scientific reports

Check for updates

OPEN

Pipeline validation for the identification of antimicrobial-resistant genes in carbapenem-resistant *Klebsiella pneumoniae*

Andressa de Almeida Vieira[1], Bruna Candia Piccoli[1], Thaís Regina y Castro[1], Bruna Campestrini Casarin[1], Luiza Funck Tessele[1], Roberta Cristina Ruedas Martins[2], Alexandre Vargas Schwarzbold[3] & Priscila de Arruda Trindade[1]✉

Antimicrobial-resistant *Klebsiella pneumoniae* is a global threat to healthcare and an important cause of nosocomial infections. Antimicrobial resistance causes prolonged treatment periods, high mortality rates, and economic impacts. Whole Genome Sequencing (WGS) has been used in laboratory diagnosis, but there is limited evidence about pipeline validation to parse generated data. Thus, the present study aimed to validate a bioinformatics pipeline for the identification of antimicrobial resistance genes from carbapenem-resistant *K. pneumoniae* WGS. Sequences were obtained from a publicly available database, trimmed, de novo assembled, mapped to the *K. pneumoniae* reference genome, and annotated. Contigs were submitted to different tools for bacterial (Kraken2 and SpeciesFinder) and antimicrobial resistance gene identification (ResFinder and ABRicate). We analyzed 201 K. *pneumoniae* genomes. In the bacterial identification by Kraken2, all samples were correctly identified, and in SpeciesFinder, 92.54% were correctly identified as *K. pneumoniae*, 6.96% erroneously as *Pseudomonas aeruginosa*, and 0.5% erroneously as *Citrobacter freundii*. ResFinder found a greater number of antimicrobial resistance genes than ABRicate; however, many were identified more than once in the same sample. All tools presented 100% repeatability and reproducibility and > 75% performance in other metrics. Kraken2 was more assertive in recognizing bacterial species, and SpeciesFinder may need improvements.

Widespread use of antimicrobials has generated microorganisms' selective pressure[1,2]. The emergence and spread of antimicrobial-resistant bacteria become a threat to public health[3]. One of the most worrying pathogens is *Klebsiella pneumoniae*. This microorganism belongs to the *Enterobacterales* order and *Enterobacteriaceae* family, which are composed of gram-negative encapsulated, non-spore-forming, and rod-shaped bacteria[4–6]. In human hosts, it can constitute the normal enteric microbiota. It can also infect the respiratory system, endocardium, surgical site wounds, reach the bloodstream, and cause sepsis[7]. Neonates, the elderly, and immunocompromised hospitalized patients present a worse prognosis[8,9]. It is capable of causing serious community-acquired infections especially due to hypervirulent strains[7].

β-lactam antimicrobials (carbapenems, cephalosporins, and monobactams) present a β-lactam ring in their molecular structure, which inhibits the transpeptidases. Consequently, they inhibit cell wall synthesis, leading to bacterial death[10]. *K. pneumoniae*'s accessory genome acquired genes encoding β-lactamases as a resistance mechanism to hydrolyze the β-lactam ring[7,11]. The first reported gene was Carbapenem-hydrolyzing beta-lactamase KPC (bla_{KPC}) in 1996[12,13]. bla_{KPC} became stable in the accessory genome of some *K. pneumoniae*

[1]Laboratório de Biologia Molecular e Bioinformática Aplicada à Microbiologia Clínica, Programa de Pós-Graduação em Ciências Farmacêuticas, Universidade Federal de Santa Maria, Santa Maria 97105-900, Brazil. [2]Laboratório de Parasitologia Médica (LIM-46), Departamento de Doenças Infecciosas e Parasitárias, Instituto de Medicina Tropical da Universidade de São Paulo, Faculdade de Medicina da Universidade de São Paulo, São Paulo 01246-903, Brazil. [3]Departamento de Clínica Médica, Universidade Federal de Santa Maria, Rio Grande do Sul, Brazil. ✉email: priscila.trindade@ufsm.br

strains[7,11,12]. Since then, other genes encoding β-lactamases have been identified, such as oxacillinases (bla_{OXA}), and metallo-β-lactamases (bla_{NDM}, bla_{IMP}, and bla_{VIM})[7,11,14].

Antimicrobial resistance is complex, multifactorial, and causes prolonged treatment periods, high mortality rates, and economic impacts[1,15]. Available molecular tests are unable to detect emerging genetic characteristics of pathogens. To ensure successful treatment, recovery, and patient safety, the identification and characterization of microorganisms causing infections are essential[16,17]. Whole Genome Sequencing (WGS) has the ability to replace traditional molecular techniques as it provides benefits in terms of higher resolution, speed, reduced cost, and numerous additional information such as species, strain type, resistance, and virulence profiles[18,19]. Analyzing and interpreting genome-scale data pose challenges due to the volume and complexity of the data[20]. Thus, the objective of this study is to validate a bioinformatics pipeline for in silico analysis of WGS of carbapenem-resistant *K. pneumoniae* isolates to produce standardized data that will enable interlaboratory comparisons.

Results

We analyzed 201 K. *pneumoniae* genomes to validate the pipeline for predicting antimicrobial resistance genes, especially carbapenems. For this purpose, we took advantage of seven BioProjects with carbapenem-resistant *K. pneumoniae* SRAs available on the National Center for Biotechnology Information (NCBI) platform. *K. pneumoniae* strain ATCC 35657 (PRJNA279657), lacking carbapenem-resistance genes, was used as a negative control. We trimmed, de novo assembled, ordered, and annotated the SRAs. De novo assembly and mapping quality metrics are listed in Table 1. A high percentage of genome coverage (mean of 93.8%) and depth (mean of 125.5x) were obtained.

Kraken2 and SpeciesFinder tools were used for bacterial identification. For Kraken2, all samples (100%) were identified correctly, and for SpeciesFinder, 92.54% (186) were identified as *K. pneumoniae*, 6.96% (14) as *Pseudomonas aeruginosa*, and 0.5% (1) as *Citrobacter freundii* (Fig. 1 and Table S1). Both tools obtained 100% reproducibility and repeatability (Table 2). The other validation metrics could not be calculated due to the lack of adequate definitions for the analysis.

ResFinder and ABRicate tools were used for identifying antimicrobial resistance genes. We evaluated 273 antimicrobial resistance genes, among them twelve are specific to carbapenems, *i.e.*, bla_{KPC-2}, bla_{KPC-3}, bla_{NDM-1}, bla_{NDM-7}, bla_{OXA-48}, $bla_{OXA-162}$, $bla_{OXA-181}$, $bla_{OXA-232}$, $bla_{OXA-245}$, bla_{VIM-1}, bla_{VIM-19}, and bla_{VIM-27} (Table S2). ResFinder identified a higher number of antimicrobial resistance genes, corresponding to 23.27 ± 0.56, compared to 15.85 ± 0.39 (ABRicate) (Fig. 2A and Table S3). Of these, 55% were found by both tools. It is important to note

Metric	1	2	3	4	5	6	7
Reads	1,667,674.91	1,091,649.50	1,593,738.60	1,719,599.30	1,385,878.40	9,052,915.20	5,313,384
Coverage (%)	94.6	94.5	94.8	96.2	92.9	92.6	90.9
Depth (x)	67.9	45.6	61.9	75.9	48.4	327.5	251.2
Contig number	142	99.6	110	96.5	79.3	47.3	17
Length of the longest contig (nt)	482,279.78	670,750.10	663,956.60	623,432.60	674,034.90	1,155,637	2,052,661
Total length (nt)	5,634,605.77	5,707,949.80	5,679,227.60	5,732,894.70	5,613,962.30	5,648,880	5,349,908
GC content (%)	57.1	57	57	57	57.1	57.1	57.4
N50	165,681.46	241,806	223,144	222,369	298,608	453,211	847,522

Table 1. De novo assembly quality metrics. Results were shown as mean.

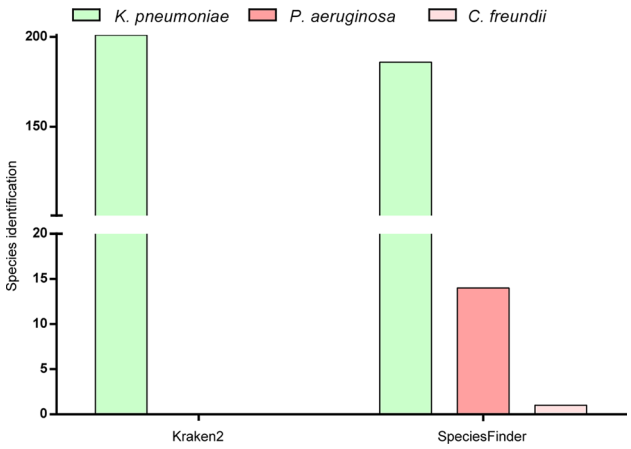

Figure 1. Bacteria identified by Kraken and SpeciesFinder databases.

BioProject	Kraken2		SpeciesFinder	
	Repeatability (%)	Reproducibility	Repeatability (%)	Reproducibility
1	100	–	100	–
2	100	100%	100	100%
3	100	–	100	–
4	100	–	100	–
5	100	–	100	–
6	100	–	100	–
7	100	–	100	–

Table 2. Repeatability and reproducibility of bacterial identification from Kraken2 and SpeciesFinder tools.

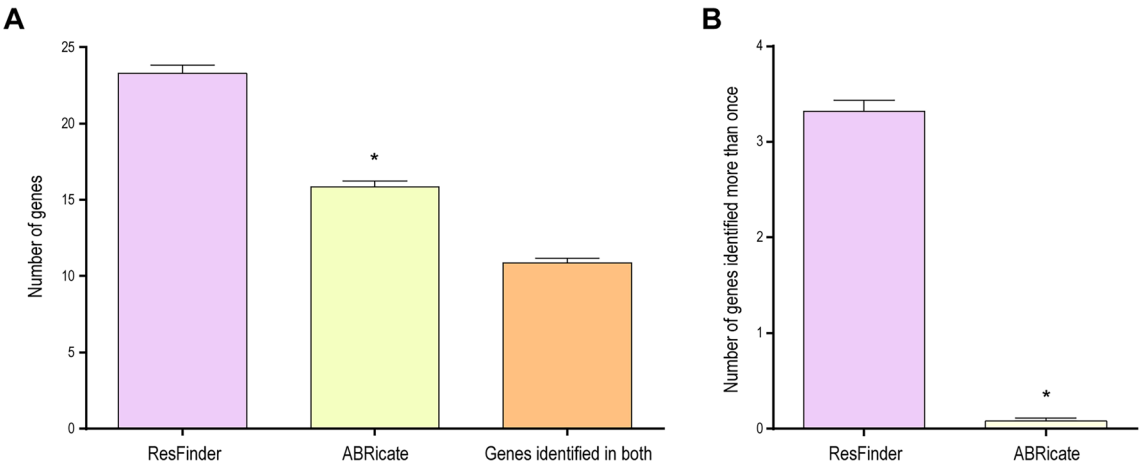

Figure 2. Resistance genes found by ResFinder and ABRicate databases in 201 SRAs (**A**). Same gene was indicated more than once in each sample (**B**). Results were presented as mean ± SEM and analyzed by Student's t test. * means statistical difference from the ResFinder group ($p \leq 0.05$).

that, in all samples, ResFinder indicated up to 6 × the same gene (Fig. 2B). ABRicate only showed duplicated genes in eight samples. Although ResFinder found a greater number of genes, this value was distorted due to gene duplication.

The genes most frequently identified by ResFinder in the 201 samples were *oqxA* and *oqxB* genes (394 times) (Fig. 3). Differently, *fosA6* gene, followed by *sul1* gene, were the genes most identified by ABRicate. Among the

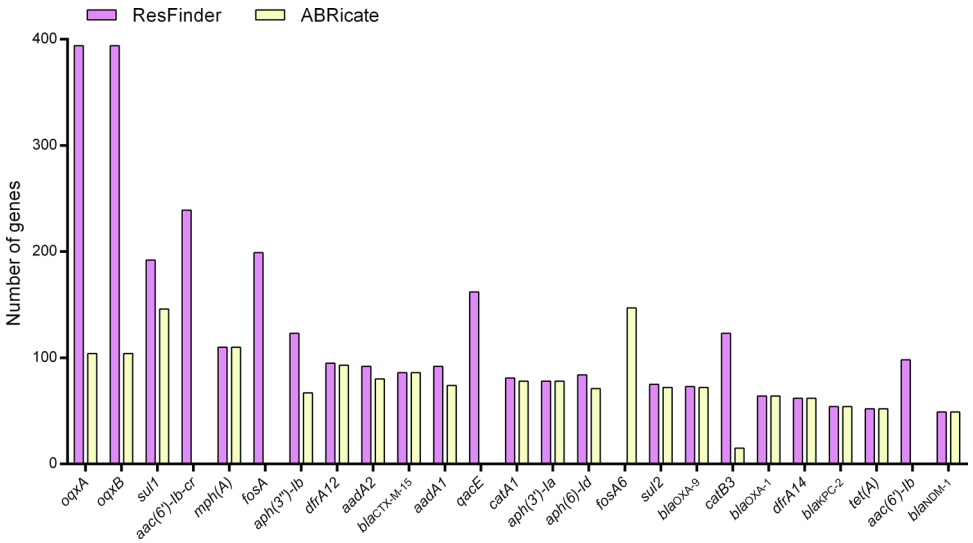

Figure 3. Twenty-five genes most frequently identified by ResFinder and ABRicate databases.

25 genes most frequently identified by the tools, *fosA6* gene was found only by ABRicate, and *aac(6')-Ib-cr*, *fosA*, *qacE* gene, and *aac(6')-Ib* gene were found only by ResFinder. We only found one carbapenem resistance gene (bla_{KPC-2}).

Carbapenem-resistant genes identified by ResFinder and ABRicate showed similar coverage and identity percentages (Fig. 4). When we consider all antimicrobial resistance genes identified, ABRicate had the highest coverage percentage [$t(7165) = 22.6$; $p < 0.0001$] and identity [$t(7165) = 3.784$; $p = 0.0002$)]. These results indicate that, probably, genes were present in the samples and were correctly identified with greater reliability by ABRicate.

Pipeline validation metrics for ABRicate and ResFinder tools, highlighting carbapenem resistance genes and all antimicrobial resistance genes, are shown in Table 3. Sequences were analyzed in triplicate on the same day to determine repeatability. Samples from BioProjects PRJNA292902/PRJNA292904, which had more than one technical replicate, were evaluated on alternate days to calculate reproducibility. Accuracy, precision, sensitivity, and specificity calculations were performed by comparing the results obtained with the reference sequence (RefSeq). ABRicate presented lower precision and sensitivity in BioProject 1 (PRJEB28660) when considering only the carbapenem resistance genes. However, when all antimicrobial resistance genes were evaluated, ResFinder showed lower percentages in 17 parameters (mainly related to accuracy, precision, sensitivity, and specificity) in five different BioProjects, compared to four parameters of ABRicate. These results indicate that ABRicate seems to be more suitable for antimicrobial resistance gene identification.

We compared the number of genes identified by the samples assembled in this study with their respective RefSeqs (Fig. 5). As expected, no carbapenem resistance gene was identified in the negative control (PRJNA279657) (Fig. 5A). A higher number of carbapenem resistance genes were found in the RefSeqs of the BioProjects PRJNA292902/PRJNA292904 and PRJNA392824 than in the samples assembled using the pipeline described in this study, as identified by both tools (Fig. 5A). Similarly, more antimicrobial resistance genes were found in the RefSeqs of the PRJEB28660 and PRJNA292902/PRJNA292904 BioProjects, as shown in Fig. 5B. These results corroborate the lower sensitivity found in these BioProjects (Table 3). Performing a manual curation, we detected that, in the RefSeq, a greater number of genes were found because the same gene (same name and accession) was identified in the sample in more than one contig; in the same contig, but in different *loci*; or in the same contig and at the same *locus*, but with different accessions. These results indicated a high number of false negatives (FN), which affected the tool sensitivities.

We additionally evaluated the influence of the default parameters of Basic Local Alignment Search Tool (BLAST) on the performance of ABRicate and ResFinder. We identified antimicrobial resistance genes using ABRicate with parameters set at 90% identity and 60% coverage (default parameters of ResFinder), and for ResFinder, we employed parameters set at 80% identity and coverage (default parameters of ABRicate) (Fig. 6). ResFinder identified a greater number of antimicrobial resistance genes compared to ABRicate under both parameter settings, considering our assembly and the RefSeq dataset. When applying the criteria of 80% sequence identity and 80% coverage, ResFinder identified a reduced number of antimicrobial resistance genes in samples assembled using the pipeline described in this study [$t(399) = 3.286$; $p = 0.0011$]. However, the results were similar when using the RefSeq dataset ($p > 0.05$). ABRicate exhibited a statistically similar antimicrobial resistance gene number under both BLAST parameter settings.

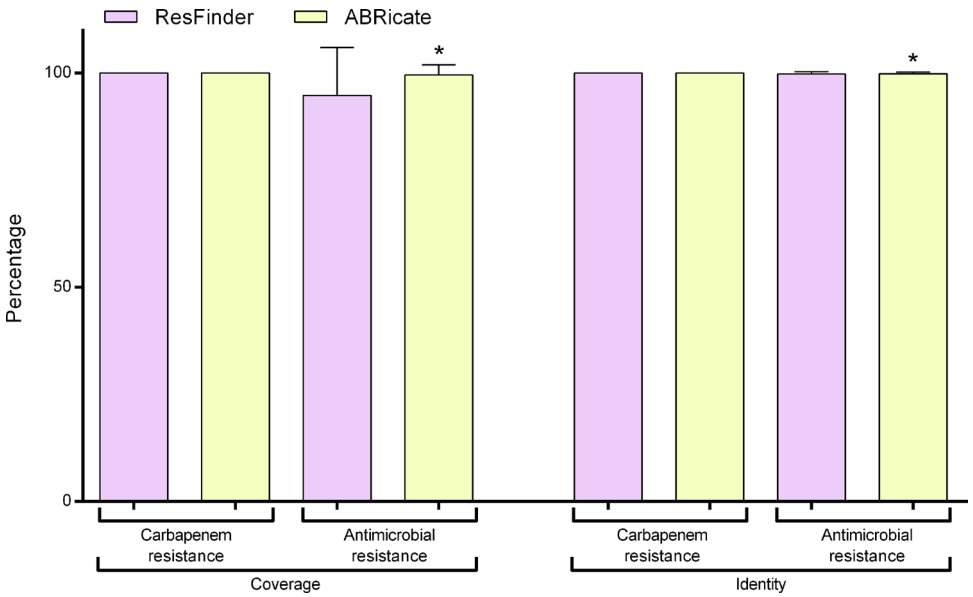

Figure 4. Percent coverage and identity of antimicrobial resistance genes found by ResFinder and ABRicate databases. Results were presented as mean ± SEM and analyzed by Student's t test. * means statistical difference from the ResFinder group ($p \leq 0.05$).

Tool	Gene	Metric	1	2	3	4	5	6	7
ABRicate	Carbapenem resistance genes	Repeatability	100%	100%	100%	100%	100%	100%	100%
		Reproducibility	–	100%	–	–	–	–	–
		Accuracy	99.99%	99.99%	100%	100%	100%	99.98%	100%
		Precision	90.19%	95.45%	100%	100%	100%	100%	0
		Sensitivity	92.00%	82.89%	100%	100%	100%	55.55%	0
		Specificity	99.99%	99.99%	100%	100%	100%	100%	100%
	All antimicrobial resistance genes	Repeatability	100%	100%	100%	100%	100%	100%	100%
		Reproducibility	–	44.92%	–	–	–	–	–
		Accuracy	99.88%	99.88%	99.97%	99.94%	99.97%	99.85%	100%
		Precision	93.79%	97.76%	96.85%	95.95%	97.39%	100%	100%
		Sensitivity	74.28%	79.59%	96.63%	90.97%	97.90%	72.36%	100%
		Specificity	99.98%	99.99%	99.98%	99.98%	99.98%	100%	100%
ResFinder	Carbapenem resistance genes	Repeatability	100%	100%	100%	100%	100%	100%	100%
		Reproducibility	–	100%	–	–	–	–	–
		Accuracy	99.99%	99.99%	100%	100%	100%	99.98%	100%
		Precision	91.07%	95.45%	100%	100%	100%	100%	0
		Sensitivity	92.72%	82.89%	100%	100%	100%	55.55%	0
		Specificity	99.99%	99.99%	100%	100%	100%	100%	100%
	All antimicrobial resistance genes	Repeatability	100%	100%	100%	100%	100%	100%	100%
		Reproducibility	–	36.23%	–	–	–	–	–
		Accuracy	99.80%	99.79%	99.94%	99.86%	99.97%	99.84%	100%
		Precision	90.44%	89.12%	92.14%	90.96%	97.79%	100%	100%
		Sensitivity	72.88%	79.36%	96.75%	85.93%	98.51%	76.92%	100%
		Specificity	99.95%	99.93%	99.95%	99.94%	99.98%	100%	100%

Table 3. Validation metrics of ABRicate and ResFinder tools for resistance genes.

Discussion

In this study, we validated a bioinformatics pipeline for *K. pneumoniae* identification and the prediction of antimicrobial resistance genes in sequenced samples obtained from humans infected with this pathogen. The *K. pneumoniae* genome has approximately two thousand conserved genes[11,21]. It also presents an accessory genome consisting of genes located on chromosomes and plasmids that vary among isolates. *K. pneumoniae* has, on average, five to six thousand accessory genes[11]. These genes are acquired through horizontal transfer, as evidenced by the presence of genomic islands and mobile genetic elements. Accessory genes could encode virulence factors, enzymes, and antimicrobial resistance mechanisms, potentially worsening the prognosis of infected individuals[11]. Thus, identifying the infecting microorganism and its resistance genes is crucial for patient diagnosis and treatment.

We used the pipeline validation protocol described by Bogaerts et al.[19]. The authors performed the first bioinformatics pipeline validation for microbiological sequence isolates using *Neisseria meningitidis* as a model. Traditional metrics of repeatability, reproducibility, precision, sensitivity, and specificity were evaluated, adapted for WGS data. The dataset consisted of 131 sequences, divided into two subsets: the main subset (composed of 67 samples sequenced in triplicate) and the extended subset (composed of 64 sequenced samples publicly available on NCBI). In our study, we used 201 sequenced samples. Among them, 132 were single replicates used to calculate the repeatability, and 69 comprised three or four technical replicates, considered for both repeatability and reproducibility calculations.

Due to the range of bioinformatic approaches used to manipulate the data, three stages of analysis can lead to discrepant results: (i) sequencing quality, (ii) databases, or (iii) software used. Sample quality control is critical to improving sensitivity. High coverage (at least 90%) and depth (at least 30x) are also recommended. Values below the recommended thresholds can generate false positive (FP) results[22]. To minimize erroneous results, the pipeline contains a trimming step to remove poorly sequenced nucleotides, adapters, and short reads. The remaining reads were mapped against the reference genome, resulting in > 90% coverage and 45 × depth (Table 1).

After ensuring the read quality and optimal coverage and depth values, sequences were submitted to Kraken2 and SpeciesFinder to identify their bacterial species. Both tools showed high repeatability and reproducibility. Kraken2 correctly identified all sequences. SpeciesFinder identified 92.54% of the sequences as *K. pneumoniae* and the rest, erroneously, as *Pseudomonas aeruginosa* and *Citrobacter freundii*. The bacteria *C. freundii* and *K. pneumoniae* belong to the same family (*Enterobacteriaceae*)[23]. However, *P. aeruginosa* only shares the same class[24], and it is counterintuitive that *K. pneumoniae* sequences were identified as *P. aeruginosa*. SpeciesFinder maps the contigs against the 16S rRNA sequence using the BLAST. The 16S rRNA corresponds to 0.1% of the microbial genome coding sequence[25]. We hypothesize that *P. aeruginosa* and *C. freundii* were identified in *K. pneumoniae* SRAs because mapping occurred in a small region of the genome, although the 16S rRNA is considered a

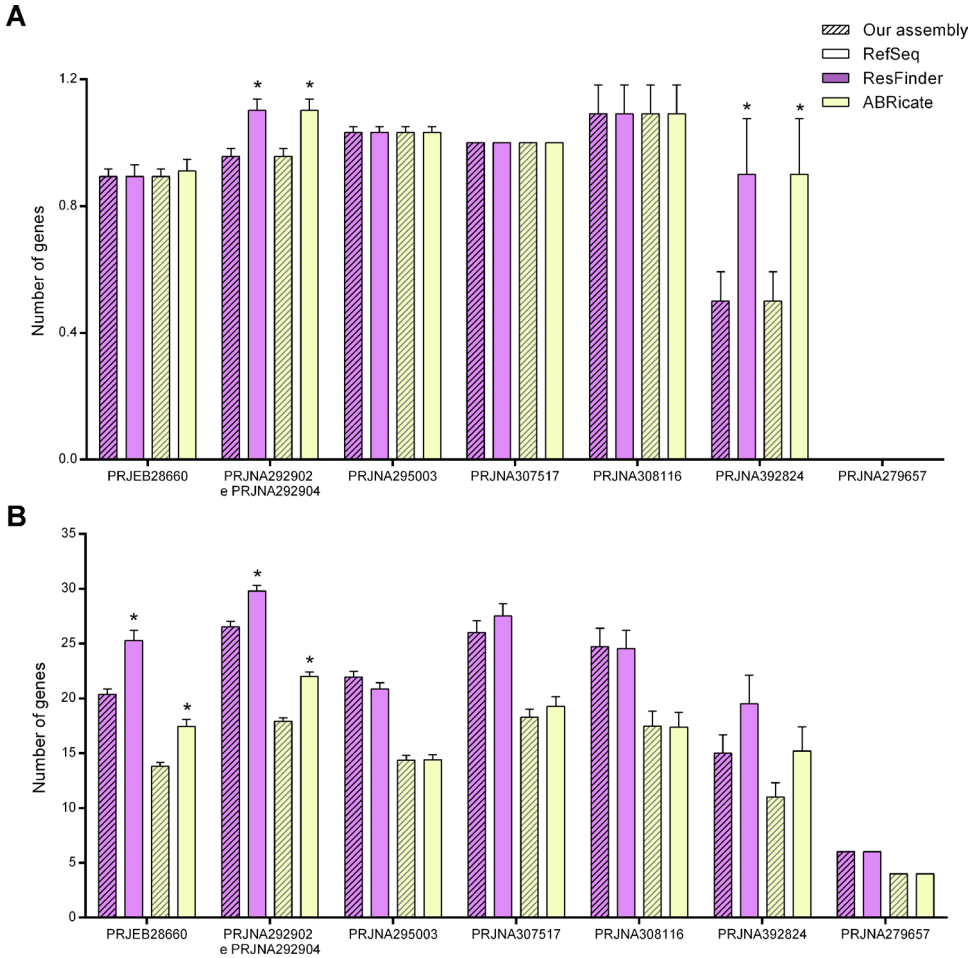

Figure 5. Resistance genes identified by ResFinder and ABRicate databases using the samples assembled using the pipeline described in this study and their RefSeq. Carbapenem resistance genes (**A**) and all antimicrobial resistance genes identified (**B**) by the databases in each bioproject. Results were presented as mean ± SEM and analyzed by Student's t test. * means statistical difference from our assembly ($p \leq 0.05$).

highly conserved gene. Kraken2 performs a comprehensive genome analysis, mapping short genomic sequences (k-mers) in genomes present in its database and comparing them to a taxonomic tree to identify the common ancestor[26,27]. This could justify Kraken2's assertiveness in identifying species.

ResFinder and ABRicate were used to identify antimicrobial resistance genes. ResFinder identified a wide range of resistance genes in the analyzed sequences; however, ResFinder provides up to six copies of the same gene (Fig. 2A,B). These tools are composed of different gene variants and/or isoforms. Thus, the high percentage of identity among the sequences (> 90%) guarantees the correct gene identification[22]. In our study, we achieved > 99.8% identity and > 94.8% genomic coverage (Fig. 4). Doyle et al.,[22], also found disagreements in the total number of genes associated with antimicrobial resistance, as well as in gene variants of pathogens resistant to carbapenems. These results show that the choice of a resistance gene identification tool can significantly impact the results.

ResFinder and ABRicate showed high repeatability and reproducibility when considering only the carbapenem resistance genes. Reproducibility was reduced to 44.92% (ABRicate) and 36.23% (ResFinder) when evaluating all antimicrobial resistance genes. Reproducibility is calculated by sequencing the same sample under different conditions. In this study, we used publicly available SRAs, some of which contained technical replicates. However, the exact sequencing conditions are not known, which is a limitation of our in silico study since we were unable to sequence the samples. The other performance metrics, including accuracy, precision, sensitivity, and specificity, were similar for both tools in the identification of carbapenem resistance genes. When we evaluated these parameters for the identification of all antimicrobial resistance genes, ABRicate showed better accuracy (mean of 97.39%) than ResFinder (mean of 93.88%). Bogaerts et al.[19] found a performance of 100% in all metrics evaluated for ResFinder and NDARO tools. The identification of other resistance genes was also done, and the metrics showed > 70% performance, except for reproducibility (36.23%).

Sensitivity presented the lowest percentages (< 55%). It is calculated by comparing the number of genes found in the RefSeq with the number found in the consensus sequences. Resistance gene identification tools (ResFinder and ABRicate) found a greater number of genes in RefSeq than in the consensus sequences assembled by our

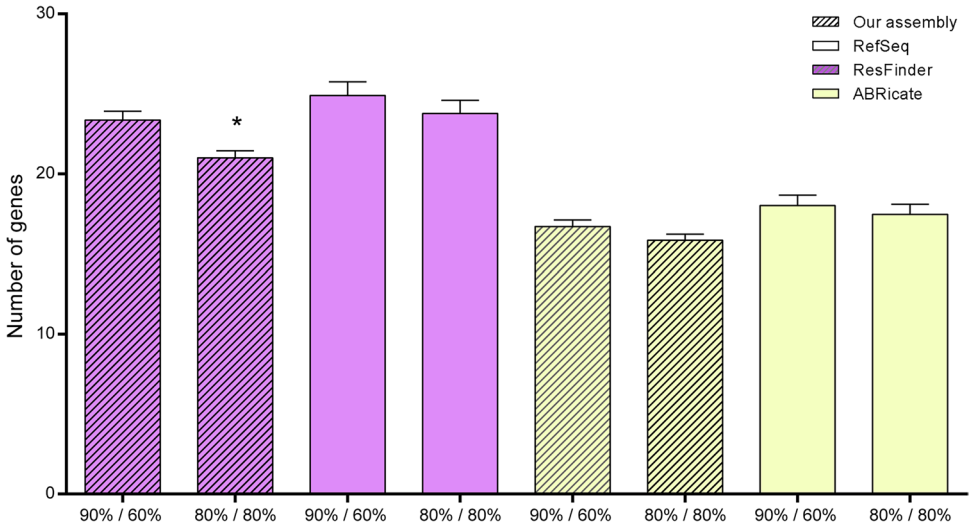

Figure 6. Resistance genes identified by ResFinder with BLAST parameters set at 80% identity and coverage (default parameters of ABRicate) and ABRicate with BLAST parameters set at 90% identity and 60% coverage (default parameters of ResFinder), using the samples assembled using the pipeline described in this study and their RefSeq. Results were presented as mean ± SEM and analyzed by Student's t test. * means statistical difference from 90% identity and 60% coverage ($p \leq 0.05$).

pipeline. After performing manual curation, we realized that this higher number was related to gene duplication. Similarly, Kozyreva et al.[28] used reference sequences from the US Food and Drug Administration (FDA)-CDC Antimicrobial Resistance (AR) Isolate Bank, previously evaluated with the ResFinder database. The authors found discrepancies in the detection of resistance genes between reference sequences and those assembled by them, leading to FP. The RefSeqs were trimmed and assembled differently from what was proposed by the pipeline, which may have influenced the identification of antimicrobial resistance genes. The difference in assembly software can alter or make it infeasible to identify a gene if it is divided into one or more contigs[29,30]. Also, the presence of duplicate genes in the tools leads to an overestimation of these genes[31]. After this manual curation, we considered that the de novo assembly proposed by our pipeline is adequate, as well as the sensitivity of the tools. It is important to notice the different BLAST default parameter settings between ABRicate and ResFinder. In both tools, default settings were employed to enhance the user-friendliness and accessibility of the pipeline, catering to operators with limited expertise in bioinformatics. Furthermore, adhering to these default parameters prevents the introduction of biases that could potentially alter diagnostic outcomes, thereby preserving the integrity of results and maintaining consistency in both intra- and inter-laboratory reproducibility.

The importance of standardized methodologies and pipelines used in WGS in microbiology laboratories is evident[28]. Therefore, the validation strategy suggested by Bogaerts et al.[19] and performed in our study can be extended to other sequencing technologies and pathogens for use in laboratory routine. Since bioinformatics expertise is one of the main challenges in WGS, it is essential to have bioinformatics professionals permanently employed in clinical laboratories to provide expert interpretation. Additionally, the generation of a centralized and standardized database, as well as computational reproducibility, is of paramount importance[19,22].

In summary, we validated a bioinformatics pipeline for *K. pneumoniae* identification and its antimicrobial resistance genes. This pipeline can be used in laboratory routine to identify the infecting microorganisms and their antimicrobial resistance mechanisms. Using this pipeline, infected patients could receive more individualized treatment, leading to a reduction in hospitalization duration and mortality rates. Kraken2, as a species identifier, proved to be more accurate, while ABRicate was more effective in identifying antimicrobial resistance genes. SpeciesFinder and ResFinder may need updates. Given the variety of bioinformatics tools and resistance determinant databases available, the validation strategy used in our study can be applied to different bioinformatic pipelines and tools to ensure standardization of intra- and inter-laboratory validation.

Methodology

Dataset. Search for carbapenem-resistant *K. pneumoniae* BioProjects was performed in NCBI database (https://www.ncbi.nlm.nih.gov/sra/). Three criteria were used to select the BioProjects: (i) to have carbapenem-resistant *K. pneumoniae* samples isolated from human hosts, (ii) to have been sequenced by Illumina MiSeq technology, and (iii) to present genome assembly as the RefSeq. Seven BioProjects (PRJEB28660, PRJNA292902, PRJNA292904, PRJNA295003, PRJNA307517, PRJNA308116, and PRJNA392824) and 201 SRA met these criteria (Table 4). In addition, a negative control sample was selected. SRAs were downloaded with the fastq-dump tool v. 2.10.9 from SRAToolkit, capable of converting SRA to fastq files.

No.	BioProject	Sample	Total no. used	Ref
1	PRJEB28660	Carbapenemase-producing *Klebsiella pneumoniae*	56	–
2	PRJNA292902 and PRJNA292904	Carbapenemase-producing *Klebsiella pneumoniae*	69	–
3	PRJNA295003	Carbapenemase-producing *Klebsiella pneumoniae*	31	32 33
4	PRJNA307517	Carbapenemase-producing *Klebsiella pneumoniae*	23	34 35
5	PRJNA308116	Carbapenemase-producing *Klebsiella pneumoniae*	11	–
6	PRJNA392824	Carbapenemase-producing *Klebsiella pneumoniae*	10	36
7	PRJNA279657	*Klebsiella pneumoniae* ATCC 35,657	1	–

Table 4. BioProjects used for pipeline validation.

Bacterial genome assembly, annotation, and species identification. Raw sequencing data were evaluated using the FastQC v0.11.9 program with default settings at the Babraham Institute, Cambridge, UK. Subsequently, the samples were subjected to trimming in Trimmomatic v0.39[37], removing adapter residues, bases with Q-score < 3 at the beginning and end of reads, and Q-score < 15 in a four-base sequence. De novo assembly of the genomes was performed using SPAdes v3.13.1 with the –careful option enabled to reduce the number of mismatches[38]. For mapping, Bowtie2 v2.3.0 was employed, utilizing the *K. pneumoniae* reference genome (NC_016845)[39]. The de novo assembly and mapping statistics were assessed through the online interface of QUAST[40] and SAMtools[41], respectively. The generated contigs were then sorted by the ABACAS v1.3.1 program, following the *K. pneumoniae* reference genome (NC_016845)[42], and subsequently annotated using Prokka v1.14.5[43] (Fig. 7).

Species identification. Species identification was performed using the Kraken tool v2.1.1[26] and SpeciesFinder 2.0[44] (Fig. 7).

Identificaction of antimicrobial resistance genes. Identification of antimicrobial resistance genes was performed using ResFinder v4.1[45] and ABRicate v1.0.1[46] under default parameters. ABRicate uses the NCBI database by default, while the BLAST tool is configured with an 80% identity and 80% coverage threshold. On the other hand, ResFinder employs the BLAST tool with parameters set at 90% identity and 60% coverage. The bioinformatics pipeline used in the study is shown in Fig. 7.

Evaluation criteria. Performance analysis, as well as pipeline validation, was performed according to Bogaerts et al.[19] with adaptations. The following metrics were evaluated: repeatability, reproducibility, accuracy,

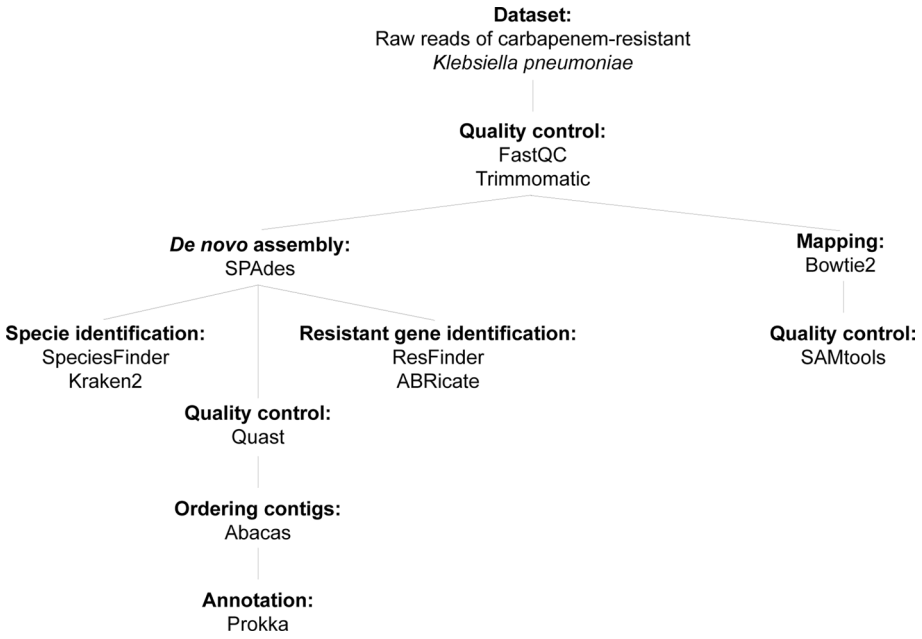

Figure 7. Bioinformatics pipeline used in the work.

Metrics	Definition	Formula for calculation
Repeatability	Concordancy based in replicates on the same run in the same assay	Repeatability = (number of intra-assay concordant repetitions)/(total number of repetitions) × 100%
Reproducibility	Concordance based on results generated by different runs for the same sample	Reproducibility = (number of concordant repetitions between assays)/(total number of repetitions) × 100%
Accuracy	Probability that the assay results are correct	Accuracy = (TP + TN)/ (TP + TN + FP + FN) × 100%
Precision	Probability that the detected results are truly positive	Precision = TP/(TP + FP) × 100%
Sensibility	Probability that the result will be correctly detected in the assay when present	Sensibilidade = TP/(TP + FN) × 100%
Specificity	Probability that a result will not be falsely detected in an assay when absent	Specificity = TN/(TN + FP) × 100%

Table 5. Parameters evaluated in the performance analysis and pipeline validation. TP = true positive; TN = true negative; FP = false positive; FN = false negative.

precision, sensitivity, and specificity (Table 5). For the repeatability calculation, the bioinformatics pipeline was run on the same day using the same dataset. For the reproducibility calculation, the PRJNA292902 and PRJNA292904 BioProjects were selected, which had more than one technical replicate. The pipeline was run on alternate days to evaluate the intra-run reproducibility. Results were considered in agreement when genes were present or absent in both runs. To evaluate accuracy, precision, sensitivity, and specificity, results were categorized as true positive (TP), false positive (FP), true negative (TN), or false negative (FN). TP indicates a gene found by our pipeline and in the reference genome; FP indicates a gene found by our pipeline but absent in the reference genome; TN indicates a gene not found by our pipeline nor in the reference genome, and FN indicates a gene absent from our pipeline but present in the reference genome (Table 5). Some metrics were not evaluated for all bioinformatic assays, as suitable definitions cannot always be found in the context of the specific analysis[19,47].

Data availability

The SRAs are available at NCBI under BioProject ID PRJEB28660, PRJNA292902 and PRJNA292904, PRJNA295003, PRJNA307517, PRJNA308116, PRJNA392824, and PRJNA279657. The SRAs used are listed in detail in Table S4.

Received: 21 March 2023; Accepted: 6 September 2023
Published online: 14 September 2023

References

1. Schürch, A. C. & Van Schaik, W. Challenges and opportunities for whole-genome sequencing–based surveillance of antibiotic resistance. *Ann. N.Y. Acad. Sci.* **1388**(1), 108–120. https://doi.org/10.1111/nyas.13310 (2017).
2. van Camp, P. J., Haslam, D. B. & Porollo, A. Prediction of antimicrobial resistance in gram-negative bacteria from whole-genome sequencing data. *Front. Microbiol.* **11**, 1–13. https://doi.org/10.3389/fmicb.2020.01013 (2020).
3. Magiorakos, A. P. *et al.* Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: An international expert proposal for interim standard definitions for acquired resistance. *Clin. Microbiol. Infect.* **18**(3), 268–281. https://doi.org/10.1111/j.1469-0691.2011.03570.x (2012).
4. Merla, C. *et al.* Description of Klebsiella spallanzanii sp. nov. and of Klebsiella pasteurii. *Front. Microbiol.* **10**, 1–9. https://doi.org/10.3389/fmicb.2019.02360 (2019).
5. Patro, L. P. P. & Rathinavelan, T. Targeting the sugary armor of Klebsiella species. *Front. Cell. Infect. Microbiol.* **9**, 1–23. https://doi.org/10.3389/fcimb.2019.00367 (2019).
6. Podschun, R. & Ullmann, U. Klebsiella spp as Nosocomial Pathogens: Epidemiology, Taxonomy, Typing Methods, and Pathogenicity Factors. *Clin. Microbiol. R* **11**(4), 589–603 (1998).
7. Hennequin, C. & Robin, F. Correlation between antimicrobial resistance and virulence in Klebsiella pneumoniae. *Eur. J. Clin. Microbiol. Infect. Dis.* **35**(3), 333–341. https://doi.org/10.1007/s10096-015-2559-7 (2016).
8. Bengoechea, J. A. & Sa Pessoa, J. Klebsiella pneumoniae infection biology: Living to counteract host defences. *FEMS Microbiol. Rev.* **43**(2), 123–144. https://doi.org/10.1093/femsre/fuy043 (2019).
9. Choby, J. E., Howard-Anderson, J. & Weiss, D. S. Hypervirulent Klebsiella pneumoniae – clinical and molecular perspectives. *J. Internal Med.* **287**(3), 283–300. https://doi.org/10.1111/joim.13007 (2020).
10. Lima, L. M. *et al.* β-lactam antibiotics: An overview from a medicinal chemistry perspective. *Eur. J. Med. Chem.* **208**, 112829. https://doi.org/10.1016/j.ejmech.2020.112829 (2020).
11. Martin, R. M. & Bachman, M. A. Colonization, Infection, and the Accessory Genome of Klebsiella pneumoniae. *Front. Cell. Infect. Microbiol.* **8**, 1–15. https://doi.org/10.3389/fcimb.2018.00004 (2018).
12. Pitout, J. D. D.,. Multiresistant Enterobacteriaceae: New threat of an old problem. *Expert Rev. Anti-Infect. Therapy* **6**(5), 657–669. https://doi.org/10.1586/14787210.6.5.657 (2008).
13. Yigit, H. *et al.* Novel Carbapenem-Hydrolyzing B-Lactamase, KPC-1, from a Carbapenem-Resistant Strain of Klebsiella pneumoniae. *Antimicrob. Agents Chemother.* **45**(4), 1151–1161. https://doi.org/10.1128/AAC.45.4.1151 (2001).
14. Lee, C. R. *et al.* Global dissemination of carbapenemase-producing Klebsiella pneumoniae: Epidemiology, genetic context, treatment options, and detection methods. *Front. Microbiol.* **7**, 1–30. https://doi.org/10.3389/fmicb.2016.00895 (2016).
15. Angers-Loustau, A. *et al.* The challenges of designing a benchmark strategy for bioinformatics pipelines in the identification of antimicrobial resistance determinants using next generation sequencing technologies. *F1000Research* **7**, 459. https://doi.org/10.12688/f1000research.14509.1 (2018).
16. Deurenberg, R. H. *et al.* Application of next generation sequencing in clinical microbiology and infection prevention. *J. Biotechnol.* **243**, 16–24. https://doi.org/10.1016/j.jbiotec.2017.03.035 (2017).
17. Mitchell, S. L. & Simner, P. J. Next-generation sequencing in clinical microbiology: Are we there yet?. *Clin. Lab. Med.* **39**(3), 405–418. https://doi.org/10.1016/j.cll.2019.05.003 (2019).

18. Besser, J. *et al.* Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin. Microbiol. Infect.* **24**(4), 335–341. https://doi.org/10.1016/j.cmi.2017.10.013 (2018).
19. Bogaerts, B. *et al.* Validation of a bioinformatics workflow for routine analysis of whole-genome sequencing data and related challenges for pathogen typing in a European national reference center: Neisseria meningitidis as a Proof-of-Concept. *Front. Microbiol.* **10**, 1–19. https://doi.org/10.3389/fmicb.2019.00362 (2019).
20. Timme, R. E. *et al.* Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. *PeerJ* **5**, 1–13. https://doi.org/10.7717/peerj.3893 (2017).
21. Holt, K. E. *et al.* Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. *Proc. Natl. Acad. Sci.* **112**(27), 3574–3581. https://doi.org/10.1073/pnas.1501049112 (2015).
22. Doyle, R. M. *et al.* Discordant bioinformatic predictions of antimicrobial resistance from whole-genome sequencing data of bacterial isolates: An inter-laboratory study. *Microbial. Genom.* **6**(2), 1–13. https://doi.org/10.1099/mgen.0.000335 (2020).
23. Liu, L. H. *et al.* Citrobacter freundii bacteremia: Risk factors of mortality and prevalence of resistance genes. *J. Microbiol. Immunol. Infect.* **51**(4), 565–572. https://doi.org/10.1016/j.jmii.2016.08.016 (2018).
24. Jackson, J. D., Kuzel, T. M. & Shafikhan, S. H. Pseudomonas aeruginosa: Infections, Animal Modeling, and Therapeutics. *Princ. Regener. Med.* **5349**(2), 191–204. https://doi.org/10.1016/B978-0-12-809880-6.00013-8 (2019).
25. Prabaa, M. S. D. *et al.* Identification of nonserotypeable Shigella spp using genome sequencing: A step forward. *Fut. Sci. OA* **3**(4), 1–11. https://doi.org/10.4155/fsoa-2017-0063 (2017).
26. Wood, D. E. & Salzberg, S. L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**(3), 1–12. https://doi.org/10.1186/gb-2014-15-3-r46 (2014).
27. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**(1), 1–13. https://doi.org/10.1186/s13059-019-1891-0 (2019).
28. Kozyreva, V. K. *et al.* Validation and implementation of clinical laboratory improvements act-compliant whole-genome sequencing in the public health microbiology laboratory. *J. Clin. Microbiol.* **55**(8), 2502–2520. https://doi.org/10.1128/JCM.00361-17 (2017).
29. Clausen, P. T. L. C. *et al.* Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. *J. Antimicrob. Chemother.* **71**(9), 2484–2488. https://doi.org/10.1093/jac/dkw184 (2016).
30. Hendriksen, R. S. *et al.* Using genomics to track global antimicrobial resistance. *Front. Public Health* **7**, 1–17. https://doi.org/10.3389/fpubh.2019.00242 (2019).
31. Papp, M. & Solymosi, N. Review and comparison of antimicrobial resistance gene databases. *Antibiotics* **11**(3), 1–12. https://doi.org/10.3390/antibiotics11030339 (2022).
32. Samuelsen, O. *et al.* Molecular and epidemiological characterization of carbapenemase- producing Enterobacteriaceae in Norway, 2007 to 2014. *PLoS ONE* **12**(11), 1–17. https://doi.org/10.1371/journal.pone.0187832 (2017).
33. Samuelsen, Ø. *et al.* Molecular characterization of VIM-producing Klebsiella pneumoniae from Scandinavia reveals genetic relatedness with international clonal complexes encoding transferable multidrug resistance. *Clin. Microbiol. Infect.* **17**(12), 1811–1816. https://doi.org/10.1111/j.1469-0691.2011.03532.x (2011).
34. Pitt, M. E. *et al.* Multifactorial chromosomal variants regulate polymyxin resistance in extensively drug-resistant Klebsiella pneumoniae. *Microbial. Genom.* **4**(3), 1. https://doi.org/10.1099/mgen.0.000158 (2018).
35. Elliott, A. G. *et al.* Complete genome sequence of Klebsiella quasipneumoniae subsp. similipneumoniae strain ATCC 700603. *Genome Announc.* **4**(3), 3–4. https://doi.org/10.1128/genomeA.00438-16 (2016).
36. Simner, P. J. *et al.* Antibiotic pressure on the acquisition and loss of antibiotic resistance genes in Klebsiella pneumoniae. *J. Antimicrob. Chemother.* **73**(7), 1796–1803. https://doi.org/10.1093/jac/dky121 (2018).
37. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170 (2014).
38. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**(5), 455–477. https://doi.org/10.1089/cmb.2012.0021 (2012).
39. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**(4), 357–359. https://doi.org/10.1038/nmeth.1923 (2012).
40. Gurevich, A. *et al.* QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086 (2013).
41. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 (2009).
42. Assefa, S. *et al.* ABACAS: Algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25**(15), 1968–1969. https://doi.org/10.1093/bioinformatics/btp347 (2009).
43. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153 (2014).
44. Larsen, M. V. *et al.* Benchmarking of methods for genomic taxonomy. *J. Clin. Microbiol.* **52**(5), 1529–1539. https://doi.org/10.1128/JCM.02981-13 (2014).
45. Zankari, E. A. *et al.* Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67**(11), 2640–2644. https://doi.org/10.1093/jac/dks26 (2012).
46. Seemann, T. ABRicate: Mass screening of contigs for antimicrobial resistance or virulence genes. https://github.com/tseemann/abricate. Acesso em: 22 março de 2019.
47. Aziz, N. *et al.* College of American pathologists laboratory standards for next-generation sequencing clinical tests. *Arch. Pathol. Lab. Med.* **139**(4), 481–493. https://doi.org/10.5858/arpa.2014-0250-CP (2015).

Acknowledgements

Authors would like to thank all the participants in this study, the Universidade Federal de Santa Maria, and the financial support of the Brazilians' development Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), grant number 88882.461702/2019-01.

Author contributions

A.A.V., R.C.R.M. and P.A.T. designed the study, A.A.V., T.R.C., B.C.C., and R.C.R.M. compiled the database, A.A.V., B.C.P., and P.A.T. analyzed the data, A.A.V., B.C.P., and L.F.T. wrote the draft manuscript, A.V.S and P.A.T. reviewed the manuscript, A.V.S and P.A.T. funding acquisition. All authors read and approved the final version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-42154-6.

Correspondence and requests for materials should be addressed to P.d.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.