Comparative analysis of Klebsiella pneumoniae genomes identifies a phospholipase D family protein as a novel virulence factor


Published in:
BMC Biology

Document Version:
Publisher's PDF, also known as Version of record

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© 2014 Lery et al. licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.
The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.
Comparative analysis of *Klebsiella pneumoniae* genomes identifies a phospholipase D family protein as a novel virulence factor

Letícia MS Lery¹²¹⁰*, Lionel Frangeul³, Anna Tomas⁴⁵, Virginie Passet⁶⁷, Ana S Almeida¹²¹¹, Suzanne Bialek-Davenet⁶⁷, Valérie Barbe⁸, José A Bengoechea⁴⁵¹², Philippe Sansonetti¹²⁹, Sylvain Brisse⁶⁷ and Régis Tournebize¹²¹³*

**Abstract**

**Background:** *Klebsiella pneumoniae* strains are pathogenic to animals and humans, in which they are both a frequent cause of nosocomial infections and a re-emerging cause of severe community-acquired infections. *K. pneumoniae* isolates of the capsular serotype K2 are among the most virulent. In order to identify novel putative virulence factors that may account for the severity of K2 infections, the genome sequence of the K2 reference strain Kp52.145 was determined and compared to two K1 and K2 strains of low virulence and to the reference strains MGH 78578 and NTUH-K2044.

**Results:** In addition to diverse functions related to host colonization and virulence encoded in genomic regions common to the four strains, four genomic islands specific for Kp52.145 were identified. These regions encoded genes for the synthesis of colibactin toxin, a putative cytotoxin outer membrane protein, secretion systems, nucleases and eukaryotic-like proteins. In addition, an insertion within a type VI secretion system locus included sel1 domain containing proteins and a phospholipase D family protein (PLD1). The *pld1* mutant was avirulent in a pneumonia model in mouse. The *pld1* mRNA was expressed in vivo and the *pld1* gene was associated with *K. pneumoniae* isolates from severe infections. Analysis of lipid composition of a defective *E. coli* strain complemented with *pld1* suggests an involvement of PLD1 in cardiolipin metabolism.

**Conclusions:** Determination of the complete genome of the K2 reference strain identified several genomic islands comprising putative elements of pathogenicity. The role of PLD1 in pathogenesis was demonstrated for the first time and suggests that lipid metabolism is a novel virulence mechanism of *K. pneumoniae*.

**Keywords:** Genome sequencing, Host-microbe interactions, Bacterial pathogenesis, Lipid metabolism

**Background**

*Klebsiella pneumoniae* is responsible for a variety of diseases in humans and animals. As a prominent nosocomial pathogen it is mainly responsible for urinary tract, respiratory tract or blood infections [1-3]. In addition, because of the acquisition of extended-spectrum β-lactamases and carbapenemases, such as the recently described NDM-1 [4], multi, extremely or pan-drug resistant clinical strains are more frequently isolated [5,6]. In addition, *K. pneumoniae* has re-emerged as a cause of community-acquired infections including pneumonia and the characteristic syndrome of pyogenic liver abscess, with possible complications including endophthalmitis or meningitis [7,8]. *K. pneumoniae* is, thus, an important virulent pathogen able to cause serious infections in ambulatory, otherwise healthy hosts and to spread within patients [5,9] that requires a better understanding of the molecular mechanisms underlying the various forms of its pathogenesis.

Major *K. pneumoniae* virulence factors include the capsule, the lipopolysaccharide, iron scavenging systems and adhesins [3,10-14]. The capsule is one of the most important virulence determinants, protecting against serum bactericidal activity, antimicrobial peptides and phagocytosis [11,15-18]. At least 77 capsular (K) types can be distinguished, but types K1 and K2 are prominent...
by their virulence in murine models of infection [19-21] and by their epidemiological prevalence [9,18,21,22]. However, not all K1/K2-type strains are necessarily virulent, as distinct clonal groups of K1 and K2 differ sharply by their virulence [23]. Reference strain Kp52.145 (derived from B5055, the reference strain of serotype K2) is a highly virulent strain from which important virulence factors, including the large virulence plasmid harboring the regulator of mucoid phenotype (rmpA) and the aerobactin cluster, were identified [10,11,21]. Even though a virulence plasmid-cured strain is less virulent than the parental strain, it remained more virulent than other isolates that do not harbor this plasmid [21], showing that factors other than capsule overexpression and aerobactin account for the higher pathogenesis of this strain. Therefore, although not all K1/K2-type strains are necessarily virulent, as distinct clonal groups of K1 and K2 differ sharply by their virulence factors, the higher pathogenesis of this strain is a powerful approach to identify putative virulence genes. Several draft or complete genomes of clinical isolates of K. pneumoniae have been published so far, but only the virulent serotype K1 strain NTUH-K2044 [24] has been described in detail. In order to identify new K. pneumoniae K2 virulence factors, we sequenced the genome of the virulent strain Kp52.145, as well as two additional strains of low virulence, SB2390 (serotype K2) and SB3193 (serotype K1). By comparing these three novel genomes to the publicly available genomes of the virulent K1 strain NTUH-K2044 and reference strain MGH 78578, we identified in Kp52.145 putative virulence genes and analyzed their distribution within a diverse collection of K. pneumoniae strains. We demonstrate that a gene coding for a phospholipase D family protein (PLD1), located within a type VI secretion system locus, is expressed in vivo, is involved in controlling bacterial membrane lipid composition, and is a new virulence factor.

**Results and Discussion**

**Genome assembly and annotation**

The genomes of Kp52.145, SB2390 and SB3193 were sequenced by a combination of 454 and Illumina technologies using single and paired-end libraries. Finishing efforts resulted in the complete genome sequence of K. pneumoniae Kp52.145 (one chromosome + two plasmids), comprising 5.45 Mbp and 5,314 protein coding genes (Figure 1). SB2390 and SB3193 genomes were assembled in 11 and 17 scaffolds, respectively. The GC% of these three genomes ranged from 55.6% to 56.7%. The general features of the K. pneumoniae sequenced genomes are summarized in Table 1.

Because Kp52.145 is a highly virulent strain, our analyses were focused on comparing its genome with the genomes of K. pneumoniae strains SB2390, SB3193, NTUH-K2044 and MGH78578. According to SEED subsystems annotations [25], about 60% of protein-coding genes for each K. pneumoniae genome had predicted functions. More specifically, the largest percentage of annotated genes is involved in the metabolism of carbohydrates (approximately 20%), of amino acids and their derivatives (approximately 10%) and of cofactors, vitamins and prosthetic groups (approximately 8%) [see Additional file 1].

**Common genome**

To define the common genome of the five strains, we used stringent BlastClust parameters of at least 90% identity and at least 80% coverage. This analysis identified 3,587 coding sequences common to the five genomes. The majority of proteins are involved in metabolic processes, such as energy metabolism and transporters, supporting the general concept that the core genome encompasses essential functions required for survival of the microorganism. The K. pneumoniae core genome comprised several sets of genes whose functions are related to bacterial survival in the environment or interaction with its host, and possibly virulence. This was the case, for example, of genes involved in quorum-sensing and biofilm formation, adhesins, and secretion systems, for which examples are detailed below.

Genes encoding for autoinducer-2 and type III fimbria, involved in biofilm formation in K. pneumoniae [26-28], were present in all sequenced strains. In addition, genes coding for synthesis and transport of the poly-β-1,6-N-acetyl-D-glucosamine (PGA) adhesin (KpST66_4915 to KpST66_4918 in Kp52.145 genome) which is required for the structural stability of Escherichia coli biofilms [29], and YidE (KpST66_0019), which mediates the hyperadherence phenotype of E. coli [30], were also found in the core genome of K. pneumoniae strains.

Moreover, the five K. pneumoniae genomes contained the genes barA/uvrY (respectively KPST66_0986 and KpST66_3517) and ycfX/ycfF (KpST66_2441 and KpST66_2442) that may be involved in bacterial fitness and virulence. The two-component system BarA/UvrY (KpST66_3517 and KpST66_0986 in Kp52.145 genome) contributes to biofilm formation in Salmonella enterica and is a virulence determinant of urinary tract E. coli infections [31,32]. The E. coli YcfF protein is expressed in a septicemia murine model of infection in which the ycfF mutant is attenuated, thus suggesting its implication in the in vivo survival/multiplication of the bacteria [33].

Several putative secretion systems were identified as part of the common genome of the K. pneumoniae strains, including one type I secretion system (T1SS) and one type II secretion system (T2SS). T2SS is composed of the pullulanase related genes pulA-O that are involved in the pathogenesis of several bacteria [34,35]. Streptococcus
Figure 1 K. pneumoniae Kp52.145 genome. A) Chromosome representation: Outermost layers in gray indicate the position of positive and negative strand CDSs. tRNAs are represented in green, while the four virulence-related genomic islands (GI) are in red, the locus coding for anaerobic metabolism in orange, the T6SS gene clusters in blue and capsular synthesis region in yellow. Inner circle represents G+C%. Detail of the T6SS locus III region containing the putative phospholipase (pld1 gene) (orange) and sel1 (pink) genes is shown enlarged.

B) Plasmids representation: plasmid maintenance genes and IS sequences are shown in blue, proteins with unknown functions in gray, known functions in green, toxin-antitoxin systems in yellow and mspA in red. CDSs, coding sequences.
PulA binds to host lung glycogen leading to a strong interaction with alveolar type II cells [36]. Similarly, the alpha-amylase AmyA degrades glycogen into cyclic maltodextrins, which increases the transepithelial translocation of Streptococcus [37]. Both amyA and pulA-O genes are encoded in Klebsiella genomes, but their functions in this bacterium remain to be characterized.

Type VI secretion system (T6SS) putative genes were located in at least three different loci of the Klebsiella genomes, in accordance with a previous in-silico study [38]. T6SS clusters are usually found within pathogenicity islands or on chromosomal regions presenting virulence or host survival biases. Additionally, T6SS has been suggested to assist colonization and infection. Indeed, in a screen that identified Klebsiella mutants failing to colonize mice [39], two of them were mutants in genes coding for proteins annotated as hypothetical, that have been subsequently re-annotated as T6SS proteins [38]. Type III secretion systems were not found in Klebsiella, but type IV secretion systems, possibly corresponding to conjugation apparatus, were present only in some strains (Kp52.145, SB2390 and NTUH-K2044) (Table 2).

Furthermore, the core genome presented a large genomic region located between gly-tRNA and phe-tRNA loci in Kp52.145 genome, containing frdABCD genes coding for the fumarate reductase enzymatic complex responsible for fumarate respiration under anaerobic growth of bacteria. This complex is a virulence determinant for Helicobacter pylori, Mycobacterium tuberculosis, Actinobacillus pleuropneumoniae and S. enterica, as mutants on these genes are attenuated [40-43]. The ability to grow anaerobically allows bacterial pathogens to persist in host tissues, including in the lungs. Curiously, this genomic locus encoded in Klebsiella at least one more protein involved in anaerobic metabolism, the anaerobic C4-dicarboxylate transporter DcuA (KpST66_4904), supporting the idea that this GI provides Klebsiella advantages to grow under anaerobic conditions, possibly favoring infection. Additional proteins that might be involved in bacterial fitness to environmental stress conditions were encoded in this region. For instance, the putative small multidrug resistance protein SugE (KpST66_4882) has been shown to regulate biofilm formation and capsule expression [44]. A lipocalin-2 bacterial protein Bcl (KpST66_4881) is also encoded in this island of the genome. Eukaryotic lipocalins are small extracellular proteins that bind hydrophobic ligands and fulfill numerous biological functions including regulation of

### Table 1 General features of the K. pneumoniae genomes analyzed

<table>
<thead>
<tr>
<th>K-type</th>
<th>Strain bank ID</th>
<th>Scaffolds</th>
<th>Size (Mb)</th>
<th>% GC</th>
<th>Total PCG</th>
<th>rRNA operons</th>
<th>tRNA genes</th>
<th>% coding</th>
<th>Plasmids</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>K2</td>
<td>3341/Kp52.145</td>
<td>3</td>
<td>5.45</td>
<td>56.4</td>
<td>5,314</td>
<td>8</td>
<td>85</td>
<td>88</td>
<td>2</td>
<td>This manuscript</td>
</tr>
<tr>
<td>K2</td>
<td>2390</td>
<td>11</td>
<td>5.62</td>
<td>56.7</td>
<td>5,367</td>
<td>6</td>
<td>77</td>
<td>86.8</td>
<td>?</td>
<td>This manuscript</td>
</tr>
<tr>
<td>K1</td>
<td>3193</td>
<td>16</td>
<td>5.01</td>
<td>55.6</td>
<td>4,793</td>
<td>10</td>
<td>73</td>
<td>92</td>
<td>?</td>
<td>This manuscript</td>
</tr>
<tr>
<td>K1</td>
<td>NTUH (ST23)</td>
<td>2</td>
<td>5.24</td>
<td>57.6</td>
<td>4,992</td>
<td>19</td>
<td>76</td>
<td>89</td>
<td>1</td>
<td>Wu et al. [63]</td>
</tr>
<tr>
<td>K5</td>
<td>MGH (SB107)</td>
<td>6</td>
<td>5.31</td>
<td>57</td>
<td>4,776</td>
<td>10</td>
<td>56</td>
<td>85</td>
<td>5</td>
<td>(NC_009648)</td>
</tr>
</tbody>
</table>

The three genomes sequenced in this study are compared to strains NTUH and MGH, previously published. PCG, protein coding genes.

### Table 2 Distribution of virulence-related factors among K. pneumoniae genomes analyzed

<table>
<thead>
<tr>
<th>Virulence-factor</th>
<th>SB3341</th>
<th>SB2390</th>
<th>SB3193</th>
<th>NTUH</th>
<th>MGH</th>
<th>Functional role</th>
</tr>
</thead>
<tbody>
<tr>
<td>rpmA</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>++</td>
<td>-</td>
<td>Regulator of capsule expression</td>
</tr>
<tr>
<td>Aerobactin</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>Siderophore</td>
</tr>
<tr>
<td>Enterobactin</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Siderophore</td>
</tr>
<tr>
<td>Yersiniabactin</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>Siderophore</td>
</tr>
<tr>
<td>Colibactin</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Genotoxin</td>
</tr>
<tr>
<td>T4SS (virB)</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>Conjugal machinery/protein secretion</td>
</tr>
<tr>
<td>T2SS</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Protein secretion</td>
</tr>
<tr>
<td>T6SS</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Protein secretion</td>
</tr>
<tr>
<td>Pld-family</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Lipid metabolism</td>
</tr>
<tr>
<td>Sell1 lipoproteins</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Unknown</td>
</tr>
<tr>
<td>cOMP</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Putative cytotoxin</td>
</tr>
<tr>
<td>Igg-like</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Binding to extra cellular matrix compounds</td>
</tr>
<tr>
<td>SEFIR-domain</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Potentially hijack IL17R signaling pathways</td>
</tr>
<tr>
<td>Bcl</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>Binding to hydrophobic ligands / putative regulation of homeostasis and immunity</td>
</tr>
</tbody>
</table>
cellular homeostasis and immunity and are regulators of antibacterial defense [45]. Lipocalin2 is for instance a siderophore scavenger for several bacteria, including K. pneumoniae [46], as well as a negative regulator of inflammatory response during Streptococcus pneumoniae pneumonia [47]. However, the role of bacterial lipocalins is not yet known.

Finally, a phosphatidylserine decarboxylase (KpST66_4873) might be important for the integrity of the bacterial membrane composition, as phospholipid biosynthetic pathways play crucial roles in the virulence of several pathogens [48,49].

**Accessory genome**

The accessory genome (genes absent in at least one of the five strains) included a large number of specific coding sequences (CDSs): 743 genes were found only in Kp52.145, 608 in NTUH-K2044, 806 in SB3193, 635 in SB2390 and 488 in MGH78578. About 50% of these genes code for hypothetical proteins or proteins of unknown function. The distribution of the putative virulence-related genes of K. pneumoniae among the sequenced strains is summarized in Table 2. The specific regions or genes of Kp52.145 are detailed below.

**Kp52.145 plasmids**

Strain Kp52.145 possessed two plasmids (Figure 1B). The first plasmid of 121 Kb carried the aerobactin cluster and the regulator of mucoid phenotype rmpA genes. The presence of this plasmid was previously correlated with the virulence of K. pneumoniae K1 and K2 isolates [21,50]. Additionally, a strain cured from this plasmid showed a 6 × 10^4-fold reduction in virulence, establishing the link between this plasmid and bacterial virulence [21]. In addition to aerobactin and rmpA, this plasmid contained genes coding for F-pilin, purine metabolism, insertion sequences and proteins of unknown functions. Kp52.145 also carried a second, previously non-described, plasmid of about 90 Kb. Interestingly, it contained rmpA2, a homologue of the regulator of mucoid phenotype rmpA, which seems to be involved in capsule expression regulation [51,52]. F-pilin genes, a subtilisin-related serine protease, an AAA + ATPase, the UV protection system UmuD/UmuC and the genes encoding the toxin-antitoxin systems RelE/orf4 and VagD/VagC (Figure 1B) were also found on this plasmid. However, the potential role of these genes in virulence remains to be investigated.

**Genomic islands identified in the genome of strain Kp52.145**

In addition to the capsule synthesis operon and the iron-acquisition systems (yellow and orange boxes in Figure 1A, respectively), known to be involved in K. pneumoniae virulence, four additional regions of the Kp52.145 genome presented several characteristics of pathogenicity islands (red boxes, Figure 1A). These regions are defined by a different GC content in comparison to the average of the genome, represented large chromosomal regions (often > 30 Kb), were associated with tRNA genes or with the presence of insertion sequences, integrases and transposases, and were present in pathogenic strains while less frequent in less-virulent strains [53,54]. The four GIs present in the Kp52.145 genome were present or partially present in NTUH, but none of them was present in SB2390, SB3193 and MGH 78578, indicating that their occurrence is specific to pathogenic genomes. Figure 2 shows the four GIs identified, highlighting the putative virulence related genes.

**Genomic island 1 (GI-I): ICE-Kp1-like region**

The largest GI found in the Kp52.145 genome comprised 133,679 bp, presented a GC content of 52%, was inserted in an asn-tRNA locus and encoded 92 CDSs (Figure 2A). Most of the protein-coding genes found in this region were described as part of the IceKpI GI of NTUH-K2044 [55,56], although several differences were found. Kp52.145 GI-I begins at asn tRNA locus followed by several uncharacterized proteins and insertion sequence elements. GI-I encoded for the synthesis of two polyketide/nonribosomal peptides (yersiniabactin and colibactin) and for the conjugative transfer machinery (T4SS) that allows horizontal transfer of the island [56]. In contrast to the previous description of IceKpI [56], the Kp52.145_GI-I carried colibactin and did not contain the region coding for vagC-vagD, iroN-iroB-iroC-iroD and rmpA genes which are carried only by the 121 kb plasmid in Kp52.145.

**Genomic island 2 (GI-II)**

We describe here a novel GI (GI-II). It is a 29,829 bp island, with a GC content of 49% and coding for 133 genes which is inserted in a leu-tRNA locus (Figure 2B). Potential pathogenesis-related genes coded for a putative cytotoxic outer membrane protein (cOMP, KpST66_4736) and a polyamine ABC transport system (KpST66_4729 to KpST66_4732). cOMP closest known homologue (34% identity) was a Plesiomonas shigelloides predominant virulence factor proposed to trigger cell death in host cells following infection [57]. Polyamine biosynthesis and transport mechanisms were intricatedly linked to fitness, survival, biofilm formation and pathogenesis, for instance in S. pneumoniae and Yersinia pestis [58,59]. Additionally, this GI encoded a 4-phytase gene (KpST66_4736), ugpQ3 (KpST66_4728) and xylA, xynT, xynB, xylR (KpST66_4724 to KpST66_4727) that are involved in xylose metabolism.

**Genomic island 3 (GI-III)**

The third GI is characterized by a 49,657 bp region presenting a G + C content of 51% and contained 66 CDSs,
Figure 2 (See legend on next page.)
most of them coding for phage structural proteins (Figure 2C). This GI included genes coding for proteins with homologues that were shown to play a role in bacterial adhesion and immune system escape [60-62]. These proteins encoded for an immunoglobulin domain-containing protein (KpST66_1506), a peptidase S24-like protein (KpST66_1511), two HNH family endonucleases (KpST66_1468 and 1486) and an exonuclease (locus 1464).

Genomic island 4 (GI-IV)

The forth island (GI-IV) was mainly comprised of phage-related genes. Among the 42 CDSs encoded within this genomic region, one gene coded for a SEFIR-domain containing protein (KpST66_1945; Figure 2D). A SEFIR domain is usually found in IL17 receptors and SEF proteins, acting in eukaryotes signaling pathways. Very little is known about prokaryotic SEFIR-containing proteins. Structural analyses suggested that these bacterial SEFIR domains share structural and electrostatic similarity with their mammalian homologues and, thereby, could potentially subvert host immunity by hijacking the IL17R signaling pathways [63]. Notably, local production of IL-17 is a significant factor in effective host defense against Gram-negative bacteria, including K. pneumoniae [64]. Therefore, further studies are required to elucidate whether KpST66_1945 is implicated in K. pneumoniae pathogenesis.

Distribution of GIs among K. pneumoniae strains

To investigate the distribution of the described GIs in K. pneumoniae, the presence of the putative virulence-related genes was searched using BlastN in 171 genomes, including 119 publicly available and 52 unpublished genomes (Bialek-Davenet, Brisse et al., unpublished work) representing 47 different sequence types (STs). Whereas the SEFIR-domain containing protein gene of GI-IV was only found in two (1.2%) isolates of sequence type ST15, the three other GIs described herein were more distributed among K. pneumoniae isolates (Table 3). GI-II genes were present in a total of 11 (6.4%) isolates, most of which belonged to ST375, ST65 and ST25, which were associated with severe infections caused by isolates of capsular serotype K2 [9,65]. The genes of GI-II were always found in synteny. GI-III genes were observed in only seven (4.1%) isolates dispersed in several unrelated STs. The distribution of the ICEKpI element (similar to GI-I) has been previously analyzed [55].

**T6SS locus III insertion**

Recently, three different T6SS loci were defined in K. pneumoniae [38]. Within these loci, three putative valine-glycine repeat (Vgr) proteins and two hemolysin-coregulated proteins (Hcp) were described as potential effector proteins, through their sequence similarities to Vibrio cholerae and Pseudomonas aeruginosa effector proteins [65-68]. Accordingly, the Kp52.145 genome also presented three conserved T6SS loci syntenic to those previously described. The first two loci were identical in composition and orientation to the previously described ones. The third one, locus III, had conservation of adjacency limited to the imcF/impG/impH and impJ/ompA/vgrG gene clusters, as a region with nine genes was inserted between these two clusters. This insertion encoded for one hypothetical protein, five putative Sel-1 repeat containing lipoproteins and three putative phospholipase D family proteins.

### Table 3 Prevalence of the genomic islands among K. pneumoniae isolates

<table>
<thead>
<tr>
<th>Genomic island</th>
<th>Virulence-related features</th>
<th>Prevalence (%)</th>
<th>Remarkable STs</th>
</tr>
</thead>
<tbody>
<tr>
<td>GI-I</td>
<td>Colibactin</td>
<td>3.5</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Yersiniabactin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GI-II</td>
<td>cOMP</td>
<td>5.8</td>
<td>3775,65,25</td>
</tr>
<tr>
<td></td>
<td>4-phytase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GI-III</td>
<td>Iggl-like</td>
<td>8.1</td>
<td>Dispersed</td>
</tr>
<tr>
<td></td>
<td>Exonuclease</td>
<td>57.6</td>
<td>375,65</td>
</tr>
<tr>
<td>GI-IV</td>
<td>Sefir-domain</td>
<td>1.7</td>
<td>-</td>
</tr>
<tr>
<td>T6SS insertion</td>
<td>PLD1</td>
<td>7.0</td>
<td>380,35</td>
</tr>
<tr>
<td></td>
<td>Sel 1</td>
<td>7.0</td>
<td>380,35</td>
</tr>
</tbody>
</table>

*aaccording to ref. [53]. Prevalence of virulence-related features encoded in GI-II, GI-III, GI-IV and T6SS insertion was based on >90% identity and >50% coverage in the length of the genes, using a database of 171 Klebsiella genomes representative of 47 different STs. STs, sequence types.*
(Figure 1A). Flanking sequences suggesting how this region was inserted were not found.

Sel1 lipoproteins are poorly characterized and there was no evidence for their function in *K. pneumoniae*, but they are essential in *Legionella pneumophila* for invasion of host cells where they influence vacuolar trafficking of bacteria [69]. The three open reading frames encoding putative phospholipase D proteins in strain Kp52.145 encoded one full length protein (KpST66_3368, 623 aminoacids), one C-terminal region (KpST66_3371, 187 aa) and one N-terminal region (KpST66_3372, 317 aa). Phospholipase D family proteins have been described as important for host cell invasion, bacterial dissemination and disease progression [70-73]. The bacterial phospholipase D family comprises at least four classes of proteins with distinct functions: true phospholipase D, cardiolipin synthase, phosphatidylserine synthase and endonuclease [74]. Full length *K. pneumoniae* PLD1 and its closest homologs all presented the conserved motif HXK(X)4Da in e. r i n e (Figure 3), but no other conserved domain was described in each family, thus making it difficult to infer protein function only by sequence analysis.

In order to obtain evidence that the product of these three genes coding either for full length or partial PLD-family proteins are important for bacterial survival *in vivo*, we checked by RT-PCR for their mRNA expression. We observed that these genes are expressed both in bacteria grown for four hours in Trypto Casein Soy broth (GTCS) medium, as well as in the lungs of mice infected for 24 hours (data not shown). These results prompted us to further check for a putative involvement of the full length PLD gene, called *pld1*, on *K. pneumoniae* virulence.

**Involvement of the phospholipase D family protein gene *pld1* in KP 52.145 virulence**

As PLD-family proteins have been shown to be involved in virulence [75-77], we decided to characterize the role of the full length PLD1 protein. We first tested a *pld* mutant strain in a *K. pneumoniae* murine model. Mice were infected intranasally with 10⁸ of either the wild-type bacteria, a *pld1* mutant, or the mutant strain complemented with a plasmid expressing the putative PLD1 protein (pPLD), and their survival was monitored for seven days. Interestingly, the mutant strain appeared avirulent in a mouse model of acute pneumonia while mice infected with the wild-type and the complemented strain succumbed in less than one week (Figure 4). However, the wild-type, the mutant and the complemented mutant strains grew equally well in Luria-Bertani broth (LB) broth (data not shown). These results indicated that the *pld1* mutant was strongly attenuated *in vivo*, thus showing an important role for PLD1 in virulence.

To analyze the frequency and clonal distribution of the *pld1* gene in *K. pneumoniae*, the 171 genomes were analyzed using BlastN. We observed that besides ST66, represented by strain Kp52.145, *pld1* was present in 10 strains (6.4%) belonging to ST380, ST67, ST67 (*K. pneumoniae* subsp. *rhinoscleromatis*) and ST35, but in none of the other isolates. It is interesting to note that ST380 was associated with severe *K. pneumoniae* infections [9,65] and that *K. pneumoniae subsp. rhinoscleromatis* is the only *Klebsiella* strain to be able to survive intracellularly in macrophages [78].

**Functional characterization of PLD1**

In order to demonstrate the phospholipase activity of PLD1 and characterize its involvement in lipid metabolism, the lipid composition of wild-type and mutant strains was analyzed by thin-layer chromatography (TLC). A remarkable lipid spot was absent from the *pld1* mutant in comparison with the complemented strain, suggesting that the putative PLD1 is involved in lipid metabolism [see Additional file 1]. To reinforce this result, a plasmid carrying the *pld1* gene was inserted into *E. coli* strain SD9 [79]. This strain is deficient in phosphatidylserine and cardiolipin, thus presenting a simpler lipid composition than its parental strain and Kp52.145. Lipid profiles of SD9 and complemented strains had a different lipid composition. Notably, the PLD1-expressing strain contained an additional lipid spot in comparison to the SD9 strain, suggesting that PLD1 is responsible for this difference (Figure 5A). SD9 wild-type strain also presented an extra lipid spot in comparison to the PLD1-expressing strain, possibly representing the PLD1 substrate (Figure 5A). Densitometric analysis of iodine-stained lipids on TLC plates revealed that this lipid spot corresponded to 21% of the total amount of lipids in SD9 strain, but only 6% of SD9 complemented with a plasmid expressing *pld1*. Mass spectrometry (MS) analysis of total lipid extract was carried out to identify such lipid. Comparing lipid profiles by MS, we found a lipid of mass 788.4 present only in the PLD1-deficient strain (Figure 5B) and identified it as phosphatidyl glycerol (PG) using the LipidMaps database.

As mentioned above, the bacterial PLD-family proteins can be classified in four subfamilies. One of them, the cardiolipin synthase is able to convert two PG molecules into glycerol and cardiolipin, or to catalyze the opposite reaction, leading to PG formation [80]. Our results suggest that PLD1 belongs to the cardiolipin synthase subfamily and that it plays a role in balancing the PG and cardiolipin content.

It has been shown that humans and mice with bacterial pneumonia have markedly elevated amounts of cardiolipin in lung fluid and that it impairs surfactant function, lung mechanics, modulation of cell survival and cytokine networks and lung consolidation [81].
Figure 3 Sequence alignment of Kp52.145 (Kp) PLD1 and its closest related sequences: putative phospholipase D family protein from Pseudomonas syringae (PS) and cardiolipin synthases from P. putida (PP), Staphylococcus aureus (SA), Bacillus subtilis (BS) and E. coli (EC). The rectangles indicate the phospholipase D active site regions and the asterisk points to the corresponding transposon insertion site in pld1 mutant strain.
There is evidence that bacteria are able to adjust their relative concentrations of phosphatidylethanolamine and PGs when subjected to environmental stresses. Such an alteration in headgroup composition seems to be a means for changing membrane permeability and, hence, preserving stability [82]. Therefore, we hypothesize that PLD1 alters the membrane composition and charge, affecting bacterial interaction with the host environment.

Recently, Russel *et al.* demonstrated that diverse phospholipase proteins encoded within the T6SS loci of several prokaryotic genomes are antibacterial effectors, conferring competitive advantages on the donor strain during interbacterial interactions [83]. These proteins are generally designated as ‘T6SS Lipase Effectors’ (Tle) and classified in five sub-families, according to the sequence conservation and number of catalytic motifs present. Kp52.145 PLD-family protein could be considered a Tle5 member, as it presents two conserved HxKxxxxD motifs. However, in the Kp52.145 genome we did not identify any gene similar to the cognate immunity genes - a hallmark of the genomic islands described by Russel *et al.* Moreover, we did not observe such an antibacterial effect of Kp52.145 or its PLD1 mutant strain upon competition with *E. coli* [see Additional file 1]. These results showed that pld1 is implicated in virulence without being an anti-bacterial factor and is, so far, unique.

**Conclusions**

This study presents a comparative analysis of the complete genome sequence of the high virulence reference strain Kp52.145, a derivative of the K2 reference strain B5055. It revealed five genomic regions possibly involved in bacterial virulence. One gene, *pld1*, was shown to be involved in virulence in a mouse model of pneumonia and revealed a novel implication of lipid metabolism in *K. pneumoniae* pathogenesis. Future analysis of additional putative virulence factors such as Sel1 lipoproteins, VgrG, Hcp, Bcl, cOMP and Sefir-domain containing protein are required for a comprehensive understanding of *K. pneumoniae* core virulence genes.

**Methods**

**Selection of isolates for genome sequencing/bacterial strains**

Strain Kp25.145 (a derivative of B5055, the reference strain of serotype K2) is a laboratory strain used to study *K. pneumoniae* pathogenesis [11,21] and was chosen as the focus of this work. SB2390 (cur15505, isolated in Curaçao, 2002, urinary tract infection; belongs to ST14) and SB3193 (IPEUC-744, isolated from a metritis case in a mare, 1981 in France; belonging to ST82) [23] are non-virulent strains that were sequenced to allow the comparison between virulent and non-virulent strains.

The *pld1* transposon mutant was isolated by Anna Tomas and Jose Bengoecheoa during a screen of a *K. pneumoniae* mutant library made by Tn5 transposon insertion (manuscript in preparation). In this mutant, the transposon was inserted at position 1,625 of *pld1* gene.

The complementation of the mutant strain was achieved through bacterial transformation using electrocompetent cells and a plasmid carrying *pld1* gene. *pld1* gene was amplified by PCR and cloned at the multiple
Figure 5 (See legend on next page.)
cloning site of puc18 plasmid. Cloning was confirmed by DNA sequencing.

**Genome sequencing and assembly**

*K. pneumoniae* strains were sequenced using a combination of 454 and Illumina reads. Single and paired-end 454 reads with an average of 400 nucleotides were assembled into contigs and scaffolds by Newbler. Illumina reads of about 76 or 36 nucleotides were aligned to scaffolds in order to confirm and correct possible homopolymer errors in the 454 reads. Coverage was as follows: Kp52.145 genome: 170 X using GAIIx (76 nt) + 13.8 X using MP titanium + 18 X using SR titanium; SB2390 genome: 81 X using GAIIx (36 nt) + 6.4 X using MP titanium + 22 X using SR titanium; SB3193 genome: 209 X using GAIIx (76 nt) + 5.4 X using MP titanium + 20 X using SR titanium.

Following the primary assembly of the genomes, an *in silico* finishing approach, based on the methods described by Pop et al. [84], was performed in order to identify small and identical repeats on the genomes. In such cases and without high quality bases discrepancies, the contigs containing the small repeats were manually duplicated and added to the assembly.

Scaffolds were aligned to experimentally determined BgiII optical maps generated by OpGen company from purified chromosomal DNA, using MapSolver version 3.1 software. Such alignments were used to check for the quality of the assemblies. Additionally, specific pairs of primers were designed in order to close all remaining gaps. PCR products were purified in NucleoFast 96 plates (Macherey Nagel, Düren, Germany) and aliquots were used for sequencing reactions with the BigDye Terminator Cycle Sequencing Ready reaction Kit (Applied Biosystems, Foster City, CA, USA). The resulting sequences were added to the previous assembly using Phred/Phrap/Consed.

The sequences of *K. pneumoniae* genomes have been deposited to the European Nucleotide Archive and are accessible under the accession numbers: FO834904, FO834905 and FO834906 (strain Kp52.145), CCB00000000 (strain SB2390) and CCCQ0000000000 (strain SB3193).

**Functional annotation of genomes**

In order to gain functional insights about the genome sequences, protein-coding genes were predicted and annotated using the CAAT-box genome browser [85], using a combination of GeneMark predictions and Blastx results against the Uniprot database. All the putative open reading frames (ORF) longer than 120 nucleotides presenting similarity to sequences of the Uniprot database or positive GeneMark result were considered for further analysis. The final set of CDSs underwent a manual annotation process based on description of similarity. Pfam and COG database searches, as well as SignalP, TMHMM and PredTMBB predictions were performed to improve the degree of annotation confidence, if necessary. CDSs were described as ‘highly similar to’, ‘similar to’ or ‘weakly similar to’ if they presented more than 70%, between 50% and 69% or less than 50% similarity to the protein hit sequence. Additionally, information on partial homology was included. The start codon for each CDS was automatically chosen and manually validated, based on a combination of GeneMark results and Blast alignments. RAST was used to classify proteins in functional categories. Structural RNAs were searched using trRNAscan.

The common genome of the five strains was determined using the BlastClust algorithm [86] using minimum parameters of 90% identity and 80% length coverage for proteins to be included in the same cluster.

**RT-PCR**

Lungs of control or Kp52.145-infected mice were homogenized in cold TRI reagent (Sigma, Gillingham, Dorset, UK) using a Precells lysing kit (Precells, Saint Quentin en Yvelines, France). Total lung and bacterial mRNAs were extracted according to the manufacturer’s instructions. RNA (2 μg) was reversed transcribed in cDNA using Superscript II enzyme (Invitrogen, Foster City, CA, USA). Aliquots were used in PCR reactions using specific primers [see Additional file 1].

**Animal experiments**

BALB/c mice were purchased from Janvier (Le Genest St. Isle, France). Mice were housed under standard conditions of feeding, light and temperature with access to food and water *ad libidium*. Experiments were performed according to the national and Institut Pasteur guidelines for laboratory animal experiments. Protocols were approved by the Institut Pasteur animal care and use committee (protocol 05–59) and the Direction des Services Vétérinaire de Paris (permit 75–713 to RT). Six- to eight-week-old mice were anesthetized with acepromazine (Calmivet, 1.5 mg/kg, Vetoquinol, Lure, France), and ketamine (Imalgene, 31.25 mg/kg, Merial, Lyon, France), and then infected intranasally with 20 μl...
bacterial suspension. The virulence of \textit{K. pneumoniae} strains was tested on six-week-old BALB/c mice, as previously described [23]. Seven mice per test condition were infected with 10\(^8\) bacteria. Mice were followed every day for one week. Experiments were performed at least twice.

**Thin-layer chromatography and mass spectrometry of lipids**

Bacterial lipids were extracted by the method of Bligh and Dyer [87]. Briefly, bacterial stationary-phase cultures were concentrated 10 times and mixed with chloroform: methanol. After centrifugation at 1,000 rpm for five minutes, the organic phase was recovered. Lipid profiles were analyzed by two-dimensional TLC using TLC Silica gel 60 F\textsubscript{254} plates (Merck, Whitehouse Station, NJ, USA). as the stationary phase and a chloroform-9:1 methanol mixture as the mobile phase in both dimensions. Staining was performed by iodine vapor. TLC calibrated images were acquired in ImageScanner using LabScan v5.0 software. The relative intensity of each spot was calculated in ImageMaster two-dimensional Platinum v7.0 software. Alternatively, lipid extracts were analyzed by MS and MS/MS in an ESI-Q-ToF Micro (Waters), in positive ion mode. Resolution was typically lower than 10 ppm.

**Anti-bacterial competition assays**

Competition assays were performed as previously described [88]. Briefly, \textit{K. pneumoniae} Kp52.145 or its pld1 mutant cells grown overnight on an agar plate were re-suspended in LB, normalized to OD\textsubscript{600} of 0.5 and mixed at a ratio of 5:1 with a spontaneous nalidixic acid (nal) resistant mutant strain of \textit{E. coli} MG1655. The mixture was incubated for four hours on a prewarmed agar plate. Recovered cells were plated out on antibiotic selective media and viable cells were reported as the total number recovered per co-culture spot. \textit{Serratia marcescens} DB10 was used as a positive control.

**Additional file**

Additional file 1: 1) Distribution of \textit{K. pneumoniae} Kp52.145 genes, according to RAST categories. Each functional category is represented in a different color. The total number of genes per category is shown. 2) pld PCR assay. A collection of 42 virulent and non virulent clones was screened by PCR for pld gene. Strains containing pld gene are indicated by (+) and the absence of pld gene is (-). 3) TLC lipid profiles of \textit{K. pneumoniae} Kp52.145 wild-type (left panel) and pld mutant strains (right panel). Black circles indicate differentially expressed lipids. 4) Bacterial competition assay. Anti-bacterial activity was measured as the number of \textit{E. coli} cells recovered after the co-culture with \textit{K. pneumoniae} Kp52.145 wild-type and pld mutant strains. \textit{S. marcescens} was used as a positive control strain. 5) List of primers used for RT-PCR analysis.

**Competing interests**

The authors declare they have no competing interests.

**Authors’ contributions**

LMSL participated in the genome sequencing and assembly, carried out the annotation and comparative genomic analysis, performed RT-PCR, lipid analysis and animal experiments and drafted the manuscript. LF carried out the genome assembly and participated in genome annotation. AT and JB generated the mutant strain and carried out the bacterial competition assay. VP, ASA and VB participated in the genome sequencing. SB participated in the design of the study and the genome sequencing and assembly. PS participated in the study design and coordination. RT conceived the study, participated in the genome sequencing and assembly and carried out animal experiments. SBD analyzed the presence of genes within other \textit{K. pneumoniae} genomes. LMLS, JB, PJS, SB and RT wrote the manuscript. All authors read and approved the final manuscript.

**Acknowledgements**

LMSL, AT and this work were supported by an ERA-Net Pathogenomic 2008 grant allocated to both P. Sansonetti and J. Bengooecha. ASA is a recipient of a fellowship from the Fundação para a Ciência e a Tecnologia from Portugal. The authors thank Stephane Salamone and David Touboul for their help with lipid analysis.

**Published**

29 May 2014

Received: 29 May 2014

**References**


54. Gal-Mor O, Finlay BB.

56. Lin TL, Lee CZ, Hsieh PF, Tsai SF, Wang JT.


52. Hsu CR, Lin TL, Chen YC, Chou HC, Wang JT.

63. Wu B, Gong J, Liu L, Li T, Wei T, Bai Z.

60. Choy HA, Kelley MM, Chen TL, Moller AK, Matsunaga J, Haake DA.


67. Shrivastava S, Mande SS.

70. McNamara PJ, Bradley GA, Songer JG.

82. Bligh EG, Dyer WJ.


84. Nagarajan N, Cook C, Di Bonaventura M, Ge H, Richards A, Bishop-Lilly KA, SN, Mougous JD.


87. Bligh EG, Dyer WJ.


