Investigation of next-generation sequencing data of *Klebsiella pneumoniae* using web-based tools

Eva Brhelova,1,2 Mariya Antonova,1,2 Filip Pardy,3 Iva Kocmanova,4 Jiri Mayer,1,2,3 Zdenek Racil1,2,3 and Martina Lengerova1,2,3,*

**Abstract**

**Purpose.** Rapid identification and characterization of multidrug-resistant *Klebsiella pneumoniae* strains is necessary due to the increasing frequency of severe infections in patients. The decreasing cost of next-generation sequencing enables us to obtain a comprehensive overview of genetic information in one step. The aim of this study is to demonstrate and evaluate the utility and scope of the application of web-based databases to next-generation sequenced (NGS) data.

**Methodology.** The whole genomes of 11 clinical *Klebsiella pneumoniae* isolates were sequenced using Illumina MiSeq. Selected web-based tools were used to identify a variety of genetic characteristics, such as acquired antimicrobial resistance genes, multilocus sequence types, plasmid replicons, and identify virulence factors, such as virulence genes, cps clusters, urease-nickel clusters and efflux systems.

**Results.** Using web-based tools hosted by the Center for Genomic Epidemiology, we detected resistance to 8 main antimicrobial groups with at least 11 acquired resistance genes. The isolates were divided into eight sequence types (ST11, 23, 37, 323, 433, 495 and 562, and a new one, ST1646). All of the isolates carried replicons of large plasmids. Capsular types, virulence factors and genes coding AcrAB and OqxAB efflux pumps were detected using BIGSdb-Kp, whereas the selected virulence genes, identified in almost all of the isolates, were detected using CLC Genomic Workbench software.

**Conclusion.** Applying appropriate web-based online tools to NGS data enables the rapid extraction of comprehensive information that can be used for more efficient diagnosis and treatment of patients, while data processing is free of charge, easy and time-efficient.

**INTRODUCTION**

*Klebsiella pneumoniae* is a member of the family *Enterobacteriaceae* and is one of the most common pathogens causing community- and hospital-acquired infections, including pneumonia, bacteraemia, urinary tract infections and pyogenic liver abscesses [1–3]. Nosocomial *K. pneumoniae* isolates often display highly resistant phenotypes, with these showing resistance to drugs such as oximinocephalosporins (ESBLs), aminoglycosides, macrolides and quinolones [1], as well as rapidly increasing resistance to carbapenem [4]. Highly drug-resistant clones complicate the antibiotic treatment of patients [5]. In addition, a combination of several virulence factors resulted in hypervirulent variants of *K. pneumoniae* that affected the previously healthy population, causing community-acquired infections [6, 7].

In current routine practice, bacteria are mostly identified and characterized using well-established and standardized phenotypic and molecular typing methods. Varieties of specific methods such as pulsed-field gel electrophoresis or multilocus sequence typing are used for the further characterization and outbreak investigation of selected isolates. Unfortunately, these methods are time-consuming, laborious and expensive [8]. The time requirements, cost and affordability have been compared in previous works [9, 10].
Rapid genotyping is essential for the accurate and targeted outbreak characterization of severe life-threatening infections. As the cost of next-generation sequencing (NGS) has decreased in recent years, it has become a useful method for outbreak detection and surveillance in a single step [11]. Proof-of-concept studies have shown that NGS holds significant potential for microbial identification from primary human specimens through both targeted amplicon sequencing of rRNA genes and whole-genome sequencing (WGS) [12–14]. WGS can also act as a precursor to generate specific diagnostic tests for timely case definition [15–19].

As NGS provides a large amount of sequencing data, a rapid and automatic conversion tool is required that enables the results to be interpreted easily by health professionals, who usually have limited informatics skills [20]. For this purpose, the Center for Genomic Epidemiology provides publicly available web-based tools for the rapid handling and extraction of relevant information for identification, diagnosis and outbreak investigation (http://www.genomic epidemiology.org/). Another well-known web-based database, Institut Pasteur MLST, provides access to genotypic data for virulence-associated genes (http://bigshares.web.pasteur.fr/klebsiella/klebsiella.html).

In this study we demonstrated ways of obtaining relevant genetic information from bacteria that differ from classic molecular methods. In particular, we explored the application, utility and scope of web-based tools with respect to NGS data for rapid typing, antimicrobial resistance genes and virulence-associated gene detection of extended-spectrum beta-lactamase (ESBL)-producing *Klebsiella pneumoniae* clinical strains.

**METHODS**

**Study design and bacterial isolates**

*K. pneumoniae* isolates were collected for the purpose of screening for colonization and sepsis in a larger study in the tertiary-care hospital focused on ESBL-resistant isolates [21]. From these isolates, 11 from different patients were selected for genome sequencing in this study. The isolates were stored at −70 °C.

**Screening for ESBL and antimicrobial susceptibility testing**

All of the isolates were cultured on selective chromogenic media chromID ESBL (BioMérieux) and the ESBL producers were confirmed using the double-disk synergy test (Oxoid, Thermo Scientific, USA). An antimicrobial susceptibility test that employed the disk diffusion method was performed to determine the resistance patterns of the isolates to 13 antibiotics (Oxoid, Thermo Scientific, USA): amikacin, cefepime, ciprofloxacin, chloramphenicol, colistin, ertapenem, fosfomycin, gentamicin, meropenem, piperacillin/tazobactam, sulfamethoxazole/trimethoprim, tetracycline and tigecycline. Antimicrobial testing was performed on Mueller–Hinton agar and the breakpoints were interpreted according to the European Committee on Antimicrobial Susceptibility Testing (EUCAST) guidelines (http://www.eucast.org/).

**Genomic DNA extraction and next-generation sequencing**

Bacterial DNA was extracted from 0.2 ml of 3 McFarland suspension using the UltraClean microbial DNA isolation kit (MO BIO Laboratories, Inc., USA) according to the manufacturer’s instructions, and the DNA concentration was measured using a NanoDrop 2000 spectrophotometer (Thermo Scientific, USA). The genomic DNA was stored at −20 °C. One microgram of genomic DNA was fragmented by sonication on the Covaris S2 system. The library was prepared with NEBNext kits (NEBNext End Repair Module, NEBNext dA-tailing module and NEBNext quick ligation module) (New England Biolabs, Inc., USA), according to the manufacturer’s instructions. In brief, fragmented DNA was end-repaired, dA-tailed, ligated to ion-compatible adapters, size-selected for optimum insert length, enriched and quantified. Sequencing was carried out with the MiSeq reagent kit 300 v2 (Illumina, USA). Next-generation sequencing was performed on the Illumina MiSeq, generating 250 bp paired-end reads.

**Analysis of sequenced data**

The flow of NGS data, including expected required times, is presented in Fig. 1. Raw data generated on the Illumina MiSeq sequencing platform were uploaded onto web-based tools hosted by the Center for Genomic Epidemiology (http://www.genomic epidemiology.org/), as of March 2016, and also imported into CLC Genomics Workbench software, version 6.0.4 (CLC Bio, Aarhus, Denmark). CLC Genomics Workbench is a commercial application that provides numerous tools for the analysis and visualization of NGS data. Although the average size of the raw data files containing paired-end reads ranged between 1–1.4 Gb, the import into CLC Genomics Workbench using the import tool was fast and easy. Sequence quality trimming was carried out using the default settings, i.e. the quality score limit was set at 0.05, the maximum number of ambiguities was set at 2 and reads that were less than 20 nucleotides in length were discarded. To obtain longer reads with higher accuracy, a CLC Genomics Workbench NGS core tool named ‘merge overlapping pairs’ was employed. The trimmed and merged data were assembled into contigs using the CLC Genomics Workbench De Novo assembly tool with the following default settings: 20 nucleotide-long words and 50 nucleotide-long bubbles; minimum contig length set at 200 nucleotides; auto detection of paired distances and scaffolding enabled. None of the operations required specific programming or computational skills, as the software employed (CLC Genomics Workbench) has an intuitive and well-organized graphical user interface (GUI).

The main evaluation of raw pair-end reads was carried out via the freely available interactive online tools MLST 1.8 [22], ResFinder 2.1 [23] and PlasmidFinder 1.3 [24],
hosted by the Center for Genomic Epidemiology (http://www.genomicepidemiology.org/), as of March 2016. ResFinder 2.1 and PlasmidFinder 1.3 use their own databases, which are constantly updated, while the MLST 1.8 allele sequence and profile data are obtained from PubMLST databases.

The paired-end reads were run through an MLST tool for sequence type (ST) detection. An unknown allele from one isolate was examined for coverage depth and after the required coverage of at least >50x the length of the MLST genes was confirmed, a newly discovered MLST type was submitted to the Institut Pasteur MLST Database (http://bigdb.web.pasteur.fr/perl/bigdb/bigdb.pl?db=pubmlst_klebsiella_seqdef_public&page=sequenceQuery).

The identification of acquired antimicrobial resistance genes using ResFinder was run with thresholds set at 98 % for sequence identity (% ID) and 60 % for minimum length. An identity threshold of 80 % was used in order to detect both large and small plasmids using PlasmidFinder.

Reference sequences of the selected virulence genes (ureA, fimH, kfu, uge, wabG, wzy_K1, mrkD, allS, rmpA and cf29a) known to be associated with K. pneumoniae virulence [6, 7, 25] were run through the NCBI nucleotide database, downloaded and imported into CLC Genomics Workbench software. Raw pair-end reads were thereafter mapped against the corresponding reference virulence genes. The presence or absence of a particular virulence gene was concluded on the basis of average mapping coverage. To conclude that a
particular virulence gene was present, it had to be covered over at least 45–50× of its length. For the purpose of using web-based tools, virulence factor analysis was performed using BIGSdb-Kp hosted by the Institut Pasteur MLST database (http://bigdweb.pasteur.fr/perl/bigsdb/bigsdb.pl?db=pubmlst_klebsiella_seqdef_public&page=sequenceQuery). The assembled contigs were run through selected schemes: efflux systems and regulators, virulence genes, cps cluster genes and urease-nickel cluster genes. Raw sequencing data were submitted to the NCBI Sequence Read Archive (SRA); the corresponding NCBI BioProject accession number is PRJNA359577.

RESULTS AND DISCUSSION

NGS summary and data processing

The main quantitative and qualitative characteristics of the genomic data are presented in Table 1. According to an estimated PHRED score of 37, the average base call accuracy was as high as 99.98%. The coverage presents the number of sequences that support the individual base position. The average coverage value varied from 76 to 95%, depending on the particular isolate. The GC content calculated was around 55–57%. During sequence quality trimming 0.01% of raw data were discarded. The trimmed and merged pair-end reads were assembled into contigs. A technical description of the NGS data is not useful for clinicians and has no impact on treatment. However, these numbers demonstrate that the input data were of good quality (base call accuracy, coverage and a small percentage of discarded data).

The increasing use of NGS means there is a need to introduce valid tools for rapid data processing on a large scale. These approaches were used for clean bacterial cultures. For the purpose of the rapid identification of pathogens from clinical samples, there is need to separate pathogens’ nucleic acids from the preponderant host background. Various types of contamination in sequencing runs have been reported, e.g. sequences obtained from impure samples, contamination of the chemistry for nucleic acid preparations or the identification and removal of human contamination from microbial metagenomes [26] as well as mitigation techniques [27]. Several bioinformatics analysis pipelines were developed for the recovery of pathogen sequences from NGS data in time frames that are relevant for clinical use [28]. Unbiased metagenomic NGS in clinical settings offers the possibility of identifying pathogens’ nucleic acids in clinical samples without a priori knowledge of the target. For this purpose, a rapid cloud-compatible bioinformatics pipeline was developed: sequence-based ultra-rapid pathogen identification (SURPI) [29] is able to detect any pathogen sequence present in the GenBank reference database. Other online analysis pipelines that have been developed are the PathoSystems Resource Integration Center (PATRIC) [30], Computational Pathogen Sequence IDentification (CaPSID) [31], PathSeq [32], SPAdes [33] and Prokka [34]. Bioinformatics tools that are focused solely on gene detection were also developed: Antibiotic Resistance Genes Online (ARGO) [35], the microbial database of protein toxins, virulence factors and antibiotic resistance genes (MvirDB) [36], the Antibiotic Resistance Genes Database (ARDB) [37], the Comprehensive Antibiotic Resistance Database (CARD) [38] or the Antibiotic Resistance Gene-ANNOTation (ARG-ANNOT) [39]. The software packages that are often used for NGS data analyses were summarized previously [16, 40].

Multilocus sequence typing characterization

Ten of 11 isolates were classified as previously identified MLST types (Table 2). These isolates were from seven different sequence types: ST11, ST23, ST37, ST323, ST433, ST495 and ST562. The last isolate was characterized as an unknown sequence type. An unknown combination of ST495 and ST562. The last isolate was characterized as an unknown sequence type. An unknown combination of alleles was verified by Sanger sequencing [41] and set as a new sequence type named ST1646 (gapA-10, infB-3, mdh-4, pgi-97, phoE-12, rpoB-1 and tonB-39) on the K. pneumoniae MLST database. Four sequence types had previously been identified sporadically: ST323 [42–46], ST433 [47–49], ST495 [44] and ST562 [50, 51].

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Reads per pair</th>
<th>Average coverage (%)</th>
<th>GC content (%)</th>
<th>No. of contigs</th>
<th>Bases in contigs</th>
<th>Maximum contig size</th>
<th>Average contig size</th>
<th>N50 contig size</th>
<th>N75 contig size</th>
<th>NCBI accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1343 030</td>
<td>-95</td>
<td>56</td>
<td>269</td>
<td>5,609,272</td>
<td>360,300</td>
<td>20,852</td>
<td>154,536</td>
<td>947,10</td>
<td>SRA2454479</td>
</tr>
<tr>
<td>2</td>
<td>2344 634</td>
<td>-93</td>
<td>55</td>
<td>516</td>
<td>5,800,100</td>
<td>431,351</td>
<td>11,241</td>
<td>129,258</td>
<td>76,775</td>
<td>SRA2454484</td>
</tr>
<tr>
<td>3</td>
<td>2108 350</td>
<td>-95</td>
<td>56</td>
<td>506</td>
<td>6,112,699</td>
<td>868,571</td>
<td>11,883</td>
<td>275,808</td>
<td>134,725</td>
<td>SRA2454482</td>
</tr>
<tr>
<td>4</td>
<td>1728 130</td>
<td>-86</td>
<td>56</td>
<td>304</td>
<td>5,551,496</td>
<td>771,644</td>
<td>18,262</td>
<td>248,869</td>
<td>125,665</td>
<td>SRA2454487</td>
</tr>
<tr>
<td>5</td>
<td>1845 098</td>
<td>-92</td>
<td>56</td>
<td>524</td>
<td>5,936,585</td>
<td>549,951</td>
<td>11,329</td>
<td>274,321</td>
<td>130,584</td>
<td>SRA2454485</td>
</tr>
<tr>
<td>6</td>
<td>1741 518</td>
<td>-89</td>
<td>57</td>
<td>326</td>
<td>5,489,236</td>
<td>533,215</td>
<td>16,838</td>
<td>254,296</td>
<td>101,072</td>
<td>SRA2454483</td>
</tr>
<tr>
<td>7</td>
<td>2015 966</td>
<td>-76</td>
<td>57</td>
<td>445</td>
<td>5,632,844</td>
<td>623,085</td>
<td>12,658</td>
<td>304,503</td>
<td>134,768</td>
<td>SRA2454478</td>
</tr>
<tr>
<td>8</td>
<td>2113 305</td>
<td>-83</td>
<td>57</td>
<td>457</td>
<td>5,638,923</td>
<td>462,795</td>
<td>12,503</td>
<td>283,415</td>
<td>100,876</td>
<td>SRA2454481</td>
</tr>
<tr>
<td>9</td>
<td>1273 658</td>
<td>-83</td>
<td>57</td>
<td>406</td>
<td>5,666,117</td>
<td>304,274</td>
<td>13,956</td>
<td>139,360</td>
<td>43,471</td>
<td>SRA2454488</td>
</tr>
<tr>
<td>10</td>
<td>1922 150</td>
<td>-80</td>
<td>55</td>
<td>206</td>
<td>6,017,466</td>
<td>507,436</td>
<td>29,211</td>
<td>168,138</td>
<td>91,252</td>
<td>SRA2454480</td>
</tr>
<tr>
<td>11</td>
<td>2561 046</td>
<td>-82</td>
<td>57</td>
<td>453</td>
<td>5,526,004</td>
<td>904,590</td>
<td>12,199</td>
<td>277,622</td>
<td>133,759</td>
<td>SRA2454486</td>
</tr>
<tr>
<td>Isolate</td>
<td>ST</td>
<td>β-lactamase genes</td>
<td>PMQR genes</td>
<td>Aminoglycoside genes</td>
<td>Additional AR genes</td>
<td>Virulence genes by CLC genomic Workbench</td>
<td>Virulence gene clusters by BIGSdb-Kp* ‡  †  †</td>
<td>cps cluster</td>
<td>Efflux pump ‡  †  †</td>
<td>Replicons</td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
<td>-------------------</td>
<td>------------</td>
<td>----------------------</td>
<td>---------------------</td>
<td>--------------------------------------</td>
<td>-----------------------------------------------</td>
<td>-------------</td>
<td>-------------------</td>
<td>----------</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>blacrTX-M10.1-1</td>
<td>aac(6’)-Ib, cr A, opxA, opB, QnrB4</td>
<td>aac(3)-Ila, aac(6’)-Ib-cr, aadA1, aadA2</td>
<td>dfrA14, dfrA12, sul1, catA1, catB3</td>
<td>allS, mkrD, fimH, ureA, wabG, uge</td>
<td>mkr, yersiniabactin</td>
<td>wzi-75</td>
<td>OqxAB, AcrAB</td>
<td>FIIR(K), FIB(K), FIA (H1), R</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>blacrTX-M10.1-1</td>
<td>aac(6’)-Ib, cr A, opxA, opB, QnrB4</td>
<td>aac(3)-Ila, aac(6’)-Ib-cr, aadA2</td>
<td>dfrA14, dfrA12, sul1, catA1, catB3</td>
<td>allS, mkrD, fimH, ureA, wabG, uge</td>
<td>mkr, yersiniabactin</td>
<td>wzi-75</td>
<td>OqxAB, AcrAB</td>
<td>FIIR(K), FIB(K), FIA (H1), R</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>blacrTX-M10.1-1</td>
<td>aac(6’)-Ib, cr A, opxA, opB, QnrB4</td>
<td>aac(3)-Ila, aac(6’)-Ib-cr, strA, strB</td>
<td>dfrA14, sul2, catB3, tet(A)</td>
<td>allS, mkrD, fimH, ureA, wabG, uge</td>
<td>mkr, yersiniabactin, allantoinase, aerobactin, salmochelin, kfu, rmpA, microcin E492, colibactin</td>
<td>mkr wzi-1, wzc-1, K1</td>
<td>OqxAB, AcrAB</td>
<td>FIIR(K), FIB(K), H1B</td>
</tr>
<tr>
<td>4</td>
<td>37</td>
<td>blacrTX-M10.1-1</td>
<td>aac(6’)-Ib, cr A, opxA, opB, QnrB4</td>
<td>aac(3)-Ila, aac(6’)-Ib-cr, strA, strB</td>
<td>dfrA14, sul1, sul2, catB3, tet(A)</td>
<td>allS, mkrD, fimH, ureA, wabG, uge</td>
<td>mkr wzi-14, wzc-15, K14</td>
<td>OqxAB</td>
<td>FIIR(K), FIB(K)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>323</td>
<td>blacrTX-M10.1-1</td>
<td>aac(6’)-Ib, cr A, opxA, opB, QnrB4</td>
<td>aac(3)-Ila, aac(6’)-Ib-cr, strA, strB</td>
<td>dfrA14, sul2, catB3</td>
<td>allS, mkrD, fimH, ureA, wabG, uge</td>
<td>mkr wzi-87, wzc-22, K21</td>
<td>OqxAB</td>
<td>FIIR(K), FIB(K), FIB (pKPH51), H2A</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>323</td>
<td>blacrTX-M10.1-1</td>
<td>aac(6’)-Ib, cr A, opxA, opB, QnrB4</td>
<td>aac(3)-Ila, aac(6’)-Ib-cr, strA, strB</td>
<td>dfrA14, sul2</td>
<td>allS, mkrD, fimH, ureA, wabG, uge</td>
<td>mkr wzi-87, wzc-22, K21</td>
<td>OqxAB</td>
<td>FIIR(K), FIB(K)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>433</td>
<td>blacrTX-M10.1-1</td>
<td>aac(6’)-Ib, cr A, opxA, opB, QnrB4</td>
<td>aac(3)-Ila, aac(6’)-Ib-cr, strA, strB</td>
<td>dfrA14, sul2, catB3, tet(A), fosA</td>
<td>allS, mkrD, fimH, ureA, wabG, uge</td>
<td>mkr wzi-173, OqxAB</td>
<td>AcrAB</td>
<td>FIIR(K), FIB(K)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>433</td>
<td>blacrTX-M10.1-1</td>
<td>aac(6’)-Ib, cr A, opxA, opB, QnrB4</td>
<td>aac(3)-Ila, aac(6’)-Ib-cr, strA, strB</td>
<td>dfrA14, sul1, sul2, catB3, tet(A), fosA</td>
<td>allS, mkrD, fimH, ureA, wabG, uge</td>
<td>mkr wzi-173, OqxAB</td>
<td>AcrAB</td>
<td>FIIR(K), FIB(K), FIB (pKPH51), H1B, R</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>495</td>
<td>blacrTX-M10.1-1</td>
<td>aac(6’)-Ib, cr A, opxA, opB, QnrB4</td>
<td>aac(3)-Ila, aac(6’)-Ib-cr, strA, strB</td>
<td>dfrA14, sul2, catB3</td>
<td>allS, mkrD, fimH, ureA, wabG, uge</td>
<td>mkr wzi-174, OqxAB</td>
<td>AcrAB</td>
<td>FIIR(K), FIB(K), FIB (pKPH51), H2A, Col(RN1221)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>562</td>
<td>blacrTX-M10.1-1</td>
<td>aac(6’)-Ib, cr A, opxA, opB, QnrB4</td>
<td>aac(3)-Ila, aac(6’)-Ib-cr, strA, strB</td>
<td>dfrA14, sul1, sul2, catA1, catB3</td>
<td>allS, mkrD, fimH, ureA, wabG, uge</td>
<td>mkr wzi-18, wzc-19, K18</td>
<td>OqxAB</td>
<td>AcrAB</td>
<td>FIIR(K), FIB(K), R, H2A, H2B, Col(RN1221), Q2 FIIR(K)</td>
</tr>
</tbody>
</table>

*Yersiniabactin cluster: ybt, irp and fyuA genes.
†Aerobactin cluster: lac and lut genes.
‡Allantoinase cluster: all, fdr, glx, ybb, YP1_1364, KPI_1371 and hyy genes.
§AcAB efflux pump: acr, mar, sox, rob and ram genes.
||OqxAB efflux pump: opr and rar genes.
types are globally disseminated high-risk clones (ST11 and ST37) or clones associated with pyogenic liver abscess (ST23) [6, 52–55].

Characterizing the sequence type is essential for epidemiological surveillance in hospitals in particular and in the corresponding geographical area in general. The contigs were analysed using an MLST tool. The results for all of the alleles (tested at once) and an appropriate ST were obtained in minutes according to the web response. In the case of isolates with an unknown ST, the allele numbers were identified and then used to subscribe the new ST to the MLST Pasteur web.

Antimicrobial susceptibility testing and resistance genes

All of the sequenced K. pneumoniae isolates were ESBL producers. All of the isolates were susceptible to colistin, fosfomycin and meropenem. The phenotypic resistance to other antimicrobials is shown in Table 3.

Eleven to 20 acquired resistance genes of 8 antimicrobial agent families were identified by ResFinder among all the isolates: aminoglycoside, beta-lactamase, fluoroquinolone, fosfomycin, phenicol, sulphonamide, tetracycline and trimethoprim (Table 2). Agreement was found between in silico detected beta-lactamase, aminoglycoside, sulphonamide and trimethoprim genes and phenotypically tested antimicrobials. The isolates for which fluoroquinolone resistance genes had been detected were phenotypically resistant, except isolate 5, which lacks the QnrB gene. Isolates 3, 4, 5, 7, 8 and 9 were phenotypically susceptible to chloramphenicol, but the phenicol-coding resistance gene catB3 was identified. This gene was shorter (442 bp) in all of the detected cases than in the corresponding sequence from the database (633 bp). A similar disagreement between phenotype and genotype has previously been described [23], with this being explained by the fact the found gene was shorter than the resistance gene, as in our case. Three fosA genes were identified in isolates 7, 8 and 9, but all of them were phenotypically susceptible. We found no evidence in the literature for carriage of the fosA gene while the phenotype is susceptible. Isolates 1 and 2 (both ST11) carried the bldDJA gene for AmpC cephalosporinases. Cefoxitin disks were used to screen AmpC producers in these isolates with the outcome of phenotypically susceptible. Isolates 1 and 10 showed an intermediate susceptibility to ertapenem, and neither of them were found to contain a carbapenemase-coding gene matching this phenotype. Carbapenem resistance with a lack of carbapenemases is usually explained by a combination of impaired outermembrane permeability or porin gene nonexpression (porin changes/loss) along with ESBL production [56, 57]. Seven isolates were tetracycline-resistant, but one of them (isolate 10) was found not to contain a tetracyclin resistance gene with default 98 % ID settings. This discrepancy was caused by the presence of the tet(A) gene, but with a lower % identity (94.92 % ID).

According to Zankari et al. [58], high concordance (99.74 %) between phenotypic and predicted antimicrobial susceptibility was observed in Salmonella, Escherichia and Enterococcus. According to Stoesser et al. [59], the sensitivity of the genotype for predicting resistance in Escherichia and K. pneumoniae was 96 %. In the PRIMERS I and II studies diagnostic molecular platforms were used to detect resistance against beta-lactams, with varying degrees of success among the tested antibiotics. The imipenem, cefazidime and cefepime resistance sensitivities were >95 %, and the piperacillin/tazobactam resistance sensitivities were ≤80 % [60]. In our limited cluster of isolates, a high number of discrepancies were observed between the phenotypes and the acquired resistance genes detected using web tools. Thus, validation with larger datasets is necessary. In the literature, phenotypic testing is usually performed first, and then only phenotypically resistant isolates are tested by PCR for the presence of resistance genes. The discordance between antimicrobial resistance genotype/phenotype is complex and, in addition to being caused by antimicrobial resistance genes, it can be caused by other known resistance mechanisms, such as porin genes or efflux pumps [59]. Another explanation for this discordance might be the presence of a currently undescribed allelic variant associated with resistance, allelic variants associated with susceptibility, or as yet undiscovered genes.

These genotype/phenotype mismatches might result in diagnostic errors and inappropriate treatment. In such cases, clinicians would over-treat patients needlessly. The consequences of this may be further propagation of resistance through the application of undesired selective pressure, increased costs and possibly increased toxicity. On the other hand, failing to identify resistance has more direct consequences for the patient, who may receive ineffective or suboptimal antibiotic treatment. The failure of ineffective treatments may result in death, especially in critically ill patients [60, 61].

The advantage of NGS is that it allows us to detect all known resistance genes present in the genome, and not only those selected and tested by PCR according to our expectations based on phenotypic testing. ResFinder allows us to detect acquired resistance genes, not chromosomal ones. All antimicrobial groups or a selected antimicrobial group can be set for testing at once. The settings for minimal length and % identity between the gene in the database and the corresponding sequence in the genome can be the default ones or they can be changed as required. The turnaround time for acquired resistance gene detection is calculated in minutes or dozens of minutes, depending on the server workload.

Virulence-associated gene detection

Although the Center for Genomic Epidemiology enables virulence factor identification, it is only available for Escherichia coli, Enterococcus and Staphylococcus aureus, and not for K. pneumoniae. The genes selected for virulence typing by the CLC Genomic Workbench were associated...
with virulent and hypervirulent phenotypes in *K. pneumoniae* strains [6, 7]. Therefore the virulence genes were tested using the web-based BIGSdb-Kp database. The genes detected by CLC Genomic Workbench and BIGSdb-Kp are shown in Tables 2 and 4.

Another basic virulence factor of *K. pneumoniae* exhibiting structural diversity, which translates into different antigenic properties relevant to bacterial virulence, is capsular polysaccharide (CPS) [62]. Several capsular serotypes (K types) were defined according to the detected genes *wzi* and *wzc*, which are connected with K-type identification (Table 2).

Multidrug efflux pumps are an important mechanism of antimicrobial resistance in Gram-negative pathogens. In addition to contributing to the multidrug resistance phenotype, the efflux pumps represent a virulence factor that is required for *K. pneumoniae* to resist the lung’s innate immune defence mechanisms, thus facilitating the onset of pneumonia [63, 64]. The genes contributing to the AcrAB efflux pump and/or the OqxAB efflux pump were found (Table 2).

The 10 selected virulence genes associated with *K. pneumoniae* were detected by commercial CLC Genomic Workbench software and thus the results were compared to the BIGSdb-Kp database for detected genes. Four of these genes (*kfu, mrk, all* and *rmp*) were detected by both BIGSdb-Kp and CLC Genomic workbench software, but the other four genes (*ure, fim, uge* and *wab*) identified in almost all of the isolates, as well as the *wzy* _K1_ gene, were only found using CLC Genomic Workbench. Despite BIGSdb-Kp only being able to identify half of the genes searched for, the advantage of database tools and NGS in general is the detection of a huge variety of virulence genes and, moreover, genes selected and searched for by PCR reactions. This is shown in the example of isolate 3, where many additional virulence genes were found. By using NGS and BIGSdb-Kp, we were also able to characterize genes associated with polysaccharide capsule production, and we identified the capsule type without needing to use the serotyping method in six isolates. The BIGSdb-Kp provides information about the presence of gene-coding efflux pumps and their regulator genes, which can contribute to antimicrobial resistance as well as virulence. The response time for virulence factor detection in BIGSdb-Kp is several minutes. Another virulence database exists, the Virulence Factor Database (VFDB) [65],

### Table 3. Phenotypic antibiograms of *K. pneumoniae* isolates

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Susceptibility</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TET</td>
</tr>
<tr>
<td>1</td>
<td>S</td>
</tr>
<tr>
<td>2</td>
<td>S</td>
</tr>
<tr>
<td>3</td>
<td>R</td>
</tr>
<tr>
<td>4</td>
<td>R</td>
</tr>
<tr>
<td>5</td>
<td>S</td>
</tr>
<tr>
<td>6</td>
<td>S</td>
</tr>
<tr>
<td>7</td>
<td>R</td>
</tr>
<tr>
<td>8</td>
<td>R</td>
</tr>
<tr>
<td>9</td>
<td>R</td>
</tr>
<tr>
<td>10</td>
<td>R</td>
</tr>
<tr>
<td>11</td>
<td>R</td>
</tr>
</tbody>
</table>

AMK, amikacin; CIP, ciprofloxacin; CMP, chloramphenicol; COL, colistin; ETP, ertapenem; FEP, cefepime; FOS, fosfomycin; GEN, gentamicin; MER, meropenem; SXT, trimethoprim/sulfamethoxazole; TZP, piperacillin/tazobactam; TET, tetracycline; TIG, tigecycline; S, susceptible; I, intermediate; R, resistant.

### Table 4. Virulence factor description of *K. pneumoniae* isolates

<table>
<thead>
<tr>
<th>Locus/cluster</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allantoinase cluster</td>
<td>Allantoin anaerobic assimilation</td>
<td>[68]</td>
</tr>
<tr>
<td>Yersiniabactin cluster</td>
<td>Iron acquisition system</td>
<td>[69]</td>
</tr>
<tr>
<td>Aerobactin cluster</td>
<td>Iron acquisition system</td>
<td>[70]</td>
</tr>
<tr>
<td>Ferric uptake kfu cluster</td>
<td>Iron acquisition system</td>
<td>[71]</td>
</tr>
<tr>
<td>Microcin E492 mec cluster</td>
<td>Bacteriocin production</td>
<td>[72]</td>
</tr>
<tr>
<td>Colibacin clb cluster</td>
<td>Toxin production</td>
<td>[73]</td>
</tr>
<tr>
<td>mrk cluster</td>
<td>Mannose-resistant <em>Klebsiella</em>-like (type III) fimbriae</td>
<td>[74]</td>
</tr>
<tr>
<td>fim cluster</td>
<td>Type I fimbriae</td>
<td>[74]</td>
</tr>
<tr>
<td>c29a</td>
<td>Nonfimbrial adhesin</td>
<td>[75]</td>
</tr>
<tr>
<td>rmpA</td>
<td>Regulator of mucoid phenotype</td>
<td>[76]</td>
</tr>
<tr>
<td>wzy, K1</td>
<td>Associated with hypermucoviscosity phenotype</td>
<td>[25]</td>
</tr>
<tr>
<td>wabG</td>
<td>Outer-core lipopolysaccharide biosynthesis</td>
<td>[77]</td>
</tr>
<tr>
<td>uge</td>
<td>Capsule and lipopolysaccharide formation</td>
<td>[78]</td>
</tr>
<tr>
<td>ureA</td>
<td>Urease synthesis</td>
<td>[79]</td>
</tr>
</tbody>
</table>
but it is not specific for *Klebsiella* or as user-friendly as BIGSdb-Kp.

**In silico plasmid characterization**

By using PlasmidFinder, 10 plasmid replicon types of both small and large plasmids were detected (Table 2). A replicon of large plasmid FII(K) was found in all of the isolates. With the exception of isolate 11, replicon FIB(K) was also found in these isolates. In addition to these replicons, isolates 1, 2 and 8 also contained the FIA replicon. The identified replicons were from plasmid incompatibility groups IncF, IncHI1 and IncHI2. The IncHI1B replicon was only detected in isolate 3, and the IncHI2 replicon was detected in isolates 5 and 10. Two detected replicon sequences Col (RNAS) and Q2 were predicted to belong to the small-size plasmid groups. Col plasmids are related to colicin production by and against *Escherichia coli* and related bacteria [66]. Plasmids from the Q2 family are small (<20 kb), have a broad host range and are highly promiscuous [67]. PlasmidFinder and the rest of the tools at the Center for Genomic Epidemiology provide results in several minutes, according to server response, using raw data or contigs.

As we expected, all of the highly virulent and resistant isolates carried replicons of large plasmids that allow the horizontal spread of resistance genes and virulence factors. For precise plasmid characterization and carried gene detection, plasmid DNA isolation and sequencing is needed.

**Summary**

The web-based tools we selected enabled us to extract a large amount of information from NGS data. Although the data obtained have no direct impact on treatment, they are relevant for local and/or global epidemiology. The interface of these tools is user-friendly and does not require advanced informatics skills, such as knowledge of a programming language, while the data processing is very time-efficient. The time range from receipt of the isolates to complete genomic analysis varied according to the methods and commercial kits employed. Following the steps described below, the DNA isolation and library preparation take 2 working days, while sequencing itself takes 1 day. Raw data analysis takes 1–2 days and the time required for subsequent data analysis using web-based tools varies from minutes to hours, according to the server response. Commercial kits that enable one to shorten the library preparation to several hours are available. Sequence reads that are longer than those we have described could be obtained with some commercial kits, but this would be at the cost of prolonged sequencing (up to 2 days). All of the sequenced isolates were ESBL producers in which we identified acquired resistance genes to major antibiotic groups. In addition, these strains contained virulence factors allowing adherence to and invasion of human tissues, efflux pumps contributing to multidrug resistance and virulence providing a selective advantage under the host or environmental state. All of the isolates contained replicons of large plasmids. From 11 randomly selected isolates, 4 belonged to globally disseminated STs: 2 isolates of ST11, 1 isolate of ST23 and 1 isolate of ST37. One isolate was newly identified as ST1646.

The web-based tools provided by the Center for Genomic Epidemiology and the Pasteur MLST database had a turn-around time of a couple of minutes. Using ResFinder, we did not obtain such a high concordance between phenotype and detected genotype as was observed in previous studies. The BIGSdb-Kp was not able to identify all of the selected genes found by the CLC Genomic Workbench software. We found virulence factors we had not expected, but which are significant for bacterial virulence. If we had only used PCR reactions for the most common genes, without NGS, these genes would have remained undetected. Once an isolate’s genome sequence is available, it can be reassessed rapidly for additional resistance gene mechanisms or other required information without the need for further laboratory work. A limitation associated with data analysis might be the database itself. Databases are filled by users and curated by administrators. The obtained sequencing data are compared to the data available in the database and in cases of poorly annotated references or those where information is lacking, the results might be distorted. Therefore, users should be mindful of this. Alternatively, a comparison of different databases is suitable. In cases of discordance in, for example, antimicrobial resistance genotype/phenotype it is necessary to consider undescribed allelic variants or undiscovered genes as well as multi-factorial genomic components rather than a single gene.

**Conclusions**

NGS is a modern method to further characterize bacterial samples and outbreak detection, and provides a comprehensive overview of genetic information in a single step, instead of through many specific PCR reactions with a narrow spectrum of preselected genes. Current freely available web-based tools enable us to obtain a large amount of information from NGS data, which can be used for more effective diagnosis and treatment of patients. Access to genomic data can help us utilize the information obtained through basic research for practical application. NGS and web-based tools can serve as a link between scientists undertaking basic research and microbiologists working with an emphasis on practice, e.g. the data concerning bacterial resistance and virulence could be used for the implementation of new examination methods.

The role of NGS in clinical microbiology will increase during the next years, not only for research, but also for molecular diagnostics, infection prevention and outbreak investigation. However, further studies are required to improve the workflow for NGS, shorten the library preparation turnaround times and NGS platform run times, reduce costs, develop standardized operational protocols and data analysis pipelines, improve reference databases, and establish proficiency testing and quality control measures.

Our study demonstrates that processing NSG data with web-based tools is easy and time-efficient, while it produces
comprehensive results. The main challenge is to choose and evaluate information from the generated outputs that is relevant for a particular task.

Funding information
This work was supported by the project ‘New approaches in research, diagnostics and therapy of hematological malignancies II’, MUNI/A-1106/2016, with the support of the Specific University Research Grant, as provided by the Ministry of Education, Youth and Sports of the Czech Republic and project TE02000058, provided by Technology Agency of The Czech Republic.

Acknowledgements
Our thanks go to Marketa Petrovska for her assistance in the laboratory and Matej Medvecky for usefulinformatics advice. We thank the team of curators at the Institut Pasteur MLST and whole-genome MLST databases for curating the data and making them publicly available at http://bigudb.web.pasteur.fr/.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References


