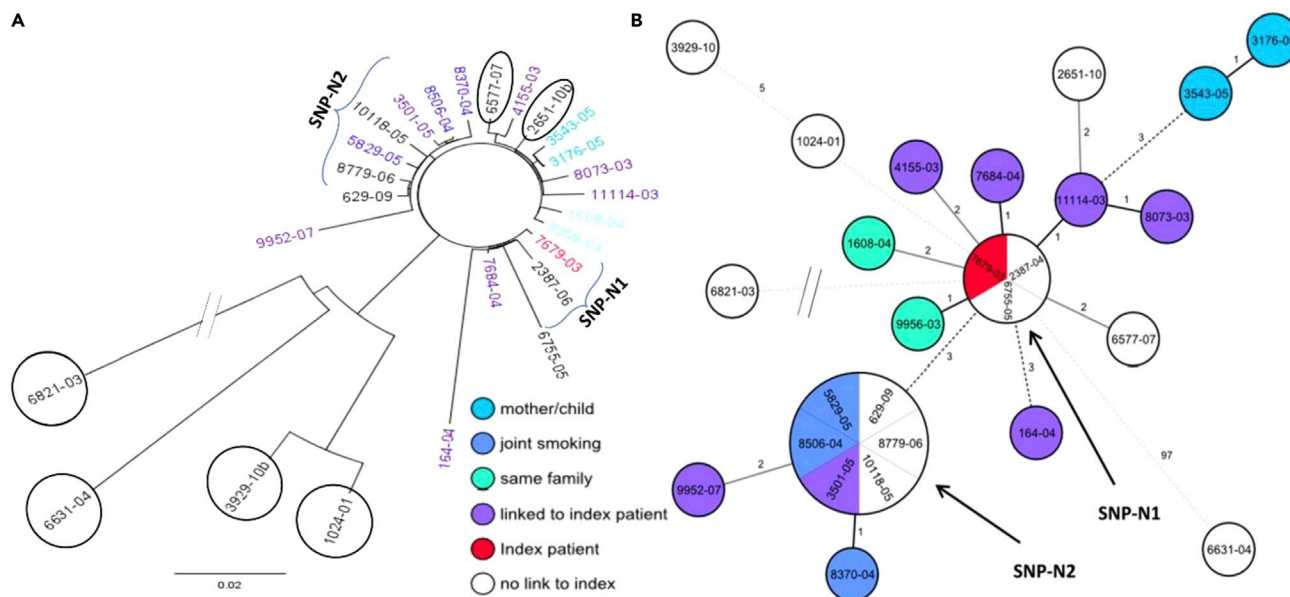
Figure 8. Comparison of Phylogenetic Analysis

Phylogenetic maximum likelihood tree of *M. tuberculosis* core-genome SNPs generated through BacPipe and visualized by TreeView tool (A) and a minimum spanning tree of concatenated sequences of the 322 SNPs of the same data from Kohl et al (Kohl et al., 2014) (B). See also Data S1.

underwent WGS using MiSeq (Illumina), reads were mapped to the H37Rv reference genome using the exact alignment program SARUMAN, and SNPs were extracted from the mapped reads by customized Perl scripts (Kohl et al., 2014). Raw reads are available on Bioproject accession number PRJEB6276.

Using BacPipe, we confirmed that 22 isolates were grouped into one major cluster while four were outliers. Within the primary cluster, we also confirmed two sub-groups compromising of three and six strains, SNP-N1 and SNP-N2, respectively (Figures 8A and 8B).

Additional data not reported in this study but generated through BacPipe were an assessment of antibiotic resistance where all isolates harbored *aac(2′)-Ic* and subclass B1 beta-lactamase genes conferring aminoglycoside and beta-lactam resistance, respectively (Supplemental Information).

Characterization of Foodborne Outbreaks of *Salmonella enterica* Serovar Enteritidis with Whole Genome Sequencing for Surveillance and Outbreak Detection

Taylor et al. described the application of whole genome sequencing for the detection of *S. enterica* serovar Enteritidis outbreaks from isolates previously characterized by PFGE in Minnesota and Ohio between 2001 and 2014 (Taylor et al., 2015). The cohort contained 28 isolates from seven epidemiologically confirmed foodborne outbreaks and 27 non-epidemiologically linked sporadic isolates that were assessed by WGS (MiSeq). Reads were mapped to the reference genome using BWA-MEM tools, which were later sorted and de-duplicated by the Picard tool. The variant call file (VCF) was produced with BCF tool, and the maximum-likelihood phylogenetic tree was calculated with PhyML. Raw reads are available on accession number: PRJNA237212.

BacPipe was able to retrieve the same phylogenetic tree, confirming that all isolates within the same outbreak were closely related (ranging from 2–7 isolates per outbreak) (Figure 9A). We derived the same conclusion as the original study, that the serovar Enteritidis shows little genetic diversity in the host over time, from investigating isolates MDH-2014-00222, MDH-2014-00223, MDH-2014-00225, and MDH-2014-00228 that were isolated from an individual over a five-week period (see outbreak 2 in Figure 9B).

Additional data not reported in this study but generated through BacPipe were the assignment of all strains as ST11 and identification of *IncFIB(S)* and *IncFII(S)* plasmids within all isolates, with exception of
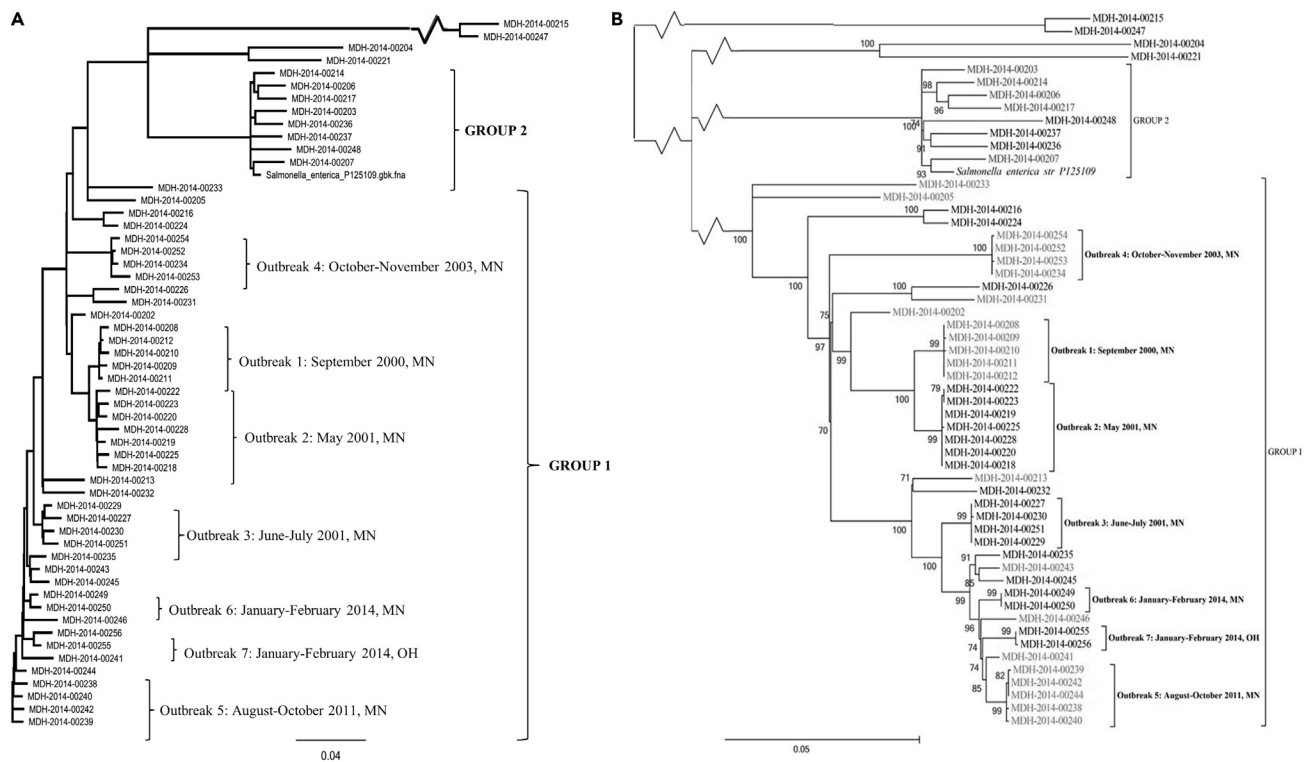
Figure 9. Comparison of Phylogenetic Analysis

Maximum-likelihood tree of *S. enteritidis* produced by SNP analysis showing outbreak clusters and time frame (month[s] and year) and the State from where each isolate originated. The phylogenetic analysis generated through BacPipe and visualized by TreeView tool (A) and tree reproduced from Taylor et al. (Taylor et al., 2015) (B). See also Data S1.

MDH-2014-00232, MDH-2014-00245, and outbreak 6/7 isolates—where no plasmids were detected—and MDH-2014-00215 and MDH-2014-00247 isolates—where *IncHI1B*, *Incl1*, *IncHI1A*, and *IncFIA(HI1)* plasmids were identified. Similarly, from MDH-2014-00215 and MDH-2014-00247, antibiotic resistance genes such as *blaTEM-1B*, *catA1*, *sul1*, *tet(B)*, and *dfrA7* were identified, whereas none were found in the remaining isolates (Data S1).

Conclusion

Here we have presented BacPipe, a bacterial whole genome sequencing analysis pipeline and demonstrated its robustness in handling diverse genomes of clinically important pathogens characterized by different sizes, GC content, and presence of repeat regions that are challenging for downstream data analysis. Along with being comprehensive and modular, BacPipe has the advantage of being a fast-on account of parallel computing, and requiring computationally low-resource as pipeline functionality does not require an internet connection or high-end computers. This would also allow the analysis to be performed locally, which is highly advantageous to hospitals that are mandated to comply with data protection guidelines.

A graphical interface makes it very user-friendly; a user can specify the tools to allow visualization of the included in the analysis and adjust the database/parameters from a drop-down list or buttons. BacPipe can be run either with raw reads from various sequencing platforms or can pick-up the analysis from any step throughout the workflow giving tremendous flexibility. The open-source nature and GNU license would allow more expert users to modify and adapt the software to their preferences.

The endpoint of the analysis provides various levels of details, from an overview comparing the results across all analyzed samples, to an Excel file with all of the compiled tool results, to very detailed

folders dedicated for each tool output and log files. Additionally, the output of BacPipe can easily be used to study the pan-genome and perform comparative genome analysis, define gene acquisition/loss through horizontal gene transfer, and perform functional analysis through the KEGG ortholog database. Finally, although prior publications have utilized numerous heterogeneous tools to delineate hospital or community-based pathogen transmission, we demonstrated that the collection of tools within BacPipe could reproduce the entire analyses as a "one-stop" platform in less than an hour. Future development of BacPipe entails expansion of tools to enable identification of prophages, IS (insertion sequence) elements, CRISPR-Cas elements, and, depending upon their open-access status, whole-/core-/pan-genome MLST schemes. We believe this fully automated pipeline will help to overcome one of the primary barriers to analyzing and interpreting WGS data, facilitating applications for routine patient care in hospitals and public health and infection control monitoring.

Limitations of the Study

Although Bacpipe is a fully automated pipeline with a user-friendly GUI that requires minimal user intervention, the heavy reliance on open-access data resources makes it imperative that these are well-curated, up-to-date, and comprehensive.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

DATA AND CODE AVAILABILITY

BacPipe can be obtained through https://hub.docker.com/r/mahmed/bacpipe (docker image) https://github.com/wholeGenomeSequencingAnalysisPipeline/BacPipe (Github approach).

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2019.100769.

AUTHOR CONTRIBUTIONS

This work was conceptualized by S.M.K. The study was designed by S.M.K. and B.B.X. The pipeline was developed and validated by M.M., B.B.X., M.B., and C.L. B.R.G. and J.A.C. contributed to the dockerization of the platform. B.T.F.A. and P.H. integrated and validated tool at EBI-SELECTA. The manuscript was drafted by B.B.X., M.M., S.K.S., G.C., H.G., and S.M.K. and was reviewed by all authors.

DECLARATION OF INTERESTS

None declared.

REFERENCES

Afgan, E., Sloggett, C., Goonasekera, N., Makunin, I., Benson, D., Crowe, M., Gladman, S., Kowsar, Y., Pheasant, M., Horst, R., et al. (2015). Genomics virtual laboratory: a practical bioinformatics workbench for the cloud. PLoS One *10*, e0140829.

Akgün, M., Bayrak, A.O., Ozer, B., and Sağıroğlu, M.Ş. (2015). Privacy preserving processing of genomic data: a survey. J. Biomed. Inform. *56*, 103–111.

Arnold, C. (2015). Outbreak breakthrough: using whole-genome sequencing to control hospital infection. Environ. Health Perspect. *123*, A281–A286.

Bertels, F., Silander, O.K., Pachkov, M., Rainey, P.B., and van Nimwegen, E. (2014). Automated reconstruction of whole-genome phylogenies from short-sequence reads. Mol. Biol. Evol. *31*, 1077–1088.

Chen, L., Zheng, D., Liu, B., Yang, J., and Jin, Q. (2016). VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. Nucleic Acids Res. *44*, D694–D697.

Cuccuru, G., Orsini, M., Pinna, A., Sbardellati, A., Soranzo, N., Travaglione, A., Uva, P., Zanetti, G., and Fotia, G. (2014). Orione, a web-based framework for NGS analysis in microbiology. Bioinformatics *30*, 1928–1929.

Deurenberg, R.H., Bathoorn, E., Chlebowicz, M.A., Couto, N., Ferdous, M., García-Cobos, S., Kooistra-Smid, A.M.D., Raangs, E.C., Rosema, S., Veloo, A.C.M., et al. (2017). Application of next generation sequencing in clinical microbiology and infection prevention. J. Biotechnol. *243*, 16–24.

Didelot, X., Bowden, R., Wilson, D.J., Peto, T.E.A., and Crook, D.W. (2012). Transforming clinical microbiology with bacterial genome sequencing. Nat. Rev. Genet. *13*, 601.

Jia, H., Du, P., Yang, H., Zhang, Y., Wang, J., Zhang, W., Han, G., Han, N., Yao, Z., Wang, H.,

et al. (2016). Nosocomial transmission of Clostridium difficile ribotype 027 in a Chinese hospital, 2012–2014, traced by whole genome sequencing. BMC Genomics *17*, 405.

Joensen, K.G., Scheutz, F., Lund, O., Hasman, H., Kaas, R.S., Nielsen, E.M., and Aarestrup, F.M. (2014). Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. J. Clin. Microbiol. *52*, 1501–1510.

Kohl, T.A., Diel, R., Harmsen, D., Rothgänger, J., Walter, K.M., Merker, M., Weniger, T., and Niemann, S. (2014). Whole-genome-based Mycobacterium tuberculosis surveillance: a standardized, portable, and expandable approach. J. Clin. Microbiol. *52*, 2479–2486.

Koser, C.U., Ellington, M.J., Cartwright, E.J., Gillespie, S.H., Brown, N.M., Farrington, M., Holden, M.T., Dougan, G., Bentley, S.D., Parkhill, J., et al. (2012). Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog. *8*, e1002824.

Kwong, J.C., McCallum, N., Sintchenko, V., and Howden, B.P. (2015). Whole genome sequencing in clinical and public health microbiology. Pathology *47*, 199–210.

Li, R., Yu, C., and Li, Y. (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics *25*, 1966–1967.

Lüth, S., Kleta, S., and Al Dahouk, S. (2018). Whole genome sequencing as a typing tool for foodborne pathogens like Listeria monocytogenes – the way towards global harmonisation and data exchange. Trends Food Sci. Technology *73*, 67–75.

Moran-Gilad, J. (2017). Whole genome sequencing (WGS) for food-borne pathogen surveillance and control - taking the pulse. Euro Surveill. *22*, 30547.

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D.J., Salichos, L., Zhang, J., Weinstock, G.M., Isaacs, F.,

Rozowsky, J., et al. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol. *17*, 53.

Punina, N.V., Makridakis, N.M., Remnev, M.A., and Topunov, A.F. (2015). Whole-genome sequencing targets drug-resistant bacterial infections. Hum. Genomics *9*, 19.

Sabat, A.J., Hermelijn, S.M., Akkerboom, V., Juliana, A., Degener, J.E., Grundmann, H., and Friedrich, A.W. (2017). Complete-genome sequencing elucidates outbreak dynamics of CA-MRSA USA300 (ST8-spa t008) in an academic hospital of Paramaribo, Republic of Suriname. Scientific Rep. *7*, 41050.

Snitkin, E.S., Zelazny, A.M., Thomas, P.J., Stock, F., Henderson, D.K., Palmore, T.N., and Segre, J.A. (2012). Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. Sci. Transl Med. *4*, 148ra116.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E.P., Zaslavsky, L., Lomsadze, A., Pruitt, K.D., Borodovsky, M., and Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. *44*, 6614–6624.

Taylor, A.J., Lappi, V., Wolfgang, W.J., Lapierre, P., Palumbo, M.J., Medus, C., and Boxrud, D. (2015). Characterization of foodborne outbreaks of Salmonella enterica serovar enteritidis with whole-genome sequencing single nucleotide polymorphism-based analysis for surveillance and outbreak detection. J. Clin. Microbiol. *53*, 3334–3340.

Thomsen, M.C.F., Ahrenfeldt, J., Cisneros, J.L.B., Jurtz, V., Larsen, M.V., Hasman, H., Aarestrup, F.M., and Lund, O. (2016). A bacterial analysis platform: an integrated system for analysing bacterial whole genome sequencing data for clinical diagnostics and surveillance. PLoS One *11*, e0157718.

Supplemental Information

BacPipe: A Rapid, User-Friendly

Whole-Genome Sequencing Pipeline

for Clinical Diagnostic Bacteriology

Basil B. Xavier, Mohamed Mysara, Mattia Bolzan, Bruno Ribeiro-Gonçalves, Blaise T.F. Alako, Peter Harrison, Christine Lammens, Samir Kumar-Singh, Herman Goossens, João A. Carriço, Guy Cochrane, and Surbhi Malhotra-Kumar

Materials and Methods

DEVELOPMENT OF BACPIPE

We first reviewed the available open-access tools that were potential candidates for inclusion within BacPipe for each stage of analyses. These included tools required for quality control check, assembly, as well as specialized tools for bacterial and plasmid typing, and for resistance and virulence gene predictions. As some of these specialized tools required annotation files (gbk/gff) as input, these were divided within BacPipe into those that required assembly (post-assembly tools) or annotation (post-annotation analysis) (Figure 1). The optimal and widely used tools were selected for inclusion in BacPipe as described below. Furthermore, to increase the user-friendliness of the pipeline, we also integrated a graphical user interface (GUI) (Figure 2). The GUI was developed using AppJar package (http://appjar.info/), split into four tabs: Settings (inputs and tools parameters), Progress (percentage finished and log file), Results (overall summary files are shown), Help (information regarding the input/outputs), and Citation (for all tools included). BacPipe is also modular i.e., depending on the analyses required, the user can select for a particular tool or a set of tools that will also further speed up the time-to-result (Figure 2). Additionally, users can choose to directly analyse raw sequencing reads, contigs, or scaffolds (for instance, published sequences).

Reads quality filtering and adapter trimming

Quality control and processing of raw reads are the initial steps and are extremely crucial for robust downstream analysis. For quality control, we opted for Trim Galore (v0.6.2, https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) that removes sequencing technology-specific sequences and trims low-quality reads. This tool consists of FastQC and

Cutadapt, the former to check the quality of the reads and the latter to trim the sequencing-specific (adapter and index) sequences.

Generating contigs and scaffolds from reads: Genome assembly

To obtain reliable information on the genetic context of genes, an error-free assembly of the genome sequences is a must. Here, we utilized SPAdes (v3.13.0) that can automatically optimize the k-mer based on read length in combination with a scaffolding step (16). The tool produces reasonably large scaffolds and with higher N50 values compared to other existing tools. Also, there are multiple options in SPAdes for different assembly purposes, such as plasmid and hybrid assembly, and output data from diverse sequencing technologies. Additionally, our previous work validating various assemblers using whole genome mapping (17), showed that SPAdes produces larger scaffolds that are free from misassemblies and, was the best performing assembler among the evaluated assembly tools (i.e., Velvet and IDBA).

Pathogen typing, detection of plasmids, virulence and resistance genes, and core genome phylogenetic analysis: Post assembly analysis

The post-assembly analysis tools include those for multi-locus sequence typing (MLST, v2.0) (18), plasmid incompatibility typing (PlasmidFinder, v2.1) (19), antibiotic resistance gene predictions (Resfinder, v3.2) (20), and for virulence gene predictions (VirulenceFinder, v2.0) (6). PlasmidFinder can classify plasmids into various incompatibility types based on some plasmid-specific genes (19). VirulenceFinder can predict putative virulence genes in the scaffolds, however, currently, gene predictions are only possible for select bacterial species (*E. coli, S. aureus*, *Enterococcus spp.* and *Listeria*) (6). All post-assembly detection tools in BacPipe, described above, work with a local BLAST (v2.9) search utilizing the databases downloaded from

the Center for Genomic Epidemiology (CGE) (https://bitbucket.org/account/user/genomicepidemiology/projects/DB).

Additionally, as a first step to adding pathogen-specific tools, we have added the option of utilizing the *S. pyogenes*-specific *emm* typing (21) tool in BacPipe. This tool also uses a local BLAST search against the curated database (identity 100%; minimum length 90%) downloaded from https://www.cdc.gov/streplab/m-proteingene-typing.html (Center for Disease Control, CDC, United States). BacPipe also includes a module for SNP detection. This module utilizes ParSNP (v1.2), a tool using MUMmer (mummer.sourceforge.net/) for comparison of scaffolds/genomes and generates a SNP-based core genome phylogeny. This module allows the user to choose the index or reference strain (GenBank or FASTA file) to compare. The analysis produces the output variant calling file (VCF), tree file and multiple alignment files (22). The Newick file or tree file can then be visualized using Gingr, Figtree (http://tree.bio.ed.ac.uk/software/figtree/), or other online tools (www.phylogeny.fr).

Genome annotation

Correct annotation of a gene/genome is a requisite for correct biological interpretation of sequencing data, and to study the genomic features, structural variations and evolutionary relationships. It is also used to assess the quality of the assembly and required sequencing depth by identified complete or core genes of the sequenced organism. Prokka (Prokaryotic annotation tool) is used in Bacpipe as an annotation tool to predict protein coding genes, Barranp (v0.8) for ribosomal RNA (5S, 23S, 16S) predictions and ARAGORN for transfer RNA (tRNA) and transfer messenger RNA (tmRNA) predictions (23). As an output, generic output file formats gbk, gff3 and sqn (sequin) are generated as these file formats are required for the downstream analysis and for uploading to public databases such as Genbank/ENA/DDBJ for submission.

Post-annotation analysis

Post-annotation tools in BacPipe include those that require protein sequences as queries such as the Comprehensive Antibiotic Resistance Database (CARD, v3.0.5) that predicts resistance genes using local blastp (minimum identity 80%; minimum length of 80%) (24, 25). CARD adds to the ResFinder (v3.2) database as the former also identifies resistance genes with a chromosomal origin. Another resistance gene prediction tool integrated into BacPipe is Resfams (v1.2), a curated database of protein families that uses hidden Markov model profiles of resistance genes to screen for distantly-related or novel resistance genes (26).

For virulence gene predictions, we have also integrated Virulence Factor Database (VFDB) (with 80% identity and minimum query length for blastp), an extensive database of virulence genes in pathogenic bacteria, as VirulenceFinder is restricted to a few organisms (27). Finally, all these tools were linked through a master python script, that finally produce/arrange the output reports.

Parallelization of tools

BacPipe is designed to run multiple tools simultaneously which considerably reduces the time-to-result. Tools that accept multi-threading, such as SPAdes and Prokka, are parsed the number of cores allocated, while those that can only run with a single core, as in post-assembly/post-annotation tools run in parallel maximizing the utilization of the resources. To demonstrate the impact of bacterial genome size on the computational time required, we processed three in house sequenced bacterial strains that vary considerably in size, i.e., *Streptococcus pyogenes* (~1.8 Mb)*, Escherichia coli* (~5.2 Mb)*,* and *Pseudomonas aeruginosa* (~6.8 Mb). Whole genome sequences of all three pathogens were normalized to the same 70-fold-coverage before running them through BacPipe. Additionally, we also studied the impact of the sequencing depth on computational time

by subsampling and analyzing the *E. coli* sequences at 50, 70, 100 and 200-fold coverage. The normalization was done using mothur (v1.39) *sub.sample* command.

Output reports

The results from each tool output for each strain analysed are saved as separate folders. In addition, a summary of the results for each strain are also saved as an Excel file in the overall 'summary' folder. Furthermore, if multiple strains are being run in parallel, then results of all tested strains are available in a comprehensive table after the run is finished, which is latter presented as in Supplementary Figure 1.

SOFTWARE AVAILABILITY

BacPipe is available via a Software containerisation or through local download and installation. For the former, we propose using docker image (28) that allows the creation of light container with all the necessary software and dependencies included. This platform independent docker image would grant BacPipe to be used across Windows and Unix would allow preserving a standardized image of Bacpipe and its dependencies with reproducible results and performance across various users/hospitals (Supplementary File 2). The latter approach (through GitHub), uses manual download and local installation in various platforms (unix, mac or windows) through a virtual machine. An automatic installation script (for the non-expert) as well as a detailed installation guidance are available.

BACPIPE INTEGRATION INTO EMBL-EUROPEAN BIOINFORMATICS INSTITUTE SELECTA

The Bacterial whole genome sequencing analysis pipeline (BacPipe 1.2.6) has been fully integrated into the EMBL-European Bioinformatics Institute SELECTA framework. SELECTA is a rule-based computational workflow engine developed within the COMPARE initiative.

SELECTA automates, the selection of data to be processed, the analysis of the data through the dedicated pipeline and the submission of generated results to the European Nucleotide Archive (ENA) for subsequent discovery and retrieval, (Supplementary Figure 2). The submitted analysis data is also made available through the COMPARE pathogens portal (https://www.ebi.ac.uk/ena/pathogens/; Supplementary Figure 3).

VALIDATION OF BACPIPE'S FUNCTIONALITY USING PRIOR PUBLISHED DATA

In order to compare the results of the BacPipe pipeline, we have reanalyzed five previously published whole genome sequenced datasets that spanned the most important multi-drug resistant and virulent pathogens causing infections in hospitals and the community. These included datasets of methicillin-resistant *S. aureus* (MRSA) (29), carbapenem-resistant *K. pneumoniae* (30), *C. difficile* (31), *M. tuberculosis* (32), and *S. enterica* (33).

Glossary

N50: Parameter to define the quality of the genome assembly by the size and a number of contigs or scaffolds produced by the assembler.

Genome assembly: Raw sequencing reads are stitched into larger contiguous sequences known as 'contigs' and extended contigs called 'scaffolds.'

Hybrid assembly: Raw sequencing reads from second generation (short read), and third generation (long read) technologies are used to make larger contiguous sequences like contigs and scaffolds is called a hybrid assembly.

k mer: *k*-mer is a subset of a sequence length of k

Genome Annotation: Demarcation of a gene or protein coding sequences, and other genetic features such as tRNA, and rRNA in a raw DNA sequence of genome.

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats, which are the indication of a bacterial defense system.

Insertion sequence: Insertion sequence is a short DNA sequence flanked by inverted repeats and act as a transposable element.

				Results		
	09-0549-Day-1-ETA-SA-pink_S1_L001_R1_001	25-1555-W2S1-ETA-SA_S5_L001_R1_001	25-1105-Day-7-ETA-SA_S4_L001_R1_001	09-0939-W3S1-ETA-SA-pink_S2_L001_R1_001	24-2394-W3S1-ETA-SA_S3_L001_R1_001	
	saureus:ST737	saureus:ST582	saureus:ST15	saureus:ST25	saureus:ST111	
	100%	100%	100%	100%	100%	
	100%	100%	100%	100%	100%	
	100%	100%	100%	100%	100%	
	100%	100%	100%	100%	100%	
	100%	100%	100%	100%	100%	
	100%	100%	100%	99%	100%	
	09-0549-Day-1-ETA-SA-pink_S1_L001_R1_001	25-1555-W2S1-ETA-SA_S5_L001_R1_001	25-1105-Day-7-ETA-SA_S4_L001_R1_001	09-0939-W3S1-ETA-SA-pink_S2_L001_R1_001	24-2394-W3S1-ETA-SA_S3_L001_R1_001	
	100%	-	-	-	-	
	09-0549-Day-1-ETA-SA-pink_S1_L001_R1_001	25-1555-W2S1-ETA-SA_S5_L001_R1_001	25-1105-Day-7-ETA-SA_S4_L001_R1_001	09-0939-W3S1-ETA-SA-pink_S2_L001_R1_001	24-2394-W3S1-ETA-SA_S3_L001_R1_001	
	99%	-	-	99%	99%	
	98%	100%	100%	99%	99%	
	100%	-	-	99%	100%	
	99%	-	-	99%	99%	
	99%	-	-	100%	100%	
	99%	-	-	100%	100%	

Supplementary Figure 1

Supplementary Figure 2

Supplementary Figure 3

Supplementary Table 1

Study_id	Tax id	Scientific name
PRJEB11543	1639	Listeria monocytogenes
PRJEB13576	562	Escherichia coli
PRJEB13885	562	Escherichia coli
PRJEB14086	562	Escherichia coli
PRJEB14641	562	Escherichia coli
PRJEB18587	28901	Salmonella enterica
PRJEB18587	562	Escherichia coli
PRJEB18618	108619	Salmonella enterica subsp. enterica serovar Newport
PRJEB18618	149385	Salmonella enterica subsp. enterica serovar Hadar
PRJEB18618	149539	Salmonella enterica subsp. enterica serovar Enteritidis
PRJEB18618	179997	Salmonella enterica subsp. enterica serovar Havana
PRJEB18618	192954	Salmonella enterica subsp. enterica serovar Mbandaka
PRJEB18618	192955	Salmonella enterica subsp. enterica serovar Kentucky
PRJEB18618	28901	Salmonella enterica
PRJEB18618	48409	Salmonella enterica subsp. enterica serovar Virchow
PRJEB18618	54388	Salmonella enterica subsp. enterica serovar Paratyphi A
PRJEB18618	562	Escherichia coli
PRJEB18618	57045	Salmonella enterica subsp. enterica serovar Paratyphi B
PRJEB18618	595	Salmonella enterica subsp. enterica serovar Infantis

PRJEB18618	611	Salmonella enterica subsp. enterica serovar Heidelberg
PRJEB18618	90371	Salmonella enterica subsp. enterica serovar Typhimurium
PRJEB21546	28901	Salmonella enterica
PRJEB21546	562	Escherichia coli
PRJEB21631	28901	Salmonella enterica
PRJEB22091	562	Escherichia coli
PRJEB23082	28901	Salmonella enterica
PRJEB27555	149539	Salmonella enterica subsp. enterica serovar Enteritidis
PRJEB27556	149539	Salmonella enterica subsp. enterica serovar Enteritidis
PRJEB27557	149539	Salmonella enterica subsp. enterica serovar Enteritidis
PRJEB27558	149539	Salmonella enterica subsp. enterica serovar Enteritidis
PRJEB27559	149539	Salmonella enterica subsp. enterica serovar Enteritidis
PRJEB27560	149539	Salmonella enterica subsp. enterica serovar Enteritidis

Methods S1

To run the image of Bacpipe with the GUI, the following steps can be followed:

1 - Install Docker for your operating system. (https://docs.docker.com/install/)

2 - Pull the Bacpipe main image (takes 15-25 minutes to download and install all dependencies)

https://hub.docker.com/r/mahmed/bacpipe

docker pull mahmed/bacpipe

Run X window System (X11) to allow the visualization of the Bacpipe GUI run by docker.

For mac

```
IP=$(ifconfig en0 | grep inet | awk '$1=="inet" {print $2}')
If you are connected via wifi, use en1 instead of en0
xhost + $IP
docker run -it -e DISPLAY=$IP:0 -v /tmp/.X11-unix:/tmp/.X11-unix \
-v <local_data_folder>:<container_data_folder>    mahmed/bacpipe python ./Pipeline.py unix
```

```
#   -e connects the container X11 with the local machine
#   -v maps the required folders from the local machine to the docker container.

#   Install XQuartz from https://www.xquartz.org/
#   For more information: https://sourabhbajaj.com/blog/2017/02/07/gui-applications-docker-mac/
```

For unix

```
docker run -it --hostname "YOUR_HOST_ID" --net=host -e DISPLAY=${DISPLAY} \
-v ${HOME}/.Xauthority:/root/.Xauthority -v <local_data_folder>:<container_data_folder> \
mahmed/bacpipe python ./Pipeline.py unix
```

```
#   --hostname add you unix machine name (using: hostnamectl command)
#   -e connects the container X11 with the local machine
#   -v maps the required folders from the local machine to the docker container.

#   For more information: http://wangkejie.me/2018/01/08/remote-gui-app-in-docker/
#   This was tested with remote accessing of a unix server from windows machine using Putty, \
    X11 needs to be allowed in Putty's setting "under SSH"
```

For Windows

```
docker run -it --rm -e DISPLAY="YOUR_IP_ADDRESS:0.0" -v
<local_data_folder>:<container_data_folder> \
mahmed/bacpipe python ./Pipeline.py unix
```

```
#   -e connects the container X11 with the local machine, you need to put your IP address (using
ipconfig)
#   -v maps the required folders from the local machine to the docker container.
```

BacPipe software can be downloaded from the release section here
(https://github.com/wholeGenomeSequencingAnalysisPipeline/BacPipe/releases)

Supplemental information

Supplementary Figure 1: Screenshot of summary of results in BacPipe. Related to Figure 2.

Supplementary Figure 2: BacPipe analysis discovery and retrieval from ENA web browser (https://www.ebi.ac.uk/ena/) and BacPipe results processed by the SELECTA framework are automatically submitted to the public archives. Related to Figure 4.

Supplementary Figure 3: BacPipe analysis discovery and retrieval from the COMPARE pathogen portal https://www.ebi.ac.uk/ena/pathogens/. Related to Figure 4.

Supplementary Table 1: List of bacterial genomes analysed with BacPipe in the EMBL-EBI SELECTA framework and submitted to the European Nucleotide Archive for subsequent discovery and retrieval from the Pathogen Portal (https://www.ebi.ac.uk/ena/pathogens/) and the ENA browser (https://www.ebi.ac.uk/ena/). Related to Figure 4

Data S1: Detailed overview of results generated in BacPipe from the five prior published WGS datasets utilized for validation (29-33). Related to Figure 5, Figure 6, Figure 7, Figure 8 and Figure 9.

Methods S1: Steps to install and run the BacPipe using docker image. Related to Figure 2.