Gene Composition as a Potential Barrier to Large Recombinations in the Bacterial Pathogen *Klebsiella pneumoniae*

Francesco Comandatore[1], Davide Sassera[2], Sion C. Bayliss[3], Erika Scaltriti[4], Stefano Gaiarsa[5], Xiaoli Cao[6], Ana C. Gales[7], Ryoichi Saito[8], Stefano Pongolini[4], Sylvain Brisse[9], Edward J. Feil[3], and Claudio Bandi[10,*]

[1]Sky Net UNIMI Platform - Pediatric Clinical Research Center Romeo ed Enrica Invernizzi, Dipartimento di Scienze Biomediche e Cliniche Luigi Sacco, Università degli Studi di Milano, Italy

[2]Department of Biology and Biotechnology L. Spallanzani, Universita' degli Studi di Pavia, Italy

[3]The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, United Kingdom

[4]Risk Analysis and Genomic Epidemiology Unit, Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia-Romagna, Parma, Italy

[5]U.O.C. Microbiologia e Virologia, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy

[6]Department of Laboratory Medicine, Nanjing Drum Tower Hospital, The Affiliated Hospital of Nanjing University Medical School, Nanjing, People's Republic of China

[7]Laboratório ALERTA, Division of Infectious Diseases, Department of Internal Medicine, Escola Paulista de Medicina - EPM, Universidade Federal de São Paulo (UNIFESP), Brazil

[8]Department of Molecular Microbiology, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan

[9]Institut Pasteur, Biodiversity and Epidemiology of Bacterial Pathogens, Paris, France

[10]Sky Net UNIMI Platform - Pediatric Clinical Research Center Romeo ed Enrica Invernizzi, Dipartimento di Bioscienze, Università degli Studi di Milano, Italy

*Corresponding author: E-mail: claudio.bandi@unimi.it.

Accepted: October 25, 2019

Data deposition: Thirteen novel *Klebsiella pneumoniae* genomes are produced in this work. The accession numbers are listed in supplementary table S2, Supplementary Material online. This project has been deposited at NCBI bioproject PRJEB25631.

Abstract

Klebsiella pneumoniae (Kp) is one of the most important nosocomial pathogens worldwide, able to cause multiorgan infections and hospital outbreaks. One of the most widely disseminated lineage of Kp is the clonal group 258 (CG258), which includes the highly resistant "high-risk" sequence types ST258 and ST11. Genomic investigations revealed that very large recombination events have occurred during the emergence of Kp lineages. A striking example is provided by ST258, which has undergone a recombination event that replaced over 1 Mb of the genome with DNA from an unrelated Kp donor. Although several examples of this phenomenon have been documented in Kp and other bacterial species, the significance of these very large recombination events for the emergence of either hypervirulent or resistant clones remains unclear. Here, we present an analysis of 834 Kp genomes that provides data on the frequency of these very large recombination events (defined as those involving >100 kb), their distribution within the genome, and the dynamics of gene flow within the Kp population. We note that very large recombination events occur frequently, and in multiple lineages, and that the majority of recombinational exchanges are clustered within two overlapping genomic regions, which have been involved by recombination events with different frequencies. Our results also indicate that certain lineages are more likely to act as donors to CG258. Furthermore, comparison of gene content in CG258 and non-CG258 strains agrees with this pattern, suggesting that the success of a large recombination depends on gene composition in the exchanged genomic portion.

Key words: *Klebsiella pneumoniae*, large recombination, clonal group 258.

Introduction

Recent genomic studies revealed that some bacterial species are able to exchange large (>5 kb) or even very large (>100 kb, up to over 1 Mb) genomic portions (Croucher and Klugman 2014) causing sudden and extended changes into the recipient genome, for example, the acquisition and/or loss of several genes and multiple nucleotide variations (Hanage 2016). In particular, genomic studies published in 2014 and 2015 described large and very large recombination events (of up to 20% of the entire genome) involving epidemiologically relevant strains of *Klebsiella pneumoniae* (Kp), an important nosocomial pathogen (Chen et al. 2014; Deleo et al. 2014; Bowers et al. 2015; Gaiarsa et al. 2015; Wyres et al. 2015).

Kp is a Gram-negative bacterium and member of the *Enterobacteriaceae* family. This species has a diverse ecology: In addition to being a common colonizer of the guts of humans and other mammals (including livestock), it is also associated with invertebrates, plants, and multiple niches in the environment including soil and water (Struve and Krogfelt 2004; Holt et al. 2015). Kp has been divided into three major phylogroups, named KpI, KpII, and KpIII. These three phylogroups are now recognized as three different species, that is, *K. pneumoniae* sensu stricto (KpI), *K. quasipneumoniae* (KpII), and *K. variicola* (KpIII) (Rosenblueth et al. 2004; Brisse et al. 2014). For simplicity, we will still refer to them as KpI–KpIII. KpI strains are often isolated from hospitalized human patients, whereas KpII strains are frequently associated with healthy carriers, and KpIII strains are mainly found in the environment or in association with other mammals or plants (Holt et al. 2015). Kp can behave as an opportunistic pathogen, especially in immuno-compromised hosts, causing multi-organ infections and sepsis. These infections are particularly difficult to treat when caused by multidrug-resistant strains, which are common as Kp is able to acquire resistance to most antibiotic classes, including extended spectrum beta-lactams and carbapenems. Several distinct multidrug-resistant Kp clones have been isolated in hospitals worldwide, making this bacterium a major public healthcare burden. Kp strains harboring blaKPC, a plasmid gene encoding for the *Klebsiella pneumoniae* carbapenemase (KPC), pose a particularly high risk to public health, and recently a Kp strain that is resistant to all 26 antibiotics licensed in the United States has been isolated (Pitout et al. 2015; Chen et al. 2017). The most widespread KPC producers are isolates from clonal group 258 (CG258), which is known to have spread throughout the Americas, Europe, Asia, and elsewhere (Cantòn et al. 2012; Lascols et al. 2013; Cao et al. 2014; Saito et al. 2014). CG258 belongs to KpI and includes sequence types reported as cause of nosocomial outbreaks, such as ST258, ST11, ST512, and ST340 (Bialek-Davenet et al. 2014).

Whole genome sequencing based studies allowed to describe three large recombination events that occurred within the CG258, each spanning at least 100 kb and up to 1.5 Mb (Chen et al. 2014; Deleo et al. 2014; Gaiarsa et al. 2015; Wyres et al. 2015). ST258 emerged after a >1 Mb recombination between an ST11-like recipient and an ST442-like donor (Chen et al. 2014). However, it is currently unclear to what extent these large recombination events are actually associated with the epidemiological success of the clinically important CG258 clones (Bowers et al. 2015). More broadly, questions remain concerning the patterns of gene flow within the Kp population, and whether certain lineages are more or less likely to act as either donors or recipients. Although the data are currently limited, Holt et al. (2015) observed that large recombinations tend to be less common between unrelated Kp phylogroups and argued that gene flow in the Kp population is likely to be structured by biological and/or ecological barriers (Holt et al. 2015). In order to investigate the large recombination phenomenon in Kp, we analyzed over 800 genomes from KpI, KpII, and KpIII, performing recombination analysis, donor identification, and gene presence/absence analysis. We show that large recombination events in Kp CG258 led to the emergence of several epidemiologically relevant lineages and provide evidence that suggests that donor gene composition may affect the successfulness of the hybrid strains.

Materials and Methods

Genome Sequencing

Thirteen Kp hospital isolates were obtained from Brazil (8 isolates), China (3 isolates), Japan (1 isolate), and France (1 isolate), based on their MultiLocus Sequence Typing (MLST) profiles: 12 ST11 and 1 ST442. DNA was extracted using a QIAamp DNA mini-kit (Qiagen) following the manufacturer's instructions. Whole genomic DNA was sequenced using an Illumina Miseq platform with a 2 by 250 paired-end run after Nextera XT paired-end library preparation. Paired-end genomic reads were assembled using MIRA 4.0 software (Chevreux et al. 1999).

Reconstruction of Kp Species and CG258 Representative Databases

Six hundred and sixty-three genome assemblies and 158 genome reads data sets were retrieved from the NCBI and Patric databases, for a total of 821 strains collecting of the entire known KpI, KpII, and KpIII genomics variability at May 2015 (see details on supplementary table S1, Supplementary Material online). The downloaded reads were assembled using Velvet software (Zerbino 2010) and the assemblies were then merged in a 834-genome data set. Single nucleotide polymorphisms (SNPs) calling was performed, as described below, using the 30660/NJST258_1 strain genome as reference: It was chosen because it belongs to the CG258 and it has been already used by Deleo et al. (2014) and Chen et al. (2014) as reference in the works in which they describe the

large recombination events that originated the emergence of the ST258 (Chen et al. 2014) and its division in two subclades (Deleo et al. 2014). Each of the 834 genomes was aligned against the reference genome using progressiveMauve (Darling et al. 2004) and the multigenome alignment and the core SNP alignments were obtained using the in-house pipeline described by Gaiarsa et al. (2015). The core SNP alignment was then subjected to phylogenetic analysis using RAxML version 8.0.0 (Stamatakis 2014) using the ASC_GTRGAMMA model and 100 bootstrap replicates.

In order to remove oversampled clones and clades and to obtain a data set of manageable size while maintaining the information of the entire genomic variability of the species, a Kp species representative genome database (from now on referred to as "species database") was constructed using the following procedure: 1) SNPs' distance matrix among the strains was obtained using the R library Ape (Popescu et al. 2012; R Development Core Team 2016) and 2) a recursive process of strain selection was performed, removing strains with less than five SNPs distance from others. The strains belonging to CG258 were then manually extracted from the species database, thus obtaining a representative selection of CG258 strains (from now "CG258 database").

Recombination Analysis

The CG258 database and two outgroup strains (18PV and K102An) were subjected to reference-based genome alignment, SNP calling, and phylogenetic analysis, as above. The obtained tree was rooted on the outgroups and the outgroups were then removed to obtain a representative CG258 tree. Recombination analysis was performed using the ClonalFrameML software (Didelot and Wilson 2015), setting the transition/transversion rate as calculated by PhyML (Guindon et al. 2009). Recombinations sized over 100 kb were clustered on the basis of their localization on the genomic alignment. More in detail, the clustering analysis was computed as follows: 1) the squared distance matrix of the recombinations was generated computing Manhattan distance using the start and the end positions of the recombinations (i.e., given the recombinations "rec1" and "rec2" it was calculated as $|start_{rec1} - start_{rec2}| + |end_{rec1} - end_{rec2}|$) and 2) the recombinations were then clustered on the basis of the distance matrix using the hierarchical clustering with *P* value support algorithm implemented in the Pvclust R library (Suzuki and Shimodaira 2016). The highly supported clusters were identified setting the approximately unbiased index threshold at 0.99. The analyses were performed using R (R Development Core Team 2016).

Identification of the Donors of the Recombinations

In order to identify the Kp donors of the recombinations, we performed an ad hoc phylogenetic analysis. The genomes of the species database were aligned to the reference genome and core SNPs were called, subjected to ML phylogeny using FastTree software (Price et al. 2009) with 100 bootstraps (we will refer to the obtained tree as "species database tree"). For each recombination, the core SNPs called within the recombined region were extracted and subjected to phylogenetic analysis, using FastTree software (Price et al. 2009) with 100 bootstraps. Each resulting tree was manually analyzed as follows: 1) the CG258 recipient(s) of the recombination was (were) identified on the tree and 2) when a highly supported (>75 bootstrap) monophylum including all the recipients and one or more non-CG258 strains was detected, the non-CG258 strains were considered as donors of the recombination.

Recombinations identified on the same branch of the CG258 tree, localized on adjacent genomic regions, and sharing the same donors, were merged, as likely originated from a single recombination event. For these recombinations, the donor identification procedure was repeated considering a novel recombined region, ranging from the beginning of the first recombination to the end of the second one.

When more than one strain was identified as donor of a recombination, the reliability of the identified donors was assessed by testing if they belong to a monophyletic clade on the smaller global tree. Indeed, in this case, we argued that the donor of the recombination was the common ancestor of the clade.

Identification of Prophages and Repeated Regions on the Reference Genome

Prophages were searched on the reference genome assembly using on line Phast tool (Zhou et al. 2011). Repeated genomic regions were searched using MUMer tool (Delcher et al. 2002) and their genomic localizations were studied using R. Repeated sequences were then annotated by BlastN search against RefSeq (O'Leary et al. 2016) database and ISFinder (Siguier et al. 2006).

Gene Presence/Absence Analysis

The CG258 belongs to the phylogroup KpI. To study the relationship between gene composition and large recombinations in CG258, we decided to consider a genome data set including strains belonging to the phylogroups KpI, KpII, and KpIII. The three phylogroups are phylogenetically related but genetically isolated (Holt et al. 2015) and we included the KpII and KpIII strains to investigate if a different gene composition can be one of the causes for this isolation.

The 834 genomes included into the global genome database were subjected to Open Reading Frame calling using Prodigal (Hyatt et al. 2010), and then to orthology analysis using Roary software (Page et al. 2015) after annotation with PROKKA (Seemann 2014). The obtained gene presence/absence matrix was then analyzed as follows. We split the matrix into two submatrices: the first one including CG258 strains

only and other one non-CG258 ones. From each submatrix, we classified the genes in the categories "common" (present in >=95% of the strains), "accessory" (<95% and >=5%), and "rare" (<5%). Then, we classified the CG258 common genes as "common-common" if classified as "common" among non-CG258 strains, "common-accessory" if classified as "accessory," and "common-rare" if "rare." The positions of the CG258 "common," "accessory," and "rare" genes present on the reference genome were obtained merging the information from the Roary output and the reference annotation file, using an *in-house* Perl script. As explained above, the non-CG258 data set contains strains which belong to KpI, KpII, and KpIII phylogroups. In order to evaluate the effect of KpII and KpIII strains on the identification of "common-common," "common-accessory," and "common-rare" genes, we repeated the gene classification excluding KpII and KpIII strains. We will refer to the non-CG258 data set which includes KpI, KpII, and KpIII as "ALL non-CG258" and to the non-CG258 data set including KpI strains only as "KpI non-CG258."

For each of the CG258 strains, the positions of "rare" genes absent in the reference genome were inferred also as follows: 1) the list of the genes shared between the strain and the reference were obtained from the gene presence/absence matrix; 2) the positions of the shared genes on the strain genome (including the contig name) and on the reference genome were obtained from PROKKA outputs (see above); 3) for each rare gene present on the strain genome, the closest shared gene on the same contig was identified and the distance between the two genes was computed; and 4) the position of the identified shared gene on the reference genome was then computed using the information retrieved at the point 2 and 3. Once all the strains had been analyzed, the position of each rare gene on the reference genome was estimated as the average of the positions obtained from all the strains in which the gene was present.

For each gene category, the Kernel density of the genes positions along the reference genome was plotted using R (R Development Core Team 2016). In Kernel density plots, the area under the curve between two *x* values represents the probability to find a value within that range.

For the genes of the categories "common" and "accessory" (among the CG258 strains), the distances from the origin of replication (ORI) on the reference genome were computed using an *in-house* Perl script and the median distances were compared with Wilcoxon test, using R. Furthermore, the preferential localization near the ORI of common and accessory genes was assessed using a bootstrapping approach, as follows: 1) the number of genes (*N*) classified in the category and their median distance (Dist-cat) from the ORI were obtained, 2) *N* genes were randomly selected from the reference genome and their median distance from the ORI (Dist-step) was computed, 3) the step 2 was repeated for 10,000 generations, and 4) the number of generations for which Dist-cat < Dist-step and number for which Dist-cat > Dist-step was obtained.

Finally, we compared gene composition of Kp lineages and recombined genomic regions using Wilcoxon and chi-square tests using R (R Development Core Team 2016).

Results

The Newly Sequenced Genomes

The genomes of 12 ST11 isolates and 1 ST442 isolate, collected from France, Brazil, China, and Japan between 1997 and 2014, were sequenced and assembled (supplementary table S2, Supplementary Material online). These 13 novel genomes were added to a collection of 821 genomes, to represent the genomic diversity of Kp (supplementary table S1, Supplementary Material online). A SNP-based phylogeny of this sample was consistent with previous studies (Holt et al. 2015; supplementary fig. S1, Supplementary Material online). The vast majority of the isolates corresponded to a large radial expansion within KpI, with KpII and KpIII clearly distinct at the end of long branches.

The 834-genome data set represents the diversity within the publicly available data sets, which present a clear bias toward clinical isolates. In order to correct this bias, that could determine on overrepresentation of specific clusters of isolates, we subsampled 394 Kp representative genomes based on pairwise SNP distances, such that no two genomes sharing fewer than five SNP differences were kept (see Materials and Methods). A subset of 60 CG258 strains was then extracted from the 394-genome data set. Core SNPs were called independently for the global (394 strains) and CG258 (60 strains) data sets, and these SNPs were used for phylogenetic reconstruction using maximum likelihood (figs. 1 and 2). A total of 44,276 core SNPs were obtained from the 834-genome data set, whereas 65,884 from the 394 strains data set and 65,884 from the 60 strains data set.

Patterns of Recombination in CG258

Recombination was detected within the 60 CG258 genomes data set using ClonalFrameML (Didelot and Wilson 2015). A total of 119 recombination events were detected, 65 on internal nodes and 54 on terminal branches of the tree (fig. 1). All the genomes were shown to contain signals of recombinations (fig. 1) with the exception of three ST11 strains: US-MD-2006, JM45, and CHS_24 (fig. 1). Thus, these three genomes could contain more genomic features similar to the ST11 ancestor in comparison to the other ST11 strains in the data set (for this reason, we will refer to these three strains as "ST11-ancestor-like").

The 119 recombination events encompassed 63% ($n = 3,344,516$) of the 5,263,229 sites in the genomic alignment. More in detail, 2,195,715 of these positions were involved in one recombination event only, whereas 391,724

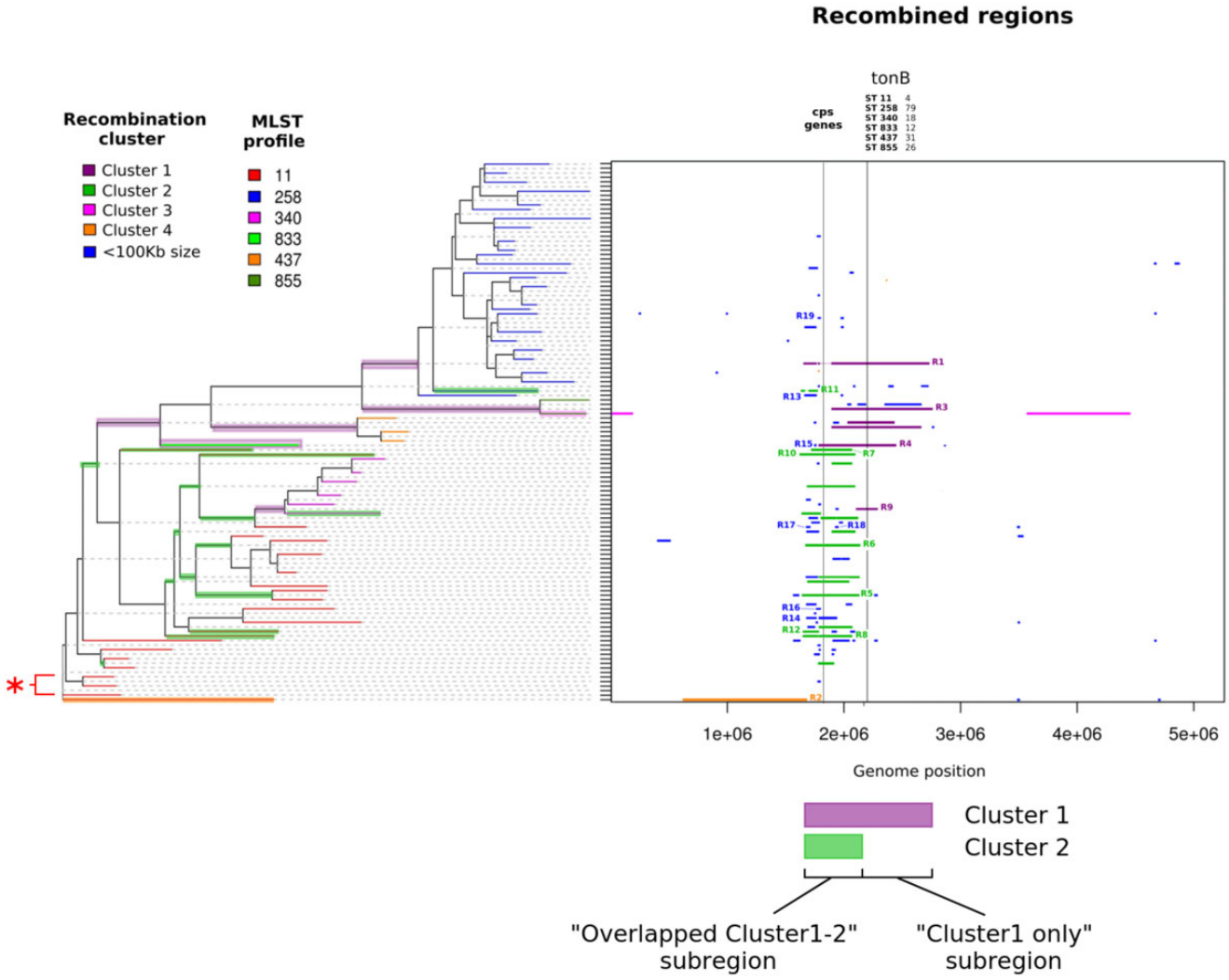

Fig. 1.—Recombination analysis, performed with ClonalFrameML. On the left, the phylogenetic tree obtained with RaxML with branches colored on the basis of the MLST profiles. Branches relative to large recombinations are highlighted with the color of the clusters identified by Pvclust. On the right, the plot of the recombined regions, colored on the basis of the involved genomic regions as grouped by Pvclust (see Materials and Methods). The positions on tonB gene and cps operon are reported by vertical lines. On the bottom, the positions of Cluster 1 and Cluster 2 genomic regions are graphically represented by colored horizontal bars (violet and green, respectively). Finally, the positions of the "Overlapped Cluster 1-2" and "Cluster 1 only" subregions are indicated on the graph.

were involved at least 10 different events. Thirty of the 119 (25%) recombinations were sized >100 kb (~2% of the reference genome size) and, among them, six have >500 kb size (~10% of the reference genome) and two >1 Mb (~20% of the genome). Thus, these initial observations confirmed that large recombination events have occurred relatively frequently within CG258.

Figure 1 clearly illustrates that the recombination events are not randomly distributed across the genome but are highly clustered. Ninety-five of the 119 (80%) total predicted recombination events occurred in a 1,185,000 sized region (23% of the genome), between positions 1,575,000 and 2,760,000 of the reference genome, and the same region contains 27 of the 30 large recombinations (>100 kb) localized between the positions 1,660,631 and 2,750,819 (from here, this region will be called "highly recombined region"). In fact, bootstrap-based hierarchical clustering analysis reveals two highly supported clusters corresponding to two partially overlapped genomic regions (supplementary fig. S2, Supplementary Material online), as shown by green and purple lines in figure 1. Cluster 1 is composed of 7 recombination events spanning the region from 1,660,631 to 2,750,819 (1,090,188 bp), whereas cluster 2 includes 20 recombination events from 1,629,115 to 2,131,208 (502,93 bp). Thus, we divided the highly recombined region in two subregions: the "overlapped Cluster1-2" subregion, from 1,629,115 to

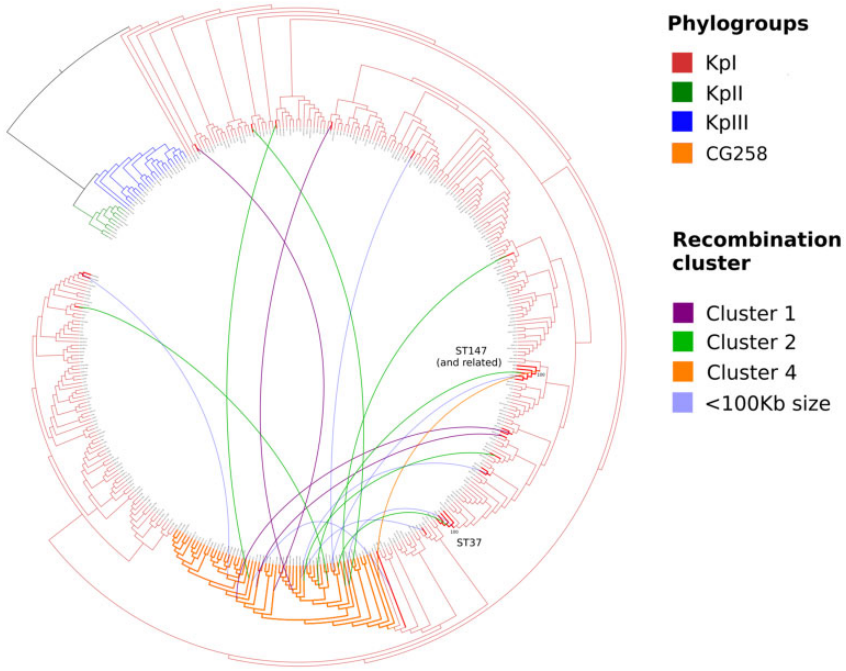

Fig. 2.—RaxML tree of 394 *Klebsiella pneumoniae* strains, selected from the global genome database to be representative of the genetic variability of the species. The branches of the tree are colored on the basis of the phylogroup: red for KpI, green for KpII, and blue for KpIII. The CG258 clade, on the bottom, is highlighted in red. The edges connect donors and recipients as identified in this work. The edge color corresponds to the recombination cluster (see Materials and Methods). The branches relative to the CG258 clade (the recipient) are colored in orange.

2,131,208, involving both "Cluster1" and "Cluster2" recombinations, and the "Cluster1 only," from 2,131,209 to 2,750,819, involving only "Cluster1" recombinations (see fig. 1).

Identification of Recombination Donors

In order to identify the donors of the recombination events detected within CG258, we used a phylogenetic approach based on the core SNPs within each of the recombined regions. Using this method, we were able to robustly identify the donors in 19 recombination events, that will be referred to as "Recombination with Identified Donor" or RID1-RID19 (see supplementary table S3 and fig. S3, Supplementary Material online, for information about RID donors and recipients; the tree used to identify the donor in RID1 is shown as an example in supplementary fig. S4, Supplementary Material online). RID1 and RID19 correspond to previously identified donor/recipient pairs (Chen et al. 2014; Deleo et al. 2014), thus confirming the soundness of our approach.

In order to visualize the flow of large recombinations toward CG258, we plotted the global tree of Kp (including all the strains used for recombination analysis or donor identification) and connected donors and recipients with links (fig. 2). We found that only two highly supported lineages, ST147 and ST37, had been involved in multiple recombination events.

The Emergence of New Sequence Types by Large Recombinations

Our analysis on the 60 representative genomes of CG258 revealed that ST340, ST437, ST833, and ST855 have emerged from a ST11-like ancestor following the acquisition of genomic regions of at least 100 kb (fig. 1). All these STs share with ST11 the same MLST variant for all the MLST genes but tonB, which is located within the recombined region (fig. 1). Furthermore, for the recombinations that originated the ST340, ST833, and ST855, the donors were identified (supplementary table S3, Supplementary Material online) and they resulted to harbor the same tonB allele of the relative emerged ST.

Identification of Prophage Sequences and Repeated Regions in the Reference Genome

Phast tool identified eight prophage sequences in the reference genome assembly, five classified as intact, one as incomplete, and two as questionable (supplementary fig. S5, Supplementary Material online). None of the identified prophage sequences was localized within the highly recombined region. A total of 121 repeated genomic regions were identified using MUMmer, with an average size of 1,638 pb; 29 out of the 121 were localized within the highly recombined region, and 23 of them within the "overlapped Cluster1-2" region (see supplementary fig. S6, Supplementary Material

online). The repeated sequences localized within the highly recombined region were annotated by BlastN search against RefSeq and ISFinder databases. All the sequences but one did not give match. The only that give match was annotated as IS5 and was localized within the "overlapped Cluster1-2" region.

Gene Presence/Absence Analysis

We investigated whether divergent gene compositions between donors and recipients can represent a genetic barrier for large recombinations between non-CG258 (possible donors) and CG258 strains (possible recipients). We compared the genomic localizations of large recombinations and the positions of the genes commonly present among CG258 but less frequent among non-CG258 strains. More specifically, the 834 genomes included in the global data set were subjected to orthology analysis, and genes were classified as "common" (present in >=95% of the strains), "accessory" (<95% and >=5%), and "rare" (<5%), considering separately the CG258 and non-CG258 strains (classification of genes as "common" and "accessory" is according to Holt et al. [2015]). Subsequently, the CG258 "common" genes were divided into "common-common" if classified as "common" also among non-CG258 strains, as "common-accessory" if classified as "accessory" in non-CG258, and as "common-rare" if "rare" in non-CG258 strains. Considering the observed high frequency of recombination, we expected that gene copresence among recombination donors could introduce biases in the classification of CG258 "common" genes. Thus, we considered only CG258 "common" genes likely present into the ST11 ancestor, that is, those shared among the three ST11-ancestor-like strains described above.

Within CG258, a total of 2,453 common, 6,786 accessory, and 22,539 rare genes were identified. In order to study the genomic localization of these gene categories, we considered only the genes present in the reference genome (2,404/2,453 common, 2,738/6,786 accessory, and 61/22,523 rare genes). The median distance from the ORI on the reference genome resulted significantly lower in common genes than in accessory genes (Wilcoxon test, P value < 0.001). Furthermore, the bootstrapping analyses of the positions of the common genes and the accessory genes revealed a strong bias in their distribution: None of the 10,000 randomly generated gene data sets showed a median distance from the ORI lower than median distance observed for common genes, at the same manner none showed a median distance greater than that observed for accessory genes.

Among the 2,453 genes classified as common in CG258, 2,385 were present in all the ST11-ancestor-like strains. The presence/absence of these 2,453 common genes among the strains of the ALL non-CG258 data set (including KpI–KpIII strains) was used to classify 1,327 genes as "common-common," 977 as "common-accessory," and 81 as "common-rare." The analysis repeated on the KpI non-CG258 data set (including only KpI) identified 1,626 "common-common," 795 "common-accessory," and 32 "common-rare" genes. The genomic positions of the genes classified using the ALL non-CG258 data set present in the reference genome (1,309/1,397 common-common genes, 948/977 common-accessory, and 80/81 common-rare genes) and KpI non-CG258 data set (1,619/1,626 common-common, 789/795 common-accessory, and 32/32 common-rare) were then retrieved and compared with the positions of large recombinations.

The distributions of common-common, common-accessory, and common-rare genes was investigated either on the ALL non-CG258 or on the KpI non-CG258 data sets. The results for these two data sets are reported in Figure 3*b* and *c*. Both graphs show a similar pattern: Within "Cluster 1 only," that is, the less frequently recombined genomic region, common-common genes frequency reaches its minimum and common-accessory genes frequency show a local maximum (fig. 3*b* and *c*). Furthermore, only one common-rare gene was found within the highly recombined region (fig. 3*b* and *c*).

Out of the 948 common-accessory genes found in the ALL non-CG258 data set, 183 are localized within the highly recombined region. Non-CG258 strains show a variable pattern of presence of these genes (supplementary fig. S7, Supplementary Material online), and statistical analyses revealed that 1) strains involved as donors in multiple recombinations (ST147 and ST37) harbored significantly more of these genes that the other donor strains (Wilcoxon test, P value < 0.05, boxplot in supplementary fig. S8, Supplementary Material online); 2) within the highly recombined region, genes localized within the less frequently recombined "Cluster 1 only" subregion were harbored by significantly fewer non-CG258 strains in comparison to those localized within the more frequently recombined "overlapped Cluster1-2" subregion (Wilcoxon test, P value < 0.01—boxplot in supplementary fig. S9, Supplementary Material online). On the other hand, the heatmap representation of the 81 common-rare genes among the non-CG258 strains (supplementary fig. S10, Supplementary Material online) shows that these genes are particularly frequent in some lineages, but an evident pattern with donors is not detectable.

Common-accessory and common-rare genes present in the reference genome were then annotated against Clusters of Orthologous Groups (COG) and pie charts of COG pathways abundances were plotted (supplementary figs. S11 and S12, Supplementary Material online, respectively). Finally, COG pathways abundances of common-accessory genes localized inside and outside the highly recombined region were compared and no significant difference was found (chi-square test, P value > 0.05).

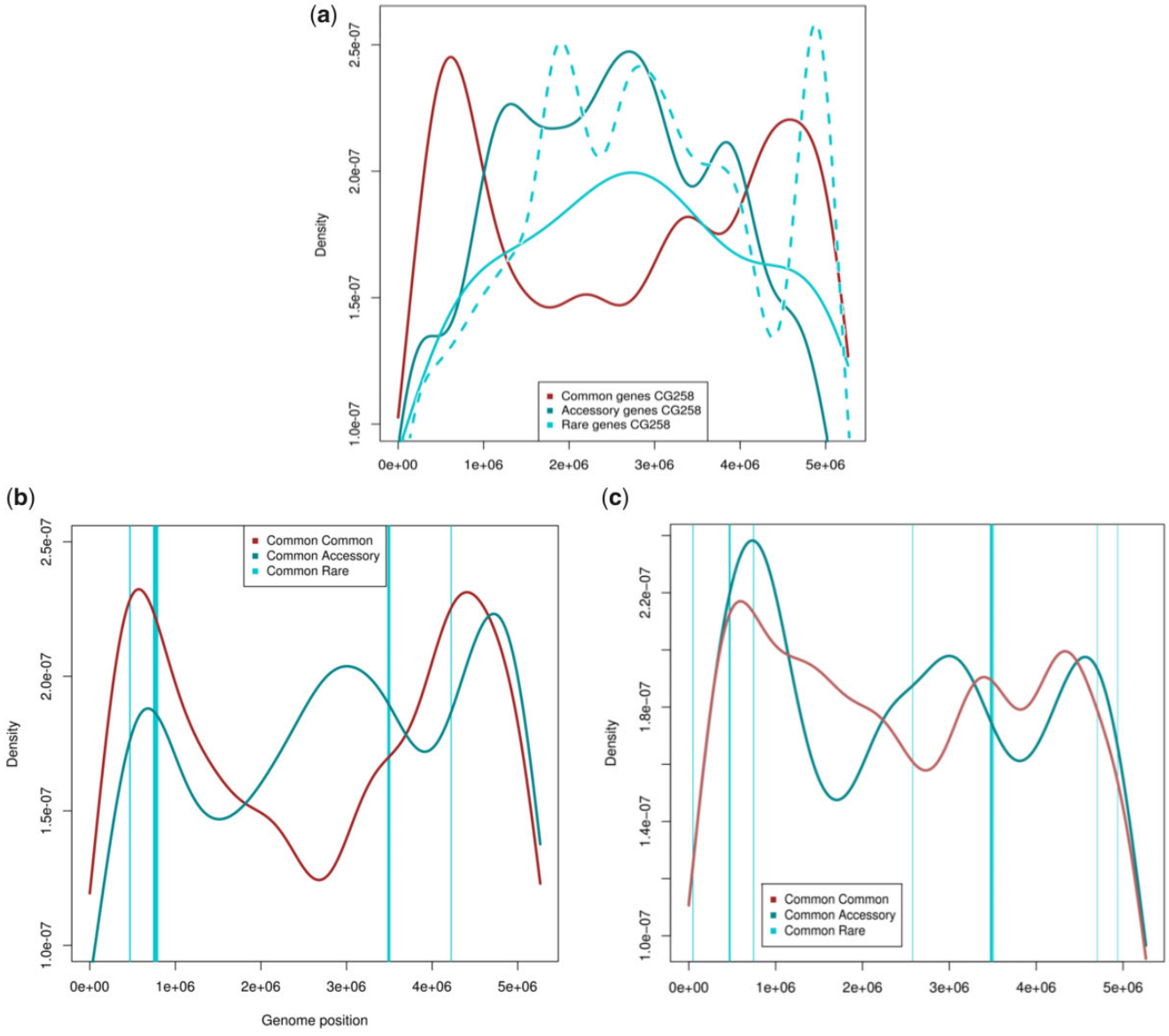

Fig. 3.—(*a*) Kernel density plot of the distribution of CG258 common (in red), accessory (in blue), and rare (in light blue) genes present of the reference. The Kernel density plot of the distribution of the rare genes absent in the reference and for which the positions were inferred (see Materials and Methods) are reported with a dotted light blue line. (*b*) Kernel density plot of the position on the reference of common-common (in red) and common-accessory (in blue) genes. The positions of the common-rare (in light blue) genes are reported with vertical lines. The classification was performed using a non-CG258 data set including strains which belong to KpI, KpII, and KpIII phylogroups (see Materials and Methods). (*c*) Kernel density plot of the position on the reference of common-common (in red) and common-accessory (in blue) genes. The positions of the common-rare (in light blue) genes are reported with vertical lines. The classification was performed using a non-CG258 data set including strains which belong to the KpI only phylogroup (see Materials and Methods).

Discussion

Whole genome sequencing has revealed an unprecedented degree of genome plasticity in Kp, both in terms of the rates of horizontal gene transfer affecting the pan-genome, and in terms of the rates of homologous recombinations in the core genome. Most strikingly, this species undergoes very large recombination events, affecting up to 20% of the genome (Chen et al. 2014; Deleo et al. 2014; Gaiarsa et al. 2015; Holt et al. 2015; Wyres et al. 2015). However, it remains unclear to what degree these large recombination events are responsible for the epidemiological success of lineages such as CG258, nor whether gene flow is in some way structured within the broader Kp population.

In order to investigate the relationship between gene composition and large recombinations in Kp, we performed a comparative genomic analysis on lineage CG258. In particular, we identified the donors and recipients of large recombinations and investigated their positioning in relation with

gene composition in the recipient genomes. This provided an overview of the large recombination phenomenon in the lineage, and thus novel epidemiological and the evolutionary insights.

Large Recombinations as a Persistent Process of Genetic Diversification

Our analyses reveal that large recombination events (>100 kb) occur commonly in Kp, strongly highlighting them as a persistent mechanism of diversification in the CG258 clade. The emergence of three CG258 lineages due to large recombination events have been described, that is, ST11 (Gaiarsa et al. 2015), ST258 (Chen et al. 2014) and the separation between ST258 clade1 and clade 2 (Deleo et al. 2014). In this work, we found that other four epidemiologically relevant lineages emerged after large recombination events: ST340, ST437, ST833, and ST855 (see supplementary table S3, Supplementary Material online, and fig. 1). This suggests that large recombinations have an important role in the emerging of epidemiologically relevant clones.

Positions of Large Recombinations in the Genome

The high number of recombinations identified within the CG258 clade is not randomly distributed along the genome. As shown in figure 1, most of them are localized in a genomic portion sized ~1 Mb (from position 1,660,631 to 2,750,819, here called "highly recombined region"), suggesting that this region of the Kp genome can experience recombination events with higher successful rate that the others. Within the "highly recombined region," cluster analysis revealed the existence of two partially overlapping, but well-delimited, subregions (here called "overlapped Cluster1-2" and "Cluster1 only" regions, see fig. 1). The "overlapped Cluster1-2" region showed a recombination frequency higher than the "Cluster1 only" region. The "overlapped Cluster1-2" region contains the capsular (cps) operon, which includes genes that encode for proteins that are recognized by the human immune system. Wyres et al. (2015) already described that the genomic portion around cps genes is a recombination hotspot within the CG258. In this work, we describe two genomic subregions ("overlapped Cluster1-2" and "Cluster1 only") characterized by different recombination rates. The higher frequency of recombination observed in the genomic region around the capsule genes can be explained as consequence of a diversifying selection on the capsule (Wyres et al. 2015). On the other hand, this selective pressure cannot explain the discontinuity in recombination frequency observed between the two subregions. For this reason, we focus on this phenomenon, investigating selective forces that could have caused this discontinuity.

Repeated Sequences and the Positions of Large Recombinations

First, we determined if the region at the limits of the "overlapped Cluster1-2" contains genomic elements that could explain its higher recombination rate, such as prophages or repeated sequences. No prophagic sequences were found within the highly recombined region (see supplementary fig. S5, Supplementary Material online). This result, interestingly, suggests that a relationship between large recombinations and the genomic integration of phages could exist, but the possible mechanisms still remain to be investigated. Despite this, it does not provide any information about the two subregions of the highly recombined region. Furthermore, we did not find any pattern in the presence of repeated sequences that distinguish the highly recombined region from the rest of the genome (supplementary fig. S6, Supplementary Material online). Indeed, the two peaks of the Kernel density of repeated sequences are localized outside the highly recombined region (supplementary fig. S6, Supplementary Material online). On the other hand, we found that the "overlapped Cluster1-2" was more rich in repeated sequences than the "Cluster1 only" (supplementary fig. S6, Supplementary Material online); we also found repeated sequences located at the border between "overlapped Cluster1-2" and "Cluster1 only" subregions. Thus, the different frequency of recombination in these two subregions might be related with the difference in the presence of repeated sequences.

CG258 Common/Accessory/Rare Genes and the Positions of Large Recombinations

In addition to the search for repeated sequences and prophages, we investigated the entire data set for gene composition. We found that CG258 common genes (i.e., present in >95% of the 401 CG258 strains analyzed) are more frequently localized around the ORI of the genome, outside the highly recombined region. Contrary, the CG258 accessory (present in >5% and <95%) and rare genes (<5%) are frequently localized within the highly recombined region (see fig. 3*a*), distant from the ORI.

After a large recombination, an entire region of the recipient genome is replaced and this can cause gene composition changes (Hanage 2016). We can expect that this phenomenon is particularly important in bacterial species with highly variable gene composition, such as Kp (Holt et al. 2015). We can argue that a gene present in >95% of the CG258 strains should have an important role in their organismal function, thus we will refer to these genes as "putative essential genes." Thus, two different explanations can be proposed for the gene presence/absence pattern described above: 1) large recombinations increase gene content variability and 2) the highly recombined region contains few essential genes and thus can be replaced more frequently. In order to

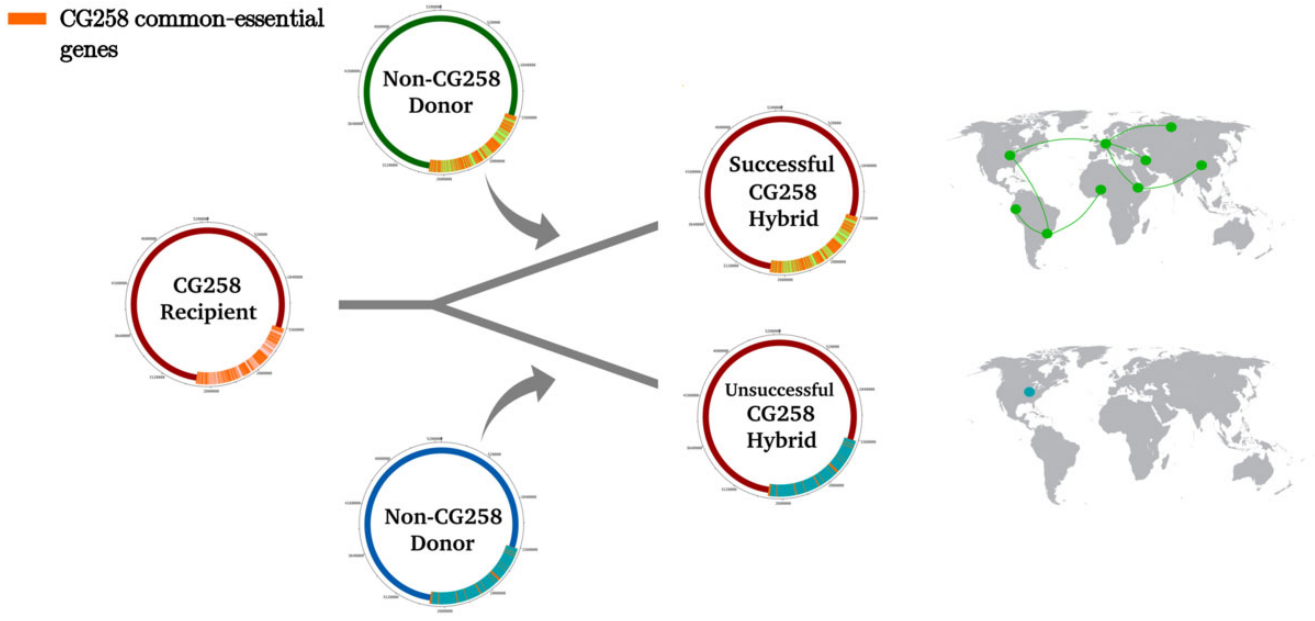

Fig. 4.—Graphical representation of the hypothesis proposed in this work: After a large recombination, the success of the emerged hybrid strain depends on how many survival genes of the recipient are present within the genomic region provided by the donor. The recipient strain genome is represented as a red circle (on the left), the genomic region that is replaced is in light red, and the survival genes within this region are represented as orange lines. Two possible events are represented: 1) a large recombination involving a donor (in green) harboring many survival genes produces a successful hybrid strain able to spread worldwide and 2) a large recombination involving a donor (in blue) harboring a few survival genes produces an unsuccessful hybrid strain.

investigate these two explanations, we studied the common genes more in depth.

Common Genes and the Positions of Large Recombinations

We divided the CG258 common genes in three categories: common-common (present also in >95% of the non-CG258), common-accessory (present in the range from 5% to 95% of the non-CG258), and common-rare (present in <5% of the non-CG258). As shown in figure 3*b* and *c*, most of the common-common genes are localized outside the highly recombined region and only one common-rare gene is localized within this region. This result gives a first evidence that the frequency of the recipient core genes (the CG258 core genes) among the possible donor strains (the non-CG258 strains) can have an effect of the successfulness of large recombinations. Furthermore, within the highly recombined region, we found that "Cluster1 only" region is richer in common-accessory genes and less rich in common-common genes than the "overlapped Cluster1-2" (see fig. 3*b* and *c*).

Based on the above results, we propose that gene compositions of the recipient and the donor strains (and the positions on the genes on the genomes) have a role in the determination of the successfulness of large recombinations. Indeed, if a recombination event causes the loss of one or more essential genes in the recipient strain genome, we can expect that an hybrid strain with reduced fitness will emerge. Let us assume that in a given strain a genomic region is particularly rich in essential genes that are not frequently present in the possible donor strains (i.e., common-accessory genes); we expect that only a few large recombinations will be successful in this region. Contrary, we expect that a genomic region rich in essential genes frequently present in the possible donors (i.e., common-common genes) will experience a higher rate of successful large recombinations. In coherence with this model, most of the common-accessory genes localized within the highly recombined region belong to fundamental pathways, highlighting how a large recombination can affect several compartments of the recipient strain metabolism. According to this model, the donor of a successful large recombination must possess, within the exchanged genomic portion, genes essential to the recipient strain (genes already present within the replaced portion of the recipient genome). Thus, the absence, in the transferred genome region, of these essential genes would represent a genetic barrier to large recombinations. Following this model, the number of successful donor–recipient combinations is likely limited, as well as the number of possible emerging hybrid strains. We graphically illustrate the proposed model in figure 4.

We emphasize that we did not find a phylogenetic pattern for the 19 identified donors: All of them belong to the KpI phylogroup, but they span all the phylogenetic variability of

this group (fig. 2). Within KpI, only two lineages were found to be donors in more than one recombination event: ST147 and ST37 (fig. 2). The strains of these two lineages harbor more common-accessory genes than the other donor strains (supplementary fig. S8, Supplementary Material online). In addition, no associations were observed between the localization of the large recombinations and the phylogenetic position (fig. 2) or ecological origin of the donors (see supplementary table S3, Supplementary Material online). In conclusion, our work reveals that large recombinations occurred with higher frequency in specific Kp lineage pairs, and that the prevalence of these events is not evenly distributed across the Kp genome. Here, we propose that this pattern could be explained if we consider the different gene content between recipients and donors as a barrier for large recombinations. This reconstruction of the network of large recombination events in Kp provides a novel point of view on this phenomenon, highlighting the importance of such an approach for investigating the evolution of recombinogenic bacterial species.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgment

Work supported by the Romeo ed Enrica Invernizzi Foundation.

Literature Cited

Bialek-Davenet S, et al. 2014. Genomic definition of hypervirulent and multidrug-resistant *Klebsiella pneumoniae* clonal groups. Emerg Infect Dis. 20(11):1812–1820.

Bowers JR, et al. 2015. Genomic analysis of the emergence and rapid global dissemination of the clonal group 258 *Klebsiella pneumoniae* pandemic. PLoS One 10(7):e0133727.

Brisse S, Passet V, Grimont P. 2014. Description of *Klebsiella quasipneumoniae* sp. nov., isolated from human infections, with two subspecies, *Klebsiella quasipneumoniae* subsp. *quasipneumoniae* subsp. nov. and *Klebsiella quasipneumoniae* subsp. *similipneumoniae* subsp. nov., and demonstration that *Klebsiella singaporensis* is a junior heterotypic synonym of *Klebsiella variicola*. Int J Syst Evol Microbiol. 64(Pt 9):3146–3152.

Cantòn R, et al. 2012. Rapid evolution and spread of carbapenemases among Enterobacteriaceae in Europe. Clin Microbiol Infect. 18(5):413–431.

Cao X, et al. 2014. Molecular characterization of clinical multidrug-resistant *Klebsiella pneumoniae* isolates. Ann Clin Microbiol Antimicrob. 13(1):16.

Chen L, Mathema B, Pitout JDD, DeLeo FR, Kreiswirth BN. 2014. Epidemic *Klebsiella pneumoniae* ST258 is a hybrid strain. MBio 5(3):e01355-14.

Chen L, Todd R, Kiehlbauch J, Walters M, Kallen A. 2017. Pan-resistant New Delhi Metallo-beta-lactamase-producing *Klebsiella pneumoniae*—Washoe County, Nevada, 2016. MMWR Morb Mortal Wkly Rep. 66:33.

Chevreux B, Wetter T, Suhai S. 1999. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99, pp. 45–56. Hannover, Germany, October 4–6, 1999. http://mira-assembler.sourceforge.net/docs/DefinitiveGuideToMIRA.html.

Croucher NJ, Klugman KP. 2014. The emergence of bacterial "hopeful monsters." MBio 5(4):e01550-14.

Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14(7):1394–1403.

Delcher AL, Phillippy A, Carlton J, Salzberg SL. 2002. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 30(11):2478–2483.

Deleo FR, et al. 2014. Molecular dissection of the evolution of carbapenem-resistant multilocus sequence type 258 *Klebsiella pneumoniae*. Proc Natl Acad Sci U S A. 111(13):4988–4993.

Didelot X, Wilson DJ. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol. 11(2):e1004041.

Gaiarsa S, et al. 2015. 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(1):389–396.

Guindon S, Delsuc F, Dufayard JF, Gascuel O. 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol. 537:113–137.

Hanage WP. 2016. Not so simple after all: bacteria, their population genetics, and recombination. Cold Spring Harb Perspect Biol. 8(7):a018069.

Holt KE, et al. 2015. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. Proc Natl Acad Sci U S A. 112(27):E3574–E3581.

Hyatt D, et al. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11(1):119.

Lascols C, Peirano G, Hackel M, Laupland KB, Pitout J. 2013. Surveillance and molecular epidemiology of *Klebsiella pneumoniae* isolates that produce carbapenemases: first report of OXA-48-like enzymes in North America. Antimicrob Agents Chemother. 57(1):130–136.

O'Leary NA, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44(D1):D733–D745.

Page AJ, et al. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31(22):3691–3693.

Pitout JDD, Nordmann P, Poirel L. 2015. Carbapenemase-producing *Klebsiella pneumoniae*, a key pathogen set for global nosocomial dominance. Antimicrob Agents Chemother. 59(10):5873–5884.

Popescu AA, Huber KT, Paradis E. 2012. Ape 3.0: new tools for distance-based phylogenetics and evolutionary analysis in R. Bioinformatics 28(11):1536–1537.

Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 26(7):1641–1650.

R Development Core Team. 2016. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Rosenblueth M, Martínez L, Silva J, Martínez-Romero E. 2004. *Klebsiella variicola*, a novel species with clinical and plant-associated isolates. Syst Appl Microbiol. 27(1):27–35.

Saito R, et al. 2014. First report of KPC-2 Carbapenemase-producing Klebsiella pneumoniae in Japan. Antimicrob Agents Chemother. 58(5):2961–2963.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30(14):2068–2069.
Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 34(90001):D32–D36.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.
Struve C, Krogfelt KA. 2004. Pathogenic potential of environmental *Klebsiella pneumoniae* isolates. Environ Microbiol. 6(6):584–590.
Suzuki R, Shimodaira H. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22(12):1540–1542.
Wyres KL, et al. 2015. Extensive capsule locus variation and large-scale genomic recombination within the *Klebsiella pneumoniae* clonal group 258. Genome Biol Evol. 7(5):1267–1279.
Zerbino DR. 2010. Using the Velvet de novo assembler for short-read sequencing technologies. Curr Protoc Bioinformatics Chapter 11:Unit 11.5.
Zhou Y, Liang Y, Lynch K, Dennis JJ, Wishart DS. 2011. PHAST: a fast phage search tool. Nucleic Acids Res. 39(Suppl):W347–W352.

Associate editor: Rachel Whitaker