Genetic diversity of capsular polysaccharide biosynthesis in *Klebsiella pneumoniae* clinical isolates

Hung-Yu Shu,^{1,2†} Chang-Phone Fung,^{3,4†} Yen-Ming Liu,⁵
Keh-Ming Wu,^{5,6} Ying-Tsong Chen,⁵ Ling-Hui Li,⁵ Tze-Tze Liu,²
Ralph Kirby⁷ and Shih-Feng Tsai^{2,5,6,7}

Correspondence

Hung-Yu Shu
hyshu@mail.cjcu.edu.tw

¹Department of Bioscience Technology, Chang Jung Christian University, Tainan County, Taiwan, ROC

²Genome Research Center, National Yang-Ming University, Taipei, Taiwan, ROC

³Division of Infectious Diseases, Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan, ROC

⁴Institute of Tropical Medicine, School of Medicine, National Yang-Ming University, Taipei, Taiwan, ROC

⁵Division of Molecular and Genomic Medicine, National Health Research Institutes, Zhunan, Miaoli County, Taiwan, ROC

⁶Institute of Biomedical Informatics, National Yang-Ming University, Taipei, Taiwan, ROC

⁷Department of Life Sciences and Institute of Genome Sciences, National Yang-Ming University, Taipei, Taiwan, ROC

Klebsiella pneumoniae is an enteric pathogen causing community-acquired and hospital-acquired infections in humans. Epidemiological studies have revealed significant diversity in capsular polysaccharide (CPS) type and clinical manifestation of *K. pneumoniae* infection in different geographical areas of the world. We have sequenced the capsular polysaccharide synthesis (*cps*) region of seven clinical isolates and compared the sequences with the publicly available *cps* sequence data of five strains: NTUH-K2044 (K1 serotype), Chedid (K2 serotype), MGH78578 (K52 serotype), A1142 (K57 serotype) and A1517. Among all strains, six genes at the 5' end of the *cps* clusters that encode proteins for CPS transportation and processing at the bacterial surface are highly similar to each other. The central region of the *cps* gene clusters, which encodes proteins for polymerization and assembly of the CPS subunits, is highly divergent. Based on the collected sequence, we found that either the *wbaP* gene or the *wcaJ* gene exists in a given *K. pneumoniae* strain, suggesting that there is a major difference in the CPS biosynthesis pathway and that the *K. pneumoniae* strains can be classified into at least two distinct groups. All isolates contain *gnd*, encoding gluconate-6-phosphate dehydrogenase, at the 3' end of the *cps* gene clusters. The *rmIBADC* genes were found in CPS K9-positive, K14-positive and K52-positive strains, while *manC* and *manB* were found in K1, K2, K5, K14, K62 and two undefined strains. Our data indicate that, while overall genomic organization is similar between different pathogenic *K. pneumoniae* strains, the genetic variation of the sugar moiety and polysaccharide linkage generate the diversity in CPS molecules that could help evade host immune attack.

Received 10 March 2009

Revised 15 August 2009

Accepted 4 September 2009

†These authors contributed equally to this work.

Abbreviations: CPS, capsular polysaccharide; GT, glycosyltransferase.

The GenBank/EMBL/DDBJ accession numbers for the capsular polysaccharide synthesis sequences of the *Klebsiella pneumoniae* strains examined in this paper are AB371289–AB371295.

Three supplementary tables, listing primers used in this study, genes encoding membrane or periplasmic proteins in the *cps* gene clusters identified using the EZ-Tn5 <blaM/R6K;Ori> transposon, and predicted *cps* genes and CPS structures, are available with the online version of this paper.

INTRODUCTION

Klebsiella pneumoniae is one of the common pathogens among the Gram-negative bacteria that cause respiratory and urinary tract infections, liver abscesses and bacteraemia (Podschun & Ullmann, 1998; Fang *et al.*, 2007). *K. pneumoniae* can produce capsular polysaccharides (CPSs, K antigen) and lipopolysaccharides (LPSs, O antigen), both of which are important virulence factors (Podschun & Ullmann, 1998). The functions of CPSs are serum resistance and antiphagocytosis (Podschun & Ullmann,

1998). Among the 77 distinct K serotypes of *K. pneumoniae* (Mori *et al.*, 1989), K1 and K2 are common causative agents isolated from patients with *K. pneumoniae* infections in Singapore, Taiwan and South Africa (Fung *et al.*, 2002; Yu *et al.*, 2007).

K. pneumoniae is phylogenetically related to *Escherichia coli*. Both *K. pneumoniae* and *E. coli* produce a variety of capsules (Podschun & Ullmann, 1998; Jann & Jann, 1997). The genetic organization of the capsular polysaccharide synthesis (*cps*) region and CPS assembly in *E. coli* have been well studied (reviewed by Whitfield, 2006; Whitfield & Roberts, 1999). *E. coli* K antigens are divided into four groups and the group 1 CPSs are synthesized by the Wzy-dependent polymerization pathway. Previous studies indicate that the mechanism of synthesis of *K. pneumoniae* capsules and *K. pneumoniae cps* gene clusters is similar to that of the *E. coli* group I CPSs (Arakawa *et al.*, 1995; Rahn *et al.*, 1999; Chuang *et al.*, 2006; Pan *et al.*, 2008). The genes in the group I *cps* gene clusters can be classified into three groups, (i) genes for sugar nucleotide synthesis, (ii) genes for capsule repeat-unit synthesis and (iii) genes for capsular repeat-unit assembly and export. Moreover, it is well documented that genes encoding proteins for capsule repeat-unit synthesis, assembly and export form an operon (Drummelsmith & Whitfield, 1999).

The synthesis of the *E. coli* K30 (group I CPS) antigen serves as a model for CPS biogenesis in *K. pneumoniae* (Whitfield & Paiment, 2003). Four glycosyltransferases (GTs) are involved in the synthesis of K30 lipid-linked antigen repeat units. WbaP, a membrane protein, transfers the Gal from UDP-Gal to an undecaprenol diphosphate (UndP) to form Gal-p-UndP. Three other GTs (WbaZ, WcaO and WcaN) then transfer three sugar moieties to Gal-p-UndP sequentially to produce the capsule repeat unit. GTs play a significant role in the biosynthesis of CPS and O antigen repeat units. GTs can be classified into two structural superfamilies and 65 families based on different criteria (Franco & Rigden, 2003; Coutinho *et al.*, 2003). The diverse enzymic activities of GTs provide the remarkable diversity of K and O antigens. The membrane proteins encoded by *wza* and *wzc* are located at the 5' end of *E. coli* group I and *K. pneumoniae cps* gene clusters (Arakawa *et al.*, 1995; Rahn *et al.*, 1999; Chuang *et al.*, 2006; Pan *et al.*, 2008). Wza is an integral outer membrane lipoprotein and Wzc functions as an inner membrane tyrosine autokinase. Wza and Wzc form a channel complex that acts as a K antigen transporter (Dong *et al.*, 2006; Whitfield, 2006). It has been suggested that Wzb, the cognate phosphatase of Wzc, regulates the function of Wzc through its phosphatase activity (Whitfield, 2006). Previous studies (Rahn *et al.*, 1999; Chuang *et al.*, 2006; Pan *et al.*, 2008) have revealed that the genes responsible for K antigen repeat-unit biosynthesis and polymerization located in the central region (between *wzc* and *gnd*) of *cps* gene clusters are highly diverse.

The full-length *cps* gene clusters of the *K. pneumoniae* strains NTUH-K2044 (Chuang *et al.*, 2006; Wu *et al.*,

2009), Chedid (Arakawa *et al.*, 1995) and MGH78578 (http://genome.wustl.edu/genomes/view/klebsiella_pneumoniae/) have been sequenced. Twenty, 18 and 22 ORFs have been predicted in the K1, K2 and K52 *cps* gene clusters, respectively. Recently, the *cps* gene clusters (from *galF* to *gnd*) of *K. pneumoniae* A1142 and A1517, each consisting of 16 ORFs, have also been sequenced and characterized (Pan *et al.*, 2008).

In this study, we sequenced seven *cps* gene clusters from clinical isolates of *K. pneumoniae*. Comparative analysis reveals that these *cps* gene clusters of *K. pneumoniae* are highly diverse. The number of GT genes and the DNA sequence of these GT genes (except *wbaP* and *wcaJ*) are specific to each *cps* gene cluster. Notably, the first GT gene that transfers sugar from activated sugar nucleotides to UndP present in each *cps* gene cluster is either *wbaP* or *wcaJ*. The information provided by this study could be used to investigate the evolution of *cps* gene clusters in *K. pneumoniae*.

METHODS

The *cps* gene clusters of 12 *K. pneumoniae* strains, NTUH-K2044, Chedid, MGH78578, NK8, NK29, NK245, VGH404, VGH484, VGH916, VGH698, A1142 and A1517, are named in this paper as *cps*_{KpK1}, *cps*_{KpK2}, *cps*_{KpK52}, *cps*_{KpA}, *cps*_{KpB}, *cps*_{KpC}, *cps*_{KpK5}, *cps*_{KpK9}, *cps*_{KpK14}, *cps*_{KpK62}, *cps*_{KpK57} and *cps*_{KpKpA1517}, respectively.

The sequences of *cps*_{KpK1}, *cps*_{KpA}, *cps*_{KpB} and *cps*_{KpC} were determined by whole-genome shotgun sequencing (Wu *et al.*, 2009; K. M. Wu and others, unpublished data). The *cps* sequence of *K. pneumoniae* NTUH-K2044 has been independently reported (Chuang *et al.*, 2006). The genomic segments containing *cps*_{KpK5}, *cps*_{KpK9}, *cps*_{KpK14} and *cps*_{KpK62} were obtained by fosmid library construction, PCR screening and PCR cloning to generate genomic sequences of the corresponding clones. Sequences for *cps*_{KpK2}, *cps*_{KpK57} and *cps*_{KpKpA1517} were downloaded from GenBank, and the sequence of *cps*_{KpK52} is available at http://genome.wustl.edu/genomes/view/klebsiella_pneumoniae/.

Bacterial strains, plasmids and growth conditions. *K. pneumoniae* strains and plasmids used in this study are listed in Table 1. *K. pneumoniae* NK8, NK29 and NK245 were collected from the Department of Pathology, National Cheng Kung University Hospital, Tainan, Taiwan (Wu *et al.*, 2009). *K. pneumoniae* VGH525, VGH404, VGH484, VGH916 and VGH698 were collected from Veterans General Hospital, Taipei, Taiwan (Fung *et al.*, 2000). *E. coli* EPI300 (Epicentre) was used to construct fosmid libraries. *E. coli* and *K. pneumoniae* were cultured in Luria broth (LB) or on LB agar plates (Miller, 1972) at 37 °C.

Fosmid library construction and PCR screening method. For fosmid library construction, the genomic DNA of *K. pneumoniae* strains NK8, NK29, NK245, VGH404, VGH484, VGH916 and VGH698 was prepared using the Wizard Genomic DNA Purification System (Promega) following the manufacturer's instructions. The fosmid libraries were constructed with a CopyControl Fosmid Library Production kit (Epicentre) according to the manufacturer's instructions. Briefly, genomic DNA was first sheared into approximately 40 kb fragments. The sheared DNA was end-repaired to generate blunt, 5'-phosphorylated ends and then size-selected by and recovered from a low-melting-point agarose gel. Finally, the size-selected DNA was ligated with the cloning-ready CopyControl pCC1FOS vector, packaged using ultra-high-efficiency

Table 1. Bacterial strains and plasmids used in this study

Strain or plasmid	Description	Source or reference
<i>K. pneumoniae</i> strains		
Serotype/site of infection/date of isolation		
NK8	Undefined/liver abscess/9 September 1996	Wu <i>et al.</i> (2009)
NK29	Undefined/urinary tract infection/9 February 1999	Wu <i>et al.</i> (2009)
NK245	Undefined/urinary tract infection/14 January 2002	Wu <i>et al.</i> (2009)
VGH525	K2/pneumonia/1 December 1992	Fung <i>et al.</i> (2000)
VGH404	K5/wound infection/4 November 1992	Fung <i>et al.</i> (2000)
VGH484	K9/wound infection/21 November 1992	Fung <i>et al.</i> (2000)
VGH916	K14/bacteraemia/13 May 1993	Fung <i>et al.</i> (2000)
VGH698	K62/pneumonia/4 February 1993	Fung <i>et al.</i> (2000)
<i>E. coli</i> strain		
EPI300	[F ⁻ <i>mcrA</i> Δ (<i>mrr-hsdRMS-mcrBC</i>) ϕ 80 Δ lacZAM15 Δ lacX74 <i>recA1 endA1 araD139</i> Δ (<i>ara, leu</i>)7697 <i>galU galK</i> λ ⁻ <i>rpsL nupG trfA tonA</i>]	Epicentre
Plasmids		
pCC1FOS	Fosmid cloning vector; Cm ^r	Epicentre
pCCK5CPS	Fosmid clone containing the partial <i>cps</i> gene cluster of strain VGH404; insert size 35 915 bp	This study
pCCK9CPS	Fosmid clone containing the full <i>cps</i> gene cluster of strain VGH484; insert size 36 834 bp	This study
pCCK14CPS	Fosmid clone containing the partial <i>cps</i> gene cluster of strain VGH916; insert size 38 343 bp	This study
pCCK62CPS	Fosmid clone containing the full <i>cps</i> gene cluster of strain VGH698; insert size 31 663 bp	This study
pCCKPA	Fosmid clone containing the full <i>cps</i> gene cluster of strain NK8; insert size 43 617 bp	This study
pCCKPB	Fosmid clone containing the full <i>cps</i> gene cluster of strain NK29; insert size 39 210 bp	This study
pCCK2	Clone of PCR product of PR01 and PR02 in pCC1FOS; insert size 4913 bp	This study
pCCK5-1	Clone of PCR product of PR03 and PR04 in pCC1FOS; insert size 7634 bp	This study
pCCK5-2	Clone of PCR product of PR05 and PR06 in pCC1FOS; insert size 5428 bp	This study
pCCK14-1	Clone of PCR product of PR07 and PR04 in pCC1FOS; insert size 13 150 bp	This study
pCCK14-2	Clone of PCR product of PR08 and PR09 in pCC1FOS; insert size 1426 bp	This study
pCCK14-3	Clone of PCR product of PR01 and PR06 in pCC1FOS; insert size 7949 bp	This study

MaxPlax Lambda Packaging Extracts, and used to transfect *E. coli* EPI300. Transfected cells were plated on agar plates containing chloramphenicol (12.5 µg ml⁻¹).

Fosmid clones were then picked and placed in LB medium containing chloramphenicol in 96-well microtitre plates (stock plates) and incubated overnight at 37 °C with shaking. The overnight bacterial culture from eight wells down a column of a 96-well microtitre plate (10 µl per well) was pooled (column pool) and stored in a single well of 96-well microtitre plates (pool plates). A 2 µl sample of bacterial culture from each well of the pool plates was used for PCR amplification. WZI1 and WZI2 primers (sequences listed in Supplementary Table S1) were used to identify fosmid clones containing *cps* gene clusters. Herculase DNA polymerase (Stratagene) was used for the PCR. The PCR products were examined in a 1% agarose gel. Individual fosmids from a PCR-positive well of the pool plate were purified and fosmid end-sequence pairs were sequenced to determine which fosmid clone contains the *cps* region.

Fosmids pCCKPA, pCCKPB, pCCK9CPS and pCCK62CPS containing full-length *cps* were identified by PCR screening of fosmid libraries of *K. pneumoniae* NK8 (*cps*_{KPA}), NK29 (*cps*_{KPB}), VGH484 (*cps*_{K9}) and VGH698 (*cps*_{K62}), respectively (Fig. 1). Fosmid pCCK5CPS, and pCCK14CPS containing partial *cps* (5' end), were identified from fosmid libraries of *K. pneumoniae* VGH404 (*cps*_{K5}) and VGH916 (*cps*_{K14}) (Fig. 1).

Cloning the 3' region of *cps* of *K. pneumoniae* strains VGH404, VGH916 and VGH525. After PCR screening, fosmid clone pCCK5CPS was identified from the fosmid library of *K. pneumoniae* VGH404 (serotype K5). pCCK5CPS contains a partial *cps*_{K5} gene cluster. Therefore, in order to get the full-length of the *cps*_{K5} gene cluster, the primer pairs PR03/PR04 and PR05/PR06 (Supplementary

Table S1) were used to amplify *cps*_{K5} from *K. pneumoniae* VGH404 genomic DNA. The amplified DNA fragments were then cloned into linearized pCC1FOS (Epicentre) to generate the plasmids pCCK5-1 and pCCK5-2. Overlapping of the inserts from pCCK5CPS, pCCK5-1 and pCCK5-2 made up the full-length of the *cps*_{K5} gene cluster (Fig. 1).

Fosmid clone pCCK14CPS identified by PCR screening contained partial *cps*_{K14}. To get full-length *cps*_{K14}, primer pairs PR07/PR04, PR08/PR09 and PR01/PR06 (Supplementary Table S1) were used to amplify the fragment from the *K. pneumoniae* VGH916 (serotype K14) genome. The PCR products were then ligated with pCC1FOS to generate pCCK14-1, pCCK14-2 and pCCK14-3. As a result, overlapping of the inserts from pCCK14CPS, pCCK14-1, pCCK14-2 and pCCK14-3 made up the full-length of the *cps*_{K14} gene cluster (Fig. 1).

Though the *cps* gene cluster sequence of *K. pneumoniae* K2 (*cps*_{K2}) has been reported by Arakawa *et al.* (1995), the 3' end of *cps*_{K2} has not been determined. We used primers PR01 and PR02 (Supplementary Table S1) to amplify the 3' region of *cps*_{K2} from *K. pneumoniae* VGH525 (serotype K2), and the PCR product was ligated with pCC1FOS to generate pCCK2 (Fig. 1).

Plasmid and fosmid DNA purification and sequencing. For plasmid and fosmid DNA purification, bacteria were grown according to the instructions of the CopyControl Fosmid Library Production kit (Epicentre). Plasmid and fosmid DNA from 1 ml bacterial culture was purified with a Montage Plasmid Miniprep96 kit (Millipore).

To generate sequencing templates, plasmids and fosmids were tagged with the EZ-Tn5 <KAN-2> transposon according to the manufac-

Fig. 1. Recombinant clones containing *cps* gene clusters of *K. pneumoniae*. Horizontal lines represent the inserts of plasmids. The vertical lines indicate the boundary of the *cps* gene cluster (from *galF* to *ugd*). The dashed line represents the DNA sequence from K2 CPS of *K. pneumoniae* Chedid (accession number D21242). The *cps* gene cluster of *K. pneumoniae* VGH404 (*cps*_{kpK5}) and VGH916 (*cps*_{kpK14}) is represented by three (pCCK5CPS, pCCK5-1 and pCCK5-2) and four (pCCK14CPS, pCCK14-1, pCCK14-2 and pCCK14-3) clones, respectively.

turer's instructions (Epicentre). Tagged plasmids and fosmids were electroporated into *E. coli* EPI300 and purified with the Montage Plasmid Miniprep96 kit.

The sequencing reactions were carried out with BigDye Terminator v3.1 following the standard procedure and the products were analysed with an ABI 3730xl DNA analyser (Applied Biosystems). The Phred/Phrap/Consed package (Ewing & Green, 1998; Gordon *et al.*, 1998) was used for sequence assembling and editing.

Coding gene prediction and sequence analysis. The protein-coding genes were predicted using Glimmer 2.13, GeneMark 2.4 and GeneMark.hmm 2.1 (Delcher *et al.*, 1999; Borodovsky & McIninch, 1993; Lukashin & Borodovsky, 1998) and those shorter than 30 translated amino acids were abandoned. The gene name, description, probable function and Clusters of Orthologous Groups (COG) group of each predicted gene were assigned according to the results of BLASTP (E-value $<10^{-5}$ and identity $>40\%$ and matched length $>30\%$) against RefSeq.microbial (Pruitt *et al.*, 2007), non-redundant protein and COG databases from NCBI (<http://www.ncbi.nlm.nih.gov>). Subsequent manual inspections were applied to correct errors from automatic annotation and information retrieval. Those genes without matched gene names within the *cps* variable region were denominated with the corresponding serotype following the alphabet, except those in strains NK8, NK29 and NK245, since their serotypes are not yet identified.

The protein family of each protein was predicted by searching the Pfam protein families database (Finn *et al.*, 2006; Bateman *et al.*, 2004), and the program MacVector (Olson, 1994) was used for pairwise sequence comparison.

Identification of novel membrane and periplasmic proteins. An EZ-Tn5 β -Lactamase Fusion kit (Epicentre) was used to identify genes encoding membrane and periplasmic proteins. Fosmids pCCKPA, pCCKPB, pCCK9CPS and pCCK62CPS were tagged with the EZ-Tn5 $\langle blaM/R6K\gamma ori \rangle$ transposon *in vitro*. Tagged fosmids were transformed into *E. coli* Epi300 by electroporation and plated on an LB plate containing $50 \mu\text{g}$ ampicillin ml^{-1} . The fosmid DNA of transformants was isolated and the $\langle blaM/R6K\gamma ori \rangle$ transposon insertion site of each fosmid was determined by sequencing using primer BLA-RP2 (5'-CTTCGGGGCGAAAACTCTCAAGGATC-3'). Ninety-six samples of each fosmid (pCCKPA, pCCKPB, pCCK9CPS and pCCK62CPS) were sequenced.

RESULTS AND DISCUSSION

General features of *cps* gene clusters of *K. pneumoniae*

Seven newly sequenced and five already sequenced *cps* of *K. pneumoniae* were analysed in this paper. The *cps* of *K. pneumoniae* NTUH-K2044, Chedid, MGH78578, NK8, NK29, NK245, VGH404, VGH484, VGH916, VGH698, A1142 and A1517 are named as *cps*_{kpK1}, *cps*_{kpK2}, *cps*_{kpK52}, *cps*_{kpA}, *cps*_{kpB}, *cps*_{kpC}, *cps*_{kpK5}, *cps*_{kpK9}, *cps*_{kpK14}, *cps*_{kpK62}, *cps*_{kpK57} and *cps*_{kpKpA1517}, respectively. The genetic organization of the *K. pneumoniae* *cps* gene clusters are shown in Fig. 2(a). In general, the *cps* region of *K. pneumoniae* begins at *galF* and ends at *ugd*. The length of the *K. pneumoniae* *cps* gene clusters ranges from 21 to 30 kb (Table 2). The number of ORFs found in the *cps* region of 12 *K. pneumoniae* *cps* gene clusters ranges from 16 to 25 (Table 2). The transcriptional direction of all ORFs is from *galF* to *ugd*, except for *kpc7* of *cps*_{kpC} (Fig. 2a). The GC content of each ORF is also shown in Fig. 2(a). The GC content of K antigen gene clusters is lower than the mean GC content (57.7%) of the *K. pneumoniae* NTUH-K2044 genome (Wu *et al.*, 2009), especially in the region between *wzc* and *gnd* (Fig. 2a).

All of the *cps* gene clusters in *K. pneumoniae* contain six conserved genes at the 5' end of the *cps* region in the following order: *galF*, *orf2*, *wzi*, *wza*, *wzb* and *wzc*. Genes encoding glucose-6-phosphate dehydrogenase (*gnd*) and UDP-glucose dehydrogenase (*ugd*) were found downstream of all *cps* gene clusters except *cps*_{kpK57} and *cps*_{kpA1517}, in which the *ugd* is still unknown (Fig. 2a). Our data also reveal that most *gnd* and *ugd* are separated by *manCB*, *rmlBADC* or *manCB**rmlBADC* genes (Fig. 2a).

Sugar nucleotide synthesis genes in the *cps* clusters

Mannose, rhamnose and glucuronic acid are common components of capsular and O antigens. The genes

Fig. 2. Organization of *cps* gene clusters of 12 *K. pneumoniae* isolates (a) and putative promoters and transcriptional units of the *K. pneumoniae cps* (b). (a) The arrows represent the direction of transcription and dark-grey arrows represent putative GT genes identified by Pfam. The percentage below each ORF indicates the G + C content of individual genes. (b) The thick lines represent transcriptional units of *K. pneumoniae cps*. The polycistronic mRNAs are driven by promoters P1, P2 and P3. The polycistronic mRNA driven by P1 consists of *galF* and *orf2*. The polycistronic mRNA driven by P2 includes genes of capsule repeat unit synthesis and polymerization as well as surface assembly. The transcriptional direction of *ugd* is opposite to that of *uge*. For clarity, only four conserved genes of *K. pneumoniae cps* are shown. ♦, JUMPStart sequence; T, stem-loop structure.

(Gmd) and GDP-4-keto-6-deoxy-D-mannose 3,5-epimerase/reductase (WcaG) are needed for the conversion of GDP-D-mannose to GDP-L-fucose, the precursor of fucose. The colanic acid of the capsule of both *E. coli* and *K. pneumoniae* K1 contains fucose. The *gmd* and *wcaG* genes are located in the gene clusters responsible for *E. coli* colanic acid (Stevenson *et al.*, 1996; Andrianopoulos *et al.*, 1998) and *K. pneumoniae* K1 capsule biosynthesis (Fig. 2a) (Chuang *et al.*, 2006). The enzyme homologous to WcaG is also found in humans (formerly called FX protein) (Tonetti *et al.*, 1996). The WcaG involved in the *K. pneumoniae* K1 capsule synthesis shows 29 % identity to FX protein (accession number AAC50786) and 77 % identity to the corresponding gene product (accession number AAC77843) of the *E. coli* colanic acid biosynthesis gene cluster.

The protein encoded by *glf* found in *K. pneumoniae* VGH916 (K14) shows 42 % identity to Glf (accession number AAC98417) of *K. pneumoniae* serotype O1 and 63 % identity to the Glf (accession number AAB88403) of *E. coli* K-12. Glf catalyses the conversion of UDP-galactopyranose into UDP-galactofuranose (Nassau *et al.*, 1996; Lee *et al.*, 1996). The enzymic activity (Nassau *et al.*, 1996; Lee *et al.*, 1996) and crystal structure (Sanders *et al.*, 2001) of Glf of *E. coli* have been characterized. The presence of a *glf* homology suggested that *K. pneumoniae* K14 capsule contains galactofuranose sugar. Owing to the absence of Glf in humans, Glf is useful in drug development (Lee *et al.*, 1996).

Membrane and periplasmic proteins in the *cps* gene clusters of *K. pneumoniae*

Membrane proteins are important in the biosynthesis of *E. coli* group 1 and *K. pneumoniae* CPSs. Wzi, Wza, Wzc, Wzx, Wzy, WbaP and WcaJ are well-characterized membrane proteins (Stevenson *et al.*, 1996; Whitfield & Roberts, 1999; Rahn & Whitfield, 2003; Beis *et al.*, 2004; Collins *et al.*, 2006; Dong *et al.*, 2006; Whitfield, 2006). The gene organization and sequence of *wzi*, *wza* and *wzc* are conserved among *E. coli* group 1 and *K. pneumoniae cps* gene clusters (Fig. 2a). On the other hand, the sequences of *wzx* and *wzy* in the *cps* gene clusters are more diverse. Therefore, the transposon containing the leaderless and promoterless β -lactamase gene was used to characterize the genes encoding Wzx, Wzy and novel membrane proteins of *K. pneumoniae cps*. The EZ-Tn5 <*blaM*/R6K γ ori> transposon was used to identify the genes encoding membrane and periplasmic proteins in *cps*_{KpA}, *cps*_{KpB}, *cps*_{KpK9} and *cps*_{KpK62}. Fosmids pCCKPA, pCCKPB, pCCK9CPS and pCCK62CPS containing the full-length *cps* gene cluster of *cps*_{KpA}, *cps*_{KpB}, *cps*_{KpK9} and *cps*_{KpK62}, respectively, were tagged with the EZ-Tn5 <*blaM*/R6K γ ori> transposon *in vitro*, transformed into *E. coli* Epi300, and then plated on LB plates with ampicillin. Genes encoding membrane or periplasmic proteins fused in-frame with β -lactamase allow *E. coli* to survive on plates with ampicillin. Ninety-six

transformants of each fosmid were purified and sequenced. Sequencing analysis revealed that 79, 80, 78 and 80 transposon insertion sites were successfully identified for pCCKPA, pCCKPB, pCCK9CPS and pCCK62CPS, respectively (Supplementary Table S2). By using *blaM*, we identified eight genes (*wzi*, *wza*, *wzc*, *wbaP*, *wzy*, *orfZ*, *wzx* and *manC*) of *cps*_{KpA}, six genes (*wzi*, *wza*, *wzc*, *wbaP*, *kpb2* and *gnd*) of *cps*_{KpB}, six genes (*wzi*, *wza*, *wzc*, *kp9D*, *wzx* and *rmlA*) of *cps*_{KpK9}, and nine genes (*orf2*, *wzi*, *wza*, *wzc*, *wbaP*, *kp62C*, *wzy*, *kp62F* and *kp62I*) of *cps*_{KpK62} that encoded membrane or periplasmic proteins (Supplementary Table S2). Wzi and Wza were identified more frequently than the other genes. WbaP is a membrane protein with four hydrophobic segments at the N terminal (Stevenson *et al.*, 1996; Jiang *et al.*, 2001; Whitfield, 2006). In *cps*_{KpA}, *wzx* and *wzy* were identified in four out of 79 transformants. *wbaP* was also identified in three out of four *cps* including *cps*_{KpA}, *cps*_{KpB} and *cps*_{KpK62} at low frequency (Supplementary Table S2).

wzx and *wzy* genes of *K. pneumoniae cps* gene clusters

Synthesis of *E. coli* groups 1 and 4 and *K. pneumoniae* CPSs are Wzy-dependent (Whitfield, 2006). Wzx functions as a flippase and exports repeat-unit oligosaccharides to the periplasmic region. Wzy in turn polymerizes these repeating units (Whitfield & Roberts, 1999). It is known that the sequence similarity of *wzx* and *wzy* is very low (Jiang *et al.*, 2001), and because of this low similarity, *wzx* and *wzy* genes should be serotype-specific. Therefore, serotyping by PCR targeting *wzx* and *wzy* for *E. coli* O antigen and *Streptococcus pneumoniae* K antigen has been developed (DebRoy *et al.*, 2005; Kong *et al.*, 2005). Due to the low-level of similarity of *wzx* and *wzy* genes among different gene clusters, sequence similarity and transmembrane helix profiles have been used in *wzx* and *wzy* identification (Jiang *et al.*, 2001; Mazur *et al.*, 2005; Nakhamchik *et al.*, 2007; Cunneen & Reeves, 2008).

In this study, protein families of putative Wzx and Wzy in *K. pneumoniae cps* gene clusters were analysed by searching the Pfam database (Finn *et al.*, 2006; Bateman *et al.*, 2004). Additionally, the sequence similarity of Wzx and Wzy from *K. pneumoniae cps* was analysed by BLASTP and PSI-BLAST (Altschul *et al.*, 1997), while transmembrane helices were predicted using HMMTOP (Tusnády & Simon, 1998) and TMPred (Hofmann & Stoffel, 1993) (Table 3). Proteins listed in Table 3 were found to contain a number of transmembrane helices ranging from nine to 14 (predicted by HMMTOP). The Pfam protein family and homologue of each protein were determined by searching the Pfam and GenBank databases. Proteins that belong to the Pfam families Polysacc_synt (PF01943) and Wzy_C (PF04932), were named as Wzx and Wzy, respectively. Other Wzx and Wzy proteins with no significant match to a Pfam family were identified according to their homologues. We found that nine *cps* gene clusters that contained ORFs encoding Wzx belonged to the Polysacc_synt family (Table 3).

Table 3. Predicted genes for *wzx* and *wzy* of *K. pneumoniae cps*

<i>cps</i>	ORF	Product size (aa)	Homologue (accession number)	Sequence identity*	Pfam family (Pfam number)†	Number of transmembrane helices‡
<i>cps_{kpK1}</i>	<i>wzx</i>	429	O-antigen flippase of <i>Francisella tularensis</i> (YP_170390)	20 %	No significant matches	12
	<i>magA (wzy)</i>	408	Putative polysaccharide polymerase of <i>Staphylococcus saprophyticus</i> (BAE17214)	25 %	Wzy_C (PF04932)	12
<i>cps_{kpA}</i>	<i>wzy</i>	410	Repeat unit polymerase Wzy of <i>E. coli</i> (AAD21566)	99 %	Wzy_C (PF04932)	11
<i>cps_{kpB}</i>	<i>wzx</i>	419	Repeat unit exporter Wzx of <i>E. coli</i> (AAD21572)	99 %	Polysacc_synt (PF01943)	12
	<i>kpb2</i>	432	Hypothetical membrane-spanning protein of <i>Fusobacterium nucleatum</i> (AAL93806)	24 %	No significant matches	13
<i>cps_{kpC}</i>	<i>wzx</i>	475	O-antigen flippase of <i>Shigella boydii</i> (AAL27344)	40 %	Polysacc_synt (PF01943)	11
	<i>kpc4</i>	500	Virulence factor MVIN-like of <i>Kineococcus radiotolerans</i> SRS30216 (EAM74568)	30 %	MVIN (PF03023)	13
<i>cps_{kpK5}</i>	<i>wzy</i>	425	O-antigen polymerase of <i>Kineococcus radiotolerans</i> SRS30216 (EAM74567)	26 %	Wzy_C (PF04932)	12
	<i>kp5A</i>	403	Hypothetical protein of <i>Shewanella amazonensis</i> SB2B (EAN38540)	23 %	No significant matches	11
<i>cps_{kpK9}</i>	<i>wzx</i>	480	Polysaccharide biosynthesis protein of <i>Geobacter metallireducens</i> GS-15 (ABB31724)	22 %	Polysacc_synt (PF01943)	14
	<i>kp9D</i>	390	No hits found		No significant matches	12
<i>cps_{kpK14}</i>	<i>wzx</i>	420	Repeat unit exporter Wzx of <i>E. coli</i> K30 (AAD21572)	43 %	Polysacc_synt (PF01943)	12
	<i>wzx</i>	508	Putative transmembrane protein of <i>Bacteroides fragilis</i> NCTC 9343 (YP_212406)	40 %	Polysacc_synt (PF01943)	14
<i>cps_{kpK62}</i>	<i>kp14E</i>	402	Hypothetical protein of <i>Bacteroides ovatus</i> ATCC 8483 (EDO11962)	25 %	No significant matches	11
	<i>wzy</i>	388	Oligosaccharide repeat unit polymerase Wzy of <i>Streptococcus pneumoniae</i> (CAI32727)	24 %	Wzy_C (PF04932)	12
<i>cps_{kpK2}</i>	<i>kp62I</i>	437	Virulence factor MVIN-like of <i>Kineococcus radiotolerans</i> SRS30216 (EAM74568)	25 %	MVIN (PF03023)	13
	<i>orf10 (wzy)</i>	429	Wzy of <i>E. coli</i> (ACA24882)	26 %	No significant matches	11
<i>cps_{kpK52}</i>	<i>orf11 (wzx)</i>	485	Colanic acid exporter of <i>Salmonella typhimurium</i> LT2 (NP_461047)	47 %	Polysacc_synt (PF01943)	14
	<i>wzy</i>	374	O-antigen polymerase of <i>S. enterica</i> (YP_217085)	26 %	No significant matches	9
<i>cps_{kpK57}</i>	<i>wzx</i>	481	Wzx of <i>Shigella boydii</i> (AAS98036)	48 %	Polysacc_synt (PF01943)	14
	<i>wzx</i>	487	Polysaccharide biosynthesis protein of <i>Serratia proteamaculans</i> 568 (YP_001477823)	49 %	Polysacc_synt (PF01943)	13
<i>cps_{kpA1517}</i>	<i>wzy</i>	392	O-antigen polymerase family of <i>Rhodobacterales</i> bacterium Y4I (YP_002697020)	25 %	Wzy_C (PF04932)	11
	<i>orf12 (wzy)</i>	390	Wzy of <i>E. coli</i> (ACA24803)	23 %	No significant matches	12
	<i>orf14 (wzx)</i>	438	Wzx of <i>E. coli</i> (AAV74382)	28 %	Polysacc_synt (PF01943)	12

*Determined by BLASTP.

†Determined by searching Pfam.

‡Predicted by HMMTOP.

Although Wzx of *cps_{kpK1}* did not belong to the Polysacc_synt family, it showed 20 % identity to O-antigen flippase of *Francisella tularensis* (YP_170390) (Chuang *et al.*, 2006). Moreover, Wzx of *cps_{kpK9}* and Wzy of *cps_{kpK62}* were indeed membrane proteins characterized using <blaM/R6K_{Yori}> transposon insertion (Supplementary Table S2). In addition, Kpb2 and Kp62I were identified using transposon insertion (Supplementary Table S2).

Wzx protein, a flippase, is essential for capsule biosynthesis and should be present in *K. pneumoniae cps*, but no Wzx-like protein was identified in *cps_{kpC}* and *cps_{kpK62}* gene clusters. Kpc4 of *cps_{kpC}* and Kp62I of *cps_{kpK62}* belonged to the MVIN (PF03023) protein family, and both MVIN and Polysacc_synt belonged to the MviN_MATE integral membrane protein superfamily. Since no other Wzx-like protein was identified in *cps_{kpC}* and *cps_{kpK62}*, Kpc4 and Kp62I are likely candidates for Wzx.

Synthesis of *K. pneumoniae* capsule is a Wzy-dependent process (Whitfield, 2006). Five out of 12 *cps* gene clusters contain ORFs encoding Wzy which belongs to structural motif Pfam04932 (Wzy_C), including *magA* of *cps*_{KpK1}. Wzy of *cps*_{KpK2}, *cps*_{KpK52} and *cps*_{KpA1517} have low similarity to Wzy of *E. coli* and *Salmonella enterica* (Table 3). Due to the sequence diversity of Wzy, we did not identify Wzy in *cps*_{KpB}, *cps*_{KpK5}, *cps*_{KpK9} and *cps*_{KpK14}. The number of transmembrane helices predicted by HMMTOP in Kpb2, Kp5A, Kp9D and Kp14E was similar to that of Wzy(s) identified in *K. pneumoniae cps* gene clusters (Table 3). Our data suggest that the similarity of *wzx* and *wzy* is extremely low in *K. pneumoniae cps*; therefore, these two genes and GTs can be used for serotyping by PCR.

Transferase genes in *K. pneumoniae* capsule biosynthesis gene clusters

Both WbaP and WcaJ are first GTs that transfer sugar to the lipid acceptor and initiate the synthesis of the capsular repeat (Whitfield, 2006). Either the *wbaP* gene or the *wcaJ* gene existed in a given *K. pneumoniae cps* (Fig. 2a, Table 1). Unlike other GTs, WbaP and WcaJ are membrane proteins with four hydrophobic segments at the N terminal (Stevenson *et al.*, 1996; Jiang *et al.*, 2001; Whitfield, 2006). EZ-Tn5 <*blaM/R6Kγori*> transposon insertion analysis also suggests that the WbaPs are indeed membrane proteins (Supplementary Table S2). Moreover, according to transposon insertion analysis, none of the other putative GTs identified by searching the Pfam database is a membrane protein.

It has been shown that WbaP is the UDP-Gal:undecaprenolphosphate galactose-1-phosphate transferase of *S. enterica* B and *E. coli* K30 (Liu *et al.*, 1993; Drummelsmith & Whitfield, 1999). The WbaPs of *K. pneumoniae* NK8, NK29, NK245, VGH484, VGH698, MGH78578, A1142 and A1517 share 98, 71, 74, 74, 72, 76, 77 and 74 % identity with WbaP of *E. coli* K30 (AAD21565), respectively. A previous study has suggested that WcaJ (AAC77848) functions as a UDP-glucose lipid carrier transferase in the biosynthesis of colanic acid in *E. coli* K-12 (Stevenson *et al.*, 1996). The WcaJs of *K. pneumoniae* NTUH-K2044, VGH404, VGH916 and Chedid share 63, 61, 63 and 62 % identity with WcaJ of *E. coli* K-12 (AAC77848), respectively.

GTs of the *K. pneumoniae cps* gene clusters were identified by searching the Pfam protein families database (Finn *et al.*, 2006; Bateman *et al.*, 2004). Putative GTs determined by searching Pfam are listed in Table 4. The results show that *cps*_{KpK1} contains three, *cps*_{KpA} four, *cps*_{KpB} four, *cps*_{KpC} six, *cps*_{KpK5} two, *cps*_{KpK9} four, *cps*_{KpK14} three, *cps*_{KpK62} five, *cps*_{KpK2} three, *cps*_{KpK52} four, *cps*_{KpK57} four and *cps*_{KpA1517} four genes that encode GTs (Table 4). Earlier studies indicate that the DXD motif is conserved in some GT families (Liu & Mushegian, 2003; Breton *et al.*, 1998; Wiggins & Munro, 1998) and is essential for their enzymic activities (Li *et al.*, 2001; Götting *et al.*, 2004; Klutts *et al.*,

2007; Persson *et al.*, 2007). Amino acid sequence analysis of GTs in *K. pneumoniae cps* gene clusters revealed that most GTs identified by Pfam HMM contained the DXD motifs (Table 4). All WbaPs in the *cps* gene clusters of *K. pneumoniae* were found to contain a conserved DXD motif in the C-terminal region. Moreover, the sequence of the DXD motif was DVD, except in WbaP of *K. pneumoniae* A1142, where it was DID (Table 4). In contrast to WbaP, two out of four WcaJs identified in *K. pneumoniae* VGH404 and Chedid were found to contain a DVD motif (Table 4).

Since the structure of the *K. pneumoniae* K1 capsule is $\rightarrow 4$ -[2,3-(S)-pyruvate]- β -D-GlcpA-(1 $\rightarrow$ 4)- α -L-Fucp-(1 $\rightarrow$ 3)- β -D-Glcp(1 $\rightarrow$) (Chuang *et al.*, 2006), we expect that a pyruvyl transferase (Kp1A) and three GTs (Kp1B, WcaI and WcaJ) are involved in repeating-unit synthesis. We also found that ORF15 of *cps*_{KpA1517} was a putative pyruvyl transferase which belonged to the Pfam PS_pyruv_trans family. In addition to GTs and pyruvyl transferase, acyltransferases were also identified in *K. pneumoniae cps*. Kp1C of *cps*_{KpK1}, ORF13 of *cps*_{KpK2}, Kp5E of *cps*_{KpK5} and OrfY of *cps*_{KpA} belonged to Pfam family Hexapep (PF00132). Some bacterial acetyl and acyltransferases contain Hexapep repeats, including DapD, GlmU and LacA (Williams & Raetz, 2007). In addition, sequence analysis indicated that Kp52I of *cps*_{KpK52} and Kp62F of *cps*_{KpK62} belonged to Pfam family Acyl_transf_3. The Kp52I of *cps*_{KpK52} showed 34 % identity to acyltransferase 3 of *Ralstonia metallidurans* CH34 (ZP_00593471). ORF14 of *cps*_{KpK57} showed 30 % identity to acyltransferase family protein of *Pseudomonas fluorescens* (YP_261319). However, acylation of *Klebsiella* K52 CPS was not noted in an earlier study (Stenutz *et al.*, 1997). In contrast to K52, *Klebsiella* K5 CPS is acetylated at glucopyranose (Cescutti *et al.*, 1995). In this study we were unable to clarify whether *cps*_{KpK5} contains gene(s) responsible for *Klebsiella* K5 CPS acetylation. The relationship between genes and acylation in *K. pneumoniae* CPS biosynthesis still requires further investigation.

Lateral gene transfer and recombination

Both *E. coli* group 1 *cps* and the *cps* gene clusters in *K. pneumoniae* contain six (*galF*, *orf2*, *wzi*, *wza*, *wzb* and *wzc*) and two conserved genes (*gnd* and *ugd*) at the 5' end and 3' end, respectively. The average pairwise sequence identities of *galF*, *orf2*, *wzi*, *wza*, *wzb*, *wzc*, *gnd* and *ugd* genes were 97.7 ± 2.0 , 95.0 ± 3.2 , 87.0 ± 3.7 , 80.1 ± 2.6 , 65.5 ± 4.3 , 66.2 ± 4.3 , 92.9 ± 3.2 and 97.3 ± 0.6 %, respectively (Fig. 4). Average pairwise sequence identities of *galF*, *orf2*, *wzi*, *wza*, *wzb*, *wzc* and *gnd* were results from 12 *cps* gene clusters of *K. pneumoniae*, whereas the value for *ugd* was the result from 10 *cps* gene clusters, not including *cps*_{KpK57} and *cps*_{KpKpA1517} of *K. pneumoniae* A1142 and A1517, respectively. The average pairwise sequence identities of the 5'-end conserved genes (*galF*, *orf2*, *wzi*, *wza*, *wzb* and *wzc*) gradually decreased toward the centre of the *cps* gene clusters (Fig. 4). A similar phenomenon has also

Table 4. GT genes of *K. pneumoniae cps*

<i>cps</i>	Gene	Pfam family/accession number	E-value*	DXD motif sequence (positions†)
<i>cps_{kpK1}</i>	<i>kp1B</i>	Glycos_transf_1/PF00534	1.7E-15	DRD (27-29)
	<i>wcaI</i>	Glycos_transf_1/PF00534	4.5E-10	
	<i>wcaJ</i>	Bac_transf/PF02397	7.2E-42	
<i>cps_{kpA}</i>	<i>wbaP</i>	Bac_transf/PF02397	1.2E-139	DVD (177-179), DVDYD (436-440)
	<i>wcaN</i>	Glycos_transf_2/PF00535	3.8E-28	DDDD (91-94), DYD (175-177)
	<i>wcaO</i>	Glycos_transf_1/PF00534	3.3E-18	DRD (272-274), DSD (335-337)
	<i>wbaZ</i>	Glycos_transf_1/PF00534	7.8E-36	DAD (133-135)
<i>cps_{kpB}</i>	<i>wbaP</i>	Bac_transf/PF02397	1.7E-137	DVD (432-434)
	<i>kpb3</i>	Glycos_transf_2/PF00535	3.5E-31	DDDD (90-93)
	<i>kpb4</i>	Glycos_transf_2/PF00535	6.1E-25	DTD (102-104), DDD (136-138)
	<i>Kpb6</i>	Glycos_transf_1/PF00534	1.7E-41	
<i>cps_{kpC}</i>	<i>wbaP</i>	Bac_transf/PF02397	2.0E-135	DVD (435-437)
	<i>kpc2</i>	Glycos_transf_1/PF00534	2.0E-45	DLD (358-360)
	<i>kpc3</i>	Glycos_transf_1/PF00534	1.9E-27	DTD (196-198), DIDKD (215-219)
	<i>kpc5</i>	Glycos_transf_1/PF00534	0.00063	DWD (27-29)
	<i>kpc6</i>	Glycos_transf_1/PF00534	6.9E-40	DAD (51-53)
	<i>kpc10</i>	Glycos_transf_1/PF00534	7.0E-31	DRD (47-49), DAD (184-186)
<i>cps_{kpK5}</i>	<i>kp5H</i>	Glyco_tran_28_C/PF04101	3.7E-05	DFD (124-126)
	<i>wcaJ</i>	Bac_transf/PF02397	1.5E-39	DID (184-186)
<i>cps_{kpK9}</i>	<i>wbaP</i>	Bac_transf/PF02397	7.3E-136	DID (194-199), DVD (435-437)
	<i>kp9B</i>	Glycos_transf_2/PF00535	2.4E-23	
	<i>kp9C</i>	Glycos_transf_2/PF00535	3.1E-14	DGD (37-39), DDDD (92-95), DID (189-191)
	<i>kp9E</i>	Glycos_transf_2/PF00535	3.0E-05	DQD (81-83)
<i>cps_{kpK14}</i>	<i>kp14F</i>	Glycos_transf_2/PF00535	6.1E-18	DDD (92-94), DED (97-99)
	<i>kp14J</i>	Glycos_transf_2/PF00535	1.8E-39	DPD (92-94)
	<i>wcaJ</i>	Bac_transf/PF02397	7.3E-38	
<i>cps_{kpK62}</i>	<i>wbaP</i>	Bac_transf/PF02397	3.30E-136	DVD (435-437)
	<i>kp62B</i>	Glycos_transf_1/PF00534	1.7E-40	
	<i>kp62D</i>	Glycos_transf_1/PF00534	1.3E-35	DDD (234-236)
	<i>kp62G</i>	Glycos_transf_1/PF00534	3.8E-27	DTD (15-17)
	<i>kp62H</i>	Glycos_transf_1/PF00534	1.80E-24	
<i>cps_{kpK2}</i>	<i>orf7</i>	Glycos_transf_1/PF00534	4.0E-29	
	<i>orf8</i>	Glycos_transf_1/PF00534	4.9E-12	DID (340-342)
	<i>wcaJ</i>	Bac_transf/PF02397	7.3E-42	DFD (104-106)
	<i>wbaP</i>	Bac_transf/PF02397	4.1E-133	DPD (197-199), DSD (347-349), DVD (435-437)
<i>cps_{kpK52}</i>	<i>kp52B</i>	Glycos_transf_2/PF00535	1.8E-20	DVD (54-56)
	<i>kp52C</i>	Glycos_transf_1/PF00534	3.8E-13	DQD (108-110), DND (174-176), DID (278-280)
	<i>kp52F</i>	Glycos_transf_2/PF00535	1.1E-33	DDDD (95-98), DYD (178-181)
	<i>wbaP</i>	Bac_transf/PF02397	2.8E-132	DIDYD (436-440)
<i>cps_{kpK57}</i>	<i>orf9</i>	Glycos_transf_1/PF00534	3.0E-32	DND (106-108)
	<i>orf10</i>	Glycos_transf_1/PF00534	3.0E-34	DKD (227-229), DVD (264-266)
	<i>wbaZ</i>	Glycos_transf_1/PF00534	3.40E-33	
	<i>wbaP</i>	Bac_transf/PF02397	3.9E-129	DND (195-197), DKD (347-349), DVD (435-437)
<i>cps_{kpA1517}</i>	<i>orf8</i>	Glycos_transf_2/PF00535	3.9E-20	
	<i>orf9</i>	Glycos_transf_2/PF00535	3.6E-36	DDDD (94-97), DWD (181-183)
	<i>orf10</i>	Glycos_transf_2/PF00535	6.9E-28	DAD (86-88), DQD (94-96)

*E-value was determined by Pfam HMM (Finn *et al.*, 2006; Bateman *et al.*, 2004). E-values less than 0.05 are significant.

†Amino acid position in protein.

been observed in the *Streptococcus pneumoniae* CPS gene clusters (Jiang *et al.*, 2001).

Sequence analysis showed that the pairwise sequence divergence of 12 *wzi*, *wza*, *wzb* and *wzc* genes in *K.*

pneumoniae cps was significantly higher than that of the *gnd* gene. The *gnd* gene, located at the 3' end of the *cps*, encoding an NADP-dependent 6-phosphogluconate dehydrogenase, is involved in the pentose phosphate pathway. 6-Phosphogluconate dehydrogenase converts 6-phospho-

Fig. 4. Average pairwise sequence identities of eight conserved genes in 12 *cps* gene clusters of *K. pneumoniae*. Average pairwise sequence identities of *galF*, *orf2*, *wzi*, *wza*, *wzb*, *wzc* and *gnd* were results from 12 *cps* gene clusters, whereas *ugd* was the result from 10 *cps* gene clusters, not including *cps*_{KpK57} and *cps*_{KpKpA1517} of A1142 and A1517, respectively. The average pairwise sequence identities of *galF*, *orf2*, *wzi*, *wza*, *wzb*, *wzc*, *gnd* and *ugd* were 97.7 ± 2.0, 95.0 ± 3.2, 87.0 ± 3.7, 80.1 ± 2.6, 65.5 ± 4.3, 66.2 ± 4.3, 92.9 ± 3.2 and 97.3 ± 0.6%, respectively.

gluconate into ribulose 5-phosphate and produces NADPH (Nelson & Selander, 1994). The pentose phosphate pathway provides NADPH and ribose 5-phosphate for reductive biosynthesis and nucleotide biosynthesis. The fact that the diversity of *wzi*, *wza*, *wzb*, *wzc* and *gnd* genes was higher than that of other housekeeping genes indicated that the diversity could be due to recombination (Nelson & Selander, 1994). Previously, the capsule of *K. pneumoniae* K20 has been shown to be identical to that of *E. coli* K30 (Laakso *et al.*, 1988; Homonylo *et al.*, 1988), and the capsule of *E. coli* K42 is serologically identical to that of *Klebsiella* K63 (Niemann *et al.*, 1978). In the present study, sequence analysis indicated that the *cps* gene cluster of *K. pneumoniae* of NK8 is identical to that of *E. coli* K30 (accession number AF104912). This suggests that gene substitution has occurred in an existing *cps* gene cluster. These clues and the low GC content (Fig. 2a) strongly suggest that the *K. pneumoniae cps* gene cluster was acquired from a low-GC organism via lateral gene transfer and recombination (Rahn *et al.*, 1999).

K. pneumoniae cps gene expression

Transcriptional regulation of the *cps* gene cluster is important for capsule synthesis and pathogenicity of *K. pneumoniae* (Lai *et al.*, 2003). RmpA2, RcsAB and RfaH modulate the differential expression of the *cps* gene clusters (Lai *et al.*, 2003; Rahn & Whitfield, 2003). Loss of the conjugative plasmid pLVPK, which carries the *rmpA2* gene, dramatically reduces the mucoidy and pathogenicity of *K. pneumoniae* CG43 (Lai *et al.*, 2003). Previous studies suggest that *cps*_{KpK2} includes three promoters (Arakawa *et al.*, 1995; Lai *et al.*, 2003), named as P1, P2 and P3 in this study (Fig. 2b). Since the 12 *K. pneumoniae cps* have conserved genetic organization, the transcriptional organization and regulation mechanism would be similar to that of *cps*_{KpK2}. Promoters P1 and P2 are located upstream of *galF* and *wzi*, respectively (Fig. 2b). The transcript driven by promoter P1 consisted of *galF* and *orf2*. The polycistronic mRNA driven by promoter P2 from *wzi* to *gnd* consisted of genes of capsule repeat unit synthesis and polymerization, as well as surface assembly (Fig. 2b).

An earlier study indicates that a transcriptional terminator structure possibly exists just downstream of *gnd* (Arakawa *et al.*, 1995), suggesting the existence of P3 downstream of *gnd* (Fig. 2b). *uge*, encoding uridine diphosphate galacturonate 4-epimerase, is located downstream of *ugd* in the *K. pneumoniae* NTUH-K2044 and MGH78578 genomes. The transcriptional direction of *ugd* is opposite to that of *uge*, suggesting that transcription by P3 could terminate at a site downstream of *ugd* (Fig. 2b). Since *manCB* and *rmlBADC* could be dispersed between *gnd* and *ugd* in *K. pneumoniae cps*, the possibility of another promoter located downstream of P3 cannot be ruled out.

RmpA2, an activator of capsule biosynthesis, can bind to P1 and P2 and activate gene expression (Lai *et al.*, 2003). An RcsAB box (TAAGATTATTCTCA) was found immediately upstream of the *galF* gene in the *K. pneumoniae* NTUKH2044 (K1) and MGH78578 (K52) genomes. An RcsAB box was also identified in the *K. pneumoniae* Chedid (K2) (Arakawa *et al.*, 1995) and *E. coli* K30 group 1 capsule gene clusters (Rahn & Whitfield, 2003). Meanwhile, *rscA* (KP3552), *rscB* (KP3872) and *rscC* (KP3873) genes are present in the genome of *K. pneumoniae* NTUKH2044 (Wu *et al.*, 2009). We speculate that both RmpA2 and RcsAB are required for high-level capsule synthesis of *K. pneumoniae*.

In *E. coli* K30 *cps*, a transcriptional terminator with a hairpin loop structure has been identified between *wzc* and *wbaP*. Transcription will be influenced by this terminator. RfaH, an antiterminator, is involved in distal genes expression (Rahn & Whitfield, 2003). The JUMPStart sequence was identified in all 12 *K. pneumoniae cps* immediately upstream of *wzi*. Moreover, *rfaH* (KP0198) is present in the genome of *K. pneumoniae* NTUKH2044 (Wu *et al.*, 2009). Therefore, the JUMPStart-RfaH antitermination mechanism could modulate *K. pneumoniae cps* expression.

CPSs of *K. pneumoniae*

Much of the information about CPS synthesis in *K. pneumoniae* is modelled after the *E. coli* pathways. We have predicted gene functions of the available sequences of the

K. pneumoniae cps clusters and correlated them with the CPS structures (Supplementary Table S3). The structural repeating unit of CPSs of *E. coli* K30 consists of $\rightarrow 2$ - α -D-Manp-(1 $\rightarrow$ 3)- β -D-Galp-(1 $\rightarrow$ chains carrying β -D-GlcpA-(1 $\rightarrow$ 3)- α -D-Galp-(1 $\rightarrow$ branches at position 3 of the mannoses (Chakraborty *et al.*, 1980). The tetrasaccharide repeating unit from the K30 capsule of *E. coli* is synthesized on UndP by four GTs (WbaP, WbaZ, WcaO and WcaN) (Whitfield & Paiment, 2003). The repeating units containing mannose are synthesized by the products of *manA*, *manB* and *manC*, of which only *manB* and *manC* are present in *cps*_{KpA}. Although *cps*_{KpK1} also contains *manB* and *manC*, mannose is not included in the structural unit of the K1 capsule (Chuang *et al.*, 2006). However, it consists of fucose, which is synthesized from GDP-D-mannose by Gmd and WcaG. These findings may reinforce the correlation of genes and capsule components. Furthermore, colanic acid of *E. coli* contains two fucose units, and the gene cluster responsible for colanic acid synthesis indeed contains *gmd* and *wcaG* genes (Stevenson *et al.*, 1996; Andrianopoulos *et al.*, 1998).

Since the K1 capsule also contains pyruvate, synthesis of *K. pneumoniae* NTUH-K2044 capsule requires pyruvyl transferase. An earlier study shows that *kp1A* of *cps*_{KpK1} encodes a putative pyruvyl transferase (Chuang *et al.*, 2006). In addition to pyruvyl transferase, three more GTs found in *cps*_{KpK1} (Fig. 2a, Table 3) may be engaged in the synthesis of K1 capsule repeating units by sequential addition of pyruvate and three sugars to UndP.

The structural unit of K2 polysaccharide of *K. pneumoniae* has been characterized as $\rightarrow 3$)- β -D-Glcp-(1 $\rightarrow$ 4)- β -D-Manp-(1 $\rightarrow$ 4)- α -D-Glcp-(1 $\rightarrow$ chains carrying α -D-GlcpA-(1 $\rightarrow$ branches at position 3 of the mannoses (Corsaro *et al.*, 2005). The capsular hexasaccharide repeating unit from *Klebsiella* type 52 contains two rhamnoses (Stenutz *et al.*, 1997). GDP-D-mannose and dTDP-L-rhamnose biosynthesis genes are present in *cps*_{KpK2} and *cps*_{KpK52}, respectively (Fig. 2a). The presence of *manCB* and *rmlBADC* could reflect the existence of mannose and rhamnose in the capsule of K2 and K52. Moreover, the structural unit of K5 polysaccharide of *K. pneumoniae* has been characterized as $\rightarrow 4$)- β -D-GlcpA-(1 $\rightarrow$ 4)- β -D-Glcp-(1 $\rightarrow$ 3)- β -D-Manp-(1 $\rightarrow$ with a pyruvic acetal group acetylated at the C-2 position of the glucopyranose (Cescutti *et al.*, 1995). The GDP-D-mannose biosynthesis gene was found in *cps*_{KpK5} (Fig. 2a). *manCB* and GDP-L-fucose biosynthesis genes (*gmd* and *wcaG*) are both found in *cps*_{KpK1}; however, only fucose is incorporated into the repeat unit of the K1 capsule. In contrast to other *cps* gene clusters of *K. pneumoniae*, except *cps*_{KpK57} and *cps*_{KpA1517}, *cps*_{KpB} does not contain mannose and rhamnose sugar nucleotide biosynthesis genes. Therefore, the capsule of *K. pneumoniae* NK29 might not contain mannose and rhamnose.

The repeating units of capsules usually include uronic acids (Corsaro *et al.*, 2005), and glucuronic acid is found in the repeating unit of K1, K2, K5 and K52 isolated from *K.*

pneumoniae (Cescutti *et al.*, 1995; Stenutz *et al.*, 1997; Corsaro *et al.*, 2005; Chuang *et al.*, 2006). The presence of *ugd* could explain the prevalence of glucuronic acids in the capsules. Moreover, Wzc could increase Ugd activity by phosphorylation (Grangeasse *et al.*, 2003), and phosphorylated Ugd could produce UDP-glucuronic acids from UDP-glucose (Fig. 4). However, galacturonic acid is present in *Klebsiella* K57 CPS instead of glucuronic acid (Pan *et al.*, 2008).

Conclusions

Serotyping has been widely used for the epidemiological study of *K. pneumoniae*. However, the procedure of serotyping is laborious and time-consuming. Recently, methods for the molecular typing of *Klebsiella* species have been developed (Brisse *et al.*, 2004; Fang *et al.*, 2007; Pan *et al.*, 2008) to supplant the existing serotyping method. In principle, molecular typing of *K. pneumoniae* is easier to perform than conventional serotyping. Nevertheless, successful typing requires a panel of probes that can easily detect all common strain types. Our study provides seven new *cps* gene clusters of *K. pneumoniae* strains that will facilitate the development of PCR-based identification methods.

ACKNOWLEDGEMENTS

We are grateful to Dr June Hsieh Wu for critical reading of the manuscript and helpful discussion about the capsule structures. This study was supported by a grant from the National Science Council of Taiwan (NSC97-3112-B-400-005). The sequencing services were provided by the Sequencing Core Facility of the National Research Program for Genomic Medicine supported by a grant from the National Science Council, Taiwan.

REFERENCES

- Albermann, C., Distler, J. & Piepersberg, W. (2000). Preparative synthesis of GDP- β -L-fucose by recombinant enzymes from enterobacterial sources. *Glycobiology* **10**, 875–881.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.
- Andrianopoulos, K., Wang, L. & Reeves, P. R. (1998). Identification of the fucose synthetase gene in the colanic acid gene cluster of *Escherichia coli* K-12. *J Bacteriol* **180**, 998–1001.
- Arakawa, Y., Wacharotayankun, R., Nagatsuka, T., Ito, H., Kato, N. & Ohta, M. (1995). Genomic organization of the *Klebsiella pneumoniae* cps region responsible for serotype K2 capsular polysaccharide synthesis in the virulent strain Chedid. *J Bacteriol* **177**, 1788–1796.
- Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S. & other authors (2004). The Pfam protein families database. *Nucleic Acids Res* **32**, D138–D141.
- Beis, K., Collins, R. F., Ford, R. C., Kamis, A. B., Whitfield, C. & Naismith, J. H. (2004). Three-dimensional structure of Wza, the protein required for translocation of group 1 capsular polysaccharide

- across the outer membrane of *Escherichia coli*. *J Biol Chem* **279**, 28227–28232.
- Borodovsky, M. & McIninch, J. (1993). GENMARK: Parallel gene recognition for both DNA strands. *Comput Chem* **17**, 123–133.
- Breton, C., Bettler, E., Joziassse, D. H., Geremia, R. A. & Imberty, A. (1998). Sequence–function relationships of prokaryotic and eukaryotic galactosyltransferases. *J Biochem* **123**, 1000–1009.
- Brisse, S., Issenhuth-Jeanjean, S. & Grimont, P. A. (2004). Molecular serotyping of *Klebsiella* species isolates by restriction of the amplified capsular antigen gene cluster. *J Clin Microbiol* **42**, 3388–3398.
- Cescutti, P., Osman, S. F., Fett, W. F. & Weisleder, D. (1995). The structure of the acidic exopolysaccharide produced by *Pseudomonas* “gingeri” strain Pf9. *Carbohydr Res* **275**, 371–379.
- Chakraborty, A. K., Friebohn, H. & Stirm, S. (1980). Primary structure of the *Escherichia coli* serotype K30 capsular polysaccharide. *J Bacteriol* **141**, 971–972.
- Chuang, Y. P., Fang, C. T., Lai, S. Y., Chang, S. C. & Wang, J. T. (2006). Genetic determinants of capsular serotype K1 of *Klebsiella pneumoniae* causing primary pyogenic liver abscess. *J Infect Dis* **193**, 645–654.
- Collins, R. F., Beis, K., Clarke, B. R., Ford, R. C., Hulley, M., Naismith, J. H. & Whitfield, C. (2006). Periplasmic protein–protein contacts in the inner membrane protein Wzc form a tetrameric complex required for the assembly of *Escherichia coli* group 1 capsules. *J Biol Chem* **281**, 2144–2150.
- Corsaro, M. M., De Castro, C., Naldi, T., Parrilli, M., Tomas, J. M. & Regue, M. (2005). ^1H and ^{13}C NMR characterization and secondary structure of the K2 polysaccharide of *Klebsiella pneumoniae* strain 52145. *Carbohydr Res* **340**, 2212–2217.
- Coutinho, P. M., Deleury, E., Davies, G. J. & Henrissat, B. (2003). An evolving hierarchical family classification for glycosyltransferases. *J Mol Biol* **328**, 307–317.
- Cunneen, M. M. & Reeves, P. R. (2008). Membrane topology of the *Salmonella enterica* serovar Typhimurium group B O-antigen translocase Wzx. *FEMS Microbiol Lett* **287**, 76–84.
- DebRoy, C., Fratamico, P. M., Roberts, E., Davis, M. A. & Liu, Y. (2005). Development of PCR assays targeting genes in O-antigen gene clusters for detection and identification of *Escherichia coli* O45 and O55 serogroups. *Appl Environ Microbiol* **71**, 4919–4924.
- Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**, 4636–4641.
- Dong, C., Beis, K., Nesper, J., Brunkan-Lamontagne, A. L., Clarke, B. R., Whitfield, C. & Naismith, J. H. (2006). Wza the translocon for *E. coli* capsular polysaccharides defines a new class of membrane protein. *Nature* **444**, 226–229.
- Drummelsmith, J. & Whitfield, C. (1999). Gene products required for surface expression of the capsular form of the group 1 K antigen in *Escherichia coli* (O9a:K30). *Mol Microbiol* **31**, 1321–1332.
- Ewing, B. & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186–194.
- Fang, C. T., Lai, S. Y., Yi, W. C., Hsueh, P. R., Liu, K. L. & Chang, S. C. (2007). *Klebsiella pneumoniae* genotype K1: an emerging pathogen that causes septic ocular or central nervous system complications from pyogenic liver abscess. *Clin Infect Dis* **45**, 284–293.
- Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A. & other authors (2006). Pfam: clans, web tools and services. *Nucleic Acids Res* **34**, D247–D251.
- Franco, O. L. & Rigden, D. J. (2003). Fold recognition analysis of glycosyltransferase families: further members of structural superfamilies. *Glycobiology* **13**, 707–712.
- Fung, C. P., Hu, B. S., Chang, F. Y., Lee, S. C., Kuo, B. I., Ho, M., Siu, L. K. & Liu, C. Y. (2000). A 5-year study of the seroepidemiology of *Klebsiella pneumoniae*: high prevalence of capsular serotype K1 in Taiwan and implication for vaccine efficacy. *J Infect Dis* **181**, 2075–2079.
- Fung, C. P., Chang, F. Y., Lee, S. C., Hu, B. S., Kuo, B. I., Liu, C. Y., Ho, M. & Siu, L. K. (2002). A global emerging disease of *Klebsiella pneumoniae* liver abscess: is serotype K1 an important factor for complicated endophthalmitis? *Gut* **50**, 420–424.
- Giraud, M. F. & Naismith, J. H. (2000). The rhamnose pathway. *Curr Opin Struct Biol* **10**, 687–696.
- Gordon, D., Abajian, C. & Green, P. (1998). Consed: a graphical tool for sequence finishing. *Genome Res* **8**, 195–202.
- Götting, C., Müller, S., Schöttler, M., Schön, S., Prante, C., Brinkmann, T., Kuhn, J. & Kleesiek, K. (2004). Analysis of the DXD motifs in human xylosyltransferase I required for enzyme activity. *J Biol Chem* **279**, 42566–42573.
- Grangeasse, C., Obadia, B., Mijakovic, I., Deutscher, J., Cozzone, A. J. & Doublet, P. (2003). Autophosphorylation of the *Escherichia coli* protein kinase Wzc regulates tyrosine phosphorylation of Ugd, a UDP-glucose dehydrogenase. *J Biol Chem* **278**, 39323–39329.
- Graninger, M., Kneidinger, B., Bruno, K., Scheberl, A. & Messner, P. (2002). Homologs of the Rml enzymes from *Salmonella enterica* are responsible for dTDP- β -L-rhamnose biosynthesis in the Gram-positive thermophile *Aneurinibacillus thermoaerophilus* DSM 10155. *Appl Environ Microbiol* **68**, 3708–3715.
- Hofmann, K. & Stoffel, W. (1993). TMbase – a database of membrane spanning protein segments. *Biol Chem Hoppe Seyler* **374**, 166.
- Homonylo, M. K., Wilmot, S. J., Lam, J. S., MacDonald, L. A. & Whitfield, C. (1988). Monoclonal antibodies against the capsular K antigen of *Escherichia coli* (O9:K30(A):H12): characterisation and use in analysis of K antigen organisation on the cell surface. *Can J Microbiol* **34**, 1159–1165.
- Jann, K. & Jann, B. (1997). Capsules of *Escherichia coli*. In *Escherichia coli: Mechanisms of Virulence*, pp. 113–143. Edited by M. Sussman. Cambridge, UK: Cambridge University Press.
- Jiang, S. M., Wang, L. & Reeves, P. R. (2001). Molecular characterization of *Streptococcus pneumoniae* type 4, 6B, 8, and 18C capsular polysaccharide gene clusters. *Infect Immun* **69**, 1244–1255.
- Klutts, J. S., Lavery, S. B. & Doering, T. L. (2007). A β -1,2-xylosyltransferase from *Cryptococcus neoformans* defines a new family of glycosyltransferases. *J Biol Chem* **282**, 17890–17899.
- Kong, F., Wang, W., Tao, J., Wang, L., Wang, Q., Sabanathan, A. & Gilbert, G. L. (2005). A molecular-capsular-type prediction system for 90 *Streptococcus pneumoniae* serotypes using partial *cpsA-cpsB* sequencing and *wzy*- or *wzx*-specific PCR. *J Med Microbiol* **54**, 351–356.
- Köplin, R., Wang, G., Hötte, B., Priefer, U. B. & Pühler, A. (1993). A 3.9-kb DNA region of *Xanthomonas campestris* pv. *campestris* that is necessary for lipopolysaccharide production encodes a set of enzymes involved in the synthesis of dTDP-rhamnose. *J Bacteriol* **175**, 7786–7792.
- Laakso, D. H., Homonylo, M. K., Wilmot, S. J. & Whitfield, C. (1988). Transfer and expression of the genetic determinants for O and K antigen synthesis in *Escherichia coli* O9:K(A)30 and *Klebsiella* sp. O1:K20, in *Escherichia coli* K12. *Can J Microbiol* **34**, 987–992.
- Lai, Y. C., Peng, H. L. & Chang, H. Y. (2003). RmpA2, an activator of capsule biosynthesis in *Klebsiella pneumoniae* CG43, regulates K2 *cps* gene expression at the transcriptional level. *J Bacteriol* **185**, 788–800.
- Lee, R., Monsey, D., Weston, A., Duncan, K., Rithner, C. & McNeil, M. (1996). Enzymatic synthesis of UDP-galactofuranose and an assay for

- UDP-galactopyranose mutase based on high-performance liquid chromatography. *Anal Biochem* **242**, 1–7.
- Li, J., Rancour, D. M., Allende, M. L., Worth, C. A., Darling, D. S., Gilbert, J. B., Menon, A. K. & Young, W. W., Jr (2001). The DXD motif is required for GM2 synthase activity but is not critical for nucleotide binding. *Glycobiology* **11**, 217–229.
- Liu, J. & Mushegian, A. (2003). Three monophyletic superfamilies account for the majority of the known glycosyltransferases. *Protein Sci* **12**, 1418–1431.
- Liu, D., Haase, A. M., Lindqvist, L., Lindberg, A. A. & Reeves, P. R. (1993). Glycosyl transferases of O-antigen biosynthesis in *Salmonella enterica*: identification and characterization of transferase genes of groups B, C2, and E1. *J Bacteriol* **175**, 3408–3413.
- Lukashin, A. V. & Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* **26**, 1107–1115.
- Ma, Y., Stern, R. J., Scherman, M. S., Vissa, V. D., Yan, W., Jones, V. C., Zhang, F., Franzblau, S. G., Lewis, W. H. & McNeil, M. R. (2001). Drug targeting *Mycobacterium tuberculosis* cell wall synthesis: genetics of dTDP-rhamnose synthetic enzymes and development of a microtiter plate-based screen for inhibitors of conversion of dTDP-glucose to dTDP-rhamnose. *Antimicrob Agents Chemother* **45**, 1407–1416.
- Marolda, C. L. & Valvano, M. A. (1996). The GalF protein of *Escherichia coli* is not a UDP-glucose pyrophosphorylase but interacts with the GalU protein possibly to regulate cellular levels of UDP-glucose. *Mol Microbiol* **22**, 827–840.
- Mazur, A., Marczak, M., Krol, J. E. & Skorupska, A. (2005). Topological and transcriptional analysis of *pssL* gene product: a putative Wzx-like exopolysaccharide translocase in *Rhizobium leguminosarum* bv. *trifolii* TA1. *Arch Microbiol* **184**, 1–10.
- Miller, J. H. (1972). *Experiments in Molecular Genetics*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.
- Mori, M., Ohta, M., Agata, N., Kido, N., Arakawa, Y., Ito, H., Komatsu, T. & Kato, N. (1989). Identification of species and capsular types of *Klebsiella* clinical isolates, with special reference to *Klebsiella planticola*. *Microbiol Immunol* **33**, 887–895.
- Nakhmchik, A., Wilde, C. & Rowe-Magnus, D. A. (2007). Identification of a Wzy polymerase required for group IV capsular polysaccharide and lipopolysaccharide biosynthesis in *Vibrio vulnificus*. *Infect Immun* **75**, 5550–5558.
- Nassau, P. M., Martin, S. L., Brown, R. E., Weston, A., Monsey, D., McNeil, M. R. & Duncan, K. (1996). Galactofuranose biosynthesis in *Escherichia coli* K-12: identification and cloning of UDP-galactopyranose mutase. *J Bacteriol* **178**, 1047–1052.
- Nelson, K. & Selander, R. K. (1994). Intergeneric transfer and recombination of the 6-phosphogluconate dehydrogenase gene (*gnd*) in enteric bacteria. *Proc Natl Acad Sci U S A* **91**, 10227–10231.
- Niemann, H., Chakraborty, A. K., Friebolin, H. & Stirm, S. (1978). Primary structure of the *Escherichia coli* serotype K42 capsular polysaccharide and its serological identity with the *Klebsiella* K63 polysaccharide. *J Bacteriol* **133**, 390–391.
- Olson, S. A. (1994). MacVector: an integrated sequence analysis program for the Macintosh. *Methods Mol Biol* **25**, 195–201.
- Pan, Y. J., Fang, H. C., Yang, H. C., Lin, T. L., Hsieh, P. F., Tsai, F. C., Keynan, Y. & Wang, J. T. (2008). Capsular polysaccharide synthesis regions in *Klebsiella pneumoniae* serotype K57 and a new capsular serotype. *J Clin Microbiol* **46**, 2231–2240.
- Persson, M., Letts, J. A., Hosseini-Maaf, B., Borisova, S. N., Palcic, M. M., Evans, S. V. & Olsson, M. L. (2007). Structural effects of naturally occurring human blood group B galactosyltransferase mutations adjacent to the DXD motif. *J Biol Chem* **282**, 9564–9570.
- Podschun, R. & Ullmann, U. (1998). *Klebsiella* spp. as nosocomial pathogens: epidemiology, taxonomy, typing methods, and pathogenicity factors. *Clin Microbiol Rev* **11**, 589–603.
- Pruitt, K. D., Tatusova, T. & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61–D65.
- Rahn, A. & Whitfield, C. (2003). Transcriptional organization and regulation of the *Escherichia coli* K30 group 1 capsule biosynthesis (*cps*) gene cluster. *Mol Microbiol* **47**, 1045–1060.
- Rahn, A., Drummelsmith, J. & Whitfield, C. (1999). Conserved organization in the *cps* gene clusters for expression of *Escherichia coli* group 1 K antigens: relationship to the colanic acid biosynthesis locus and the *cps* genes from *Klebsiella pneumoniae*. *J Bacteriol* **181**, 2307–2313.
- Sanders, D. A., Staines, A. G., McMahon, S. A., McNeil, M. R., Whitfield, C. & Naismith, J. H. (2001). UDP-galactopyranose mutase has a novel structure and mechanism. *Nat Struct Biol* **8**, 858–863.
- Stenutz, R., Erbing, B., Widmalm, G., Jansson, P. E. & Nimmich, W. (1997). The structure of the capsular polysaccharide from *Klebsiella* type 52, using the computerised approach CASPER and NMR spectroscopy. *Carbohydr Res* **302**, 79–84.
- Stevenson, G., Andrianopoulos, K., Hobbs, M. & Reeves, P. R. (1996). Organization of the *Escherichia coli* K-12 gene cluster responsible for production of the extracellular polysaccharide colanic acid. *J Bacteriol* **178**, 4885–4893.
- Tonetti, M., Sturla, L., Bisso, A., Benatti, U. & De Flora, A. (1996). Synthesis of GDP-L-fucose by the human FX protein. *J Biol Chem* **271**, 27274–27279.
- Tusnády, G. E. & Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol* **283**, 489–506.
- Whitfield, C. (2006). Biosynthesis and assembly of capsular polysaccharides in *Escherichia coli*. *Annu Rev Biochem* **75**, 39–68.
- Whitfield, C. & Paiment, A. (2003). Biosynthesis and assembly of group 1 capsular polysaccharides in *Escherichia coli* and related extracellular polysaccharides in other bacteria. *Carbohydr Res* **338**, 2491–2502.
- Whitfield, C. & Roberts, I. S. (1999). Structure, assembly and regulation of expression of capsules in *Escherichia coli*. *Mol Microbiol* **31**, 1307–1319.
- Wiggins, C. A. & Munro, S. (1998). Activity of the yeast MNN1 α -1,3-mannosyltransferase requires a motif conserved in many other families of glycosyltransferases. *Proc Natl Acad Sci U S A* **95**, 7945–7950.
- Williams, A. H. & Raetz, C. R. (2007). Structural basis for the acyl chain selectivity and mechanism of UDP-N-acetylglucosamine acyltransferase. *Proc Natl Acad Sci U S A* **104**, 13543–13550.
- Wu, K. M., Li, L. H., Yan, J. J., Tsao, N., Liao, T. L., Tsai, H. C., Fung, C. P., Chen, H. J., Liu, Y. M. & other authors (2009). Genome sequencing and comparative analysis of *Klebsiella pneumoniae* NTUH-K2044, a strain causing liver abscess and meningitis. *J Bacteriol* **191**, 4492–4501.
- Yu, V. L., Hansen, D. S., Ko, W. C., Sagnimeni, A., Klugman, K. P., von Gottberg, A., Goossens, H., Wagener, M. M. & Benedi, V. J. (2007). Virulence characteristics of *Klebsiella* and clinical manifestations of *K. pneumoniae* bloodstream infections. *Emerg Infect Dis* **13**, 986–993.

Edited by: S. D. Bentley