BMC Bioinformatics

RESEARCH

Open Access

# CamPype: an open-source workflow for automated bacterial whole-genome sequencing analysis focused on *Campylobacter*

Irene Ortega-Sanz[1], José A. Barbero-Aparicio[2*], Antonio Canepa-Oneto[2], Jordi Rovira[1] and Beatriz Melero[1*]

*Correspondence:
jabarbero@ubu.es;
bmelero@ubu.es

[1] Department of Biotechnology and Food Science, University of Burgos, 09006 Burgos, Spain
[2] Department of Computer Science, University of Burgos, 09006 Burgos, Spain

## Abstract

**Background:** The rapid expansion of Whole-Genome Sequencing has revolutionized the fields of clinical and food microbiology. However, its implementation as a routine laboratory technique remains challenging due to the growth of data at a faster rate than can be effectively analyzed and critical gaps in bioinformatics knowledge.

**Results:** To address both issues, CamPype was developed as a new bioinformatics workflow for the genomics analysis of sequencing data of bacteria, especially *Campylobacter*, which is the main cause of gastroenteritis worldwide making a negative impact on the economy of the public health systems. CamPype allows fully customization of stages to run and tools to use, including read quality control filtering, read contamination, reads extension and assembly, bacterial typing, genome annotation, searching for antibiotic resistance genes, virulence genes and plasmids, pangenome construction and identification of nucleotide variants. All results are processed and resumed in an interactive HTML report for best data visualization and interpretation.

**Conclusions:** The minimal user intervention of CamPype makes of this workflow an attractive resource for microbiology laboratories with no expertise in bioinformatics as a first line method for bacterial typing and epidemiological analyses, that would help to reduce the costs of disease outbreaks, or for comparative genomic analyses. CamPype is publicly available at https://github.com/JoseBarbero/CamPype.

**Keywords:** Pipeline, Comparative genomics, Genome analysis, Bacterial typing, Genome annotation, Virulence genes, Antimicrobial resistance genes

## Background

Since the Human Genome Project was completed in 2003 [1], Whole-Genome Sequencing (WGS) costs are substantially decreasing over time, which has led to the emergence of new sequencing technologies that empower Next-Generation Sequencing (NGS) based on Sequencing by Synthesis (SBS), Sequencing by Oligo Ligation Detection (SOLiD), Single-Molecule Real-Time (SMRT) sequencing and nanopore-based DNA sequencing [2, 3]. Among them, Illumina sequencing remains one of the most prevalent sequencing technologies providing high accuracy and coverage with low error rates

(< 1%), compared to Pacific Biosciences or Oxford Nanopore technologies, that can afford much longer read lengths but with higher error rates and lower accuracy [4, 5].

The development of WGS has revolutionized microbiology research practices by replacing many traditional time-consuming and labor-intensive techniques [6]. Genome sequences can be obtained in a matter of hours, compared to the days or weeks required for the conventional laboratory methods and tests for completion, including Pulse-Field Gel Electrophoresis (PFGE), serotyping and phenotypic tests [7, 8]. In clinical microbiology, patient diagnosis time has been significantly reduced providing wider diagnostics repertoire. Current applications of WGS in this field include clinical identification from primary samples, infection control actions, antimicrobial stewardship, outbreak detection and intervention and pathogen discovery [9]. In food safety, EFSA (European Food Safety Authority) and FDA (Food and Drug Administration) are applying WGS of foodborne pathogens for microbial risk assessments and regulatory purposes [10, 11]. Even more, WGS is being widely used to study the microbial ecology of food products and environments along the food supply chain [12].

The implementation of WGS in clinical and food microbiology laboratories has led to the establishment of large public databases comprising thousands of genomes available [13]. The vast amounts of data produced by NGS require advanced bioinformatics skills for efficient WGS analysis, which normally are not acquired by researchers. This is one of the main bottlenecks for every microbiology laboratory in the application of WGS as a routine laboratory technique [14]. To overcome this obstacle, bioinformatics workflows are constantly being developed for the automatic analysis of genome sequences and many of them are designed for researchers without bioinformatics expertise, such as TORMES [15], BacPipe [16], ASA$^3$P [17] and Bactopia [18]. However, none is specially intended in the genera *Campylobacter*, that is the main cause of gastroenteritis worldwide [19]. Most campylobacteriosis cases (> 90%) are caused by *Campylobacter jejuni,* while *Campylobacter coli* is responsible for almost 10%. These bacteria are ubiquitous and live in the intestinal tract of poultry, pigs and cattle, but they may also be found in the feces [20]. Their genome is relatively small with ~ 1.6–1.8 Mbp length and a G + C content around 30–32% and encodes a rich inventory of virulence genes and antibiotic resistance markers responsible for their pathogenicity [21]. Moreover, campylobacteriosis is estimated to cost the EU public health systems around 2.4 billion euros per year [22]. Thus, an automated workflow for *Campylobacter* spp. would accelerate epidemiological studies through the different sequencing-based typing methods that have arisen since the first *Campylobacter* genome was published in 2000 [23], such as ribosomal Multilocus Sequence Typing (rMLST), core-genome MLST (cgMLST) or whole-genome MLST (wgMLST), that ultimately would help to reduce the costs of campylobacteriosis outbreaks.

In this work, we present CamPype, an open-source workflow for the WGS analysis of paired-end Illumina reads from *C. jejuni* and *C. coli*, although any other bacterial genus can be analyzed as well. CamPype includes a fully customizable configuration, leading to the specific results the researchers want and saving time running steps they do not need. The entire workflow can be run using one single command, making it easy to use for researchers that are not familiar with the command line. Also, CamPype provides conda environments (https://docs.anaconda.com/) with Bioconda packages [24] for all the

dependencies it needs, avoiding incompatibilities between them and making the installation as easy as possible. Finally, CamPype integrates a graphical HTML report that includes the results of every tool in the workflow shown in a more illustrative manner, allowing the researchers to access the results of the analysis easily and at a simple glance.

## Implementation

### CamPype analysis workflow

The CamPype workflow comprises three main stages (read quality control, genome assembly and genome characterization) that include several processes conducted by different tools. CamPype can take raw reads or assembled genomes in contigs as inputs. Instructions to set up the input files and workflow configuration are addressed in the CamPype repository (https://github.com/JoseBarbero/CamPype). Users can skip certain processes and adjust the configuration of parameters and databases from among the different options included for each stage in the *campype_config.py* file. An overview of the structure of CamPype is summarized in Fig. 1.

#### *Read quality control*

Previously to CamPype, a sequencing data analysis can be performed with FastQC (S. Andrews, http://www.bioinformatics.babraham.ac.uk/projects/fastqc) to assess the quality of raw reads and optimize the step of read quality control. For fast visualization of this analysis, MultiQC [25] combines these results into a single interactive HTML. Then, sequencing adaptors can be trimmed from raw sequencing reads through Trimmomatic [26] by using the *trim_adaptors* option. CamPype includes all possible Illumina adaptor sequences in the file indicated in the option *adapters_reference_file*, although users can include their own. Then, reads are quality filtered and trimmed by PRINSEQ [27] according to specific parameters. These are the minimum read length (*min_len*), minimum read quality (*min_qual_mean*), quality threshold score from the 3'-end to trim

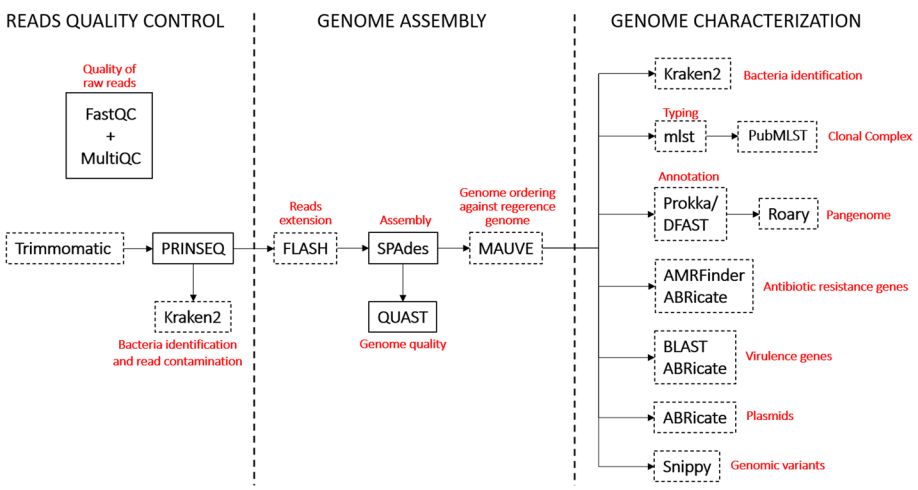


**Fig. 1** Summary of the CamPype analysis workflow. Evaluation of sequencing raw data can be performed independently and previously to CamPype. Bacteria identification is performed on the filtered fastq reads when raw reads are provided or after genome assembly when contigs are used. Software or databases are indicated in boxes. Discontinuous boxes indicate tools that users can deactivate

sequence by (*trim_qual_right*) and sliding window size (*trim_qual_window*). The reads that pass the quality control can then be used for bacterial identification and read contamination by Kraken2 [28] using the *species_identification* option.

#### *Genome assembly*

The reads that pass the quality control can be extended using FLASH [29] (*merge_reads*) and merged and unmerged reads are further de novo assembled using SPAdes [30]. The assembly mode and k-mer size(s) to be used can be selected using the *mode* and *k* options, respectively. Quality assembly of genomes is evaluated with QUAST [31] and contigs below the minimum length provided by *min_contig_len* are discarded. Resulted contigs are ordered using progressiveMauve [32] against a reference genome when specified in the options included under the *reference_genome* block.

#### *Genome characterization*

Draft genomes (ordered or not) can be further characterized through different tools, which can be selected or disabled by using the corresponding options. These include software for taxonomic classification, Multi-Locus Sequence Typing (MLST), genome annotation, detection of antibiotic resistance genes, virulence genes and plasmids, pangenome construction and identification of SNPs. For bacterial identification, Kraken2 [28] assigns taxonomic labels to draft genomes (*species_identification*) when assembled genomes are used as inputs. For subtyping purposes, MLST is performed through mlst (T. Seemann, https://github.com/tseemann/mlst) using the *run_mlst* option, and Clonal Complexes (CCs) are assigned with the *Campylobacter jejuni/coli* PubMLST scheme [33] using the *include_cc* option. Prokka [34] or DFAST [35] can be used to annotate genomes by using the *annotator* option, although this stage can be disabled by using the *run_annotation* option. A reference genome annotation in GenBank format to first annotate from can be used in Prokka [34] through the *reference_annotation* option. Keeping the raw product annotation (*rawproduct*) in Prokka [34] is highly encouraged to reduce number of hypothetical proteins. DFAST [35] includes pseudo/frameshifted gene prediction and conserved domain search. The "gff" files generated are used by Roary [36] to construct the pangenome based on the presence/absence of predicted genes. Pangenome summary figures are created using the *roary_plots.py* script by Marco Galardini (https://github.com/sanger-pathogens/Roary/blob/master/contrib/roary_plots/roary_plots.py) with minor modifications to show isolate labels and CCs (when possible) in the presence/absence accessory genome tree. Paralogs split can be disabled by the option *split_paralogs* and minimum percentage of identity for blastp can be selected by using the *minid* option. Moreover, pangenome analysis can be skipped by using the *run_pangenome* tool. Antibiotic resistance genes can be searched using protein alignments with AMRFinderPlus [37] against the NCBI Bacterial Antimicrobial Resistance Reference Gene Database (BioProject PRJNA313047), that will also identify resistance-associated point mutations only for *Campylobacter* spp., or/and using nucleotide alignments with ABRicate (T. Seemann, https://github.com/tseemann/abricate) against any of the databases provided by this tool (*antimicrobial_resistance_databases*), such as ARG-ANNOT [38], CARD [39], MEGARes [40], the NCBI Bacterial Antimicrobial Resistance Reference Gene Database (BioProject PRJNA313047) or/and ResFinder

[41]. Antibiotic resistance genes searching can be skipped using the *run_antimicrobial_resistance_genes_prediction* option and specific tools can be selected in the *antimicrobial_resistance_genes_predictor_tool* option. Draft genomes can be also screened for virulence genes using tBLASTn against an in-house database (*proteins_reference_file*), or/and BLASTn with ABRicate (T. Seemann, https://github.com/tseemann/abricate) against any of the databases provided by this tool (*virulence_factors_databases*), such as the Virulence Factors Database (VFDB) [42]. Users are encouraged to increase the size of the in-house *Campylobacter* spp. database provided with more sequences of interest or create a new one for other species, while checking the databases available in ABRicate at its repository (T. Seemann, https://github.com/tseemann/abricate). Activation of *soft_masking* is highly encouraged to find initial matches when using tBLASTn. Virulence genes search can be skipped using the *run_virulence_genes_prediction* option and specific tools can be selected in the *virulence_genes_predictor_tool* option. Minimum identity (*minid*) and coverage (*mincov*) can be selected within each tool for considering an antibiotic resistance gen and virulence gen as present. Plasmids are searched using BLASTn and ABRicate (T. Seemann, https://github.com/tseemann/abricate) against the PlasmidFinder database [43], although the analysis can be disabled by using the *run_plasmid_prediction* option. Genetic variants identification is performed through snippy (T. Seemann, https://github.com/tseemann/snippy) using the reference genome indicated in the *file* option below the *reference genome* options, as mentioned before.

Last, a summary HTML report is generated to resume the results of CamPype and can be displayed on any web browser. The report is generated in R environment (https://www.R-project.org/) using the following R packages: ape [44], complexheatmap [45], dplyr (https://CRAN.R-project.org/package=dplyr), DT (https://CRAN.R-project.org/package=DT), ggplot2 [46], ggtree [47], pander (https://CRAN.R-project.org/package=pander), plotly [48], rjson (https://CRAN.R-project.org/package=rjson), rmarkdown (https://rmarkdown.rstudio.com) and tidyverse [49]. The report includes data summary, interactive tables and figures that can be copied or downloaded. An example of an analysis report can be found at https://josebarbero.github.io/CamPype/example_report/CamPype_Report_long_first_case_study.html.

The results of CamPype are stored in specific directories for each stage and tool, with separate folders for each isolate, and include log files for analysis tracking and results standardization across different users together with files that compare the results across all analyzed samples. The location and name of the CamPype output directory can be set using the options *output_directory* and *custom_output_name*, although by default date and time of execution will be added to the directory name for managing analyses. Moreover, CamPype allows the use of multiple threads to accelerate the analysis (*n_threads*).

#### Hardware and software setup

The CamPype workflow was developed using a combination of python v3.7.8, GNU bash v5.0.17 (https://www.gnu.org/software/bash/) and R v4.1.3. CamPype is freely available at https://github.com/JoseBarbero/CamPype with a detailed instruction manual for its installation and use on any UNIX operating system. The CamPype workflow, including all required tools and dependencies, can be automatically installed using the conda environment provided. The execution of CamPype requires enough storage space. It is

recommended to have available at least three times the size of the input data for a successful complete execution when raw reads are taken as inputs and one or two GB of free space in hard disk when contigs are taken as inputs (although the genomic variant calling also requires ~ 250 MB per genome). For the study case reported here, we used a high computational capability (28 CPU cores and 64 GB RAM), even though CamPype can be run in any standard computer.

### Validation of CamPype's functionality: two case studies

#### *Analysis of Campylobacter jejuni and Campylobacter coli strains (input: raw reads)*

Ten previously published and WGS analyzed (raw reads) *C. jejuni* (5) and *C. coli* (5) strains isolated from faeces of *Bos taurus* and *Ovis aries* [50] were used to test CamPype workflow. DNA was extracted using the NZY Microbial gDNA Isolation kit (NZYtech) and sequenced using Illumina NovaSeq6000 with the NEBNext Ultra™ II FS DNA Library Prep Kit (Illumina, San Diego, CA, USA). Raw reads deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession numbers SAMN17214749 (strain C0430), SAMN17214753 (strain C0455), SAMN17214754 (strain C0459), SAMN17214757 (strain C0538), SAMN17214765 (strain C0551), SAMN17214771 (strain C0561), SAMN17214781 (strain C0582), SAMN17214797 (strain C0642), SAMN17214804 (strain C0663) and SAMN17214806 strain (strain C0669), were directly analyzed using CamPype with default configuration.

#### *Analysis of Escherichia coli genomes (input: contigs)*

A total of 44 assembled genomes of *Escherichia coli* randomly selected from the RefSeq database were used as input for CamPype: GCF_003017915.1 (strain 2014C-3051), GCF_003018035.1 (strain 2015C-4944), GCF_003018055.1 (strain 2013C-3252), GCF_003018135.1 (strain 2014C-3050), GCF_003018315.1 (strain 2013C-3513), GCF_003018455.1 (strain 97–3250), GCF_003018555.1 (strain 2013C-4225), GCF_003018575.1 (strain 2013C-4538), GCF_003018795.1 (strain 2012C-4606), GCF_003018895.1 (strain 2014C-3057), GCF_003019175.1 (strain 2013C-4187), GCF_004010675.1 (strain 2010C-3347), GCF_004010715.1 (strain 08–3914), GCF_025995195.1 (strain F690), GCF_025995255.1 (strain F765), GCF_025995315.1 (strain H52_982342), GCF_025995355.1 (strain 8_140198), GCF_025995415.1 (strain 26_141088), GCF_025995475.1 (strain 27_141091), GCF_025995535.1 (strain 53_142304), GCF_025995615.1 (strain 57_142493), GCF_025995675.1 (strain 61_150228), GCF_025995735.1 (strain 93_161312), GCF_025995895.1 (strain CEC96047), GCF_025996315.1 (strain CEC13091), GCF_025996495.1 (strain CEC08123), GCF_025996555.1 (strain CEC03102), GCF_025996675.1 (strain CEC13002), GCF_025996735.1 (strain CEC13004), GCF_027925505.1 (strain 2313), GCF_027925565.1 (strain EH031), GCF_027925625.1 (strain H19), GCF_027925685.1 (strain 20–1), GCF_027925745.1 (strain EH2252), GCF_027925765.1 (strain 98E11), GCF_027925785.1 (strain NIID080884), GCF_027925805.1 (strain PV0838), GCF_027925825.1 (strain 10,153), GCF_027925845.1 (strain 02E060), GCF_008926165.1 (strain ERL06-2442), GCF_005221885.1 (strain 143), GCF_008931135.1 (strain ERL04-3476), GCF_005221505.1 (strain 150) and GCF_008926185.1 (strain ERL05-1306). The default configuration of CamPype was modified as follows. The genome and annotation

of *Escherichia coli* strain K-12 from NCBI (NZ_CP047127) were used as reference (*reference_genome*), the *assembled_genomes* option was set to True, the *include_cc* option was set to False, ABRicate was used for virulence genes screening (*virulence_genes_predictor_tool*), and variant calling was set to True (*run_variant_calling*).

## Results

Here, the analysis of ten *C. jejuni* and *C. coli* isolate sequences is reported to validate CamPype workflow. The analysis took 5.4 h using 28 CPUs and generated a result directory of 17.2 GB (from 9 GB of compressed input data). The results of the raw reads quality control can be found in https://josebarbero.github.io/CamPype/example_report/multiqc_report_first_case_study.html, and the report with the summarized results generated by CamPype can be visualized in https://josebarbero.github.io/CamPype/example_report/CamPype_Report_long_first_case_study.html. A total of 13.0 M ± 2.5 M reads per sample were directly submitted to CamPype and reduced to 12.2 M ± 2.3 M reads per sample by the quality control stage; i.e., 97% of reads survived overall and 30% were then merged (Additional file 1). The assembly yielded 1.6–1.7 Mbp-long draft genomes fragmented into 10 to 41 contigs, corresponding to an average coverage of 1102 ± 207X with mean N50 of 242 ± 103 kbp and overall GC content of 30.4% in *C. jejuni* and 31.4% in *C. coli*. MLST revealed that isolates belonged to 7 defined Sequence Types (STs) (ST-21, ST-53, ST-441, ST-827, ST-1055, ST-3769 and ST-6775) that were grouped into Clonal Complexes (CCs) CC21 (*C. jejuni*) and CC828 (*C. coli*). Most isolates (100% *C. jejuni* and 60% *C. coli*) harbored a $bla_{OXA}$ gen and *tet(O)*, conferring resistance to beta-lactams and tetracyclines, respectively. *C. jejuni* strains harbored $bla_{OXA\text{-}193}$ or $bla_{OXA\text{-}611}$, whereas *C. coli* strains C0551, C0561 and C0663 harbored $bla_{OXA\text{-}489}$ (Fig. 2A). No antibiotic resistance gen was found in *C. coli* strain C0430, while *C. coli* strain C0538 presented only the *tet(O)* gen. Moreover, resistance to aminoglycosides (*aadE* or *aadE-Cc*) was only found in *C. coli* (60%). The efflux systems CmeABC and CmeDEF and the *CmeR* repressor were present in all isolates (Fig. 2B). The point mutation *gyrA* p.T86I conferring resistance to quinolones was present in 80% of both species, and the point mutation *rpsL* p.K43R conferring resistance to streptomycin was only found in *C. jejuni* strain C0642, while the 50S rRNA L22-A103V point mutation was only found in *C. coli* (80%) (Fig. 2C). TBLASTn against an in-house database of 76 sequences was used for virulence genes searching and 52 to 57 virulence genes were found among all isolates (Fig. 3). Differences among the ten isolates were found for the following ten genes: *capA, cdtA, cdtC, cfrB, cheY, cst-III, flaA, flaB, htrB* and *wlaN*. The genes *virB11*, *ggt*, *cgtB*, *cst-II* and the 13 genes of the Type VI Secretion System (T6SS) were not found in any of the isolates. No plasmids were found in any of the isolates. A total of 1657–1806 Coding DNA Sequences (CDS) were annotated among all isolates (Additional file 1) and were further grouped in 3575 gen clusters in the pangenome, of which 314 genes (9%) were present in all isolates (Fig. 4).

CamPype reproduced the results included in the publication of Ocejo et al. [50], that were MLST, antibiotic resistance determinants and plasmids screening in the ten assembled *Campylobacter* spp. genomes in 31 to 63 contigs of such publication.

Besides, the analysis of the 44 genomes of *E. coli* took 5.5 h and produced a directory of 13.6 GB (from 74.7 MB of compressed input data), of which 10.4 GB constituted

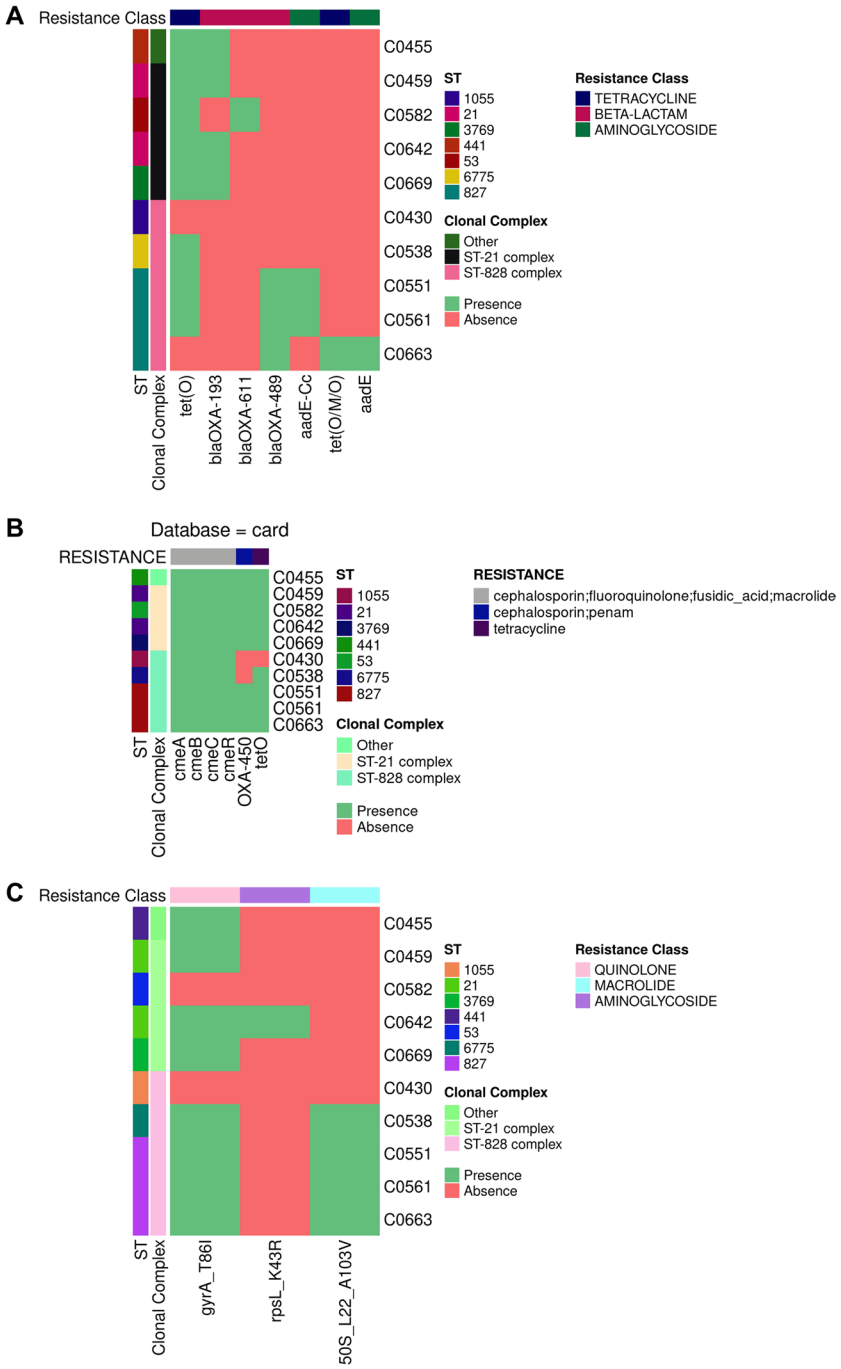


**Fig. 2** Antibiotic resistance markers identified in the *Campylobacter jejuni* and *Campylobacter coli* isolates included in the first case study. Prevalence of antibiotic resistance genes were determined with the NCBI database and AMRFinder **A**, and the CARD database and ABRicate **B**. Point mutations conferring antibiotic resistance **C** were determined with the NCBI database and AMRFinder

the genomic variants calling directory. The CamPype's HTML report for the analysis of the *E. coli* genomes can be found in https://josebarbero.github.io/CamPype/example_report/CamPype_Report_short_second_case_study.

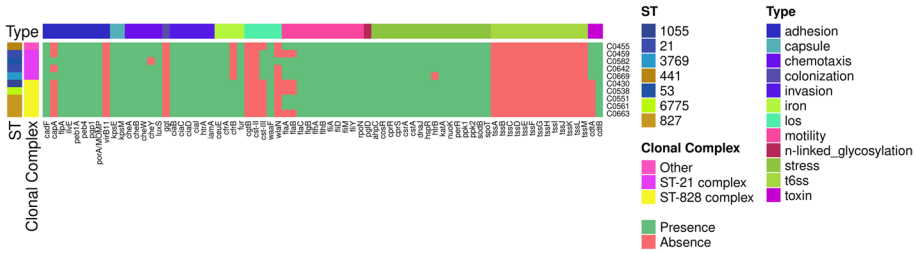


**Fig. 3** Virulence genes identified in the *Campylobacter jejuni* and *Campylobacter coli* isolates included in the first case study. Prevalence of the 76 genes comprising the inhouse database of virulence factors was evaluated with tBLASTn. For each isolate, the Sequence Type (ST) and Clonal Complex (CC) is indicated. For each gene, the virulence factor category is indicated


**Fig. 4** Gene presence/absence analysis of the *Campylobacter jejuni* and *Campylobacter coli* isolates included in the first case study. Roary was used to create the pangenome and the binary presence/absence of accessory genes was used to construct the tree. Genes (columns) coloured in grey are present in each isolate (rows), whereas genes coloured in white are absent in each isolate

## Discussion

Advances in NGS has transformed the fields of clinical and food microbiology [51, 52]. The impact is such that WGS is now routinely applied as the reference standard for infection control and epidemiology and pathogen typing [53]. WGS allows the most detailed characterization possible of bacteria to date by enabling a resolution unattainable compared to conventional laboratory typing methods with much higher level of certainty [54]. However, the large data sets generated from sequencing technologies require advanced bioinformatics training to properly use the tools and interpret the results obtained [55]. Even so, automated workflows are a rapid solution for microbiologists to allow fast and efficient analysis of data [56]. Here, the robustness of CamPype to handle *Campylobacter* WGS reads obtained from Illumina paired-end sequencing technologies is demonstrated through two different scenarios. Ten previously published *C. jejuni* and *C. coli* genomes were analyzed from the sequencing raw data using CamPype in a single command and produced same results to that of the multi-stage analyses included in the publication of Ocejo et al. [50], and even with reduced number of assembled contigs. Additional data not reported in the aforementioned study was also generated through CamPype to complement the WGS analysis, including extended statistics of reads, assembly and annotation, detection of virulence genes, and pangenome construction,

all of which were showed in an attractive HTML report. In addition, CamPype behaved efficiently for assembled genomes of different species, demonstrating the successful performance of this workflow for processing varying amounts of genomic sequencing data from diverse origin. For that second scenario, the default configuration was properly adjusted to analyze 44 previously published *E. coli* genomes using contigs as input and results were accurately reported for each genome, including bacterial typing (MLST), assembly analysis and genome annotation, searching for antibiotic resistance genes, virulence genes and plasmids, pangenome construction and identification of nucleotide variants against *E. coli* str. K-12 as reference genome. The most outstanding and promising tools hitherto for WGS are available for the users to include in the analysis, and their parameters can also be adjusted to meet their preferences. CamPype integrates various alternatives to identify antibiotic resistance genes and virulence genes since there is no single standardized and open-access database for antimicrobial resistance targets or virulence factors identification, so that the supplementary use of sequence databases generates the most complete results possible [57]. The combination of diverse data sources with different records is an excellent strategy to get partial but complementary information [58]. This is the starting point toward advancing in precision medicine for effective target therapies, as more information becomes available through the use of WGS approaches [59]. Moreover, certain analysis can be skipped to generate results in shorter times providing faster turnaround times, which has the advantage of favoring therapeutic decision-making as well. Additionally, the output of CamPype can easily be used to study epidemiological outbreaks through phylogenetic analyses of genomic variants. CamPype can handle either raw reads or assembled contigs, giving great flexibility for users and broadening its application not only for clinical diagnostic and food safety laboratories, but also towards epidemiology and comparative genomics studies.

CamPype is specially developed for *C. jejuni* and *C. coli* as they are the main responsible of gastroenteritis in humans with a frequency of about 3–4 times higher than in *Salmonella* or *E. coli* [60]. The possibility of grouping *C. jejuni* and *C. coli* Sequence Types into Clonal Complexes while providing a specific virulence genes database of this genus were not found in any of the existing microbial analysis pipelines to date, such as TORMES [15], BacPipe [16], ASA$^3$P [17] and Bactopia [18]. Moreover, opposite to these currently available pipelines, CamPype offers the option to evaluate the quality of sequencing data for optimal read quality filtering. Nonetheless, isolates from any genera and origin can be analyzed as well using CamPype. The routine use of WGS as a primary prevention is an economic favorable priority for the control of foodborne infection and other serious hospital-associated infections [61]. A mathematical simulation modelling study highlighted the direct hospital cost savings and outbreaks sizes reduction of using WGS compared to standard medical care practices [62]. The web-like report generated in CamPype provides a quick insight into antibiotic resistance targets and virulence genes facilitating a faster and accurate response in time-critical situations with lower healthcare costs [53]. Thus, CamPype would definitely help in *Campylobacter* infection control actions to minimize adverse patient outcome and in outbreak investigation. Besides, the workflow has been already used for the characterization of *Campylobacter jejuni*-associated with perimyocarditis [63] and also for comparative genomics analysis of hundreds of *Campylobacter* spp. isolated from Spain (*in prep.*).

CamPype was developed with the needs of microbiology laboratories in mind and obstacles that restrict the use of WGS for clinical/public health microbiology investigations [56]. Along with being user-friendly and customizable, CamPype is a comprehensive workflow that is capable of performing a very detailed automated analysis of large numbers of genomes in a single process without previous specific knowledge and bioinformatics skills, by using simple commands. The open-source nature allows collaborative coding between users and developers with the intention to fulfill users' needs and be improved through as many suggestions as proposed by the community to make CamPype an outstanding workflow. The analysis is performed locally, which means the user is the owner of the data in every moment without needing internet connection. Other free resources, such as Galaxy [64] and PATRIC [65], integrate attractive and interactive user interfaces, but require a fast and consistent internet connection for importing data to the server that can lead to privacy and security issues with data protection policies varying between countries [66]. Moreover, the performance of analysis depends not only on the number of raw reads but also on the hardware of the computer used, with reduced execution time when more CPU processors are available, whereas web-served based analyses take indeterminate execution times that vary on the server workloads, which is unreliable for patient care emergency situations [67].

## Conclusions

Implementing WGS in clinical and food microbiology laboratories has led to an increase in the amount of raw data and genomes publicly available. However, the use of WGS as a routine method is unfeasible without the application of bioinformatics resources and remains a challenge due to the required specific skill set. CamPype is a reliable solution for integration WGS into routinely use and overcome these barriers because it enables easy and automated analysis of large genome datasets, providing a quick visualization of results that facilitates data interpretation.

**Abbreviations**

| | |
|---|---|
| CC | Clonal complex |
| CDS | Coding DNA sequence |
| cgMLST | Core-genome MLST |
| EFSA | European Food Safety Authority |
| ENA | European Nucleotide Archive |
| FDA | Food and Drug Administration |
| GC | Guanine-cytosine |
| MLST | Multilocus sequence typing |
| NGS | Next-generation sequencing |
| PFGE | Pulse-field gel electrophoresis |
| rMLST | Ribosomal MLST |
| SBS | Sequencing by synthesis |
| SMRT | Single-molecule real-time |
| SNP | Single-nucleotide polymorphism |
| SOLiD | Sequencing by oligo ligation detection |
| ST | Sequence type |
| T6SS | Type VI secretion system |
| VFDB | Virulence factors database |
| wgMLST | Whole-genome MLST |
| WGS | Whole-genome sequencing |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05414-w.

**Additional file 1:** Genomic characteristics of the Campylobacter jejuni and Campylobacter coli isolates included in the first case study.

**Acknowledgements**
The authors thank the anonymous reviewers for their constructive comments.

**Availability and requirements**
*Project name*: CamPype; *Project home page*: https://github.com/JoseBarbero/CamPype; *Operating system(s)*: Linux; *Programming language*: Python, R, bash; *Other requirements*: Conda; *License*: GPL-3.0; *Any restrictions to use by non-academics*: None.

**Author contributions**
IOS was responsible for conceptualization and design of the study, participated in coding, carried out the two case studies and drafted the manuscript. JABA was responsible for conceptualization of study and creation of the workflow, validation of its functionality, and participated in the review and editing of the original draft manuscript. ACO was responsible for conceptualization of study and participated in coding, and review and editing of the original draft manuscript. JR was responsible for conceptualization of study, co-directed the study and participated in the design of the study and revision of the final manuscript. BM was responsible for conceptualization of study, managed the project and coordinated the study, acquired resources and funding, and revised the final manuscript. All authors read and approved the final manuscript.

**Funding**
The project leading to these results received funding from "La Caixa" Foundation and Caja Burgos Foundation, under agreement LCF/PR/PR18/51130007. Irene Ortega-Sanz received a predoctoral grant from the Junta of Castile and León, cofinanced by the Ministry of Education of the Government of Castile and León and the European Social Fund. José A. Barbero-Aparicio was founded through a pre-doctoral grant from the University of Burgos. The funders played no roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

**Availability of data and materials**
The datasets supporting the conclusions of this article are included within the article (and its additional file). The CamPype software is available at https://github.com/JoseBarbero/CamPype and an example of the analysis report is provided in https://josebarbero.github.io/CamPype/example_report/CamPype_Report_long_first_case_study.html. The raw reads and assembled genomes used to test CamPype can be found in https://zenodo.org/record/7999130.

## Declarations

**Ethics approval and consent to participate**
Not applicable

**Consent for publication**
Not applicable

**Competing interests**
The authors declare that they have no competing interests.

Received: 20 March 2023 Accepted: 14 July 2023
Published online: 20 July 2023

**References**

1. Collins FS, Green ED, Guttmacher AE, Guyer MS. US National Human Genome Research Institute. A vision for the future of genomics research. Nature. 2003;422(6934):835–47.
2. Slatko BE, Gardner AF, Ausubel FM. Overview of next-generation sequencing technologies. Curr Protoc Mol Biol. 2018;122(1):e59.
3. Furlani B, Kouter K, Rozman D, Videtič PA. Sequencing of nucleic acids: from the first human genome to next generation sequencing in {COVID}-19 pandemic. Acta Chim Slov. 2021;68(2):268–78.
4. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333–51.
5. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: an overview. Hum Immunol. 2021;82(11):801–11.
6. Applications of Clinical Microbial Next-Generation Sequencing. Report on an American Academy of Microbiology Colloquium held in Washington, DC, in April 2015. Washington: American Society for Microbiology; 2016.
7. Nouws S, Bogaerts B, Verhaegen B, Denayer S, Crombé F, De Rauw K, et al. The benefits of whole genome sequencing for foodborne outbreak investigation from the perspective of a national reference laboratory in a smaller country. Foods. 2020;9(8):1030.

8. Dookie N, Khan A, Padayatchi N, Naidoo K. Application of next generation sequencing for diagnosis and clinical management of drug-resistant tuberculosis: updates on recent developments in the field. Front Microbiol. 2022;13:775030.
9. Motro Y, Moran-Gilad J. Next-generation sequencing applications in clinical bacteriology. Biomol Detect Quantif. 2017;14:1–6.
10. Van Hoorde K, Butler F. Use of next-generation sequencing in microbial risk assessment. EFSA J. 2018;16(Suppl 1):e16086.
11. Brown E, Dessai U, McGarry S, Gerner-Smidt P. Use of whole-genome sequencing for food safety and public health in the United States. Foodborne Pathog Dis. 2019;16(7):441–50.
12. García-Sánchez L, Melero B, Jaime I, Hänninen ML, Rossi M, Rovira J. *Campylobacter jejuni* survival in a poultry processing plant environment. Food Microbiol. 2017;65:185–92.
13. Carrillo CD, Blais BW. Whole-genome sequence datasets: a powerful resource for the food microbiology laboratory toolbox. Front Sustain Food Syst. 2021;5:754988.
14. Afolayan AO, Bernal JF, Gayeta JM, Masim ML, Shamanna V, Abrudan M, et al. Overcoming data bottlenecks in genomic pathogen surveillance. Clin Infect Dis. 2021;73(Suppl 4):S267–74.
15. Quijada NM, Rodríguez-Lázaro D, Eiros JM, Hernández M. TORMES: an automated pipeline for whole bacterial genome analysis. Bioinformatics. 2019;35(21):4207–12.
16. Xavier BB, Mysara M, Bolzan M, Ribeiro-Gonçalves B, Alako BTF, Harrison P, et al. BacPipe: a rapid, user-friendly whole-genome sequencing pipeline for clinical diagnostic bacteriology. iScience. 2020;23(1):100769.
17. Schwengers O, Hoek A, Fritzenwanker M, Falgenhauer L, Hain T, Chakraborty T, et al. ASA3P: an automatic and scalable pipeline for the assembly, annotation and higher-level analysis of closely related bacterial isolates. PLoS Comput Biol. 2020;16(3):e1007134.
18. Petit RA 3rd, Read TD. Bactopia: a flexible pipeline for complete analysis of bacterial genomes. mSystems. 2020;5(4):e00190-e220.
19. European Food Safety Authority; European Centre for Disease Prevention and Control. The European Union One Health 2021 Zoonoses Report. EFSA J. 2022;20(12):e07666.
20. Kaakoush NO, Castaño-Rodríguez N, Mitchell HM, Man SM. Global epidemiology of campylobacter infection. Clin Microbiol Rev. 2015;28(3):687–720.
21. Bunduruș IA, Balta I, Ștef L, Ahmadi M, Peț I, McCleery D, et al. Overview of virulence and antibiotic resistance in campylobacter spp. Livestock Isolates Antibiotics. 2023;12(2):402.
22. EFSA. Scientific opinion on *Campylobacter* in broiler meat production: control options and performance objectives and/or targets at different stages of the food chain. EFSA J. 2011;9(4):2105.
23. Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D, et al. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature. 2000;403(6770):665–8.
24. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018;15(7):475–6.
25. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047–8.
26. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
27. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27(6):863–4.
28. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20(1):257.
29. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011;27(21):2957–63.
30. 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.
31. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5.
32. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS ONE. 2010;5(6):e11147.
33. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. 2018;3:124.
34. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–9.
35. Tanizawa Y, Fujisawa T, Nakamura Y. DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. Bioinformatics. 2018;34(6):1037–9.
36. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691–3.
37. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, et al. AMRFinderPlus and the reference gene catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021;11(1):12728.
38. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother. 2014;58(1):212–20.
39. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, et al. The comprehensive antibiotic resistance database. Antimicrob Agents Chemother. 2013;57(7):3348–57.
40. Lakin SM, Dean C, Noyes NR, Dettenwanger A, Ross AS, Doster E, et al. MEGARes: an antimicrobial resistance database for high throughput sequencing. Nucleic Acids Res. 2017;45(D1):D574–80.
41. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012;67(11):2640–4.
42. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, et al. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 2005;33:D325–8.

43. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother. 2014;58(7):3895–903.
44. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35(3):526–8.
45. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32(18):2847–9.
46. Wickham H. ggplot2: elegant graphics for data analysis. 1st ed. New York: Springer; 2016.
47. Yu G, Smith DK, Zhu H, Guan Y, Lam TT. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8(1):28–36.
48. Sievert C. Interactive web-based data visualization with R, plotly, and shiny. 1st ed. Chapman and Hall/CRC Florida; 2020.
49. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R. Welcome to the tidyverse. J Open Source Softw. 2019;4(43):1686.
50. Ocejo M, Oporto B, Lavín JL, Hurtado A. Whole genome-based characterisation of antimicrobial resistance and genetic diversity in *Campylobacter jejuni* and *Campylobacter coli* from ruminants. Sci Rep. 2021;11(1):8998.
51. Jagadeesan B, Gerner-Smidt P, Allard MW, Leuillet S, Winkler A, Xiao Y, et al. The use of next generation sequencing for improving food safety: translation into practice. Food Microbiol. 2019;79:96–115.
52. Mitchell SL, Simner PJ. Next-generation sequencing in clinical microbiology: are we there yet? Clin Lab Med. 2019;39(3):405–18.
53. Bogaerts B, Winand R, Van Braekel J, Hoffman S, Roosens NHC, De Keersmaecker SCJ, et al. Evaluation of WGS performance for bacterial pathogen characterization with the Illumina technology optimized for time-critical situations. Microb Genom. 2021;7(11):000699.
54. Pightling AW, Pettengill JB, Luo Y, Baugher JD, Rand H, Strain E. Interpreting whole-genome sequence analyses of foodborne bacteria for regulatory applications and outbreak investigations. Front Microbiol. 2018;10(9):1482.
55. Barretto C, Rincón C, Portmann AC, Ngom-Bru C. Whole genome sequencing applied to pathogen source tracking in food industry: key considerations for robust bioinformatics data analysis and reliable results interpretation. Genes. 2021;12(2):275.
56. Wyres KL, Conway TC, Garg S, Queiroz C, Reumann M, Holt K, et al. WGS Analysis and Interpretation in Clinical and Public Health Microbiology Laboratories: what are the requirements and how do existing tools compare? Pathogens. 2014;3(2):437–58.
57. Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C, et al. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. Clin Microbiol Infect. 2017;23(1):2–22.
58. Bazeley P. Complementary analysis of varied data sources. In: Seaman J, editor. Integrating analyses in mixed methods research. SAGE Publications Ltd.; 2018. p. 91–125.
59. Suwinski P, Ong C, Ling MHT, Poh YM, Khan AM, Ong HS. Advancing personalized medicine through the application of whole exome sequencing and big data analytics. Front Genet. 2019;10:49.
60. Facciolà A, Riso R, Avventuroso E, Visalli G, Delia SA, Laganà P. *Campylobacter:* from microbiology to prevention. J Prev Med Hyg. 2017;58(2):E79–92.
61. Gordon LG, Elliott TM, Forde B, Mitchell B, Russo PL, Paterson DL, et al. Budget impact analysis of routinely using whole-genomic sequencing of six multidrug-resistant bacterial pathogens in Queensland, Australia. BMJ Open. 2021;11(2):e041968.
62. Lee XJ, Elliott TM, Harris PNA, Douglas J, Henderson B, Watson C, et al. Clinical and economic outcomes of genome sequencing availability on containing a hospital outbreak of resistant *Escherichia coli* in Australia. Value Health. 2020;23(8):994–1002.
63. Ortega-Sanz I, García M, Bocigas C, Megías G, Melero B, Rovira J. Genomic characterization of *Campylobacter jejuni* associated with perimyocarditis: a family case report. Foodborne Pathog Dis. 2023;20:8.
64. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018;46(W1):W537–44.
65. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 2014;42:D581–91.
66. Molnár-Gábor F, Korbel JO. Genomic data sharing in Europe is stumbling-could a code of conduct prevent its fall? EMBO Mol Med. 2020;12(3):e11421.
67. Miller NA, Farrow EG, Gibson M, Willig LK, Twist G, Yoo B, et al. A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. Genome Med. 2015;7:100.


## Publisher's Note