SCIENTIFIC REPORTS

natureresearch

OPEN

EasyPrimer: user-friendly tool for pan-PCR/HRM primers design. Development of an HRM protocol on *wzi* gene for fast *Klebsiella pneumoniae* typing

Matteo Perini[1], Aurora Piazza[1], Simona Panelli[1], Domenico Di Carlo[1], Marta Corbella[2], Floriana Gona[3], Francesca Vailati[4], Piero Marone[2], Daniela Maria Cirillo[3], Claudio Farina[4], Gianvincenzo Zuccotti[1,5] & Francesco Comandatore[1]*

In this work we present EasyPrimer, a user-friendly online tool developed to assist pan-PCR and High Resolution Melting (HRM) primer design. The tool finds the most suitable regions for primer design in a gene alignment and returns a clear graphical representation of their positions on the consensus sequence. EasyPrimer is particularly useful in difficult contexts, e.g. on gene alignments of hundreds of sequences and/or on highly variable genes. HRM analysis is an emerging method for fast and cost saving bacterial typing and an HRM scheme of six primer pairs on five Multi-Locus Sequence Type (MLST) genes is already available for *Klebsiella pneumoniae*. We validated the tool designing a scheme of two HRM primer pairs on the hypervariable gene *wzi* of *Klebsiella pneumoniae* and compared the two schemes. The *wzi* scheme resulted to have a discriminatory power comparable to the HRM MLST scheme, using only one third of primer pairs. Then we successfully used the *wzi* HRM primer scheme to reconstruct a *Klebsiella pneumoniae* nosocomial outbreak in few hours. The use of hypervariable genes reduces the number of HRM primer pairs required for bacterial typing allowing to perform cost saving, large-scale surveillance programs.

Most methods used for the identification and typing of prokaryotes are based on DNA amplification and sequencing. Indeed, the sequence of specific genes can harbour enough information to classify bacteria at species, subspecies or also to a clonal level. For instance, Multi-Locus Sequence Typing (MLST) is one of the most used methods for bacterial typing and it is based on the amplification and sequencing of few housekeeping genes[1]. During the last ten years, the analysis of the entire bacterial genome by Whole Genome Sequencing (WGS) approach revolutionized the field, drastically increasing the typing precision[1].

The reconstruction of nosocomial outbreaks is one of the most important clinical applications of bacterial typing. A nosocomial outbreak occurs when the number of patients infected by a pathogen increases above the expected in a limited time[2]. In these situations, it is fundamental to determine the clonality of bacteria causing disease in the patients to define the proper strategy to handle the emergency. Pulsed-Field Gel Electrophoresis (PFGE), MLST and WGS are the most frequently applied molecular methods in outbreak investigation[1].

During a nosocomial outbreak, clinicians need bacterial typing information in the shortest time possible. Despite the high potential of WGS in outbreak reconstruction, the sequencing of a complete genome requires two to four working days, introducing an important time lag. Similarly, PFGE typing requires five days and also MLST needs few days. During the last years, the High Resolution Melting (HRM) assay has emerged as a low-cost and fast method for bacterial typing, particularly promising for epidemiological applications[3–6]. HRM is a single-step

[1]Department of Biomedical and Clinical Sciences "L. Sacco", Università di Milano, Pediatric Clinical Research Center "Romeo and Enrica Invernizzi", Milan, 20157, Italy. [2]S.C. Microbiologia e Virologia, Fondazione IRCCS Policlinico San Matteo, Pavia, 27100, Italy. [3]Emerging Bacterial Pathogens Unit, Division of Immunology, Transplantation and Infectious Diseases, IRCCS San Raffaele Scientific Institute, Milan, 20132, Italy. [4]Microbiology Institute, A.S.S.T. "Papa Giovanni XXIII", Bergamo, 24127, Italy. [5]Department of Pediatrics, V. Buzzi Childrens' Hospital, Università di Milano, Milan, Italy. *email: francesco.comandatore@unimi.it
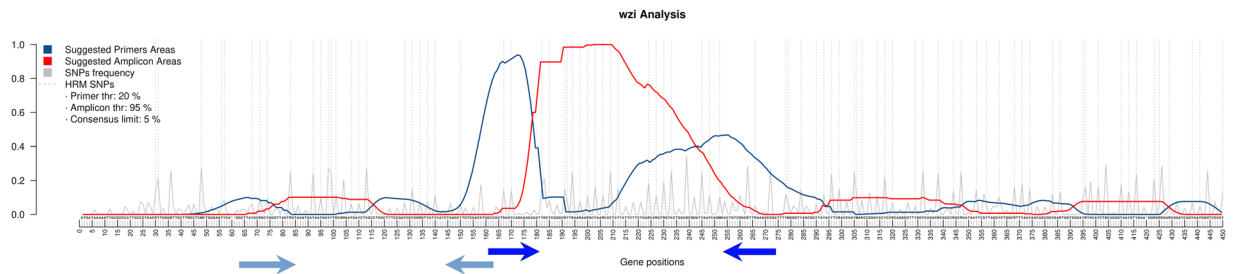
   1

Figure 1. EasyPrimer output for the *wzi gene* (563 alleles on https://bigsdb.pasteur.fr). The consensus sequence calculated from the gene alignment is reported on the x-axis. Residues under the peaks of the blue curve are highly conserved and thus suitable for primer design. Conversely the red curve increases over the highly variable regions suggested to be amplified. The grey peaks represent all the Single Nucleotide Polymorphisms (SNPs) with their own frequency. The dotted lines are used to highlight the "HRM-detectable" SNPs, i.e. the ones causing a change in the GC content. The blue arrows were manually added to show the positions of the two primer pairs designed on *wzi* in this work.

procedure for the discrimination of sequence variants on the basis of their melting temperature. This method allows to perform bacterial typing in less than five hours[7].

To develop a novel HRM-based typing procedure, it is necessary to: i) select one or more core genes; ii) design a primer pair in conserved regions flanking a gene portion where the melting temperature varies among the strains.

Andersson and colleagues have developed the "MinimumSNP" tool[8], which identifies, in a gene alignment, the variable positions that can lead to a melting temperature change (called informative SNPs). MinimumSNP identifies single informative positions, not necessarily grouped in a single portion of the gene. In other words, it does not indicate which regions are more suitable for primer design: two low-variable regions neighbouring a SNP-rich informative stretch. Thus, the user of MinimumSNP has to choose one (or few) SNPs and then design primers around it (or around them).

Herein, we present EasyPrimer, a web-based tool for the identification of the gene regions suitable for primer design to perform HRM studies and any kind of pan-PCR experiments. Moreover, we validated EasyPrimer by designing HRM primers for the discrimination of clinical isolates of *Klebsiella pneumoniae*, an important opportunistic pathogen frequently cause of infections in humans and animals[9].

This article was submitted to an online preprint archive[10].

Results

EasyPrimer: a tool for primers design. EasyPrimer is a user-friendly open-source tool developed to assists primer design in difficult contexts, e.g. on an alignment of hundreds of sequences and/or on hypervariable genes. The tool uses as input a nucleotide multi-fasta file and identifies the best regions for primer design: two low variable regions flanking a highly variable one. The on-line and the stand-alone versions of the tool are freely available at https://skynet.unimi.it/index.php/tools/easyprimer.

Primers design. We downloaded *pgi*, *gapA* and *wzi* gene sequences from BigsDB database (https://bigsdb.pasteur.fr) and we run EasyPrimer to identify the best regions for primer design. The EasyPrimer output for the *wzi* gene is reported in Fig. 1, while the outputs relative to *pgi* and *gapA* genes are reported in Supplementary Figs. S1 and S2, respectively. Then we designed a total of four novel primer pairs: one for *pgi*, one for *gapA* and two for *wzi* (reported in Table 1).

High-resolution melting analysis. In this work we considered three clinical strains collections:

- the "background" collection, which includes 17 *K. pneumoniae* strains belonging to 17 different Sequence Types (STs);
- the "outbreak" collection, which includes 11 *K. pneumoniae* strains isolated during a nosocomial outbreak;
- the "validation" collection that includes 54 *K. pneumoniae* strains belonging to six of the most epidemiologically relevant STs (i.e. ST258, ST512, ST11, ST101, ST15 and ST307)[11,12].

The strains of the background and outbreak collections were analysed using all the ten primer pairs listed in Table 1. The strains of the validation collection were subjected to HRM experiments using only the two primer pairs designed on *wzi* gene (*wzi-3* and *wzi-4* primer pairs). Four out of the ten primer pairs were newly designed in this work (see above), while the remaining six were already available in literature[7]. The obtained melting temperatures ("Tm") of the three HRM replicates and their relative average temperature ("aTm") values are reported in Supplementary Table S1.

Primer pairs and schemes comparison. For each of the ten primer pairs we calculated the strain distance matrix among the background collection strains based on the aTm values (see Methods). The calculated aTm distances ranged from zero to three degrees, and the median distances varied among the genes (as shown in Fig. 2). In particular, the two *wzi* primer pairs showed median distance values significantly higher than those obtained for many of the other primer pairs (see Supplementary Table S2 for details).

Gene name	Primer name	Primer pair name - in this work	Primer scheme name	Sequence (5′ - 3′)	Amplicon length (bp)	Reference
infB	infB729-F	infB-1	MLST6, MLST8	CCTGCCGGAAGAGTGG	50	Andersson et al. 2012
	infB729-R		MLST6, MLST8	TCGCGGAAACGTGGAC		Andersson et al. 2012
mdh	mdh1197-F	mdh-1	MLST6, MLST8	ATTGCCGACCTGACTAAACG	58	Andersson et al. 2012
	mdh1197 R		MLST6, MLST8	CTTTCGCTTCCACGACTTC		Andersson et al. 2012
phoE	phoE2013-F	phoE-1	MLST6, MLST8	GAAGGGGTGGGGAGTGA	78	Andersson et al. 2012
	phoE2013-R		MLST6, MLST8	GGCGTTCATGTTTTTGTTGA		Andersson et al. 2012
rpoB	rpoB2227-F	rpoB-1	MLST6, MLST8	TGATTAACTCCCTGTCCGTGT	132	Andersson et al. 2012
	rpoB2227-R		MLST6, MLST8	CGTAGTTGCCTTCTTCGATAGC		Andersson et al. 2012
tonB	tonB2693-F	tonB-1	MLST6, MLST8	GTTGAACCCGAACCTGAGC	101	Andersson et al. 2012
	tonB2693-R		MLST6, MLST8	GGTTTGGGCTTCGGCTTA		Andersson et al. 2012
tonB	tonB2886-F	tonB-2	MLST6, MLST8	AAAAGGTTGAACAGCCGAAG	120	Andersson et al. 2012
	tonB2886-R		MLST6, MLST8	CCGCTGCTGTCGAGGT		Andersson et al. 2012
gapA	gapA_F	gapA-1	MLST8	AAAGTCGTTCTGACTGGC	95	This work
	gapA_R		MLST8	TTRAAACGATGTCCTGGC		This work
pgi	pgi_F	pgi-1	MLST8	CCAAAATGGTACCGTGCGATT	156	This work
	pgi_R		MLST8	CCTGATCGCGRTATTCCTGCT		This work
wzi	wzi3_F	wzi-3	wzi	GCTTAYGCRGCYGGGGTTAGTRGT	114	This work
	wzi3_R		wzi	GGCCASGTCGACARGCTCAG		This work
wzi	wzi4_F	wzi-4	wzi	GCCGCTRAGYCAGGAAGAGAT	101	This work
	wzi4_R		wzi	GACTGTCWGCBTTRAAAGCSGA		This work

Table 1. Primer pairs used in this work.

We also compared the aTm distance matrices of the following schemes:

- "MLST6": which includes the HRM-MLST primer pairs already present in literature;
- "MLST8": which includes the MLST6 primer pairs and the two newly designed;
- "wzi": which includes the two primer pairs designed for the *wzi* gene.

The median pairwise distance did not significantly change among the three schemes (Wilcoxon test with Holm post-hoc correction, p-value > 0.05) and the relative boxplot graphs are reported in Supplementary Fig. S3.

Furthermore, we compared the aTm distance matrices of wzi and MLST8 schemes for each strain pair of the background collection, subtracting the two matrices (see Fig. 3). We found that, among all the 136 possible strain pairs, 66 (48.5%) showed a higher distance for wzi scheme than MLST8 scheme. More in detail, the ST258 *wzi_29* (also known as ST258 Clade 1[13]) was better discriminated from the ST512 *wzi_154* (which is part of the ST258 Clade 2[14]) by wzi scheme than MLST8 (see Fig. 3). ST307 was better or equally discriminated from all the other strains by wzi scheme than MLST8 (except for the ST512 and ST147 strains). Similarly, ST15 (*wzi_89*) strain was better discriminated from all the background strains by wzi scheme than MLST8, apart from the ST147 strain (see Fig. 3). Furthermore, the ST101 strain was better discriminated by wzi than MLST8 scheme from ST258 (*wzi_29*), ST307, ST512 and ST15 (*wzi_89*). ST11 (*wzi_75*) was better discriminated by wzi scheme from ST15 (*wzi_89*), ST307, ST258 and ST512. Conversely, ST11 (*wzi_75*) and ST101 were better discriminated by MLST8 than wzi. Lastly, ST147 was the only ST better discriminated by MLST8 scheme than wzi for all the strains pairs (see Fig. 3).

Whole genome sequencing-based strain typing.

A total of 82 *K. pneumoniae* strains have been subjected to WGS-based typing in this work.

24 out of 82 strains have been previously subjected to NGS sequencing as part of two published works:

- 12 from Gaiarsa and colleagues[14] (10/12 for the background and 2/12 for the validation collection).
- 12 from Gona and colleagues[15] (8/12 included in the background collection and 4/12 included in the validation collection).

For these strains, the NGS sequences were retrieved from public database.

For the remaining 58 out of 82 *K. pneumoniae* strains, the reads and genomic sequences were obtained in this work (see Supplementary Table S3):

- 11/58 strains of the outbreak collection (all from Papa Giovanni XXIII hospital).
- 47/58 strains of the validation collection (45/47 from San Raffaele hospital and 2/47 from Papa Giovanni XXIII hospital).

The *wzi* allele, the ST, the K-type as well as the accession numbers of these 82 *K. pneumoniae* strains are reported in Supplementary Table S3.

WGS-based outbreak reconstruction.

Ten out of the 11 outbreak isolates belonged to the ST512 while the isolate "BG-Kpn-22–18" belonged to the ST307 (see Supplementary Table S3). An alignment of 66 core-SNPs
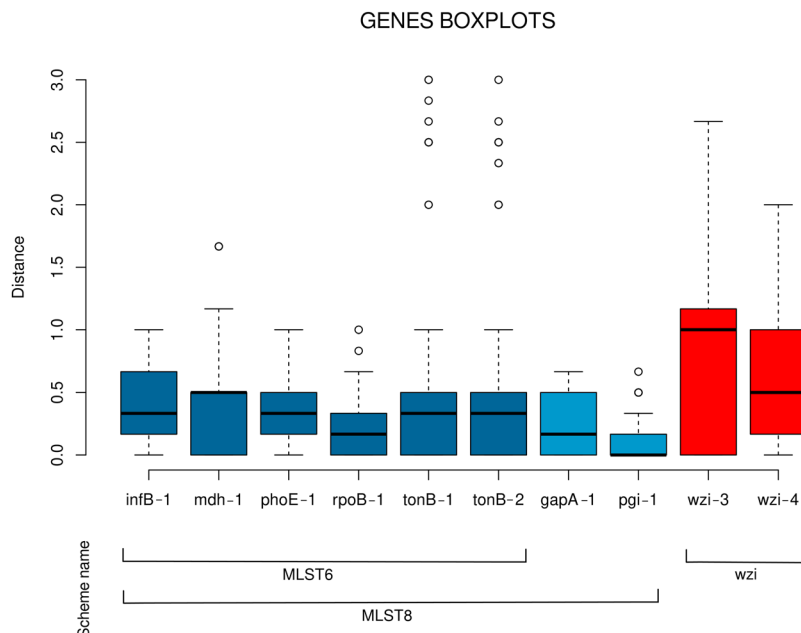
GENES BOXPLOTS

Figure 2. Distribution of the average melting temperature differences among the 17 *Klebsiella pneumoniae* strains for the ten primer pairs. Boxes are the 25th and 75th quartiles divided by the medians, whiskers are 1.5x the interquartile ranges and dots are outliers. The lines in the bottom show the composition of the three primer schemes used in this work.
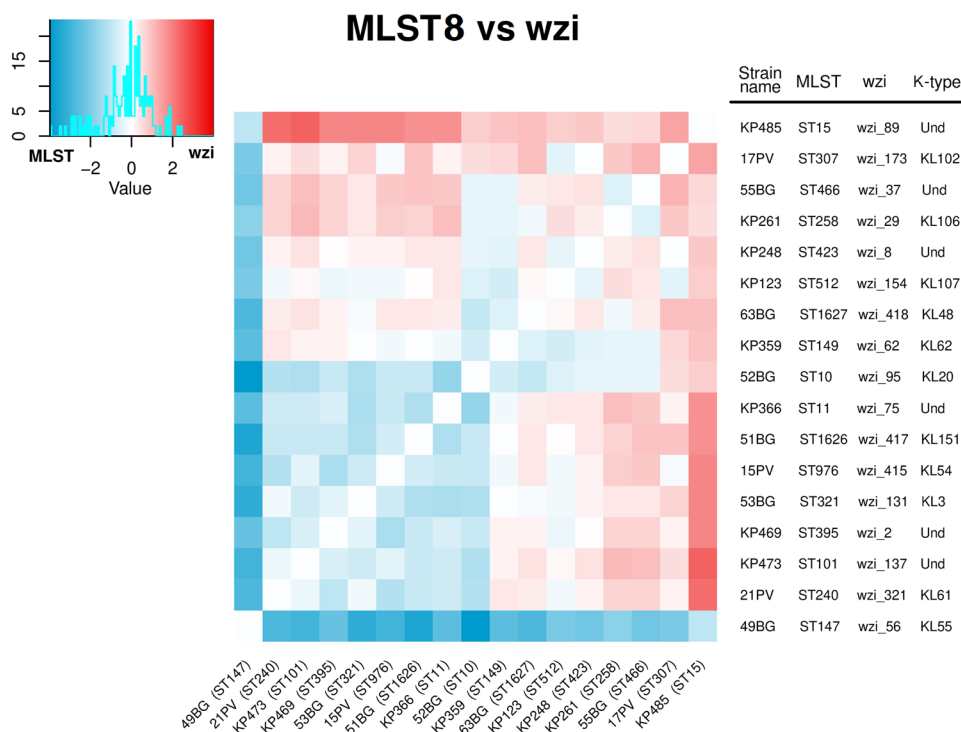

Figure 3. Arithmetic difference between the average melting temperature distance matrices computed among the 17 *Klebsiella pneumoniae* strains (selected to belong to 17 different STs) using the MLST8 scheme (eight primer pairs on seven genes) and wzi scheme (two primer pairon one gene). The heatmap colours range from blue to white to red: if the temperature distance between two strains is greater for the MLST8 than the wzi scheme the relative position on the heatmap is coloured in blue, otherwise in red.
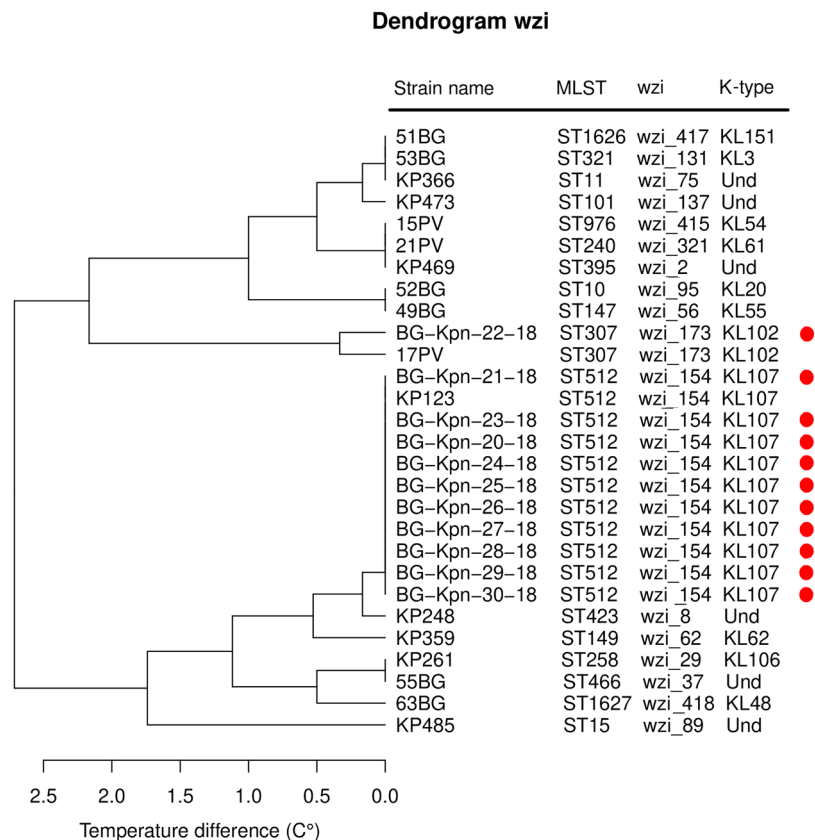
Figure 4. Dendrogram of the hierarchical clustering analysis on the average temperature distance matrix obtained using the *wzi* primer scheme. The 17 "background" strains belonging to 17 different MLSTs are written in black, while the 11 strains isolated during the nosocomial outbreak are highlighted by a red dot.

was obtained from the 11 outbreak strains. The relative Maximum Likelihood phylogenetic tree is reported in Supplementary Fig. S4. The ST512 strains have SNP distances ranging from zero to four SNPs. Conversely, the SNP distances among the ST512 strains and the ST307 strain ranged from 63 to 66 SNPs.

HRM-based outbreak reconstruction. The three dendrograms obtained by hierarchical clustering on the aTms strain distances for the schemes MLST6, MLST8 and wzi are reported in Supplementary Figs. S5 and S6 and Fig. 4, respectively. All the schemes correctly discriminated the outbreak ST512 strains from the ST307 one. Notably, only the wzi scheme correctly clustered the outbreak strains with the background strain of the same ST.

Repeatability of the *wzi* HRM protocol. To validate the repeatability of the *wzi* HRM typing protocol, we included in the analysis 54 additional strains belonging to the most epidemiologically relevant clades (ST258 Clade 1, ST258 Clade 2, ST307, ST101, ST11 and ST15)[11,12], for a total of 82 strains (see Supplementary Table S3).

For each of these clades, the aTms obtained for *wzi-3* and *wzi-4* primer pairs varied in a range $< = 0.5\,°C$, corresponding to the sensitivity of the machine used for the HRM experiments (see Table 2).

Clustering analysis on *wzi-3* and *wzi-4* melting temperatures grouped the strains in seven clusters (see Figs. 5 and 6). Among the most epidemiologically relevant clades, all the strains of the ST258 Clade 1 (ST258 *wzi_29*), ST258 Clade 2 (ST258 *wzi_154* and ST512 *wzi_154*) and ST307 were correctly clustered; while ST11 and ST101 strains fell in the same cluster. The three ST15 strains fell in two different clusters coherently to their *wzi* alleles: the one harbouring *wzi_89* clustered alone, while the other two strains, both harbouring *wzi_24*, fell in the ST11/ST101 cluster (see Figs. 5 and 6).

Discussion

High Resolution Melting (HRM) is a real-time PCR analysis for the detection of mutations and polymorphisms[3,4], also applicable for fast bacterial typing in hospital surveillance and real-time nosocomial outbreak detection[5]. Several works applied HRM to bacterial typing, exploiting Multi Locus Sequence Type (MLST) genes[7,16,17] which have been considered the gold standard genes for bacterial typing for almost 20 years[18]. These genes have been selected to be housekeeping therefore they display low variability. In this work we show that it is possible to increase HRM discriminatory power using hypervariable genes.

On the other hand, the identification of the regions suitable for primer design can be challenging when the number of aligned sequences is high or when the gene is hypervariable. Thus, we developed EasyPrimer, a tool for the identification of the best regions for primer design for HRM analysis and, more in general, for any kind of

Clade	MLST profile	*wzi* allele	Min-Max *wzi-3* T° (°C)	Min-Max *wzi-4* T° (°C)	Range *wzi-3* T° (°C)	Range *wzi-4* T° (°C)	# of strains
ST258 Clade 2	ST512	*wzi_154*	83.50–84.00	83.33–83.50	0.50	0.17	23
	ST258	*wzi_154*	83.83–83.83	83.50–83.50	0	0	3
ST258 Clade 1	ST258	*wzi_29*	83.50–84.00	84.33–84.67	0.50	0.34	8
ST307	ST307	*wzi_173*	84.00–84.50	82.33–82.83	0.50	0.50	19
ST101	ST101	*wzi_137*	84.83–85.00	83.33–83.83	0.17	0.50	12
ST11	ST11	*wzi_24*	85.00–85.00	83.50–83.50	0	0	2
ST15	ST15	*wzi_24*	85.00–85.17	83.83–83.83	0.17	0	2
	ST15	*wzi_89*	82.33–82.33	84.00–84.00	0	0	1

Table 2. Ranges of *wzi-3* and *wzi-4* melting temperatures of the most epidemiologically relevant clades and Sequence Types.

pan-PCR study. EasyPrimer shows, with an easy-to-read graphical output, which are the best regions for primer design: two conserved regions flanking a highly variable one. The on-line and the stand-alone versions of the tool are freely available at https://skynet.unimi.it/index.php/tools/.

We validated the tool designing HRM primers for the nosocomial pathogen *Klebsiella pneumoniae*. A scheme including six HRM primer pairs for five out of the seven *K. pneumoniae* MLST genes was already available in literature[7] (MLST6 scheme). Thus, we used EasyPrimer to design the primers for the remaining two MLST genes (*pgi* and *gapA*), obtaining a larger scheme with eight primer pairs (MLST8 scheme). Furthermore, we designed two HRM primer pairs for the hypervariable capsular gene *wzi* (wzi scheme). We tested the discriminatory power of these schemes on 17 *K. pneumoniae* strains belonging to 17 different STs (background collection) and we used the HRM approach to study an outbreak occurred in an Italian hospital.

Notably, most of the epidemiologically relevant *K. pneumoniae* clades (and/or STs) emerged after large recombination events that involved the capsule locus (which includes *wzi* gene), often leading to K-type change. For this reason, the emergence of a novel clade/ST is often associated to *wzi* allele change[19]. This makes *wzi* gene particularly suitable for *K. pneumoniae* typing. Our analyses on the background collection showed a good discriminatory power for both the MLST-based and *wzi*-based HRM assays: both schemes successfully discriminated most of the analysed strains. Wzi scheme discriminated better than MLST8 scheme all the highly epidemiologically relevant clades (ST258 Clade 1, ST258 Clade 2, ST307, ST11, ST101 and ST15), except for the pairs ST258 Clade 2 - ST307 and ST11 - ST101 (see Fig. 3). Nonetheless, wzi scheme is able to discriminate the ST258 Clade 2 from ST307 as well (see Fig. 4). Conversely, *wzi* scheme does not discriminate ST11 from ST101 (see Figs. 5 and 6). The latter may represent a minor flaw of the *wzi* HRM protocol as the two STs are mostly isolated in geographically distant areas of the globe (namely: ST11 in Asia[20], ST101 in Europe-Africa[21]).

Clustering analysis on the 82 strains, including multiple strains from the same clade (see Materials), allowed to evaluate with more precision the discriminatory power of the *wzi* HRM protocol. The analysis clearly showed that the protocol is able to discriminate five of the six most epidemiologically relevant *K. pneumoniae* clades, discriminating ST258 Clade 1, ST258 Clade 2, ST307, ST11/ST101 and ST15 (see Figs. 5 and 6). In our dataset we found strains of the ST15 and ST11 harbouring different *wzi* alleles. This is not surprising, considering that the capsule locus (which contain the *wzi* gene) is a recombinational hotspot in ST11[19]. Wzi scheme discriminated among the different ST15 strains present in the collections, according to the different *wzi* alleles they harbour (see Figs. 5 and 6). This highlights the benefits of using hypervariable genes instead of MLST genes in typing methods: e.g. *wzi* HRM protocol can rule out an ST15 outbreak of strains harbouring different *wzi* alleles, any MLST-based protocol cannot.

As stated above, *K. pneumoniae* clades are often associated to specific *wzi* alleles and K-types. Despite the *wzi* HRM protocol was designed on *wzi* gene, it correctly discriminates most of the epidemiological relevant clades. Furthermore, we found that every K-types correspond to a specific *wzi* allele (see Supplementary Table S3).

Moreover, the analysis of the 82 strains clearly showed that the *wzi* HRM protocol is highly repeatable: *wzi-3* and *wzi-4* aTms ranged within 0.5 °C (the machine sensitivity) among the strains of the same clade (see Table 2). For this study, dozens of independent HRM experiments have been performed in different months by two different operators (M.P. and A.P.). The observed stability of HRM aTms for each clade clearly shows that the results of *wzi* HRM protocol are portable. This makes the method suitable for studies involving several isolates, such as large hospital surveillance programs.

Additionally, we want to highlight that the observed HRM discriminatory power was obtained using a BioRad CFX Connect real-time PCR instrument (BioRad, Hercules, California): a machine not specifically designed for HRM experiments but for real-time PCR, with a melting temperature sensitivity of 0.5 °C (i.e. a lower sensitivity compared to HRM machines).

We applied the wzi scheme to the reconstruction of a nosocomial outbreak occurred in an Italian hospital. During the outbreak, 11 patients resulted colonized or infected by *K. pneumoniae* and the WGS typing revealed that the isolates belonged to two different clones. These clones were identified on the basis of core SNP distance (SNP distance < 5) and MLST profile (one isolate belongs to the ST307 and ten isolates to the ST512). As shown in Fig. 4, the wzi scheme not only correctly discriminated the outbreak isolates of the two clones but it also clustered them with the background isolates of the corresponding ST profile.

During the last years, WGS has revolutionized clinical microbiology, allowing the precise description of bacterial genomic features in few days (including the presence of resistance and/or virulence factors). Despite this, its application during real-time outbreak reconstructions still shows some limits: the time required to be
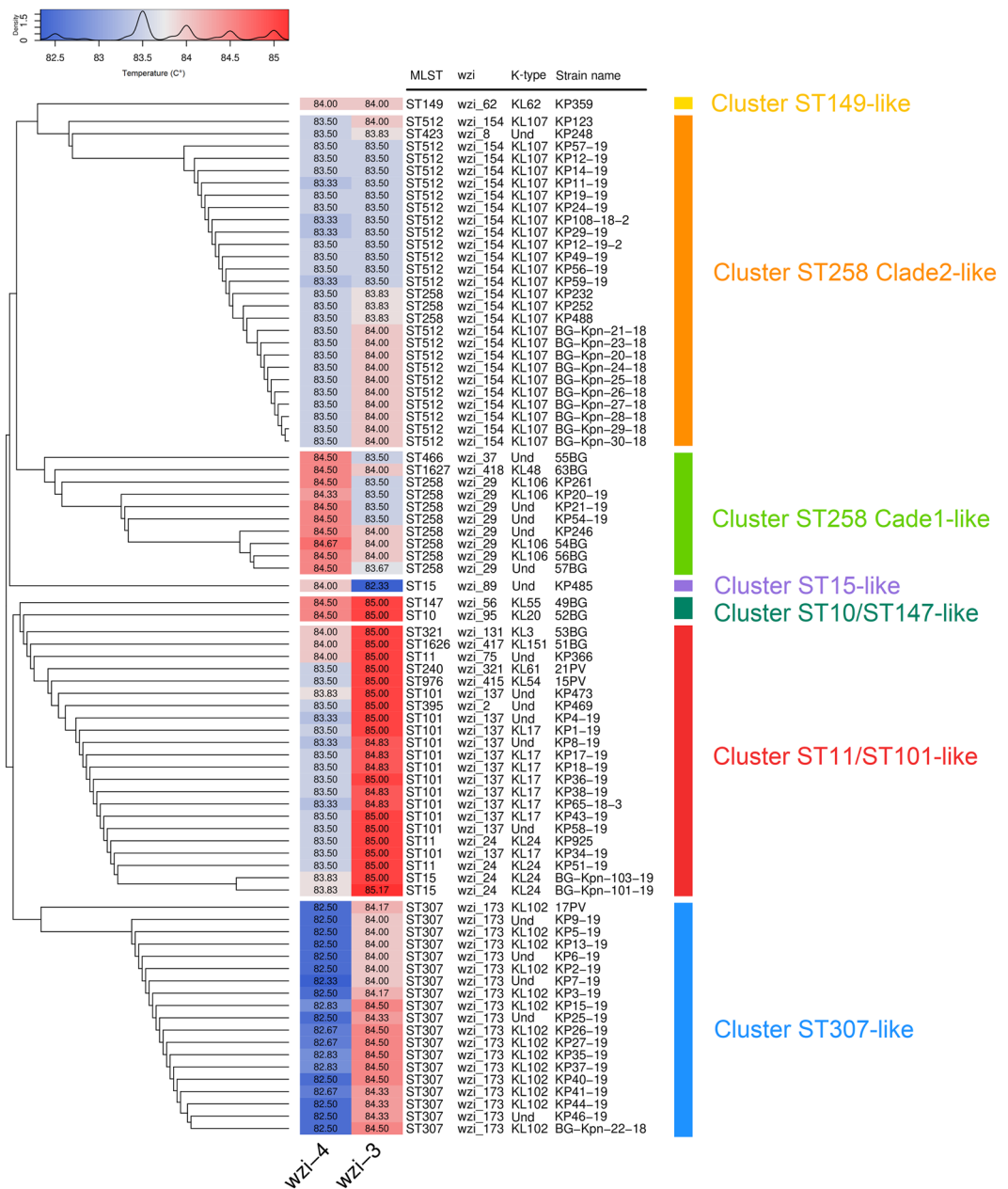
Figure 5. Strain-to-strain network of the 82 isolates analysed, generated on the basis of *wzi-3* and *wzi-4* HRM melting temperatures. Two strains were connected if both *wzi-3* and *wzi-4* gave difference in melting temperature < 0.5 °C. Clusters were identified as separated sub-networks on the strain-to-strain network and they were named from the major Sequence Type they include.

completed, the cost and the necessity of qualified personnel for library preparation, bioinformatic analyses and results interpretation. Indeed, the complete sequencing of a bacterial strain genome costs at least ~100 euros (using an Illumina MiSeq machine) and requires one or two days for library preparation and 5–36 hours for sequencing. During the first days of a nosocomial outbreak the number of cases still increases slowly. In this time frame, it is crucial to quickly understand if the bacterial strains are genetically related, and if the clone is spreading in the nosocomial environment. In this situation HRM is a "first-line" typing technology to figure out when an

Figure 6. On the left the dendrogram obtained from the strain-to-strain network reported in Fig. 5. In the middle, the average melting temperatures of *wzi-3* and *wzi-4* primer pairs, the Sequence Type, the *wzi* allele and the K-type of the 82 isolates analysed. On the right, the names of the clusters identified by network analysis, corresponding to the clusters in Fig. 5.

outbreak is starting. Indeed, HRM is less precise than WGS but it can be reliable for a fast, preliminary bacterial typing, fundamental in the first days of a nosocomial outbreak. If the outbreak is identified, WGS could be used to further investigate the transmission dynamics. HRM assay represents a fast, simple and time/cost saving approach for bacterial typing, allowing to analyse several bacterial samples per days. Furthermore, this technique does not require advanced skills in molecular biology and the results can be analysed without the use of any specific software. This method can be useful also in veterinary and dairy farming settings: *K. pneumoniae* is a relevant veterinary pathogen and one of the most frequent cause of mastitis in dairy cattle[9].

We found that the discriminatory power of an HRM scheme does not strictly depend on the number of genes but also on their genetic variability. Indeed, comparing the MLST6 and MLST8 schemes, we found that the median distance among the strains did not change significantly. Wzi scheme contains two primer pairs and this reduces drastically the amount of time and costs required for typing. For instance, using only two primer pairs on a 96-well PCR plate, it is possible to type 15 isolates per run (five hours, including DNA extraction, HRM run and analysis of results) with a cost of ~5 euros each. This makes the HRM a feasible method for real-time surveillance and for a preliminary typing step in large epidemiological studies. Lately, Multi-Drug Resistance (MDR) *K. pneumoniae* strains have become a major burden for public health worldwide. Despite WGS represents an important

tool for precise bacterial typing, it remains too demanding for developing countries healthcare systems. Low cost and simple protocols, as the *wzi* HRM typing proposed here, represent a real opportunity for surveillance programs.

The use of hypervariable genes in HRM-based bacterial typing, such as *wzi* in *K. pneumoniae*, can drastically increase the discriminatory power of the method. With the large number of genomes available in databases, it is now possible to find the most variable genes for a species. Unfortunately, it is not easy to identify the best regions to design primers in such hypervariable genes, particularly when hundreds of different alleles are available. EasyPrimer can represent a useful tool to overcome this limit.

Methods

Isolates collections. We considered three strain collections: the background, the outbreak and the validation collections. The background collection includes 17 strains belonging to 17 different STs retrieved from two previously WGS typed bacterial collections: nine strains from Gaiarsa and colleagues[14] and eight strains form Gona and colleagues[15] (for details see Supplementary Table S3). The outbreak collection includes 11 *K. pneumoniae* isolates gathered during a 16 days nosocomial outbreak occurred in April 2018, in the Papa Giovanni XXIII hospital (Bergamo) (For details see Supplementary Tables S3 and S4). The validation collection includes 54 *K. pneumoniae* isolates belonging to six of the most epidemiologically relevant clades[11,12]:

- 17 strains belonging to ST307.
- 15 strains belonging to ST258 Clade 2, including ST258 *wzi_154* and ST512 *wzi_154*.
- seven strains belonging to ST258 Clade 1 (ST258 *wzi_29*).
- 11 strains belonging to ST101.
- two strains belonging to ST11.
- two strains belonging to ST15.

45/54 strains were isolated at San Raffaele hospital (Milan), 2/54 strains were isolated at Papa Giovanni XXIII hospital (Bergamo), 4/54 retrieved from the *K. pneumoniae* collection of Gona and colleagues[15] 3/54 retrieved from the *K. pneumoniae* collection of Gaiarsa and colleagues[14].

Neither ethics committee approval, nor informed consent were required as all collected data are fully anonymized, there was no contact with patients and/or their families and no interventions or changes to treatment and management were made, in accordance with institutional guidelines.

DNA extraction and whole-genome sequencing. The genomic DNA of the 45 strains isolated from San Raffaele hospital (Milan) were extracted using Maxwell 16 Cell DNA purification kit. The extracted DNA was sequenced using the NextSeq. 500 platform with 2×150 bp paired-ends runs, after Nextera XT library preparation.

The genomic DNA of the 13 strains isolated from Papa Giovanni XXIII hospital (Bergamo) was extracted using the DNeasy blood and tissue kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. The extracted DNA was sequenced using the Illumina Miseq platform with a 2×250 bp paired-end run, after Nextera XT library preparation.

The genomic DNA of the 12 strains previously sequenced by Gaiarsa and colleagues[14] were extracted using QIAsymphony Virus/Pathogen minikit, version 1 (Qiagen, Hilden, Germany) with the automated instrument QIAsimphony (Qiagen, Hilden, Germany) according to manufacturer's instructions.

The genomic DNA of the 12 strains previously sequenced by Gona and colleagues[15] were extracted using the DNeasy blood and tissue kit (Qiagen, Hilden, Germany) following the manufacturer's instructions.

Details about strains hospital isolation are reported in Supplementary Table S3.

High resolution melting primer design using EasyPrimer. The EasyPrimer tool was developed for the identification of the most suitable regions for primer design in HRM and, more in general, in pan-PCR experiments. Briefly, the tool starts from gene sequences in multi-fasta format. The sequences are considered as not aligned by default and they are aligned by Muscle software[22] as the first step of the pipeline (see Supplementary Fig. S7 and Supplementary Note S1). The user can also decide to submit aligned sequences (in multi-fasta format) and skip the alignment step. EasyPrimer evaluates the amount of genetic variation for each position of the alignment and identifies the most reliable regions for primer design. EasyPrimer flags as good candidates for primer design two conserved regions flanking a highly variable one (taking into consideration, in advance, the optimal lengths of primers and amplicon). The user can decide either to evaluate the variability of the amplicon considering HRM-detectable SNPs only (the best option for HRM primer design) or all the SNPs (the best setting for pan-PCR experiments). A detailed description of the algorithm is reported in the Supplementary Note S1.

To develop an HRM-based protocol for *K. pneumoniae* typing, we focused on the seven MLST genes and on the hypervariable capsular gene *wzi*[23]. The HRM primer pairs for five out of the seven *K. pneumoniae* MLST genes were already available in literature[7] (*infB, mdh, phoE, rpoB* and two pairs on *tonB*). For the remaining two MLST genes (*pgi* and *gapA*) and for the *wzi* capsular gene (two primer pairs) the primers were designed using EasyPrimer. For each gene, the sequences were downloaded from the BigsDB database (https://bigsdb.pasteur.fr, 218 alleles for *pgi*, 183 for *gapA* and 563 for *wzi*), EasyPrimer was run and primer pairs were designed on the basis of its output.

High-resolution melting assays. We performed HRM assays using the genomic DNA extracted from each of the 82 *K. pneumoniae* strains included in this work. On the strains of the background and outbreak collections we used each of the ten primer pairs mentioned above. On the validation collection strains we used *wzi-3*

and *wzi-4* primer pairs only. HRM analyses were performed on the BioRad CFX Connect real-time PCR System (BioRad, Hercules, California). Each 10 µl reaction contained: 5 µl of 2x SsoAdvanced Universal SYBR® Green Supermix (BioRad, Hercules, California), 0.4 µl of each primer (0.4 µM) and 1 µl of template DNA (25–50 ng/µl). The thermal profile was as follows: 98 °C for 2 min, 40 cycles of [95 °C for 7 s, 61 °C for 7 s, and 72 °C for 15 s], 95 °C for 2 min, followed by HRM ramping from 70–95 °C with fluorescence data acquisition at 0.5 °C increments. Three technical replicates were performed for each strain and for each gene analysed. Negative controls were added in every run and for each gene.

Comparison of the HRM primer pairs and schemes. We compared the discriminatory power of the ten HRM primer pairs on the 17 background collection strains. For each primer pair we calculated the average melting temperatures (aTms) of the three replicates for each strain and the relative strain distance matrix based on the obtained aTms. Thus, we compared the discriminatory power of the different primer pairs by comparing the relative distance matrix values using Wilcoxon test with Holm post-hoc correction.

Furthermore, we grouped the primer pairs in three schemes (MLST6, MLST8 and wzi) and we compared the relative strain distance matrices using Wilcoxon test with Holm post-hoc correction. The scheme compositions were as follows: the MLST6 scheme included the six primer pairs proposed by Andersson and colleagues[7] for five MLST genes (with two primer pairs for *tonB*); the MLST8 included all the MLST6 primer pairs, the primers for *pgi* and *gapA* (newly designed in this work using the EasyPrimer tool); the wzi scheme included the two primer pairs for the *wzi* gene (newly designed in this work). For details see Table 1 and Fig. 2.

Then, we compared the discriminatory power of MLST8 and wzi schemes by subtracting the relative distance matrices (wzi – MLST8) and studying the obtained matrix with a heatmap.

All these analyses were performed using R (https://www.r-project.org/) and the R libraries Ape and Gplots.

HRM-based outbreak reconstruction. From the aTms of the outbreak and background collections we calculated the distance matrices for MLST6, MLST8 and wzi primer schemes (for more details see above) and clustered the strains using the hierarchical clustering method implemented in the Hclust function in R.

Repeatability of the *wzi* HRM protocol. To test the repeatability of the *wzi* HRM typing protocol we analysed the wzi scheme aTms for all the strains of the three collections (17 background strains, 11 outbreak and 54 validation, see above). This allowed to compare the *wzi-3* and *wzi-4* aTms of multiple strains for each of the most epidemiologically relevant clades[11,12] (eight ST258 clade 1, 26 ST258 Clade 2, 19 ST307, 12 ST101, two ST11 and three ST15 strains) (see Table 2 and Supplementary Table S3).

We clustered the strains on the basis of wzi scheme aTms. Given the 0.5 °C sensitivity of the machine, we considered the strains with differences both in *wzi-3* and *wzi-4* aTms < 0.5 as indistinguishable. Thus, we built a strain-to-strain network, in which the indistinguishable strains pairs were connected. Clusters were extracted from the network using the decompose igraph R function (https://www.r-project.org/). Lastly, the strain-to-strain network was converted to a dendrogram and merged to *wzi* aTms in a heatmap plot, using R (https://www.r-project.org/).

WGS-based strain typing. We retrieved the genome assemblies of the 12 *K. pneumoniae* strains previously WGS-typed by Gaiarsa and colleagues[14] from NCBI database and the reads files of the 12 strains by Gona and colleagues[15] using fastq-dump tool (accession numbers are reported in Supplementary Table S3).

We performed *de novo* assembly on the reads obtained from the 58 strains sequenced in this work and on the reads of the 12 strains retrived from Gona and colleagues[15] using SPAdes software[24].

We retrieved the sequences of the *K. pneumoniae* MLST gene alleles and the relative scheme tables from the BigsDB database. Thus, we determined the MLST profiles using an in-house Blastn-based Perl script.

We retrieved the *wzi* allele sequences from BigsDB database and we annotated the *wzi* allele present in each of the 82 genome assemblies included in the study by Blastn search and manual curation of the results.

We annotated the K-type of the 82 strains using Kaptive[23] on the genome assemblies.

Core-SNP-based phylogenetic reconstruction on outbreak strains. We aligned the reads obtained from the 11 outbreak strains against the NTUH_K2044 reference genome (accession number NC_016845.1), and performed the SNPs calling following the GATK best practice procedure. We masked SNPs localized within repeated regions, identified using MUMmer[25], or prophages, identified using PhiSpy[26], and we called the core-SNPs among the strains using an in-house Python script. Thus, we subjected the core-SNPs alignment to phylogenetic analysis as follows: the best evolutionary model was assessed by ModelTest-NG and phylogenetic reconstruction was performed using the selected best model, with RAxML8 software[27]. We evaluated the core-SNPs distances among the strains using the R Ape library (https://www.r-project.org/).

Data availability

We deposited all Illumina sequence data from the 58 strains in NCBI's Short Read Archive under BioProject ID ERP119329 and all Illumina data were deposited under BioProject ID PRJEB36171.

References

1. Nutman, A. & Marchaim, D. How to: molecular investigation of a hospital outbreak. *Clinical Microbiology and Infection* **25**, 688–695 (2019).
2. Mehta, Y. *et al*. Guidelines for prevention of hospital acquired infections. *Indian J. Crit. Care Med.* **18**, 149–163 (2014).
3. Naze, F. *et al*. *Pseudomonas aeruginosa* Outbreak Linked to Mineral Water Bottles in a Neonatal Intensive Care Unit: Fast Typing by Use of High-Resolution Melting Analysis of a Variable-Number Tandem-Repeat Locus. *Journal of Clinical Microbiology* **48**, 3146–3152 (2010).
4. Mayerhofer, B. *et al*. Improved protocol for rapid identification of certain spa types using high resolution melting curve analysis. *PLoS One* **10**, e0116713 (2015).
5. Tamburro, M. & Ripabelli, G. High Resolution Melting as a rapid, reliable, accurate and cost-effective emerging tool for genotyping pathogenic bacteria and enhancing molecular epidemiological surveillance: a comprehensive review of the literature. *Ann. Ig.* **29**, 293–316 (2017).
6. Wong, Y. P., Chua, K. H. & Thong, K. L. One-step species-specific high resolution melting analysis for nosocomial bacteria detection. *J. Microbiol. Methods* **107**, 133–137 (2014).
7. Andersson, P., Tong, S. Y. C., Bell, J. M., Turnidge, J. D. & Giffard, P. M. Minim Typing – A Rapid and Low Cost MLST Based Typing Tool for *Klebsiella pneumoniae*. *PLoS ONE* **7**, e33530 (2012).
8. Price, E. P., Inman-Bamber, J., Thiruvenkataswamy, V., Huygens, F. & Giffard, P. M. Computer-aided identification of polymorphism sets diagnostic for groups of bacterial and viral genetic variants. *BMC Bioinformatics* **8**, 278 (2007).
9. Klaas, I. C. & Zadoks, R. N. An update on environmental mastitis: Challenging perceptions. *Transbound. Emerg. Dis.* **65**(Suppl 1), 166–185 (2018).
10. Perini, M. *et al*. Fast Klebsiella pneumoniae typing for outbreak reconstruction: an highly discriminatory HRM protocol on *wzi* capsular gene developed using EasyPrimer tool. *bioRxiv* (2019).
11. the EuSCAPE Working Group. *et al*. Epidemic of carbapenem-resistant *Klebsiella pneumoniae* in Europe is driven by nosocomial spread. *Nat Microbiol* **4**, 1919–1929 (2019).
12. Wyres, K. L. *et al*. Emergence and rapid global dissemination of CTX-M-15-associated *Klebsiella pneumoniae* strain ST307. *J. Antimicrob. Chemother.* **74**, 577–581 (2019).
13. Deleo, F. R. *et al*. Molecular dissection of the evolution of carbapenem-resistant multilocus sequence type 258 *Klebsiella pneumoniae*. *Proc. Natl. Acad. Sci. USA* **111**, 4988–4993 (2014).
14. Gaiarsa, S. *et al*. Genomic epidemiology of *Klebsiella pneumoniae* in Italy and novel insights into the origin and global evolution of its resistance to carbapenem antibiotics. *Antimicrob. Agents Chemother.* **59**, 389–396 (2015).
15. Gona, F. *et al*. Comparison of core genome MLST, coreSNP and PFGE methods for *Klebsiella pneumoniae* cluster analysis. (Paper under revision on Microbial Genomics)
16. Jeremiah, C. J. *et al*. Differing epidemiology of two major healthcare-associated meticillin-resistant *Staphylococcus aureus* clones. *J. Hosp. Infect.* **92**, 183–190 (2016).
17. Larssen, K. W., Nor, A. & Bergh, K. Rapid discrimination of *Staphylococcus epidermidis* genotypes in a routine clinical microbiological laboratory using single nucleotide polymorphisms in housekeeping genes. *Journal of Medical Microbiology* **67**, 169–182 (2018).
18. Chan, M. S., Maiden, M. C. & Spratt, B. G. Database-driven multi locus sequence typing (MLST) of bacterial pathogens. *Bioinformatics* **17**, 1077–1083 (2001).
19. Comandatore, F. *et al*. Gene Composition as a Potential Barrier to Large Recombinations in the Bacterial Pathogen *Klebsiella pneumoniae*. *Genome Biol. Evol.* **11**, 3240–3251 (2019).
20. Qi, Y. *et al*. ST11, the dominant clone of KPC-producing *Klebsiella pneumoniae* in China. *J. Antimicrob. Chemother.* **66**, 307–312 (2011).
21. Roe, C. C., Vazquez, A. J., Esposito, E. P., Zarrilli, R. & Sahl, J. W. Diversity, Virulence, and Antimicrobial Resistance in Isolates From the Newly Emerging ST101 Lineage. *Front. Microbiol.* **10**, 542 (2019).
22. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792–1797 (2004).
23. Wyres, K. L. *et al*. Identification of capsule synthesis loci from whole genome data. *Microb Genom* **2**, e000102 (2016).
24. Bankevich, A. *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
25. Delcher, A. L. *et al*. Alignment of whole genomes. *Nucleic Acids Research* **27**, 2369–2376 (1999).
26. Akhter, S., Aziz, R. K. & Edwards, R. A. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res.* **40**, e126 (2012).
27. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

Acknowledgements

Author contributions

M.P. developed the tool, performed the H.R.M. experiments and drafted the paper. A.P. performed the H.R.M. experiments and revised the manuscript. S.P. designed the primers and revised the manuscript. D.D.C. implemented the tool online. M.C., F.G., F.V. collected the samples and extracted the D.N.A. P.M., D.M.C., C.F. collected the samples. G.V.Z. wrote the paper. F.C. conceived and designed the experiments and wrote the paper. All authors read, revised and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-57742-z.

Correspondence and requests for materials should be addressed to F.C.

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.