Studies on the Serotypes and Genetic Structure of Capsular Polysaccharide Biosynthesis Loci of the Fish Pathogen *Streptococcus parauberis*

July, 2015

Graduate School of Fisheries and Environmental Sciences

Nagasaki University

Chuandeng Tu

Abstract

Streptococcus parauberis, a Gram-positive coccus, was reported as a fish pathogen for the first time in Spain in 1996. In Japan, *S. parauberis* has been recognized as a pathogen of Japanese flounder *Paralichthys olivaceus* since 2002. According to the results of slide agglutination test with rabbit antisera, the isolates from flounder have been divided into two serotypes, termed serotypes I and II. In streptococci, a variety of serotypes are generally due to differences of the capsular polysaccharide (CPS) antigens on their cell-surface, which are synthesized by the *cps* gene cluster on the chromosome. Since 2008, *S. parauberis* isolates that did not agglutinate with anti-serotype I or II sera were occasionally found, suspecting that there exist new serotypes in *S. parauberis*. Accurate identification of pathogens is important for the epidemiological study, pathogenic study, and therapy of the infection. This study was aimed to reexamine the serological relationship among the stocked *S. parauberis* strains including the non-agglutinating ones and also to investigate the correlation between the serotypes and structure of *cps* loci of *S. parauberis*.

In Chapter 1, the serotypes of stocked *S. parauberis* strains were reexamined using a newly prepared rabbit antiserum against a representative of non-agglutinating strains (NUF1071) as well as previously prepared anti-serotype I and II sera, and also pulsed-field gel electrophoresis (PFGE) was conducted to clarify the genetic relationship among the strains. An antiserum cross-absorption test and microtiter agglutination test revealed that the serotype I was divided into three subserotypes, designated Ia, Ib and Ic, of which the non-agglutinating strains belonged to the subserotype Ic. Of the 104 serotype I strains, 6, 91 and 7 strains belonged to subserotypes Ia, Ib and Ic, respectively. Formalin-killed cells (FKC) of subserotype Ia and Ic strains showed high agglutination titers against the anti-Ia (so far being used as an anti-serotype I serum) and Ic sera, respectively. Subserotype Ib strains agglutinated with the both sera. In PFGE analysis, the stocked 188 *S. parauberis* strains were

classified into three clusters; clusters I, II and III consisted of subserotypes Ib/Ic and Ia and serotype II, respectively. Non-typeable (NT) strains which agglutinated with both anti-serotype I and II sera were thought to be non-capsulated and belonged to cluster I or III. Differentiation of *S. parauberis* by serotyping and genotyping is useful for epidemiological studies.

In Chapter 2, using inverse-PCR technique, the nucleotide sequences of the *cps* loci of six representative strains were determined. The length of the *cps* loci of *S. parauberis* KRS02083 (Ia), NUF1003 (Ib), NUF1071 (Ic), NUF1032 (II), 2007-1 (NT, Cluster I) and NUF1095 (NT, Cluster III) was 18,910, 19,334, 19,334, 19,920, 19,334 and 19,920 bp, respectively. The genes involved in CPS biosynthesis comprised a gene cluster with cassette-like structure, and type-specific genes were flanked by conserved genes. The upper conserved region of the *cps* locus consisted of five regulatory genes and one processing gene, which were common to all the six strains. The type-specific region contained the genes that encoded an initial glycosyltransferase, polysaccharide polymerase (*wzy*), flippase (*wzx*), glycosyltransferases, acetyltransferases and aminotransferases and modifying enzymes. These structures suggest that the capsule of *S. parauberis* is synthesized by the Wzy-dependent pathway. Although the type-specific region was varied among KRS02083, NUF1003 and NUF1032, several common genes existed between KRS02083 and NUF1003. This would explain the cross reactivity in agglutination test between subserotypes Ia and Ib. The type-specific regions of NUF1071/2007-1 and NUF1095 had the same genetic structure as NUF1003 and NUF1032, respectively. There were only a few base substitutions found among them. Three of seven subserotype Ic strains harbored an IS (insertion sequence) element in the *cps* locus. This and a few base substitutions in the genes of *cps* locus would cause the change of amino acid and protein structure which resulted in the loss of enzyme activity and, hence, would cause the diversification of capsular polysaccharide structure or non-capsulation .

In Chapter 3, a multiplex PCR method for the detection of *S. parauberis* serotypes was developed based on the studies in the previous chapter. Serotype-specific primer sets were designed from the *wzy* gene sequences of *S. parauberis* with the expected product length of 213, 303 and 413 bp for subserotypes Ia and Ib/Ic and serotype II, respectively. The test results were consistent with those of agglutination test using antisera. Moreover, NT strains could be classified into serotype Ib/Ic or II, which was correlated with PFGE clusters. None of the other streptococcal species or the other pathogens of Japanese flounder showed positive reaction. Accordingly, the present multiplex PCR method was considered as an alternative to agglutination test for identification of *S. parauberis* and determination of the serotypes simultaneously.

In this study, reexamination of the serological relationship among the *S. parauberis* strains demonstrated that there were five serological phenotypes (subserotypes Ia, Ib and Ic, serotype II and nontypeable). The research on the genetic structure of *cps* loci brought about the foundational knowledge on the capsular polysaccharide biosynthesis pathway and the reasons for the presence of serological variation in *S. parauberis*, and, in addition, an accurate, rapid detection method for this pathogen was developed. The current study would help understand the pathogen and contribute to the future study for prevention and treatment of the disease.

Contents

Introduction

Streptococcal infection in aquaculture is considered to be a reemerging disease around the world and was initially reported in cultured rainbow trout *Oncorhynchus mykiss* in Japan (Romalde and Toranzo, 1999). There are at least six different species are known as significant pathogens of fish streptococcal infection including: *Lactococcus garvieae*, *Lactococcccus piscium*, *Streptococcus iniae*, *Streptococcus agalactiae*, *Streptococcus parauberis* and *Vagococcus salmoninarum*. Among these fish streptococci, *S. parauberis* was initially reported in Spain with physiological, biochemical and 16S rRNA gene sequence analysis of the isolates from turbot *Scophthalmus maximus* (Doménech *et al*., 1996). Since then, this pathogen caused high economic losses in cultured turbot in Spain and Japanese flounder *Paralichthys olivaceus* in Korea (Toranzo *et al*., 2005; Baeck *et al*., 2006). The diseased fish showed exophthalmia, hemorrhages of the eyes and jaw, and abdominal distension. Traditional treatment with frequently using antibiotics often resulted in occurrence of bacterial resistance and remnants of antibiotics, thus, vaccination strategies were increasingly becoming important.

In Japan, *S. parauberis* has become a major pathogen of Japanese flounder instead of *S. iniae* since 2002 (Kanai *et al*., 2009). According to the slide agglutination test with rabbit antisera, the isolates from flounder have been divided into two serotypes, termed serotypes I and II (Kanai *et al*., 2009). Reliable methods for strain differentiation are important for epidemiological studies and diagnoses of disease. Although serotyping continues to be a useful tool for epidemiologic surveillance, with the development of molecular techniques, the methods such as plasmid fingerprinting (Tenover, 1985), ribotyping (Stull *et al*.,1988), PCR-based methods (van Belkum, 1994) and pulsed-field gel electrophoresis (PFGE) can be used simultaneously for supplementing strain typing. Using several molecular methods for typing, Liebana *et al*. (2001) demonstrated that the isolates in the genus *Salmonella* from farm animals in the United Kingdom could be differentiated to five selected serovars by PFGE analysis. Fakhr *et al*. (2005) indicated that PFGE was superior to multilocus sequence typing in identifying strains of *S. enterica* serovar Typhimurium. PFGE exhibits exceptional discriminatory power and is acknowledged as „„gold standard"" for assessing isolate interrelationships. As a trend, traditional methods combined with molecular techniques for identification of strains are thought to be useful for epidemiological studies of fish pathogens.

In streptococci, capsular polysaccharides (CPS) are considered as virulent factors and components of vaccines. They are surface antigens most likely related to the serotypes. The diversity of polysaccharide structures due to variety of sugars and glycosidic linkages is reflected in the different serotypes among species. There reported at least ninety-one capsular serotypes in *S. pneumoniae* (Yu *et al.*, 2008) and thirty-three for *S. suis* (Liu *et al.*, 2013). Hence, the studies that can lead to understanding the construction of CPS are required. Generally, CPS biosynthesis is responsible by genes which are cluster in a single locus on the chromosome (Roberts, 1996) with a rare exception (Patrick *et al.*, 2010). The biosynthesis of CPS requires a complex pathway, where there are two major pathways termed Wzy-dependent and synthase-dependent pathways in Gram-positive bacteria. The genetic loci of Wzy-dependent pathway are more complex than those of synthase-dependent pathway (Yother, 2011). In Wzy-dependent pathway, the genes that encode the initial glycosyltransferase, Wzy polymerase and Wzx flippase are essential, and CPS biosynthesis may be performed as followings: an initial sugar is linked to a membrane-associated lipid carrier (undecaprenyl phosphate) by the initial glycosyltransferase to form the Und-P-P-Glc, to which glycosyltransferases link further monosaccharides sequentially to generate repeat units. Then, the repeat units are transported to the outer surface of the cytoplasmic membrane by Wzx flippase, and each repeat unit is polymerized to form the lipid-linked CPS by Wzy polymerase. Finally, mature CPS is translocated to the peptidoglycan by the membrane protein complex (Bentley *et al.*, 2006). In synthase-dependent pathway, polysaccharides usually consist of only one or two sugars, and the genes that encode the synthases and UDP-Glc dehydrogenases are essential for synthesis of these polymers. In all the serotypes of *S. pneumoniae*, only the serotype 3 and 37 are synthesized by the synthase-dependent pathway and the others are synthesized by the Wzy-dependent pathway (Yother, 2011).

In fish streptococci, it was demonstrated that *L. garvieae* exhibited two serogroups base on the presence (phenotype KG^{-}) or absence (KG^{+}) of a capsule (Yoshida *et al*., 1996). In *S. iniae*, two serotypes (I and II) with different capsule composition were clarified (Bachrach *et al*., 2001). Although Kanai *et al*., (2009) demonstrated that there exist two serotypes (I and II) in *S. parauberis*, upon the diagnosis, some *S. parauberis* isolates that did not agglutinate with anti-serotype I or II sera were occasionally found. These may give a serious suggestion that new serotypes exist in this pathogen. So the epidemiologic surveillance thinks to be desperate to predict the potential outbreak. On the other hand, in order to well explain the phenotypic diversity of serotypes in *S. parauberis*, the closely associated genetic study is also required. Studies on the genetic loci of CPS biosynthesis may lead to a better view to understand the mechanism of serotype diversity and contribute to development of a novel molecular serotyping method which needs less time and expense. Above all, in this study, serological and genetic typing (Chapter 1), characterization of *cps* gene clusters (Chapter 2) and a PCR molecular serotyping method (Chapter 3) of *S. parauberis* are described, respectively.

Chapter 1

Serotyping and genotyping of *S. parauberis* strains isolated from Japanese flounder

1.1 Introduction

Previous study demonstrated that there were two serotypes in *S. parauberis* (Kanai *et al*., 2009). However, some *S. parauberis* isolates which did not agglutinate with anti-serotype I or II sera were found occasionally from 2008, suggesting new serotypes may exist in *S. parauberis*. In this chapter, serological relationship among the non-agglutinating and agglutinating strains was investigated using rabbit antisera raised against the non-agglutinating strain as well as serotype I and II strains. Pulsed-field gel electrophoresis (PFGE) was also conducted to clarify the genetic relationship among the strains.

1.2 Materials and Methods

Bacteria

A total of 188 strains of *S. parauberis* isolated from diseased Japanese flounder between 2002 and 2012 were used (Table 1.1). They were stocked at -80℃ in Todd Hewitt broth (TH; Difco) supplemented with 10% glycerol. For preparation of formalin-killed cells (FKC), bacterial cells of each strain grown in TH broth or agar at 27℃ for 24 h were inactivated with 0.5% formalin for 2 days, washed with sterilized phosphate-buffered saline, pH 7.2, (PBS) and resuspended in PBS at 100 mg/mL. FKC suspension was added with 0.1% formalin as preservative and stored at 4℃ .

Identification of S. parauberis *by PCR*

The strains were identified as *S. parauberis* by PCR targeted the partial 23S rRNA gene with the primer pair: S.para-F (5'TTTCGTCTGAGGCAATGTTG3') and S.para-R (5'GCTTCATATATCGCTATACT3') (Mata *et al*., 2004). PCR reaction mixture contained colonies of *S. parauberis*, 0.25 µM of each primer, 2.5 µL of 10 × Ex Taq buffer (Mg^{2+} plus), dNTP mixture (0.2 mM each) and 1.25 U of Ex Taq DNA polymerase (Hot Start Version, Takara) to a total volume of 25 µL by adding with deionized distilled water. PCR amplification was conducted in a thermal cycler and started by denaturation at 95℃ for 3 min. Then, the PCR mixture was subjected to 25 cycles of denaturation at 98℃ for 10 s, annealing at 55℃ for 30 s and extension at 72℃ for 45 s. A final extension was performed at 72℃ for 7 min. The PCR products were analyzed by electrophoresis in 1.2% agarose gel and stained with ethidium bromide.

Antisera

Rabbit antisera were raised against FKCs of *S. parauberis* NUF1003 (serotype I) and NUF1071 (a non-agglutinating strain) according to the previous report (Kanai *et al*., 2009). The antisera against the strains KRS02083 (serotype I) and KRS02109 (serotype II) prepared in the previous study (Kanai *et al.*, 2009) were also used. The strains KRS02083 and KRS02109 were isolated in Kagawa Prefecture in 2002, NUF1003 was isolated in Shimane Prefecture in 2004, and NUF1071 was isolated in Nagasaki Prefecture in 2008.

Microtiter agglutination test

A two-fold serial dilution of rabbit antisera in PBS was carried out in a 96-well microtiter plate. To 25 μL of serially diluted antisera, the same volume of FKC suspension (approximately 2 mg/mL in PBS) of each strain was added, and the mixture was agitated thoroughly and incubated at 4℃ overnight. Agglutination titer was expressed by the reciprocal of the highest dilution of antiserum at which more than 90% of FKC agglutinated.

Preparation of absorbed antisera

For antiserum cross-absorption test, absorbed antisera were prepared by mixing each antiserum with the FKC (200 mg/mL of antiserum) of an absorbing strain and incubating at 4℃ overnight. After centrifugation (13,000 ×g, 2 min), the supernatant was designated the absorbed antiserum. The absorption procedure was repeated until the agglutination titer against the absorbing FKC became <4.

PFGE analysis

Bacterial cells grown on TH agar at 27℃ overnight were washed twice and suspended at 4.5×10^9 CFU/mL in TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0). Gel plugs were made by mixing 110 μL of the bacterial suspension with the same volume of 1.6% low melting-point agarose (Bio-Rad) in TE buffer using PFGE molds (Bio-Rad). Each solidified plug was placed into 380 μL of lysozyme solution (10 mM Tris-HCl, 50 mM NaCl, 5 mg/mL lysozyme, pH 7.2) and incubated at 37℃ for 3 h with gentle shaking. After the lysozyme solution was discarded, 780 μL of proteinase K solution (0.1 M EDTA, pH 8.0, 1% sodium N-lauroylsarcosinate, 0.5% SDS, 0.25 mg/mL proteinase K) was added and incubated at 50℃ for 20 h. Plugs were washed 3 times in TE buffer supplemented with 1 mM phenylmethylsulfonyl fluoride for 1 h at room temperature and twice in 10 mM Tris-HCl (pH 8.0) at 4℃ for 20 min with gentle shaking. *Sma*I (Takara), at a final concentration of 1 U/mL, was used for restriction endonuclease digestion at 30℃ for 40 h according to the manufacturer‟s instructions. PFGE was performed in 1% agarose gel (Pulsed Field Certified Agarose, Bio-Rad) using a CHEF-DR III system (Bio-Rad) at 6 V/cm for 18 h at 14℃ with pulse time switched from 0.5 s to 25 s under included angle 120°. The gel was stained with ethidium bromide for 30 min, destained in distilled water and photographed under UV light. The banding patterns were compared using ATTO Lane & Spot Analyzer ver 6.0 software. Dendrogram was constructed based on the unweighted pair group method with arithmetic averages (UPGMA) with MEGA ver 6.0 software.

1.3 Results

Identification of S. parauberis *strains*

All of the strains used were identified as *S. parauberis,* giving expected 718-bp PCR amplification product of 23 S rRNA gene specific for *S. parauberis*.

Serotyping of S. parauberis *strains*

From the agglutination titers of FKCs against four kinds of antiserum, 188 *S. parauberis* strains were divided into five groups; three subtypes of serotype I (Ia, Ib and Ic), serotype II and non-typeable (Table 1.2). Strains of subserotypes Ia and Ic were distinguished each other using anti-subserotype Ia (KRS02083) and Ic (NUF1071) sera, that is, subserotypes Ia and Ic showed high titers against anti-Ia and Ic sera, respectively. Subserotype Ib strains agglutinated with these antisera showing titers at an equal level. Although the agglutination titers of anti-Ib (NUF1003) serum were low in comparison with the other antisera, serotype II strains reacted weakly with it. Out of 104 serotype I strains, 87.5% were classified into subserotype Ib. Non-typeable strains, which showed relatively high titers against anti-subserotype Ia (KRS02083) and serotype II (KRS02109) sera, accounted for 11.7% of the total strains (Table 1.2).

Antiserum cross-absorption test

Antiserum cross-absorption test demonstrated the relationship among subserotypes Ia, Ib and Ic more clearly (Table 1.3). In this test it was shown that subserotypes Ia and Ic were completely separated, since cross-absorption between them did not lower the agglutination titers of anti-Ia and Ic sera. On the other hand, subserotype Ib could possess surface-exposed antigens common to Ia and Ic, since the agglutination titers of these antisera were reduced after absorbed with Ib FKC.

PFGE analysis

PFGE analysis of *sma*I-digested genomic DNAs revealed that *S. parauberis* strains were divided into 40 pulsotypes and classified into three clusters according to the distance value (>0.35) (Fig. 1.1). By clustering the pulsotypes, subserotypes Ib and Ic belonged to the same group (cluster I) and were not separated each other. Subserotype Ia and serotype II constructed respective clusters (clusters II and III). Pulsotypes S3 and S13, and S34, S36, S38 and S39 were the main types for subserotype Ib and serotype II, respectively. Non-typeable strains belonged to cluster I or III.

1.4 Discussion

In this chapter the existence of three subtypes of serotype I (Ia, Ib and Ic) in *S. parauberis* derived from Japanese flounder was described. Until now, serotype I has been thought to be a single entity, and anti-serotype I (anti-KRS02083) (designated Ia in this study) and II (anti-KRS02109) sera have been used for serodiagnosis of *S. parauberis* in Japan (Kanai *et al*., 2009). Now, using anti-subserotype Ic serum as well as anti-subserotype Ia serum, strains that had been classified as serotype I could be distinguished to subserotypes Ia and Ib. Because serotype Ic has seldom been isolated, unidentifiable cases of *S. parauberis* would scarcely be encountered even if one uses only anti-serotype I (Ia) serum. However, upon serodiagnosis one should be aware of the existence of the subserotypes.

Although there are few investigations on the antigens that participate in the agglutination reaction of *S. parauberis* with antisera, Kanai *et al*. (2009) suggested that the CPS were involved in the serological variation of *S. parauberis*. Therefore, differences in agglutination titers among the three subtypes of serotype I may come from the structural variation in CPS. Kanai *et al*. (2009) also reported the serotype-specific precipitating reactions of autoclave-extracted cellular antigens of serotype I and II strains in an agar double-diffusion test, but the test could not distinguish between subserotypes Ia and Ib. Thus subserotypes Ia and Ib may have a common CPS structure in part. Similarly, the same thing would be said between subserotypes Ib and Ic, because subserotype Ib strains agglutinate with anti-serotype Ic serum. In other streptococci, CPS synthetic gene clusters were varied according to serotypes (Shibata *et al*., 2003; Mavroidi *et al*., 2007; Thurlow *et al*., 2009; Okura *et al*., 2013), indicating the relationship between CPS structure and serotypes. However, although CPS is most likely involved in serotype specificity of *S. parauberis*, antigenic cell-surface proteins or other polysaccharides such as teichoic acids and rhamno polysaccharides (Hancock and Gilmore, 2002) may also be involved.

PFGE analysis of the *S. parauberis* strains revealed the relationship between pulsotypes and serotypes. This suggests that the serotypes are genetically separated. However, subserotypes Ib and Ic were not separated in PFGE analysis. As the reason for this, minor differences in genetic structure or expression of the CPS genes might exist between the two subserotypes. Non-typeable strains were distributed to clusters I (subserotypes Ib and Ic) and III (serotype II). Kanai *et al*. (2009) found one non-typeable strain of *S. parauberis*, which agglutinated with both anti-serotype I and II sera but did not show precipitation reaction between its autoclave-extracts and the antisera, suggesting that the non-typeable strain was non-capsulated. The similar phenomena were observed in the KG$^+$ phenotype of *Lactococcus garvieae* (Yoshida *et al*., 1997) and K$^-$ phenotype of *S. iniae* (Kanai *et al*., 2006). If the non-typeable strains were non-capsulated, their CPS synthetic gene(s) may be lost or mutated (Shutou *et al*., 2007; Morita *et al*., 2011).

Meng *et al*. (2009) reported five serotype I strains that harbored an *ermB* gene (erythromycin resistance) on their chromosome and a *tet*(S) gene (tetracycline resistance) on an 11-kbp conjugative plasmid. These strains isolated in 2002 were originated from one prefecture. In the present study, these strains and two additional strains isolated in 2007 and 2010 at another prefecture were revealed to be classified to subserotype Ia. The latter two strains were shown to possess the 11-kbp plasmid, but did not harbor the *ermB* gene (data not shown). Therefore, there is a risk of dissemination of subserotype Ia strains possessing the R-plasmid in the western part of Japan.

Serotypes of pathogens are a matter of concern not only for serodiagnosis but also for vaccine development. At present, a commercially available combined vaccine for *S. parauberis* I/II and *S. iniae* infections in Japanese flounder is used at aquaculture farms with high efficacy. Our preliminary experiment showed a low protective effect in fish vaccinated with FKC of subserotype Ic followed by challenged with a virulent subserotype Ia strain, but a high protective effect was achieved in these fish challenged with serotype Ib or in fish vaccinated with FKC of subserotype Ib followed by challenged with the subserotype Ia strain. These results suggest that a vaccine made from subserotype Ib cells are effective against the infections caused by the three subtypes of serotype I, because subserotype Ib cells possess common cell-surface antigens among them.

1.5 Conclusion

1. *S. parauberis* isolates from Japanese flounder were divided into five serological phenotypes; subserotypes Ia, Ib and Ic, serotype II and non-typeable.
2. Subserotype Ib strains possess common cell-surface antigens among serotype I.
3. By PFGE analysis of *Sma*I-digested genomic DNA, *S. parauberis* strains exhibited 40 pulsotypes and generated three clusters.
4. Pulsotypes S3, S13, S34, S36, S38 and S39 were the main genetypes.
5. Serotypes Ib/Ic, Ia and II corresponded to cluster I, II and III, respectively, and non-typeable strains belonged to cluster I or III.
6. PFGE can be used in classification of *S. parauberis* and it is a useful supplementary method for traditional serotyping of *S. parauberis*.

Table 1.1. Source of *S. parauberis* strains used in this study

Year	No. of strains from							Total
	Kagawa	Shimane	Ehime	Oita	Kumamoto	Kagoshima	Nagasaki	
2002	11							11
2003	1			1				2
2004	3	1	2			3		9
2005	4		14	7			1	26
2006			1	6		2	2	11
2007	3		11	13				27
2008	2		12	10	1		9	34
2009			16	11			9	36
2010	2		7	8			10	27
2011							2	2
2012							3	3
Total	26	1	63	56	1	5	36	188

Table 1.2. Serotyping of *S. parauberis* based on agglutination titers of FKCs against rabbit antisera

Antiserum	Agglutination titer of strains belong to				
	Subserotype Ia (n=6)	Subserotype Ib (n=91)	Subserotype Ic (n=7)	Serotype II (n=62)	Non-typeable (n=22)
Anti-KRS02083 (Ia)	128～256	32～128	8～16	<4	64～>65,536
Anti-NUF1003 (Ib)	<4～32	8～64	4～16	<4～32	32～8,192
Anti-NUF1071 (Ic)	<4	16～128	64～256	<4	64～4,096
Anti-KRS02109 (II)	<4	<4	<4	256～2,048	16～32,768

Table 1.3. Agglutination titers of representative serotype I strains belonged to three subserotypes against unabsorbed and absorbed antisera

Antiserum	Absorbed with FKC of	Agglutination titer of FKC of		
		KRS02083 (Ia)	NUF1003 (Ib)	NUF1071 (Ic)
Anti-KRS02083	Unabsorbed	128	32	4
(Type Ia)	KRS02083 (3)*	<4	<4	<4
	NUF1003 (4)	16	<4	<4
	NUF1071 (2)	128	32	<4
Anti-NUF1003	Unabsorbed	16	16	4
(Type Ib)	KRS02083 (3)	<4	<4	<4
	NUF1003 (2)	<4	<4	<4
	NUF1071 (3)	<4	4	<4
Anti-NUF1071	Unabsorbed	<4	64	128
(Type Ic)	KRS02083 (2)	<4	64	128
	NUF1003 (4)	<4	<4	4
	NUF1071 (3)	<4	<4	<4

*No. in the parentheses indicates no. of repetition of absorption procedure.

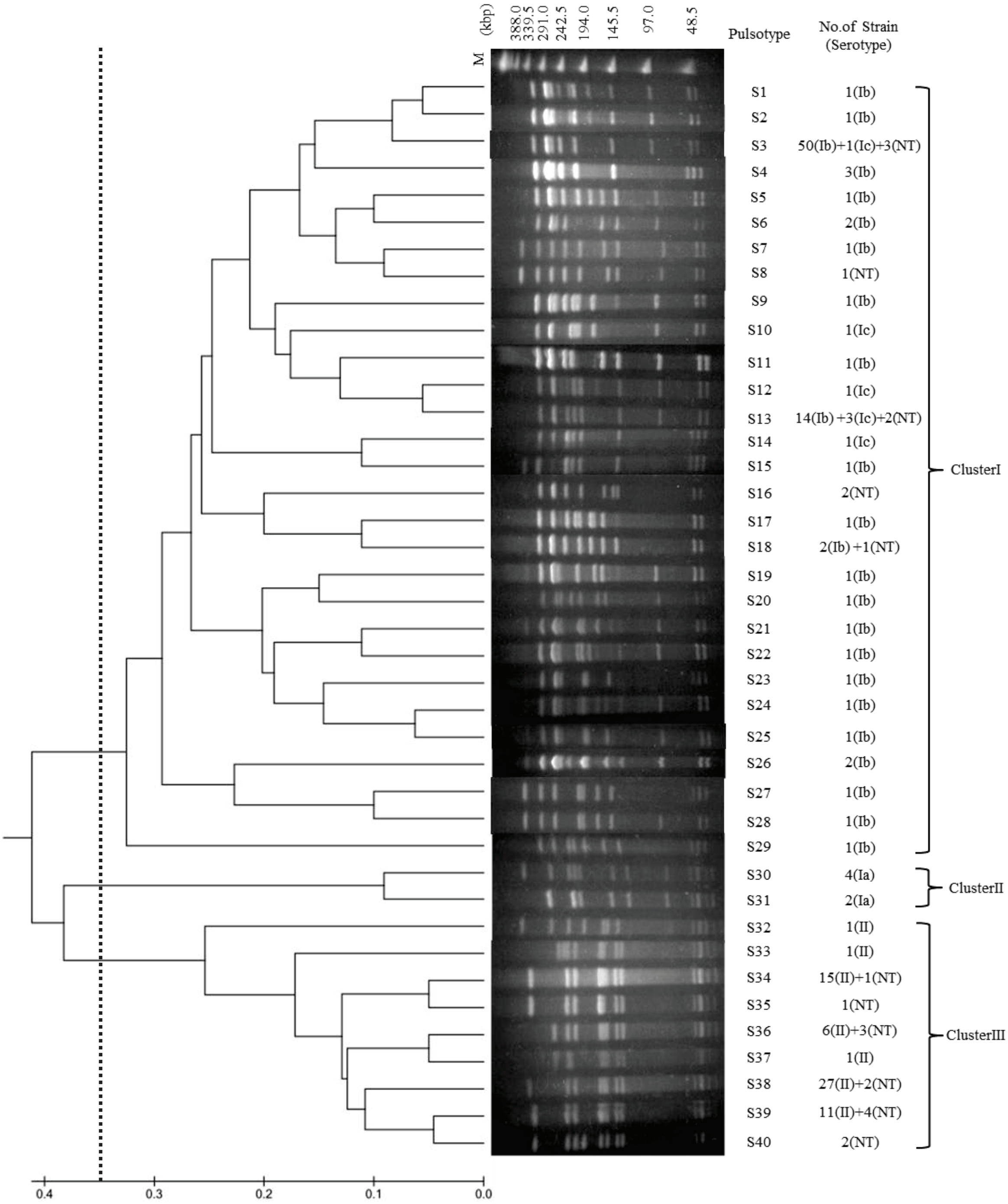


Fig. 1.1. Dendrogram of *S. parauberis* strains constructed by UPGMA cluster analysis of the PFGE patterns of *Sma*I-digested genomic DNA. The number of serotyped strains in each pulsotype was showed on the right. All the strains were classified into three clusters according to the distance value (>0.35) indicated by dotted line across the scale bar. M, Lambda Ladder PFG Marker (New England BioLabs); NT, non-typeable.

Chapter 2

Characterization of genetic loci for capsular polysaccharide biosynthesis of *S. parauberis*

2.1 Introduction

Polysaccharide capsules are particular structures found on the cell surface of many bacterial species. They are usually regarded as crucial virulent factors and have antigenicity, which can be used for serotyping of bacteria. The diversity of capsular polysaccharide (CPS) structures due to variety of sugars and glycosidic linkages is reflected in the differences of serotypes (Okura *et al*., 2013). Generally, the CPS biosynthetic genes are clustered at a single locus on the chromosome (Roberts, 1996) with a rare exception (Patrick *et al*., 2010). The biosynthesis of CPS requires a complex pathway, where there are two major pathways termed synthase-dependent and Wzy-dependent pathways, which are responsible for polymerization of individual sugars in a processive reaction and discrete repeat units in a nonprocessive reaction, respectively, in Gram-positive bacteria (Yother, 2011).

In Chapter 1, it was demonstrated that the serotype I was subdivided into three subserotypes, designated Ia, Ib and Ic, while non-typeable strains which agglutinated with both serotypes I and II antisera were also found. The existence of serological diversity suggests the presence of distinct features of *cps* locus according to each serotype of this pathogen. In this chapter, to investigate the correlation between the serotypes and structure of *cps* loci of *S. parauberis*, the *cps* loci of the representative strain of each serotype and two non-typeable strains were sequenced and analyzed. The possible biosynthetic pathway and mechanisms by which the diversity arose in *S. parauberis* were also discussed.

2.2 Materials and Methods

Bacterial strains and genomic DNA isolation

For DNA sequencing, four *S. parauberis* strains, KRS02083, NUF1003, NUF1071 and NUF1032, were selected as the representatives of subserotypes Ia, Ib and Ic and serotype II, respectively. Two non-typeable (NT) strains, 2007-1 and NUF1095, which belonged to the clusters I and III in *Sma*I-PFGE analysis, respectively, were also included (Table 2.1). For PCR scanning (verification of the structure of the *cps* locus by amplifying the parts of the entire *cps* locus), 182 remaining *S. parauberis* strains were used (Table 1.1). All the strains were cultured on Todd Hewitt (Difco) agar at 27℃ for 24 h, and genomic DNA was prepared using Wizard Genomic DNA Purification kit (Promega) according to the manufacturer‟s instructions.

Sequencing of the cps *loci of* S. parauberis *NUF1003 and NUF1032*

Initial PCR amplification and sequencing: The primer pair SP1F (5‟ TATYAGYGGTATTGATAC) and SP2R (5‟ GGTGGRGTATCRATGAT) was designed from the conserved region of *cps* loci of *S. agalactiae* serotype Ia (GenBank accession no., AB028896) and III (AF163833), *S. thermophilus* (DQ393658) and *S. iniae* (AY904444), and the initial PCR was conducted to amplify a partial sequence of the *cps* loci of *S. parauberis* NUF1003 and NUF1032. The PCR was performed on C1000 Thermal Cycler (Bio-Rad) using *TaKaRa Ex Taq* Hot Start Version (Takara). The PCR reaction mixture (a total volume of 50 μL) contained 5 μL of ×10 Ex Taq buffer (Mg^{2+} plus), 4 μL of dNTP mixture (0.2 mM each), SP1F and SP2R primers (0.5 μM each), 1 μL of template DNA and 1.25 U of Ex Taq DNA polymerase. The condition of PCR amplification was denaturation at 95°C for 5 min, 30 cycles of denaturation at 95°C for 30 s, annealing at 55°C for 30 s and extension at 72°C for 3 min. The PCR products were cloned into pGEM-T Easy vector using pGEM-T Easy Vector System (Promega) with *Escherichia coli* JM109 competent cells (Promega), and three clones per strain were sequenced using BigDye Terminator v3.1 Cycle Sequencing kit and ABI PRISM 3130xl Genetic Analyzer (Applied Biosystems).

Inverse PCR and sequencing: The adjacent upstream and downstream regions of the sequenced site of the *cps* locus were amplified and sequenced successively using inverse PCR technique (Ochman *et al*., 1988). Briefly, genomic DNA was digested with each of restriction enzymes (*Bam*HI, *Bgl*II, *Eco*RI, *Eco*RV, *Hin*dIII, *Kpn*I, *Pst*I, *Pvu*II, *Sph*I) and self-ligated using DNA Ligation kit ver. 2.1 (Takara). Inverse PCR was performed using *TaKaRa Ex Taq* Hot Start Version with the self-ligated (looped)

DNA and oppositely directed primer pair designed from the sequenced region of the *cps* locus by the PCR condition mentioned above, except that the extension time was changed to 5 min. The PCR products obtained were cloned and sequenced as above. Inverse PCR was repeated using primers newly designated from the sequenced region until the entire sequence data of *cps* locus was obtained. DNASIS program (Hitachi Software Engineering) was used for the DNA sequence assembling.

Sequencing of the cps *loci of* S. parauberis *KRS02083, NUF1071, 2007-1 and NUF1095*

Primer pairs were designed from the sequence data of NUF1003 and NUF1032, and PCRs were performed to amplify the parts of entire *cps* loci of KRS02083, NUF1071, 2007-1 and NUF1095 using PrimeSTAR HS DNA Polymerase (Takara). The PCR products were sequenced directly using BigDye Terminator v3.1 Cycle Sequencing kit and ABI PRISM 3130xl Genetic Analyzer (Applied Biosystems). In the case of KRS02083, inverse PCR technique was also applied, since several parts could not be amplified.

Analysis of sequence data

Open reading frames (ORFs) were predicted using NCBI ORF finder (http://www.ncbi.nlm.nih.gov/gorf/orfig.html) which could link to the BLAST network server (http://blast.ncbi.nlm.nih.gov/Blast.cgi). All predicted genes were named in alphabetical order in each serotype (e. g. *cpsA-cpsR*). ClustalW2.1 network version on DDBJ (http://clustalw.ddbj.nig.ac.jp/index.php?lang=en) was used for DNA multiple sequence alignment.

PCR scanning

To verify whether the structure of *cps* loci of the remaining 182 *S. parauberis* strains was the same as the representative strains of respective serotype, PCR that spanned the entire region of *cps* locus was conducted with primer pairs listed in Table

2.2. The primers were designed so that the neighbored PCR products overlapped each other. The PCR was performed on C1000 Thermal Cycler (Bio-Rad) using *TaKaRa Ex Taq* Hot Start Version (Takara) and the condition mentioned above, but the annealing temperature listed in Table 2.2 and extension time of 1 min/kb were employed.

Nucleotide sequence accession numbers

The nucleotide sequences of six *S. parauberis cps* loci obtained in this study were deposited in the DDBJ/EMBL/GenBank databases under accession numbers from LC060252 to LC060257 for KRS02083, NUF1003, NUF1071, NUF1032, 2007-1 and NUF1095, respectively.

2.3 Results

General features of the cps *loci of the representative strains*

By PCR amplification, cloning and sequencing, the constitution of *cps* loci of six representative strains became apparent (Fig. 2.1). The length, number of ORF and G+C content of the loci are shown in Table 2.1. The function of each gene was predicted based on the homology of its deduced amino acid sequence with those on the database (Tables 2.3 and 2.4). In all the six strains, predicted genes were orientated in the same direction except the first gene, and cassette-like structure was observed, i.e., conserved regions flanked serotype-specific genes. At the upper conserved region of the *cps* locus, there were five regulatory genes, *lysR* and *cpsABCD* (also known as *wzg*, *wzh*, *wzd* and *wze*), and one processing (epimerase/short-chain dehydrogenase) gene (*cpsE*) sequentially, which were conserved with high sequence identity among all the six strains (>99%) (S1) (S= supplemental data listed in this thesis). Each of the two genes, *cpsQ* and *cpsR*, at the lower conserved region also showed high identity among the strains (>96%) (S2). In the serotype-specific region, genes that encode the initial glycosyltransferase, polysaccharide polymerase (*wzy*) and flippase (*wzx*) were present at different locations along with genes encoding enzymes such as glycosyltransferase, acetyltransferase, aminotransferase and modifying enzymes (Fig. 2.1). As shown in Tables 2.3 and 2.4, most of the genes in the conserved region showed the highest identity with those of *Streptococcus uberis*, while the genes in the serotype-specific region did not necessarily show the high identity with them.

Comparison of the six cps *loci*

The nucleotide sequences of serotype-specific regions of NUF1003 (subserotype Ib), NUF1071 (subserotype Ic) and 2007-1 (NT type, *Sma*I-PFGE Cluster I) were almost identical and those of NUF1032 (serotype II) and NUF1095 (NT type, *Sma*I-PFGE Cluster III) was also almost identical. The difference between NUF1003 and NUF1071 was only two single-base substitution found in *cps1bM* and *cps1bN*. Similarly, the difference between NUF1003 and 2007-1 and between NUF1032 and NUF1095 were three and two single-base substitution in *cpsC*, *cps1bG* and *cps1bM* and in *cpsC* and *cps2G*, respectively (S11-S15). From these results it is suggested that there are three kinds of *cps* locus among serotypes and subserotypes in *S. parauberis*, namely the loci for subserotype Ia, subserotypes Ib/Ic and serotype II (Fig. 2.1).

Sequence homology of the deduced amino acid sequences of the genes in the serotype-specific regions is shown in Table 2.5 (S3-S5, S9, S10, S16-S18). In many genes the homology values between subserotype Ia and Ib/Ic were higher than those between subserotype Ia and serotype II or between subserotypes Ib/Ic and serotype II (S6-S10, S16-S18).

PCR scanning of cps *loci*

Expected sizes of amplification products listed in Table 2.2 were obtained in 182 *S. parauberis* strains except three strains belonged to subserotype Ic, suggesting all the strains possess the same genetic structure of *cps* loci as representative strains of respective serological phenotypes. In the three subserotype Ic strains one of PCR products was longer than that obtained in NUF1071 (S20). Sequencing of the PCR product revealed that an *ISSdy1*-like element (GenBank accession no., AJ250837) was inserted in *cps1bN* (S19).

2.4 Discussion

Many bacterial pathogens have a number of antigenic variants due to differences in the polysaccharides present at their cell surface. In streptococci, the cell-surface polysaccharides are most likely known as CPS. The biosynthesis of CPS always needs some essential enzymes with a complex pathway. The CPS of *Streptococcus pneumoniae* has been well studied and are known to be generally synthesized by the Wzy-dependent pathway (Bentley *et al*., 2006; Yother, 2011). In this study, the genes in the *cps* loci were deduced to encode the initial glycosyltransferase (Undecaprenyl-phosphate galactose phosphotransferase), Wzy polymerase and Wzx flippase, which were needed for the Wzy-dependent pathway. Capsule biosynthesis may be performed as followings: an initial galactose is linked as a galactose phosphate to a membrane-associated lipid carrier (undecaprenyl phosphate) by the initial glycosyltransferase to form the Und-P-P-Gal, to which glycosyltransferases link further monosaccharides sequentially to generate repeat units. Then, Wzx flippase transports the repeat units to the outer surface of the cytoplasmic membrane, and each repeat unit is polymerized to form the lipid-linked CPS by Wzy polymerase. Finally, mature CPS is translocated to the peptidoglycan by the membrane protein complex (Bentley *et al*., 2006).

The genes involved in CPS biosynthesis of *S. parauberis* comprised a gene cluster of cassette-like structure, i.e., conserved genes flanked serotype-specific genes, which is also found in other streptococcal species, such as *S. pneumoniae*, *S. agalactiae* and *S. thermophilus* (Wessels, 1997). *lysR* encoding a transcriptional regulator that belongs to the type 2 periplasmic binding fold protein superfamily can control the synthesis of virulence factors and may play the regulatory role in the CPS biosynthetic operon. CpsA is similar to LytR of *Bacillus subtilis* and could be a transcriptional regulator. *cpsA*-deleted mutants of *S. pneumoniae* and *S. agalactiae* have been proved to produce less capsule (Cieslewicz *et al.*, 2001). CpsBCD are considered to be a tyrosine kinase phosphoregulatory system which can affect the polymer length. CpsD with the activity of initial phosphorylation requires intracellular interaction with CpsC, while CpsB can block the initial phosphorylation of CpsD, as a kinase inhibitor (Bender and Yother, 2001). The CpsBCD system can be affected by oxygen levels, and reduced aeration conditions were found to correlate with increases in both CpsD phosphorylation levels and capsule production (Weiser *et al.* 2001). *cpsE* identified in *Campylobacter jejuni* encodes sugar-nucleotide-modifying enzyme involved in the biosynthesis of surface carbohydrates (Creuzenet, 2004). It can affect the synthesis of branched chain. In the serotype-specific region, the genes encode the enzymes to construct the repeat units, including an initial glycosyltransferase, additional glycosyltransferases, Wzx flippase, Wzy polymerase and enzymes to modify the repeat units (Yother, 2004). At the lower conserved region, there are two genes, one of which (*cpsQ*) encodes a protein similar to WbpA, an enzyme involved in the biosynthesis of unusual di-N-acetyl-d-mannosaminuronic acid-derived sugar nucleotides found in the O antigen of *Pseudomonas aeruginosa* PAO1 (Miller *et al.*, 2004), while the other one encodes hypothetical protein.

The result of PCR scanning suggests that the *cps* locus of *S. parauberis* is generally conserved among different strains of the respective serotypes. The G+C contents of *cps* loci (31.07%−31.51%) were lower than those of whole-genome (35.48%−35.60%) (Nho *et al.*, 2013) indicated that the genes of *cps* loci may be imported from other sources. With comparing the structure of the *cps* locus, the distinct features of the *cps* locus were observed in serologically different phenotypes except subserotypes Ib and Ic. Between subserotypes Ib and Ic and between nontypeable strains and serotypes Ib/II, there are several single-base substitution, which may lead to the change of amino acid or the appearance of earlier stop codon (S1, S11-15), and insertion events in the case of subserotype Ic strains (S19). They could affect the biosynthesis and structure of CPS and lead to reflect in the generation of nontypeable phenotype and subserotypes (Wang *et al.*, 2011a; Lakkitjaroen *et al.*, 2014). High similarities of genes in the serotype-specific region of subserotypes Ia and Ib/Ic could participate in construction of common CPS structure, which resulted in antigenic cross reactivity between them (Kanai *et. al*., 2015).

In this chapter it was demonstrated that the three kinds of serotype-associated *cps* locus existed in *S. parauberis*. In order to better understand the mechanism of diversity arisen among the serological phenotypes of this pathogen, studies on the CPS structure and activity of each gene product will be necessary by biochemical and genetic research such as mutant construction. In this regard, the present findings will be an important base for further studies. On the other hand, current works may also contribute to the development of a novel molecular method which can diagnose serotypes rapidly and less expensively.

2.5 Conclusion

1. The length of *cps* loci for NUF934, NUF1003, NUF1071, NUF1032, 2007-1 and NUF1095 was 18,910, 19,334, 19,334, 19,920, 19,334 and 19,920 bp, respectively.
2. The genes involved in CPS biosynthesis of *S. parauberis* comprised a gene cluster of cassette-like structure and the biosynthesis of CPS in *S. parauberis* may be operated by the Wzy-dependent pathway.
3. In *S. parauberis*, there were three distinctly different *cps* loci corresponded to five serological phenotypes. The loci of subserotypes Ib and Ic were almost identical. The loci of non-typeable strains which belonged to cluster I and III of *Sma*I-PFGE analysis were also almost identical to the subserotype Ib/Ic and II, respectively.
4. Single-base substitution or insertion event may affect the biosynthesis and structure of CPS.
5. The *cps* locus is generally conserved among different strains of the respective serotypes.

Table 2.1. Representative *Streptococcus parauberis* strains used for *cps* locus analyses

Strain	Origin		Serotype*	Affiliation in *Sma*I-PFGE analysis*	*cps* locus			
	Prefecture	Year			Length (bp)	No. of ORF	G+C content (%)	Accession number
KRS02083	Kagawa	2002	Ia	Cluster II	18,910	19	31.14	LC060252
NUF1003	Shimane	2004	Ib	Cluster I	19,334	18	31.07	LC060253
NUF1071	Nagasaki	2008	Ic	Cluster I	19,334	18	31.07	LC060254
NUF1032	Nagasaki	2006	II	Cluster III	19,920	19	31.51	LC060255
2007-1	Oita	2007	NT	Cluster I	19,334	18	31.07	LC060256
NUF1095	Nagasaki	2009	NT	Cluster III	19,920	19	31.51	LC060257

*: Kanai *et al*., 2015; NT, non-typeable; All the strains were isolated from diseased Japanese flounder.

Table 2.2. Primers used for PCR scanning of *cps* locus

Primer	Sequence (5'–3')	Region	Approximate size of amplicon (bp)	Annealing temperature (℃)
Sp-cps21	GACATCACGGTTATAGTC	*LysR-cpsB*	2,600	55
Sp-cpsR	GTTCGATATGCGCGAC			
5sp-IP6	CTAGATGATCAGAGTGTC	*cpsA-cpsD*	1,600	51
3sp-IP1	GATGTTAGTCCGAATGGAG			
I-1-3sp2	GACCGCGTGAAACGCCCAG	*cpsC-cpsE*	2,300	61
Sp-cps26	TGAATGACAAGACGACTAGC			
3sp-IP5	GTACTCGATTAAGCGAC	*cpsE-cps1aG*	3,000	49
Sp-cps47	CTGATAATTCCCGATCAATG			
Sp-cps32	ACTAGCAGCTGAGAATGG	*cps1aG-cps1aI*	1,900	55
Sp-cps48	GCCATCAGCAATAGAAATAC			
Sp-cps38	ATCCTTCTGTAGTATTAGGC	*cps1aH-cps1aK*	3,300	51
Sp-cps49	GAACCTTGTATAACGTCAAC			
Sp-cps41	CTAACGAAATCGGAGCTG	*cps1aK-cps1aM*	2,200	55
Sp-cps50	TGAGACTACTGAACCTGCAG			
Sp-cps45	TACGTGGGATTCTAATC	*cps1aL-cps1aN*	2.900	51
Sp-cps42	GATTTTACGAATTCTTGTTC			
Sp-cps51	ATTGCAACAGCAGCTATTTC	*cps1aN-cps1aP*	2,300	55
Sp-cps52	ACACCAGCAGAAATTAGTTC			
Sp-cps36	TAGCAGCACAAGCAGGAG	*cps1aP-cpsR*	2,600	55
3sp-IP85	CACAATGATTTTCAAAGATG			
3sp-IP7	CATGGTCTCAACTGAC	*cpsE-cps1bG*	2,300	50
3sp-IP38	AGGTCTAACTTCATGTCTAC			
3sp-IP17	GAAAGAGGACTTTGCTTGCC	*cps1bF-cps1bI*	2,200	59
3sp-IP36	GTCAAAGTGTCATCGACAATCTC			
3sp-IP23	TACGGTGCAATAGTTGCC	*cps1bI-cps1bK*	2,800	56
3sp-IP45	GTTTTTACCTAATCTATCCTTACTA			
3sp-IP39	GAACATCATCATATTGATCG	*cps1bK-cps1bM*	2,700	54
3sp-IP53	TACATACCACAAACATAATCC			
3sp-IP50	AGTGCTCTCGTAATAGCTCTGG	*cps1bM-cps1bO*	2,800	56
Sp-cps52	ACACCAGCAGAAATTAGTTC			
3sp-IP58	GAATGACTATTACGATGTCTC	*cps1bO-cpsR*	2,700	54
3sp-IP85	CACAATGATTTTCAAAGATG			
3sp-IP5	GTACTCGATTAAGCGAC	*cpsE-cps2F*	2,200	53
Sp-cps7	GGACATACCATTTGCAGTA			
3sp-IP14	GAATCAGGAATCAGACCTGG	*cpsE-cps2G*	1,900	58
3sp-IP72	TGAACATTGCAGGCTATACC			
Sp-cps8	TAGCTGATGGTGCTCATG	*cps2G-cps2I*	1,500	55
3sp-IP67	GTACTTGCCGAAACTATTGTG			
3sp-IP61	GGTAGAGAAATGAATGAATATCG	*cps2H-cps2J*	2,100	55
3sp-IP68	TTCAGGTCCACCACCGATG			
3sp-IP65	GGAGATTGTAATGAGGTAAATC	*cps2J-cps2L*	2,200	55
3sp-IP66	CACCAATATACACAGTCTAAATC			
Sp-cps23	GCAAATGATTCAGATGA	*cps2L-cps2N*	2,300	55
Sp-cps53	CTGGTATGCCACCAACAATC			
3sp-IP78	TGAGCATCGAACAGATACAG	*cps2N-cps2P*	2,900	55
Sp-cps12	TTTTGAATCATTGAAGTCG			
3sp-IP86	CTTAATGATACATTTGCAACAG	*cps2P-cpsR*	2,000	55
3sp-IP85	CACAATGATTTTCAAAGATG			

Table 2.3. Proteins of other bacterial species homologous to deduced amino acid sequences of genes in the conserved region

Gene	Category	Protein showing the highest identity			Identity
		Organism	Product	Accession no.	
lysR	Regulator	*Streptococcus uberis*	LysR family transcriptional regulator	WP_037627683	91%
cpsA		*Streptococcus uberis*	LytR family transcriptional regulator	WP_037627066	79%
cpsB		*Streptococcus uberis*	Tyrosine protein phosphatase	WP_037627064	92%
cpsC		*Streptococcus uberis*	Capsular biosynthesis protein CpsC	WP_037627063	87%
cpsD		*Streptococcus porcinus*	Tyrosine-protein kinase	WP_003085167	81%
cpsE	Epimerase	*Streptococcus pseudoporcinus*	Short-chain dehydrogenase	WP_007893078	80%
cpsQ		*Streptococcus uberis*	UDP-glucose 6-dehydrogenase	WP_046393482	86%
cpsR	Hypothetical	*Streptococcus uberis*	Hypothetical protein	WP_037627032	89%

Table 2.4. Proteins of other bacterial species homologous to the deduced amino acid sequences of genes in the serotype-specific region

Gene	Category	Protein showing the highest identity			Identify
		Organism	Product	Accession no.	
cps1aF	Initial glycosyltransferase	*Streptococcus uberis*	UDP-phosphate galactose phosphotransferase	WP_012658534	88%
cps1bG		*Lachnospiraceae*bacterium	UDP-galactose phosphate transferase	WP_009261689	67%
cps2F		*Streptococcus suis*	initial sugar transferase	FAA00872	60%
cps1aL	Polymerase	*Megasphaera sp.*	O-antigen ligase	WP_023052747	25%
cps1bL		*Clostridium sp.*	O-antigen polymerase	WP_022113405	25%
cps2O		*Streptococcus pseudoporcinus*	oligosaccharide repeat unit polymerase	WP_007893424	75%
cps1aN	Flippase	*Lactobacillus paraplantarum*	flippaseWzx	CDF77686	41%
cps1bM					42%
cps2P					36%
cps1aH	Glycosyltransferase	*Streptococcus uberis*	glycosyltransferase	WP_012658532	81%
cps1bI					61%
cps1aJ		*Acetobacteriumwoodii*	glycosyltransferase family 1	WP_041668735	48%
cps1bJ					46%
cps1aK		*Fibrobacter succinogenes*	glycosyltransferase family 2	WP_014546699	38%
cps1bK					39%
cps2I		*Streptococcus pseudoporcinus*	glycosyltransferase family 1	WP_007893181	77%
cps2J		*Streptococcus pseudoporcinus*	glycosyltransferase family 1	WP_007893376	70%
cps2M		*Bacteroides sp.*	glycosyltransferase, group 2 family	EFI06947	49%

cps1aI	Acetyltransferase	*Streptococcus uberis*	maltose O-acetyltransferase	WP_046388669	64%
cps1aM		*Clostridium akagii*	acetyltransferase	WP_035795827	49%
cps1bH		*Ruminococcusalbus*	acetyltransferase	WP_024858893	56%
cps2N		*Streptococcus pseudoporcinus*	acetyl transferase	WP_007893375	65%
cps2H		*Streptococcus didelphis*	GNAT family acetyltransferase	WP_018366072	63%
cps2L		*Streptococcus pseudoporcinus*	GNAT family acetyltransferase	WP_007893003	80%
cps1aG	Epimerase	*Streptococcus uberis*	NAD-dependent epimerase	WP_046392145	72%
cps1aP		*Streptococcus porcinus*	NAD dependent epimerase/dehydratase family protein	WP_003085154	83%
cps1bO					84%
cps1bF	Aminotransferase	*Streptococcus suis*	aminotransferase DegT	WP_029175971	81%
cps2G		*Streptococcus suis*	aminotransferase	AKE79842	68%
cps2K		*Butyrivibrioproteoclasticus*	aminotransferase	WP_026662038	69%
cps1aO	Hypothetical	*Streptococcus equinus*	hypothetical protein	WP_033152408	50%
cps1bN					45%

Table 2.5. Homology of deduced amino acid sequences between the genes in the serotype-specific region of the *cps* loci

Category	Predicted protein	Identity (%)	Category	Predicted protein	Identity (%)
Initial glycosyltransferase	Cps1aF – Cps1bG	40	Polymerase Wzy	Cps1aL – Cps1bL	31
	Cps1aF – Cps2F	18		Cps1aL – Cps2O	10
	Cps1bG – Cps2F	23		Cps1bL – Cps2O	11
Glycosyltransferase	Cps1aH – Cps1bI	60	Flippase Wzx	Cps1aN – Cps1bM	72
	Cps1aH – Cps2I	9		Cps1aN – Cps2P	37
	Cps1aH – Cps2J	7		Cps1bM – Cps2P	40
	Cps1aH – Cps2M	8	Hypothetical	Cps1aO – Cps1bN	61
	Cps1bI – Cps2I	10	Epimerase	Cps1aP – Cps1bO	93
	Cps1bI – Cps2J	12	Aminotransferase	Cps1bF – Cps2G	11
	Cps1bI – Cps2M	6		Cps1bF – Cps2K	13
	Cps1aJ – Cps1bJ	49	Acetyltransferase	Cps1aI – Cps1bH	12
	Cps1aJ – Cps2I	9		Cps1aI – Cps2H	9
	Cps1aJ – Cps2J	13		Cps1aI – Cps2L	6
	Cps1aJ – Cps2M	8		Cps1aI – Cps2N	10
	Cps1bJ – Cps2I	8		Cps1aM – Cps1bH	7
	Cps1bJ – Cps2J	10		Cps1aM – Cps2H	7
	Cps1bJ – Cps2M	9		Cps1aM – Cps2L	7
	Cps1aK – Cps1bK	53		Cps1aM – Cps2N	28
	Cps1aK – Cps2I	11		Cps1bH – Cps2H	7
	Cps1aK – Cps2J	7		Cps1bH – Cps2L	7
	Cps1aK – Cps2M	5		Cps1bH – Cps2N	8
	Cps1bK – Cps2I	5			
	Cps1bK – Cps2J	7			
	Cps1bK – Cps2M	10			

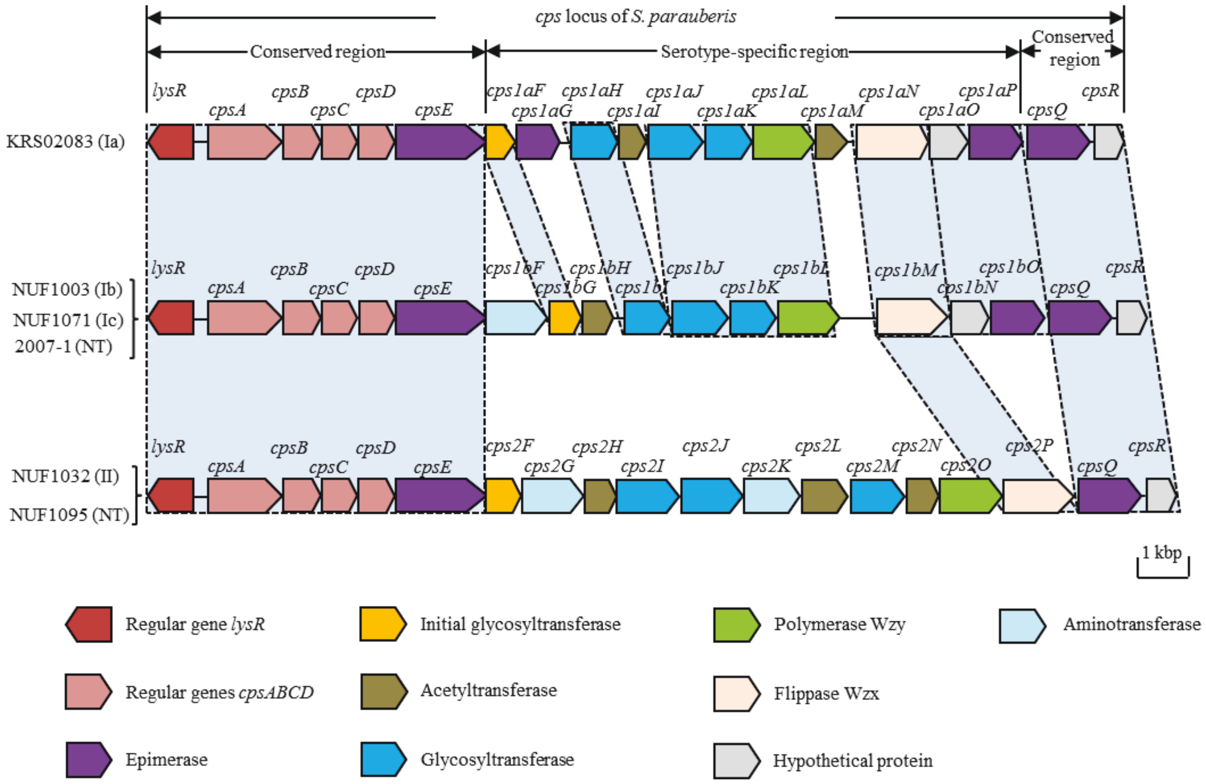


Fig. 2.1. Structure of *cps* loci of *S. parauberis* representative strains. Putative functions of genes are differentiated using different colors. Shaded regions indicate the genes whose deduced amino acid sequences showed over 30% identity among them.

Chapter 3

A multiplex PCR assay for differentiation of *S. parauberis* serotypes

3.1 Introduction

Serotyping of bacteria is very important for bacterial epidemiology. In chapter 1, I demonstrated that the isolates of *S. parauberis* in Japan were divided into five serological phenotypes (subserotypes Ia, Ib and Ic, serotype II and non-typeable) according to the surface polysaccharide capsular antigens. However, the identification of *S. parauberis* serotypes by slide agglutination test is sometime difficult, because most institutes and laboratories in the world do not have the antisera for serotyping. Thus, an alternative diagnostic method is required. In chapter 2, it was demonstrated that there were three distinct *cps* loci in *S. parauberis*, which corresponded to subserotypes Ia and Ib/Ic and serotype II, respectively. Moreover, the non-typeable strains of Cluster I and III in *Sma*I-PFGE had similar *cps* loci to subserotype Ib/Ic and serotype II, respectively. Base on these findings, in this chapter, a multiplex PCR method for the detection of *S. parauberis* serotypes was investigated.

3.2 Materials and Methods

Bacteria

A total of 188 *S. parauberis* strains used in the previous study was used. To verify the specificity of the multiplex PCR, the *S. parauberis* strain ATCC13386 derived from cow and the other Gram-positive cocci, *S. agalactiae* JCM5671, *S. thermophilus* JCM17834*, S. uberis* ATCC19436*, S. iniae* NUF631 from the Japanese flounder*, S. dysgalactiae* NUF1015 from greater amberjack *Seriola dumerili, Enterococcus faecalis* JCM5803*, Lactococcus lactis* subsp. *lactis* JCM5805 and *L. garvieae* NUF699 from the Japanese flounder; and the Gram-negative pathogens, *Edwardsiella tarda* NUF251from the Japanese flounder*, Vibrio anguillarum* NUF652 from greater amberjack and *Photobacterium damselae* subsp. *piscicida* NUF89 from yellowtail *Seriola quinqueradiata*, were used as reference strains. Strains of JCM were provided by Japan Collection of Microorganisms, RIKEN BRC which is participating in the National BioResource Project of the MEXT. For the cultivation of Gram-positive cocci and Gram-negative pathogens, Todd Hewitt agar and Heart

Infusion agar (Difco), respectively, were used. The bacteria were cultured at 27°C for 24 h.

Genomic DNA preparation

The genomic DNA of the 188 *S. parauberis* strains prepared in the previous chapter were also used in this study as templates. The genomic DNAs of the other bacterial strains were prepared using Wizard Genomic DNA Purification kit (Promega) according to the manufacturer‟s instructions.

Primer design

Based on the multiple sequence alignments of polysaccharide polymerase gene (*wzy*) in the type-specific region of *S. parauberis*, three pairs of serotype-specific primers were designed using the Primer-BLAST program (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) in order to allow simultaneous amplification in the same conditions with specificity (Fig. 3.1). The designed primer pairs were named For-Ia (5'-ATTGTTAGTCATTCAGTTGT-3') and Rev-Ia (5'-AAT TATAGTCAACAGTCCAG-3'), For-Ib/Ic (5'-ATTTCTACCAGGTTACTTTG-3') and Rev-Ib/Ic (5'-ACATCTCGAAACTTCATATT-3') and For-II (5'-GAACTACTT AGGTTTAGCAT-3') and Rev-II (5'-AACTTGTAAATAGGATTGCT-3'). They will give the amplification products of 213, 303 and 413 bp for subserotypes Ia and Ib/Ic and serotype II, respectively.

PCR amplification

The PCR was performed on C1000 Thermal Cycler (Bio-Rad) with *TaKaRa Ex Taq* Hot Start Version (Takara). The PCR reaction mixture (a total volume of 50 μL) contained 5 μL of 10× Ex Taq buffer (Mg^{2+} plus), 4 μL of dNTP mixture (0.2 mM each), six primers (0.5 μM each), 1 μL of template DNA and 1.25 U of Ex Taq DNA polymerase. The condition of PCR amplification was denaturation at 95°C for 3 min, 30 cycles of denaturation at 98°C for 10 s, annealing at 55°C for 30 s and extension at

72°C for 30 s and a final extension at 72°C for 7 min. The PCR products were electrophoresed in 2.0% agarose gel followed by staining with ethidium bromide.

3.3 Results

Using the designed primer set, PCR products with the expected size specific for subserotypes Ia, Ib and Ic and serotype II were obtained (Fig. 3.2). In all 188 *S. parauberis* strains from the Japanese flounder, the results were consistent with those of agglutination test using rabbit antisera, except that the subserotypes Ib and Ic could not be distinguished each other by this PCR method. The NT strains could be differentiated into subserotype Ib/Ic or serotype II. All the reference strains including the *S. parauberis* ATCC13386 derived from cow were negative. The test results are summarized in Table. 3.1.

3.4 Discussion

In streptococci, the PCR assays developed for serotype differentiation were focused on the *cps* locus, because in many cases the construction and nucleotide sequence of *cps* locus were different according to serotypes (Smith *et al*., 1999; Wang *et al*., 2011b). In the previous chapter, it was shown that three genes encoding initial glycosyltransferase, polysaccharide polymerase (Wzy) and flippase (Wzx) were present in the *cps* loci of all *S. parauberis* serotypes. A multiple sequence alignment of these genes showed that the *wzy* gene was less similar among the serotypes than the other two genes (S16-S18). Accordingly, the primers for the multiplex PCR were designed from the *wzy* gene. However, the primer pair for differentiation between subserotypes Ib and Ic could not be designed, because the sequences of *cps* loci of two subserotypes were almost identical (chapter 2). The same case was found in *S. suis* (Liu *et al.,* 2013). Antisera may be required in such situation to aid distinction.

The NT strains of *S. parauberis* which agglutinate with both serotype I and II rabbit antisera are thought to be noncapsulated (Kanai *et al*., 2015). About one-third of them were derived from apparently healthy fish. We suspected that their *cps* loci were different from those of capsulated strains, but the sequencing of the *cps* loci revealed that the NT strains had almost the same sequences as subserotype Ib/Ic or serotype II (Tu *et al*., 2015). Accordingly, it is suggested that the NT isolates are originally subserotype Ib/Ic or serotype II *S. parauberis* and that their capsule production is suppressed by mutation in the gene(s) for capsule production. (chapter 2).

Although the multiplex PCR developed in this study could not differentiate the subserotypes Ib and Ic, the other serotypes including non-typeable strains could be differentiated accurately and specifically, and no positive amplification was detected in the reference strains used. So, it would be a useful tool for diagnosis and serotyping of fish-pathogenic *S. parauberis*.

3.5 Conclusion

1. The developed multiplex PCR assay could differentiate the serotypes accurately.
2. The non-typeable strains could be differentiated by this multiplex PCR and the result corresponded with PFGE.
3. This multiplex PCR assay is specific for diagnosis and serotyping of fish-pathogenic *S. parauberis*.

Table 3.1. The results of the multiplex PCR assay for differentiation of the *Streptococcus parauberis* serotypes

Bacteria	Primer pair for amplification		
(no. of strain)	For/Rev-Ia	For/Rev-Ib/Ic	For/Rev-II
S. parauberis from flounder			
subserotype Ia (6)	＋	－	－
subserotype Ib (91)	－	＋	－
subserotype Ic (7)	－	＋	－
serotype II (62)	－	－	＋
NT/Cluster I[*1] (9)	－	＋	－
NT/Cluster III[*1](13)	－	－	＋
S. parauberis from cow			
ATCC13386	－	－	－
Other species(11)[*2]	－	－	－

＋, positive amplification;－, negative amplification; NT, non-typeable.

[*1] Clusters in *Sma*I-PFGE analysis (Kanai *et al.*, 2015).

[*2] *S. agalactiae* JCM5671, *S. thermophilus* JCM17834, *S. uberis* ATCC19436, *S. iniae* NUF631, *S. dysgalactiae* NUF1015, *Enterococcus faecalis* JCM5803, *Lactococcus lactis*s ubsp. *lactis* JCM5805, *L. garvieae* NUF699, *Edwardsiella tarda* NUF251, *Vibrio anguillarum* NUF652 and *Photobacterium damselae* subsp. *piscicida* NUF89.

```
KRS02083(Ia)       ATGA-------------ATA------ATTCTCAAATATTTGA----TAAATCTCACAATA 37
NUF1003 (Ib)       ATGG-------------ATA------ACTTTAAAATAGGTGA----TTCTTCATTGAATA 37
NUF1032 (II)       ATGACACCAATTATTTTATATTCCATATTATTAGGTATTTTACTTTTATACTGTCTAATA 60
                   ***              ***      * * * *  **  * *    *         ****

KRS02083(Ia)       A-TA-------TTCTCACAACTAGG-ATTA-CA----CTTGTTTTATTAGTTATATTTTT 83
NUF1003 (Ib)       A-AAAAACGAATCATCATAATAAAGTATTAATA----TTTAGCTTATTATTAATAATTTT 92
NUF1032 (II)       AATAAAACGAACTACTTAGGTTTAGCATTAGTAGCAGTCTATTTTATTATTT-CACTGAT 119
                   *  *                    * ****  *      *   ****** *   * *  *

KRS02083(Ia)       GACAATATATGGTTCCGGA--GA----ATTTAAAGTTCCTTCT--TTCAGTATAAACGT- 134
NUF1003 (Ib)       TATCATTAATTTTACTGGTTTAG----ACATAAAATTCATCCCAGTTC--TTCAAATGTC 146
NUF1032 (II)       TACTCTAT-TAGTACCAACAAGAGGTTATTTGAAATACTCTTATGTTTTAGATGGAAGTA 178
                    *   *   *  * *            *  * ** * *       **        * **

KRS02083(Ia)       -----AGCTGTAAAATA-TATCATTT----TAACTT-TTGGGCTAT--TTTTTTCAGTTT 181
NUF1003 (Ib)       A----AATTTTAATAAG-TATATTTC----TATTTT-ATTAATAATAGTTTTTTCAATTT 196
NUF1032 (II)       ATATTACTTATGGAATAGTATTATTTATGGTGTTTTCATTTGCTAT--TTTTTTCTTTCC 236
                        *  * *   *   ***  **     *   **  *     **  *******  *

KRS02083(Ia)       -TTTTTATTAAAGTAAAAAAATTTTCAAT------AA--TAAAGTTAGAT----ATAATT 228
NUF1003 (Ib)       -TTCTT-TCAATATTTCAAGA----CAAT------AAACTTAAGTTTGAT----CAAATT 240
NUF1032 (II)       GTTTCTGATAAAA-AAGAAAAGTTTCAATGTTAAAAATATTGAGCTAAATGTTCATGAAA 295
                    **  *   **      ** *    ****      **  *  ** *  **       *

KRS02083(Ia)       TCTCAGTTGTTATTGATTCATTTATTGTATAAT----ACAATAATTATTGTATTAAATTC 284
NUF1003 (Ib)       TCAATAATACTTTTTTGTCAGCTTTTATACTATTTTTGCGTT--TTATTAT-TTAG--TA 295
NUF1032 (II)       ACTACATTATTTTTGCCAAAATATTTATATTTTT--TGCACTGATTACTGTTCTCG---- 349
                    *     *  * **     *    ** **   *     *  *  *** * *  *

KRS02083(Ia)       GTCCGATA-ATTACGTTTTATCGTACTTTACATATGTATTTGCATTCTTAGGTTATTTAT 343
NUF1003 (Ib)       GCTCTACAGACTATTCTTTATTATTTTTCTCATATACTTTTTCTTTCTTTGCTTACATAT 355
NUF1032 (II)       --TTTATA---TGGTCAAAATTATTCCTCTTATACAG----------TCAGGTCA---AT 391
                        * *   *       **  *   *   ***             *  * * *   **

KRS02083(Ia)       GGGGAAGTAACAGCCTAATTGACAAAAAAACGT-TCAAAATATTCTTATCAAGTTTTTAC 402
NUF1003 (Ib)       GGATTAATAATAACAAAAT-ATCAAAAGATCATCTTTTAAT-TTTTTCTCGAGTCTTTT- 412
NUF1032 (II)       GGGCTAGTAATCGATTAGT-----AATGAAC------------TCTGATCAAGCCATAAT 434
                   **   * ***      * *     **  * *            * *  ** **   *

KRS02083(Ia)       TATATCCTTATTA-----TTCAAACATTATTTACCGTATA-CATTGTTAGTCATTCAGTT 456
NUF1003 (Ib)       --TATCAATATTAATAATTTCTATTATCATT-ACAGTATA-CA--ATTTATC--TAAATT 464
NUF1032 (II)       ----TCCATATAA--------CAATATCTTTGAAAAAATAGCA------ATC--CTATTT 474
                       **  *** *         *  **  ** *    *** **       **    * **

KRS02083(Ia)       GTGGAT---ATGTATTTGTTTAAAAATGGTATTGTTATT--CCTATAGGAGCATCAAATG 511
NUF1003 (Ib)       ACAAATTCCACTTTTTCTTTTTAAATCTCAAATAATATTA-CCAATTGGGTCTTCAAATG 523
NUF1032 (II)       ACAAGTTATACTCAATTACTAGCAATTGTAGTTGGTTTTAATTTACTGAG--AACAAAT- 531
                        *   *     *   *   **       *  * **     *  *      *****

KRS02083(Ia)       GGATTACAACATTTATTGTTATGATT---TTTCCAATATTGTATAAATTAAGCAATAGTA 568
NUF1003 (Ib)       CTATTACAACTTATACATTTTTACTT---TTACCTATAGTATATTACTTGGATAAAAAT- 579
NUF1032 (II)       ------CAACAT-TATATGCTTGGCTCAATTGTCT-TATTA-ATGGTTGGAGCAACAGA- 581
                         **** * **      *   *   ** *  **  *  **   *     ** *

KRS02083(Ia)       GAACTTCAC---AAT--ATTTTTTAACTATATTTACTATGATTTTTGCTGTATTATCTCG 623
NUF1003 (Ib)       --ACTTTAAGAAAATCAATTTTTTTGTTTGG--TACCTTCATATTAGTTCTTTTATCAAG 635
NUF1032 (II)       --GTTTTGT----ATTGATATGTATGTATCATCTCGGGGGATGCTAGCCATATTT----- 630
                       **       **  ** * *          *      **  * *   * **

KRS02083(Ia)       TTCAAATTCTGGA-CTGTTGACTATAATTGCTATAATTCTTATTCTGTTTATGCAGGAA- 681
NUF1003 (Ib)       ATCCAACTCGGGAATTATTG-TTATGGGAACAATGATCTTTTATATGTTGAT-TAAAAAT 693
NUF1032 (II)       -----ATTCT--ATTTATTGGTT-----CACTATA---TTTATTCT-TTTATAAAGAAAT 674
                        * **   *  * ***  *       * **     **  * * ** **  *  **
```

```
KRS02083(Ia)       --AAAAAATATAAACT--AATACGTGGGATTCTA--ATCTTTTTAATATTTCTATTAT-- 733
NUF1003 (Ib)       TCAAAAAATAAATGGTTGAATATATTATTTTCTGCGGTATGTTTATTGCTAATATTAT-- 751
NUF1032 (II)       T-AATAAAAAAAGTCGGAAATATATTGATTT------TATTTTTATTTTAACTATAGTAA 727
                     ** *** * *      ****  *    **      * * **** *     ***  *

KRS02083(Ia)       ------TTTTATATTTAATCGGTAAGTATTCTCCTGGATATCTTTCAAGATTTTCTTC-- 785
NUF1003 (Ib)       ------TTTTG-ACTTCAC-----AATTTCTACCAGGTTACTTTGAAAGATTTTCAAA-- 797
NUF1032 (II)       CAAGTGTTCCATACTTGAT--------TTCGGTAACAGTATC----AAGATTTTCATCCT 775
                         **    * ** *          *         **      *********

KRS02083(Ia)       -TACATTGCAGTCATTAATTACAGATCAAAGTA---CTAACCAAACTAAAGCAATGAATG 841
NUF1003 (Ib)       -TGTAATAAATACCATAACTGGAAATAGTAGTATTTCTAATAATATTGATGCTTTAAATG 856
NUF1032 (II)       CTGCTATTAGTTCATTGATT--------TATTATT-TTGGGCAAAC----ACCATATATG 822
                    *    *     *  * * *         * **    *    * *      *  *  ***

KRS02083(Ia)       GTAGGATT-GAAGT--GTTCCATAGCGCCTTATATCATATAAAGAATCATTTTTTCATTG 898
NUF1003 (Ib)       GTCGAGCT-CAGGTTTATTCGAT--TGCTAGAAGTTTAATTAGCAAGGATTTTCTTTTTG 913
NUF1032 (II)       TTCTCATTAGAAGT----------------TAAATCAT-TGAAAAAACTTATGTTTGGTG 865
                    *     *  * **                 *  *    * *  **   * *  *   **

KRS02083(Ia)       GAAATGGATTTGGA-TATAG--AGAAAGAATGCCTAGCTATTTAATGACTCATAA--TTG 953
NUF1003 (Ib)       GAATTGGTTTTGTT-TATCG--AAGTTATATGCCAAGTCTGCTAATGACTCATAA--CTG 968
NUF1032 (II)       AATATGGATTTGGTGCACTGGCAGGTGGGATG--AAATTTTCAGATGAACTAGGAATTTG 923
                    *  *** ****    *  *  *      ***   *        ****   *  *   **

KRS02083(Ia)       GTTATTAGAATATCTAATTACAGGTGGAATAATATCATTTTTACTAAAAAT--ATTT--A 1009
NUF1003 (Ib)       GATTTTAGAATCTTTGATTACAGGTGGTATTTTTGCATTAATGATAAGAAT--AAT---A 1023
NUF1032 (II)       GGTTCGTGGATTTTATACTTTTCTTGGT-TGGTTATATGCCGACTGGGGATTTGTTGGTA 982
                   * *    * ** *   * *     ***  *  *   **      *    **    *   *

KRS02083(Ia)       TCTTTATT-TTACAATT---T--TTAAAACTTGTAACTGTCAAAAATAATGCACTAAAAC 1063
NUF1003 (Ib)       CTTTTATG-CCAAATTT---TGTTTGAAAATATGAAGTTTCGAGA----TGTTCAAATAC 1075
NUF1032 (II)       TTTTTATTATTGGAATTGCTTGCTTGATATTCTTCACTATAATGA----TAAATAAAC-C 1037
                     *****      * **   *  ** * * *    * * *    *    *     **  *

KRS02083(Ia)       AAGGCTT--AATTATT---------TCATT--T---GTATTTG----TTCTAATTCAAGG 1103
NUF1003 (Ib)       AAAATTG--TATTAGTG--------TGATTATT---ATATTTT----CATTAATCCAAGG 1118
NUF1032 (II)       AAAGCTTGATATTAGTGATACTTTTTTACTACTTGGATATTATAAATTACTCATTCAGGG 1097
                   **   *    **** *         * * *  *    ****         * ** ** **

KRS02083(Ia)       TTT----------AGTTGAACCATCATTTGGTTCTCCATTATTTGAGTTGATTTTTGCAT 1153
NUF1003 (Ib)       GCT----------GGTTGAACCATCTTTTGGAGGACCGGTTTTCGAGTTAGTTTTTTGGC 1168
NUF1032 (II)       TGTCTTCACCATGGGTAGAACAAAGATATATTCAATC-TTAATCAGTTTGATTATTTATG 1156
                     *           ** **** *   * *       *  *  *    **  ** **

KRS02083(Ia)       TAATAATTGGATTT----------GGTACAAATACGCTTTACGAGGAG-AAAAATATTTA 1202
NUF1003 (Ib)       TAATCATGAGCTTTTT--------AATTCAAAT-CGGTAGAAAACCAGTAGGAATGTTCA 1219
NUF1032 (II)       TATTAATAAAATTTATAGTCGAGAGATTTAGATTTATTATAAGA--AGGAAACAGGTTTA 1214
                   ** * **    ***            *  * **    *  *  *  ** *   *  ** *

KRS02083(Ia)       -TGATTAA--- 1209
NUF1003 (Ib)       ATCAGCAATAA 1230
NUF1032 (II)       CTAA------- 1218
                    * *
```

Fig. 3.1. Multiple sequence alignment of *wzy* gene in the serotype-specific region of *S. parauberis cps* loci. The primers used for multiplex PCR and their directions were indicated under the sequences with red arrows.

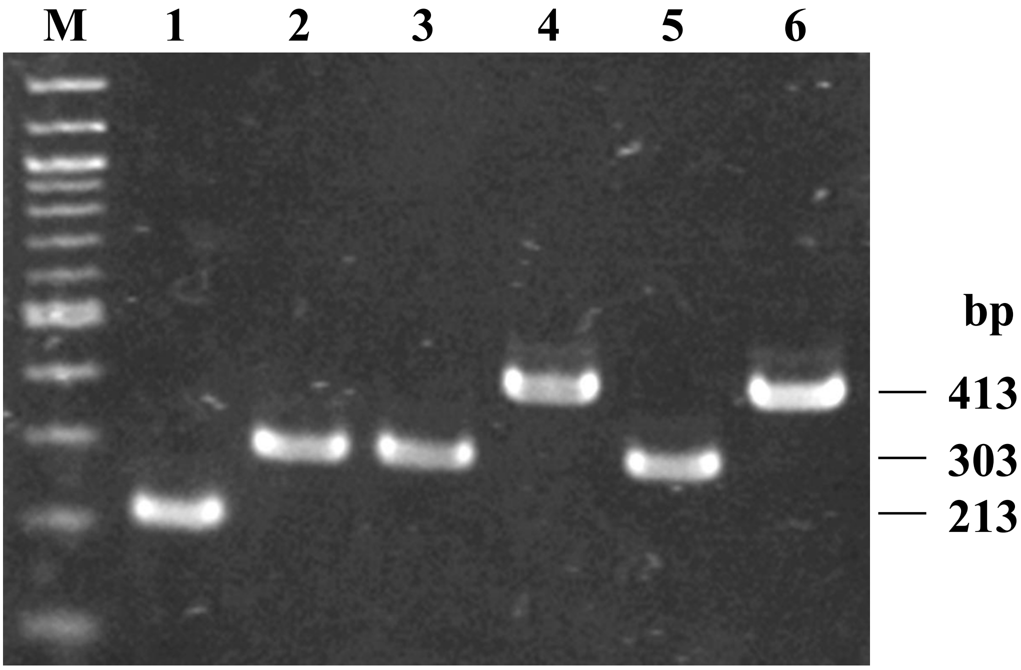


Fig. 3.2. Multiplex PCR products of *Streptococcus parauberis*. PCR products were electrophoresed on 2% agarose gel, stained with ethidium bromide and photographed under UV light. M, 100-bp DNA ladder markers;1, subserotype Ia; 2, subserotype Ib; 3, subserotype Ic; 4, serotype II; 5 and 6, nontypeable strains of clusters I and III in *Sma*I-PFGE analysis, respectively.

General discussion

Streptococcus parauberis emerged as an important cause of streptococcal infection in cultured Japanese flounder since 2002 and steadily spread across the western districts of Japan. Accurate and rapid identification of this pathogen is important for the epidemiological and pathogenesis studies and chemotherapy of the infection.

As a standard method, serotyping has been used for typing of bacteria. According to the result of previous study, it was considered that there were only two serotypes in *S. parauberis* strains isolated from Japanese flounder (Kanai *et al*., 2009). However, there veritably existed three subtypes (subserotypes Ia, Ib and Ic) in serotype I (Table 1.2). Although the serotyping method has some limitation such that it could not adequately reflect the clonal diversity of bacterial strains and that some strains are nontypable with it, it can exhibit the relationship among strains visually. The PFGE genotyping method used in this study exhibited robust subdivision power and well corresponded to serotyping. Hence, identification of pathogens with both serotyping and genotyping is more useful for epidemiological study.

The serotyping investigation in this study not only clarified the existence of subserotypes in serotype I but also indicated the relationship among them. However, the reason why cross-reactions occur among them was unknown. It is important to know the common antigens that cause the cross-reactions. For this purpose, the capsular polysaccharides of subserotypes should be purified and analyzed by biochemical methods, and the monosaccharide composition should also be determined. The possible correlation between the monosaccharide composition and cross-reactions may be clarified. The similar cross-reactions were found in some *S. suis* serotypes (Higgins and Gottschalk, 1990).

Polysaccharide capsules are crucial virulence factors and the active ingredients of vaccines and closely correlative to bacterial serotypes. They are generally synthesized by genes which are clustered in a single locus on the chromosome (Roberts, 1996). Thus, studies on the genetic loci for biosynthesis of CPS are very significant. At the start of this study, it was considered that there would be distinct structure of *cps* loci corresponding to respective serotype. However, the result showed that subserotypes Ib and Ic had a common *cps* locus with only few single-base substitution found between them. The similar case was also found between capsulated and noncapsulated types. These minor changes in sequences might cause mutation events that affect the biosynthesis and structure of CPS and lead to the difference of phenotypes. Lakkitjaroen *et al*. (2014) investigated the noncapsulation mechanisms in 25 endocarditis isolates of *S. suis* and found that a single-nucleotide substitution or frameshift mutation in two glycosyltransferase genes (*cps2E* and *cps2F*) were the main causes of the capsule loss. In future studies, clarifying the mechanisms that can switch between subserotypes Ib and Ic and between capsulated and noncapsulated types in *S. parauberis* are required.

The biosynthesis of CPS requires a complex pathway. The present works preliminary presumed that the biosynthesis pathway in *S. parauberis* may be Wzy-dependent pathway, because some principle genes (initial glycosyltransferase gene, polysaccharide polymerase *wzy* and flippase *wzx*) for Wzy-dependent pathway were present in the *cps* loci of *S. parauberi*s. Nakhamchik *et al*. (2007) used a new mini-Tn10 delivery vector, pNKTXI-SceI, to generate a mutant library and identify genes essential for CPS biosynthesis. The research demonstrated that Wzy polysaccharide polymerase was required for CPS biosynthesis in *Vibrio vulnificus.* In *S. parauberi*s, it is necessary to clarify whether the initial glycosyltransferase, flippase and polysaccharide polymerase are required for biosynthesis of CPS and determine the biosynthesis pathway.

The genes of *cps* loci individually or correlatively play the vital role in the biosynthesis of CPS. In *S. parauberis,* the functions of most *cps* genes were not verified yet. In order to better understand the mechanism of CPS biosynthesis in this pathogen, studies on the function and activity of each gene product will be necessary by the biochemical and genetic research such as mutant construction. Lakkitjaroen *et al*. (2014) also investigated mutations in the genes involved in side-chain formation (*cps2J* and *cps2N*), polymerase (*cps2I*) and flippase (*cps2O*) appeared to be lethal and these lethal effects were relieved by mutations in the *cps2EF* region.

Streptococcosis is responsible for extensive losses in the worldwide aquaculture industry. Han *et al*. (2011) demonstrated that a combined vaccine containing formalin-inactivated *Edwardsiella tarda*, *Streptococcus iniae* and *S. parauberis* had the protective efficacy in farmed Japanese flounder in Korea. In this study, four serotypes in *S. parauberis* were demonstrated. Therefore, the combined vaccine may not be effective for all these serotypes. For prevention and treatment of the infection caused by *S. parauberis*, a vaccine common to all serotypes was required, although Aviles *et al*. (2013) indicated that a conserved surface M-protein, SiMA, of *S. iniae* is not effective as a cross-protective vaccine against different capsular serotypes in farmed fish. But, research in *Salmonella enterica* serovar Typhimurium infection indicated that immunization with the *S.* Typhimurium *wecA*-mutant strain confered protection against lethal challenge with wild type *S.* Typhimurium and *S.* Enteritidis (Bridge *et al*. 2015). This may provide to ponder for vaccine strategy of streptococcal infection in farmed Japanese flounder.

In this study, by investigating the serological relationship among *S. parauberis* isolates, it was demonstrated that there were five serological phenotypes (subserotypes Ia, Ib and Ic, serotype II and nontypeable) in it. The research about genetic structure of *cps* loci can provide the foundational knowledge for the synthesis pathway of CPS and the reasons of variation in serotypes of *S. parauberis* and serve for development of an accurate, rapid detection method for this pathogen. Above all, the current study would take a better understanding of this pathogen and contribute to the future study for prevention and treatment of this disease.

References

Aviles, F., M. M. Zhang, J. Chan, J. Delamare-Deboutteville, T. J. Green, C. Dang and A. C. Barnes (2013): The conserved surface M-protein SiMA of *Streptococcus iniae* is not effective as a cross-protective vaccine against differing capsular serotypes in farmed fish. *Vet. Microbiol.*, **162**, 151-159.

Bachrach, G., A. Zlotkin, A. Hurvitz, D. L. Evans and A. Eldar (2001): Recovery of *Streptococcus iniae* from diseased fish previously vaccinated with a *Streptococcus* vaccine. *Appl. Environ. Microbiol.*, **67**, 3756-3758.

Baeck, G. W., J. H. Kim, D. K. Gomez and S. C. Park (2006): Isolation and characterization of *Streptococcus* sp. from diseased flounder (*Paralichthys olivaceus*) in Jeju Island. *J. Vet. Sci.*, **7**, 53-58.

Bentley, S. D., D. M. Aanensen, A. Mavroidi, D. Saunders, E. Rabbinowitsch, M. Collins, K. Donohoe, D. Harris, L. Murphy, M. A. Quail, G. Samuel, I. C. Skovsted, M. S. Kaltoft, B. Barrell, P. R. Reeves, J. Parkhill and B. G. Spratt (2006): Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet.*, **2**, e31.

Bridge, D. R., J. M. Whitmire, J. J. Gilbreath, E. S. Metcalf and D. S. Merrell (2015): An enterobacterial common antigen mutant of *Salmonella enterica* serovar Typhimurium as a vaccine candidate. *Int. J. Med. Microbiol.* (in press)

Cieslewicz M. J., D. L. Kasper, Y. Wang and M. R. Wessels (2001): Functional analysis in type Ia group B *Streptococcus* of a cluster of genes involved in extracellular polysaccharide production by diverse species of streptococci. *J. Biol. Chem.*, **276**, 139-146.

Creuzenet, C. (2004): Characterization of CJ1293, a new UDP-GlcNAc C6 dehydratase from *Campylobacter jejuni. FEBS Lett.*, **559**, 136-140.

Doménech, A., J. F. Fernández-Garayzábal, C. Pascual, J. A. Garcia, M. T. Cutuli, M. A. Moreno, M. D. Collins and L. Dominguez (1996): Streptococcosis in cultured turbot, *Scopthalmus maximus*(L,) associated with *Streptococcus parauberis*. *J. Fish Dis.*, **19**, 33-38.

Fakhr, M. K., L. K. Nolan and C. M. Logue (2005): Multilocus sequence typing lacks the discriminatory ability of pulsed-field gel electrophoresis for typing *Salmonella enterica* serovar Typhimurium. *J. Clin. Microbiol.*, **43**, 2215-2219.

Han, S. Y., B. K. Kang, B. J. Kang, J. M. Kim, J. E. Han, J. H. Kim, C. Choresca Jr, S. P. Shin, J. W. Jun and S. C. Park (2011): Protective efficacy of a combined vaccine against *Edwardsiella tarda*, *Streptococcus iniae* and *Streptococcus parauberis* in farmed olive flounder *Paralichthys olivaceus*. *Fish Pathol.*, **46**, 108-111.

Hancock, L. E. and M. S. Gilmore (2002): The capsular polysaccharide of *Enterococcus faecalis* and its relationship to other polysaccharides in the cell wall. *Proc. Natl. Acad. Sci. USA.*, **99**, 1574-1579.

Higgins, R. and M. Gottschalk (1990): An update on *Streptococcus suis* identification. *J. Vet. Diagn. Invest.*, **2**, 249-252.

Kanai, K., M. Notohara, T. Kato, K. Shutou and K. Yoshikoshi (2006): Serological characterization of *Streptococcus iniae* strains isolated from cultured fish in Japan. *Fish Pathol.*, **41**, 57-66.

Kanai, K., M. Yamada, F. Meng, I. Takahashi, T. Nagano, H. Kawakami, A. Yamashita, S. Matsuoka, Y. Fukuda, Y. Miyoshi, I. Takami, H. Nakano, T. Hirae, K. Shutou and T. Honma (2009): Serological differentiation of *Streptococcus parauberis* strains isolated from cultured Japanese flounder in Japan. *Fish Pathol.*, **44**, 33-39. (In Japanese with English abstract)

Kanai, K., C. Tu, N. Katayama and K. Suga (2015): Existence of subserotypes in *Streptococcus parauberis* serotype I. *Fish Pathol.*, **50**, 75-80.

Lakkitjaroen, N., D. Takamatsu, M. Okura, M. Sato, M. Osaki and T. Sekizaki (2014): Capsule loss or death: The position of mutations among capsule genes sways the destiny of *Streptococcus suis*. *FEMS Microbiol. Lett.*, **354**, 46-54.

Liebana, E., D. Guns, L. Garcia-Migura, M. J. Woodward, F. A. CliftonHadley and R. H. Davies (2001): Molecular typing of *Salmonella* serotypes prevalent in animals in England: assessment of methodology. *J. Clin. Microbiol.*, **39**, 3609-3616.

Liu, Z, H. Zheng, M. Gottschalk, X. Bai, R. Lan, S. Ji, H. Liu and J. Xu (2013): Development of multiplex PCR assays for the identification of the 33 serotypes of

Streptococcus suis. *PLoS One*, **8**, e72070.

Mata, A. I., A. Gibello, A. Casamayor, M. M. Blanco, L. Dominguez and J. F. Fernández-Garayzábal (2004): Multiplex PCR assay for detection of bacterial pathogens associated with warm-water streptococcosis in fish. *Appl. Environ. Microbiol.*, **70**, 3183-3187.

Mavroidi, A., D. M. Aanensen, D. Godoy, I. C. Skovsted, M. S. Kaltoft, P. R. Reeves, S. D. Bentley and B. G. Spratt (2007): Genetic relatedness of the *Streptococcus pneumoniae*capsular biosynthetic loci. *J. Bacteriol.*, **189**, 7841-7855.

Meng, F., K. Kanai and K. Yoshikoshi (2009): Characterization of drug resistance in *Streptococcus parauberis* isolated from Japanese flounder. *Fish Pathol.*, **44**, 40-46.

Miller, W. L., C. Q. Wenzel, C. Daniels, S. Larocque, J. R. Brisson and J. S. Lam (2004): Biochemical characterization of WbpA, a UDP-N-acetyl-D-glucosamine 6-dehydrogenase involved in O-antigen biosynthesis in *Pseudomonas aeruginosa* PAO1. *J. Biol. Chem.*, **279**, 37551-37558.

Morita, H., H. Toh, K. Oshima, M. Yoshizaki, M. Kawanishi, K. Nakaya, T. Suzuki, E. Miyauchi, Y. Ishii, S. Tanabe, M. Murakami and M. Hattori (2011): Complete genome sequence and comparative analysis of the fish pathogen *Lactococcus garvieae*. *PLoS One*, **6**, e23184.

Nakhamchik, A., C. Wilde and D. A. Rowe-Magnus (2007): Identification of a Wzy polymerase required for group IV capsular polysaccharide and lipopolysaccharide biosynthesis in *Vibrio vulnificus*. *Infect. Immun.*, **75**, 5550-5558.

Nho, S. W., J. Hikima, S. B. Park, H. B. Jang, I. S. Cha , M. Yasuike , Y. Nakamura, A. Fujiwara, M. Sano, K. Kanai, H. Kondo, I. Hirono, H. Takeyama, T. Aoki and T. S. Jung (2013): Comparative genomic characterization of three *Streptococcus parauberis* strains in fish pathogen, as assessed by wide-genome analyses. *PLoS One*, **8**, e80395.

Ochman, H., A. S. Gerber and D. L. Hartl (1988): Genetic applications of an inverse polymerase chain reaction. *Genetics*, **120**, 621-623.

Okura, M., D. Takamatsu, F. Maruyama, T. Nozawa, I. Nakagawa, M. Osaki, T. Sekizaki, M. Gottschalk, Y. Kumagai and S. Hamada (2013): Genetic analysis of capsular polysaccharide synthesis gene clusters from all serotypes of *Streptococcus suis*: potential mechanisms for generation of capsular variation. *Appl. Environ. Microbiol.*, **79**, 2796-2806.

Patrick, S., G. W. Blakely, S. Houston, J. Moore, V. R. Abratt, M. Bertalan, A. M. Cerdeño-Tárraga, M. A. Quail, N. Corton, C. Corton, A. Bignell, A. Barron, L. Clark, S. D. Bentley and J. Parkhill (2010): Twenty-eight divergent polysaccharide loci specifying within- and amongst-strain capsule diversity in three strains of *Bacteroides fragilis*. *Microbiology*, **156**, 3255-3269.

Roberts, I. S. (1996): The biochemistry and genetics of capsular polysaccharide production in bacteria. *Annu. Rev. Microbiol.,* **50**, 285-315.

Romalde, J. L. and A. E. Toranzo (1999): Streptococcosis of marine fish. In: Olivier, G. (Ed.), ICES Identification Leaflets for Diseases and Parasites of Fish and Shellfish. No. 56. International Council for the Exploration of the Sea. Copenhagen, Denmark, pp. 1 – 8.

Shibata, Y., K. Ozaki, M. Seki, T. Kawato, H. Tanaka, Y. Nakano and Y. Yamashita (2003): Analysis of loci required for determination of serotype antigenicity in *Streptococcus mutans* and its clinical utilization. *J. Clin. Microbiol.*, **41**, 4107-4112.

Shutou, K., K. Kanai and K. Yoshikoshi (2007): Virulence attenuation of capsular polysaccharide-deleted mutants of *Streptococcus iniae* in Japanese flounder *Paralichthys olivaceus*. *Fish Pathol.*, **42**, 41-48.

Smith, H. E., V. Veenbergen, J. van der Velde, M. Damman, H. J. Wisselink and M. A. Smits (1999): The *cps* genes of *Streptococcus suis* serotypes 1, 2, and 9: development of rapid serotype-specific PCR assays. *J. Clin. Microbiol.*, **37**, 3146-3152.

Stull, T. L., J. J. LiPuma and T. D. Edlind (1988): A broad-spectrum probe for molecular epidemiology of bacteria: ribosomal RNA. *J. Infect. Dis.*, **157**, 280-286.

Tenover, F. C. (1985): Plasmid fingerprinting: A tool for bacterial strain identification and surveillance of nosocomial and community-acquired infections. *Clin. Lab. Med.*, **5**, 413-436.

Thurlow, L. R., V. C. Thomas and L. E. Hancock (2009): Capsular polysaccharide production in *Enterococcus faecalis* and contribution of CpsF to capsule serospecificity. *J. Bacteriol.*, **191**, 6203-6210.

Toranzo, A. E., B. Magariños and J. L. Romalde (2005): A review of the main bacterial fish diseases in mariculture systems. *Aquaculture*, **246**, 37-61.

van Belkum, A. L. E. X. (1994): DNA fingerprinting of medically important microorganisms by use of PCR. *Clin. Microbiol. Rev.*, **7**, 174-184.

Wang, K., W. Fan, L. Cai, B. Huang and C. Lu (2011a): Genetic analysis of the capsular polysaccharide synthesis locus in 15 *Streptococcus suis* serotypes. *FEMS Microbiol. Lett.*, **324**, 117-124.

Wang, K., W. Fan, H. Wisselink and C. Lu (2011b): The *cps* locus of *Streptococcus suis* serotype 16: Development of a serotype-specific PCR assay. *Vet. Microbiol.*, **153**, 403-406.

Weiser, J. N., D. Bae, H. Epino, S. B. Gordon, M. Kapoor, L. A. Zenewicz and M. Shchepetov (2001): Changes in availability of oxygen accentuate differences in capsular polysaccharide expression by phenotypic variants and clinical isolates of *Streptococcus pneumoniae*. *Infect. Immun.*, **69**, 5430-5439.

Wessels, M. R. (1997): Biology of streptococcal capsular polysaccharides. *Soc. Appl. Bacteriol. Symp.*, **83**, 20-31.

Yoshida, T., T. Eshima, Y. Wada, Y. Yamada, E. Kakizaki, M. Sakai, T. Kitao and V. Inglis (1996): Phenotypic variation associated with an antiphagocytic factor in the bacterial fish pathogen *Enterococcus seriolicida*. *Dis. Aquat. Org.*, **25**, 81-86.

Yoshida, T., M. Endo, M. Sakai and V. Inglis (1997): A cell capsule with possible involvement in resistance to opsonophagocytosis in *Enterococcus seriolicida* isolated from yellowtail *Seriola quinqueradiata. Dis. Aquat. Org.*, **29**, 233-235.

Yother, J. (2004): Capsules. In "The pneumococcus" (ed. by E. I. Tuomanen, T. J.

Mitchell, D. Morrison and B. G. Spratt). ASM Press, Washington D.C., pp. 30-48.

Yother, J. (2011): Capsules of *Streptococcus pneumoniae* and other bacteria: paradigms for polysaccharide biosynthesis and regulation. *Annu. Rev. Microbiol.*, **65**, 563-581.

Yu, J., M. da G. S. Carvalho, B. Beall and M. H. Nahm (2008): A rapid pneumococcal serotyping system based on monoclonal antibodies and PCR. *J. med. microbiol.*, **57**, 171-178.

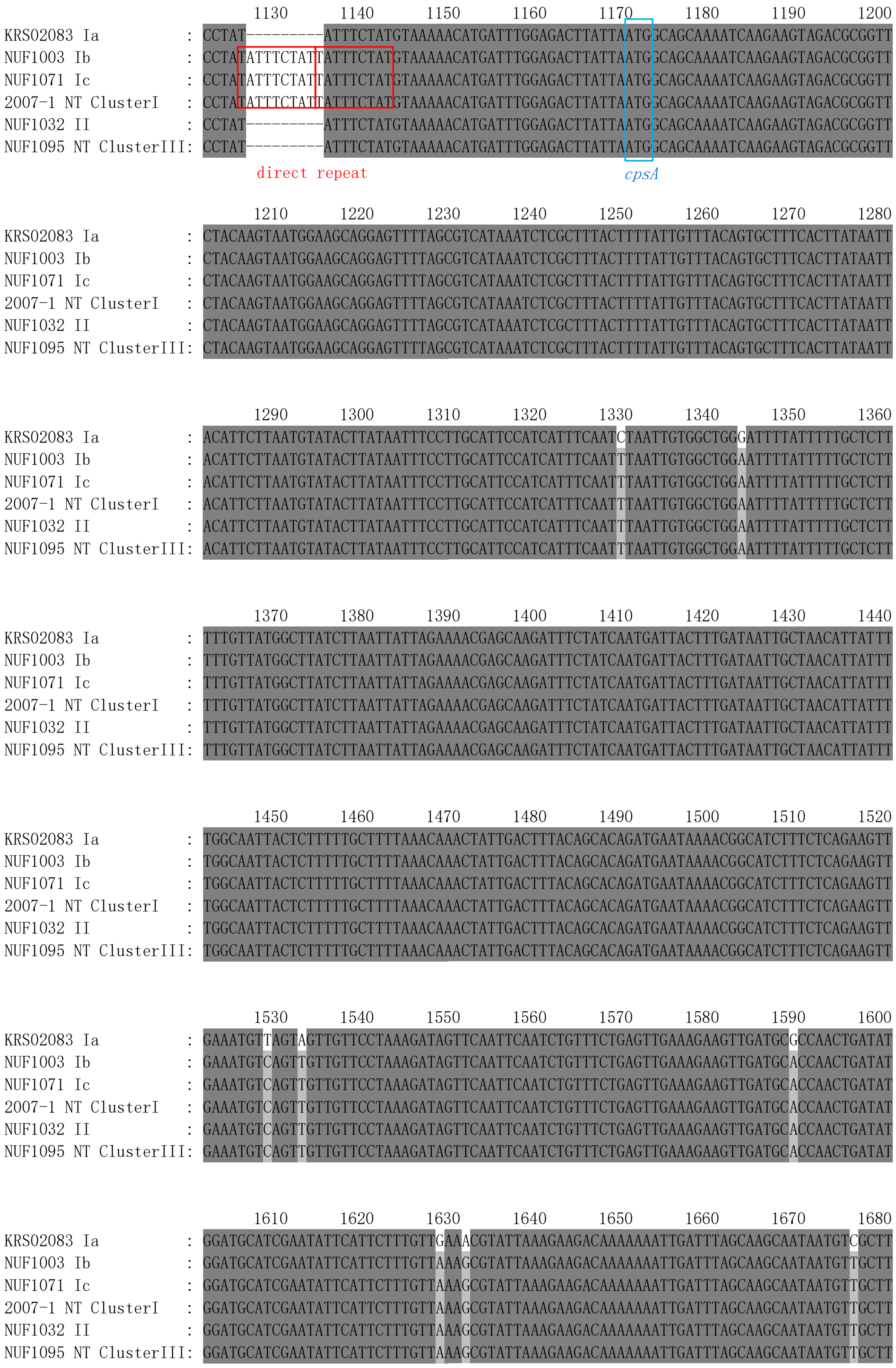
Supplement

Contents

S1.1 Multiple nucleotide sequence alignment of *cps* loci upper conserved region in six representative strains.(*lysR*,*cpsA-E*)

```
                               10        20        30        40        50        60        70        80
KRS02083 Ia          : TTAAAATTGAACCTCTTTGATCAAATAATCAATAAAGGCTTGGCCCATTTTTGATAAATTTGCCTTTTCGTGTGTGATAT
NUF1003 Ib           : TTAAAATTGAACCTCTTTGATCAAATAATCAATAAAGGCTTGGCCCATTTTTGATAAATTTGCCTTTTCGTGTGTGATAT
NUF1071 Ic           : TTAAAATTGAACCTCTTTGATCAAATAATCAATAAAGGCTTGGCCCATTTTTGATAAATTTGCCTTTTCGTGTGTGATAT
2007-1 NT ClusterI   : TTAAAATTGAACCTCTTTGATCAAATAATCAATAAAGGCTTGGCCCATTTTTGATAAATTTGCCTTTTCGTGTGTGATAT
NUF1032 II           : TTAAAATTGAACCTCTTTGATCAAATAATCAATAAAGGCTTGGCCCATTTTTGATAAATTTGCCTTTTCGTGTGTGATAT
NUF1095 NT ClusterIII: TTAAAATTGAACCTCTTTGATCAAATAATCAATAAAGGCTTGGCCCATTTTTGATAAATTTGCCTTTTCGTGTGTGATAT

                               90       100       110       120       130       140       150       160
KRS02083 Ia          : ATACAATATCGATAATGTCAGGAACATCAAGTGGGATTGCAACAATCCCATCTCCGTTTAAATTACTGTTCAGAATGCCA
NUF1003 Ib           : ATACAATATCGATAATGTCAGGAACATCAAGTGGGATTGCAACAATCCCATCTCCGTTTAAATTACTGTTCAGAATGCCA
NUF1071 Ic           : ATACAATATCGATAATGTCAGGAACATCAAGTGGGATTGCAACAATCCCATCTCCGTTTAAATTACTGTTCAGAATGCCA
2007-1 NT ClusterI   : ATACAATATCGATAATGTCAGGAACATCAAGTGGGATTGCAACAATCCCATCTCCGTTTAAATTACTGTTCAGAATGCCA
NUF1032 II           : ATACAATATCGATAATGTCAGGAACATCAAGTGGGATTGCAACAATCCCATCTCCGTTTAAATTACTGTTCAGAATACCA
NUF1095 NT ClusterIII: ATACAATATCGATAATGTCAGGAACATCAAGTGGGATTGCAACAATCCCATCTCCGTTTAAATTACTGTTCAGAATACCA

                              170       180       190       200       210       220       230       240
KRS02083 Ia          : CTGGCGATGGTATAACCATCTAAACCAATCATAAGATTAAACAAGGTTGCTCGGTCACTGACCACGATGGATTTGGGGTG
NUF1003 Ib           : CTGGCGATGGTATAACCATCTAAACCAATCATAAGATTAAACAAGGTTGCTCGGTCACTGACCACGATGGATTTGGGGTG
NUF1071 Ic           : CTGGCGATGGTATAACCATCTAAACCAATCATAAGATTAAACAAGGTTGCTCGGTCACTGACCACGATGGATTTGGGGTG
2007-1 NT ClusterI   : CTGGCGATGGTATAACCATCTAAACCAATCATAAGATTAAACAAGGTTGCTCGGTCACTGACCACGATGGATTTGGGGTG
NUF1032 II           : CTGGCGATGGTATAACCATCTAAACCAATCATAAGATTAAACAAGGTTGCTCGGTCACTGACCACGATGGATTTGGGGTG
NUF1095 NT ClusterIII: CTGGCGATGGTATAACCATCTAAACCAATCATAAGATTAAACAAGGTTGCTCGGTCACTGACCACGATGGATTTGGGGTG

                              250       260       270       280       290       300       310       320
KRS02083 Ia          : ACTGATTTCTGACAACATTTCTTCCGAAAAGTAGAAGGAGTTGTGTAAGCCTTGATCATAACTGAGATAAGGGTATGGTT
NUF1003 Ib           : ACTGATTTCTGACAACATTTCTTCCGAAAAGTAGAAGGAGTTGTGTAAGCCTTGATCATAACTGAGATAAGGGTATGGTT
NUF1071 Ic           : ACTGATTTCTGACAACATTTCTTCCGAAAAGTAGAAGGAGTTGTGTAAGCCTTGATCATAACTGAGATAAGGGTATGGTT
2007-1 NT ClusterI   : ACTGATTTCTGACAACATTTCTTCCGAAAAGTAGAAGGAGTTGTGTAAGCCTTGATCATAACTGAGATAAGGGTATGGTT
NUF1032 II           : ACTGATTTCTGACAACATTTCTTCCGAAAAGTAGAAGGAGTTGTGTAAGCCTTGATCATAACTGAGATAAGGGTATGGTT
NUF1095 NT ClusterIII: ACTGATTTCTGACAACATTTCTTCCGAAAAGTAGAAGGAGTTGTGTAAGCCTTGATCATAACTGAGATAAGGGTATGGTT

                              330       340       350       360       370       380       390       400
KRS02083 Ia          : TTAAATCATTGAAGGTTAAACGTGTATGCTTGGCAAGCGGATGATTTCGACTGATAAAAATGTGAGGGTGAGCCTTGAAG
NUF1003 Ib           : TTAAATCATTGAAGGTTAAACGTGTATGCTTGGCAAGCGGATGATTTCGACTGATAAAAATGTGAGGGTGAGCCTTGAAG
NUF1071 Ic           : TTAAATCATTGAAGGTTAAACGTGTATGCTTGGCAAGCGGATGATTTCGACTGATAAAAATGTGAGGGTGAGCCTTGAAG
2007-1 NT ClusterI   : TTAAATCATTGAAGGTTAAACGTGTATGCTTGGCAAGCGGATGATTTCGACTGATAAAAATGTGAGGGTGAGCCTTGAAG
NUF1032 II           : TTAAATCATTGAAGGTTAAACGTGTATGCTTGGCAAGCGGATGATTTCGGCTGATAAAAATGTGAGGGTGAGCCTTGAAG
NUF1095 NT ClusterIII: TTAAATCATTGAAGGTTAAACGTGTATGCTTGGCAAGCGGATGATTTCGGCTGATAAAAATGTGAGGGTGAGCCTTGAAG

                              410       420       430       440       450       460       470       480
KRS02083 Ia          : AGTGGTGTCGCAATTAAGTGGTCATCCTCCATCATCTTGCTCAGGACATCACGGTTATAGTCGTTTAAAAATAAGACACC
NUF1003 Ib           : AGTGGTGTCGCAATTAAGTGGTCATCCTCCATCATCTTGCTCAGGACATCACGGTTATAGTCGTTTAAAAATAAGACACC
NUF1071 Ic           : AGTGGTGTCGCAATTAAGTGGTCATCCTCCATCATCTTGCTCAGGACATCACGGTTATAGTCGTTTAAAAATAAGACACC
2007-1 NT ClusterI   : AGTGGTGTCGCAATTAAGTGGTCATCCTCCATCATCTTGCTCAGGACATCACGGTTATAGTCGTTTAAAAATAAGACACC
NUF1032 II           : AGTGGTGTCGCAATTAAGTGGTCATCCTCCATCATCTTGCTCAGGACATCACGGTTATAGTCGTTTAAAAATAAGACACC
NUF1095 NT ClusterIII: AGTGGTGTCGCAATTAAGTGGTCATCCTCCATCATCTTGCTCAGGACATCACGGTTATAGTCGTTTAAAAATAAGACACC

                              490       500       510       520       530       540       550       560
KRS02083 Ia          : AATTTCAGATCGGAAGTTTTTAACATCATCTAAGATTTCCCAAGTTCTTGTTTCACGAAGGAAAAGTTCATACTGGGTCA
NUF1003 Ib           : AATTTCAGATCGGAAGTTTTTAACATCATCTAAGATTTCCCAAGTTCTTGTTTCACGAAGGAAAAGTTCATACTGGGTCA
NUF1071 Ic           : AATTTCAGATCGGAAGTTTTTAACATCATCTAAGATTTCCCAAGTTCTTGTTTCACGAAGGAAAAGTTCATACTGGGTCA
2007-1 NT ClusterI   : AATTTCAGATCGGAAGTTTTTAACATCATCTAAGATTTCCCAAGTTCTTGTTTCACGAAGGAAAAGTTCATACTGGGTCA
NUF1032 II           : AATTTCAGATCGGAAGTTTTTAACATCATCTAAGATTTCCCAAGTTCTTGTTTCACGAAGGAAAAGTTCATACTGGGTCA
NUF1095 NT ClusterIII: AATTTCAGATCGGAAGTTTTTAACATCATCTAAGATTTCCCAAGTTCTTGTTTCACGAAGGAAAAGTTCATACTGGGTCA
```

```
                              570       580       590       600       610       620       630       640
KRS02083 Ia          : TATCGGTTTCTTTTAATAGTGAGACAAAAGCATTAACAACAAAGGCGTAGTGCTGTGAAGACACACTAAAAAGTTCTCGT
NUF1003 Ib           : TATCGGTTTCTTTTAATAGTGAGACAAAAGCATTAACAACAAAGGCGTAGTGCTGTGAAGACACACTAAAAAGTTCTCGT
NUF1071 Ic           : TATCGGTTTCTTTTAATAGTGAGACAAAAGCATTAACAACAAAGGCGTAGTGCTGTGAAGACACACTAAAAAGTTCTCGT
2007-1 NT ClusterI   : TATCGGTTTCTTTTAATAGTGAGACAAAAGCATTAACAACAAAGGCGTAGTGCTGTGAAGACACACTAAAAAGTTCTCGT
NUF1032 II           : TATCGGTTTCTTTTAATAGTGAGACAAAAGCATTAACAACAAAGGCGTAGTGCTGTGAAGACACACTAAAAAGTTCTCGT
NUF1095 NT ClusterIII: TATCGGTTTCTTTTAATAGTGAGACAAAAGCATTAACAACAAAGGCGTAGTGCTGTGAAGACACACTAAAAAGTTCTCGT

                              650       660       670       680       690       700       710       720
KRS02083 Ia          : TTAGAATTGAGATTTTTATATCTGTCTTCTAAGAGGGATGTTTGCTCAACAACTTGGCGTGCATAAGATAAAAATTCGAC
NUF1003 Ib           : TTAGAATTGAGATTTTTATATCTGTCTTCTAAGAGGGATGTTTGCTCAACAACTTGGCGTGCATAAGATAAAAATTCGAC
NUF1071 Ic           : TTAGAATTGAGATTTTTATATCTGTCTTCTAAGAGGGATGTTTGCTCAACAACTTGGCGTGCATAAGATAAAAATTCGAC
2007-1 NT ClusterI   : TTAGAATTGAGATTTTTATATCTGTCTTCTAAGAGGGATGTTTGCTCAACAACTTGGCGTGCATAAGATAAAAATTCGAC
NUF1032 II           : TTAGAATTGAGATTTTTATATCTGTCTTCTAAGAGGGATGTTTGCTCAACAACTTGGCGTGCATAAGATAAAAATTCGAC
NUF1095 NT ClusterIII: TTAGAATTGAGATTTTTATATCTGTCTTCTAAGAGGGATGTTTGCTCAACAACTTGGCGTGCATAAGATAAAAATTCGAC

                              730       740       750       760       770       780       790       800
KRS02083 Ia          : ACCATCCTTTGTTAGGGTGATACCCTTTGGATTTCTATTAAAGATAGTAATTCCCATCTCGTTTTCTAAATCCTTAACAG
NUF1003 Ib           : ACCATCCTTTGTTAGGGTGATACCCTTTGGATTTCTATTAAAGATAGTAATTCCCATCTCGTTTTCTAAATCCTTAACAG
NUF1071 Ic           : ACCATCCTTTGTTAGGGTGATACCCTTTGGATTTCTATTAAAGATAGTAATTCCCATCTCGTTTTCTAAATCCTTAACAG
2007-1 NT ClusterI   : ACCATCCTTTGTTAGGGTGATACCCTTTGGATTTCTATTAAAGATAGTAATTCCCATCTCGTTTTCTAAATCCTTAACAG
NUF1032 II           : ACCATCCTTTGTTAGGGTGATACCCTTTGGATTTCTATTAAAGATAGTAATTCCCATCTCGTTTTCTAAATCCTTAACAG
NUF1095 NT ClusterIII: ACCATCCTTTGTTAGGGTGATACCCTTTGGATTTCTATTAAAGATAGTAATTCCCATCTCGTTTTCTAAATCCTTAACAG

                              810       820       830       840       850       860       870       880
KRS02083 Ia          : CATTTGAGAGGCTAGGTTGAGTAATGAAAAGTTGTTTTGCTGCGACGTTCATCGAGCCACATTCAACAACTTTTATAATG
NUF1003 Ib           : CATTTGAGAGGCTAGGTTGAGTAATGAAAAGTTGTTTTGCTGCGACGTTCATCGAGCCACATTCAACAACTTTTATAATG
NUF1071 Ic           : CATTTGAGAGGCTAGGTTGAGTAATGAAAAGTTGTTTTGCTGCGACGTTCATCGAGCCACATTCAACAACTTTTATAATG
2007-1 NT ClusterI   : CATTTGAGAGGCTAGGTTGAGTAATGAAAAGTTGTTTTGCTGCGACGTTCATCGAGCCACATTCAACAACTTTTATAATG
NUF1032 II           : CATTTGAGAGGCTAGGTTGAGTAATGAAAAGTTGTTTTGCTGCGACGTTCATCGAGCCACATTCAACAACTTTTATAATG
NUF1095 NT ClusterIII: CATTTGAGAGGCTAGGTTGAGTAATGAAAAGTTGTTTTGCTGCGACGTTCATCGAGCCACATTCAACAACTTTTATAATG

                                          lysR
                              890       900       910       920       930       940       950       960
KRS02083 Ia          : TAATGTAATTGTTGTAATCTCATACTTTCAGTCTAACTTAATATGATTATTCTTTCAAGATGAAAAGTAGATTTACAAAA
NUF1003 Ib           : TAATGTAATTGTTGTAATCTCATACTTTCAGTCTAACTTAATATGATTATTCTTTCAAGATGAAAAGTAGATTTACAAAA
NUF1071 Ic           : TAATGTAATTGTTGTAATCTCATACTTTCAGTCTAACTTAATATGATTATTCTTTCAAGATGAAAAGTAGATTTACAAAA
2007-1 NT ClusterI   : TAATGTAATTGTTGTAATCTCATACTTTCAGTCTAACTTAATATGATTATTCTTTCAAGATGAAAAGTAGATTTACAAAA
NUF1032 II           : TAATGTAATTGTTGTAATCTCATACTTTCAGTCTAACTTAATATGATTATTCTTTCAAGATGAAAAGTAGATTTACAAAA
NUF1095 NT ClusterIII: TAATGTAATTGTTGTAATCTCATACTTTCAGTCTAACTTAATATGATTATTCTTTCAAGATGAAAAGTAGATTTACAAAA

                              970       980       990      1000      1010      1020      1030      1040
KRS02083 Ia          : TTAAGTTTTTGAAAAATTATAAAAAGTTAGATGGGCTCAAAATAGACTTAATTAAAACAATTCATTAGCAAAGGAAACAC
NUF1003 Ib           : TTAAGTTTTTGAAAAATTATAAAAAGTTAGATGGGCTCAAAATAGACTTAATTAAAACAATTCATTAGCAAAGGAAACAC
NUF1071 Ic           : TTAAGTTTTTGAAAAATTATAAAAAGTTAGATGGGCTCAAAATAGACTTAATTAAAACAATTCATTAGCAAAGGAAACAC
2007-1 NT ClusterI   : TTAAGTTTTTGAAAAATTATAAAAAGTTAGATGGGCTCAAAATAGACTTAATTAAAACAATTCATTAGCAAAGGAAACAC
NUF1032 II           : TTAAGTTTTTGAAAAATTATAAAAAGTTAGATGGGCTCAAAATAGACTTAATTAAAACAATTCATTAGCAAAGGAAACAC
NUF1095 NT ClusterIII: TTAAGTTTTTGAAAAATTATAAAAAGTTAGATGGGCTCAAAATAGACTTAATTAAAACAATTCATTAGCAAAGGAAACAC

                             1050      1060      1070      1080      1090      1100      1110      1120
KRS02083 Ia          : TTACATAGCTAAAAAGCGGACTGTTAACATAATCTTAATATTCAAAAATTGCAAAAAATGTGAAAATTTGGTATAATTTT
NUF1003 Ib           : TTACATAGCTAAAAAGCGGACTGTTAACATTATCTTAATATTCAAAAATTGCAAAAAATGTGAAAATTTGGTATAATTTT
NUF1071 Ic           : TTACATAGCTAAAAAGCGGACTGTTAACATTATCTTAATATTCAAAAATTGCAAAAAATGTGAAAATTTGGTATAATTTT
2007-1 NT ClusterI   : TTACATAGCTAAAAAGCGGACTGTTAACATTATCTTAATATTCAAAAATTGCAAAAAATGTGAAAATTTGGTATAATTTT
NUF1032 II           : TTACATAGCTAAAAAGCGGACTGTTAACATAATCTTAATATTCAAAAATTGCAAAAAATGTGAAAATTTGGTATAATTTT
NUF1095 NT ClusterIII: TTACATAGCTAAAAAGCGGACTGTTAACATAATCTTAATATTCAAAAATTGCAAAAAATGTGAAAATTTGGTATAATTTT
```



```
                             1130      1140      1150      1160      1170      1180      1190      1200
KRS02083 Ia          : CCTAT---------ATTTCTATGTAAAAACATGATTTGGAGACTTATTAATGGCAGCAAAATCAAGAAGTAGACGCGGTT
NUF1003 Ib           : CCTATATTTCTATTATTTCTATGTAAAAACATGATTTGGAGACTTATTAATGGCAGCAAAATCAAGAAGTAGACGCGGTT
NUF1071 Ic           : CCTATATTTCTATTATTTCTATGTAAAAACATGATTTGGAGACTTATTAATGGCAGCAAAATCAAGAAGTAGACGCGGTT
2007-1 NT ClusterI   : CCTATATTTCTATTATTTCTATGTAAAAACATGATTTGGAGACTTATTAATGGCAGCAAAATCAAGAAGTAGACGCGGTT
NUF1032 II           : CCTAT---------ATTTCTATGTAAAAACATGATTTGGAGACTTATTAATGGCAGCAAAATCAAGAAGTAGACGCGGTT
NUF1095 NT ClusterIII: CCTAT---------ATTTCTATGTAAAAACATGATTTGGAGACTTATTAATGGCAGCAAAATCAAGAAGTAGACGCGGTT
                            direct repeat                             cpsA

                             1210      1220      1230      1240      1250      1260      1270      1280
KRS02083 Ia          : CTACAAGTAATGGAAGCAGGAGTTTTAGCGTCATAAATCTCGCTTTACTTTTATTGTTTACAGTGCTTTCACTTATAATT
NUF1003 Ib           : CTACAAGTAATGGAAGCAGGAGTTTTAGCGTCATAAATCTCGCTTTACTTTTATTGTTTACAGTGCTTTCACTTATAATT
NUF1071 Ic           : CTACAAGTAATGGAAGCAGGAGTTTTAGCGTCATAAATCTCGCTTTACTTTTATTGTTTACAGTGCTTTCACTTATAATT
2007-1 NT ClusterI   : CTACAAGTAATGGAAGCAGGAGTTTTAGCGTCATAAATCTCGCTTTACTTTTATTGTTTACAGTGCTTTCACTTATAATT
NUF1032 II           : CTACAAGTAATGGAAGCAGGAGTTTTAGCGTCATAAATCTCGCTTTACTTTTATTGTTTACAGTGCTTTCACTTATAATT
NUF1095 NT ClusterIII: CTACAAGTAATGGAAGCAGGAGTTTTAGCGTCATAAATCTCGCTTTACTTTTATTGTTTACAGTGCTTTCACTTATAATT

                             1290      1300      1310      1320      1330      1340      1350      1360
KRS02083 Ia          : ACATTCTTAATGTATACTTATAATTTCCTTGCATTCCATCATTTCAATCTAATTGTGGCTGGGATTTTATTTTTGCTCTT
NUF1003 Ib           : ACATTCTTAATGTATACTTATAATTTCCTTGCATTCCATCATTTCAATTTAATTGTGGCTGGAATTTTATTTTTGCTCTT
NUF1071 Ic           : ACATTCTTAATGTATACTTATAATTTCCTTGCATTCCATCATTTCAATTTAATTGTGGCTGGAATTTTATTTTTGCTCTT
2007-1 NT ClusterI   : ACATTCTTAATGTATACTTATAATTTCCTTGCATTCCATCATTTCAATTTAATTGTGGCTGGAATTTTATTTTTGCTCTT
NUF1032 II           : ACATTCTTAATGTATACTTATAATTTCCTTGCATTCCATCATTTCAATTTAATTGTGGCTGGAATTTTATTTTTGCTCTT
NUF1095 NT ClusterIII: ACATTCTTAATGTATACTTATAATTTCCTTGCATTCCATCATTTCAATTTAATTGTGGCTGGAATTTTATTTTTGCTCTT

                             1370      1380      1390      1400      1410      1420      1430      1440
KRS02083 Ia          : TTTGTTATGGCTTATCTTAATTATTAGAAAACGAGCAAGATTTCTATCAATGATTACTTTGATAATTGCTAACATTATTT
NUF1003 Ib           : TTTGTTATGGCTTATCTTAATTATTAGAAAACGAGCAAGATTTCTATCAATGATTACTTTGATAATTGCTAACATTATTT
NUF1071 Ic           : TTTGTTATGGCTTATCTTAATTATTAGAAAACGAGCAAGATTTCTATCAATGATTACTTTGATAATTGCTAACATTATTT
2007-1 NT ClusterI   : TTTGTTATGGCTTATCTTAATTATTAGAAAACGAGCAAGATTTCTATCAATGATTACTTTGATAATTGCTAACATTATTT
NUF1032 II           : TTTGTTATGGCTTATCTTAATTATTAGAAAACGAGCAAGATTTCTATCAATGATTACTTTGATAATTGCTAACATTATTT
NUF1095 NT ClusterIII: TTTGTTATGGCTTATCTTAATTATTAGAAAACGAGCAAGATTTCTATCAATGATTACTTTGATAATTGCTAACATTATTT

                             1450      1460      1470      1480      1490      1500      1510      1520
KRS02083 Ia          : TGGCAATTACTCTTTTTGCTTTTAAACAAACTATTGACTTTACAGCACAGATGAATAAAACGGCATCTTTCTCAGAAGTT
NUF1003 Ib           : TGGCAATTACTCTTTTTGCTTTTAAACAAACTATTGACTTTACAGCACAGATGAATAAAACGGCATCTTTCTCAGAAGTT
NUF1071 Ic           : TGGCAATTACTCTTTTTGCTTTTAAACAAACTATTGACTTTACAGCACAGATGAATAAAACGGCATCTTTCTCAGAAGTT
2007-1 NT ClusterI   : TGGCAATTACTCTTTTTGCTTTTAAACAAACTATTGACTTTACAGCACAGATGAATAAAACGGCATCTTTCTCAGAAGTT
NUF1032 II           : TGGCAATTACTCTTTTTGCTTTTAAACAAACTATTGACTTTACAGCACAGATGAATAAAACGGCATCTTTCTCAGAAGTT
NUF1095 NT ClusterIII: TGGCAATTACTCTTTTTGCTTTTAAACAAACTATTGACTTTACAGCACAGATGAATAAAACGGCATCTTTCTCAGAAGTT

                             1530      1540      1550      1560      1570      1580      1590      1600
KRS02083 Ia          : GAAATGTTAGTAGTTGTTCCTAAAGATAGTTCAATTCAATCTGTTTCTGAGTTGAAAGAAGTTGATGCGCCAACTGATAT
NUF1003 Ib           : GAAATGTCAGTTGTTGTTCCTAAAGATAGTTCAATTCAATCTGTTTCTGAGTTGAAAGAAGTTGATGCACCAACTGATAT
NUF1071 Ic           : GAAATGTCAGTTGTTGTTCCTAAAGATAGTTCAATTCAATCTGTTTCTGAGTTGAAAGAAGTTGATGCACCAACTGATAT
2007-1 NT ClusterI   : GAAATGTCAGTTGTTGTTCCTAAAGATAGTTCAATTCAATCTGTTTCTGAGTTGAAAGAAGTTGATGCACCAACTGATAT
NUF1032 II           : GAAATGTCAGTTGTTGTTCCTAAAGATAGTTCAATTCAATCTGTTTCTGAGTTGAAAGAAGTTGATGCACCAACTGATAT
NUF1095 NT ClusterIII: GAAATGTCAGTTGTTGTTCCTAAAGATAGTTCAATTCAATCTGTTTCTGAGTTGAAAGAAGTTGATGCACCAACTGATAT

                             1610      1620      1630      1640      1650      1660      1670      1680
KRS02083 Ia          : GGATGCATCGAATATTCATTCTTTGTTGAAACGTATTAAAGAAGACAAAAAAATTGATTTAGCAAGCAATAATGTCGCTT
NUF1003 Ib           : GGATGCATCGAATATTCATTCTTTGTTAAAGCGTATTAAAGAAGACAAAAAAATTGATTTAGCAAGCAATAATGTTGCTT
NUF1071 Ic           : GGATGCATCGAATATTCATTCTTTGTTAAAGCGTATTAAAGAAGACAAAAAAATTGATTTAGCAAGCAATAATGTTGCTT
2007-1 NT ClusterI   : GGATGCATCGAATATTCATTCTTTGTTAAAGCGTATTAAAGAAGACAAAAAAATTGATTTAGCAAGCAATAATGTTGCTT
NUF1032 II           : GGATGCATCGAATATTCATTCTTTGTTAAAGCGTATTAAAGAAGACAAAAAAATTGATTTAGCAAGCAATAATGTTGCTT
NUF1095 NT ClusterIII: GGATGCATCGAATATTCATTCTTTGTTAAAGCGTATTAAAGAAGACAAAAAAATTGATTTAGCAAGCAATAATGTTGCTT
```

```
                             1690      1700      1710      1720      1730      1740      1750      1760
KRS02083 Ia          : CATATAAGGAAGCTTATGACAAAATGATGATTGGTTCTTCCAAAGCGATGGTTTTAAATAGTGCCTATGGTTCATTGATT
NUF1003 Ib           : CATATAAGGAAGCTTATGACAAAATGATGATTGGTTCTTCAAAGGCGATGGTCTTAAATAGTGCCTATGGTTCATTGATT
NUF1071 Ic           : CATATAAGGAAGCTTATGACAAAATGATGATTGGTTCTTCAAAGGCGATGGTCTTAAATAGTGCCTATGGTTCATTGATT
2007-1 NT ClusterI   : CATATAAGGAAGCTTATGACAAAATGATGATTGGTTCTTCAAAGGCGATGGTCTTAAATAGTGCCTATGGTTCATTGATT
NUF1032 II           : CATATAAGGAAGCTTATGACAAAATGATGATTGGTTCTTCAAAGGCGATGGTTTTAAATAGTGCCTATGGTTCATTGATT
NUF1095 NT ClusterIII: CATATAAGGAAGCTTATGACAAAATGATGATTGGTTCTTCAAAGGCGATGGTTTTAAATAGTGCCTATGGTTCATTGATT

                             1770      1780      1790      1800      1810      1820      1830      1840
KRS02083 Ia          : GAGCAATCAGATGCCAACTATAGGGATAAGGTTAAGACTATTTACACTTACAAAGTCAAAAAAGCAATCAAATCGTCAAA
NUF1003 Ib           : GAGCAATCAGATGCCAACTATAGGGATAAGGTTAAGACTATTTACACTTACAAAGTCAAAAAAGCAATCAAATCGTCAAA
NUF1071 Ic           : GAGCAATCAGATGCCAACTATAGGGATAAGGTTAAGACTATTTACACTTACAAAGTCAAAAAAGCAATCAAATCGTCAAA
2007-1 NT ClusterI   : GAGCAATCAGATGCCAACTATAGGGATAAGGTTAAGACTATTTACACTTACAAAGTCAAAAAAGCAATCAAATCGTCAAA
NUF1032 II           : GAGCAATCAGATGCCAACTATAGGGATAAGGTTAAGACTATTTACACTTACAAAGTCAAAAAAGCAATCAAATCGTCAAA
NUF1095 NT ClusterIII: GAGCAATCAGATGCCAACTATAGGGATAAGGTTAAGACTATTTACACTTACAAAGTCAAAAAAGCAATCAAATCGTCAAA

                             1850      1860      1870      1880      1890      1900      1910      1920
KRS02083 Ia          : TAAACTGTCATCAAACAAGGATTCCTTTAATGTCTATATCAGTGGTATTGATACTTATGGCGCAATTTCTACAGTTTCAC
NUF1003 Ib           : TAAACTGTCATCAAACAAGGATTCCTTTAATGTCTATATCAGTGGTATTGATACTTATGGCGCAATTTCTACAGTTTCAC
NUF1071 Ic           : TAAACTGTCATCAAACAAGGATTCCTTTAATGTCTATATCAGTGGTATTGATACTTATGGCGCAATTTCTACAGTTTCAC
2007-1 NT ClusterI   : TAAACTGTCATCAAACAAGGATTCCTTTAATGTCTATATCAGTGGTATTGATACTTATGGCGCAATTTCTACAGTTTCAC
NUF1032 II           : TAAACTGTCATCAAACAAGGATTCCTTTAATGTCTATATCAGTGGTATTGATACTTATGGCGCAATTTCTACAGTTTCAC
NUF1095 NT ClusterIII: TAAACTGTCATCAAACAAGGATTCCTTTAATGTCTATATCAGTGGTATTGATACTTATGGCGCAATTTCTACAGTTTCAC

                             1930      1940      1950      1960      1970      1980      1990      2000
KRS02083 Ia          : GTTCAGACGTGAACATCATTTTGACTGCCAATACTAAGACCCACAAAGTTCTGTTAACAACAACACCTCGTGATTCTTAT
NUF1003 Ib           : GTTCAGACGTGAACATCATTTTGACTGCCAATACTAAGACCCACAAAGTTCTGTTAACAACAACACCTCGTGATTCTTAT
NUF1071 Ic           : GTTCAGACGTGAACATCATTTTGACTGCCAATACTAAGACCCACAAAGTTCTGTTAACAACAACACCTCGTGATTCTTAT
2007-1 NT ClusterI   : GTTCAGACGTGAACATCATTTTGACTGCCAATACTAAGACCCACAAAGTTCTGTTAACAACAACACCTCGTGATTCTTAT
NUF1032 II           : GTTCAGACGTGAACATCATTTTGACTGCCAATACTAAGACCCACAAAGTTCTGTTAACAACAACACCTCGTGATTCTTAT
NUF1095 NT ClusterIII: GTTCAGACGTGAACATCATTTTGACTGCCAATACTAAGACCCACAAAGTTCTGTTAACAACAACACCTCGTGATTCTTAT

                             2010      2020      2030      2040      2050      2060      2070      2080
KRS02083 Ia          : GTGAAAATTCCTGGCGGGGGAGGAGATCAGTACGATAAGTTAACTCATGCTGGAATCTATGGTGTAGAGACTTCTATGGC
NUF1003 Ib           : GTGAAAATTCCTGGCGGCGGAGGAGATCAGTACGATAAGTTAACTCATGCTGGAATCTATGGTGTAGAGACTTCTATGGC
NUF1071 Ic           : GTGAAAATTCCTGGCGGCGGAGGAGATCAGTACGATAAGTTAACTCATGCTGGAATCTATGGTGTAGAGACTTCTATGGC
2007-1 NT ClusterI   : GTGAAAATTCCTGGCGGCGGAGGAGATCAGTACGATAAGTTAACTCATGCTGGAATCTATGGTGTAGAGACTTCTATGGC
NUF1032 II           : GTGAAAATTCCTGGCGGCGGAGGAGATCAGTACGATAAGTTAACTCATGCTGGAATCTATGGTGTAGAGACTTCTATGGC
NUF1095 NT ClusterIII: GTGAAAATTCCTGGCGGCGGAGGAGATCAGTACGATAAGTTAACTCATGCTGGAATCTATGGTGTAGAGACTTCTATGGC

                             2090      2100      2110      2120      2130      2140      2150      2160
KRS02083 Ia          : AACTTTAGAAAGTCTATATGACATCAAGATTAATAACTATGCAAGAATTAATTTCTCGACCTTCATGGATCTAATTGATT
NUF1003 Ib           : AACTCTAGAAAGTCTATATGACATCAATATTAATAACTATGCAAGAATTAATTTCTCGACCTTCATGGATCTAATTGATT
NUF1071 Ic           : AACTCTAGAAAGTCTATATGACATCAATATTAATAACTATGCAAGAATTAATTTCTCGACCTTCATGGATCTAATTGATT
2007-1 NT ClusterI   : AACTCTAGAAAGTCTATATGACATCAATATTAATAACTATGCAAGAATTAATTTCTCGACCTTCATGGATCTAATTGATT
NUF1032 II           : AACTCTAGAAAGTCTATATGACATCAATATTAATAACTATGCAAGAATTAATTTCTCGACCTTCATGGATCTAATTGATT
NUF1095 NT ClusterIII: AACTCTAGAAAGTCTATATGACATCAATATTAATAACTATGCAAGAATTAATTTCTCGACCTTCATGGATCTAATTGATT

                             2170      2180      2190      2200      2210      2220      2230      2240
KRS02083 Ia          : TATTAGGTGGCATTGAAGTTAATAATGACCAAGCTTTTAGTGCAAATGGTTATGACTTCCCACAAGGTCGGATTGCTCTG
NUF1003 Ib           : TATTAGGTGGCATTGAAGTTAATAATGATCAAGCTTTTAGTGCAAATGGTTATGACTTCCCACAAGGTCGGATTGCTCTG
NUF1071 Ic           : TATTAGGTGGCATTGAAGTTAATAATGATCAAGCTTTTAGTGCAAATGGTTATGACTTCCCACAAGGTCGGATTGCTCTG
2007-1 NT ClusterI   : TATTAGGTGGCATTGAAGTTAATAATGATCAAGCTTTTAGTGCAAATGGTTATGACTTCCCACAAGGTCGGATTGCTCTG
NUF1032 II           : TATTAGGTGGCATTGAAGTTAATAATGATCAAGCTTTTAGTGCAAATGGTTATGACTTCCCACAAGGTCGGATTGCTCTG
NUF1095 NT ClusterIII: TATTAGGTGGCATTGAAGTTAATAATGATCAAGCTTTTAGTGCAAATGGTTATGACTTCCCACAAGGTCGGATTGCTCTG
```

```
                              2250      2260      2270      2280      2290      2300      2310      2320
KRS02083 Ia          : AATTCAAAACAAGCTCTAACATTTGTTCGCGAACGTCATTCACTACAAGGTGGAGATAACGATCGTGGTAAAAATCAAGA
NUF1003 Ib           : AATTCAAAACAAGCTCTAACATTTGTTCGCGAACGTCATTCACTACAAGGTGGAGATAACGATCGTGGTAAAAATCAAGA
NUF1071 Ic           : AATTCAAAACAAGCTCTAACATTTGTTCGCGAACGTCATTCACTACAAGGTGGAGATAACGATCGTGGTAAAAATCAAGA
2007-1 NT ClusterI   : AATTCAAAACAAGCTCTAACATTTGTTCGCGAACGTCATTCACTACAAGGTGGAGATAACGATCGTGGTAAAAATCAAGA
NUF1032 II           : AATTCAAAACAAGCTCTAACATTTGTTCGCGAACGTCATTCACTACAAGGTGGAGATAACGATCGTGGTAAAAATCAAGA
NUF1095 NT ClusterIII: AATTCAAAACAAGCTCTAACATTTGTTCGCGAACGTCATTCACTACAAGGTGGAGATAACGATCGTGGTAAAAATCAAGA

                              2330      2340      2350      2360      2370      2380      2390      2400
KRS02083 Ia          : GAAGGTCATCTCGGCAGTTATTCAAAAATTATCAACAATTAAATCACCAGCACAATTTACATCAATTGTAACGGGATTAC
NUF1003 Ib           : GAAGGTCATCTCGGCAGTTATTCAAAAATTATCAACAATTAAATCACCAGCTCAATTTACATCAATTGTAACGGGATTAC
NUF1071 Ic           : GAAGGTCATCTCGGCAGTTATTCAAAAATTATCAACAATTAAATCACCAGCTCAATTTACATCAATTGTAACGGGATTAC
2007-1 NT ClusterI   : GAAGGTCATCTCGGCAGTTATTCAAAAATTATCAACAATTAAATCACCAGCTCAATTTACATCAATTGTAACGGGATTAC
NUF1032 II           : GAAGGTCATCTCGGCAGTTATTCAAAAATTATCAACAATTAAATCACCAGCTCAATTTACATCAATTGTAACGGGATTAC
NUF1095 NT ClusterIII: GAAGGTCATCTCGGCAGTTATTCAAAAATTATCAACAATTAAATCACCAGCTCAATTTACATCAATTGTAACGGGATTAC

                              2410      2420      2430      2440      2450      2460      2470      2480
KRS02083 Ia          : AAAATTCAATTCAAACTAATTTATCATTAAATCAATTGATGACATTAGCAAATTCACAGGTAGCTGATAAATCATCATAT
NUF1003 Ib           : AAAATTCAATTCAAACTAATTTATCATTAAATCAATTGATGACATTAGCAAATTCACAGGTAGCTGATAAATCATCATAT
NUF1071 Ic           : AAAATTCAATTCAAACTAATTTATCATTAAATCAATTGATGACATTAGCAAATTCACAGGTAGCTGATAAATCATCATAT
2007-1 NT ClusterI   : AAAATTCAATTCAAACTAATTTATCATTAAATCAATTGATGACATTAGCAAATTCACAGGTAGCTGATAAATCATCATAT
NUF1032 II           : AAAATTCAATTCAAACTAATTTATCATTAAATCAATTGATGACATTAGCAAATTCACAGGTAGCTGATAAATCATCATAT
NUF1095 NT ClusterIII: AAAATTCAATTCAAACTAATTTATCATTAAATCAATTGATGACATTAGCAAATTCACAGGTAGCTGATAAATCATCATAT

                              2490      2500      2510      2520      2530      2540      2550      2560
KRS02083 Ia          : ACAGTAACCTCACAAGATGTTACTGGTTCAGGGTCAACTGGTGAATTACCATCTTATGCAATGCCAGGTTCTGCTCTATA
NUF1003 Ib           : ACAGTAACGTCACAAGATGTTACTGGTTCAGGGTCAACTGGTGAATTACCATCTTATGCAATGCCAGGTTCTGCGCTATA
NUF1071 Ic           : ACAGTAACGTCACAAGATGTTACTGGTTCAGGGTCAACTGGTGAATTACCATCTTATGCAATGCCAGGTTCTGCGCTATA
2007-1 NT ClusterI   : ACAGTAACGTCACAAGATGTTACTGGTTCAGGGTCAACTGGTGAATTACCATCTTATGCAATGCCAGGTTCTGCGCTATA
NUF1032 II           : ACAGTAACGTCACAAGATGTTACTGGTTCAGGGTCAACTGGTGAATTACCATCTTATGCAATGCCAGGTTCTGCTCTATA
NUF1095 NT ClusterIII: ACAGTAACGTCACAAGATGTTACTGGTTCAGGGTCAACTGGTGAATTACCATCTTATGCAATGCCAGGTTCTGCTCTATA

                              2570      2580      2590      2600      2610      2620      2630      2640
KRS02083 Ia          : CATGTTAAAACTAGATGATCAGAGTGTCGAAACTGCAAAAGAAGCAATAAAAAATACCATGGAGGGGAACTAAATGATTG
NUF1003 Ib           : CATGTTAAAACTAGATGATCAGAGTGTCGAAACTGCAAAAGAAGCAATAAAAAATACCATGGAGGGGAACTAAATGATTG
NUF1071 Ic           : CATGTTAAAACTAGATGATCAGAGTGTCGAAACTGCAAAAGAAGCAATAAAAAATACCATGGAGGGGAACTAAATGATTG
2007-1 NT ClusterI   : CATGTTAAAACTAGATGATCAGAGTGTCGAAACTGCAAAAGAAGCAATAAAAAATACCATGGAGGGGAACTAAATGATTG
NUF1032 II           : CATGTTAAAACTAGATGATCAGAGTGTCGAAACTGCAAAAGAAGCAATAAAAAATACCATGGAGGGGAACTAAATGATTG
NUF1095 NT ClusterIII: CATGTTAAAACTAGATGATCAGAGTGTCGAAACTGCAAAAGAAGCAATAAAAAATACCATGGAGGGGAACTAAATGATTG
                                                                                                      cpsB

                              2650      2660      2670      2680      2690      2700      2710      2720
KRS02083 Ia          : ATATCCATTCCCACATCATTTTCGATGTGGATGATGGTCCGCTAACAATAGATGAAAGTTTATCGTTACTCGAGGAAAGT
NUF1003 Ib           : ATATCCATTCCCACATCATTTTCGATGTGGATGATGGTCCGCTAACAATAGATGAAAGTTTATCGTTACTCGAGGAAAGT
NUF1071 Ic           : ATATCCATTCCCACATCATTTTCGATGTGGATGATGGTCCGCTAACAATAGATGAAAGTTTATCGTTACTCGAGGAAAGT
2007-1 NT ClusterI   : ATATCCATTCCCACATCATTTTCGATGTGGATGATGGTCCGCTAACAATAGATGAAAGTTTATCGTTACTCGAGGAAAGT
NUF1032 II           : ATATCCATTCCCACATCATTTTCGATGTGGATGATGGTCCGCTAACAATAGATGAAAGTTTATCGTTACTCGAGGAAAGT
NUF1095 NT ClusterIII: ATATCCATTCCCACATCATTTTCGATGTGGATGATGGTCCGCTAACAATAGATGAAAGTTTATCGTTACTCGAGGAAAGT

                              2730      2740      2750      2760      2770      2780      2790      2800
KRS02083 Ia          : TATAAGCAAGGTGTTCGTACAATTGTATCAACATCACATCGCAGAAAAGGAATGTTTGAAACACCAGAAGATGACATTTT
NUF1003 Ib           : TATAAGCAAGGTGTTCGTACAATTGTATCAACATCACATCGCAGAAAAGGAATGTTTGAAACACCAGAAGATGATATTTT
NUF1071 Ic           : TATAAGCAAGGTGTTCGTACAATTGTATCAACATCACATCGCAGAAAAGGAATGTTTGAAACACCAGAAGATGATATTTT
2007-1 NT ClusterI   : TATAAGCAAGGTGTTCGTACAATTGTATCAACATCACATCGCAGAAAAGGAATGTTTGAAACACCAGAAGATGATATTTT
NUF1032 II           : TATAAGCAAGGTGTTCGTACAATTGTATCAACATCACATCGCAGAAAAGGAATGTTTGAAACACCAGAAGATGATATTTT
NUF1095 NT ClusterIII: TATAAGCAAGGTGTTCGTACAATTGTATCAACATCACATCGCAGAAAAGGAATGTTTGAAACACCAGAAGATGATATTTT
```

```
                              2810      2820      2830      2840      2850      2860      2870      2880
KRS02083 Ia         : AAATAAATTTAAGCATGTTAAACGGGAAGCTGCAGATAAGTTTCCAGACTTAACTTTGTTATACGGTGGTGAATTATACT
NUF1003 Ib          : AAATAAATTTAAGCATGTTAAACGGGAAGCTGCAGATAAGTTTCCAGACTTAACTTTGTTATACGGTGGTGAATTATACT
NUF1071 Ic          : AAATAAATTTAAGCATGTTAAACGGGAAGCTGCAGATAAGTTTCCAGACTTAACTTTGTTATACGGTGGTGAATTATACT
2007-1 NT ClusterI  : AAATAAATTTAAGCATGTTAAACGGGAAGCTGCAGATAAGTTTCCAGACTTAACTTTGTTATACGGTGGTGAATTATACT
NUF1032 II          : AAATAAATTTAAGCATGTTAAACGGGAAGCTGCAGATAAGTTTCCAGACTTAACTTTGTTATACGGTGGTGAATTATACT
NUF1095 NT ClusterIII: AAATAAATTTAAGCATGTTAAACGGGAAGCTGCAGATAAGTTTCCAGACTTAACTTTGTTATACGGTGGTGAATTATACT

                              2890      2900      2910      2920      2930      2940      2950      2960
KRS02083 Ia         : TCACAGCAGACATTCTAAAAAAACTGGAAAATAATGAAATTCCTCGCATGAACGATACACGTTTTGCTTTAATTGAATTT
NUF1003 Ib          : TCACAGCAGACATTCTAAAAAAACTGGAAAATAATGAAATTCCTCGCATGAACGATACACGTTTTGCTTTAATTGAATTT
NUF1071 Ic          : TCACAGCAGACATTCTAAAAAAACTGGAAAATAATGAAATTCCTCGCATGAACGATACACGTTTTGCTTTAATTGAATTT
2007-1 NT ClusterI  : TCACAGCAGACATTCTAAAAAAACTGGAAAATAATGAAATTCCTCGCATGAACGATACACGTTTTGCTTTAATTGAATTT
NUF1032 II          : TCACAGCAGACATTCTAAAAAAACTGGAAAATAATGAAATTCCTCGCATGAACGATACACGTTTTGCTTTAATTGAATTT
NUF1095 NT ClusterIII: TCACAGCAGACATTCTAAAAAAACTGGAAAATAATGAAATTCCTCGCATGAACGATACACGTTTTGCTTTAATTGAATTT

                              2970      2980      2990      3000      3010      3020      3030      3040
KRS02083 Ia         : AGCCAGATGACCCCATGGAAGGATATCCATTTAGCTTTATCTCAAGTGTTAATGTTAGGAATTACACCAATTGTCGCGCA
NUF1003 Ib          : AGCCAGATGACCCCATGGAAGGATATCCATTTAGCTTTATCTCAAGTGTTAATGTTAGGAATTACACCAATTGTCGCGCA
NUF1071 Ic          : AGCCAGATGACCCCATGGAAGGATATCCATTTAGCTTTATCTCAAGTGTTAATGTTAGGAATTACACCAATTGTCGCGCA
2007-1 NT ClusterI  : AGCCAGATGACCCCATGGAAGGATATCCATTTAGCTTTATCTCAAGTGTTAATGTTAGGAATTACACCAATTGTCGCGCA
NUF1032 II          : AGCCAGATGACCCCATGGAAGGATATCCATTTAGCTTTATCTCAAGTGTTAATGTTAGGAATTACACCAATTGTCGCGCA
NUF1095 NT ClusterIII: AGCCAGATGACCCCATGGAAGGATATCCATTTAGCTTTATCTCAAGTGTTAATGTTAGGAATTACACCAATTGTCGCGCA

                              3050      3060      3070      3080      3090      3100      3110      3120
KRS02083 Ia         : TATCGAACGCTACGCAGCTCTTGAATTTAATGGGGACCGTGTGCAGGAGTTAATTAATATGGGTTGTTACACGCAAGTCA
NUF1003 Ib          : TATCGAACGCTACGCAGCTCTTGAATTTAATGGGGACCGTGTGCAGGAGTTAATTAATATGGGTTGTTACACGCAAGTCA
NUF1071 Ic          : TATCGAACGCTACGCAGCTCTTGAATTTAATGGGGACCGTGTGCAGGAGTTAATTAATATGGGTTGTTACACGCAAGTCA
2007-1 NT ClusterI  : TATCGAACGCTACGCAGCTCTTGAATTTAATGGGGACCGTGTGCAGGAGTTAATTAATATGGGTTGTTACACGCAAGTCA
NUF1032 II          : TATCGAACGCTACGCAGCTCTTGAATTTAATGGGGACCGTGTGCAGGAGTTAATTAATATGGGTTGTTACACGCAAGTCA
NUF1095 NT ClusterIII: TATCGAACGCTACGCAGCTCTTGAATTTAATGGGGACCGTGTGCAGGAGTTAATTAATATGGGTTGTTACACGCAAGTCA

                              3130      3140      3150      3160      3170      3180      3190      3200
KRS02083 Ia         : ATAGTGCCCATGTCTTAAAAGCGAAGCTATTTGGTGATAAACTTAAAAACTTTAAAAAACGTGCCAAATACTTTTTGGAT
NUF1003 Ib          : ATAGTGCCCATGTCTTAAAAGCGAAGCTATTTGGTGATAAACTTAAAATCTTTAAAAAACGTACCAAATACTTTTTGGAT
NUF1071 Ic          : ATAGTGCCCATGTCTTAAAAGCGAAGCTATTTGGTGATAAACTTAAAATCTTTAAAAAACGTACCAAATACTTTTTGGAT
2007-1 NT ClusterI  : ATAGTGCCCATGTCTTAAAAGCGAAGCTATTTGGTGATAAACTTAAAATCTTTAAAAAACGTACCAAATACTTTTTGGAT
NUF1032 II          : ATAGTGCCCATGTCTTAAAAGCGAAGCTATTTGGTGATAAACTTAAAATCTTTAAAAAACGTGCCAAATACTTTTTGGAT
NUF1095 NT ClusterIII: ATAGTGCCCATGTCTTAAAAGCGAAGCTATTTGGTGATAAACTTAAAATCTTTAAAAAACGTGCCAAATACTTTTTGGAT

                              3210      3220      3230      3240      3250      3260      3270      3280
KRS02083 Ia         : AAGGACTTAGTCCATTGTATATCAAGTGATATGCATAATCTAAAAAAACGTCCACCATATATGAAACAGGCTTATCAAAT
NUF1003 Ib          : AAGGACTTAGTCCATTGTATATCAAGTGATATGCATAATCTAAAAAAACGTCCACCATATATGAAACAGGCTTATCAAAT
NUF1071 Ic          : AAGGACTTAGTCCATTGTATATCAAGTGATATGCATAATCTAAAAAAACGTCCACCATATATGAAACAGGCTTATCAAAT
2007-1 NT ClusterI  : AAGGACTTAGTCCATTGTATATCAAGTGATATGCATAATCTAAAAAAACGTCCACCATATATGAAACAGGCTTATCAAAT
NUF1032 II          : AAGGACTTAGTCCATTGTATATCAAGTGATATGCATAATCTAAAAAAACGTCCACCATATATGAAACAGGCTTATCAAAT
NUF1095 NT ClusterIII: AAGGACTTAGTCCATTGTATATCAAGTGATATGCATAATCTAAAAAAACGTCCACCATATATGAAACAGGCTTATCAAAT

                              3290      3300      3310      3320      3330      3340      3350      3360
KRS02083 Ia         : CGTTGAAAAGGATTATGGAACAAGACGGGCAAGAAAACTTTTTGAAGAGAATGCTGCAAGTTTGATCAATAACGAATTTT
NUF1003 Ib          : CGTTGAAAAGGATTATGGAACAAGACGGGCAAGAAAACTTTTTGAAGAGAATGCTGCAAGTTTGATCAATAACGAATTTT
NUF1071 Ic          : CGTTGAAAAGGATTATGGAACAAGACGGGCAAGAAAACTTTTTGAAGAGAATGCTGCAAGTTTGATCAATAACGAATTTT
2007-1 NT ClusterI  : CGTTGAAAAGGATTATGGAACAAGACGGGCAAGAAAACTTTTTGAAGAGAATGCTGCAAGTTTGATCAATAACGAATTTT
NUF1032 II          : CGTTGAAAAGGATTATGGAACAAGACGGGCAAGAAAACTTTTTGAAGAGAATGCTGCAAGTTTGATCAATAACGAATTTT
NUF1095 NT ClusterIII: CGTTGAAAAGGATTATGGAACAAGACGGGCAAGAAAACTTTTTGAAGAGAATGCTGCAAGTTTGATCAATAACGAATTTT
```

```
                          3370      3380      3390      3400      3410      3420      3430      3440
KRS02083 Ia          : TATAGGAGAATTTATGAATAACATGGAAAAACCATCAATGGAAATTGATGTATTGAGTTTATTAAAGAAGTTATGGACGA
NUF1003 Ib           : TATAGGAGAATTTATGAATAACATGGAAAAACCATCAATGGAAATTGATGTATTGAGTTTATTAAAGAAGTTATGGACGA
NUF1071 Ic           : TATAGGAGAATTTATGAATAACATGGAAAAACCATCAATGGAAATTGATGTATTGAGTTTATTAAAGAAGTTATGGACGA
2007-1 NT ClusterI   : TATAGGAGAATTTATGAATAACATGGAAAAACCATCAATGGAAATTGATGTATTGAGTTTATTAAAGAAGTTATGGACGA
NUF1032 II           : TATAGGAGAATTTATGAATAACATGGAAAAACCATCAATGGAAATTGATGTATTGAGTTTATTAAAGAAGTTATGGACGA
NUF1095 NT ClusterIII: TATAGGAGAATTTATGAATAACATGGAAAAACCATCAATGGAAATTGATGTATTGAGTTTATTAAAGAAGTTATGGACGA
                                   cpsC

                          3450      3460      3470      3480      3490      3500      3510      3520
KRS02083 Ia          : AGAAGTTTTTAATTATTTTCATGGCCCTGTTTTTTGGGACTTTGGCTTTATTATCTAGTATCTTCTTAATTAAGCCCTCT
NUF1003 Ib           : AGAAGTTTTTAATTATTTTCATGGCCCTGTTTTTTGGGACTTTGGCTTTATTATCTAGTATCTTCTTAATTAAGCCCTCT
NUF1071 Ic           : AGAAGTTTTTAATTATTTTCATGGCCCTGTTTTTTGGGACTTTGGCTTTATTATCTAGTATCTTCTTAATTAAGCCCTCT
2007-1 NT ClusterI   : AGAAGTTTTTAATTATTTTCATGGCCCTGTTTTTTTGGACTTTGGCTTTATTATCTAGTATCTTCTTAATTAAGCCCTCT
NUF1032 II           : AGAAGTTTTTAATTATTTTCATGGCCCTGTTTTTTGGGACTTTGGCTTTATTATCTAGTATCTTCTTAATTAAGCCCTCT
NUF1095 NT ClusterIII: AGAAGTTTTTAATTATTTTCATGGCCCTGTTTTTTGGGACTTTGGCTTTATTATCTAGTATCTTCTTAATTAAGCCCTCT

                          3530      3540      3550      3560      3570      3580      3590      3600
KRS02083 Ia          : TACACAGCATCAACACGTATTTATGTTCTTAATAAAACACAACAAGCTGATAATTTGTCAGCAACTGACTTGCAAGCCGG
NUF1003 Ib           : TACACAGCGTCAACACGTATTTATGTTCTAAATAAGACACAACAAGCTGATAATTTGTCAGCAACTGACTTGCAAGCCGG
NUF1071 Ic           : TACACAGCGTCAACACGTATTTATGTTCTAAATAAGACACAACAAGCTGATAATTTGTCAGCAACTGACTTGCAAGCCGG
2007-1 NT ClusterI   : TACACAGCGTCAACACGTATTTATGTTCTAAATAAGACACAACAAGCTGATAATTTGTCAGCAACTGACTTGCAAGCCGG
NUF1032 II           : TACACAGCGTCAACACGTATTTATGTTCTAAATAAGACACAACAAGCTGATAATTTGTCAGCAACTGACTTGCAAGCCGG
NUF1095 NT ClusterIII: TACACAGCGTCAACACGTATTTATGTTCTAAATAAGACACAACAAGCTGATAATTTGTCAGCAACTGACTTGCAAGCCGG

                          3610      3620      3630      3640      3650      3660      3670      3680
KRS02083 Ia          : TGGTTTATTAGTTAATGACTATAAGGAAATTATTACATCTCGTGATGTTATGAAAGATGTCATTGCTAATGACGGTGTTT
NUF1003 Ib           : TGGTTTATTAGTTAATGACTATAAGGAAATTATAACATCTCGTGATGTTATGAAAGATGTCATTGCTAATGACGGTGTTT
NUF1071 Ic           : TGGTTTATTAGTTAATGACTATAAGGAAATTATAACATCTCGTGATGTTATGAAAGATGTCATTGCTAATGACGGTGTTT
2007-1 NT ClusterI   : TGGTTTATTAGTTAATGACTATAAGGAAATTATAACATCTCGTGATGTTATGAAAGATGTCATTGCTAATGACGGTGTTT
NUF1032 II           : TGGTTTATTAGTTAATGACTATAAGGAAATTATAACATCTCGTGATGTTATGAAAGATGTCATTGCTAATGACGGTGTTT
NUF1095 NT ClusterIII: TGGTTTATTAGTTAATGACTATAAGGAAATTATAACATCTCGTGATGTTATGAAAGATGTCATTGCTAATGACGGTGTTT

                          3690      3700      3710      3720      3730      3740      3750      3760
KRS02083 Ia          : CAATGACACCGGAAGAATTGAGTAAGATGATTGCTGTAACAATTCCAGCAGATACACGTGTTATCTCAATTGCTGTGACA
NUF1003 Ib           : CAATGACACCGGAAGAATTGAGTAAGATGATTGCTGTAACAATTCCAGCAGATACACGTGTTATCTCAATTGCTGTGACA
NUF1071 Ic           : CAATGACACCGGAAGAATTGAGTAAGATGATTGCTGTAACAATTCCAGCAGATACACGTGTTATCTCAATTGCTGTGACA
2007-1 NT ClusterI   : CAATGACACCGGAAGAATTGAGTAAGATGATTGCTGTAACAATTCCAGCAGATACACGTGTTATCTCAATTGCTGTGACA
NUF1032 II           : CAATGACACCGGAAGAATTGAGTAAGATGATTGCTGTAACAATTCCAGCAGATACACGTGTTATCTCAATTGCTGTGACA
NUF1095 NT ClusterIII: CAATGACACCGGAAGAATTGAGTAAGATGATTGCTGTAACAATTCCAGCAGATACACGTGTTATCTCAATTGCTGTGACA

                          3770      3780      3790      3800      3810      3820      3830      3840
KRS02083 Ia          : AATCATGATCCTCAGGCTGCTAAAGACTTAGCAAACTCAGTTCGCGATGTTGCATCCGAAAAAATTAAATCAGTCACTAA
NUF1003 Ib           : AATCATGATCCTCAGGCTGCTAAAGACTTAGCAAACTCAGTTCGCGATGTTGCATCCGAAAAAATTAAATCAGTCACTAA
NUF1071 Ic           : AATCATGATCCTCAGGCTGCTAAAGACTTAGCAAACTCAGTTCGCGATGTTGCATCCGAAAAAATTAAATCAGTCACTAA
2007-1 NT ClusterI   : AATCATGATCCTCAGGCTGCTAAAGACTTAGCAAACTCAGTTCGCGATGTTGCATCCGAAAAAATTAAATCAGTCACTAA
NUF1032 II           : AATCATGATCCTCAGGCTGCTAAAGACTTAGCAAACTCAGTTCGCGATGTTGCATCCAAAAAAATTAAATCAGTCACTAA
NUF1095 NT ClusterIII: AATCATGATCCTCAGGCTGCTAAAGACTTAGCAAACTCAGTTCGCGATGTTGCATCCAAAAAAATTAAATCAGTCACTAA

                          3850      3860      3870      3880      3890      3900      3910      3920
KRS02083 Ia          : GGTCCAAGATGTGACTCCGCTTGAAAAAGCAGTTATGCCGACAAGTCCATCTTCACCAAATATCAAACGTAACACACTTG
NUF1003 Ib           : GGTCCAAGATGTGACTCCGCTTGAAAAAGCAGTTATGCCGACAAGTCCATCTTCACCAAATATCAAACGTAACACACTTG
NUF1071 Ic           : GGTCCAAGATGTGACTCCGCTTGAAAAAGCAGTTATGCCGACAAGTCCATCTTCACCAAATATCAAACGTAACACACTTG
2007-1 NT ClusterI   : GGTCCAAGATGTGACTCCGCTTGAAAAAGCAGTTATGCCGACAAGTCCATCTTCACCAAATATCAAACGTAACACACTTG
NUF1032 II           : GGTCCAAAATGTAACTCCGCTTGAAAAAGCAGTTATGCCGACAAGTCCATCTTCACCAAATATCAAACGTAACACACTTG
NUF1095 NT ClusterIII: GGTCCAAGATGTAACTCCGCTTGAAAAAGCAGTTATGCCGACAAGTCCATCTTCACCAAATATCAAACGTAACACACTTG
```

```
                             3930      3940      3950      3960      3970      3980      3990      4000
KRS02083 Ia          : TTGGTATTTTATTCGGTGCTTTCTTAACAATCGTGACAGTGATTGTCGGAGAAGTCGTTGATGACCGCGTGAAACGCCCA
NUF1003 Ib           : TTGGTATTTTATTCGGTGCTTTCTTAACAATCGTGACAGTGATTGTCGGAGAAGTCGTTGATGACCGCGTGAAACGCCCA
NUF1071 Ic           : TTGGTATTTTATTCGGTGCTTTCTTAACAATCGTGACAGTGATTGTCGGAGAAGTCGTTGATGACCGCGTGAAACGCCCA
2007-1 NT ClusterI   : TTGGTATTTTATTCGGTGCTTTCTTAACAATCGTGACAGTGATTGTCGGAGAAGTCGTTGATGACCGCGTGAAACGCCCA
NUF1032 II           : TTGGTATTTTATTCGGTGCTTTCTTAACAATCGTGACAGTGATTGTCGGAGAAGTCGTTGATGACCGCGTGAAACGCCCA
NUF1095 NT ClusterIII: TTGGTATTTTATTCGGTGCTTTCTTAACAATCGTGACAGTGATTGTCGGAGAAGTCGTTGATGACCGCGTGAAACGCCCA

                             4010      4020      4030      4040      4050      4060      4070      4080
KRS02083 Ia          : GAAGATGTTGAAGAGCTTCTTGGTATGACTTTGTTAGGTATTGTACCTAACACTAACAAGATGTAAGAGGTAAAGAATGG
NUF1003 Ib           : GAAGATGTTGAAGAGCTTCTTGGTATGACTTTGTTAGGTATTGTACCTAACACTAACAAGATGTAAGAGGTAAAGAATGG
NUF1071 Ic           : GAAGATGTTGAAGAGCTTCTTGGTATGACTTTGTTAGGTATTGTACCTAACACTAACAAGATGTAAGAGGTAAAGAATGG
2007-1 NT ClusterI   : GAAGATGTTGAAGAGCTTCTTGGTATGACTTTGTTAGGTATTGTACCTAACACTAACAAGATGTAAGAGGTAAAGAATGG
NUF1032 II           : GAAGATGTTGAAGAGCTTCTTGGTATGACTTTGTTAGGTATTGTACCTAACACTAACAAGATGTAAGAGGTAAAGAATGG
NUF1095 NT ClusterIII: GAAGATGTTGAAGAGCTTCTTGGTATGACTTTGTTAGGTATTGTACCTAACACTAACAAGATGTAAGAGGTAAAGAATGG
                                                                                                   cpsD

                             4090      4100      4110      4120      4130      4140      4150      4160
KRS02083 Ia          : CACGTTTAGAATTAGTAAAAAGTAAAAAAGAACTATATGATATCGCCGAAGAGTACTATAACTCCATTCGGACTAACATC
NUF1003 Ib           : CACGTTTAGAATTAGTAAAAAGTAAAAAAGAACTATATGATATCGCCGAAGAGTACTATAACTCCATTCGGACTAACATC
NUF1071 Ic           : CACGTTTAGAATTAGTAAAAAGTAAAAAAGAACTATATGATATCGCCGAAGAGTACTATAACTCCATTCGGACTAACATC
2007-1 NT ClusterI   : CACGTTTAGAATTAGTAAAAAGTAAAAAAGAACTATATGATATCGCCGAAGAGTACTATAACTCCATTCGGACTAACATC
NUF1032 II           : CACGTTTAGAATTAGTAAAAAGTAAAAAAGAACTATATGATATCGCCGAAGAGTACTATAACTCCATTCGGACTAACATC
NUF1095 NT ClusterIII: CACGTTTAGAATTAGTAAAAAGTAAAAAAGAACTATATGATATCGCCGAAGAGTACTATAACTCCATTCGGACTAACATC

                             4170      4180      4190      4200      4210      4220      4230      4240
KRS02083 Ia          : CAATTTAGTGGTCGAGATTTAAAAGTTATCACCTTAACATCAGTACAACCTGGTGAAGGGAAATCGACATTATCCGCAAA
NUF1003 Ib           : CAATTTAGTGGTCGAGATTTAAAAGTTATCACCTTAACATCAGTACAACCTGGTGAAGGGAAATCGACAACATCCGCAAA
NUF1071 Ic           : CAATTTAGTGGTCGAGATTTAAAAGTTATCACCTTAACATCAGTACAACCTGGTGAAGGGAAATCGACAACATCCGCAAA
2007-1 NT ClusterI   : CAATTTAGTGGTCGAGATTTAAAAGTTATCACCTTAACATCAGTACAACCTGGTGAAGGGAAATCGACAACATCCGCAAA
NUF1032 II           : CAATTTAGTGGTCGAGATTTAAAAGTTATCACCTTAACATCAGTACAACCTGGTGAAGGGAAATCGACATTATCCGCAAA
NUF1095 NT ClusterIII: CAATTTAGTGGTCGAGATTTAAAAGTTATCACCTTAACATCAGTACAACCTGGTGAAGGGAAATCGACATTATCCGCAAA

                             4250      4260      4270      4280      4290      4300      4310      4320
KRS02083 Ia          : TATTGCTATCTCATTTGCTAAAGCAGGTCTTAAAACCCTATTAATCGATGCAGACATCCGTAATTCAGTTATGTCTGGTA
NUF1003 Ib           : TATTGCTATCTCATTTGCTAAAGCAGGTCTAAAAACCCTATTAATCGATGCAGACATCCGTAATTCAGTTATGTCTGGTA
NUF1071 Ic           : TATTGCTATCTCATTTGCTAAAGCAGGTCTAAAAACCCTATTAATCGATGCAGACATCCGTAATTCAGTTATGTCTGGTA
2007-1 NT ClusterI   : TATTGCTATCTCATTTGCTAAAGCAGGTCTAAAAACCCTATTAATCGATGCAGACATCCGTAATTCAGTTATGTCTGGTA
NUF1032 II           : TATTGCTATCTCATTTGCTAAAGCAGGTCTTAAAACCCTATTAATCGATGCAGACATCCGTAATTCAGTTATGTCTGGTA
NUF1095 NT ClusterIII: TATTGCTATCTCATTTGCTAAAGCAGGTCTTAAAACCCTATTAATCGATGCAGACATCCGTAATTCAGTTATGTCTGGTA

                             4330      4340      4350      4360      4370      4380      4390      4400
KRS02083 Ia          : CATTTAAAGCTGATGAAAAGTATGAAGGTCTATCAAGTTACCTATCAGGTAATGCAGAATTATCAGCAGTTATTTCTCAT
NUF1003 Ib           : CATTTAAAGCTGATGAAAAGTATGAAGGTCTATCAAGTTACCTGTCAGGTAATGCAGAATTATCAGCAGTTATCTCTCAT
NUF1071 Ic           : CATTTAAAGCTGATGAAAAGTATGAAGGTCTATCAAGTTACCTGTCAGGTAATGCAGAATTATCAGCAGTTATCTCTCAT
2007-1 NT ClusterI   : CATTTAAAGCTGATGAAAAGTATGAAGGTCTATCAAGTTACCTGTCAGGTAATGCAGAATTATCAGCAGTTATCTCTCAT
NUF1032 II           : CATTTAAAGCTGATGAAAAGTATGAAGGTCTATCAAGTTACCTATCAGGTAATGCAGAATTATCAGCAGTTATTTCTCAT
NUF1095 NT ClusterIII: CATTTAAAGCTGATGAAAAGTATGAAGGTCTATCAAGTTACCTATCAGGTAATGCAGAATTATCAGCAGTTATTTCTCAT

                             4410      4420      4430      4440      4450      4460      4470      4480
KRS02083 Ia          : ACAAATATTGAAAACTTAATGTTGATTCCAGCAGGACATGTTCCTCCTAATCCAACAACTTTACTCCAAAATAGCAATTT
NUF1003 Ib           : ACAAATATTGAAAACTTAATGTTGATTCCAGCAGGACATGTTCCTCCTAATCCAACAACTTTACTCCAAAATAGCAATTT
NUF1071 Ic           : ACAAATATTGAAAACTTAATGTTGATTCCAGCAGGACATGTTCCTCCTAATCCAACAACTTTACTCCAAAATAGCAATTT
2007-1 NT ClusterI   : ACAAATATTGAAAACTTAATGTTGATTCCAGCAGGACATGTTCCTCCTAATCCAACAACTTTACTCCAAAATAGCAATTT
NUF1032 II           : ACAAATATTGAAAACTTAATGTTGATTCCAGCAGGACATGTTCCTCCTAATCCAACAACTTTACTCCAAAATAGCAATTT
NUF1095 NT ClusterIII: ACAAATATTGAAAACTTAATGTTGATTCCAGCAGGACATGTTCCTCCTAATCCAACAACTTTACTCCAAAATAGCAATTT
```

```
                             4490      4500      4510      4520      4530      4540      4550      4560
KRS02083 Ia          : TAATTTCATGATTGATACTGTAAAAGAGTTATTTGATTATGTGATTATCGATACCCCACCTATTGGCCTTGTTATAGACT
NUF1003 Ib           : TAATTTCATGATTGATACTGTAAAAGAGTTATTTGATTATGTGATTATCGATACCCCACCTATTGGCCTTGTTATCGACT
NUF1071 Ic           : TAATTTCATGATTGATACTGTAAAAGAGTTATTTGATTATGTGATTATCGATACCCCACCTATTGGCCTTGTTATCGACT
2007-1 NT ClusterI   : TAATTTCATGATTGATACTGTAAAAGAGTTATTTGATTATGTGATTATCGATACCCCACCTATTGGCCTTGTTATCGACT
NUF1032 II           : TAATTTCATGATTGATACTGTAAAAGAGTTATTTGATTATGTGATTATCGATACCCCACCTATTGGCCTTGTTATAGACT
NUF1095 NT ClusterIII: TAATTTCATGATTGATACTGTAAAAGAGTTATTTGATTATGTGATTATCGATACCCCACCTATTGGCCTTGTTATAGACT

                             4570      4580      4590      4600      4610      4620      4630      4640
KRS02083 Ia          : CAGCGATTATTTCACAAAAAGCTGACGCAAACATTTTAGTAACAGAAGCAGGGGCTATTAAACGACGCTTTATCCAAAAA
NUF1003 Ib           : CAGCGATTATTTCACAAAAAGCTGACGCAAACATCTTAGTAACAGAAGCTGGGGCTATTAAACGACGCTTTATCCAAAAA
NUF1071 Ic           : CAGCGATTATTTCACAAAAAGCTGACGCAAACATCTTAGTAACAGAAGCTGGGGCTATTAAACGACGCTTTATCCAAAAA
2007-1 NT ClusterI   : CAGCGATTATTTCACAAAAAGCTGACGCAAACATCTTAGTAACAGAAGCTGGGGCTATTAAACGACGCTTTATCCAAAAA
NUF1032 II           : CAGCGATTATTTCACAAAAAGCTGACGCAAACATTTTAGTAACAGAAGCAGGGGCTATTAAACGACGCTTTATCCAAAAA
NUF1095 NT ClusterIII: CAGCGATTATTTCACAAAAAGCTGACGCAAACATTTTAGTAACAGAAGCAGGGGCTATTAAACGACGCTTTATCCAAAAA

                             4650      4660      4670      4680      4690      4700      4710      4720
KRS02083 Ia          : GCAAAAGAACAAATGGAACAAAGTGGTGCCTTGTTCTTGGGTGTTATTTTAAATAAAGTAGAAGAAACACTTGATTCATA
NUF1003 Ib           : GCAAAAGAACAAATGGAACAAAGTGGTGCTTTGTTCTTGGGTGTTATTTTAAATAAAGTAGAAGAAACACTTGATTCATA
NUF1071 Ic           : GCAAAAGAACAAATGGAACAAAGTGGTGCTTTGTTCTTGGGTGTTATTTTAAATAAAGTAGAAGAAACACTTGATTCATA
2007-1 NT ClusterI   : GCAAAAGAACAAATGGAACAAAGTGGTGCTTTGTTCTTGGGTGTTATTTTAAATAAAGTAGAAGAAACACTTGATTCATA
NUF1032 II           : GCAAAAGAACAAATGGAACAAAGTGGTGCCTTGTTCTTGGGTGTTATTTTAAATAAAGTAGAAGAAACACTTGATTCATA
NUF1095 NT ClusterIII: GCAAAAGAACAAATGGAACAAAGTGGTGCCTTGTTCTTGGGTGTTATTTTAAATAAAGTAGAAGAAACACTTGATTCATA

                             4730      4740      4750      4760      4770      4780      4790      4800
KRS02083 Ia          : TGGTGGTTATGGTAGTTATGGGGCTTACGGGAATTATGGTAAGCCAGCTAAAAAGAAATCTAGAAAGAGAAGATAAAAAA
NUF1003 Ib           : TGGTGGTTATGGTAGTTATGGGGCTAACGGGAATTATGGTAAGCCAGCTAAAAAGAAATCTAGAAAGAGAAGATAAAAAA
NUF1071 Ic           : TGGTGGTTATGGTAGTTATGGGGCTAACGGGAATTATGGTAAGCCAGCTAAAAAGAAATCTAGAAAGAGAAGATAAAAAA
2007-1 NT ClusterI   : TGGTGGTTATGGTAGTTATGGGGCTAACGGGAATTATGGTAAGCCAGCTAAAAAGAAATCTAGAAAGAGAAGATAAAAAA
NUF1032 II           : TGGTGGTTATGGTAGTTATGGGGCTTACGGGAATTATGGTAAGCCAGCTAAAAAGAAATCTAGAAAGAGAAGATAAAAAA
NUF1095 NT ClusterIII: TGGTGGTTATGGTAGTTATGGGGCTTACGGGAATTATGGTAAGCCAGCTAAAAAGAAATCTAGAAAGAGAAGATAAAAAA

                             4810      4820      4830      4840      4850      4860      4870      4880
KRS02083 Ia          : TGAAAAGAAGTCAAAAAAGAGCGATTTTATATCTGATTGATGTGTTTATGATTACATTCGCACATCTGTCGGCTTTTCGG
NUF1003 Ib           : TGAAAAGAAGTCAAAAAAGAGCGATTTTATATCTGATTGATGTGTTTATGATCACATTCGCAAATCTGTCGGCTTTTCGG
NUF1071 Ic           : TGAAAAGAAGTCAAAAAAGAGCGATTTTATATCTGATTGATGTGTTTATGATCACATTCGCAAATCTGTCGGCTTTTCGG
2007-1 NT ClusterI   : TGAAAAGAAGTCAAAAAAGAGCGATTTTATATCTGATTGATGTGTTTATGATCACATTCGCAAATCTGTCGGCTTTTCGG
NUF1032 II           : TGAAAAGAAGTCAAAAAAGAGCGATTTTATATCTGATTGATGTGTTTATGATTACATTCGCACATCTGTCGGCTTTTCGG
NUF1095 NT ClusterIII: TGAAAAGAAGTCAAAAAAGAGCGATTTTATATCTGATTGATGTGTTTATGATTACATTCGCACATCTGTCGGCTTTTCGG
                       cpsE

                             4890      4900      4910      4920      4930      4940      4950      4960
KRS02083 Ia          : TTCTTACTATCATACAGTACTCGATTAAGCGACAAAGAAATCTATATCACTCTTTTTGTAACACTTTTCGTTTATACAGT
NUF1003 Ib           : TTCTTGCTATCATACAGTACTCGATTAAGCGACAAAGAAATCTATATCACTCTTTTTGTAACACTTTTCGTTTATACAGT
NUF1071 Ic           : TTCTTGCTATCATACAGTACTCGATTAAGCGACAAAGAAATCTATATCACTCTTTTTGTAACACTTTTCGTTTATACAGT
2007-1 NT ClusterI   : TTCTTGCTATCATACAGTACTCGATTAAGCGACAAAGAAATCTATATCACTCTTTTTGTAACACTTTTCGTTTATACAGT
NUF1032 II           : TTCTTACTATCATACAGTACTCGATTAAGCGACAAAGAAATCTATATCACTCTTTTTGTAACACTTTTCGTTTATACAGT
NUF1095 NT ClusterIII: TTCTTACTATCATACAGTACTCGATTAAGCGACAAAGAAATCTATATCACTCTTTTTGTAACACTTTTCGTTTATACAGT

                             4970      4980      4990      5000      5010      5020      5030      5040
KRS02083 Ia          : TTTAGGAATAAGAGCTCGAATCTTTTCCATTATCAATCGCTTTACCGATTATAAAGTCATCTTTATCCTAATAGCTAACA
NUF1003 Ib           : TTTAGGAATAAGAGCTCGAATCTTTTCCATTATCAATCGCTTTACCGATTATAAAGTCATCTTTATCCTAATAGCTAACA
NUF1071 Ic           : TTTAGGAATAAGAGCTCGAATCTTTTCCATTATCAATCGCTTTACCGATTATAAAGTCATCTTTATCCTAATAGCTAACA
2007-1 NT ClusterI   : TTTAGGAATAAGAGCTCGAATCTTTTCCATTATCAATCGCTTTACCGATTATAAAGTCATCTTTATCCTAATAGCTAACA
NUF1032 II           : TTTAGGAATAAGAGCTCGAATCTTTTCCATTATCAATCGCTTTACCGATTATAAAGTCATCTTTATCCTAATAGCTAACA
NUF1095 NT ClusterIII: TTTAGGAATAAGAGCTCGAATCTTTTCCATTATCAATCGCTTTACCGATTATAAAGTCATCTTTATCCTAATAGCTAACA
```

```
                             5050      5060      5070      5080      5090      5100      5110      5120
KRS02083 Ia          : TGTTCTTCGCATCCTTATTAAGTTATTTGGTCGACGTTCTTTTCCTAGATACATTCAGTCGTCGTTTTCTTTTCTTATCA
NUF1003 Ib           : TGTTCTTCGCATCCCTATTAAGTTATTTGGTGGACGTTCTTTTCCTAGATACATTCAGTCGTCGTTTTCTTTTCTTATCA
NUF1071 Ic           : TGTTCTTCGCATCCCTATTAAGTTATTTGGTGGACGTTCTTTTCCTAGATACATTCAGTCGTCGTTTTCTTTTCTTATCA
2007-1 NT ClusterI   : TGTTCTTCGCATCCCTATTAAGTTATTTGGTGGACGTTCTTTTCCTAGATACATTCAGTCGTCGTTTTCTTTTCTTATCA
NUF1032 II           : TGTTCTTCGCATCCTTATTAAGTTATTTGGTCGACGTTCTTTTCCTAGATACATTCAGTCGTCGTTTTCTTTTCTTATCA
NUF1095 NT ClusterIII: TGTTCTTCGCATCCTTATTAAGTTATTTGGTCGACGTTCTTTTCCTAGATACATTCAGTCGTCGTTTTCTTTTCTTATCA

                             5130      5140      5150      5160      5170      5180      5190      5200
KRS02083 Ia          : TTCCTTTTTGGAACATTTTTAATTATCCTTCCTCGGATGATTTGGCGGATGTGGCATGAACAAAATTTGTTTGTCAAACA
NUF1003 Ib           : TTCCTTTTTGGAACATTTTTAATTATCCTTCCTCGGATGATTTGGCGGATGTGGCATGAACAAAATTTGTTTGTCAAACA
NUF1071 Ic           : TTCCTTTTTGGAACATTTTTAATTATCCTTCCTCGGATGATTTGGCGGATGTGGCATGAACAAAATTTGTTTGTCAAACA
2007-1 NT ClusterI   : TTCCTTTTTGGAACATTTTTAATTATCCTTCCTCGGATGATTTGGCGGATGTGGCATGAACAAAATTTGTTTGTCAAACA
NUF1032 II           : TTCCTTTTTGGAACATTTTTAATTATCCTTCCTCGGATGATTTGGCGGATGTGGCATGAACAAAATTTGTTTGTCAAACA
NUF1095 NT ClusterIII: TTCCTTTTTGGAACATTTTTAATTATCCTTCCTCGGATGATTTGGCGGATGTGGCATGAACAAAATTTGTTTGTCAAACA

                             5210      5220      5230      5240      5250      5260      5270      5280
KRS02083 Ia          : TAATAAAAAAGACCAAAAGACAAAAATGTTGGTTGTTGGTGCCGGTGAAGGTGGTAGTGCCTTTATTCAAACAATTCTGA
NUF1003 Ib           : TAATAAAAAAGACCAAAAGACAAAAATGTTGGTTGTTGGTGCTGGTGAAGGTGGTAGTGCCTTTATTCAAACAATTCTGA
NUF1071 Ic           : TAATAAAAAAGACCAAAAGACAAAAATGTTGGTTGTTGGTGCTGGTGAAGGTGGTAGTGCCTTTATTCAAACAATTCTGA
2007-1 NT ClusterI   : TAATAAAAAAGACCAAAAGACAAAAATGTTGGTTGTTGGTGCTGGTGAAGGTGGTAGTGCCTTTATTCAAACAATTCTGA
NUF1032 II           : TAATAAAAAAGACCAAAAGACAAAAATGTTGGTTGTTGGTGCCGGTGAAGGTGGTAGTGCCTTTATTCAAACAATTCTGA
NUF1095 NT ClusterIII: TAATAAAAAAGACCAAAAGACAAAAATGTTGGTTGTTGGTGCCGGTGAAGGTGGTAGTGCCTTTATTCAAACAATTCTGA

                             5290      5300      5310      5320      5330      5340      5350      5360
KRS02083 Ia          : ATAAGAGTAAAGATATTGACATTGTCGGTATTGTTGATGCTGATATCAATAAATTAGGCACCTACTTACATGGGATTAAA
NUF1003 Ib           : ATAAGAGTAAAGATATTGACATTGTCGGTATTGTTGATGCTGATATCAATAAATTAGGCACCTACTTACACGGGATTAAA
NUF1071 Ic           : ATAAGAGTAAAGATATTGACATTGTCGGTATTGTTGATGCTGATATCAATAAATTAGGCACCTACTTACACGGGATTAAA
2007-1 NT ClusterI   : ATAAGAGTAAAGATATTGACATTGTCGGTATTGTTGATGCTGATATCAATAAATTAGGCACCTACTTACACGGGATTAAA
NUF1032 II           : ATAAGAGTAAAGATATTGACATTGTCGGTGTTGTTGATGCTGATATCAATAAATTAGGCACCTACTTACATGGGATTAAA
NUF1095 NT ClusterIII: ATAAGAGTAAAGATATTGACATTGTCGGTGTTGTTGATGCTGATATCAATAAATTAGGCACCTACTTACATGGGATTAAA

                             5370      5380      5390      5400      5410      5420      5430      5440
KRS02083 Ia          : GTACTGGGAAATAAAAATTCCATTCCAAGATTAGTAGCAGAATATGAAGTTAAACAAGTTACGATTGCCATCCCAAGTTT
NUF1003 Ib           : GTACTGGGAAATAAAAATTCCATTCCAAGATTACTAGCAGAATATGAAGTTAAACAGGTTACGATTGCCATCCCAAGTTT
NUF1071 Ic           : GTACTGGGAAATAAAAATTCCATTCCAAGATTACTAGCAGAATATGAAGTTAAACAGGTTACGATTGCCATCCCAAGTTT
2007-1 NT ClusterI   : GTACTGGGAAATAAAAATTCCATTCCAAGATTACTAGCAGAATATGAAGTTAAACAGGTTACGATTGCCATCCCAAGTTT
NUF1032 II           : GTACTGGGAAATAAAAATTCCATTCCAAGATTAGTAGCAGAATATGAAGTTAAACAAGTTACGATTGCCATCCCAAGTTT
NUF1095 NT ClusterIII: GTACTGGGAAATAAAAATTCCATTCCAAGATTAGTAGCAGAATATGAAGTTAAACAAGTTACGATTGCCATCCCAAGTTT

                             5450      5460      5470      5480      5490      5500      5510      5520
KRS02083 Ia          : ATCTGGGGAAGAACGAGAATCAATCTTAGATATTTGCCGTAACGCAAATGTTCACGTAAACAATATGCCTAGTATTGAGA
NUF1003 Ib           : ATCTGGGGAAGAACGAGAATCAATCTTAGATATCTGCCGTAACGCAAACGTTCACGTAAACAATATGCCTAGTATTGAGA
NUF1071 Ic           : ATCTGGGGAAGAACGAGAATCAATCTTAGATATCTGCCGTAACGCAAACGTTCACGTAAACAATATGCCTAGTATTGAGA
2007-1 NT ClusterI   : ATCTGGGGAAGAACGAGAATCAATCTTAGATATCTGCCGTAACGCAAACGTTCACGTAAACAATATGCCTAGTATTGAGA
NUF1032 II           : ATCTGGGGAAGAACGAGAATCAATCTTAGATATTTGCCGTAACGCAAATGTTCACGTAAACAATATGCCTAGTATTGAGA
NUF1095 NT ClusterIII: ATCTGGGGAAGAACGAGAATCAATCTTAGATATTTGCCGTAACGCAAATGTTCACGTAAACAATATGCCTAGTATTGAGA

                             5530      5540      5550      5560      5570      5580      5590      5600
KRS02083 Ia          : ATATCGTTCTAGGTAATGTGTCACTTAATAAATTTAAAGAAATTGAGATTGCTGACTTACTTGGACGAAAGGAAGTAGTA
NUF1003 Ib           : ATATCGTTCTAGGTAATGTGTCACTTAATAAATTTAAAGAAATTGAGATTGCTGACTTACTTGGACGAAAGGAAGTAGTA
NUF1071 Ic           : ATATCGTTCTAGGTAATGTGTCACTTAATAAATTTAAAGAAATTGAGATTGCTGACTTACTTGGACGAAAGGAAGTAGTA
2007-1 NT ClusterI   : ATATCGTTCTAGGTAATGTGTCACTTAATAAATTTAAAGAAATTGAGATTGCTGACTTACTTGGACGAAAGGAAGTAGTA
NUF1032 II           : ATATCGTTCTAGGTAATGTGTCACTTAATAAATTTAAAGAAATTGAGATTGCTGACTTACTTGGACGAAAGGAAGTAGTA
NUF1095 NT ClusterIII: ATATCGTTCTAGGTAATGTGTCACTTAATAAATTTAAAGAAATTGAGATTGCTGACTTACTTGGACGAAAGGAAGTAGTA
```

```
                             5610      5620      5630      5640      5650      5660      5670      5680
KRS02083 Ia         : TTAGATCAAACGTCTTTAAATTCATTCTTTAACGGGAAAACAGTTCTTGTAACTGGTGCAGGAGGATCAATTGGTTCAGA
NUF1003 Ib          : TTAGATCAAACGTCTTTAAATTCATTCTTTAACGGGAAAACAGTTCTTGTAACTGGTGCAGGAGGATCAATTGGTTCAGA
NUF1071 Ic          : TTAGATCAAACGTCTTTAAATTCATTCTTTAACGGGAAAACAGTTCTTGTAACTGGTGCAGGAGGATCAATTGGTTCAGA
2007-1 NT ClusterI  : TTAGATCAAACGTCTTTAAATTCATTCTTTAACGGGAAAACAGTTCTTGTAACTGGTGCAGGAGGATCAATTGGTTCAGA
NUF1032 II          : TTAGATCAAACGTCTTTAAATTCATTCTTTAACGGGAAAACAGTTCTTGTAACTGGTGCAGGAGGATCAATTGGTTCAGA
NUF1095 NT ClusterIII: TTAGATCAAACGTCTTTAAATTCATTCTTTAACGGGAAAACAGTTCTTGTAACTGGTGCAGGAGGATCAATTGGTTCAGA


                             5690      5700      5710      5720      5730      5740      5750      5760
KRS02083 Ia         : AATCTGTCGTCAAGTTTCTAAATTTAATCCAGCACGCATTTTACTTTTAGGGCATGGTGAAAATTCCATTTATCTGATTC
NUF1003 Ib          : AATCTGTCGTCAAGTTTCTAAATTTAATCCAGCACGCATTTTACTTTTAGGGCATGGTGAAAATTCCATTTATCTGATTC
NUF1071 Ic          : AATCTGTCGTCAAGTTTCTAAATTTAATCCAGCACGCATTTTACTTTTAGGGCATGGTGAAAATTCCATTTATCTGATTC
2007-1 NT ClusterI  : AATCTGTCGTCAAGTTTCTAAATTTAATCCAGCACGCATTTTACTTTTAGGGCATGGTGAAAATTCCATTTATCTGATTC
NUF1032 II          : AATCTGTCGTCAAGTTTCTAAATTTAATCCAGCACGCATTTTACTTTTAGGGCATGGTGAAAATTCCATTTATCTGATTC
NUF1095 NT ClusterIII: AATCTGTCGTCAAGTTTCTAAATTTAATCCAGCACGCATTTTACTTTTAGGGCATGGTGAAAATTCCATTTATCTGATTC


                             5770      5780      5790      5800      5810      5820      5830      5840
KRS02083 Ia         : ATCGTGAATTATCAGCACTGTTAAAGGGACGAATTGACATTGTCCCAATTATCGCGGACATTCAAGATCGAGACTTGATT
NUF1003 Ib          : ATCGTGAATTATCAGCACTGTTAAAGGGACGAATTGACATTGTCCCAATTATCGCGGACATTCAAGATCGAGACTTGATT
NUF1071 Ic          : ATCGTGAATTATCAGCACTGTTAAAGGGACGAATTGACATTGTCCCAATTATCGCGGACATTCAAGATCGAGACTTGATT
2007-1 NT ClusterI  : ATCGTGAATTATCAGCACTGTTAAAGGGACGAATTGACATTGTCCCAATTATCGCGGACATTCAAGATCGAGACTTGATT
NUF1032 II          : ATCGTGAATTATCAGCACTGTTAAAGGGACGAATTGACATTGTCCCAATTATCGCGGACATTCAAGATCGAGACTTGATT
NUF1095 NT ClusterIII: ATCGTGAATTATCAGCACTGTTAAAGGGACGAATTGACATTGTCCCAATTATCGCGGACATTCAAGATCGAGACTTGATT


                             5850      5860      5870      5880      5890      5900      5910      5920
KRS02083 Ia         : TTTGAAATTATGGCGAATTATCGACCTGATATAGTCTATCATGCTGCAGCACATAAACATGTGCCATTGATGGAATACAA
NUF1003 Ib          : TTTGAAATTATGGCGAATTATCGACCTGATATAGTCTATCATGCTGCAGCACATAAACATGTGCCATTGATGGAATACAA
NUF1071 Ic          : TTTGAAATTATGGCGAATTATCGACCTGATATAGTCTATCATGCTGCAGCACATAAACATGTGCCATTGATGGAATACAA
2007-1 NT ClusterI  : TTTGAAATTATGGCGAATTATCGACCTGATATAGTCTATCATGCTGCAGCACATAAACATGTGCCATTGATGGAATACAA
NUF1032 II          : TTTGAAATTATGGCGAATTATCGACCTGATATAGTCTATCATGCTGCAGCACATAAACATGTGCCATTGATGGAATACAA
NUF1095 NT ClusterIII: TTTGAAATTATGGCGAATTATCGACCTGATATAGTCTATCATGCTGCAGCACATAAACATGTGCCATTGATGGAATACAA


                             5930      5940      5950      5960      5970      5980      5990      6000
KRS02083 Ia         : TCCAAAAGAAGCTGTTAAAAATAATATCTTTGGGACAAAAAATGTGGCGGAAGCAGCCAAGGCTGCTGGGATTCCTAAAT
NUF1003 Ib          : TCCAAAAGAAGCTGTTAAAAATAATATCTTTGGGACAAAAAATGTGGCGGAAGCAGCCAAGGCTGCTGGGATTCCTAAAT
NUF1071 Ic          : TCCAAAAGAAGCTGTTAAAAATAATATCTTTGGGACAAAAAATGTGGCGGAAGCAGCCAAGGCTGCTGGGATTCCTAAAT
2007-1 NT ClusterI  : TCCAAAAGAAGCTGTTAAAAATAATATCTTTGGGACAAAAAATGTGGCGGAAGCAGCCAAGGCTGCTGGGATTCCTAAAT
NUF1032 II          : TCCAAAAGAAGCTGTTAAAAATAATATCTTTGGGACAAAAAATGTGGCGGAAGCAGCCAAGGCTGCTGGGATTCCTAAAT
NUF1095 NT ClusterIII: TCCAAAAGAAGCTGTTAAAAATAATATCTTTGGGACAAAAAATGTGGCGGAAGCAGCCAAGGCTGCTGGGATTCCTAAAT


                             6010      6020      6030      6040      6050      6060      6070      6080
KRS02083 Ia         : TTATCATGGTCTCTACTGACAAAGCAGTTAATCCGCCTAATGTGATGGGTGCAACTAAGCGCTTCGCTGAAATGATCGTC
NUF1003 Ib          : TTATCATGGTCTCAACTGACAAAGCTGTTAACCCACCTAATGTGATGGGTGCAACTAAGCGCTTCGCTGAAATGATCGTC
NUF1071 Ic          : TTATCATGGTCTCAACTGACAAAGCTGTTAACCCACCTAATGTGATGGGTGCAACTAAGCGCTTCGCTGAAATGATCGTC
2007-1 NT ClusterI  : TTATCATGGTCTCAACTGACAAAGCTGTTAACCCACCTAATGTGATGGGTGCAACTAAGCGCTTCGCTGAAATGATCGTC
NUF1032 II          : TTATCATGGTCTCTACTGACAAAGCAGTTAATCCGCCTAATGTGATGGGTGCAACTAAGCGCTTCGCTGAAATGATCGTC
NUF1095 NT ClusterIII: TTATCATGGTCTCTACTGACAAAGCAGTTAATCCGCCTAATGTGATGGGTGCAACTAAGCGCTTCGCTGAAATGATCGTC


                             6090      6100      6110      6120      6130      6140      6150      6160
KRS02083 Ia         : ACTGGCTTGAATGAAGAAGGCAAAACTCAGTTTGCAGCTGTGCGATTCGGTAACGTCCTTGGTAGTCGTGGTAGTGTGGT
NUF1003 Ib          : ACTGGCTTGAATGAAGAAGGTAAAACTCAGTTTGCAGCTGTGCGATTCGGTAACGTCCTTGGTAGTCGTGGTAGTGTGGT
NUF1071 Ic          : ACTGGCTTGAATGAAGAAGGTAAAACTCAGTTTGCAGCTGTGCGATTCGGTAACGTCCTTGGTAGTCGTGGTAGTGTGGT
2007-1 NT ClusterI  : ACTGGCTTGAATGAAGAAGGTAAAACTCAGTTTGCAGCTGTGCGATTCGGTAACGTCCTTGGTAGTCGTGGTAGTGTGGT
NUF1032 II          : ACTGGCTTGAATGAAGAAGGCAAAACTCAGTTTGCAGCTGTGCGATTCGGTAACGTCCTTGGTAGTCGTGGTAGTGTGGT
NUF1095 NT ClusterIII: ACTGGCTTGAATGAAGAAGGCAAAACTCAGTTTGCAGCTGTGCGATTCGGTAACGTCCTTGGTAGTCGTGGTAGTGTGGT
```

```
                             6170      6180      6190      6200      6210      6220      6230      6240
KRS02083 Ia          : TCCTCTCTTTAAAGAGCAAATCAAAAAAGGAGGACCTATCACAGTAACAGACTTTAGAATGACCCGTTATTTCATGACCA
NUF1003 Ib           : TCCTCTCTTTAAAGAGCAAATCAAAAAAGGAGGACCTATCACAGTAACAGACTTTAGAATGACCCGTTATTTCATGACCA
NUF1071 Ic           : TCCTCTCTTTAAAGAGCAAATCAAAAAAGGAGGACCTATCACAGTAACAGACTTTAGAATGACCCGTTATTTCATGACCA
2007-1 NT ClusterI   : TCCTCTCTTTAAAGAGCAAATCAAAAAAGGAGGACCTATCACAGTAACAGACTTTAGAATGACCCGTTATTTCATGACCA
NUF1032 II           : TCCTCTCTTTAAAGAGCAAATCAAAAAAGGAGGACCTATCACAGTAACAGACTTTAGAATGACCCGTTATTTCATGACCA
NUF1095 NT ClusterIII: TCCTCTCTTTAAAGAGCAAATCAAAAAAGGAGGACCTATCACAGTAACAGACTTTAGAATGACCCGTTATTTCATGACCA


                             6250      6260      6270      6280      6290      6300      6310      6320
KRS02083 Ia          : TTCCGGAAGCTAGTCGTCTTGTCATTCAAGCAGGATTTTTGGCTAAAGGTGGAGAGGTCTTTGTCCTTGATATGGGTGAA
NUF1003 Ib           : TTCCGGAAGCTAGTCGTCTTGTCATTCAAGCAGGATTTTTGGCTAAAGGTGGAGAGGTCTTTGTCCTTGATATGGGTGAA
NUF1071 Ic           : TTCCGGAAGCTAGTCGTCTTGTCATTCAAGCAGGATTTTTGGCTAAAGGTGGAGAGGTCTTTGTCCTTGATATGGGTGAA
2007-1 NT ClusterI   : TTCCGGAAGCTAGTCGTCTTGTCATTCAAGCAGGATTTTTGGCTAAAGGTGGAGAGGTCTTTGTCCTTGATATGGGTGAA
NUF1032 II           : TTCCGGAAGCTAGTCGTCTTGTCATTCAAGCAGGATTTTTGGCTAAAGGTGGAGAGGTCTTTGTCCTTGATATGGGTGAA
NUF1095 NT ClusterIII: TTCCGGAAGCTAGTCGTCTTGTCATTCAAGCAGGATTTTTGGCTAAAGGTGGAGAGGTCTTTGTCCTTGATATGGGTGAA


                             6330      6340      6350      6360      6370      6380      6390      6400
KRS02083 Ia          : CCAGTCAAAATTTTAGACCTTGCGAAAAAAGTAATCAAATTGAGTGGTCACACAGAAGAAGATATTAAAATTGTAGAATC
NUF1003 Ib           : CCAGTCAAAATTTTAGACCTTGCGAAAAAAGTAATCAAATTGAGTGGTCACACAGAAGAAGATATTAAAATTGTAGAATC
NUF1071 Ic           : CCAGTCAAAATTTTAGACCTTGCGAAAAAAGTAATCAAATTGAGTGGTCACACAGAAGAAGATATTAAAATTGTAGAATC
2007-1 NT ClusterI   : CCAGTCAAAATTTTAGACCTTGCGAAAAAAGTAATCAAATTGAGTGGTCACACAGAAGAAGATATTAAAATTGTAGAATC
NUF1032 II           : CCAGTCAAAATTTTAGACCTTGCGAAAAAAGTAATCAAATTGAGTGGTCACACAGAAGAAGATATTAAAATTGTAGAATC
NUF1095 NT ClusterIII: CCAGTCAAAATTTTAGACCTTGCGAAAAAAGTAATCAAATTGAGTGGTCACACAGAAGAAGATATTAAAATTGTAGAATC


                             6410      6420      6430      6440      6450      6460      6470      6480
KRS02083 Ia          : AGGAATCAGACCTGGTGAAAAACTTTATGAAGAATTGTTATCTTCAAATGAACGAGTCAGTGAACAAATACATGATAAAA
NUF1003 Ib           : AGGAATCAGACCTGGTGAAAAACTTTATGAAGAATTGTTATCTTCAAATGAACGAGTCAGTGAACAAATACATGATAAAA
NUF1071 Ic           : AGGAATCAGACCTGGTGAAAAACTTTATGAAGAATTGTTATCTTCAAATGAACGAGTCAGTGAACAAATACATGATAAAA
2007-1 NT ClusterI   : AGGAATCAGACCTGGTGAAAAACTTTATGAAGAATTGTTATCTTCAAATGAACGAGTCAGTGAACAAATACATGATAAAA
NUF1032 II           : AGGAATCAGACCTGGTGAAAAACTTTATGAAGAATTGTTATCTTCAAATGAACGAGTCAGTGAACAAATACATGATAAAA
NUF1095 NT ClusterIII: AGGAATCAGACCTGGTGAAAAACTTTATGAAGAATTGTTATCTTCAAATGAACGAGTCAGTGAACAAATACATGATAAAA


                             6490      6500      6510      6520      6530      6540      6550      6560
KRS02083 Ia          : TTTTTGTTGGTAAAGTTACTAGTAAACCACTAGCAGAAGTTGAAGCTATCATTTCAAACTTAGATCAACTGCCAACAGAC
NUF1003 Ib           : TTTTTGTTGGTAAAGTTACTAGTAAACCACTAGCAGAAGTTGAAGCTATTATTTCAAACCTAGATCAACTTCCAACAGAC
NUF1071 Ic           : TTTTTGTTGGTAAAGTTACTAGTAAACCACTAGCAGAAGTTGAAGCTATTATTTCAAACCTAGATCAACTTCCAACAGAC
2007-1 NT ClusterI   : TTTTTGTTGGTAAAGTTACTAGTAAACCACTAGCAGAAGTTGAAGCTATTATTTCAAACCTAGATCAACTTCCAACAGAC
NUF1032 II           : TTTTTGTTGGTAAAGTTACTAGTAAACCACTAGCAGAAGTTGAAGCTATCATTTCAAACTTAGATCAACTGCCAACAGAC
NUF1095 NT ClusterIII: TTTTTGTTGGTAAAGTTACTAGTAAACCACTAGCAGAAGTTGAAGCTATCATTTCAAACTTAGATCAACTGCCAACAGAC


                             6570      6580      6590      6600
KRS02083 Ia          : AATTTGAAAGAAACATTGATTCACTTTGCAAAACAGGAGTAG
NUF1003 Ib           : AATTTAAAAGAAACATTGATTCACTTTGCAAAACAGGAGTAG
NUF1071 Ic           : AATTTAAAAGAAACATTGATTCACTTTGCAAAACAGGAGTAG
2007-1 NT ClusterI   : AATTTAAAAGAAACATTGATTCACTTTGCAAAACAGGAGTAG
NUF1032 II           : AATTTGAAAGAAACATTGATTCACTTTGCAAAACAGGAGTAG
NUF1095 NT ClusterIII: AATTTGAAAGAAACATTGATTCACTTTGCAAAACAGGAGTAG
```

S1.2 Multiple amino acid sequence alignment of LysR, CpsA-E in six representative strains.

LysR

```
                                10        20        30        40        50        60        70        80
KRS02083 Ia           : MRLQQLHYIIKVVECGSMNVAAKQLFITQPSLSNAVKDLENEMGITIFNRNPKGITLTKDGVEFLSYARQVVEQTSLLED
NUF1003 Ib            : MRLQQLHYIIKVVECGSMNVAAKQLFITQPSLSNAVKDLENEMGITIFNRNPKGITLTKDGVEFLSYARQVVEQTSLLED
NUF1071 Ic            : MRLQQLHYIIKVVECGSMNVAAKQLFITQPSLSNAVKDLENEMGITIFNRNPKGITLTKDGVEFLSYARQVVEQTSLLED
2007-1 NT ClusterI    : MRLQQLHYIIKVVECGSMNVAAKQLFITQPSLSNAVKDLENEMGITIFNRNPKGITLTKDGVEFLSYARQVVEQTSLLED
NUF1032 II            : MRLQQLHYIIKVVECGSMNVAAKQLFITQPSLSNAVKDLENEMGITIFNRNPKGITLTKDGVEFLSYARQVVEQTSLLED
NUF1095 NT ClusterIII : MRLQQLHYIIKVVECGSMNVAAKQLFITQPSLSNAVKDLENEMGITIFNRNPKGITLTKDGVEFLSYARQVVEQTSLLED


                                90       100       110       120       130       140       150       160
KRS02083 Ia           : RYKNLNSKRELFSVSSQHYAFVVNAFVSLLKETDMTQYELFLRETRTWEILDDVKNFRSEIGVLFLNDYNRDVLSKMMED
NUF1003 Ib            : RYKNLNSKRELFSVSSQHYAFVVNAFVSLLKETDMTQYELFLRETRTWEILDDVKNFRSEIGVLFLNDYNRDVLSKMMED
NUF1071 Ic            : RYKNLNSKRELFSVSSQHYAFVVNAFVSLLKETDMTQYELFLRETRTWEILDDVKNFRSEIGVLFLNDYNRDVLSKMMED
2007-1 NT ClusterI    : RYKNLNSKRELFSVSSQHYAFVVNAFVSLLKETDMTQYELFLRETRTWEILDDVKNFRSEIGVLFLNDYNRDVLSKMMED
NUF1032 II            : RYKNLNSKRELFSVSSQHYAFVVNAFVSLLKETDMTQYELFLRETRTWEILDDVKNFRSEIGVLFLNDYNRDVLSKMMED
NUF1095 NT ClusterIII : RYKNLNSKRELFSVSSQHYAFVVNAFVSLLKETDMTQYELFLRETRTWEILDDVKNFRSEIGVLFLNDYNRDVLSKMMED


                               170       180       190       200       210       220       230       240
KRS02083 Ia           : DHLIATPLFKAHPHIFISRNHPLAKHTRLTFNDLKPYPYLSYDQGLHNSFYFSEEMLSEISHPKSIVVSDRATLFNLMIG
NUF1003 Ib            : DHLIATPLFKAHPHIFISRNHPLAKHTRLTFNDLKPYPYLSYDQGLHNSFYFSEEMLSEISHPKSIVVSDRATLFNLMIG
NUF1071 Ic            : DHLIATPLFKAHPHIFISRNHPLAKHTRLTFNDLKPYPYLSYDQGLHNSFYFSEEMLSEISHPKSIVVSDRATLFNLMIG
2007-1 NT ClusterI    : DHLIATPLFKAHPHIFISRNHPLAKHTRLTFNDLKPYPYLSYDQGLHNSFYFSEEMLSEISHPKSIVVSDRATLFNLMIG
NUF1032 II            : DHLIATPLFKAHPHIFISRNHPLAKHTRLTFNDLKPYPYLSYDQGLHNSFYFSEEMLSEISHPKSIVVSDRATLFNLMIG
NUF1095 NT ClusterIII : DHLIATPLFKAHPHIFISRNHPLAKHTRLTFNDLKPYPYLSYDQGLHNSFYFSEEMLSEISHPKSIVVSDRATLFNLMIG


                               250       260       270       280       290       300
KRS02083 Ia           : LDGYTIASGILNSNLNGDGIVAIPLDVPDIIDIVYITHEKANLSKMGQAFIDYLIKEVQF
NUF1003 Ib            : LDGYTIASGILNSNLNGDGIVAIPLDVPDIIDIVYITHEKANLSKMGQAFIDYLIKEVQF
NUF1071 Ic            : LDGYTIASGILNSNLNGDGIVAIPLDVPDIIDIVYITHEKANLSKMGQAFIDYLIKEVQF
2007-1 NT ClusterI    : LDGYTIASGILNSNLNGDGIVAIPLDVPDIIDIVYITHEKANLSKMGQAFIDYLIKEVQF
NUF1032 II            : LDGYTIASGILNSNLNGDGIVAIPLDVPDIIDIVYITHEKANLSKMGQAFIDYLIKEVQF
NUF1095 NT ClusterIII : LDGYTIASGILNSNLNGDGIVAIPLDVPDIIDIVYITHEKANLSKMGQAFIDYLIKEVQF
```

CpsA

```
                                10        20        30        40        50        60        70        80
KRS02083 Ia           : MAAKSRSRRGSTSNGSRSFSVINLALLLLFTVLSLIITFLMYTYNFLAFHHFNLIVAGILFLLFLLWLILIIRKRARFLS
NUF1003 Ib            : MAAKSRSRRGSTSNGSRSFSVINLALLLLFTVLSLIITFLMYTYNFLAFHHFNLIVAGILFLLFLLWLILIIRKRARFLS
NUF1071 Ic            : MAAKSRSRRGSTSNGSRSFSVINLALLLLFTVLSLIITFLMYTYNFLAFHHFNLIVAGILFLLFLLWLILIIRKRARFLS
2007-1 NT ClusterI    : MAAKSRSRRGSTSNGSRSFSVINLALLLLFTVLSLIITFLMYTYNFLAFHHFNLIVAGILFLLFLLWLILIIRKRARFLS
NUF1032 II            : MAAKSRSRRGSTSNGSRSFSVINLALLLLFTVLSLIITFLMYTYNFLAFHHFNLIVAGILFLLFLLWLILIIRKRARFLS
NUF1095 NT ClusterIII : MAAKSRSRRGSTSNGSRSFSVINLALLLLFTVLSLIITFLMYTYNFLAFHHFNLIVAGILFLLFLLWLILIIRKRARFLS


                                90       100       110       120       130       140       150       160
KRS02083 Ia           : MITLIIANIILAITLFAFKQTIDFTAQMNKTASFSEVEMLVVVPKDSSIQSVSELKEVDAPTDMDASNIHSLLKRIKEDK
NUF1003 Ib            : MITLIIANIILAITLFAFKQTIDFTAQMNKTASFSEVEMSVVVPKDSSIQSVSELKEVDAPTDMDASNIHSLLKRIKEDK
NUF1071 Ic            : MITLIIANIILAITLFAFKQTIDFTAQMNKTASFSEVEMSVVVPKDSSIQSVSELKEVDAPTDMDASNIHSLLKRIKEDK
2007-1 NT ClusterI    : MITLIIANIILAITLFAFKQTIDFTAQMNKTASFSEVEMSVVVPKDSSIQSVSELKEVDAPTDMDASNIHSLLKRIKEDK
NUF1032 II            : MITLIIANIILAITLFAFKQTIDFTAQMNKTASFSEVEMSVVVPKDSSIQSVSELKEVDAPTDMDASNIHSLLKRIKEDK
NUF1095 NT ClusterIII : MITLIIANIILAITLFAFKQTIDFTAQMNKTASFSEVEMSVVVPKDSSIQSVSELKEVDAPTDMDASNIHSLLKRIKEDK


                               170       180       190       200       210       220       230       240
KRS02083 Ia           : KIDLASNNVASYKEAYDKMMIGSSKAMVLNSAYGSLIEQSDANYRDKVKTIYTYKVKKAIKSSNKLSSNKDSFNVYISGI
NUF1003 Ib            : KIDLASNNVASYKEAYDKMMIGSSKAMVLNSAYGSLIEQSDANYRDKVKTIYTYKVKKAIKSSNKLSSNKDSFNVYISGI
NUF1071 Ic            : KIDLASNNVASYKEAYDKMMIGSSKAMVLNSAYGSLIEQSDANYRDKVKTIYTYKVKKAIKSSNKLSSNKDSFNVYISGI
2007-1 NT ClusterI    : KIDLASNNVASYKEAYDKMMIGSSKAMVLNSAYGSLIEQSDANYRDKVKTIYTYKVKKAIKSSNKLSSNKDSFNVYISGI
NUF1032 II            : KIDLASNNVASYKEAYDKMMIGSSKAMVLNSAYGSLIEQSDANYRDKVKTIYTYKVKKAIKSSNKLSSNKDSFNVYISGI
NUF1095 NT ClusterIII : KIDLASNNVASYKEAYDKMMIGSSKAMVLNSAYGSLIEQSDANYRDKVKTIYTYKVKKAIKSSNKLSSNKDSFNVYISGI
```

```
                               250       260       270       280       290       300       310       320
KRS02083 Ia           : DTYGAISTVSRSDVNIILTANTKTHKVLLTTTPRDSYVKIPGGGGDQYDKLTHAGIYGVETSMATLESLYDIKINNYARI
NUF1003 Ib            : DTYGAISTVSRSDVNIILTANTKTHKVLLTTTPRDSYVKIPGGGGDQYDKLTHAGIYGVETSMATLESLYDININNYARI
NUF1071 Ic            : DTYGAISTVSRSDVNIILTANTKTHKVLLTTTPRDSYVKIPGGGGDQYDKLTHAGIYGVETSMATLESLYDININNYARI
2007-1 NT ClusterI    : DTYGAISTVSRSDVNIILTANTKTHKVLLTTTPRDSYVKIPGGGGDQYDKLTHAGIYGVETSMATLESLYDININNYARI
NUF1032 II            : DTYGAISTVSRSDVNIILTANTKTHKVLLTTTPRDSYVKIPGGGGDQYDKLTHAGIYGVETSMATLESLYDININNYARI
NUF1095 NT ClusterIII : DTYGAISTVSRSDVNIILTANTKTHKVLLTTTPRDSYVKIPGGGGDQYDKLTHAGIYGVETSMATLESLYDININNYARI

                               330       340       350       360       370       380       390       400
KRS02083 Ia           : NFSTFMDLIDLLGGIEVNNDQAFSANGYDFPQGRIALNSKQALTFVRERHSLQGGDNDRGKNQEKVISAVIQKLSTIKSP
NUF1003 Ib            : NFSTFMDLIDLLGGIEVNNDQAFSANGYDFPQGRIALNSKQALTFVRERHSLQGGDNDRGKNQEKVISAVIQKLSTIKSP
NUF1071 Ic            : NFSTFMDLIDLLGGIEVNNDQAFSANGYDFPQGRIALNSKQALTFVRERHSLQGGDNDRGKNQEKVISAVIQKLSTIKSP
2007-1 NT ClusterI    : NFSTFMDLIDLLGGIEVNNDQAFSANGYDFPQGRIALNSKQALTFVRERHSLQGGDNDRGKNQEKVISAVIQKLSTIKSP
NUF1032 II            : NFSTFMDLIDLLGGIEVNNDQAFSANGYDFPQGRIALNSKQALTFVRERHSLQGGDNDRGKNQEKVISAVIQKLSTIKSP
NUF1095 NT ClusterIII : NFSTFMDLIDLLGGIEVNNDQAFSANGYDFPQGRIALNSKQALTFVRERHSLQGGDNDRGKNQEKVISAVIQKLSTIKSP

                               410       420       430       440       450       460       470       480
KRS02083 Ia           : AQFTSIVTGLQNSIQTNLSLNQLMTLANSQVADKSSYTVTSQDVTGSGSTGELPSYAMPGSALYMLKLDDQSVETAKEAI
NUF1003 Ib            : AQFTSIVTGLQNSIQTNLSLNQLMTLANSQVADKSSYTVTSQDVTGSGSTGELPSYAMPGSALYMLKLDDQSVETAKEAI
NUF1071 Ic            : AQFTSIVTGLQNSIQTNLSLNQLMTLANSQVADKSSYTVTSQDVTGSGSTGELPSYAMPGSALYMLKLDDQSVETAKEAI
2007-1 NT ClusterI    : AQFTSIVTGLQNSIQTNLSLNQLMTLANSQVADKSSYTVTSQDVTGSGSTGELPSYAMPGSALYMLKLDDQSVETAKEAI
NUF1032 II            : AQFTSIVTGLQNSIQTNLSLNQLMTLANSQVADKSSYTVTSQDVTGSGSTGELPSYAMPGSALYMLKLDDQSVETAKEAI
NUF1095 NT ClusterIII : AQFTSIVTGLQNSIQTNLSLNQLMTLANSQVADKSSYTVTSQDVTGSGSTGELPSYAMPGSALYMLKLDDQSVETAKEAI

KRS02083 Ia           : KNTMEGN
NUF1003 Ib            : KNTMEGN
NUF1071 Ic            : KNTMEGN
2007-1 NT ClusterI    : KNTMEGN
NUF1032 II            : KNTMEGN
NUF1095 NT ClusterIII : KNTMEGN
```

CpsB

```
                                10        20        30        40        50        60        70        80
KRS02083 Ia           : MIDIHSHIIFDVDDGPLTIDESLSLLEESYKQGVRTIVSTSHRRKGMFETPEDDILNKFKHVKREAADKFPDLTLLYGGE
NUF1003 Ib            : MIDIHSHIIFDVDDGPLTIDESLSLLEESYKQGVRTIVSTSHRRKGMFETPEDDILNKFKHVKREAADKFPDLTLLYGGE
NUF1071 Ic            : MIDIHSHIIFDVDDGPLTIDESLSLLEESYKQGVRTIVSTSHRRKGMFETPEDDILNKFKHVKREAADKFPDLTLLYGGE
2007-1 NT ClusterI    : MIDIHSHIIFDVDDGPLTIDESLSLLEESYKQGVRTIVSTSHRRKGMFETPEDDILNKFKHVKREAADKFPDLTLLYGGE
NUF1032 II            : MIDIHSHIIFDVDDGPLTIDESLSLLEESYKQGVRTIVSTSHRRKGMFETPEDDILNKFKHVKREAADKFPDLTLLYGGE
NUF1095 NT ClusterIII : MIDIHSHIIFDVDDGPLTIDESLSLLEESYKQGVRTIVSTSHRRKGMFETPEDDILNKFKHVKREAADKFPDLTLLYGGE

                                90       100       110       120       130       140       150       160
KRS02083 Ia           : LYFTADILKKLENNEIPRMNDTRFALIEFSQMTPWKDIHLALSQVLMLGITPIVAHIERYAALEFNGDRVQELINMGCYT
NUF1003 Ib            : LYFTADILKKLENNEIPRMNDTRFALIEFSQMTPWKDIHLALSQVLMLGITPIVAHIERYAALEFNGDRVQELINMGCYT
NUF1071 Ic            : LYFTADILKKLENNEIPRMNDTRFALIEFSQMTPWKDIHLALSQVLMLGITPIVAHIERYAALEFNGDRVQELINMGCYT
2007-1 NT ClusterI    : LYFTADILKKLENNEIPRMNDTRFALIEFSQMTPWKDIHLALSQVLMLGITPIVAHIERYAALEFNGDRVQELINMGCYT
NUF1032 II            : LYFTADILKKLENNEIPRMNDTRFALIEFSQMTPWKDIHLALSQVLMLGITPIVAHIERYAALEFNGDRVQELINMGCYT
NUF1095 NT ClusterIII : LYFTADILKKLENNEIPRMNDTRFALIEFSQMTPWKDIHLALSQVLMLGITPIVAHIERYAALEFNGDRVQELINMGCYT

                               170       180       190       200       210       220       230       240
KRS02083 Ia           : QVNSAHVLKAKLFGDKLKNFKKRAKYFLDKDLVHCISSDMHNLKKRPPYMKQAYQIVEKDYGTRRARKLFEENAASLINN
NUF1003 Ib            : QVNSAHVLKAKLFGDKLKIFKKRTKYFLDKDLVHCISSDMHNLKKRPPYMKQAYQIVEKDYGTRRARKLFEENAASLINN
NUF1071 Ic            : QVNSAHVLKAKLFGDKLKIFKKRTKYFLDKDLVHCISSDMHNLKKRPPYMKQAYQIVEKDYGTRRARKLFEENAASLINN
2007-1 NT ClusterI    : QVNSAHVLKAKLFGDKLKIFKKRTKYFLDKDLVHCISSDMHNLKKRPPYMKQAYQIVEKDYGTRRARKLFEENAASLINN
NUF1032 II            : QVNSAHVLKAKLFGDKLKIFKKRAKYFLDKDLVHCISSDMHNLKKRPPYMKQAYQIVEKDYGTRRARKLFEENAASLINN
NUF1095 NT ClusterIII : QVNSAHVLKAKLFGDKLKIFKKRAKYFLDKDLVHCISSDMHNLKKRPPYMKQAYQIVEKDYGTRRARKLFEENAASLINN
```

```
KRS02083 Ia           : EFL
NUF1003 Ib            : EFL
NUF1071 Ic            : EFL
2007-1 NT ClusterI    : EFL
NUF1032 II            : EFL
NUF1095 NT ClusterIII : EFL
```

CpsC

```
                                10        20        30        40        50        60        70        80
KRS02083 Ia           : MNNMEKPSMEIDVLSLLKKLWTKKFLIIFMALFFGTLALLSSIFLIKPSYTASTRIYVLNKTQQADNLSATDLQAGGLLV
NUF1003 Ib            : MNNMEKPSMEIDVLSLLKKLWTKKFLIIFMALFFGTLALLSSIFLIKPSYTASTRIYVLNKTQQADNLSATDLQAGGLLV
NUF1071 Ic            : MNNMEKPSMEIDVLSLLKKLWTKKFLIIFMALFFGTLALLSSIFLIKPSYTASTRIYVLNKTQQADNLSATDLQAGGLLV
2007-1 NT ClusterI    : MNNMEKPSMEIDVLSLLKKLWTKKFLIIFMALFFWTLALLSSIFLIKPSYTASTRIYVLNKTQQADNLSATDLQAGGLLV
NUF1032 II            : MNNMEKPSMEIDVLSLLKKLWTKKFLIIFMALFFGTLALLSSIFLIKPSYTASTRIYVLNKTQQADNLSATDLQAGGLLV
NUF1095 NT ClusterIII : MNNMEKPSMEIDVLSLLKKLWTKKFLIIFMALFFGTLALLSSIFLIKPSYTASTRIYVLNKTQQADNLSATDLQAGGLLV

                                90       100       110       120       130       140       150       160
KRS02083 Ia           : NDYKEIITSRDVMKDVIANDGVSMTPEELSKMIAVTIPADTRVISIAVTNHDPQAAKDLANSVRDVASEKIKSVTKVQDV
NUF1003 Ib            : NDYKEIITSRDVMKDVIANDGVSMTPEELSKMIAVTIPADTRVISIAVTNHDPQAAKDLANSVRDVASEKIKSVTKVQDV
NUF1071 Ic            : NDYKEIITSRDVMKDVIANDGVSMTPEELSKMIAVTIPADTRVISIAVTNHDPQAAKDLANSVRDVASEKIKSVTKVQDV
2007-1 NT ClusterI    : NDYKEIITSRDVMKDVIANDGVSMTPEELSKMIAVTIPADTRVISIAVTNHDPQAAKDLANSVRDVASEKIKSVTKVQDV
NUF1032 II            : NDYKEIITSRDVMKDVIANDGVSMTPEELSKMIAVTIPADTRVISIAVTNHDPQAAKDLANSVRDVASKKIKSVTKVQNV
NUF1095 NT ClusterIII : NDYKEIITSRDVMKDVIANDGVSMTPEELSKMIAVTIPADTRVISIAVTNHDPQAAKDLANSVRDVASKKIKSVTKVQDV

                               170       180       190       200       210       220       230
KRS02083 Ia           : TPLEKAVMPTSPSSPNIKRNTLVGILFGAFLTIVTVIVGEVVDDRVKRPEDVEELLGMTLLGIVPNTNKM
NUF1003 Ib            : TPLEKAVMPTSPSSPNIKRNTLVGILFGAFLTIVTVIVGEVVDDRVKRPEDVEELLGMTLLGIVPNTNKM
NUF1071 Ic            : TPLEKAVMPTSPSSPNIKRNTLVGILFGAFLTIVTVIVGEVVDDRVKRPEDVEELLGMTLLGIVPNTNKM
2007-1 NT ClusterI    : TPLEKAVMPTSPSSPNIKRNTLVGILFGAFLTIVTVIVGEVVDDRVKRPEDVEELLGMTLLGIVPNTNKM
NUF1032 II            : TPLEKAVMPTSPSSPNIKRNTLVGILFGAFLTIVTVIVGEVVDDRVKRPEDVEELLGMTLLGIVPNTNKM
NUF1095 NT ClusterIII : TPLEKAVMPTSPSSPNIKRNTLVGILFGAFLTIVTVIVGEVVDDRVKRPEDVEELLGMTLLGIVPNTNKM
```

CpsD

```
                                10        20        30        40        50        60        70        80
KRS02083 Ia           : MARLELVKSKKELYDIAEEYYNSIRTNIQFSGRDLKVITLTSVQPGEGKSTLSANIAISFAKAGLKTLLIDADIRNSVMS
NUF1003 Ib            : MARLELVKSKKELYDIAEEYYNSIRTNIQFSGRDLKVITLTSVQPGEGKSTTSANIAISFAKAGLKTLLIDADIRNSVMS
NUF1071 Ic            : MARLELVKSKKELYDIAEEYYNSIRTNIQFSGRDLKVITLTSVQPGEGKSTTSANIAISFAKAGLKTLLIDADIRNSVMS
2007-1 NT ClusterI    : MARLELVKSKKELYDIAEEYYNSIRTNIQFSGRDLKVITLTSVQPGEGKSTTSANIAISFAKAGLKTLLIDADIRNSVMS
NUF1032 II            : MARLELVKSKKELYDIAEEYYNSIRTNIQFSGRDLKVITLTSVQPGEGKSTLSANIAISFAKAGLKTLLIDADIRNSVMS
NUF1095 NT ClusterIII : MARLELVKSKKELYDIAEEYYNSIRTNIQFSGRDLKVITLTSVQPGEGKSTLSANIAISFAKAGLKTLLIDADIRNSVMS

                                90       100       110       120       130       140       150       160
KRS02083 Ia           : GTFKADEKYEGLSSYLSGNAELSAVISHTNIENLMLIPAGHVPPNPTTLLQNSNFNFMIDTVKELFDYVIIDTPPIGLVI
NUF1003 Ib            : GTFKADEKYEGLSSYLSGNAELSAVISHTNIENLMLIPAGHVPPNPTTLLQNSNFNFMIDTVKELFDYVIIDTPPIGLVI
NUF1071 Ic            : GTFKADEKYEGLSSYLSGNAELSAVISHTNIENLMLIPAGHVPPNPTTLLQNSNFNFMIDTVKELFDYVIIDTPPIGLVI
2007-1 NT ClusterI    : GTFKADEKYEGLSSYLSGNAELSAVISHTNIENLMLIPAGHVPPNPTTLLQNSNFNFMIDTVKELFDYVIIDTPPIGLVI
NUF1032 II            : GTFKADEKYEGLSSYLSGNAELSAVISHTNIENLMLIPAGHVPPNPTTLLQNSNFNFMIDTVKELFDYVIIDTPPIGLVI
NUF1095 NT ClusterIII : GTFKADEKYEGLSSYLSGNAELSAVISHTNIENLMLIPAGHVPPNPTTLLQNSNFNFMIDTVKELFDYVIIDTPPIGLVI

                               170       180       190       200       210       220       230
KRS02083 Ia           : DSAIISQKADANILVTEAGAIKRRFIQKAKEQMEQSGALFLGVILNKVEETLDSYGGYGSYGAYGNYGKPAKKKSRKRR
NUF1003 Ib            : DSAIISQKADANILVTEAGAIKRRFIQKAKEQMEQSGALFLGVILNKVEETLDSYGGYGSYGANGNYGKPAKKKSRKRR
NUF1071 Ic            : DSAIISQKADANILVTEAGAIKRRFIQKAKEQMEQSGALFLGVILNKVEETLDSYGGYGSYGANGNYGKPAKKKSRKRR
2007-1 NT ClusterI    : DSAIISQKADANILVTEAGAIKRRFIQKAKEQMEQSGALFLGVILNKVEETLDSYGGYGSYGANGNYGKPAKKKSRKRR
NUF1032 II            : DSAIISQKADANILVTEAGAIKRRFIQKAKEQMEQSGALFLGVILNKVEETLDSYGGYGSYGAYGNYGKPAKKKSRKRR
NUF1095 NT ClusterIII : DSAIISQKADANILVTEAGAIKRRFIQKAKEQMEQSGALFLGVILNKVEETLDSYGGYGSYGAYGNYGKPAKKKSRKRR
```

CpsE

```
                                10        20        30        40        50        60        70        80
KRS02083 Ia           : MKRSQKRAILYLIDVFMITFAHLSAFRFLLSYSTRLSDKEIYITLFVTLFVYTVLGIRARIFSIINRFTDYKVIFILIAN
NUF1003 Ib            : MKRSQKRAILYLIDVFMITFANLSAFRFLLSYSTRLSDKEIYITLFVTLFVYTVLGIRARIFSIINRFTDYKVIFILIAN
NUF1071 Ic            : MKRSQKRAILYLIDVFMITFANLSAFRFLLSYSTRLSDKEIYITLFVTLFVYTVLGIRARIFSIINRFTDYKVIFILIAN
2007-1NT ClusterI     : MKRSQKRAILYLIDVFMITFANLSAFRFLLSYSTRLSDKEIYITLFVTLFVYTVLGIRARIFSIINRFTDYKVIFILIAN
NUF1032 II            : MKRSQKRAILYLIDVFMITFAHLSAFRFLLSYSTRLSDKEIYITLFVTLFVYTVLGIRARIFSIINRFTDYKVIFILIAN
NUF1095 NT ClusterIII : MKRSQKRAILYLIDVFMITFAHLSAFRFLLSYSTRLSDKEIYITLFVTLFVYTVLGIRARIFSIINRFTDYKVIFILIAN

                                90       100       110       120       130       140       150       160
KRS02083 Ia           : MFFASLLSYLVDVLFLDTFSRRFLFLSFLFGTFLIILPRMIWRMWHEQNLFVKHNKKDQKTKMLVVGAGEGGSAFIQTIL
NUF1003 Ib            : MFFASLLSYLVDVLFLDTFSRRFLFLSFLFGTFLIILPRMIWRMWHEQNLFVKHNKKDQKTKMLVVGAGEGGSAFIQTIL
NUF1071 Ic            : MFFASLLSYLVDVLFLDTFSRRFLFLSFLFGTFLIILPRMIWRMWHEQNLFVKHNKKDQKTKMLVVGAGEGGSAFIQTIL
2007-1NT ClusterI     : MFFASLLSYLVDVLFLDTFSRRFLFLSFLFGTFLIILPRMIWRMWHEQNLFVKHNKKDQKTKMLVVGAGEGGSAFIQTIL
NUF1032 II            : MFFASLLSYLVDVLFLDTFSRRFLFLSFLFGTFLIILPRMIWRMWHEQNLFVKHNKKDQKTKMLVVGAGEGGSAFIQTIL
NUF1095 NT ClusterIII : MFFASLLSYLVDVLFLDTFSRRFLFLSFLFGTFLIILPRMIWRMWHEQNLFVKHNKKDQKTKMLVVGAGEGGSAFIQTIL

                               170       180       190       200       210       220       230       240
KRS02083 Ia           : NKSKDIDIVGIVDADINKLGTYLHGIKVLGNKNSIPRLVAEYEVKQVTIAIPSLSGEERESILDICRNANVHVNNMPSIE
NUF1003 Ib            : NKSKDIDIVGIVDADINKLGTYLHGIKVLGNKNSIPRLLAEYEVKQVTIAIPSLSGEERESILDICRNANVHVNNMPSIE
NUF1071 Ic            : NKSKDIDIVGIVDADINKLGTYLHGIKVLGNKNSIPRLLAEYEVKQVTIAIPSLSGEERESILDICRNANVHVNNMPSIE
2007-1NT ClusterI     : NKSKDIDIVGIVDADINKLGTYLHGIKVLGNKNSIPRLLAEYEVKQVTIAIPSLSGEERESILDICRNANVHVNNMPSIE
NUF1032 II            : NKSKDIDIVGVVDADINKLGTYLHGIKVLGNKNSIPRLVAEYEVKQVTIAIPSLSGEERESILDICRNANVHVNNMPSIE
NUF1095 NT ClusterIII : NKSKDIDIVGVVDADINKLGTYLHGIKVLGNKNSIPRLVAEYEVKQVTIAIPSLSGEERESILDICRNANVHVNNMPSIE

                               250       260       270       280       290       300       310       320
KRS02083 Ia           : NIVLGNVSLNKFKEIEIADLLGRKEVVLDQTSLNSFFNGKTVLVTGAGGSIGSEICRQVSKFNPARILLLGHGENSIYLI
NUF1003 Ib            : NIVLGNVSLNKFKEIEIADLLGRKEVVLDQTSLNSFFNGKTVLVTGAGGSIGSEICRQVSKFNPARILLLGHGENSIYLI
NUF1071 Ic            : NIVLGNVSLNKFKEIEIADLLGRKEVVLDQTSLNSFFNGKTVLVTGAGGSIGSEICRQVSKFNPARILLLGHGENSIYLI
2007-1NT ClusterI     : NIVLGNVSLNKFKEIEIADLLGRKEVVLDQTSLNSFFNGKTVLVTGAGGSIGSEICRQVSKFNPARILLLGHGENSIYLI
NUF1032 II            : NIVLGNVSLNKFKEIEIADLLGRKEVVLDQTSLNSFFNGKTVLVTGAGGSIGSEICRQVSKFNPARILLLGHGENSIYLI
NUF1095 NT ClusterIII : NIVLGNVSLNKFKEIEIADLLGRKEVVLDQTSLNSFFNGKTVLVTGAGGSIGSEICRQVSKFNPARILLLGHGENSIYLI

                               330       340       350       360       370       380       390       400
KRS02083 Ia           : HRELSALLKGRIDIVPIIADIQDRDLIFEIMANYRPDIVYHAAAHKHVPLMEYNPKEAVKNNIFGTKNVAEAAKAAGIPK
NUF1003 Ib            : HRELSALLKGRIDIVPIIADIQDRDLIFEIMANYRPDIVYHAAAHKHVPLMEYNPKEAVKNNIFGTKNVAEAAKAAGIPK
NUF1071 Ic            : HRELSALLKGRIDIVPIIADIQDRDLIFEIMANYRPDIVYHAAAHKHVPLMEYNPKEAVKNNIFGTKNVAEAAKAAGIPK
2007-1NT ClusterI     : HRELSALLKGRIDIVPIIADIQDRDLIFEIMANYRPDIVYHAAAHKHVPLMEYNPKEAVKNNIFGTKNVAEAAKAAGIPK
NUF1032 II            : HRELSALLKGRIDIVPIIADIQDRDLIFEIMANYRPDIVYHAAAHKHVPLMEYNPKEAVKNNIFGTKNVAEAAKAAGIPK
NUF1095 NT ClusterIII : HRELSALLKGRIDIVPIIADIQDRDLIFEIMANYRPDIVYHAAAHKHVPLMEYNPKEAVKNNIFGTKNVAEAAKAAGIPK

                               410       420       430       440       450       460       470       480
KRS02083 Ia           : FIMVSTDKAVNPPNVMGATKRFAEMIVTGLNEEGKTQFAAVRFGNVLGSRGSVVPLFKEQIKKGGPITVTDFRMTRYFMT
NUF1003 Ib            : FIMVSTDKAVNPPNVMGATKRFAEMIVTGLNEEGKTQFAAVRFGNVLGSRGSVVPLFKEQIKKGGPITVTDFRMTRYFMT
NUF1071 Ic            : FIMVSTDKAVNPPNVMGATKRFAEMIVTGLNEEGKTQFAAVRFGNVLGSRGSVVPLFKEQIKKGGPITVTDFRMTRYFMT
2007-1NT ClusterI     : FIMVSTDKAVNPPNVMGATKRFAEMIVTGLNEEGKTQFAAVRFGNVLGSRGSVVPLFKEQIKKGGPITVTDFRMTRYFMT
NUF1032 II            : FIMVSTDKAVNPPNVMGATKRFAEMIVTGLNEEGKTQFAAVRFGNVLGSRGSVVPLFKEQIKKGGPITVTDFRMTRYFMT
NUF1095 NT ClusterIII : FIMVSTDKAVNPPNVMGATKRFAEMIVTGLNEEGKTQFAAVRFGNVLGSRGSVVPLFKEQIKKGGPITVTDFRMTRYFMT

                               490       500       510       520       530       540       550       560
KRS02083 Ia           : IPEASRLVIQAGFLAKGGEVFVLDMGEPVKILDLAKKVIKLSGHTEEDIKIVESGIRPGEKLYEELLSSNERVSEQIHDK
NUF1003 Ib            : IPEASRLVIQAGFLAKGGEVFVLDMGEPVKILDLAKKVIKLSGHTEEDIKIVESGIRPGEKLYEELLSSNERVSEQIHDK
NUF1071 Ic            : IPEASRLVIQAGFLAKGGEVFVLDMGEPVKILDLAKKVIKLSGHTEEDIKIVESGIRPGEKLYEELLSSNERVSEQIHDK
2007-1NT ClusterI     : IPEASRLVIQAGFLAKGGEVFVLDMGEPVKILDLAKKVIKLSGHTEEDIKIVESGIRPGEKLYEELLSSNERVSEQIHDK
NUF1032 II            : IPEASRLVIQAGFLAKGGEVFVLDMGEPVKILDLAKKVIKLSGHTEEDIKIVESGIRPGEKLYEELLSSNERVSEQIHDK
NUF1095 NT ClusterIII : IPEASRLVIQAGFLAKGGEVFVLDMGEPVKILDLAKKVIKLSGHTEEDIKIVESGIRPGEKLYEELLSSNERVSEQIHDK

                               570       580       590       600
KRS02083 Ia           : IFVGKVTSKPLAEVEAIISNLDQLPTDNLKETLIHFAKQE
NUF1003 Ib            : IFVGKVTSKPLAEVEAIISNLDQLPTDNLKETLIHFAKQE
NUF1071 Ic            : IFVGKVTSKPLAEVEAIISNLDQLPTDNLKETLIHFAKQE
2007-1NT ClusterI     : IFVGKVTSKPLAEVEAIISNLDQLPTDNLKETLIHFAKQE
NUF1032 II            : IFVGKVTSKPLAEVEAIISNLDQLPTDNLKETLIHFAKQE
NUF1095 NT ClusterIII : IFVGKVTSKPLAEVEAIISNLDQLPTDNLKETLIHFAKQE
```

S2.1 Multiple nucleotide sequence alignment of *cps* loci lower conserved region in six representative strains.(*cpsQ,cpsR*)

```
                               10        20        30        40        50        60        70        80
KRS02083 Ia          : ATGAAAAAAATAGCAGTAGCCGGTACAGGCTATGTAGGTTTATCTATGGCTGTCTTATTGGCTCAACATCACCAAGTAAC
NUF1003 Ib           : ATGAAAAAAATAGCAGTAGCCGGTACAGGCTATGTAGGTTTATCTATGGCTGTCTTATTGGCTCAACATCACCAAGTAAC
NUF1071 Ic           : ATGAAAAAAATAGCAGTAGCCGGTACAGGCTATGTAGGTTTATCTATGGCTGTCTTATTGGCTCAACATCACCAAGTAAC
2007-1 NT ClusterI   : ATGAAAAAAATAGCAGTAGCCGGTACAGGCTATGTAGGTTTATCTATGGCTGTCTTATTGGCTCAACATCACCAAGTAAC
NUF1032 II           : ATGACAAAAATAGCAGTAGCCGGTACAGGCTATGTAGGTTTATCTATGGCTGTCTTATTGGCTCAACATCACCAAGTAAC
NUF1095 NT ClusterIII: ATGACAAAAATAGCAGTAGCCGGTACAGGCTATGTAGGTTTATCTATGGCTGTCTTATTGGCTCAACATCACCAAGTAAC
                       cpsQ

                               90       100       110       120       130       140       150       160
KRS02083 Ia          : AGCAGTAGATATTATTCCAGAAAAAGTAGACTTAATTAATGATAAAAAATCACCAATTCAAGACAACGAAATAGAAGATT
NUF1003 Ib           : AGCAGTAGATATTATTCCAGAAAAAGTAGACTTAATTAATGATAAAAAATCACCAATTCAAGACAACGAAATAGAAGATT
NUF1071 Ic           : AGCAGTAGATATTATTCCAGAAAAAGTAGACTTAATTAATGATAAAAAATCACCAATTCAAGACAACGAAATAGAAGATT
2007-1 NT ClusterI   : AGCAGTAGATATTATTCCAGAAAAAGTAGACTTAATTAATGATAAAAAATCACCAATTCAAGACAACGAAATAGAAGATT
NUF1032 II           : AGCAGTAGATATTATTCCAGAAAAAGTTGAATTAATTAATAATAAAAAATCACCAATTCAAGATAATGAAATAGAAGATT
NUF1095 NT ClusterIII: AGCAGTAGATATTATTCCAGAAAAAGTTGAATTAATTAATAATAAAAAATCACCAATTCAAGATAATGAAATAGAAGATT

                              170       180       190       200       210       220       230       240
KRS02083 Ia          : ACTTAGCTCATAAAGAGCTAAATCTCATTGCTACATTAGATGGTGAAAAGGCCTATAAAGATGTAGAATTTGTTATAATT
NUF1003 Ib           : ACTTAGCTCATAAAGAGCTAAATCTCATTGCTACATTAGATGGTGAAAAGGCCTATAAAGATGTAGAATTTGTTATAATT
NUF1071 Ic           : ACTTAGCTCATAAAGAGCTAAATCTCATTGCTACATTAGATGGTGAAAAGGCCTATAAAGATGTAGAATTTGTTATAATT
2007-1 NT ClusterI   : ACTTAGCTCATAAAGAGCTAAATCTCATTGCTACATTAGATGGTGAAAAGGCCTATAAAGATGTAGAATTTGTTATAATT
NUF1032 II           : TTTTAGCACATAAAGAGCTCAATCTCACTGCAACATTAGAAGCAGAAAAAGCTTATACTGATGCAGAGTATGTCATTATT
NUF1095 NT ClusterIII: TTTTAGCACATAAAGAGCTCAATCTCACTGCAACATTAGAAGCAGAAAAAGCTTATACTGATGCAGAGTATGTCATTATT

                              250       260       270       280       290       300       310       320
KRS02083 Ia          : GCGGCGCCAACAAATTATGACAGTAAACGTGATTTCTTTGATACAAGTGCAGTAGAAGAAGTTATTAAAACAGTTTTGAG
NUF1003 Ib           : GCGGCGCCAACAAATTATGACAGTAAACGTGATTTCTTTGATACAAGTGCAGTAGAAGAAGTTATTAAAACAGTTTTGAG
NUF1071 Ic           : GCGGCGCCAACAAATTATGACAGTAAACGTGATTTCTTTGATACAAGTGCAGTAGAAGAAGTTATTAAAACAGTTTTGAG
2007-1 NT ClusterI   : GCGGCGCCAACAAATTATGACAGTAAACGTGATTTCTTTGATACAAGTGCAGTAGAAGAAGTTATTAAAACAGTTTTGAG
NUF1032 II           : GCGGCGCCAACAAATTATGACAGTAAACGTGATTTCTTTGATACAAGTGCAGTAGAAGAAGTTATTAAAACAGTTTTGAG
NUF1095 NT ClusterIII: GCGGCGCCAACAAATTATGACAGTAAACGTGATTTCTTTGATACAAGTGCAGTAGAAGAAGTTATTAAAACAGTTTTGAG

                              330       340       350       360       370       380       390       400
KRS02083 Ia          : AGTTAACCCTGATGCAATCATGGTTATAAAATCAACAATTCCAGTAGGATTCACTGAAAGTATTCGTGAAAAATATAATA
NUF1003 Ib           : AGTTAACCCTGATGCAATCATGGTTATAAAATCAACAATTCCAGTAGGATTCACTGAAAGTATTCGTGAAAAATATAATA
NUF1071 Ic           : AGTTAACCCTGATGCAATCATGGTTATAAAATCAACAATTCCAGTAGGATTCACTGAAAGTATTCGTGAAAAATATAATA
2007-1 NT ClusterI   : AGTTAACCCTGATGCAATCATGGTTATAAAATCAACAATTCCAGTAGGATTCACTGAAAGTATTCGTGAAAAATATAATA
NUF1032 II           : AGTTAACCCTGATGCAATCATGGTTATAAAATCAACAATTCCAGTAGGATTCACTGAAAGTATTCGTGAAAAATATAATA
NUF1095 NT ClusterIII: AGTTAACCCTGATGCAATCATGGTTATAAAATCAACAATTCCAGTAGGATTCACTGAAAGTATTCGTGAAAAATATAATA

                              410       420       430       440       450       460       470       480
KRS02083 Ia          : CAACTAATATCATCTTTAGTCCAGAATTCTTAAGAGAATCTCGTGCTTTATATGACAATCTTTATCCAAGTCGGATTATT
NUF1003 Ib           : CAACTAATATCATCTTTAGTCCGGAATTCTTAAGAGAATCTCGTGCTTTATATGACAATCTTTATCCAAGTCGGATTATT
NUF1071 Ic           : CAACTAATATCATCTTTAGTCCGGAATTCTTAAGAGAATCTCGTGCTTTATATGACAATCTTTATCCAAGTCGGATTATT
2007-1 NT ClusterI   : CAACTAATATCATCTTTAGTCCGGAATTCTTAAGAGAATCTCGTGCTTTATATGACAATCTTTATCCAAGTCGGATTATT
NUF1032 II           : CAACTAATATTATCTTTAGTCCGGAATTCTTAAGAGAATCTCGTGCTTTATATGACAATCTTTATCCAAGTCGGATTATT
NUF1095 NT ClusterIII: CAACTAATATTATCTTTAGTCCGGAATTCTTAAGAGAATCTCGTGCTTTATATGACAATCTTTATCCAAGTCGGATTATT

                              490       500       510       520       530       540       550       560
KRS02083 Ia          : GTTGGCACTGATAAAGCTGATCCAGAACTCACAGCTAAGGCTGAAACATTTGCCCATTTGCTTCAAGAGGGTGCCCTGAA
NUF1003 Ib           : GTTGGCACTGATAAAGCTGATCCAGAACTTACGGCTAAGGCTGAAACATTTGCCCATTTGCTTCAAGAGGGTGCCCTAAA
NUF1071 Ic           : GTTGGCACTGATAAAGCTGATCCAGAACTTACGGCTAAGGCTGAAACATTTGCCCATTTGCTTCAAGAGGGTGCCCTAAA
2007-1 NT ClusterI   : GTTGGCACTGATAAAGCTGATCCAGAACTTACGGCTAAGGCTGAAACATTTGCCCATTTGCTTCAAGAGGGTGCCCTAAA
NUF1032 II           : GTTGGCACTGATAAAGCTGATCCAGAACTTACAGCTAAGGCTGAAACATTTGCCCATTTGCTTCAAGAGGGTGCCCTGAA
NUF1095 NT ClusterIII: GTTGGCACTGATAAAGCTGATCCAGAACTTACAGCTAAGGCTGAAACATTTGCCCATTTGCTTCAAGAGGGTGCCCTGAA
```

```
                              570       580       590       600       610       620       630       640
KRS02083 Ia          : AGAATCAATTGAGACATTAATTATGGGATTTACTGAAGCAGAAGCAGTGAAGCTCTTTTCTAACACCTATTTAGCTCTTC
NUF1003 Ib           : AGAATCAATTGAGACATTAATTATGGGATTTACTGAAGCAGAAGCAGTGAAGCTCTTTTCTAACACCTATTTAGCTCTTC
NUF1071 Ic           : AGAATCAATTGAGACATTAATTATGGGATTTACTGAAGCAGAAGCAGTGAAGCTCTTTTCTAACACCTATTTAGCTCTTC
2007-1 NT ClusterI   : AGAATCAATTGAGACATTAATTATGGGATTTACTGAAGCAGAAGCAGTGAAGCTCTTTTCTAACACCTATTTAGCTCTTC
NUF1032 II           : AGAATCAATTGAGACATTAATTATGGGATTTACTGAAGCAGAAGCAGTGAAGCTCTTTTCTAACACCTATTTAGCTCTTC
NUF1095 NT ClusterIII: AGAATCAATTGAGACATTAATTATGGGATTTACTGAAGCAGAAGCAGTGAAGCTCTTTTCTAACACCTATTTAGCTCTTC

                              650       660       670       680       690       700       710       720
KRS02083 Ia          : GCGTTTCATACTTTAATGAACTTGATACATATGCTGAAACTAAAGGTTTAGATTCGAAAGCAATTATTGATGGTGTTGGT
NUF1003 Ib           : GCGTTTCATACTTTAATGAACTTGATACATATGCTGAAACTAAAGGTTTAGATTCGAAAGCAATTATTGATGGTGTTGGT
NUF1071 Ic           : GCGTTTCATACTTTAATGAACTTGATACATATGCTGAAACTAAAGGTTTAGATTCGAAAGCAATTATTGATGGTGTTGGT
2007-1 NT ClusterI   : GCGTTTCATACTTTAATGAACTTGATACATATGCTGAAACTAAAGGTTTAGATTCGAAAGCAATTATTGATGGTGTTGGT
NUF1032 II           : GCGTTTCATACTTTAATGAACTTGATACATATGCTGAAACTAAAGGTTTAGATGCAAAAGCAATTATTGATGGTGTTGGT
NUF1095 NT ClusterIII: GCGTTTCATACTTTAATGAACTTGATACATATGCTGAAACTAAAGGTTTAGATGCAAAAGCAATTATTGATGGTGTTGGT

                              730       740       750       760       770       780       790       800
KRS02083 Ia          : TTAGACCCTAGGATCGGGAATCAATATAATAATCCTTCATTTGGCTATGGTGGTTATTGCTTACCTAAGGATACAAAGCA
NUF1003 Ib           : TTAGACCCTAGGATCGGGAATCAATATAATAATCCTTCATTTGGCTATGGTGGTTATTGCTTACCTAAGGATACAAAGCA
NUF1071 Ic           : TTAGACCCTAGGATCGGGAATCAATATAATAATCCTTCATTTGGCTATGGTGGTTATTGCTTACCTAAGGATACAAAGCA
2007-1 NT ClusterI   : TTAGACCCTAGGATCGGGAATCAATATAATAATCCTTCATTTGGCTATGGTGGTTATTGCTTACCTAAGGATACAAAGCA
NUF1032 II           : TTAGACCCTAGGATCGGGAATCAATATAATAATCCTTCATTTGGCTATGGTGGTTATTGCTTACCTAAGGATACAAAGCA
NUF1095 NT ClusterIII: TTAGACCCTAGGATCGGGAATCAATATAATAATCCTTCATTTGGCTATGGTGGTTATTGCTTACCTAAGGATACAAAGCA

                              810       820       830       840       850       860       870       880
KRS02083 Ia          : ATTACTTGCTAATTATAAAGACGTTCCTCAAAATATGATGACAGCCATCGTTGAAAGTAATCGTACAAGAAAAGATTATA
NUF1003 Ib           : ATTACTTGCTAATTATAAAGACGTTCCTCAAAATATGATGACAGCCATCGTTGAAAGTAATCGGACCCGTAAAGATTATA
NUF1071 Ic           : ATTACTTGCTAATTATAAAGACGTTCCTCAAAATATGATGACAGCCATCGTTGAAAGTAATCGGACCCGTAAAGATTATA
2007-1 NT ClusterI   : ATTACTTGCTAATTATAAAGACGTTCCTCAAAATATGATGACAGCCATCGTTGAAAGTAATCGGACCCGTAAAGATTATA
NUF1032 II           : ATTACTTGCTAATTATAAAGACGTTCCTCAAAATATGATGACAGCCATCGTTGAAAGTAATCGTACAAGAAAAGATTATA
NUF1095 NT ClusterIII: ATTACTTGCTAATTATAAAGACGTTCCTCAAAATATGATGACAGCCATCGTTGAAAGTAATCGTACAAGAAAAGATTATA

                              890       900       910       920       930       940       950       960
KRS02083 Ia          : TTGCGGACAAGGTTTTAGAGATGGCTGGAGCCTATGACGGATCAAGTGAATATAATCCTGAACTGGAAAAAGAAATTGTT
NUF1003 Ib           : TTGCCGATAGAGTACTTGAGATGGCTGGAGCTTATGAACAGTCTAGTGACTTTAACCCTGAGAAGGAGAAAGAAGTCATT
NUF1071 Ic           : TTGCCGATAGAGTACTTGAGATGGCTGGAGCTTATGAACAGTCTAGTGACTTTAACCCTGAGAAGGAGAAAGAAGTCATT
2007-1 NT ClusterI   : TTGCCGATAGAGTACTTGAGATGGCTGGAGCTTATGAACAGTCTAGTGACTTTAACCCTGAGAAGGAGAAAGAAGTCATT
NUF1032 II           : TTGCGGACAAGGTTTTGGAGATGGCTGGAGCTTATGACGGATCAAGTGAATATAATCCTGAACTGGAAAAAGAAATTGTT
NUF1095 NT ClusterIII: TTGCGGACAAGGTTTTGGAGATGGCTGGAGCTTATGACGGATCAAGTGAATATAATCCTGAACTGGAAAAAGAAATTGTT

                              970       980       990      1000      1010      1020      1030      1040
KRS02083 Ia          : ATTGGTGTCTATCGTCTGACAATGAAAAGTAATAGTGATAACTTTAGACAATCTTCAATTCAAGGTGTTATGAAACGTAT
NUF1003 Ib           : ATTGGTGTTTATCGTTTAACCATGAAGAGTAATAGTGATAATTTCAGACAATCTTCAATTCAAGGTGTTATGAAACGTAT
NUF1071 Ic           : ATTGGTGTTTATCGTTTAACCATGAAGAGTAATAGTGATAATTTCAGACAATCTTCAATTCAAGGTGTTATGAAACGTAT
2007-1 NT ClusterI   : ATTGGTGTTTATCGTTTAACCATGAAGAGTAATAGTGATAATTTCAGACAATCTTCAATTCAAGGTGTTATGAAACGTAT
NUF1032 II           : ATTGGTGTCTATCGTCTGACAATGAAAAGTAATAGTGATAACTTTAGACAATCTTCAATTCAAGGTGTTATGAAACGTAT
NUF1095 NT ClusterIII: ATTGGTGTCTATCGTCTGACAATGAAAAGTAATAGTGATAACTTTAGACAATCTTCAATTCAAGGTGTTATGAAACGTAT

                             1050      1060      1070      1080      1090      1100      1110      1120
KRS02083 Ia          : CAAAGCTAAAGGTGCAAAAGTTATTATTTTTGAACCTAGTCTAGAAAATGGAACTACATTTTTTGGTAGCCAGGTTGTTA
NUF1003 Ib           : CAAAGCTAAAGGTGCAAAAGTTATTATTTTTGAACCTAGTCTAGAAAATGGAACTACATTTTTTGGTAGCCAGGTTGTTA
NUF1071 Ic           : CAAAGCTAAAGGTGCAAAAGTTATTATTTTTGAACCTAGTCTAGAAAATGGAACTACATTTTTTGGTAGCCAGGTTGTTA
2007-1 NT ClusterI   : CAAAGCTAAAGGTGCAAAAGTTATTATTTTTGAACCTAGTCTAGAAAATGGAACTACATTTTTTGGTAGCCAGGTTGTTA
NUF1032 II           : CAAAGCTAAAGGTGCAAAAGTTATTATTTTTGAACCTAGTCTAGAAAATGGAACTACATTTTTTGGTAGCCAGGTTGTTA
NUF1095 NT ClusterIII: CAAAGCTAAAGGTGCAAAAGTTATTATTTTTGAACCTAGTCTAGAAAATGGAACTACATTTTTTGGTAGCCAGGTTGTTA
```

```
                             1130      1140      1150      1160      1170      1180      1190      1200
KRS02083 Ia          : ATGACCTTGAAGAATTTAAAAAACTATCTCAAGCAATCATTGCTAACCGCTATGATAACAGCCTTGACGATGTTGAATCA
NUF1003 Ib           : ACGACCTTGAAGAATTTAAAAACCTATCTCAAGCAATCATTGCTAACCGCTATGATAACAGCCTTGACGATGTTGAATCA
NUF1071 Ic           : ACGACCTTGAAGAATTTAAAAACCTATCTCAAGCAATCATTGCTAACCGCTATGATAACAGCCTTGACGATGTTGAATCA
2007-1 NT ClusterI   : ACGACCTTGAAGAATTTAAAAACCTATCTCAAGCAATCATTGCTAACCGCTATGATAACAGCCTTGACGATGTTGAATCA
NUF1032 II           : ATGACCTTGAAGAATTTAAAAAACTATCTCAAGCAATCATTGCTAACCGCTATGATAACAGTCTTGACGATGTTGAATCA
NUF1095 NT ClusterIII: ATGACCTTGAAGAATTTAAAAAACTATCTCAAGCAATCATTGCTAACCGCTATGATAACAGTCTTGACGATGTTGAATCA

                             1210      1220      1230      1240      1250      1260      1270      1280
KRS02083 Ia          : AAAGTTTATACGAGAGATCTATTCAGACGCGATTAACTTTCTCTTTCATCTAATTTTCAAGAAATTGAATTGCCCTTTTA
NUF1003 Ib           : AAAGTTTATACGAGAGATTTATTCAGACGCGATTAACTTTCTCTTTCATCTAATTTTCAAGAAATTGAATTGCCCTTTTA
NUF1071 Ic           : AAAGTTTATACGAGAGATTTATTCAGACGCGATTAACTTTCTCTTTCATCTAATTTTCAAGAAATTGAATTGCCCTTTTA
2007-1 NT ClusterI   : AAAGTTTATACGAGAGATTTATTCAGACGCGATTAACTTTCTCTTTCATCTAATTTTCAAGAAATTGAATTGCCCTTTTA
NUF1032 II           : AAAGTTTATACGAGAGATTTATTCAGACGCGATTAACTTTCTCTTTCATCTAATTTTCAAGAAATTGAATTGCCCTTTTA
NUF1095 NT ClusterIII: AAAGTTTATACGAGAGATTTATTCAGACGCGATTAACTTTCTCTTTCATCTAATTTTCAAGAAATTGAATTGCCCTTTTA

                             1290      1300      1310      1320      1330      1340      1350      1360
KRS02083 Ia          : ATAGTCTTGTGTTATATTTATAATATAGTTTACCATGGAAGAGTTTTCAAAGAAGGAGTGTCGTATGAGTCGTACTGAAC
NUF1003 Ib           : ATAGTCTTGTGTTATATTTATAATATAGTTTACCATGGAAGAGTTTTCAAAGAAGGAGTGTCGTATGAGTCGTACTGAAC
NUF1071 Ic           : ATAGTCTTGTGTTATATTTATAATATAGTTTACCATGGAAGAGTTTTCAAAGAAGGAGTGTCGTATGAGTCGTACTGAAC
2007-1 NT ClusterI   : ATAGTCTTGTGTTATATTTATAATATAGTTTACCATGGAAGAGTTTTCAAAGAAGGAGTGTCGTATGAGTCGTACTGAAC
NUF1032 II           : ATAGTCTTGTGTTATATTTATAATATAGTTTACCATGGAAGAGTTTTCAAAGAAGGAGTGTCGTATGAGTCGTACTGAAC
NUF1095 NT ClusterIII: ATAGTCTTGTGTTATATTTATAATATAGTTTACCATGGAAGAGTTTTCAAAGAAGGAGTGTCGTATGAGTCGTACTGAAC
                                                         cpsR

                             1370      1380      1390      1400      1410      1420      1430      1440
KRS02083 Ia          : AAAGTAGTCATAAAGTTTTAATTCAAAAATTATTGGTCAGTATTCATTATTTGACACTGTTTAAAGATGAACTCCAATTA
NUF1003 Ib           : AAAGTAGTCATAAAGTTTTAATTCAAAAATTATTGGTCAGTATTCATTATTTGACACTGTTTAAAGATGAACTCCAATTA
NUF1071 Ic           : AAAGTAGTCATAAAGTTTTAATTCAAAAATTATTGGTCAGTATTCATTATTTGACACTGTTTAAAGATGAACTCCAATTA
2007-1 NT ClusterI   : AAAGTAGTCATAAAGTTTTAATTCAAAAATTATTGGTCAGTATTCATTATTTGACACTGTTTAAAGATGAACTCCAATTA
NUF1032 II           : AAAGTAGTCATAAAGTTTTAATTCAAAAATTATTGGTCAGTATTCATTATTTGACACTGTTTAAAGATGAACTCCAATTA
NUF1095 NT ClusterIII: AAAGTAGTCATAAAGTTTTAATTCAAAAATTATTGGTCAGTATTCATTATTTGACACTGTTTAAAGATGAACTCCAATTA

                             1450      1460      1470      1480      1490      1500      1510      1520
KRS02083 Ia          : GTGGAGCGGACGCCGTCTATTTTAGGTGGCGAGTTTTCAGCTCAATCTGTTCAATCTGAGTTGGGTGAAATTGTTTCTGC
NUF1003 Ib           : GTGGAGCGGACGCCGTCTATTTTAGGTGGCGAGTTTTCGGCTCAATCTGTTCAATCTGAGTTGGGTGAAATTGTTTCTGC
NUF1071 Ic           : GTGGAGCGGACGCCGTCTATTTTAGGTGGCGAGTTTTCGGCTCAATCTGTTCAATCTGAGTTGGGTGAAATTGTTTCTGC
2007-1 NT ClusterI   : GTGGAGCGGACGCCGTCTATTTTAGGTGGCGAGTTTTCGGCTCAATCTGTTCAATCTGAGTTGGGTGAAATTGTTTCTGC
NUF1032 II           : GTGGAGCGGACGCCGTCTATTTTAGGTGGCGAGTTTTCGGCTCAATCTGTTCAATCTGAGTTGGGTGAAATTGTTTCTGC
NUF1095 NT ClusterIII: GTGGAGCGGACGCCGTCTATTTTAGGTGGCGAGTTTTCGGCTCAATCTGTTCAATCTGAGTTGGGTGAAATTGTTTCTGC

                             1530      1540      1550      1560      1570      1580      1590      1600
KRS02083 Ia          : AATTAATCAATTGGATTATCAGCAACGTTTAATTGAATCAACGTTTTGGTATGAAGAGAAAGCATTCAAATTAATGAACA
NUF1003 Ib           : AATTAATCAATTGGATTATCAGCAACGTTTAATTGAATCAACGTTTTGGTATGAAGAGAAAGCATTCAAATTAATGAACA
NUF1071 Ic           : AATTAATCAATTGGATTATCAGCAACGTTTAATTGAATCAACGTTTTGGTATGAAGAGAAAGCATTCAAATTAATGAACA
2007-1 NT ClusterI   : AATTAATCAATTGGATTATCAGCAACGTTTAATTGAATCAACGTTTTGGTATGAAGAGAAAGCATTCAAATTAATGAACA
NUF1032 II           : AATTAATCAATTGGATTATCAGCAACGTTTAATTGAATCAACGTTTTGGTATGAAGAGAAAGCATTCAAATTAATGAACA
NUF1095 NT ClusterIII: AATTAATCAATTGGATTATCAGCAACGTTTAATTGAATCAACGTTTTGGTATGAAGAGAAAGCATTCAAATTAATGAACA

                             1610      1620      1630      1640      1650      1660      1670      1680
KRS02083 Ia          : AGACTTTGCAAATTGTAGATACTTGGATTCGTGGTCTTGAGAATCTAGTTGAACTTTGCCAAAATAAAGATGTTTTTCAA
NUF1003 Ib           : AGACTTTGCAAATTGTAGATACTTGGATTCGTGGTCTTGAGAATCTAGTTGAACTTTGCCAAAATAAAGATGTTTTTCAA
NUF1071 Ic           : AGACTTTGCAAATTGTAGATACTTGGATTCGTGGTCTTGAGAATCTAGTTGAACTTTGCCAAAATAAAGATGTTTTTCAA
2007-1 NT ClusterI   : AGACTTTGCAAATTGTAGATACTTGGATTCGTGGTCTTGAGAATCTAGTTGAACTTTGCCAAAATAAAGATGTTTTTCAA
NUF1032 II           : AGACTTTGCAAATTGTAGATACTTGGATTCGTGGTCTTGAGAATCTAGTTGAACTTTGCCAAAATAAAGATGTTTTTCAA
NUF1095 NT ClusterIII: AGACTTTGCAAATTGTAGATACTTGGATTCGTGGTCTTGAGAATCTAGTTGAACTTTGCCAAAATAAAGATGTTTTTCAA
```

```
                            1690      1700      1710      1720      1730      1740      1750      1760
KRS02083 Ia          : GCAATCCTTGGAGATAAAAGAATTCGTGTTTTTGGTGTCTTAATTGATGTCTTTTCATCTTTGAAAATCATTGTGATGTC
NUF1003 Ib           : GCAATCCTTGGAGATAAAAGAATTCGTGTTTTTGGTGTCTTAATTGATGTCTTTTCATCTTTGAAAATCATTGTGATGTC
NUF1071 Ic           : GCAATCCTTGGAGATAAAAGAATTCGTGTTTTTGGTGTCTTAATTGATGTCTTTTCATCTTTGAAAATCATTGTGATGTC
2007-1 NT ClusterI   : GCAATCCTTGGAGATAAAAGAATTCGTGTTTTTGGTGTCTTAATTGATGTCTTTTCATCTTTGAAAATCATTGTGATGTC
NUF1032 II           : GCAATCCTTGGAGATAAAAGAATTCGTGTTTTTGGTGTCTTAATTGATGTCTTTTCATCTTTGAAAATCATTGTGATGTC
NUF1095 NT ClusterIII: GCAATCCTTGGAGATAAAAGAATTCGTGTTTTTGGTGTCTTAATTGATGTCTTTTCATCTTTGAAAATCATTGTGATGTC

                            1770      1780      1790      1800      1810      1820      1830      1840
KRS02083 Ia          : ACTCAAAGAAGTGCCCGTTCCTCCTGTTTTGTATGAACATATTAAAATGGTCAACATTGAAGAAGATGCTTTTATTAAGC
NUF1003 Ib           : ACTCAAAGAAGTGCCCGTTCCTCCTGTTTTGTATGAACATATTAAAATGGTCAACATTGAAGAAGATGCTTTTATTAAGC
NUF1071 Ic           : ACTCAAAGAAGTGCCCGTTCCTCCTGTTTTGTATGAACATATTAAAATGGTCAACATTGAAGAAGATGCTTTTATTAAGC
2007-1 NT ClusterI   : ACTCAAAGAAGTGCCCGTTCCTCCTGTTTTGTATGAACATATTAAAATGGTCAACATTGAAGAAGATGCTTTTATTAAGC
NUF1032 II           : ACTCAAAGAAGTGCCCGTTCCTCCTGTTTTGTATGAACATATTAAAATGGTCAACATTGAAGAAGATGCTTTTATTAAGC
NUF1095 NT ClusterIII: ACTCAAAGAAGTGCCCGTTCCTCCTGTTTTGTATGAACATATTAAAATGGTCAACATTGAAGAAGATGCTTTTATTAAGC

                            1850      1860      1870      1880      1890      1900
KRS02083 Ia          : ATTATCAAGGGCCAAAACATTCCTCACCGGAGGCACTTAGTGCAAATGAACAATTATTAAAATAA
NUF1003 Ib           : ATTATCAAGGGCCAAAACATTCCTCACCGGAGGCACTTAGTGCAAATGAACAATTATTAAAATAA
NUF1071 Ic           : ATTATCAAGGGCCAAAACATTCCTCACCGGAGGCACTTAGTGCAAATGAACAATTATTAAAATAA
2007-1 NT ClusterI   : ATTATCAAGGGCCAAAACATTCCTCACCGGAGGCACTTAGTGCAAATGAACAATTATTAAAATAA
NUF1032 II           : ATTATCAAGGGCCAAAACATTCCTCACCGGAGGCACTTAGTGCAAATGAACAATTATTAAAATAA
NUF1095 NT ClusterIII: ATTATCAAGGGCCAAAACATTCCTCACCGGAGGCACTTAGTGCAAATGAACAATTATTAAAATAA
```

S2.2 Multiple amino acid sequence alignment of CpsQ, CpsR in six representative strains.

CpsQ

```
                                10        20        30        40        50        60        70        80
KRS02083 Ia          : MKKIAVAGTGYVGLSMAVLLAQHHQVTAVDIIPEKVDLINDKKSPIQDNEIEDYLAHKELNLIATLDGEKAYKDVEFVII
NUF1003 Ib           : MKKIAVAGTGYVGLSMAVLLAQHHQVTAVDIIPEKVDLINDKKSPIQDNEIEDYLAHKELNLIATLDGEKAYKDVEFVII
NUF1071 Ic           : MKKIAVAGTGYVGLSMAVLLAQHHQVTAVDIIPEKVDLINDKKSPIQDNEIEDYLAHKELNLIATLDGEKAYKDVEFVII
2007-1 NT ClusterI   : MKKIAVAGTGYVGLSMAVLLAQHHQVTAVDIIPEKVDLINDKKSPIQDNEIEDYLAHKELNLIATLDGEKAYKDVEFVII
NUF1032 II           : MTKIAVAGTGYVGLSMAVLLAQHHQVTAVDIIPEKVELINNKKSPIQDNEIEDFLAHKELNLTATLEAEKAYTDAEYVII
NUF1095 NT ClusterIII : MTKIAVAGTGYVGLSMAVLLAQHHQVTAVDIIPEKVELINNKKSPIQDNEIEDFLAHKELNLTATLEAEKAYTDAEYVII


                                90       100       110       120       130       140       150       160
KRS02083 Ia          : AAPTNYDSKRDFFDTSAVEEVIKTVLRVNPDAIMVIKSTIPVGFTESIREKYNTTNIIFSPEFLRESRALYDNLYPSRII
NUF1003 Ib           : AAPTNYDSKRDFFDTSAVEEVIKTVLRVNPDAIMVIKSTIPVGFTESIREKYNTTNIIFSPEFLRESRALYDNLYPSRII
NUF1071 Ic           : AAPTNYDSKRDFFDTSAVEEVIKTVLRVNPDAIMVIKSTIPVGFTESIREKYNTTNIIFSPEFLRESRALYDNLYPSRII
2007-1 NT ClusterI   : AAPTNYDSKRDFFDTSAVEEVIKTVLRVNPDAIMVIKSTIPVGFTESIREKYNTTNIIFSPEFLRESRALYDNLYPSRII
NUF1032 II           : AAPTNYDSKRDFFDTSAVEEVIKTVLRVNPDAIMVIKSTIPVGFTESIREKYNTTNIIFSPEFLRESRALYDNLYPSRII
NUF1095 NT ClusterIII : AAPTNYDSKRDFFDTSAVEEVIKTVLRVNPDAIMVIKSTIPVGFTESIREKYNTTNIIFSPEFLRESRALYDNLYPSRII


                               170       180       190       200       210       220       230       240
KRS02083 Ia          : VGTDKADPELTAKAETFAHLLQEGALKESIETLIMGFTEAEAVKLFSNTYLALRVSYFNELDTYAETKGLDSKAIIDGVG
NUF1003 Ib           : VGTDKADPELTAKAETFAHLLQEGALKESIETLIMGFTEAEAVKLFSNTYLALRVSYFNELDTYAETKGLDSKAIIDGVG
NUF1071 Ic           : VGTDKADPELTAKAETFAHLLQEGALKESIETLIMGFTEAEAVKLFSNTYLALRVSYFNELDTYAETKGLDSKAIIDGVG
2007-1 NT ClusterI   : VGTDKADPELTAKAETFAHLLQEGALKESIETLIMGFTEAEAVKLFSNTYLALRVSYFNELDTYAETKGLDSKAIIDGVG
NUF1032 II           : VGTDKADPELTAKAETFAHLLQEGALKESIETLIMGFTEAEAVKLFSNTYLALRVSYFNELDTYAETKGLDAKAIIDGVG
NUF1095 NT ClusterIII : VGTDKADPELTAKAETFAHLLQEGALKESIETLIMGFTEAEAVKLFSNTYLALRVSYFNELDTYAETKGLDAKAIIDGVG


                               250       260       270       280       290       300       310       320
KRS02083 Ia          : LDPRIGNQYNNPSFGYGGYCLPKDTKQLLANYKDVPQNMMTAIVESNRTRKDYIADKVLEMAGAYDGSSEYNPELEKEIV
NUF1003 Ib           : LDPRIGNQYNNPSFGYGGYCLPKDTKQLLANYKDVPQNMMTAIVESNRTRKDYIADRVLEMAGAYEQSSDFNPEKEKEVI
NUF1071 Ic           : LDPRIGNQYNNPSFGYGGYCLPKDTKQLLANYKDVPQNMMTAIVESNRTRKDYIADRVLEMAGAYEQSSDFNPEKEKEVI
2007-1 NT ClusterI   : LDPRIGNQYNNPSFGYGGYCLPKDTKQLLANYKDVPQNMMTAIVESNRTRKDYIADRVLEMAGAYEQSSDFNPEKEKEVI
NUF1032 II           : LDPRIGNQYNNPSFGYGGYCLPKDTKQLLANYKDVPQNMMTAIVESNRTRKDYIADKVLEMAGAYDGSSEYNPELEKEIV
NUF1095 NT ClusterIII : LDPRIGNQYNNPSFGYGGYCLPKDTKQLLANYKDVPQNMMTAIVESNRTRKDYIADKVLEMAGAYDGSSEYNPELEKEIV


                               330       340       350       360       370       380       390       400
KRS02083 Ia          : IGVYRLTMKSNSDNFRQSSIQGVMKRIKAKGAKVIIFEPSLENGTTFFGSQVVNDLEEFKKLSQAIIANRYDNSLDDVES
NUF1003 Ib           : IGVYRLTMKSNSDNFRQSSIQGVMKRIKAKGAKVIIFEPSLENGTTFFGSQVVNDLEEFKNLSQAIIANRYDNSLDDVES
NUF1071 Ic           : IGVYRLTMKSNSDNFRQSSIQGVMKRIKAKGAKVIIFEPSLENGTTFFGSQVVNDLEEFKNLSQAIIANRYDNSLDDVES
2007-1 NT ClusterI   : IGVYRLTMKSNSDNFRQSSIQGVMKRIKAKGAKVIIFEPSLENGTTFFGSQVVNDLEEFKNLSQAIIANRYDNSLDDVES
NUF1032 II           : IGVYRLTMKSNSDNFRQSSIQGVMKRIKAKGAKVIIFEPSLENGTTFFGSQVVNDLEEFKKLSQAIIANRYDNSLDDVES
NUF1095 NT ClusterIII : IGVYRLTMKSNSDNFRQSSIQGVMKRIKAKGAKVIIFEPSLENGTTFFGSQVVNDLEEFKKLSQAIIANRYDNSLDDVES


                               410
KRS02083 Ia          : KVVYTRDLFRRD
NUF1003 Ib           : KVVYTRDLFRRD
NUF1071 Ic           : KVVYTRDLFRRD
2007-1 NT ClusterI   : KVVYTRDLFRRD
NUF1032 II           : KVVYTRDLFRRD
NUF1095 NT ClusterIII : KVVYTRDLFRRD
```

CpsR

```
                                10        20        30        40        50        60        70        80
KRS02083 Ia          : MEEFSKKECRMSRTEQSSHKVLIQKLLVSIHYLTLFKDELQLVERTPSILGGEFSAQSVQSELGEIVSAINQLDYQQRLI
NUF1003 Ib           : MEEFSKKECRMSRTEQSSHKVLIQKLLVSIHYLTLFKDELQLVERTPSILGGEFSAQSVQSELGEIVSAINQLDYQQRLI
NUF1071 Ic           : MEEFSKKECRMSRTEQSSHKVLIQKLLVSIHYLTLFKDELQLVERTPSILGGEFSAQSVQSELGEIVSAINQLDYQQRLI
2007-1 NT ClusterI   : MEEFSKKECRMSRTEQSSHKVLIQKLLVSIHYLTLFKDELQLVERTPSILGGEFSAQSVQSELGEIVSAINQLDYQQRLI
NUF1032 II           : MEEFSKKECRMSRTEQSSHKVLIQKLLVSIHYLTLFKDELQLVERTPSILGGEFSAQSVQSELGEIVSAINQLDYQQRLI
NUF1095 NT ClusterIII : MEEFSKKECRMSRTEQSSHKVLIQKLLVSIHYLTLFKDELQLVERTPSILGGEFSAQSVQSELGEIVSAINQLDYQQRLI
```

```
                                90       100       110       120       130       140       150       160
KRS02083 Ia           : ESTFWYEEKAFKLMNKTLQIVDTWIRGLENLVELCQNKDVFQAILGDKRIRVFGVLIDVFSSLKIIVMSLKEVPVPPVLY
NUF1003 Ib            : ESTFWYEEKAFKLMNKTLQIVDTWIRGLENLVELCQNKDVFQAILGDKRIRVFGVLIDVFSSLKIIVMSLKEVPVPPVLY
NUF1071 Ic            : ESTFWYEEKAFKLMNKTLQIVDTWIRGLENLVELCQNKDVFQAILGDKRIRVFGVLIDVFSSLKIIVMSLKEVPVPPVLY
2007-1 NT ClusterI    : ESTFWYEEKAFKLMNKTLQIVDTWIRGLENLVELCQNKDVFQAILGDKRIRVFGVLIDVFSSLKIIVMSLKEVPVPPVLY
NUF1032 II            : ESTFWYEEKAFKLMNKTLQIVDTWIRGLENLVELCQNKDVFQAILGDKRIRVFGVLIDVFSSLKIIVMSLKEVPVPPVLY
NUF1095 NT ClusterIII : ESTFWYEEKAFKLMNKTLQIVDTWIRGLENLVELCQNKDVFQAILGDKRIRVFGVLIDVFSSLKIIVMSLKEVPVPPVLY
```

```
                               170       180       190
KRS02083 Ia           : EHIKMVNIEEDAFIKHYQGPKHSSPEALSANEQLLK
NUF1003 Ib            : EHIKMVNIEEDAFIKHYQGPKHSSPEALSANEQLLK
NUF1071 Ic            : EHIKMVNIEEDAFIKHYQGPKHSSPEALSANEQLLK
2007-1 NT ClusterI    : EHIKMVNIEEDAFIKHYQGPKHSSPEALSANEQLLK
NUF1032 II            : EHIKMVNIEEDAFIKHYQGPKHSSPEALSANEQLLK
NUF1095 NT ClusterIII : EHIKMVNIEEDAFIKHYQGPKHSSPEALSANEQLLK
```

S3. Alignment of amino acid sequence of Cps1aH, Cps1aJ, Cps1aK, Cps1bI, Cps1bJ, Cps1bK, Cps2I, Cps2J and Cps2M. (Glycosyltransferase)

```
                10        20        30        40        50        60        70        80
Cps1aH : ----------------------MNPLEEYKLIQIYKNLYKEVDPSV-VLGFT--IKPNIY--GSIAARKLNIPFIANITG
Cps1bI : ----------------------MNPLSEIKLINTYKNIIKTNNPDV-ILGYT--IKPNIY--GAIVAKSFKKPFIANITG
Cps1aK : ----------------------MVEINNFKRKFTITN-STIAIIVN-YNDAQ--RTINLV--KDIIDY------QALKNV
Cps1bK : --------------------MKKDITETKDKSNIENK-KVAAVIVN-YNDAN--RTINLL--KEIVNY------ESINYV
Cps1aJ : ----------------------MV--GSLNVGGSQTMV------MNLYRNID--RSKIQF--DFIIDRPNELFFADEINS
Cps1bJ : ----------------------MKPIRVLHMIGSLNLGGSQALIMNIYRNID--RTQIQF--DFIIDHPSMNYYQDEIER
Cps2I  : MELTNNKNNIWIFCHYAQQPPYNTMLRYHNWGKELVNRDYGVTIVSASTVHN--TDVDLI--DILGKDN------DVVDS
Cps2J  : -------------MNILLITQLYPQPDDEGDNKPTKTVQYFAD-VWQNQGHN--VAIAHCP-SKFPFVFYLIPPKIKNYL
Cps2M  : ---------MVEYISFYLPQFHPVPENDEWYGKGFTEWTNVAKAKPLYPNHYQPHVPADLGFYDLRVKETR---KAQAKL

                90       100       110       120       130       140       150       160
Cps1aH : LGTAVE---Y----KSWKQPIFINL-------YKYAFKGIFQVYFQNTENRD---------------FFLNNKIIKDHYD
Cps1bI : LGTAVE---Y----KSWKQPILINL-------YKFAFKNVYKVFFQNTSNCE---------------FFVNNKIITSKYE
Cps1aK : IVVNNN---S----TDNSIEILSDFEH-----PKYLIINSEINGGYGYGNNL---------------GIKKANEIGADFV
Cps1bK : IVVNNK---S----TDNSQENLESFEH-----DKYILINAKKNGGYGYGNNI---------------GIKKSDEMDMDYT
Cps1aJ : LGGNVY---VLPTFSPKDYFKMVNE-W-----KKFLLSHNYYSAVHFHVRSYISLIIPIV-------KSFKVPVISHSHS
Cps1bJ : LGGRVY---SFPTFTGRNVRNVRNE-W-----DKFFKEHVEYSIIHFHVRSYISLLIPIA-------KRYGLITISHSHS
Cps2I  : ISYKYL---KTPKYSGNGISRIKNMLSFALKLYRLRSEKNRPDTIVVCEAYLYFFAKLFFPKIPIITDIVDLWPLSIVEY
Cps2J  : AGKTSNIFPNISSRKNTRRLEDNKLILRLPLLKSLPGKGYSHKYLRKISDQIIKILGKENFVP----DIVMGHFANPSTE
Cps2M  : AKDYGLTAFC----YWNYWFGDGVELLEQPIRDVYNDKDIDFPFCLGWANH----------------SWEKKQWDKNGTN

               170       180       190       200       210       220       230       240
Cps1aH : LIPGSG-----VNLERFTEKDYPND---D-VI------KFSFISRVMKEKGID--------------QYLDMASHIK---
Cps1bI : ILPGSG-----VNLEKFSYETFPSE---E-IV------KFSFISRIMKEKGID--------------QYLAAAEYVR---
Cps1aK : LICNPDIVFQENTLNSMIKKIGSDT---K-CSII----NARETHLGNFAWKYT--------------NTFQDIISAS---
Cps1bK : LICNPDIYFSENTLIEMMKYLEQDE---S-CALI----NAKENYLGNFAWKYT--------------SDLQDVLCTS---
Cps1aJ : ISSGSGFSSIVKSSLQFPIRYQADYFLAC-SDEA----GRWLFGEKILNKNNY--------------HTVKNAIDGN---
Cps1bJ : ISNGLGLKSKIKNLLQIPIRYQADYLFAC-SKEA----GEWLFGKAAIKKGNF--------------HIIKNAIDGK---
Cps2I  : ANVNPS-NPLIKILYSIEKWAYIKTDALVFSMEG----GPDYVKERSYSTKVN--------------FDKVFHINMG---
Cps2J  : LVSLISQKYNCKSSIVFHGDCNEVNL--SKYRIVENLQNIKAIGARSQYESKKIQKLLSLKELPFICYSGVPDQAVKEID
Cps2M  : ELLVEQ-----KYLGVEDYKKYFYS-------------YLDIFKDDRYYRVDN--------KPFFIIYSPLANEKEIISF

               250       260       270       280       290       300       310       320
Cps1aH : ---KNYPNTEFNIYGFCEQSYEEKLNRLQKE----NV-VKFHGMIKDIA-----GALESSHCLVHP------------SY
Cps1bI : ---SKYPKTEFNIYGFCEQEYENILEDLQNK----KI-VNYHGLVNNIP-----LVLSNTHCLIHP------------TY
Cps1aK : ---IIM-NKLFK-SRYYDEVYFKNKSSAIVD----VIQGSFLLISLPLM-----IKYGMYDEDFFL------------YE
Cps1bK : ---IVF-NKFFS-KRYYKNSYFENKDVVNVD----ILQGSFLLVKTDLM-----LSFGMYDEEFFL------------YE
Cps1aJ : ---LFYFNL--D-KRSEVRKKLQISEHTFVF----GNVGRLTAAKNHMFLLEVFSEL-RKKIDSKL------------LL
Cps1bJ : ---KFKFDT--R-IRTTTRKKLDISNDCLVL----GNVGRITEAKNQEFLIDILHNLISKSENIKL------------IV
Cps2I  : ---IDIKTADQNLSKNINNIPFDESKKNIVYIGSIRKANNIKQICDVALEI--QKDPSLKDVFFHIFGNGDELDSLISYC
Cps2J  : RIETHNLDDTQGLT-YLYVGSLIKRKNLISVI---EAFDQQAEMHDKLLIIGGGPEEEKIKNFISQL-----------KN
Cps2M  : INTWRELAKLEGLGDFYFVGKDMSGINKDKIL---SIGVDAVFEDNTL----------NIHHELNI------------VS

               330       340       350       360       370       380       390       400
Cps1aH : YPEG-LSNVLLE--ASAT-------GRPIITT--NRPGCREVIDD---NINGFLVKQKDTNDLI-------EKVERFINL
Cps1bI : YPEG-MSNVLLE--AAAT-------GRPAITT--NRSGCREIVDD---TLTGFIIEEQNTKDLI-------EKIEIFLNL
Cps1aK : EEKV-LYKKFID--HGYY-------ALTDLD---VNYEHHHVDSK---KITINQLV-TSKRRLL-------QSKLLYLMK
Cps1bK : EEKV-LYKKFHS--HGYY-------SKSVLT---ESYEHHHIDRK---YNYVTQFL-TTKQRLI-------DSKLLFLKK
Cps1aJ : IGDGQLKNDLLR--RAEF-------LGIKQDCIFLGDQKDVFEFY---NAMDTFIFPSLWEGLGIAVIEAETNGIQCYVS
Cps1bJ : VGDGENKSKLKQ--KIDY-------LSLNNHCLLVGSNEFAETYL---NAMDFFIFPSFWEGLGMAAIEAQANGLFCFIS
Cps2I  : DKNSIENIKFYG--RVLKEDIPSILSHSTANILTYKQVSLMKYGG---SQSKLF---DYLASAKPIICNAKFGYNLIERY
Cps2J  : VDKIHMLGRIPREEVMEH-------MSRADVFALISSNEVFGMVYIEAMLNGCITIASAKGGVDGIIIDNQNGYLCEQGN
Cps2M  : KVSQLIKRKVLKRPTVFK-------YKDAIKY--MVDETVTDEHVIPVVAPNWDHSPRSANNAMILHDAKPKYFEDLLKE
```

```
                  410       420       430       440       450       460       470
Cps1aH : THTEKI---KLGHN-------------ARLKVEKEFDRQIVVEQ-----YLKDI----KRAIHK---------
Cps1bI : TTLEKS---NMGKH-------------AREKVEREFNRNIVIRK-----YNHAI----DSIEKKK--------
Cps1aK : YRHLNK--QKLNVA-------------KLFFDLTIIEMYIYSSFLI---VIQSL----KRI-NNE--------
Cps1bK : YRNFSS--FKLSLS-------------KLFFLLTKFEMLLYSLFLI---SKDRL----GKN-NG---------
Cps1aJ : DRVPDS--VDINAGLVKFLSLNEPSEYWAEQIINKKICNRKSPVEKFKSSGYDI----DSTAKWYESFYLNIR
Cps1bJ : NTVPKE--VDIKADLVSFLPLEEGAEFWANKIINSKIVPRSDKTNYLKKSEYLI----DDTVAFFETFYKEIS
Cps2I  : DCGIVT--ENQSEAAFT----KVIKEICSLSDEKLELMGENARKAALDYDQPVL----VDKLIKVIDYVK---
Cps2J  : TKMLSSIFAKIKDM------------TPNEFELMRNKAQETAKHFSETEVATNYL---NKVLARNEE------
Cps2M  : TVKYVK--TKPSNK------------QQVIIKSWNEWGEGNHVEPDLK-YGTGYLEAIKNSLED---------
```

S4. Alignment of amino acid sequence of Cps1aI, Cps1aM, Cps1bH, Cps2H, Cps2L and Cps2N. (Acetyltransferase)

```
                10        20        30        40        50        60        70        80
Cps1aM : --------MINLLTQIFKLYLFKKKWR--SLNNHN---QT------TISKIVDL--------NRVSV----GKNTYGRIN
Cps2N  : --------MIK--------VILNRLRKILNQREWR---RL------NSNNKAVL--------GSVSNYNLISVGNNSYGV
Cps1aI : --------MNK--YSRALIDLPTALLKIISLKVVH---KN------HFKSNKLQFTSPFSEITMDKGAKLEIGNKFRQRS
Cps1bH : -----MKQWKNL-KGMNKLVIIGASGH--GKVVAE---IA------KLSGYNDII-----FLDDYSNEKLCSGYPVVGKV
Cps2H  : --------MKNNEFSIELLN--DKTLVDQVVNIHM---KAFPEFFLTFLGEKFLHTLYTGFLNHNDSGLIIARKKESNSI
Cps2L  : MYLRKLDLADSK-LMLEWMHDEDVTKDLFSNFKNKTIEDVENFITSSQVEDKNIHYAIANDSDEYMGTVSLKNVNRSDGS

                90       100       110       120       130       140       150       160
Cps1aM : M-----SVFDHSDTKLLTIGSF---------CSIAGNVQFLCGGDHFQ---NRLLNYPIE-KKFLNKDE--ATS--KG--
Cps2N  : L-----NVINHSDNYELKIGNF---------CSIATNVQFIVCGEHR----TDTVSTFPL-KVHFMGEKFEAFS--NG--
Cps1aI : H-----SRLRVRKDANLKIGNN---------ISLNHGCMI-VCRDSIS---IADGVQFGP-NVLLYDH---DHD--YK--
Cps1bH : SEI---VNFKNEDVFIAIGSSA---------VREKIAKHLKDHKIVSL---IHPAA-----VVSEKAKI--GKG--SV--
Cps2H  : V-----GFLAYSKDLSNFYKWL---------LKHKIVQFGFYSLIAAI---KSPKSIFRLIRAFLYPSQ--ANKEEDY--
Cps2L  : AEFAISVRKASMGHGYSWYGMKEILDLAFEKYDLDCVYWCVSRRNKRALRFYTKHNFHEVLDVPRDLVE--RYSSIDDLV

               170       180       190       200       210       220       230       240
Cps1aM : EIIIEDDVWIGT--------NALVLSGVTISQGAIIAAGS-----VVSKNVPAYSI-----------VAGVPAKVV----
Cps2N  : NIIVEDDVWIGT--------NAIIMSGVTLGKGSIIAAGS-----VVTKSVPAYTI-----------VGGIPAKVI----
Cps1aI : TDGGINAGLFKT--------APIEIGEN--------VWIG-----AGSIILKGTKI-----------GANS-VVAA----
Cps1bH : IMAG-AVVNPDTEIGEFCIVNTCSSVDHDCIIGDFSHVSV-----GSHVAGTVTVG-----------SHVWIGAGA----
Cps2H  : IEISSLGVLPGS--------SNGGIGSNLLSSFTSLIDTT-----GYNYIELTTDA-----------KNNHKANYF----
Cps2L  : WFSVLKGDVIDNR------DSVSGCKVVRLNTISTLGAGELSFFEGKHDLPFDIKRIYFITKVPEGIRRGYHAHKNLEQL

               250       260       270       280       290       300       310
Cps1aM : ----KYRFP-----ESMINKLIQM--DINQIDDKFINE------NMTLLTTPINDNLCEELLLKLNQTL------
Cps2N  : ----KER-------FSEEIKQLLMELDFSKLTDEFIKN------NIDLLYSDLENKDNQIALRNLVNELNKKEIK
Cps1aI : ----GSV---------VRGVFPEN--TLVYQER---KM------VSK--NY------------------------
Cps1bH : ----TII----------------N--NIETHNNICIGA------GATVINNLVDSGTYVGVPVRRIK--------
Cps2H  : ----YQK----------------NGFILNKSYKTPEGR------EMNEYRYKLKGNNGINKQQK-----------
Cps2L  : LFCPYGRIQLILEDENGREEIELSDPSIGVIIDKPIWREMLWLEKDSVLCVAASEYYDENDYIRDYNDFKEFISK
```

S5. Alignment of amino acid sequence of Cps1bF, Cps2G, and Cps2K. (Aminotransferase)

```
                 10        20        30        40        50        60        70        80
Cps2G  : MKT--MNIPFSPPDITEEEISAVSEALRSGWITTGPKTKLLESRIADYLGTKKV---VCLNSATAGLELSLRILGVGVGD
Cps2K  : MKIPFVSFLPMEKELNDELRNAFETVLTNSWYIGGNEDKKFEESFAKYCETDYC---VGVGNGLDALLLSLKALGIGEGD
Cps1bF : MEYLKNKVWLSSPTMHGDELIFMQEAYDTNWMTTAGRNINELEKMLKV--YTKSEEIVALSSGTSALHLAIKLAGVKHSD

                 90       100       110       120       130       140       150       160
Cps2G  : EVIVPAMTYTASCSVIEHVGATPVIVDIQESSHQFSFDALKNAI--TDKTKVIIPVDLAGIPCDYESLFKIVEEKSSIFK
Cps2K  : EVIVPANTYIASALAISYVGATPVLVEPELETFNIDAKKIESAI--TKNTKAIMPVHLYGLSCDMDVIQEIAQKHNLYII
Cps1bF : YVFCSDMTFSATVNPVVYEGGIPVFIDSETDTWNMDPKALEKAFEIYPNVKVVVLVHLYGVPSKIDEILEICDRHGAVLI

                170       180       190       200       210       220       230       240
Cps2G  : ANSQLQNKIGRITILADGAHAFGAKYKDKMVGQVADMTVFSF----HAVKNFTTAE-----GGAVTWFNDSLNDEDLYKQ
Cps2K  : EDCAQAHG-----------ALYKDKKVGSFGVLSGFSFYPG-----KNLGALGDAG-----GVVGNSKELIDKVRALSN-
Cps1bF : EDAAESLG-----------ATYKGKQTGTFGKHSIISFNGNKIITGSSGGALLTHSAEAANKVRKWSTQSRETASWYQHE

                250       260       270       280       290       300       310       320
Cps2G  : FQIFSLHGQTKDALEKTKIGSWEY---------DILIPGYKCNMTDIMASIGLVQLDRYPNLLEIRREIVETYNKGFQGT
Cps2K  : ----YGSD-YKYHHIYKGNNSRLDELQAAFLSAKLPILDKINKNRNEIANRYLSEIKND-ALALPSVPEGRTHVW-----
Cps1bF : EIG-YNYR-MSNVIAGVVRGQMPY------------LDEHIAQKKAIYFRYKEGLKDLPIVMNPYDGENSIPNFWL----

                330       340       350       360       370       380       390       400
Cps2G  : IVKPLPHLSETSEYES---SNHLYIVHLDGFTQSERNKVIEEMAQKGIACNVHYKPLPMLTAYK----NMGFDVNDYPNA
Cps2K  : ----------HVFAIR---SDK---RADLEEYLSEQGIS----------TNRHYPIPVHLQDCY----KDLGYKKGDFPI
Cps1bF : ----------SCLIINPEAMAEQVRSDSKALYKSENGKSSPTEILEALAKINAEGRPIWKPMHSQPIFRMNPFVTVDGYA

                410       420       430       440
Cps2G  : YHYFENTLTLPLHTKLSNEEVNYVVE-NLNSIT---------QGIQNEK
Cps2K  : AEKISDTQISIPIYYGMTTEEIDYVI-TTLNNY---------R------
Cps1bF : RANSNAYIEGSKTDVVSDIFERGLCLPSDNKMTIEEQEIIIEVIKECFK
```

S6.1 Alignment of nucleotide sequence of *cps1aH* and *cps1bI*. (Glycosyltransferase)

```
                        10        20        30        40        50        60        70        80
KRS02083 Ia : ATGAATCCACTCGAAGA-ATATAAATTAATACAAATATATAAAAATTTATATAAAGAAGTTGATCCTTCTGTAGTATTAG
NUF1003 Ib  : ATGAATCCATTAAGTGAGATA-AAACTAATCAATACTTACAAAAATATAATTAAAACAAATAATCCAGATGTTATTTTAG
              start

                        90       100       110       120       130       140       150       160
KRS02083 Ia : GCTTTACGATTAAACCAAATATCTATGGATCGATTGCTGCACGAAAATTAAATATACCATTTATAGCTAATATTACTGGT
NUF1003 Ib  : GGTATACAATTAAACCTAATATTTACGGTGCAATAGTTGCCAAATCCTTTAAAAAACCATTTATTGCAAATATTACCGGT

                       170       180       190       200       210       220       230       240
KRS02083 Ia : TTAGGAACAGCAGTAGAATACAAAAGTTGGAAGCAACCAATTTTTATCAATCTATACAAGTATGCTTTTAAAGGTATTTT
NUF1003 Ib  : TTAGGAACAGCAGTTGAATACAAAAGCTGGAAACAACCAATTTTAATTAACTTATATAAATTTGCATTTAAAAATGTATA

                       250       260       270       280       290       300       310       320
KRS02083 Ia : TCAAGTTTACTTTCAAAATACAGAAAATAGAGATTTCTTTTTAAATAATAAGATTATAA----AAGATCATTATGATTTA
NUF1003 Ib  : TAAAGTATTTTTTCAAAACACATCAAACTGCGAATTTTTTGTAAATAATAAAATTATTACTTCGAAAT-ATGAAATTTTA

                       330       340       350       360       370       380       390       400
KRS02083 Ia : ATTCCAGGGTCTGGTGTAAATTTAGAAAGATTCACAGAAAAAGATTATCCTAATGATGATGTAATTAAGTTTTCTTTCAT
NUF1003 Ib  : ---CCAGGTTCGGGGGTTAATCTTGAAAAATTTTCTTATGAAACGTTTCCTTCAGAAGAAATTGTGAAGTTTTCTTTTAT

                       410       420       430       440       450       460       470       480
KRS02083 Ia : TTCACGCGTAATGAAAGAAAAAGGGATTGATCAATACCTTGATATGGCTAGTCACATTAAGAAAAATTATCCAAATACTG
NUF1003 Ib  : TTCAAGAATAATGAAAGAAAAAGGAATTGATCAATATTTAGCTGCAGCTGAATATGTTAGAAGTAAGTATCCAAAAACGG

                       490       500       510       520       530       540       550       560
KRS02083 Ia : AATTTAATATTTACGGATTCTGTGAACAGTCGTATGAAGAAAAATTAAATAGACTT-CAGAAAGAAAATGTAGTTAAATT
NUF1003 Ib  : AATTTAATATTTATGGATTTTGTGAGCAAGAGTATGAAAATATTTTAGA-AGATTTACAAAATAAAAAAATAGTAAATTA

                       570       580       590       600       610       620       630       640
KRS02083 Ia : CCATGGGATGATAAAGGATATTGCGGGGGCTTTGGAATCAAGT---CATTGTCTTGTTCATCCTAGTTATTATCCAGAAG
NUF1003 Ib  : TCATGGTTTAGTAAATAATATACCTTTGGTTTT---ATCAAACACACATTGTTTGATTCATCCGACATATTATCCAGAAG

                       650       660       670       680       690       700       710       720
KRS02083 Ia : GTTTATCAAACGTATTACTTGAAGCTTCTGCAACTGGTAGGCCTATTATTACTACCAATAGACCAGGATGTAGAGAAGTT
NUF1003 Ib  : GGATGTCTAACGTCTTATTAGAAGCAGCAGCAACAGGTAGACCTGCAATTACAACAAATAGATCTGGCTGTAGGGAGATT

                       730       740       750       760       770       780       790       800
KRS02083 Ia : ATAGATGATAATATTAATGGATTTTTAGTAAAACAAAAAGATACCAACGATTTAATTGAAAAAGTTGAAAGATTTATTAA
NUF1003 Ib  : GTCGATGACACTTTGACAGGATTTATCATTGAAGAACAGAATACAAAAGACTTAATTGAAAAAATTGAGATATTTTTAAA

                       810       820       830       840       850       860       870       880
KRS02083 Ia : TCTAACT-CATACAGAAAAAATAAAATTAGGACATAACGCTCGTTTAAAAGTTGAAAAGGAATTCGATAGACAAATAGTT
NUF1003 Ib  : CTTAACAACATTAGAAAAAAGTAATATGGGTAAACA-TGCTCGAGAAAAAGTAGAAAGAGAATTTAATAGAAATATTGTA

                       890       900       910       920
KRS02083 Ia : GTCGAACAATAT--TTAAAAGATATAAAACGA--GCAATACATAAATAA
NUF1003 Ib  : ATTCGAAAATATAATCACGCAAT-TGATTCTATTGAAAAAAAGAAATGA
                                                    stop
```

S6.2 Alignment of amino acid sequence of Cps1aH and Cps1bI. (Glycosyltransferase)

```
                      10        20        30        40        50        60        70        80
KRS02083 Ia : MNPLEEYKLIQIYKNLYKEVDPSVVLGFTIKPNIYGSIAARKLNIPFIANITGLGTAVEYKSWKQPIFINLYKYAFKGIF
NUF1003 Ib  : MNPLSEIKLINTYKNIIKTNNPDVILGYTIKPNIYGAIVAKSFKKPFIANITGLGTAVEYKSWKQPILINLYKFAFKNVY

                      90       100       110       120       130       140       150       160
KRS02083 Ia : QVYFQNTENRDFFLNNKIIKDHYDLIPGSGVNLERFTEKDYPNDDVIKFSFISRVMKEKGIDQYLDMASHIKKNYPNTEF
NUF1003 Ib  : KVFFQNTSNCEFFVNNKIITSKYEILPGSGVNLEKFSYETFPSEEIVKFSFISRIMKEKGIDQYLAAAEYVRSKYPKTEF

                     170       180       190       200       210       220       230       240
KRS02083 Ia : NIYGFCEQSYEEKLNRLQKENVVKFHGMIKDIAGALESSHCLVHPSYYPEGLSNVLLEASATGRPIITTNRPGCREVIDD
NUF1003 Ib  : NIYGFCEQEYENILEDLQNKKIVNYHGLVNNIPLVLSNTHCLIHPTYYPEGMSNVLLEAAATGRPAITTNRSGCREIVDD

                     250       260       270       280       290       300
KRS02083 Ia : NINGFLVKQKDTNDLIEKVERFINLTHTEKIKLGHNARLKVEKEFDRQIVVEQYLKDIKRAIHK-
NUF1003 Ib  : TLTGFIIEEQNTKDLIEKIEIFLNLTTLEKSNMGKHAREKVEREFNRNIVIRKYNHAIDSIEKKK
```

S7.1 Alignment of nucleotide sequence of *cps1aJ* and *cps1bJ*. (Glycosyltransferase)

```
                                      start
                       10        20        30        40        50        60        70        80
KRS02083 Ia : ------------------------ATGGTTGGGAGTTTAAATGTTGGTGGCTCTCAGACCATGGTCATGAATCTTTACAG
NUF1003 Ib  : ATGAAACCGATAAGAGTCTTACATATGATAGGTAGTCTTAATCTTGGTGGATCCCAAGCACTAATAATGAATATTTACAG
              start

                       90       100       110       120       130       140       150       160
KRS02083 Ia : AAATATTGATCGTAGTAAGATTCAATTTGATTTTATTATCGATAGACCAA--ATGAACTTTTTTTGCTGATGAAATAAA
NUF1003 Ib  : AAATATTGATCGTACACAAATTCAATTTGATTTTATAATCGATCATCCAAGTATGAACTACTATCAG--GATGAAATAGA

                      170       180       190       200       210       220       230       240
KRS02083 Ia : CTCATTAGGAGGAAATGTTTATGTTCTTCCAACATTTTCTCCTAAAGATTATTTCAAAATGGTAAATGAATGGAAAAAAT
NUF1003 Ib  : AAGACTAGGTGGACGAGTATATTCTTTTCCAACATTTACAGGAAGAAATGTTCGTAATGTGCGTAATGAATGGGATAAGT

                      250       260       270       280       290       300       310       320
KRS02083 Ia : TTTTACTTTCTCATAATTATTATTCAGCAGTGCATTTCCATGTAAGAAGTTACATTTCTCTAATTATACCTATAGTGAAA
NUF1003 Ib  : TTTTTAAAGAACATGTAGAGTACTCGATTATTCATTTTCATGTTAGAAGTTATATTTCTTTATTAATTCCTATTGCAAAA

                      330       340       350       360       370       380       390       400
KRS02083 Ia : TCATTTAAGGTTC--CTGTTATATCTCATAGTCATAGTATTTCAAGTGGAAGTGGGTTTTCCTCAATAGTTAAAAGTTCG
NUF1003 Ib  : AGAT--ACGGTTTAATAACAATATCACATAGTCATAGTATTTCTAATGGACTAGGACTAAAATCAAAGATTAAGAATCTT

                      410       420       430       440       450       460       470       480
KRS02083 Ia : TTACAATTTCCAATAAGATATCAAGCTGATTATTTTCTTGCTTGCTCTGATGAAGCAGGAAGGTGGTTATTTGGGGAGAA
NUF1003 Ib  : TTACAAATTCCGATTAGATATCAAGCTGATTATCTTTTTGCTTGCTCTAAAGAGGCTGGAGAGTGGTTGTTTGGAAAAGC

                      490       500       510       520       530       540       550       560
KRS02083 Ia : AATACTTAACAAAAACAATTATCATACAGTAAAAAATGCAATTGATGGAAATCTGTTTTATTTTAATTTAGATAAAAGAA
NUF1003 Ib  : AGCTATAAAAAAAGGAAATTTTCATATTATAAAAAATGCTATTGACGGAAAAAAATTTAAATTTGATACGAGAATCAGGA

                      570       580       590       600       610       620       630       640
KRS02083 Ia : GTGAAGTAAGAAAAAAATTACAAATTTCTGAACACACTTTTGTTTTCGGCAATGTTGGAAGATTAACAGCGGCGAAAAAT
NUF1003 Ib  : CTACTACTAGAAAAAAACTTGACATTAGTAATGATTGTTTAGTGTTAGGAAATGTTGGTAGAATAACTGAAGCAAAAAAT

                      650       660       670       680       690       700       710       720
KRS02083 Ia : CATATGTTTCTACTTGAAGTTTTTTCTGAGTTAAG---AAAA----AAAATAGATTCGAAACTT-TTATTAATTGGAGAT
NUF1003 Ib  : CAAGAATTTCTAATTGATATTTTACATAATTTAATTTCAAAATCTGAAAATA--TT--AAACTTATTGTTG-TCGGAGAT

                      730       740       750       760       770       780       790       800
KRS02083 Ia : GGTCAACTTAAAAATGACTTATTAAGAAGAGCTGAGTTTTTAGGAATTAAACAAG-ACTGTATATTCTTAGGTGACCAAA
NUF1003 Ib  : GGTGAGAATAAATCAAAATTAAAACAAAAAATTGACTATTTATCT-TTAAATAATCACTGTCTATT---AGTTGGAAGTA

                      810       820       830       840       850       860       870       880
KRS02083 Ia : AGGATGTTTTCGAATTTTAT---AATGCAATGGATACTTTTATTTTTCCTTCACTTTGGGAAGGATTAGGAATTGCTGTA
NUF1003 Ib  : ATGAATTTGCCGAAACGTATTTAAATGCAATGGATTTCTTTATATTTCCATCGTTTTGGGAAGGTTTAGGAATGGCTGCA

                      890       900       910       920       930       940       950       960
KRS02083 Ia : ATTGAAGCAGAAACAAACGG--TATTCAGTGTTACGTTTCTGATAGAGTACCAGATTCTGTTGATATAAACGCAGGACTA
NUF1003 Ib  : ATTGAAGCCCAAGCAAATGGGTTATTT--TGTTTATTTCTAATACTGTTCCAAAAGAAGTTGATATAAAAGCTGACCTT

                      970       980       990      1000      1010      1020      1030      1040
KRS02083 Ia : GT-TAAGTTCTTAAGTCTGAATGAACCATCCGAGTATTGGGCCGAGCAAATTATCAATAAAAAGATATGTAATCGAAAGT
NUF1003 Ib  : GTATCTTTTCTTCCGTTAGAA-GAAGGGGCGGAATTTTGGGCCAACAAAATTATAAACAGCAAAAT-TGTTCCTCGAAGT
```

```
                    1050      1060      1070      1080      1090      1100      1110      1120
KRS02083 Ia : CACCTGTAGAGAA--ATTTAAATCT-TCTGGTTATGATATTGATTCTACAGCAAAATGGTACGAATCGTTTTATTTAAAT
NUF1003 Ib  : GA--TAAAACAAATTATTTAAAAAAATCTGAGTATTTAATTGATGATACAGTAGCATTTTTTGAGACTTTTTATAAGGAG


KRS02083 Ia : ATCCGATGA
NUF1003 Ib  : ATTTCTTAA
                    stop
```

S7.2 Alignment of amino acid sequence of Cps1aJ and Cps1bJ. (Glycosyltransferase)

```
                      10        20        30        40        50        60        70        80
KRS02083 Ia : MV--GSLNVGGSQTMV------MNLYRNIDRSKIQFDFIIDRPNELFFADEINSLGGNVYVLPTFSPKDYFKMVNEWKKF
NUF1003 Ib  : MKPIRVLHMIGSLNLGGSQALIMNIYRNIDRTQIQFDFIIDHPSMNYYQDEIERLGGRVYSFPTFTGRNVRNVRNEWDKF

                      90       100       110       120       130       140       150       160
KRS02083 Ia : LLSHNYYSAVHFHVRSYISLIIPIVKSFKVPVISHSHSISSGSGFSSIVKSSLQFPIRYQADYFLACSDEAGRWLFGEKI
NUF1003 Ib  : FKEHVEYSIIHFHVRSYISLLIPIAKRYGLITISHSHSISNGLGLKSKIKNLLQIPIRYQADYLFACSKEAGEWLFGKAA

                     170       180       190       200       210       220       230       240
KRS02083 Ia : LNKNNYHTVKNAIDGNLFYFNLDKRSEVRKKLQISEHTFVFGNVGRLTAAKNHMFLLEVFSEL-RKKIDSKLLLIGDGQL
NUF1003 Ib  : IKKGNFHIIKNAIDGKKFKFDTRIRTTTRKKLDISNDCLVLGNVGRITEAKNQEFLIDILHNLISKSENIKLIVVGDGEN

                     250       260       270       280       290       300       310       320
KRS02083 Ia : KNDLLRRAEFLGIKQDCIFLGDQKDVFEFYNAMDTFIFPSLWEGLGIAVIEAETNGIQCYVSDRVPDSVDINAGLVKFLS
NUF1003 Ib  : KSKLKQKIDYLSLNNHCLLVGSNEFAETYLNAMDFFIFPSFWEGLGMAAIEAQANGLFCFISNTVPKEVDIKADLVSFLP

                     330       340       350       360
KRS02083 Ia : LNEPSEYWAEQIINKKICNRKSPVEKFKSSGYDIDSTAKWYESFYLNIR
NUF1003 Ib  : LEEGAEFWANKIINSKIVPRSDKTNYLKKSEYLIDDTVAFFETFYKEIS
```

S8.1 Alignment of nucleotide sequence of *cps1aK* and *cps1bK*. (Glycosyltransferase)

```
                          start
                      10        20        30        40        50        60        70        80
KRS02083 Ia : ------------ATGGTTGAAATAAACAACTTTAAAAGAAA-ATTCACGATAACAAATTCTACAATTGCTATAATTGTGA
NUF1003 Ib  : ATGAAAAAAGATATTACTGAAACAAAAGA---TAAAAGTAATATAGAAAATAAGAAG----GTAGCTGCAGTTATTGTCA
              start

                      90       100       110       120       130       140       150       160
KRS02083 Ia : ATTATAACGATGCACAAAGAACTATTAATCTTGTAAAAGATATAATTGATTATCAGGCATTAAAAAACGTAATAGTTGTT
NUF1003 Ib  : ATTATAATGATGCAAACCGAACAATCAATTTACTTAAAGAAATAGTAAATTATGAATCTATTAACTATGTAATTGTAGTA

                     170       180       190       200       210       220       230       240
KRS02083 Ia : AATAATAACTCTACAGATAATTCAATTGAAATATTATCTGAT--TTTGAACATCCTAAATATTTAATCATTAATTCAGAA
NUF1003 Ib  : AATAACAAGTCTACAGATAATTC--TCAAGAAAACCTTGAGTCCTTTGAACATGATAAGTATATTCTGATTAATGCAAAA

                     250       260       270       280       290       300       310       320
KRS02083 Ia : ATTAATGGTGGTTATGGTTATGGAAACAATTTAGGAATAAAAAAAGCTAACGAAATCGGAGCTGACTTTGTCTTAATTTG
NUF1003 Ib  : AAAAATGGTGGATATGGCTATGGGAACAATATTGGTATAAAAAAATCAGATGAAATGGATATGGATTATACTTTAATCTG

                     330       340       350       360       370       380       390       400
KRS02083 Ia : TAATCCTGATATTGTTTTCCAAGAAAATACATTAAATAGTATGATTAAAAAGATTGGTTCTGATACAAAA--TGTTCAAT
NUF1003 Ib  : TAATCCAGACATTTATTTTTCTGAAAATACTTTGATTGAAATGAT--GAAATATTTGGAGCAAGACGAAAGTTGTGCTCT

                     410       420       430       440       450       460       470       480
KRS02083 Ia : TATCAATGCAAGAGAAACTCATCTTGGAAATTTTGCATGGAAGTATACTAATACTTTTCAGGATATCATTTCAGCAAGTA
NUF1003 Ib  : AATTAATGCTAAAGAAAATTATCTGGGAAATTTTGCTTGGAAATATACAAGTGATTTGCAAGACGTTCTCTGTACAAGCA

                     490       500       510       520       530       540       550       560
KRS02083 Ia : TAATAATGAACAAACTATTTAAAAGTAGATATTACGATGA-AGTGTACTTTAAAAATAAAAGTTCCGCAATAGTTGACGT
NUF1003 Ib  : TAGTATTTAATAAATTTTTTAGTAAAAGATATTATAAAAATAGT-TATTTTGAAAATAAAGATGTTGTGAATGTTGATAT

                     570       580       590       600       610       620       630       640
KRS02083 Ia : TATACAAGGTTCCTTTTTACTAATATCACTTCCTTTAATGATAAAGTATGGTATGTATGATGAAGACTTCTTCTTATATG
NUF1003 Ib  : TTTGCAAGGCAGTTTCTTATTAGTAAAAACGGATTTAATGTTAAGTTTTGGAATGTATGATGAAGAATTTTTTCTTTACG

                     650       660       670       680       690       700       710       720
KRS02083 Ia : AGGAAGAGAAGGTATTATATAAAAAATTT-ATAGACCATGGATATTATGCACTAACTGATTTGGATGTTAATTATGAGCA
NUF1003 Ib  : AAGAAGAAAAGTATTATATAAAAAATTTCATTCAC-ATGGATATTATTCAAAATCAGTTCTAACAGAAAGTTATGAACA

                     730       740       750       760       770       780       790       800
KRS02083 Ia : TCATCATGTAGATTCTAAAAAGATAACAA--TCAATCAATTGGTAACTTCTAAGAGAAGA-TTGTTGCAAAGTAAATTAC
NUF1003 Ib  : TCATCATATTGAT--CGAAAATACAATTACGTAACCCAATTTTTGACAACAAAACAAAGACTTATTG-ATAGTAAACTAT

                     810       820       830       840       850       860       870       880
KRS02083 Ia : TATATTTAATGAAATACAG--ACATTTAAATAAACAAAAATTAAACGTGGCAAAATTATTTTTTGATCTTACAATAATAG
NUF1003 Ib  : TATTTTTAAAAAAATATCGTAACTTTTCAA--GTTTTAAATTAAGCTTATCAAAATTATTTTTCTTACTTACAAAATTTG

                     890       900       910       920       930       940
KRS02083 Ia : AGATGTATATT-TACAGTTCATTTCTTAT-AGTAATTCAGAGTCTAAAAAGGATAAACAATGAATAA
NUF1003 Ib  : AAATGT-TACTATACAGTTTATTTTTAATTAGTA----AGGATAGATTAGGTAAAACAATGGATAA
                                                                              stop
```

S8.2 Alignment of amino acid sequence of Cps1aK and Cps1bK. (Glycosyltransferase)

```
                      10        20        30        40        50        60        70        80
KRS02083 Ia : --MVEINNFKRKFTITNSTIAIIVNYNDAQRTINLVKDIIDYQALKNVIVVNNNSTDNSIEILSDFEHPKYLIINSEING
NUF1003 Ib  : MKKDITETKDKSNIENKKVAAVIVNYNDANRTINLLKEIVNYESINYVIVVNNKSTDNSQENLESFEHDKYILINAKKNG

                      90       100       110       120       130       140       150       160
KRS02083 Ia : GYGYGNNLGIKKANEIGADFVLICNPDIVFQENTLNSMIKKIGSDTKCSIINARETHLGNFAWKYTNTFQDIISASIIMN
NUF1003 Ib  : GYGYGNNIGIKKSDEMDMDYTLICNPDIYFSENTLIEMMKYLEQDESCALINAKENYLGNFAWKYTSDLQDVLCTSIVFN

                     170       180       190       200       210       220       230       240
KRS02083 Ia : KLFKSRYYDEVYFKNKSSAIVDVIQGSFLLISLPLMIKYGMYDEDFFLYEEEKVLYKKFIDHGYYALTDLDVNYEHHHVD
NUF1003 Ib  : KFFSKRYYKNSYFENKDVVNVDILQGSFLLVKTDLMLSFGMYDEEFFLYEEEKVLYKKFHSHGYYSKSVLTESYEHHHID

                     250       260       270       280       290       300
KRS02083 Ia : SKKITINQLVTSKRRLLQSKLLYLMKYRHLNKQKLNVAKLFFDLTIIEMYIYSSFLIVIQSLKRINNE
NUF1003 Ib  : RKYNYVTQFLTTKQRLIDSKLLFLKKYRNFSSFKLSLSKLFFLLTKFEMLLYSLFLISKDRLGKNNG-
```

S9.1 Alignment of nucleotide sequence of *cps1aO* and *cps1bN*. (Hypothetical protein)

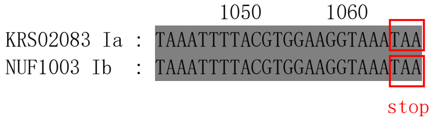
```
                      10        20        30        40        50        60        70        80
KRS02083 Ia : ATGGAAAACTTCAAAAGAACAAACTTTAAACATATTGTCTCTCTAGGCTATTTTTGTTCTATTGCCCAAGATTTAGAAAA
NUF1003 Ib  : ATGGAAAACTTAAAAAAATCACAATTCAAACATTTTATTTCTTTAGGATATTTTTGTTCAATAGCTCAAGATTTGGAAAA
              start

                      90       100       110       120       130       140       150       160
KRS02083 Ia : ATTAGGTCTTAGAGATAGGTCGTACCCATTTGATTGGTGTATCACGAATTTTGAAAAAAACATTGAACTAATTAATAATC
NUF1003 Ib  : ACTAGGTTTACGTGATAAGTCCTATCCTTTTGATTGGTGTATAACTGATTTGGAAAAAAATATTGAATTAATAAACAATA

                     170       180       190       200       210       220       230       240
KRS02083 Ia : ATTTTGATAACTTACTC--AATGAAGAATTTCTCTCACAGAGTGAAGAAATACCTCATCACTATAAAAATGAATTATATG
NUF1003 Ib  : ATTTTG--AATCTATTTTAGATATGGATCTTTTATCACAAGATAAAGAAATTCCACATCATTATCGAAATGATAAGTATG

                     250       260       270       280       290       300       310       320
KRS02083 Ia : ATTTCTACTTTTTCCATGATTTTAATGATACAGAGCCACTTTTGGAA--CAATTGATAACTGTTAAAACTAAATACGATC
NUF1003 Ib  : ACTTTTATTTTTTTCATGATTTTAATTCAGAAGAG--ACTTTGAAAGATCAATTACAAAATGTGAAAGATAAATACAAGA

                     330       340       350       360       370       380       390       400
KRS02083 Ia : GACGGATTAAGGCATTCTACGAAACAATTGAACAACCTACACTTTTTATCAGGTATATTTCTCCACAAGATATTGATTCT
NUF1003 Ib  : GGCGAATTAAAGGATTTTACAAGGATATTGAGGAACCTACATTATTTATAAGATATATATCTCCTGAAAAAATTAATAGT

                     410       420       430       440       450       460       470       480
KRS02083 Ia : TCTGGAAATTCTACTGAACTGAAATATATTGAAAATAGTAAGAATGATATACTTCAAACACTT--AAGCAGTTTAATGAA
NUF1003 Ib  : GAAGGGAAACCTTTAGAGTTAATTTATATTGAAGAAAATCATTCTCACATA--TTAGATACTTTGAAACGATTTAATCCT

                     490       500       510       520       530       540       550       560
KRS02083 Ia : AAAAATGAGATTATTTATATTTCAAATGATAATCTTATGTCAGATAAAATAAAATTATATCATGTTTCTAAAGATACTGG
NUF1003 Ib  : AAGAATGAAATAATTTATATTACTAATATTGGCTTTCAATCTGAAGTTATAAAAGTATTTAATGTCTCAAATGATGATGG

                     570       580       590       600       610       620       630       640
KRS02083 Ia : AGATATTGTTTCGAGATCGCCATTATATAAACAAAAACAATTATTTGATTATTTCAGTCGTCTTAATTTTCCCAATAGAG
NUF1003 Ib  : TGACATAGTTTCTCGCTCCCCTTTATATAAACAGAATTCTCTATTTAATTATTTTTCTAATGTTGATTACCCAAATAGAG

                     650       660       670       680       690       700       710       720
KRS02083 Ia : AAGAAAATCTAAATTTTTTTCTTAATAAAAAGAAATCAAATAATGGATTGATTTCAAAAATCAATCGAAAGATTTCAAAA
NUF1003 Ib  : AACAAAATCTTGATTTTTTTAAAAAGTAAAGATACGAAGACTAGAAAATTACGATCAAGGAT-AATAAAAAAATTCGTGA

                     730       740       750       760
KRS02083 Ia : TTATAT-AAGAAACATTATCAGCATTCCAAGAGATATAAAGAATAG
NUF1003 Ib  : CAATTTTGGGAAAAAATATACACATGTTAAACAATTCAATGATTAG
                                                 stop
```

S9.2 Alignment of amino acid sequence of Cps1aO and Cps1bN. (Hypothetical protein)

```
                      10        20        30        40        50        60        70        80
KRS02083 Ia : MENFKRTNFKHIVSLGYFCSIAQDLEKLGLRDRSYPFDWCITNFEKNIELINNHFDNLLNEEFLSQSEEIPHHYKNELYD
NUF1003 Ib  : MENLKKSQFKHFISLGYFCSIAQDLEKLGLRDKSYPFDWCITDLEKNIELINNNFESILDMDLLSQDKEIPHHYRNDKYD

                      90       100       110       120       130       140       150       160
KRS02083 Ia : FYFFHDFNDTEPLLEQLITVKTKYDRRIKAFYETIEQPTLFIRYISPQDIDSSGNSTELKYIENSKNDILQTLKQFNEKN
NUF1003 Ib  : FYFFHDFNSEETLKDQLQNVKDKYKRRIKGFYKDIEEPTLFIRYISPEKINSEGKPLELIYIEENHSHILDTLKRFNPKN

                     170       180       190       200       210       220       230       240
KRS02083 Ia : EIIYISNDNLMSDKIKLYHVSKDTGDIVSRSPLYKQKQLFDYFSRLNFPNREENLNFFLNKKKSNNGLISKINRKISKLY
NUF1003 Ib  : EIIYITNIGFQSEVIKVFNVSNDDGDIVSRSPLYKQNSLFNYFSNVDYPNREQNLDFLKSKDTKTRKLRSRIIKKIRDNF

                     250
KRS02083 Ia : KKHYQHSKRYKE
NUF1003 Ib  : GKKYTHVKQFND
```

S10.1 Alignment of nucleotide sequence of *cps1aP* and *cps1bO*. (Epimerase)

```
                      10        20        30        40        50        60        70        80
KRS02083 Ia : ATGAACAAACTTTTTAATAAGACCATATTTATTACTGGCGCAGCTGGTTTTATTGGTAGTAATTTAGTATTAGAATTAAT
NUF1003 Ib  : ATGGCTAAAGAAGATAAAAAAACAATTCTAGTTACTGGAGCAGCGGGTTTTATTGGTAGTAATTTAGTTTTAGAATTAAT
             start

                      90       100       110       120       130       140       150       160
KRS02083 Ia : GAAAGATGATGGGATTGAACATATTATTGGTTTAGATAATATGAATGACTATTATGACGTCTCATTAAAAGAATTTAGAC
NUF1003 Ib  : GAACGATAATAGTGTTGAACATATTATTGGATTGGATAATATGAATGACTATTACGATGTCTCATTAAAAGATTTTAGAC

                     170       180       190       200       210       220       230       240
KRS02083 Ia : TTGAACAAATTGAAAATTTGCAAAACGACTCCACAACCAAATGGACTTTTATTAAAGGTTCAATAGCTGATCGTGAACTA
NUF1003 Ib  : TTGAGCAAATTGAAAATTTAGAAAAAAATTCCACAACCAAGTGGACTTTTATTAAAGGTTCCATTGCTGATCGAGATTTA

                     250       260       270       280       290       300       310       320
KRS02083 Ia : GTGGAATATATCTTTAGTAAATATAAACCTGATATTGTAGTTAATTTAGCAGCACAAGCAGGAGTGAGATATTCTATTAC
NUF1003 Ib  : GTTGAAAGTATATTTAAAACATACAAACCTGAAATAGTTGTTAATTTAGCAGCACAAGCAGGGGTGCGTTATTCTATAAC

                     330       340       350       360       370       380       390       400
KRS02083 Ia : TAATCCAGATGTTTATATTGAAAGTAATATTATTGGCTTTTACAATATTCTTGAAGCTTGTAGAAAATACCCAGTAGAAC
NUF1003 Ib  : TAATCCAGATGTTTATATCGAAAGTAATATTATTGGCTTTTATAATATTCTTGAAGCTTGTAGAAATTATCCTGTTGAGC

                     410       420       430       440       450       460       470       480
KRS02083 Ia : ATTTAGTTTATGCGTCGTCATCTTCAGTTTATGGTTCCAATGCAAAAATTCCATATAGTACAGACGATAAAGTTGATTCT
NUF1003 Ib  : ATTTAGTTTATGCTTCATCTTCATCAGTATATGGTTCAAATGCAAAAATTCCTTATAGTACGGAAGATAAAGCTGATTCT

                     490       500       510       520       530       540       550       560
KRS02083 Ia : CCAGTATCATTATATGCAGCTACAAAAAAATCAAATGAACTATTAGCACATTCTTATAGTAAACTTTATAATATTCCATC
NUF1003 Ib  : CCTGTTTCATTATATGCAGCTACAAAAAAATCAAATGAACTATTAGCACATTCTTATAGTAAACTTTATAATATTCCATC

                     570       580       590       600       610       620       630       640
KRS02083 Ia : AACTGGGTTAAGATTTTTTACAGTCTACGGTCCTGCTGGACGGCCAGATATGGCTTACTTTGGTTTTACCAATAAATTAA
NUF1003 Ib  : AACTGGGTTAAGATTTTTTACAGTCTACGGTCCTGCTGGACGGCCAGATATGGCTTACTTTGGTTTTACCAATAAATTAA

                     650       660       670       680       690       700       710       720
KRS02083 Ia : TTAATGGTGAAACAATTGAAATTTTCAATTATGGTAATTGTAAAAGAGATTTCACGTATATTGATGATATTGTAAAAGGT
NUF1003 Ib  : TTAATGGTGAAACAATTGAAATTTTCAATTATGGTAATTGTAAAAGAGATTTCACGTATATTGATGATATTGTAAAAGGT

                     730       740       750       760       770       780       790       800
KRS02083 Ia : ATTAAATTAGTGATGGCATGGCCTCCAAGTCAAATTAATGGTGAAGACGGCTTACCAATTCCGCCATATTCAGTTTACAA
NUF1003 Ib  : ATTAAATTAGTGATGGCATGGCCTCCAAGTCAAATTAATGGTGAAGACGGCTTACCAATTCCGCCATATTCAGTTTACAA

                     810       820       830       840       850       860       870       880
KRS02083 Ia : TATTGGAAATAGTCATCCTGAAAATCTGCTTGAATTTGTAACAATTTTACAAGAAGAACTAATTTCTGCTGGTGTTTTAC
NUF1003 Ib  : TATTGGAAATAGTCATCCTGAAAATCTGCTTGAATTTGTAACAATTTTACAAGAAGAACTAATTTCTGCTGGTGTTTTAC

                     890       900       910       920       930       940       950       960
KRS02083 Ia : CTGCCAATTTTGACTTTGAAGGACATAAAAAAATTGGTAGCAATGCAACCAGGCGATGTTGCAGTAACATATGCTGACACT
NUF1003 Ib  : CTGCCAATTTTGACTTTGAAGGACATAAAAAAATTGGTAGCAATGCAACCAGGCGATGTTGCAGTAACATATGCTGACACT

                     970       980       990      1000      1010      1020      1030      1040
KRS02083 Ia : AGTGCTTTGCAAAAAGATTTTGGATATAAGCCTAATACTAGTTTGAGAGAAGGTTTACGTAACTTTGCAAAATGGTATTT
NUF1003 Ib  : AGTGCTTTGCAAAAAGATTTTGGATATAAGCCTAATACTAGTTTGAGAGAAGGTTTACGTAACTTTGCAAAATGGTACTT
```


1050
1060
KRS02083 Ia : TAAATTTTACGTGGAAGGTAAATAA
NUF1003 Ib : TAAATTTTACGTGGAAGGTAAATAA
stop

S10.2 Alignment of amino acid sequence of Cps1aP and Cps1bO. (Epimerase)

```
                      10        20        30        40        50        60        70        80
KRS02083 Ia : MNKLFNKTIFITGAAGFIGSNLVLELMKDDGIEHIIGLDNMNDYYDVSLKEFRLEQIENLQNDSTTKWTFIKGSIADREL
NUF1003 Ib  : MAKEDKKTILVTGAAGFIGSNLVLELMNDNSVEHIIGLDNMNDYYDVSLKDFRLEQIENLEKNSTTKWTFIKGSIADRDL

                      90       100       110       120       130       140       150       160
KRS02083 Ia : VEYIFSKYKPDIVVNLAAQAGVRYSITNPDVYIESNIIGFYNILEACRKYPVEHLVYASSSSVYGSNAKIPYSTDDKVDS
NUF1003 Ib  : VESIFKTYKPEIVVNLAAQAGVRYSITNPDVYIESNIIGFYNILEACRNYPVEHLVYASSSSVYGSNAKIPYSTEDKADS

                     170       180       190       200       210       220       230       240
KRS02083 Ia : PVSLYAATKKSNELLAHSYSKLYNIPSTGLRFFTVYGPAGRPDMAYFGFTNKLINGETIEIFNYGNCKRDFTYIDDIVKG
NUF1003 Ib  : PVSLYAATKKSNELLAHSYSKLYNIPSTGLRFFTVYGPAGRPDMAYFGFTNKLINGETIEIFNYGNCKRDFTYIDDIVKG

                     250       260       270       280       290       300       310       320
KRS02083 Ia : IKLVMAWPPSQINGEDGLPIPPYSVYNIGNSHPENLLEFVTILQEELISAGVLPANFDFEGHKKLVAMQPGDVAVTYADT
NUF1003 Ib  : IKLVMAWPPSQINGEDGLPIPPYSVYNIGNSHPENLLEFVTILQEELISAGVLPANFDFEGHKKLVAMQPGDVAVTYADT

                     330       340       350
KRS02083 Ia : SALQKDFGYKPNTSLREGLRNFAKWYFKFYVEGK
NUF1003 Ib  : SALQKDFGYKPNTSLREGLRNFAKWYFKFYVEGK
```

S11.1 One base substitution in *cps1bM* (Flippase Wzx) between NUF1003 and NUF1071.

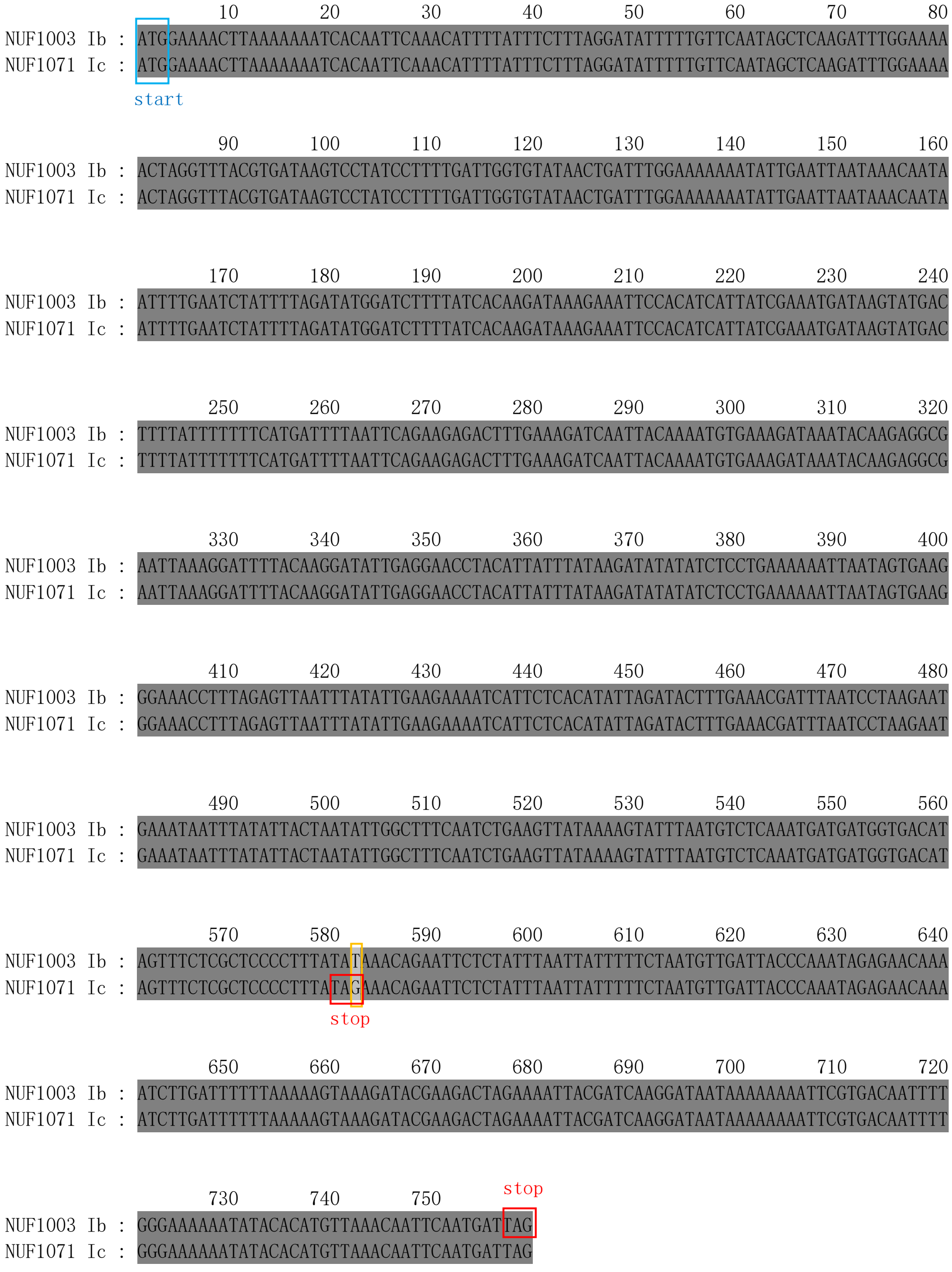
```
                     10        20        30        40        50        60        70        80
NUF1003 Ib : ATGAATTATAGAAATACAGTTAGAAGTTTTTTTGGAGTTGTAGTAAGTAATTTTTCCTCAATAGTTGCAGGAGTAATTGT
NUF1071 Ic : ATGAATTATAGAAATACAGTTAGAAGTTTTTTTGGAGTTGTAGTAAGTAATTTTTCCTCAATAGTTGCAGGAGTAATTGT
             start

                     90       100       110       120       130       140       150       160
NUF1003 Ib : TGGCTTTGCAATACCAAAAATTTTATCAGTGGAAGATTATGGTTACTTAAAGACTTTTACTTTATATGTGTCTTATATGG
NUF1071 Ic : TGGCTTTGCAATACCAAAAATTTTATCAGTGGAAGATTATGGTTACTTAAAGACTTTTACTTTATATGTGTCTTATATGG

                    170       180       190       200       210       220       230       240
NUF1003 Ib : GACTATTTAGTTTCGGAATTATTGATGGAATTGTTCTAGAATTTGGTGGAAAAAGCTATGATGAGTTAGAAAGATTGACC
NUF1071 Ic : GACTATTTAGTTTCGGAATTATTGATGGAATTGTTCTAGAATTTGGTGGAAAAAACTATGATGAGTTAGAAAGATTGACC

                    250       260       270       280       290       300       310       320
NUF1003 Ib : TTCCGTAATTACTTTAGATGGTTTTTCGTTGTAACTTTTATTAGTGCTCTCGTAATAGCTCTGGTTTCCATTATATGTTT
NUF1071 Ic : TTCCGTAATTACTTTAGATGGTTTTTCGTTGTAACTTTTATTAGTGCTCTCGTAATAGCTCTGGTTTCCATTATATGTTT

                    330       340       350       360       370       380       390       400
NUF1003 Ib : TGACAGTGACCTATCATTCATTCTTTTTGCAATTTCTTTTAATTTAATTGCAATCAATACTTCTAATTATTTTCAACAGA
NUF1071 Ic : TGACAGTGACCTATCATTCATTCTTTTTGCAATTTCTTTTAATTTAATTGCAATCAATACTTCTAATTATTTTCAACAGA

                    410       420       430       440       450       460       470       480
NUF1003 Ib : TTTCCCAAATTACTCAAAGATTTAAAGAGTATTCATTAAGAAAAATTTTACAAAGCTTTTCAAATATCTTATTAGTAATA
NUF1071 Ic : TTTCCCAAATTACTCAAAGATTTAAAGAGTATTCATTAAGAAAAATTTTACAAAGCTTTTCAAATATCTTATTAGTAATA

                    490       500       510       520       530       540       550       560
NUF1003 Ib : CTTTGTTTTATCCTTTACAAAACTCACTATGATGTAAATTATAAATTTTATATTATAATGTTAGTCCTTATCAACTTTGG
NUF1071 Ic : CTTTGTTTTATCCTTTACAAAACTCACTATGATGTAAATTATAAATTTTATATTATAATGTTAGTCCTTATCAACTTTGG

                    570       580       590       600       610       620       630       640
NUF1003 Ib : ATTATGTTTGTGGTATGTATATACTTATAGGGAAATAATTTTTGGTGATAAAGTTTCATTTTTTGATAGCAAAAATGACA
NUF1071 Ic : ATTATGTTTGTGGTATGTATATACTTATAGGGAAATAATTTTTGGTGATAAAGTTTCATTTTTTGATAGCAAAAATGACA

                    650       660       670       680       690       700       710       720
NUF1003 Ib : TAATATTTTTAATAAAGACAGGTGTTCCATTGCTGATAGCTAATATTAGTTCTGTTTTAATCGTCACCATTGATAGTCAA
NUF1071 Ic : TAATATTTTTAATAAAGACAGGTGTTCCATTGCTGATAGCTAATATTAGTTCTGTTTTAATCGTCACCATTGATAGTCAA

                    730       740       750       760       770       780       790       800
NUF1003 Ib : TTTGTAAATACACTTTTTTCTACTAGAGAATATGCAATGTATGCATTCGCTTATAATTTACTTTCATTGATAACGATTGC
NUF1071 Ic : TTTGTAAATACACTTTTTTCTACTAGAGAATATGCAATGTATGCATTCGCTTATAATTTACTTTCATTGATAACGATTGC

                    810       820       830       840       850       860       870       880
NUF1003 Ib : AACAGCAGCAATATCGACGATTTTATATCCAACTCTCAAAAGAACTGAAGAGACACGTATAAAAGATAACTATGGGAATC
NUF1071 Ic : AACAGCAGCAATATCGACGATTTTATATCCAACTCTCAAAAGAACTGAAGAGACACGTATAAAAGATAACTATGGGAATC

                    890       900       910       920       930       940       950       960
NUF1003 Ib : TAATTAGCATTTTAGAAGTATTAATATTTGCTTTTCTTATTGCTTTTTTCCCATTATCTATCTTTGTAAATTGGTTCTTA
NUF1071 Ic : TAATTAGCATTTTAGAAGTATTAATATTTGCTTTTCTTATTGCTTTTTTCCCATTATCTATCTTTGTAAATTGGTTCTTA
```

```
                    970       980       990      1000      1010      1020      1030      1040
NUF1003 Ib : CCAAACTATAGCGAATCACTTGAAATTTTTAGAATAATATTTCCAGGTGTTGCTTTAACGACGCCTATAGTAGTCATTAT
NUF1071 Ic : CCAAACTATAGCGAATCACTTGAAATTTTTAGAATAATATTTCCAGGTGTTGCTTTAACGACGCCTATAGTAGTCATTAT


                   1050      1060      1070      1080      1090      1100      1110      1120
NUF1003 Ib : GCATAACTACTATAAAACTTTAAAAAAAAGCAATCAATATTTCTTTAAAAGCATCATGGTTTTAATTTTTTCAATGATCG
NUF1071 Ic : GCATAACTACTATAAAACTTTAAAAAAAAGCAATCAATATTTCTTTAAAAGCATCATGGTTTTAATTTTTTCAATGATCG

                   1130      1140      1150      1160      1170      1180      1190      1200
NUF1003 Ib : CTAATTTTATTGCTTATTACTTCTTTAAAACAACTATAGCTATTTCAGTTGCATCAATTGTAGTTTTATTTCTTTGGTAT
NUF1071 Ic : CTAATTTTATTGCTTATTACTTCTTTAAAACAACTATAGCTATTTCAGTTGCATCAATTGTAGTTTTATTTCTTTGGTAT

                   1210      1220      1230      1240      1250      1260      1270      1280
NUF1003 Ib : GTTTATGTGGAACGAGAATTTGTTAAATCTTTTAATTATAAATCACAGCGAAATTTATCATATATTTTATTATTAATGAC
NUF1071 Ic : GTTTATGTGGAACGAGAATTTGTTAAATCTTTTAATTATAAATCACAGCGAAATTTATCATATATTTTATTATTAATGAC

                   1290      1300      1310      1320      1330      1340      1350      1360
NUF1003 Ib : TATATTTTATTGTTGTTCTTTCATAACTAATTTATACATTGGGTGTTTCGTTTATATTGTTTTATATTGTTTAGTTTCAT
NUF1071 Ic : TATATTTTATTGTTGTTCTTTCATAACTAATTTATACATTGGGTGTTTCGTTTATATTGTTTTATATTGTTTAGTTTCAT

                   1370      1380      1390      1400      1410      1420
NUF1003 Ib : TGATCTATTATAAAAAATTAATTGTTAGCATCTATAATAAGATTACTTCAAATAAACATAGGTAA
NUF1071 Ic : TGATCTATTATAAAAAATTAATTGTTAGCATCTATAATAAGATTACTTCAAATAAACATAGGTAA
                                                                           stop
```

S11.2 Alignment of amino acid sequence of Cps1bM (Flippase Wzx) between NUF1003 and NUF1071.

```
                     10        20        30        40        50        60        70        80
NUF1003 Ib : MNYRNTVRSFFGVVVSNFSSIVAGVIVGFAIPKILSVEDYGYLKTFTLYVSYMGLFSFGIIDGIVLEFGGKSYDELERLT
NUF1071 Ic : MNYRNTVRSFFGVVVSNFSSIVAGVIVGFAIPKILSVEDYGYLKTFTLYVSYMGLFSFGIIDGIVLEFGGKNYDELERLT

                     90       100       110       120       130       140       150       160
NUF1003 Ib : FRNYFRWFFVVTFISALVIALVSIICFDSDLSFILFAISFNLIAINTSNYFQQISQITQRFKEYSLRKILQSFSNILLVI
NUF1071 Ic : FRNYFRWFFVVTFISALVIALVSIICFDSDLSFILFAISFNLIAINTSNYFQQISQITQRFKEYSLRKILQSFSNILLVI

                    170       180       190       200       210       220       230       240
NUF1003 Ib : LCFILYKTHYDVNYKFYIIMLVLINFGLCLWYVYTYREIIFGDKVSFFDSKNDIIFLIKTGVPLLIANISSVLIVTIDSQ
NUF1071 Ic : LCFILYKTHYDVNYKFYIIMLVLINFGLCLWYVYTYREIIFGDKVSFFDSKNDIIFLIKTGVPLLIANISSVLIVTIDSQ

                    250       260       270       280       290       300       310       320
NUF1003 Ib : FVNTLFSTREYAMYAFAYNLLSLITIATAAISTILYPTLKRTEETRIKDNYGNLISILEVLIFAFLIAFFPLSIFVNWFL
NUF1071 Ic : FVNTLFSTREYAMYAFAYNLLSLITIATAAISTILYPTLKRTEETRIKDNYGNLISILEVLIFAFLIAFFPLSIFVNWFL

                    330       340       350       360       370       380       390       400
NUF1003 Ib : PNYSESLEIFRIIFPGVALTTPIVVIMHNYYKTLKKSNQYFFKSIMVLIFSMIANFIAYYFFKTTIAISVASIVVLFLWY
NUF1071 Ic : PNYSESLEIFRIIFPGVALTTPIVVIMHNYYKTLKKSNQYFFKSIMVLIFSMIANFIAYYFFKTTIAISVASIVVLFLWY

                    410       420       430       440       450       460       470
NUF1003 Ib : VYVEREFVKSFNYKSQRNLSYILLLMTIFYCCSFITNLYIGCFVYIVLYCLVSLIYYKKLIVSIYNKITSNKHR
NUF1071 Ic : VYVEREFVKSFNYKSQRNLSYILLLMTIFYCCSFITNLYIGCFVYIVLYCLVSLIYYKKLIVSIYNKITSNKHR
```

S12.1 One base substitution in *cps1bN* (Hypothetical protein) between NUF1003 and NUF1071.



```
                     10        20        30        40        50        60        70        80
NUF1003 Ib : ATGGAAAACTTAAAAAAATCACAATTCAAACATTTTATTTCTTTAGGATATTTTTGTTCAATAGCTCAAGATTTGGAAAA
NUF1071 Ic : ATGGAAAACTTAAAAAAATCACAATTCAAACATTTTATTTCTTTAGGATATTTTTGTTCAATAGCTCAAGATTTGGAAAA

                     90       100       110       120       130       140       150       160
NUF1003 Ib : ACTAGGTTTACGTGATAAGTCCTATCCTTTTGATTGGTGTATAACTGATTTGGAAAAAAATATTGAATTAATAAACAATA
NUF1071 Ic : ACTAGGTTTACGTGATAAGTCCTATCCTTTTGATTGGTGTATAACTGATTTGGAAAAAAATATTGAATTAATAAACAATA

                    170       180       190       200       210       220       230       240
NUF1003 Ib : ATTTTGAATCTATTTTAGATATGGATCTTTTATCACAAGATAAAGAAATTCCACATCATTATCGAAATGATAAGTATGAC
NUF1071 Ic : ATTTTGAATCTATTTTAGATATGGATCTTTTATCACAAGATAAAGAAATTCCACATCATTATCGAAATGATAAGTATGAC

                    250       260       270       280       290       300       310       320
NUF1003 Ib : TTTTATTTTTTTCATGATTTTAATTCAGAAGAGACTTTGAAAGATCAATTACAAAATGTGAAAGATAAATACAAGAGGCG
NUF1071 Ic : TTTTATTTTTTTCATGATTTTAATTCAGAAGAGACTTTGAAAGATCAATTACAAAATGTGAAAGATAAATACAAGAGGCG

                    330       340       350       360       370       380       390       400
NUF1003 Ib : AATTAAAGGATTTTACAAGGATATTGAGGAACCTACATTATTTATAAGATATATATCTCCTGAAAAAATTAATAGTGAAG
NUF1071 Ic : AATTAAAGGATTTTACAAGGATATTGAGGAACCTACATTATTTATAAGATATATATCTCCTGAAAAAATTAATAGTGAAG

                    410       420       430       440       450       460       470       480
NUF1003 Ib : GGAAACCTTTAGAGTTAATTTATATTGAAGAAAATCATTCTCACATATTAGATACTTTGAAACGATTTAATCCTAAGAAT
NUF1071 Ic : GGAAACCTTTAGAGTTAATTTATATTGAAGAAAATCATTCTCACATATTAGATACTTTGAAACGATTTAATCCTAAGAAT

                    490       500       510       520       530       540       550       560
NUF1003 Ib : GAAATAATTTATATTACTAATATTGGCTTTCAATCTGAAGTTATAAAAGTATTTAATGTCTCAAATGATGATGGTGACAT
NUF1071 Ic : GAAATAATTTATATTACTAATATTGGCTTTCAATCTGAAGTTATAAAAGTATTTAATGTCTCAAATGATGATGGTGACAT

                    570       580       590       600       610       620       630       640
NUF1003 Ib : AGTTTCTCGCTCCCCTTTATATAAACAGAATTCTCTATTTAATTATTTTTCTAATGTTGATTACCCAAATAGAGAACAAA
NUF1071 Ic : AGTTTCTCGCTCCCCTTTATAGAAACAGAATTCTCTATTTAATTATTTTTCTAATGTTGATTACCCAAATAGAGAACAAA

                    650       660       670       680       690       700       710       720
NUF1003 Ib : ATCTTGATTTTTTAAAAAGTAAAGATACGAAGACTAGAAAATTACGATCAAGGATAATAAAAAAAATTCGTGACAATTTT
NUF1071 Ic : ATCTTGATTTTTTAAAAAGTAAAGATACGAAGACTAGAAAATTACGATCAAGGATAATAAAAAAAATTCGTGACAATTTT

                    730       740       750
NUF1003 Ib : GGGAAAAAAATATACACATGTTAAACAATTCAATGATTAG
NUF1071 Ic : GGGAAAAAAATATACACATGTTAAACAATTCAATGATTAG
```

S12.2 Alignment of amino acid sequence of Cps1bN (Hypothetical protein) between NUF1003 and NUF1071.

```
                    10         20         30         40         50         60         70         80
NUF1003 Ib : MENLKKSQFKHFISLGYFCSIAQDLEKLGLRDKSYPFDWCITDLEKNIELINNNFESILDMDLLSQDKEIPHHYRNDKYD
NUF1071 Ic : MENLKKSQFKHFISLGYFCSIAQDLEKLGLRDKSYPFDWCITDLEKNIELINNNFESILDMDLLSQDKEIPHHYRNDKYD

                    90        100        110        120        130        140        150        160
NUF1003 Ib : FYFFHDFNSEETLKDQLQNVKDKYKRRIKGFYKDIEEPTLFIRYISPEKINSEGKPLELIYIEENHSHILDTLKRFNPKN
NUF1071 Ic : FYFFHDFNSEETLKDQLQNVKDKYKRRIKGFYKDIEEPTLFIRYISPEKINSEGKPLELIYIEENHSHILDTLKRFNPKN

                   170        180        190        200        210        220        230        240
NUF1003 Ib : EIIYITNIGFQSEVIKVFNVSNDDGDIVSRSPLYKQNSLFNYFSNVDYPNREQNLDFLKSKDTKTRKLRSRIIKKIRDNF
NUF1071 Ic : EIIYITNIGFQSEVIKVFNVSNDDGDIVSRSPL-----------------------------------------------

                   250
NUF1003 Ib : GKKYTHVKQFND
NUF1071 Ic : ------------
```

S13.1 One base substitution in *cps1bG* (Initial glycosyltransferase) between NUF1003 and 2007-1.

```
                             10        20        30        40        50        60        70        80
NUF1003 Ib         : ATGCAAAAAACTTCATTATACAGAACATTTTTTAAACGATTTTTTGATATTTTGTTATCAGGAACTGCATTACTTATTTT
2007-1 NT ClusterI : ATGCAAAAAACTTCATTATACAGAACATTTTTTAAACGATTTTTTGATATTTTGTTATCAGGAACTGCATTACTTATTTT
                     start

                             90       100       110       120       130       140       150       160
NUF1003 Ib         : ATCTCCTATCTTTCTGATAGTAATGTTATTAGTTAAAATCAATTTAGGATCACCAGTTATTTTTAAACAGAAAAGACCTG
2007-1 NT ClusterI : ATCTCCTATCTTTCTGATAGTAATGTTATTAGTTAAAATCAATTTAGGATCACCAGTTATTTTTAAACAGAAAAGACCTG

                            170       180       190       200       210       220       230       240
NUF1003 Ib         : GAAAAGATAGTATGGTTTTTAATATGTATAAATTTAGAACAATGACTTCTTATAAAGATCAGAATGGAAATTTATTACCT
2007-1 NT ClusterI : GAAAAGATAGTATGGTTTTTAATATGTATAAATTTAGAACAATGACTTCTTATAAAGATCAGAATGGAAATTTATTACCT

                            250       260       270       280       290       300       310       320
NUF1003 Ib         : GATAACTTAAGGCTAACTAGTTTTGGAAAATTACTTAGATCAACGAGTTTAGATGAATTACCAGAATTATGGAATATCTT
2007-1 NT ClusterI : GATAACTTAAGGCTAACTAGTTTTGGAAAATTACTTAGATCAACGAGTTTAGATGAATTACCAGAATTATGGAATATCTT

                            330       340       350       360       370       380       390       400
NUF1003 Ib         : TATAGGTGATATGAGTATAGTTGGTCCTAGACCTCTTTTAGAAAAATATTTATCGTTGTATTCTGTTGAGCAAAATCGTA
2007-1 NT ClusterI : TATAGGTGATATGAGTATAGTTGGTCCTAGACCTCTTTTAGAAAAATATTTATCGTTGGATTCTGTTGAGCAAAATCGTA

                            410       420       430       440       450       460       470       480
NUF1003 Ib         : GACATGAAGTTAGACCTGGATTAACAGGATATGCTCAGGCAAATGGACGAAATTCATTATCTTGGCAAGAAAAATTCAAA
2007-1 NT ClusterI : GACATGAAGTTAGACCTGGATTAACAGGATATGCTCAGGCAAATGGACGAAATTCATTATCTTGGCAAGAAAAATTCAAA

                            490       500       510       520       530       540       550       560
NUF1003 Ib         : ATGGATGTCCACTATGTTGATCACATTACTTTTATTGGAGATATAAAAATAATTTGGAAAACCATAGTTACGGTATTAAA
2007-1 NT ClusterI : ATGGATGTCCACTATGTTGATCACATTACTTTTATTGGAGATATAAAAATAATTTGGAAAACCATAGTTACGGTATTAAA

                            570       580       590       600       610       620
NUF1003 Ib         : ACGTGATGGCATAAGTTCTGAAACAAATGAAACAATGGAAGAATTTAAAGGGAATGAATAA
2007-1 NT ClusterI : ACGTGATGGCATAAGTTCTGAAACAAATGAAACAATGGAAGAATTTAAAGGGAATGAATAA
                                                                                  stop
```

S13.2 Alignment of amino acid sequence of Cps1bG (Initial glycosyltransferase) between NUF1003 and 2007-1.

```
                              10        20        30        40        50        60        70        80
NUF1003 Ib         : MQKTSLYRTFFKRFFDILLSGTALLILSPIFLIVMLLVKINLGSPVIFKQKRPGKDSMVFNMYKFRTMTSYKDQNGNLLP
2007-1 NT ClusterI : MQKTSLYRTFFKRFFDILLSGTALLILSPIFLIVMLLVKINLGSPVIFKQKRPGKDSMVFNMYKFRTMTSYKDQNGNLLP

                              90       100       110       120       130       140       150       160
NUF1003 Ib         : DNLRLTSFGKLLRSTSLDELPELWNIFIGDMSIVGPRPLLEKYLSLYSVEQNRRHEVRPGLTGYAQANGRNSLSWQEKFK
2007-1 NT ClusterI : DNLRLTSFGKLLRSTSLDELPELWNIFIGDMSIVGPRPLLEKYLSLDSVEQNRRHEVRPGLTGYAQANGRNSLSWQEKFK

                             170       180       190       200
NUF1003 Ib         : MDVHYVDHITFIGDIKIIWKTIVTVLKRDGISSETNETMEEFKGNE
2007-1 NT ClusterI : MDVHYVDHITFIGDIKIIWKTIVTVLKRDGISSETNETMEEFKGNE
```

S14.1 One base substitution in *cps1bM* (Flippase Wzx) between NUF1003 and 2007-1.

```
                             10        20        30        40        50        60        70        80
NUF1003 Ib         : ATGAATTATAGAAATACAGTTAGAAGTTTTTTTGGAGTTGTAGTAAGTAATTTTTCCTCAATAGTTGCAGGAGTAATTGT
2007-1 NT ClusterI : ATGAATTATAGAAATACAGTTAGAAGTTTTTTTGGAGTTGTAGTAAGTAATTTTTCCTCAATAGTTGCAGGAGTAATTGT
                     start

                             90       100       110       120       130       140       150       160
NUF1003 Ib         : TGGCTTTGCAATACCAAAAATTTTATCAGTGGAAGATTATGGTTACTTAAAGACTTTTACTTTATATGTGTCTTATATGG
2007-1 NT ClusterI : TGGCTTTGCAATACCAAAAATTTTATCAGTGGAAGATTATGGTTACTTAAAGACTTTTACTTTATATGTGTCTTATATGG

                            170       180       190       200       210       220       230       240
NUF1003 Ib         : GACTATTTAGTTTCGGAATTATTGATGGAATTGTTCTAGAATTTGGTGGAAAAAGCTATGATGAGTTAGAAAGATTGACC
2007-1 NT ClusterI : GACTATTTAGTTTCGGAATTATTGATGGAATTGTTCTAGAATTTGGTGGAAAATGCTATGATGAGTTAGAAAGATTGACC

                            250       260       270       280       290       300       310       320
NUF1003 Ib         : TTCCGTAATTACTTTAGATGGTTTTTCGTTGTAACTTTTATTAGTGCTCTCGTAATAGCTCTGGTTTCCATTATATGTTT
2007-1 NT ClusterI : TTCCGTAATTACTTTAGATGGTTTTTCGTTGTAACTTTTATTAGTGCTCTCGTAATAGCTCTGGTTTCCATTATATGTTT

                            330       340       350       360       370       380       390       400
NUF1003 Ib         : TGACAGTGACCTATCATTCATTCTTTTTGCAATTTCTTTTAATTTAATTGCAATCAATACTTCTAATTATTTTCAACAGA
2007-1 NT ClusterI : TGACAGTGACCTATCATTCATTCTTTTTGCAATTTCTTTTAATTTAATTGCAATCAATACTTCTAATTATTTTCAACAGA

                            410       420       430       440       450       460       470       480
NUF1003 Ib         : TTTCCCAAATTACTCAAAGATTTAAAGAGTATTCATTAAGAAAAATTTTACAAAGCTTTTCAAATATCTTATTAGTAATA
2007-1 NT ClusterI : TTTCCCAAATTACTCAAAGATTTAAAGAGTATTCATTAAGAAAAATTTTACAAAGCTTTTCAAATATCTTATTAGTAATA

                            490       500       510       520       530       540       550       560
NUF1003 Ib         : CTTTGTTTTATCCTTTACAAAACTCACTATGATGTAAATTATAAATTTTATATTATAATGTTAGTCCTTATCAACTTTGG
2007-1 NT ClusterI : CTTTGTTTTATCCTTTACAAAACTCACTATGATGTAAATTATAAATTTTATATTATAATGTTAGTCCTTATCAACTTTGG

                            570       580       590       600       610       620       630       640
NUF1003 Ib         : ATTATGTTTGTGGTATGTATATACTTATAGGGAAATAATTTTTGGTGATAAAGTTTCATTTTTTGATAGCAAAAATGACA
2007-1 NT ClusterI : ATTATGTTTGTGGTATGTATATACTTATAGGGAAATAATTTTTGGTGATAAAGTTTCATTTTTTGATAGCAAAAATGACA

                            650       660       670       680       690       700       710       720
NUF1003 Ib         : TAATATTTTTAATAAAGACAGGTGTTCCATTGCTGATAGCTAATATTAGTTCTGTTTTAATCGTCACCATTGATAGTCAA
2007-1 NT ClusterI : TAATATTTTTAATAAAGACAGGTGTTCCATTGCTGATAGCTAATATTAGTTCTGTTTTAATCGTCACCATTGATAGTCAA

                            730       740       750       760       770       780       790       800
NUF1003 Ib         : TTTGTAAATACACTTTTTTCTACTAGAGAATATGCAATGTATGCATTCGCTTATAATTTACTTTCATTGATAACGATTGC
2007-1 NT ClusterI : TTTGTAAATACACTTTTTTCTACTAGAGAATATGCAATGTATGCATTCGCTTATAATTTACTTTCATTGATAACGATTGC

                            810       820       830       840       850       860       870       880
NUF1003 Ib         : AACAGCAGCAATATCGACGATTTTATATCCAACTCTCAAAAGAACTGAAGAGACACGTATAAAAGATAACTATGGGAATC
2007-1 NT ClusterI : AACAGCAGCAATATCGACGATTTTATATCCAACTCTCAAAAGAACTGAAGAGACACGTATAAAAGATAACTATGGGAATC

                            890       900       910       920       930       940       950       960
NUF1003 Ib         : TAATTAGCATTTTAGAAGTATTAATATTTGCTTTTCTTATTGCTTTTTTCCCATTATCTATCTTTGTAAATTGGTTCTTA
2007-1 NT ClusterI : TAATTAGCATTTTAGAAGTATTAATATTTGCTTTTCTTATTGCTTTTTTCCCATTATCTATCTTTGTAAATTGGTTCTTA
```

```
                            970       980       990      1000      1010      1020      1030      1040
NUF1003 Ib         : CCAAACTATAGCGAATCACTTGAAATTTTTAGAATAATATTTCCAGGTGTTGCTTTAACGACGCCTATAGTAGTCATTAT
2007-1 NT ClusterI : CCAAACTATAGCGAATCACTTGAAATTTTTAGAATAATATTTCCAGGTGTTGCTTTAACGACGCCTATAGTAGTCATTAT

                           1050      1060      1070      1080      1090      1100      1110      1120
NUF1003 Ib         : GCATAACTACTATAAAACTTTAAAAAAAAGCAATCAATATTTCTTTAAAAGCATCATGGTTTTAATTTTTTCAATGATCG
2007-1 NT ClusterI : GCATAACTACTATAAAACTTTAAAAAAAAGCAATCAATATTTCTTTAAAAGCATCATGGTTTTAATTTTTTCAATGATCG

                           1130      1140      1150      1160      1170      1180      1190      1200
NUF1003 Ib         : CTAATTTTATTGCTTATTACTTCTTTAAAACAACTATAGCTATTTCAGTTGCATCAATTGTAGTTTTATTTCTTTGGTAT
2007-1 NT ClusterI : CTAATTTTATTGCTTATTACTTCTTTAAAACAACTATAGCTATTTCAGTTGCATCAATTGTAGTTTTATTTCTTTGGTAT

                           1210      1220      1230      1240      1250      1260      1270      1280
NUF1003 Ib         : GTTTATGTGGAACGAGAATTTGTTAAATCTTTTAATTATAAATCACAGCGAAATTTATCATATATTTTATTATTAATGAC
2007-1 NT ClusterI : GTTTATGTGGAACGAGAATTTGTTAAATCTTTTAATTATAAATCACAGCGAAATTTATCATATATTTTATTATTAATGAC

                           1290      1300      1310      1320      1330      1340      1350      1360
NUF1003 Ib         : TATATTTTATTGTTGTTCTTTCATAACTAATTTATACATTGGGTGTTTCGTTTATATTGTTTTATATTGTTTAGTTTCAT
2007-1 NT ClusterI : TATATTTTATTGTTGTTCTTTCATAACTAATTTATACATTGGGTGTTTCGTTTATATTGTTTTATATTGTTTAGTTTCAT

                           1370      1380      1390      1400      1410      1420
NUF1003 Ib         : TGATCTATTATAAAAAATTAATTGTTAGCATCTATAATAAGATTACTTCAAATAAACATAGGTAA
2007-1 NT ClusterI : TGATCTATTATAAAAAATTAATTGTTAGCATCTATAATAAGATTACTTCAAATAAACATAGGTAA
                                                                                 stop
```

S14.2 Alignment of amino acid sequence of Cps1bM (Flippase Wzx) between NUF1003 and 2007-1.

```
                             10        20        30        40        50        60        70        80
NUF1003 Ib         : MNYRNTVRSFFGVVVSNFSSIVAGVIVGFAIPKILSVEDYGYLKTFTLYVSYMGLFSFGIIDGIVLEFGGKSYDELERLT
2007-1 NT ClusterI : MNYRNTVRSFFGVVVSNFSSIVAGVIVGFAIPKILSVEDYGYLKTFTLYVSYMGLFSFGIIDGIVLEFGGKCYDELERLT

                             90       100       110       120       130       140       150       160
NUF1003 Ib         : FRNYFRWFFVVTFISALVIALVSIICFDSDLSFILFAISFNLIAINTSNYFQQISQITQRFKEYSLRKILQSFSNILLVI
2007-1 NT ClusterI : FRNYFRWFFVVTFISALVIALVSIICFDSDLSFILFAISFNLIAINTSNYFQQISQITQRFKEYSLRKILQSFSNILLVI

                            170       180       190       200       210       220       230       240
NUF1003 Ib         : LCFILYKTHYDVNYKFYIIMLVLINFGLCLWYVYTYREIIFGDKVSFFDSKNDIIFLIKTGVPLLIANISSVLIVTIDSQ
2007-1 NT ClusterI : LCFILYKTHYDVNYKFYIIMLVLINFGLCLWYVYTYREIIFGDKVSFFDSKNDIIFLIKTGVPLLIANISSVLIVTIDSQ

                            250       260       270       280       290       300       310       320
NUF1003 Ib         : FVNTLFSTREYAMYAFAYNLLSLITIATAAISTILYPTLKRTEETRIKDNYGNLISILEVLIFAFLIAFFPLSIFVNWFL
2007-1 NT ClusterI : FVNTLFSTREYAMYAFAYNLLSLITIATAAISTILYPTLKRTEETRIKDNYGNLISILEVLIFAFLIAFFPLSIFVNWFL

                            330       340       350       360       370       380       390       400
NUF1003 Ib         : PNYSESLEIFRIIFPGVALTTPIVVIMHNYYKTLKKSNQYFFKSIMVLIFSMIANFIAYYFFKTTIAISVASIVVLFLWY
2007-1 NT ClusterI : PNYSESLEIFRIIFPGVALTTPIVVIMHNYYKTLKKSNQYFFKSIMVLIFSMIANFIAYYFFKTTIAISVASIVVLFLWY

                            410       420       430       440       450       460       470
NUF1003 Ib         : VYVEREFVKSFNYKSQRNLSYILLLMTIFYCCSFITNLYIGCFVYIVLYCLVSLIYYKKLIVSIYNKITSNKHR
2007-1 NT ClusterI : VYVEREFVKSFNYKSQRNLSYILLLMTIFYCCSFITNLYIGCFVYIVLYCLVSLIYYKKLIVSIYNKITSNKHR
```

S15.1 One base substitution in *cps2G* (Aminotransferase) between NUF1032 and NUF1095.

```
                        start
                                10        20        30        40        50        60        70        80
NUF1032 II            : ATGAAAACTATGAATATTCCATTTTCTCCCCCAGATATTACAGAAGAAGAAATCAGCGCAGTTTCTGAAGCTTTAAGATC
NUF1095 NT ClusterIII : ATGAAAACTATGAATATTCCATTTTCTCCCCCAGATATTACAGAAGAAGAAATCAGCGCAGTTTCTGAAGCTTTAAGATC

                                90       100       110       120       130       140       150       160
NUF1032 II            : AGGATGGATTACCACTGGTCCTAAAACTAAGTTATTAGAATCTAGAATAGCTGATTATTTAGGTACTAAGAAAGTTGTAT
NUF1095 NT ClusterIII : AGGATAGATTACCACTGGTCCTAAAACTAAGTTATTAGAATCTAGAATAGCTGATTATTTAGGTACTAAGAAAGTTGTAT
                         early stop

                               170       180       190       200       210       220       230       240
NUF1032 II            : GTTTGAATTCTGCAACTGCGGGATTAGAATTATCATTAAGAATTTTGGGTGTTGGGGTAGGTGATGAAGTTATTGTTCCA
NUF1095 NT ClusterIII : GTTTGAATTCTGCAACTGCGGGATTAGAATTATCATTAAGAATTTTGGGTGTTGGGGTAGGTGATGAAGTTATTGTTCCA

                               250       260       270       280       290       300       310       320
NUF1032 II            : GCAATGACATATACAGCATCTTGTAGTGTAATTGAGCACGTCGGAGCTACTCCAGTAATCGTTGATATTCAAGAAAGCTC
NUF1095 NT ClusterIII : GCAATGACATATACAGCATCTTGTAGTGTAATTGAGCACGTCGGAGCTACTCCAGTAATCGTTGATATTCAAGAAAGCTC
                           start

                               330       340       350       360       370       380       390       400
NUF1032 II            : TCATCAATTTAGTTTTGATGCATTAAAGAATGCAATTACTGATAAGACAAAAGTTATTATTCCTGTAGATTTAGCTGGAA
NUF1095 NT ClusterIII : TCATCAATTTAGTTTTGATGCATTAAAGAATGCAATTACTGATAAGACAAAAGTTATTATTCCTGTAGATTTAGCTGGAA

                               410       420       430       440       450       460       470       480
NUF1032 II            : TTCCATGTGATTATGAATCATTATTCAAGATTGTTGAAGAAAAAAGTTCAATCTTTAAAGCAAACTCTCAATTACAAAAT
NUF1095 NT ClusterIII : TTCCATGTGATTATGAATCATTATTCAAGATTGTTGAAGAAAAAAGTTCAATCTTTAAAGCAAACTCTCAATTACAAAAT

                               490       500       510       520       530       540       550       560
NUF1032 II            : AAAATTGGCCGAATTACTATTTTAGCTGATGGTGCTCATGCATTTGGAGCAAAATATAAAGACAAAATGGTTGGACAAGT
NUF1095 NT ClusterIII : AAAATTGGCCGAATTACTATTTTAGCTGATGGTGCTCATGCATTTGGAGCAAAATATAAAGACAAAATGGTTGGACAAGT

                               570       580       590       600       610       620       630       640
NUF1032 II            : TGCAGATATGACCGTATTTTCATTTCATGCTGTTAAAAATTTCACAACAGCAGAAGGTGGAGCAGTAACATGGTTTAATG
NUF1095 NT ClusterIII : TGCAGATATGACCGTATTTTCATTTCATGCTGTTAAAAATTTCACAACAGCAGAAGGTGGAGCAGTAACATGGTTTAATG

                               650       660       670       680       690       700       710       720
NUF1032 II            : ATTCATTAAATGATGAGGATTTATATAAACAATTTCAAATTTTTTCTCTTCATGGTCAAACTAAGGACGCTTTAGAAAAA
NUF1095 NT ClusterIII : ATTCATTAAATGATGAGGATTTATATAAACAATTTCAAATTTTTTCTCTTCATGGTCAAACTAAGGACGCTTTAGAAAAA

                               730       740       750       760       770       780       790       800
NUF1032 II            : ACTAAAATTGGATCTTGGGAATACGATATTTTAATTCCAGGTTATAAATGTAACATGACTGATATTATGGCATCAATTGG
NUF1095 NT ClusterIII : ACTAAAATTGGATCTTGGGAATACGATATTTTAATTCCAGGTTATAAATGTAACATGACTGATATTATGGCATCAATTGG

                               810       820       830       840       850       860       870       880
NUF1032 II            : TTTAGTGCAACTTGATCGCTACCCAAATTTACTCGAAATAAGACGAGAAATTGTAGAAACATACAATAAAGGATTTCAAG
NUF1095 NT ClusterIII : TTTAGTGCAACTTGATCGCTACCCAAATTTACTCGAAATAAGACGAGAAATTGTAGAAACATACAATAAAGGATTTCAAG

                               890       900       910       920       930       940       950       960
NUF1032 II            : GAACAATTGTAAAACCTCTTCCTCATTTATCTGAAACTAGTGAATATGAATCTTCAAATCACTTATACATTGTTCATCTA
NUF1095 NT ClusterIII : GAACAATTGTAAAACCTCTTCCTCATTTATCTGAAACTAGTGAATATGAATCTTCAAATCACTTATACATTGTTCATCTA

                               970       980       990      1000      1010      1020      1030      1040
NUF1032 II            : GATGGTTTTACCCAAAGTGAAAGAAATAAAGTAATCGAAGAAATGGCACAAAAAGGTATAGCCTGCAATGTTCATTATAA
NUF1095 NT ClusterIII : GATGGTTTTACCCAAAGTGAAAGAAATAAAGTAATCGAAGAAATGGCACAAAAAGGTATAGCCTGCAATGTTCATTATAA
```

```
                              1050      1060      1070      1080      1090      1100      1110      1120
NUF1032 II            : GCCATTACCAATGTTGACAGCATATAAAAACATGGGATTTGATGTAAATGATTATCCAAATGCTTATCATTACTTCGAAA
NUF1095 NT ClusterIII : GCCATTACCAATGTTGACAGCATATAAAAACATGGGATTTGATGTAAATGATTATCCAAATGCTTATCATTACTTCGAAA


                              1130      1140      1150      1160      1170      1180      1190      1200
NUF1032 II            : ATACACTAACACTTCCATTGCATACTAAATTGTCTAATGAAGAAGTTAATTATGTAGTAGAAAACCTCAATAGTATCACA
NUF1095 NT ClusterIII : ATACACTAACACTTCCATTGCATACTAAATTGTCTAATGAAGAAGTTAATTATGTAGTAGAAAACCTCAATAGTATCACA


                              1210      1220
NUF1032 II            : CAGGGTATTCAAAATGAAAAATAA
NUF1095 NT ClusterIII : CAGGGTATTCAAAATGAAAAATAA
                                            stop
```

S15.2 Alignment of amino acid sequence of Cps2G (Aminotransferase) between NUF1032 and NUF1095.

```
                                10        20        30        40        50        60        70        80
NUF1032 II            : MKTMNIPFSPPDITEEEISAVSEALRSGWITTGPKTKLLESRIADYLGTKKVVCLNSATAGLELSLRILGVGVGDEVIVP
NUF1095 NT ClusterIII : --------------------------------------------------------------------------------

                                90       100       110       120       130       140       150       160
NUF1032 II            : AMTYTASCSVIEHVGATPVIVDIQESSHQFSFDALKNAITDKTKVIIPVDLAGIPCDYESLFKIVEEKSSIFKANSQLQN
NUF1095 NT ClusterIII : -MTYTASCSVIEHVGATPVIVDIQESSHQFSFDALKNAITDKTKVIIPVDLAGIPCDYESLFKIVEEKSSIFKANSQLQN

                               170       180       190       200       210       220       230       240
NUF1032 II            : KIGRITILADGAHAFGAKYKDKMVGQVADMTVFSFHAVKNFTTAEGGAVTWFNDSLNDEDLYKQFQIFSLHGQTKDALEK
NUF1095 NT ClusterIII : KIGRITILADGAHAFGAKYKDKMVGQVADMTVFSFHAVKNFTTAEGGAVTWFNDSLNDEDLYKQFQIFSLHGQTKDALEK

                               250       260       270       280       290       300       310       320
NUF1032 II            : TKIGSWEYDILIPGYKCNMTDIMASIGLVQLDRYPNLLEIRREIVETYNKGFQGTIVKPLPHLSETSEYESSNHLYIVHL
NUF1095 NT ClusterIII : TKIGSWEYDILIPGYKCNMTDIMASIGLVQLDRYPNLLEIRREIVETYNKGFQGTIVKPLPHLSETSEYESSNHLYIVHL

                               330       340       350       360       370       380       390       400
NUF1032 II            : DGFTQSERNKVIEEMAQKGIACNVHYKPLPMLTAYKNMGFDVNDYPNAYHYFENTLTLPLHTKLSNEEVNYVVENLNSIT
NUF1095 NT ClusterIII : DGFTQSERNKVIEEMAQKGIACNVHYKPLPMLTAYKNMGFDVNDYPNAYHYFENTLTLPLHTKLSNEEVNYVVENLNSIT


NUF1032 II            : QGIQNEK
NUF1095 NT ClusterIII : QGIQNEK
```

S16.1 Multiple nucleotide sequence alignment of *cps1aF*, *cps1bG* and *cps2F* (Initial glycosyltransferase gene).

```
                      10        20        30        40        50        60        70        80
KRS02083 Ia : ATG--TA-----CCCCATTAT-C---------A-AACGTAGC---TTA-----TC----------------------TAT
NUF1003 Ib  : ATG--CAAAAAACTTCATTATAC---------AGAACATTTT---TTAAACGATT----------------------TTT
NUF1032 II  : ATGAATAAAGAGTTTTTTAATTCTCAAGTAGTGAAGCATGCTGAGTCAA-TAATCGAAAACAAAAAAACAGACTTAATAT
              start

                      90       100       110       120       130       140       150       160
KRS02083 Ia : TA-------TCGTT-TCAGGA--ATTGCCATCA--TTA-TTTTA------GCACCAGTTTTACTGGGAAT--TGCTATCG
NUF1003 Ib  : TGATATT--TTGTTATCAGGA--ACTGC-ATTAC-TTA-TTTTA------TCTCCTATCTTTCTGATAGTAATGTTATT-
NUF1032 II  : TAAAACAAATTTTTGATAAAGTTGTTGCAGTTATTTTACTTTTAATACTTTCCCCTGTGTTTTTAGTTCTA--GCAATCC

                     170       180       190       200       210       220       230       240
KRS02083 Ia : CAATAAAGATTGATTCCACTGGGCCTGTGCTTTTTAAACAAAAACGGGTAGGAAAAGATAAAAGTCACTTTATGATTTAC
NUF1003 Ib  : -AGTTAAAATCAATTTAGGATCACCAGTTATTTTTAAACAGAAAAGACCTGGAAAAGATAGTATGGTTTTTAATATGTAT
NUF1032 II  : TTATAAAAAATGAGGATAATGGTCCAATTTTTTATAGACAGGAACGTATAACAAGATATGGGAAGTTTTTTCATATATAT

                     250       260       270       280       290       300       310       320
KRS02083 Ia : AAGTTCAGAAGTATGTATGTCGATGCTCCTGCAGATATGCCAAC--ACATATGTTAAAAGAT--CCTACAGTTAT-----
NUF1003 Ib  : AAATTTAGAACAATGACT--------TCTTATAAAGATCAGAATGGAAATTTATTACCTGATAACTTAAGGCTA------
NUF1032 II  : AAGTTCAGAACTATGGTT-----------TATAATGCTGATAAACTAGGGTCTTTAG-TAACGACTAAAAATGATCAAAG

                     330       340       350       360       370       380       390       400
KRS02083 Ia : GATTACAAAAGTAGGAGCTTTTCTACGAAAA-ACAAGTTTAGATGAATTGCCTCAGCTTTTTAATATTTTTAAAGGTGAA
NUF1003 Ib  : ----ACTAGTTTTGGAAAATTACTTAGATCA-ACGAGTTTAGATGAATTACCAGAATTATGGAATATCTTTATAGGTGAT
NUF1032 II  : AATTACAAAAGTTGGTTCTTTTATTAGAAAATACAGGTT-AGATGAAATTCCACAACTTTTAAATATTTTAGAAGGTGAT

                     410       420       430       440       450       460       470       480
KRS02083 Ia : ATGGCTATCGTAGGACCTAGACCTGCTTTGTGGAATCAATT--TGATTTAATC-GCCGAAA--GAGATAAATACAGAGCC
NUF1003 Ib  : ATGAGTATAGTTGGTCCTAGACCTCTTTTAGAAAAATATTTATCGTTGTATTCTGTTGAGC--AAAAT-----CGTAGAC
NUF1032 II  : ATGAGTTTCGTTGGAGCTAGACCAGAAGTAATGAAATATGTAGCTCAATATTCAGATGAGATGAAAAT---------TAC

                     490       500       510       520       530       540       550       560
KRS02083 Ia : AATGATATTC--GACCAGGTTTAACAGGCTGGGCACAAATCAAT-----GGACGTG-ATGAATTAGAAATTG--AAGACA
NUF1003 Ib  : A-TGAAGTTA--GACCTGGATTAACAGGATATGCTCAGGCAAAT-----GGACGAA-ATTCATTAT--CTTGGCAAGAAA
NUF1032 II  : ATTGTTATTACCGGCTGGAGTTACCTCAATGT-CTAGTATTGATTTTAAAGATGAGGAAGAGTTGC--TTTC--AAAATA

                     570       580       590       600       610       620       630       640
KRS02083 Ia : AATCAAGGCTAG-ATGGCTATT----------ATGTTGACAACATGTCATTTTTACTAGATCTTAAATGCTTCTTTGGAA
NUF1003 Ib  : AATTCAAAATGG-ATGTCCACT----------ATGTTGAT--CACATTACTTTTATTGGAGATATAAAAATAATTTGGAA
NUF1032 II  : TACTGCAAATGGTATGTCCCTTGATGAGGCTTATATTGAA---AAAATACTTCCAATAAAAATGAAAAAAAATATTGAA-

                     650       660       670       680       690       700       710       720
KRS02083 Ia : ---CATTCTTGAGTGTCCTTAGAAGTGATGGCGTAG---TTGAAG---GTGGTAC-A--GGACAAA---AAAAGGAG-AA
NUF1003 Ib  : AACCATAGTTACG-GTATTAAAACGTGATGGCATAAGTTCTGAAACAAATGAAAC-AATGGAAGAATTTAAAGGGAATGA
NUF1032 II  : ---TATA--TATATAAATTTAATTTTATTGA---AGATATTAAAATA-ATGATATTAACTGTTA---TTCATGTCATAAA

KRS02083 Ia : ATAA
NUF1003 Ib  : ATAA
NUF1032 II  : ATAA
               stop
```

S16.2 Multiple amino acid sequence alignment of Cps1aF, Cps1bG and Cps2F (Initial glycosyltransferase).

```
                      10        20        30        40        50        60        70        80
KRS02083 Ia : ------MYPIIKRSLSIIVSGIAIIILAPVLLGIAIAIKIDSTGPVLFKQ----KRVGKD---KSHFMIYKFRS------
NUF1003 Ib  : MQKTSLYRTFFKRFFDILLSGTALLILSPIFLIVMLLVKINLGSPVIFKQ----KRPGKD---SMVFNMYKFRT------
NUF1032 II  : MNKEFFNSQVVKHAESIIENKKTDLILKQIFDKVVAVILLLILSPVFLVLAILIKNEDNGPIFYRQERITRYGKFFHIYK


                      90       100       110       120       130       140       150       160
KRS02083 Ia : -MYVDAPADMPTH-MLKDPTVMITKVGAFLRKTSLDELPQLFNIFKGEMAIVGPRPALWNQFDLI-AERDKY----RAND
NUF1003 Ib  : -MTSY---KDQNG-NLLPDNLRLTSFGKLLRSTSLDELPELWNIFIGDMSIVGPRPLLEKYLSLY-S-VEQN----RRHE
NUF1032 II  : FRTMVYNADKLGSLVTTKNDQRITKVGSFIRKYRLDEIPQLLNILEGDMSFVGARPEVMKYVAQYSDEMKITLLLPAGVT


                     170       180       190       200       210       220       230
KRS02083 Ia : IRPGLTGWAQINGRDELEIEDKSRLDGYYVDNMSFLLDLKCFFGTFLSVLRSDGVVEGGT-----GQK----KEK-
NUF1003 Ib  : VRPGLTGYAQANGRNSLSWQEKFKMDVHYVDHITFIGDIKIIWKTIVTVLKRDGIS-SET-----NETMEEFKGNE
NUF1032 II  : SMSSIDFKDEEELLS------KYTANGMSLDEAYIEKILPIKMKKNIEYIYKFNFIEDIKIMILTVIH----VIK-
```

S17.1 Multiple nucleotide sequence alignment of *wzy*. (Polymerase)

```
                      10        20        30        40        50        60        70        80
KRS02083 Ia : ATGA-------------ATA------ATTCTCAAATATTTGA----TAAATCTCACAATAA-TA-------TTCTCACAA
NUF1003 Ib  : ATGG-------------ATA------ACTTTAAAATAGGTGA----TTCTTCATTGAATAA-AAAAACGAATCATCATAA
NUF1032 II  : ATGACACCAATTATTTTATATTCCATATTATTAGGTATTTTACTTTTATACTGTCTAATAAATAAAACGAACTACTTAGG
              start

                      90       100       110       120       130       140       150       160
KRS02083 Ia : CTAGG-ATTA-CA----CTTGTTTTATTAGTTATATTTTTGACAATATATGGTTCCGGA--GA----ATTTAAAGTTCCT
NUF1003 Ib  : TAAAGTATTAATA----TTTAGCTTATTATTAATAATTTTTATCATTAATTTTACTGGTTTAG----ACATAAAATTCAT
NUF1032 II  : TTTAGCATTAGTAGCAGTCTATTTTATTATTT-CACTGATTACTCTAT-TAGTACCAACAAGAGGTTATTTGAAATACTC

                     170       180       190       200       210       220       230       240
KRS02083 Ia : TCT--TTCAGTATAAACGT------AGCTGTAAAATA-TATCATTT----TAACTT-TTGGGCTAT--TTTTTTCAGTTT
NUF1003 Ib  : CCCAGTTC--TTCAAATGTCA----AATTTTAATAAG-TATATTTC----TATTTT-ATTAATAATAGTTTTTTCAATTT
NUF1032 II  : TTATGTTTTAGATGGAAGTAATATTACTTATGGAATAGTATTATTTATGGTGTTTTCATTTGCTAT--TTTTTTCTTTCC

                     250       260       270       280       290       300       310       320
KRS02083 Ia : -TTTTTATTAAAGTAAAAAAATTTTCAAT------AA--TAAAGTTAGAT----ATAATTTCTCAGTTGTTATTGATTCA
NUF1003 Ib  : -TTCTT-TCAATATTTCAAGA----CAAT------AAACTTAAGTTTGAT----CAAATTTCAATAATACTTTTTTGTCA
NUF1032 II  : GTTTCTGATAAAA-AAGAAAAGTTTCAATGTTAAAAATATTGAGCTAAATGTTCATGAAAACTACATTATTTTTGCCAAA

                     330       340       350       360       370       380       390       400
KRS02083 Ia : TTTATTGTATAAT----ACAATAATTATTGTATTAAATTCGTCCGATA-ATTACGTTTTATCGTACTTTACATATGTATT
NUF1003 Ib  : GCTTTTATACTATTTTTGCGTT--TTATTAT-TTAG--TAGCTCTACAGACTATTCTTTATTATTTTTCTCATATACTTT
NUF1032 II  : ATATTTATATTTTT--TGCACTGATTACTGTTCTCG------TTTATA---TGGTCAAAATTATTCCTCTTATACAG---

                     410       420       430       440       450       460       470       480
KRS02083 Ia : TGCATTCTTAGGTTATTTATGGGGAAGTAACAGCCTAATTGACAAAAAAACGT-TCAAAATATTCTTATCAAGTTTTTAC
NUF1003 Ib  : TTCTTTCTTTGCTTACATATGGATTAATAATAACAAAAT-ATCAAAAGATCATCTTTTAAT-TTTTTCTCGAGTCTTTT-
NUF1032 II  : -------TCAGGTCA---ATGGGCTAGTAATCGATTAGT-----AATGAAC------------TCTGATCAAGCCATAAT

                     490       500       510       520       530       540       550       560
KRS02083 Ia : TATATCCTTATTA-----TTCAAACATTATTTACCGTATA-CATTGTTAGTCATTCAGTTGTGGAT---ATGTATTTGTT
NUF1003 Ib  : --TATCAATATTAATAATTTCTATTATCATT-ACAGTATA-CA--ATTTATC--TAAATTACAAATTCCACTTTTTCTTT
NUF1032 II  : ----TCCATATAA--------CAATATCTTTGAAAAAATAGCA------ATC--CTATTTACAAGTTATACTCAATTACT

                     570       580       590       600       610       620       630       640
KRS02083 Ia : TAAAAATGGTATTGTTATT--CCTATAGGAGCATCAAATGGGATTACAACATTTATTGTTATGATT---TTTCCAATATT
NUF1003 Ib  : TTAAATCTCAAATAATATTA-CCAATTGGGTCTTCAAATGCTATTACAACTTATACATTTTTACTT---TTACCTATAGT
NUF1032 II  : AGCAATTGTAGTTGGTTTTAATTTACTGAG--AACAAAT-------CAACAT-TATATGCTTGGCTCAATTGTCT-TATT

                     650       660       670       680       690       700       710       720
KRS02083 Ia : GTATAAATTAAGCAATAGTAGAACTTCAC---AAT--ATTTTTTAACTATATTTACTATGATTTTTGCTGTATTATCTCG
NUF1003 Ib  : ATATTACTTGGATAAAAAT---ACTTTAAGAAAATCAATTTTTTTGTTTGG--TACCTTCATATTAGTTCTTTTATCAAG
NUF1032 II  : A-ATGGTTGGAGCAACAGA---GTTTTGT----ATTGATATGTATGTATCATCTCGGGGGATGCTAGCCATATTT-----

                     730       740       750       760       770       780       790       800
KRS02083 Ia : TTCAAATTCTGGA-CTGTTGACTATAATTGCTATAATTCTTATTCTGTTTATGCAGGAA---AAAAAATATAAACT--AA
NUF1003 Ib  : ATCCAACTCGGGAATTATTG-TTATGGGAACAATGATCTTTTATATGTTGAT-TAAAAATTCAAAAAATAAATGGTTGAA
NUF1032 II  : -----ATTCT--ATTTATTGGTT-----CACTATA---TTTATTCT-TTTATAAAGAAATT-AATAAAAAAAGTCGGAAA
```

```
                     810       820       830       840       850       860       870       880
KRS02083 Ia : TACGTGGGATTCTA--ATCTTTTTAATATTTCTATTAT--------TTTTATATTTAATCGGTAAGTATTCTCCTGGATA
NUF1003 Ib  : TATATTATTTTCTGCGGTATGTTTATTGCTAATATTAT--------TTTTG-ACTTCAC-----AATTTCTACCAGGTTA
NUF1032 II  : TATATTGATTT------TATTTTTATTTTAACTATAGTAACAAGTGTTCCATACTTGAT--------TTCGGTAACAGTA

                     890       900       910       920       930       940       950       960
KRS02083 Ia : TCTTTCAAGATTTTCTTC---TACATTGCAGTCATTAATTACAGATCAAAGTA---CTAACCAAACTAAAGCAATGAATG
NUF1003 Ib  : CTTTGAAAGATTTTCAAA---TGTAATAAATACCATAACTGGAAATAGTAGTATTTCTAATAATATTGATGCTTTAAATG
NUF1032 II  : TC----AAGATTTTCATCCTCTGCTATTAGTTCATTGATT--------TATTATT-TTGGGCAAAC----ACCATATATG

                     970       980       990      1000      1010      1020      1030      1040
KRS02083 Ia : GTAGGATT-GAAGT--GTTCCATAGCGCCTTATATCATATAAAGAATCATTTTTTCATTGGAAATGGATTTGGA-TATAG
NUF1003 Ib  : GTCGAGCT-CAGGTTTATTCGAT--TGCTAGAAGTTTAATTAGCAAGGATTTTCTTTTTGGAATTGGTTTTGTT-TATCG
NUF1032 II  : TTCTCATTAGAAGT----------------TAAATCAT-TGAAAAAACTTATGTTTGGTGAATATGGATTTGGTGCACTG

                    1050      1060      1070      1080      1090      1100      1110      1120
KRS02083 Ia : --AGAAAGAATGCCTAGCTATTTAATGACTCATAA--TTGGTTATTAGAATATCTAATTACAGGTGGAATAATATCATTT
NUF1003 Ib  : --AAGTTATATGCCAAGTCTGCTAATGACTCATAA--CTGGATTTTAGAATCTTTGATTACAGGTGGTATTTTTGCATTA
NUF1032 II  : GCAGGTGGGATG--AAATTTTCAGATGAACTAGGAATTTGGGTTCGTGGATTTTATACTTTTCTTGGT-TGGTTATATGC

                    1130      1140      1150      1160      1170      1180      1190      1200
KRS02083 Ia : TTACTAAAAAT--ATTT--ATCTTTATT-TTACAATT---T--TTAAAACTTGTAACTGTCAAAAATAATGCACTAAAAC
NUF1003 Ib  : ATGATAAGAAT--AAT---ACTTTTATG-CCAAATTT---TGTTTGAAAATATGAAGTTTCGAGA----TGTTCAAATAC
NUF1032 II  : CGACTGGGGATTTGTTGGTATTTTTATTATTGGAATTGCTTGCTTGATATTCTTCACTATAATGA----TAAATAAAC-C

                    1210      1220      1230      1240      1250      1260      1270      1280
KRS02083 Ia : AAGGCTT--AATTATT---------TCATT--T---GTATTTG----TTCTAATTCAAGGTTT----------AGTTGAA
NUF1003 Ib  : AAAATTG--TATTAGTG--------TGATTATT---ATATTTT----CATTAATCCAAGGGCT----------GGTTGAA
NUF1032 II  : AAAGCTTGATATTAGTGATACTTTTTTACTACTTGGATATTATAAATTACTCATTCAGGGTGTCTTCACCATGGGTAGAA

                    1290      1300      1310      1320      1330      1340      1350      1360
KRS02083 Ia : CCATCATTTGGTTCTCCATTATTTGAGTTGATTTTTGCATTAATAATTGGATTT----------GGTACAAATACGCTTT
NUF1003 Ib  : CCATCTTTTGGAGGACCGGTTTTCGAGTTAGTTTTTTGGCTAATCATGAGCTTTTT--------AATTCAAAT-CGGTAG
NUF1032 II  : CAAAGATATATTCAATC-TTAATCAGTTTGATTATTTATGTATTAATAAAATTTATAGTCGAGAGATTTAGATTTATTAT

                    1370      1380      1390
KRS02083 Ia : ACGAGGAG-AAAAATATTTA-TGATTAA---
NUF1003 Ib  : AAAACCAGTAGGAATGTTCAATCAGCAATAA
NUF1032 II  : AAGA--AGGAAACAGGTTTACTAA-------
```

S17.2 Multiple amino acid sequence alignment of Wzy. (Polymerase)

```
                      10        20        30        40        50        60        70        80
KRS02083 Ia : MN----NSQIFD----KSHNNILTTRITLVLLVIFL-----TIYGSGEFKVPSFSINVAVKYIILTF----GLFFSVFFI
NUF1003 Ib  : MD----NFKIGD----SSLNKKTNHHNKVLIFSLLLIIFIINFTGLDIKFIPVLQMSNFNKYISILLIIVFSIFLSIFQD
NUF1032 II  : MTPIILYSILLGILLLYCLINKTNYLGLALVAVYFI---------ISLITLLVPTRGYLKYSYVLD---------GSNIT

                      90       100       110       120       130       140       150       160
KRS02083 Ia : KVKKF-SIIKLDIISQLLLIHLLYNTIIIVLNSSDNYVLSYFTYVFAFLG---YLWGSNS-LIDKKTFKIFLSSFYYILI
NUF1003 Ib  : NKLKF-DQISIILFCQL--LYYFCVLLFSSSTDYSLLFFSYTFSFFA------YIWINNN-KISKDHLLIFSRVFLSILI
NUF1032 II  : YGIVLFMVFSFAIFFFPFLIKK-------KSFNVKNIELNVHENYIIFAKIFIFFALITVLVYMVKIIPLIQSGQ-----

                     170       180       190       200       210       220       230       240
KRS02083 Ia : IQTLFTVYIVSH----------------SVVDMYLFKNGIVIPIGASNGITTF----IVMIFPILY---KLSNSRTS-Q-
NUF1003 Ib  : ISIIITVYNLSKLQIPLFLFKSQIILPIGSSNAITTYTFLLLPIVYYLDKNTL----RKSIFLFGTFILVLLSRSNS-GI
NUF1032 II  : -----WASNRLVM---------------NSDQAIIPYNNIFEKIAILFTSYTQLLAIVVGFNLLRT---NQHYMLGSIVL

                     250       260       270       280       290       300       310       320
KRS02083 Ia : -YFLTIFTMIFAVLSRSNSGLLTII--AIILILFMQEKKYKLIRGILIFLIFLLFLYLIG-KYSPGYLS-------RFSS
NUF1003 Ib  : IVMGTMIFY--MLIKNSKNKWLNILFSAVCLLLILFLTSQFLPG------------YFERFSNVINTIT-------GNSS
NUF1032 II  : LMVGATEFCIDMYVSSRGMLAIFILFIGSLYLFFY---------------------KEIN-KKSRKYIDFIFILTIVTSV

                     330       340       350       360       370       380       390       400
KRS02083 Ia : TLQSLITDQSTNQTKAMNGRIEVF-HSALYHIKNHFFIGNGFGY-RERMPSYLMTHNWLLEYLITGGIISFLLKIFIFIL
NUF1003 Ib  : I--------SNNIDALNGRAQVY--SIARSLISKDFLFGIGFVY-RSYMPSLLMTHNWILESLITGGIFALMIRIILLCQ
NUF1032 II  : P-----YLISVTVSRFSSSAISSLIYYFGQTPYMFSLEVKSLKKLMFGEYGFGALAG----GMKFSDELGIWVRGFYTFL

                     410       420       430       440       450       460       470       480
KRS02083 Ia : QFLKLVTVKNNALKQGLIISFVFVLIQGLVEPSFGSPLFELIFALII--GFGT---NTLYEE------------------
NUF1003 Ib  : ILFENMKFRDVQIQNCISVIIIFSLIQGLVEPSFGGPVFELVFWLIM--SFLIQIGRKPV--------------------
NUF1032 II  : GWLYADWGFVGIFIIGIACLIFFTIMINKPKLDISDTFLLLGYYKLLIQGVFTMGRTKIYSILISLIIYVLIKFIVERFR

                     490
KRS02083 Ia : ----KNIYD-
NUF1003 Ib  : ----GMFNQQ
NUF1032 II  : FIIRRKQVY-
```

S18.1 Multiple nucleotide sequence alignment of *wzx*. (Flippase)

```
                         10        20        30        40        50        60        70        80
KRS02083 Ia : ATGAATTTTAAGAATATAATAAAAAATTTCCTAGGCGTTGTCATAAGTAATTTTTCATCCATTATCGCTGGTGTAATAGT
NUF1003 Ib  : ATGAATTATAGAAATACAGTTAGAAGTTTTTTTGGAGTTGTAGTAAGTAATTTTTCCTCAATAGTTGCAGGAGTAATTGT
NUF1032 II  : ATGAAATTTA---------TAAGAAATTTACTTGCAGTTTTAATGAGTAATGCATTCACGGTTTTTTCTGGTTTACTAGT
              start

                         90       100       110       120       130       140       150       160
KRS02083 Ia : GGGATTTGCTGTTCCAAAATTTTTATCTGTCGAGGAATATGGACTTTTAAAAACATTTACCTTGTATATATCTTATTTAG
NUF1003 Ib  : TGGCTTTGCAATACCAAAAATTTTATCAGTGGAAGATTATGGTTACTTAAAGACTTTTACTTTATATGTGTCTTATATGG
NUF1032 II  : TGGTCTTGTATTACCAATGATTCTTACTATTGATGACTTTGGATATTTTAAAACTTTTACATTATATCTAACATATTTGG

                        170       180       190       200       210       220       230       240
KRS02083 Ia : GATTATTTAGCTTTGGTATTATCGACGGAATAGTTTTAGAATATGGTGGAAAAAATTATGATGAACTTGATAGACCCGTA
NUF1003 Ib  : GACTATTTAGTTTCGGAATTATTGATGGAATTGTTCTAGAATTTGGTGGAAAAAGCTATGATGAGTTAGAAAGATTGACC
NUF1032 II  : GTTTATTCAGTATTGGAATTATAGATGGAATTGTTTTAAAGTATGGTGGAAATGATTTTGAAGACTTACAAAAAGAGAAC

                        250       260       270       280       290       300       310       320
KRS02083 Ia : TTTCGTAACTATTTCAGATGGTTTCTAATCGTAAGTTTAATAAGTGCATTAATAATATTATTTACTGCAACAATGTTTTT
NUF1003 Ib  : TTCCGTAATTACTTTAGATGGTTTTTCGTTGTAACTTTTATTAGTGCTCTCGTAATAGCTCTGGTTTCCATTATATGTTT
NUF1032 II  : TTCAGAAGTTACTTTAGATGGTATTTAATTGTTCATATCATTGTTTCTATTCCAATGTTGCTATTGTCAACTCTAGAATC

                        330       340       350       360       370       380       390       400
KRS02083 Ia : CAAAACTGATCAAGCGTTCATATTAATTGCGCTTGCATTTAATTTGTTAGCTATAAACATTTCAAACTATTTTCAACAAA
NUF1003 Ib  : TGACAGTGACCTATCATTCATTCTTTTTGCAATTTCTTTTAATTTAATTGCAATCAATACTTCTAATTATTTTCAACAGA
NUF1032 II  : AAATACAAATATAAGATTTATTATATTTGCACTCATAATTAATATGATTTCTTTAAATGTTATAGGTTATTTTAGACAGA

                        410       420       430       440       450       460       470       480
KRS02083 Ia : TATCTCAAATCACGCAGAGATTTAAAGAATATTCTCTCAGAAAAATTTTACAA--AGTTTTTCAAATGTTTTACTTGTAT
NUF1003 Ib  : TTTCCCAAATTACTCAAAGATTTAAAGAGTATTCATTAAGAAAAATTTTACAA--AGCTTTTCAAATATCTTATTAGTAA
NUF1032 II  : TATCAGAAATAACGCAACGATTTAAAGAATATACAATTATCAAAATCGCTCAATCAATTTTT--AATAT---ACTAACAG

                        490       500       510       520       530       540       550       560
KRS02083 Ia : TTCTCTGTTATTTTCTCTTCAAAAATAAT----CAGCACGTTAGCTATCGATTTTATATTATATTATTGGTTATTATTAA
NUF1003 Ib  : TACTTTGTTTTATCCTTTACAAAACTCAC----TATGATGTAAATTATAAATTTTATATTATAATGTTAGTCCTTATCAA
NUF1032 II  : TTGTCGGCCTGTTTTTATTGAAGAGTAACGGAATAGCA-GTAAACTTTAAAATTTATATTTTACTAGTTATTATAGCTAA

                        570       580       590       600       610       620       630       640
KRS02083 Ia : TTACATATTATGTATTTGGTATGTTTATACCTATCGAAAAATTGTTTTTGGTGAAAAAATATC-TTTTATATTA-----G
NUF1003 Ib  : CTTTGGATTATGTTTGTGGTATGTATATACTTATAGGGAAATAATTTTTGGTGATAAAGTTTCATTTTTTGATA-----G
NUF1032 II  : TTTCATTGTAACTCTGTGGTATGTATACCTTTATCAAGAAATAATTTTTGGAAAATCAACTCCACTTAATGATACATTTG

                        650       660       670       680       690       700       710       720
KRS02083 Ia : GAAAGAAAGATATTTTTTTATTAGTGAAAACAGGATTACCCTTATTAATAGCAAATATATGTTCAGTTTTAATCGTTACA
NUF1003 Ib  : CAAA-AATGACATAATATTTTTAATAAAGACAGGTGTTCCATTGCTGATAGCTAATATTAGTTCTGTTTTAATCGTCACC
NUF1032 II  : CAAC-AGTAAAAAAATTCTCT-----AAAATTGGAATTCCACTGATGTTTGCAAATCTTATTTCAACATTAATATTAACT

                        730       740       750       760       770       780       790       800
KRS02083 Ia : ATAGATAGTCAATTTGTAAATACTCTTTTTTCAACTCGAGATTATGCTATGTATGCATTTGCCTATAATTTGTTATCATT
NUF1003 Ib  : ATTGATAGTCAATTTGTAAATACACTTTTTTCTACTAGAGAATATGCAATGTATGCATTCGCTTATAATTTACTTTCATT
NUF1032 II  : CTTGATAGACAGTTTGTTAATATTTTATTTTCCAATAAAATTTATGCCATCTATGCATTTGCATTTAACTTACTTTCTAT
```

```
                     810       820       830       840       850       860       870       880
KRS02083 Ia : AATTACTATTGCAACAGCAGCTATTTCAACTGTTTTATATCCTACTTTAAAAAGAACAGATGAAACATTAATAAAACAAA
NUF1003 Ib  : GATAACGATTGCAACAGCAGCAATATCGACGATTTTATATCCAACTCTCAAAAAGAACTGAAGAGACACGTATAAAAGATA
NUF1032 II  : CTTAACTCTTGCAACAGCTGCATTTTCTACAGTGCTTTATCCAAGTTTAAAAAGATCTGATGTGACCAAAATTGGTGGAA

                     890       900       910       920       930       940       950       960
KRS02083 Ia : ATTATG-GCTATCTA-GTAAGCATC-ATTGAGATTGTTATTTTTGGAGCATTATTCGTATTCTTTCCACTTTCTATTTTT
NUF1003 Ib  : ACTATG-GGAATCTA-ATTAGCATT-TTAGAAGTATTAATATTTGCTTTTCTTATTGCTTTTTTCCCATTATCTATCTTT
NUF1032 II  : AATATAAGAAATTTACATTTTTATCTATTAGTATAGTT-TTTTTG--TTACTTAGTTTTTATTTTCCAATGAAGATACTC

                     970       980       990      1000      1010      1020      1030      1040
KRS02083 Ia : GT-AAATTGGTTTTTACCCAATTATAATGAATCTCTTGAAATTTTCAGAGTGATATTTCCTGGTGTTGCACTCACTACAC
NUF1003 Ib  : GT-AAATTGGTTCTTACCAAACTATAGCGAATCACTTGAAATTTTTAGAATAATATTTCCAGGTGTTGCTTTAACGACGC
NUF1032 II  : ATTGAGGCTATTCTT-CCAAAATATATTAGTTCTTTAGTAATTTTTAGAGTTATTTTTCCAACTTTACCTATTACCACAA

                    1050      1060      1070      1080      1090      1100      1110      1120
KRS02083 Ia : CCATTGTAGTCATTATGCATAACTATTACAAAACTTTAAAGAAAAGT--AATCTATATTTTTATAAGAGTATTATGGTTC
NUF1003 Ib  : CTATAGTAGTCATTATGCATAACTACTATAAAACTTTAAAAAAAAGC--AATCAATATTTCTTTAAAAGCATCATGGTTT
NUF1032 II  : CAATAACTGTAATTATCAATAATTATTTTAAAACTTTAGGAAAAAGTGTAATT--TACTTTAATAGAAGTATTGTGATAT

                    1130      1140      1150      1160      1170      1180      1190      1200
KRS02083 Ia : TCGTATTTTCTAT--GATTGCAAATTATATTGCATATTATCTTTTTAAAACGACAATAGCAATTTCTGCAGCTTCAATAG
NUF1003 Ib  : TAATTTTTTCAAT--GATCGCTAATTTTATTGCTTATTACTTCTTTAAAACAACTATAGCTATTTCAGTTGCATCAATTG
NUF1032 II  : TAATACTTTCGATTGGACT--CAATCTTTTGGCTTACTTAATTTGGAAAACACCAATTGCAATTTCAGCTTCATCAGTAC

                    1210      1220      1230      1240      1250      1260      1270      1280
KRS02083 Ia : TTGTTCTTTTCTTATGGTATTTATATGTTGAACAAGAATTC-GTAAAATCATTTAACTATAAAT-CTTTAAAAAATCTTT
NUF1003 Ib  : TAGTTTTATTTCTTTGGTATGTTTATGTGGAACGAGAATTT-GTTAAATCTTTTAATTATAAAT-CACAGCGAAATTTAT
NUF1032 II  : TTACAGTATTGATTTGGTATGTTGTCGTCGA-TAATTATTTAGCCAAAGAATTAAAAT-TAAATACGTCAAAAAAATTAA

                    1290      1300      1310      1320      1330      1340      1350      1360
KRS02083 Ia : CATATATTTAATTTAATGTGCTCGTTTTATCTTTGTACATTTCTTCCAAACATATATTTAGG-ATGTATTAGTTATAT
NUF1003 Ib  : CATATATTTATTATTAATGACTATATTTTATTGTTGTTCTTTCATAACTAATTTATACATTGG-GTGTTTCGTTTATAT
NUF1032 II  : TAATGATATTATTATTTACAATAATTTTTATTCGACTTCAATGATTCAAAATTGGCTGATCGGTATG-ATCATCTACAT

                    1370      1380      1390      1400      1410      1420      1430      1440
KRS02083 Ia : CATAGTTTATTTGATATTTACTTTACTTTATTTTAAAAAATTAA-TAATCA-ATATCAGACAAAAATTCACATTAAA-AA
NUF1003 Ib  : TGTTTTATATTGTTTAGTTTCATTGATCTATTATAAAAAATTAA-TTGTTA-GCATCTATAATAAGATTACTTCAAATAA
NUF1032 II  : TGTACTTT-TTGG--AGTTACATTAA-CGATCATTAATCCTGAAGTTATTACATATTTGAAAAAG----ACAAGAGG-AA

                    1450
KRS02083 Ia : ATATTAGGTAA
NUF1003 Ib  : ACAT-AGGTAA
NUF1032 II  : ATAAAGGATGA
                      stop
```

S18.2 Multiple amino acid sequence alignment of Wzx. (Flippase)

```
                      10        20        30        40        50        60        70        80
KRS02083 Ia : MNF-KNIIKNFLGVVISNFSSIIAGVIVGFAVPKFLSVEEYGLLKTFTLYISYLGLFSFGIIDGIVLEYGGKNYDELDRP
NUF1003 Ib  : MNY-RNTVRSFFGVVVSNFSSIVAGVIVGFAIPKILSVEDYGYLKTFTLYVSYMGLFSFGIIDGIVLEFGGKSYDELERL
NUF1032 II  : MKFIRNLLAVLMSNAFTVFSGLLVGL-VLPMILTIDDF---GYFKTFTLYLTYLGLFSIGIIDGIVLKYGGNDFEDLQKE

                      90       100       110       120       130       140       150       160
KRS02083 Ia : VFRNYFRWFLIVSLISALIILFTATMFFKTDQAFILIALAFNLLAINISNYFQQISQITQRFKEYSLRKILQSFSNVLLV
NUF1003 Ib  : TFRNYFRWFFVVTFISALVIALVSIICFDSDLSFILFAISFNLIAINTSNYFQQISQITQRFKEYSLRKILQSFSNILLV
NUF1032 II  : NFRSYFRWYLIVHIIVSIPMLLLSTLESNTNIRFIIFALIINMISLNVIGYFRQISEITQRFKEYTIIKIAQSIFNILTV

                     170       180       190       200       210       220       230       240
KRS02083 Ia : FLCYFLFKNNQHVSYRFYIILLVIINYILCIWYVVTYRKIVFGEK---ISFILGKKDIFLLVKTGLPLLIANICSVLIVT
NUF1003 Ib  : ILCFILYKTHYDVNYKFYIIMLVLINFGLCLWYVVTYREIIFGDK---VSFFDSKNDIIFLIKTGVPLLIANISSVLIVT
NUF1032 II  : VGLFLLKSNGIAVNFKIYILLVIIANFIVTLWYVVLYQEIIFGKSTPLNDTFATVKKF---SKIGIPLMFANLISTLILT

                     250       260       270       280       290       300       310       320
KRS02083 Ia : IDSQFVNTLFSTRDYAMYAFAYNLLSLITIATAAISTVLYPTLKRTDETLIKQNYGYLVSIIEIVIFGALFVFFPLSIFV
NUF1003 Ib  : IDSQFVNTLFSTREYAMYAFAYNLLSLITIATAAISTILYPTLKRTEETRIKDNYGNLISILEVLIFAFLIAFFPLSIFV
NUF1032 II  : LDRQFVNILFSNKIYAIYAFAFNLLSILTLATAAFSTVLYPSLKRSDVTKIGGKYKKFTFLSISIVFLLLSFYFPMKILI

                     330       340       350       360       370       380       390       400
KRS02083 Ia : NWFLPNYNESLEIFRVIFPGVALTTPIVVIMHNYYKTLKKSNLYFYKSIMVLVFSMIANYIAYYLFKTTIAISAASIVVL
NUF1003 Ib  : NWFLPNYSESLEIFRIIFPGVALTTPIVVIMHNYYKTLKKSNQYFFKSIMVLIFSMIANFIAYYFFKTTIAISVASIVVL
NUF1032 II  : EAILPKYISSLVIFRVIFPTLPITTTITVIINNYFKTLGKSVIYFNRSIVILILSIGLNLLAYLIWKTPIAISASSVLTV

                     410       420       430       440       450       460       470
KRS02083 Ia : FLWYLYVEQEFVKSFNYKSLKNLSYILILMCSFYLCTFLPNIYLGCISYIIVYLIFTLLYFKKLIINIRQKFTLKNIR
NUF1003 Ib  : FLWYVYVEREFVKSFNYKSQRNLSYILLLMTIFYCCSFITNLYIGCFVYIVLYCLVSLIYYKKLIVSIYNKITSNKHR
NUF1032 II  : LIWYVVVDNYLAKELKLNTSKNLIMILLFTIIFYSTSMIQNWLIGMIIY--IVLFGVTLTIINPEVITYLKKTRGNKG
```

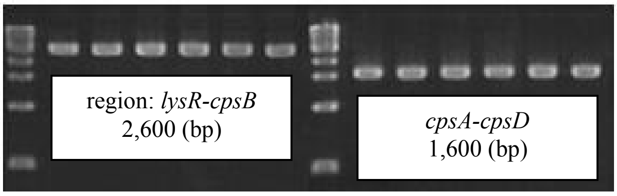
S19. *ISSdy1* insertion event in *cpsIbN* (Hypothetical protein) of NUF1117 (subtype Ic).

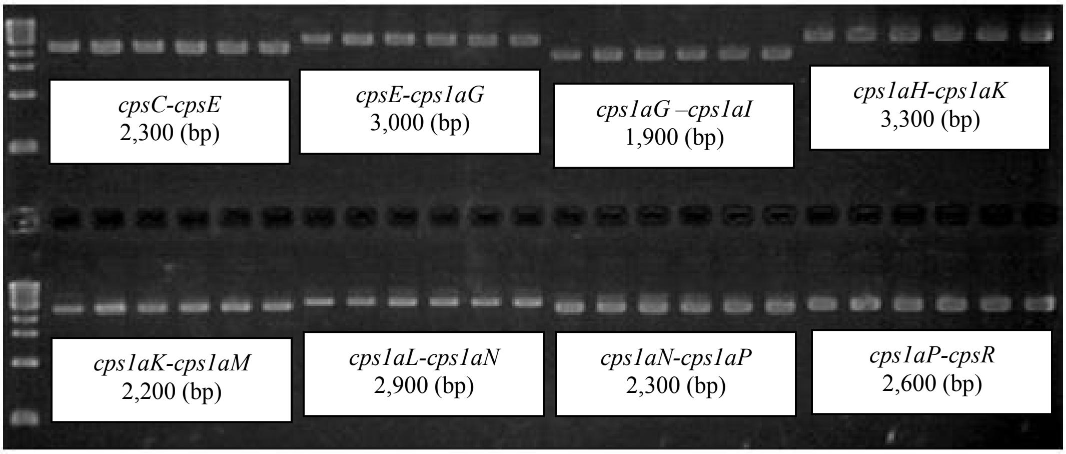
```
                     10        20        30        40        50        60        70        80
NUF1003 Ib : TATTGAATTAATAAACAATAATTTTGAATCTATTTTAGATATGGATCTTTTATCACAAGATAAAGAAATTCCACATCATT
NUF1117 Ic : TATTGAATTAATAAACAATAATTTTGAATCTATTTTAGATATGGATCTTTTATCACAAGATAAAGAAATTCCACATCATT

                     90       100       110       120       130       140       150       160
NUF1003 Ib : ATCGAAATGATAAGTATGACTTTTATTTTTTTCATGATTTTAATTCAGAAGAGACTTTGAAAGATCAATT----------
NUF1117 Ic : ATCGAAATGATAAGTATGACTTTTATTTTTTTCATGATTTTAATTCAGAAGAGACTTTGAAAGATCAATTTGAATCGGTT

                    170       180       190       200       210       220       230       240
NUF1003 Ib : --------------------------------------------------------------------------------
NUF1117 Ic : ACACTAAACTAGACAGAATTTATAAAGTGTTCTACACTAAAGAAAACAGGAGAACAGATATGTCTAGAAAAATACGTCGC
                                                                         start  ISSdy1 orfA

                    250       260       270       280       290       300       310       320
NUF1003 Ib : -------------------------------------------------------------------------------
NUF1117 Ic : CACTTCACCGATGATTTTAAGCAACAAATCGTTGACTTGCACAATGCTGGAATGAAACGAAGTGAGCTTATCAAAGAATA

                    330       340       350       360       370       380       390       400
NUF1003 Ib : CAAG----------------------------------------------------------------------------
NUF1117 Ic : TGAGTTAACGCCCTCAACCTTCGATAAGTGGGTCAGACAAGACAAACTAACGGGTTCCTTCAAGTCTGTTGATAATATGA

                    410       420       430       440       450       460       470       480
NUF1003 Ib : --------------------------------------------------------------------------------
NUF1117 Ic : CAGATGAACAACGGGAACTGATTCAACTCAGAAAACGCAATAAAGAACTCGAAATGCAATTAGACATCCTAAAGCAAGCG

                    490       500       510       520       530       540       550       560
NUF1003 Ib : --------------------------------------------------------------------------------
NUF1117 Ic : GCAGTGATTATGGCACGAAAAGACAAGTAATCACTGCAAACAAGGATAAATACAGCATTTCAGCTATGTGTCGTTGGCTG
                                        stop                                     start  ISSdy1 orfB

                    570       580       590       600       610       620       630       640
NUF1003 Ib : --------------------------------------------------------------------------------
NUF1117 Ic : AACATCCCGCGCTCAAGTTACTACTACCAAGTTGTAGATCCTGTGTCAGAGGCTGATCTCGAGGATAAAATTACCTATAT

                    650       660       670       680       690       700       710       720
NUF1003 Ib : --------------------------------------------------------------------------------
NUF1117 Ic : TTTCTTTGAGAGCAAGTCCAGATATGGGGCTAGAAAAATCAAGAAATGCTTAGAAAAGGACGGCATCATACTCTCTCGCC

                    730       740       750       760       770       780       790       800
NUF1003 Ib : --------------------------------------------------------------------------------
NUF1117 Ic : GTCGGATTCGTCGCATCATGGAGAGACTCCACTTGGTATCCGTTTATCAGAAAGCAGCCTTCAAACCGCATTCTAGAGGG

                    810       820       830       840       850       860       870       880
NUF1003 Ib : --------------------------------------------------------------------------------
NUF1117 Ic : AAGAACGAAGCTCCTATTCCAAACCGCCTAGACAGGCAATTCGACCAAGAAAGACCACTGGAAGCCCTGGTGACCGACTT

                    890       900       910       920       930       940       950       960
NUF1003 Ib : --------------------------------------------------------------------------------
NUF1117 Ic : GACTTACGTCCGTGTGGATAGGCGCTGGGCTTACGTTTGTCTCATCATTGACCTCTTTAATCGTGAAATCATAGGATTGT

                    970       980       990      1000      1010      1020      1030      1040
NUF1003 Ib : --------------------------------------------------------------------------------
```

```
NUF1117 Ic : CCGTTGGCTGGCACAAGACTGCGGAGCTCGTTAAACAAGCCATTCAAAGCATCCCTTATGCACTAACCAAGGTCAAGTTG

                   1050      1060      1070      1080      1090      1100      1110      1120
NUF1003 Ib : --------------------------------------------------------------------------------
NUF1117 Ic : TTCCATTCTGATCGTGGCAAGGAGTTTGACAATCAGCTGATTGATGAGATGCTTGAAGCTTTTGGAATCACCCGTTCTCT

                   1130      1140      1150      1160      1170      1180      1190      1200
NUF1003 Ib : --------------------------------------------------------------------------------
NUF1117 Ic : CAGTCAGGCTGGTTGTCCCTATGACAATGCCGTCGCTGAGAGTACCTATCGTTCCTTCAAACTGGAGTTTATCAACCAAG

                   1210      1220      1230      1240      1250      1260      1270      1280
NUF1003 Ib : --------------------------------------------------------------------------------
NUF1117 Ic : AAACCTTTCAATCACTAGAAGAATTAGCCCTCAAAACGAAAGATTATGTCCACTGGTGGAACTATCATCGCATTCATGGT

                   1290      1300      1310      1320      1330      1340      1350      1360
NUF1003 Ib : --------------------------------------------------------------------------------
NUF1117 Ic : AGTCTTAACTACCAAACTCCCATGACCAAACGAATTATCGCTTAAAAAGCACTTTATAAAAATTGTTCAGAAAAGTGTTG
                                                       stop

                   1370      1380      1390      1400      1410
NUF1003 Ib : -----------ACAAAATGTGAAAGATAAATACAAGAGGCGAATTAAAGGATTTT
NUF1117 Ic : CCTTTTCAATTACAAAATGTGAAAGATAAATACAAGAGGCGAATTAAAGGATTTT
```

S20. The result of PCR scanning of *cps* locus (P1-P3)

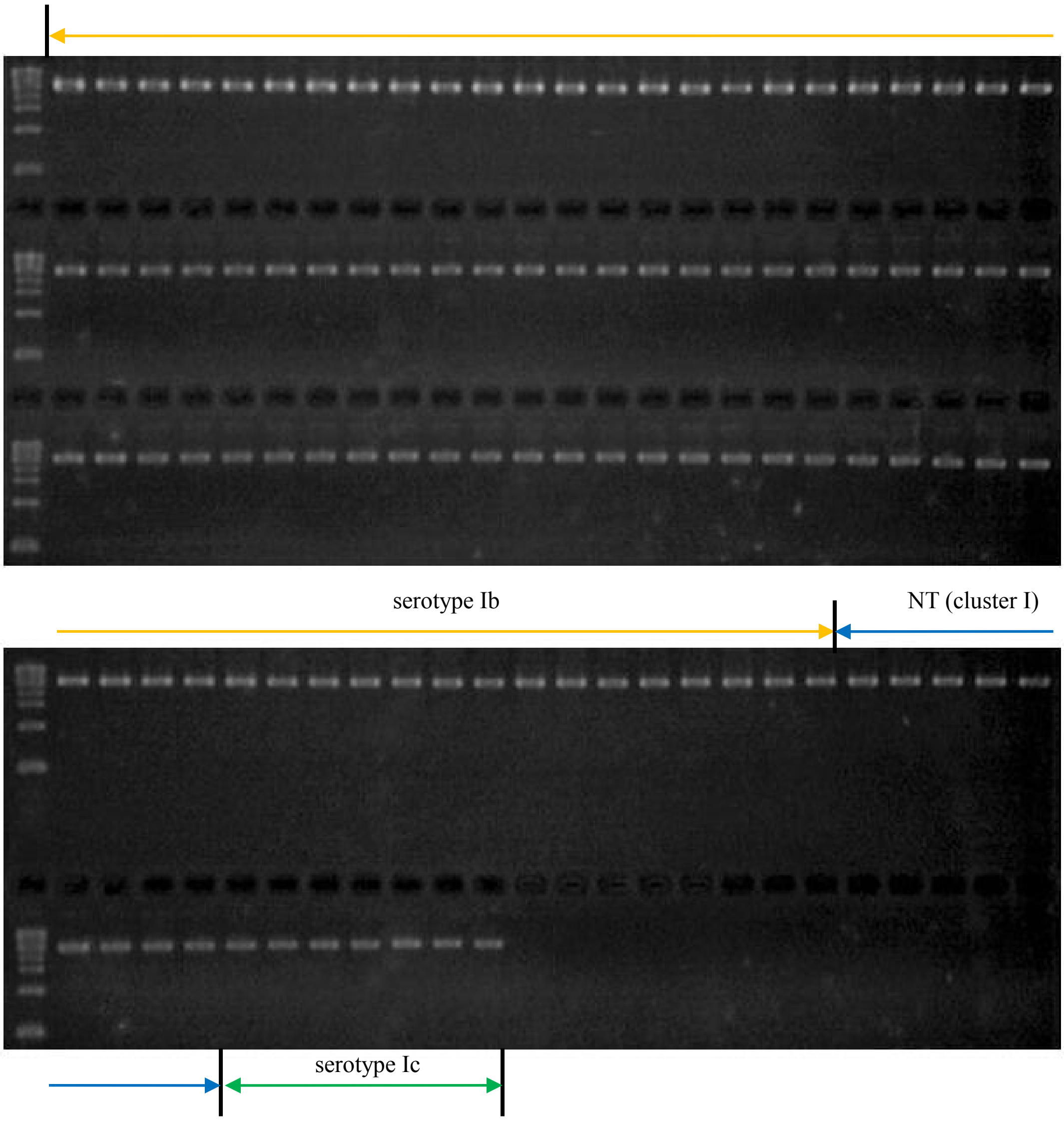
P1. PCR scanning of *cps* locus of serotype Ia. (6 strains)



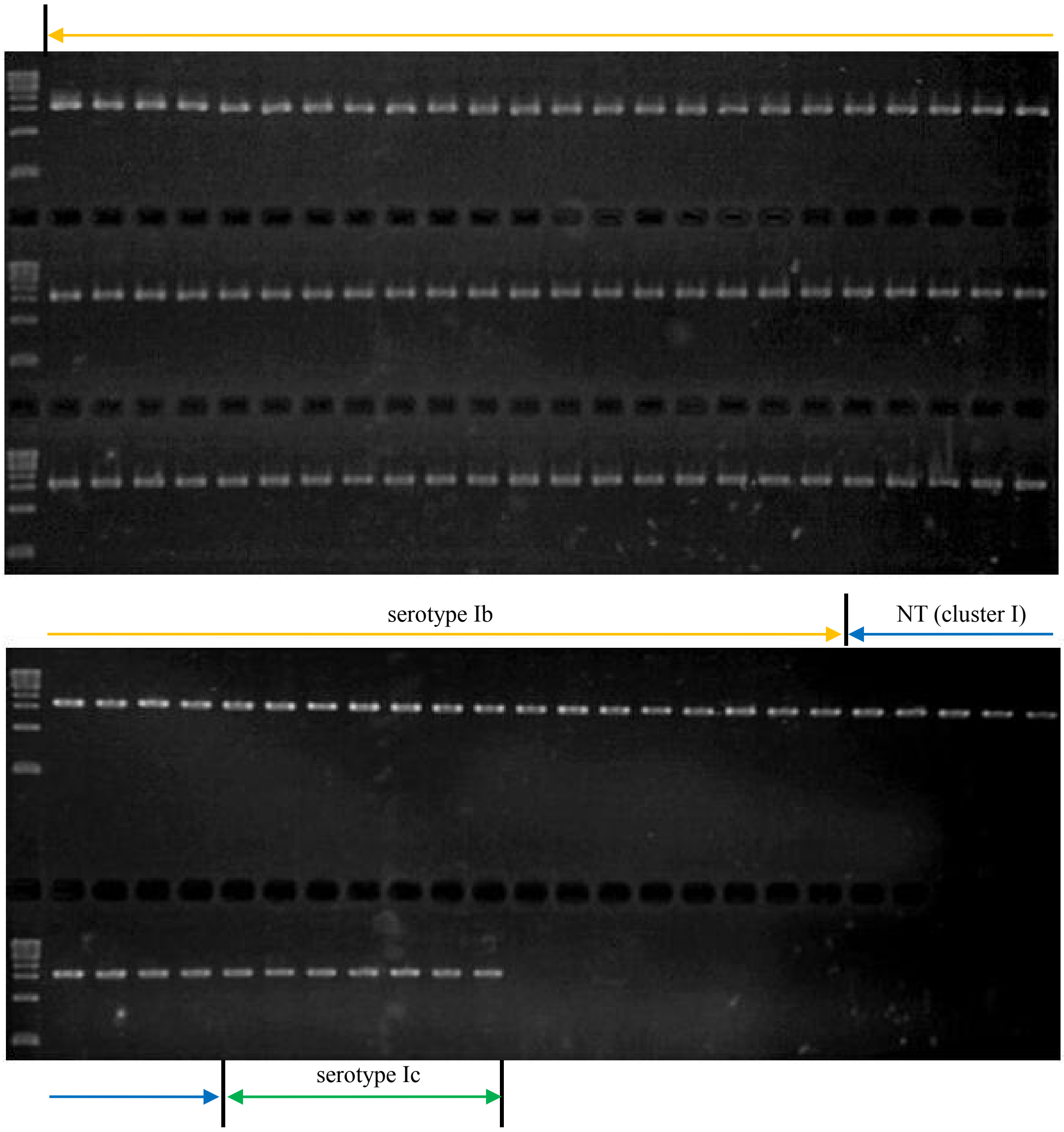


P2. PCR scanning of *cps* locus of serotype Ib, NT (cluster I) and Ic. (91+9+7 strains)

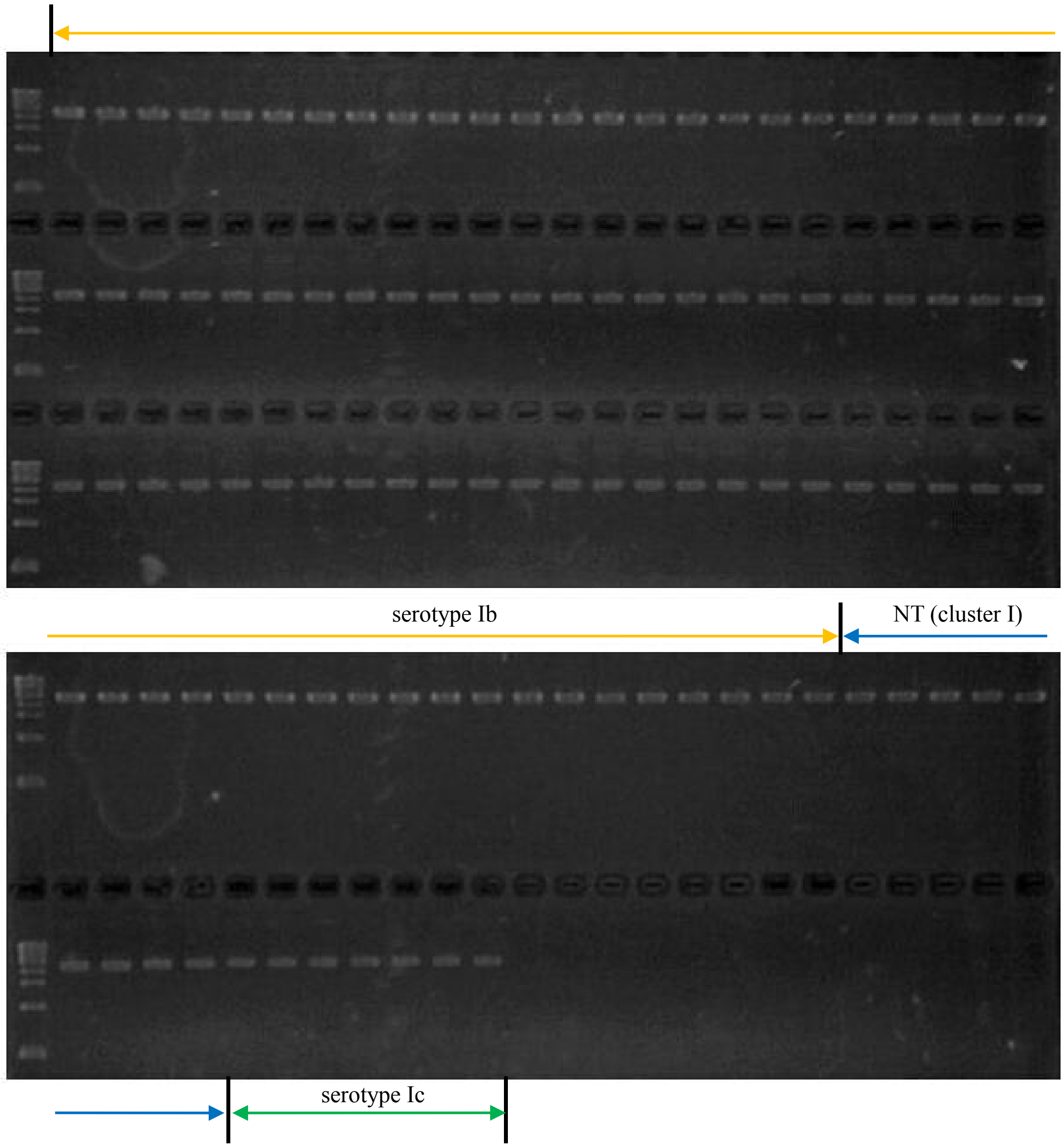
region: *lysR-cpsB* 2,600 (bp)



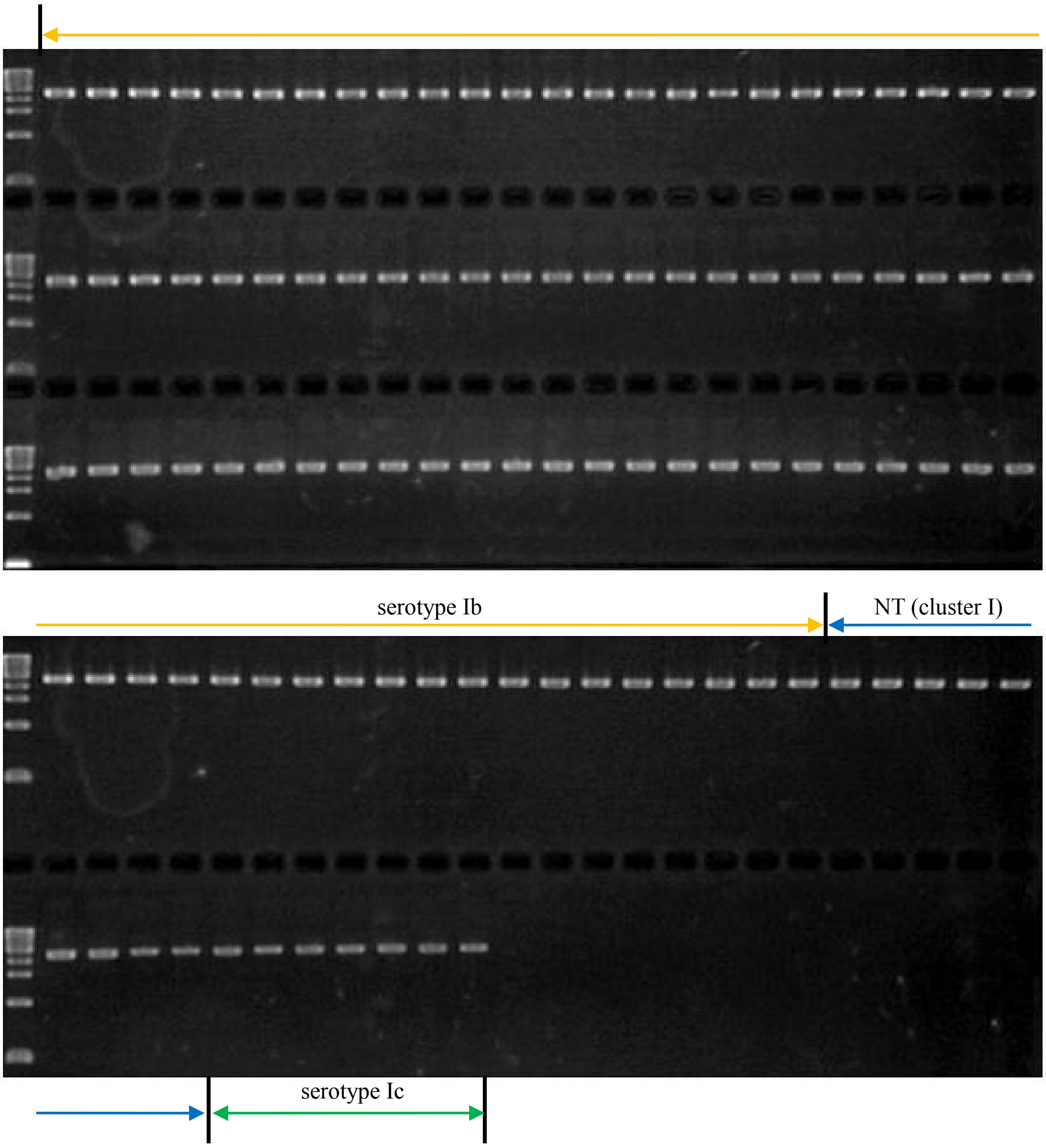
cpsA-cpsD 1,600 (bp)



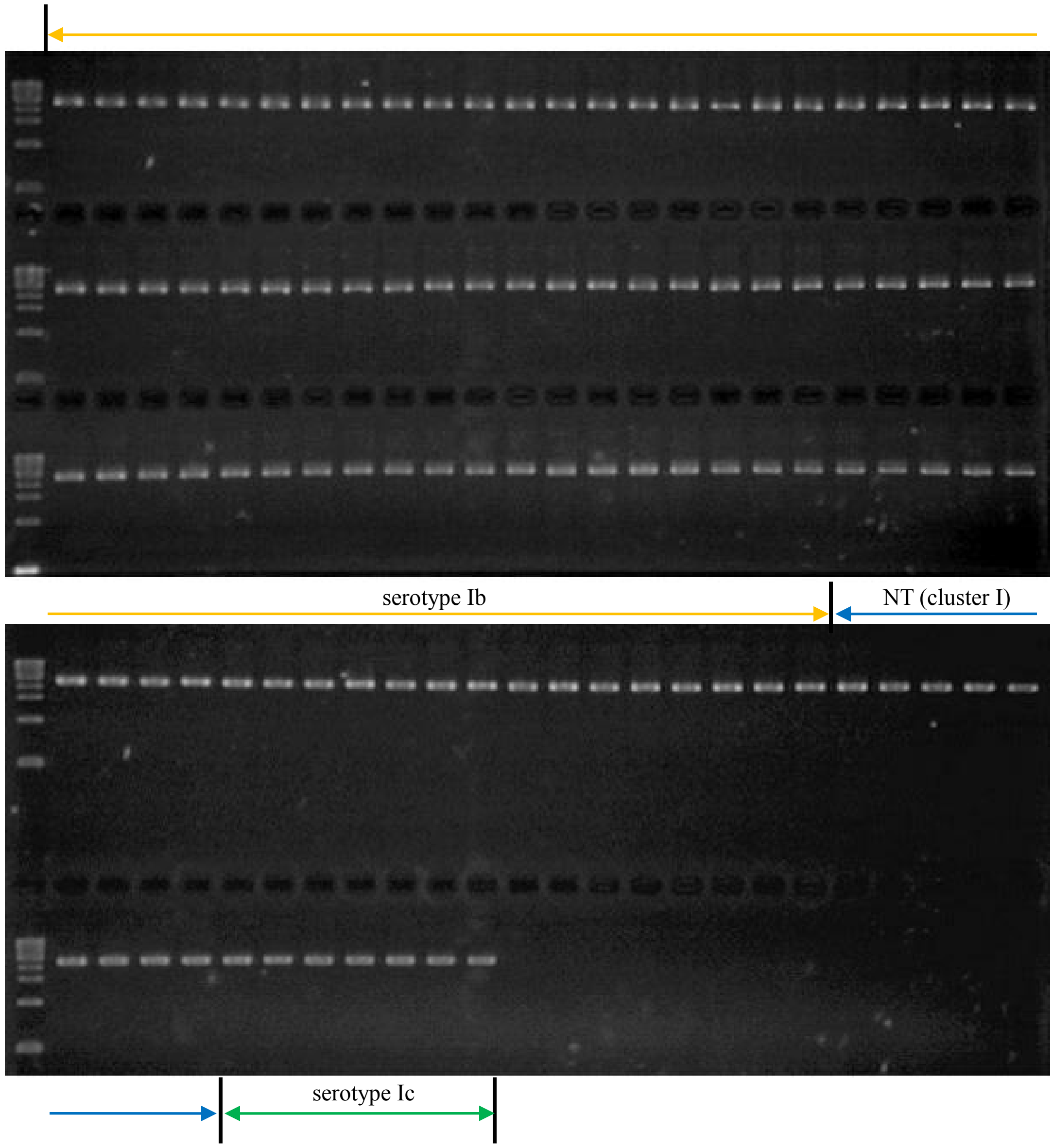
cpsC-cpsE 2,300 (bp)



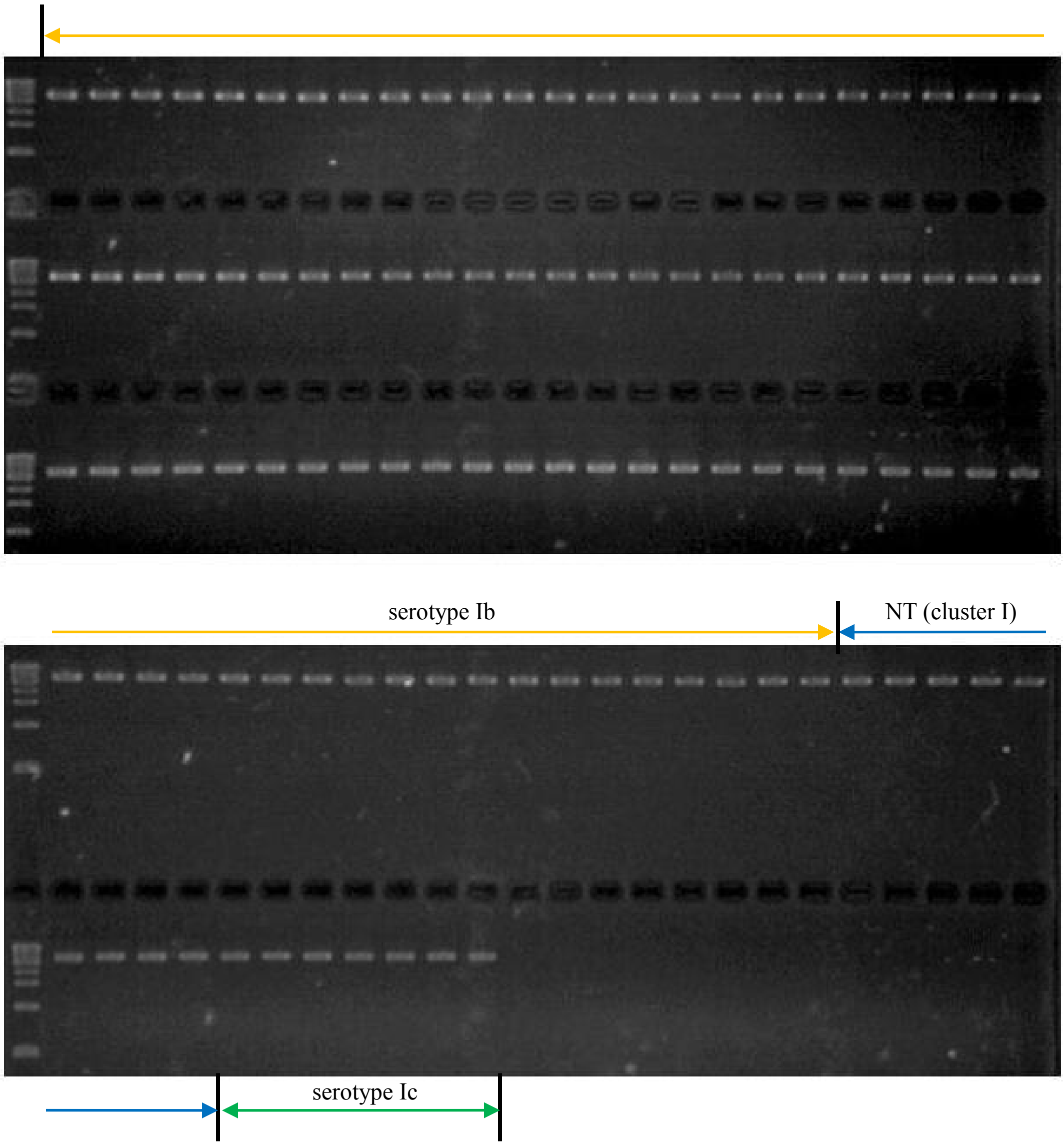
cpsE-cps1bG 2,300 (bp)



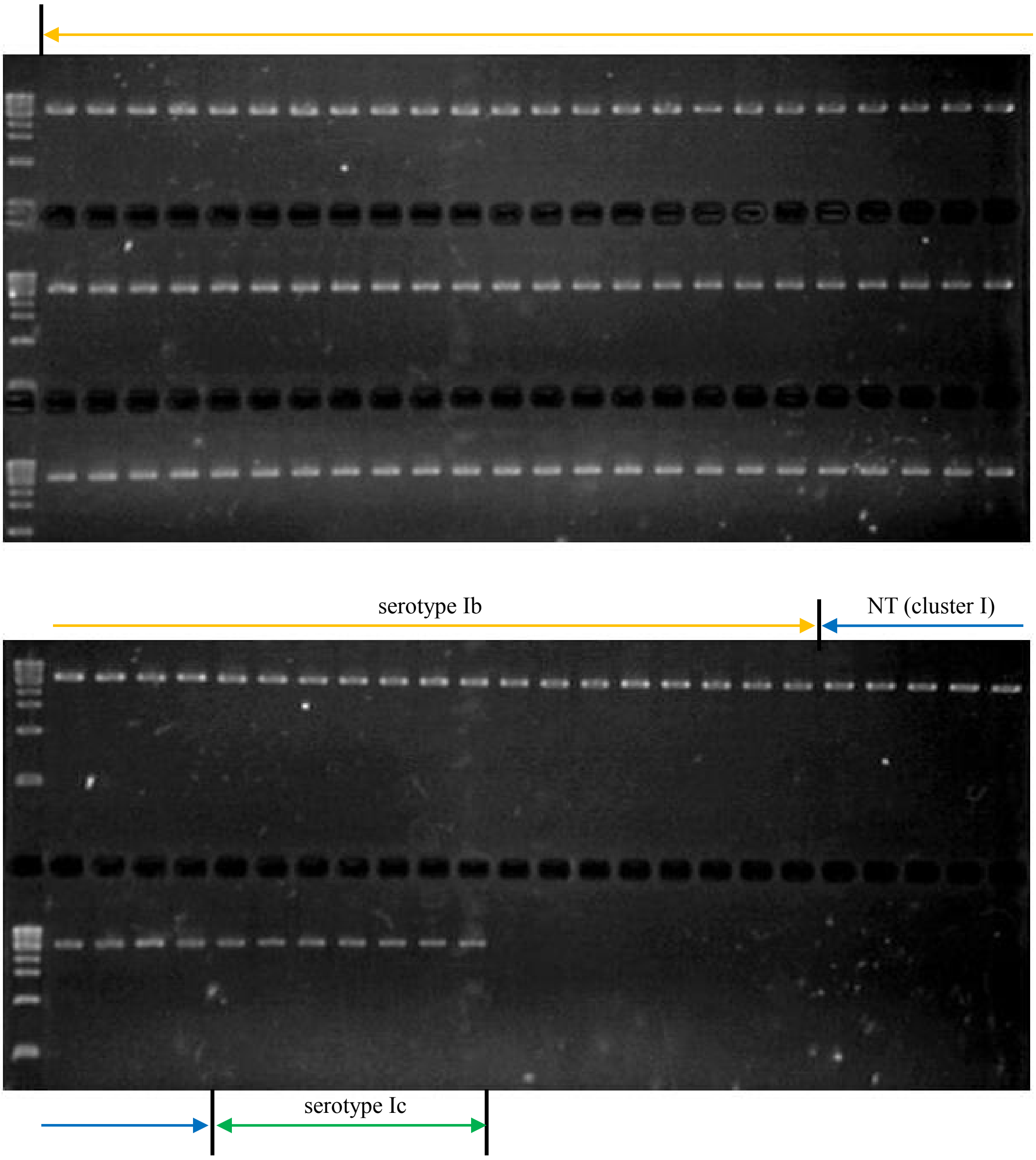
cps1bF –cps1bI 2,200 (bp)



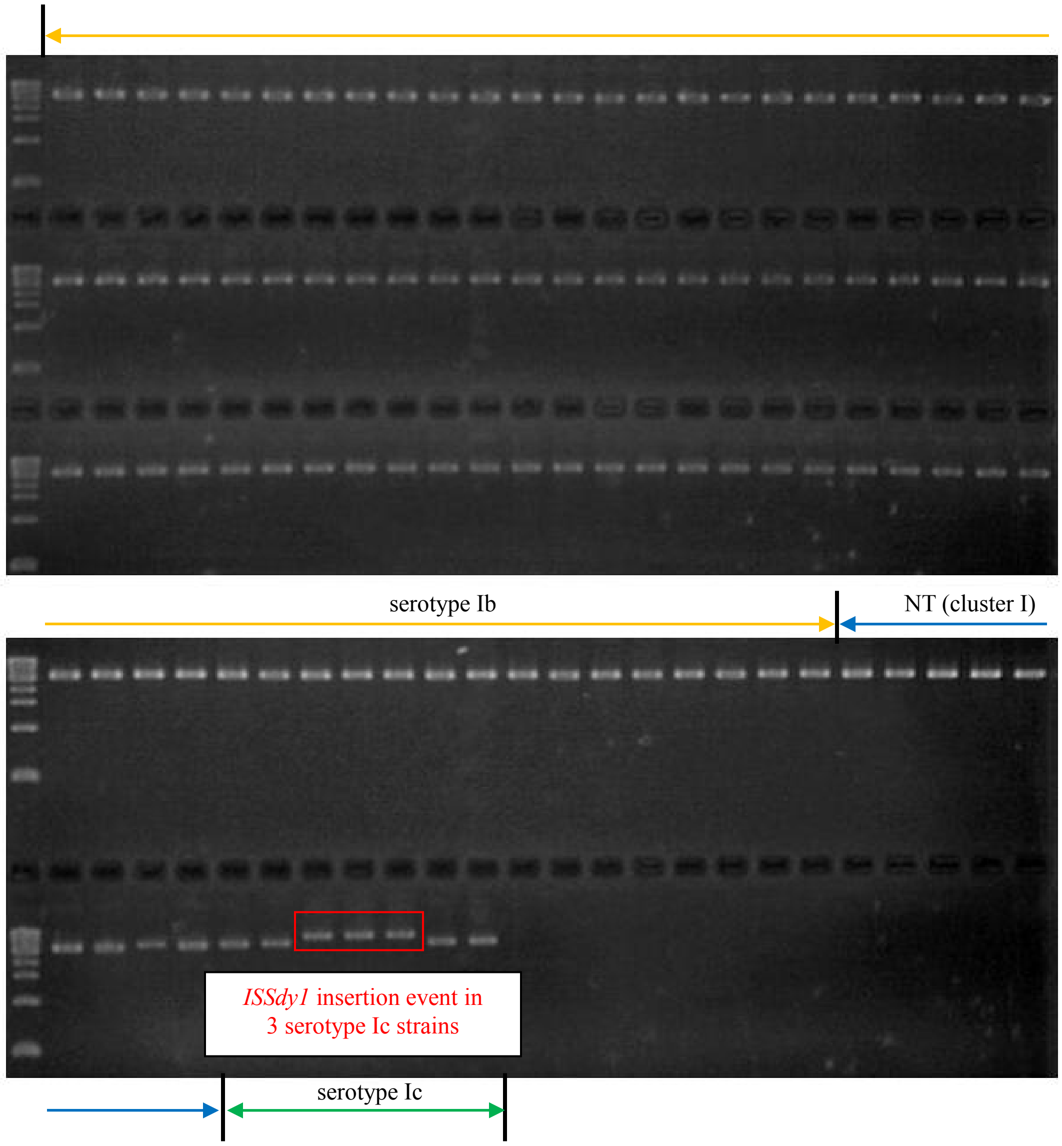
cps1bI –cps1bK 2,800 (bp)



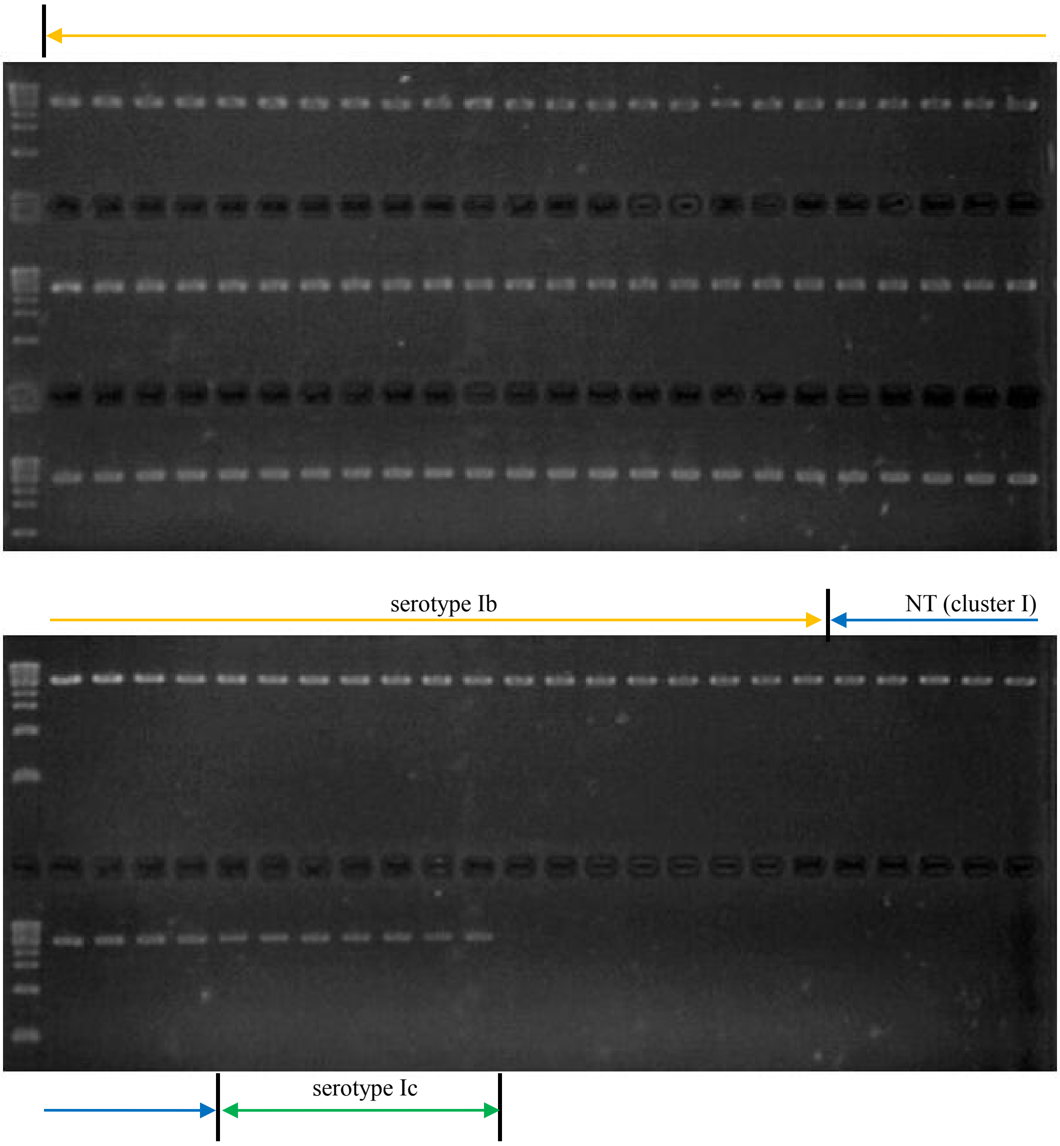
cps1bK –cps1bM 2,700 (bp)



cps1bM –*cps1bO* 2,800 (bp)

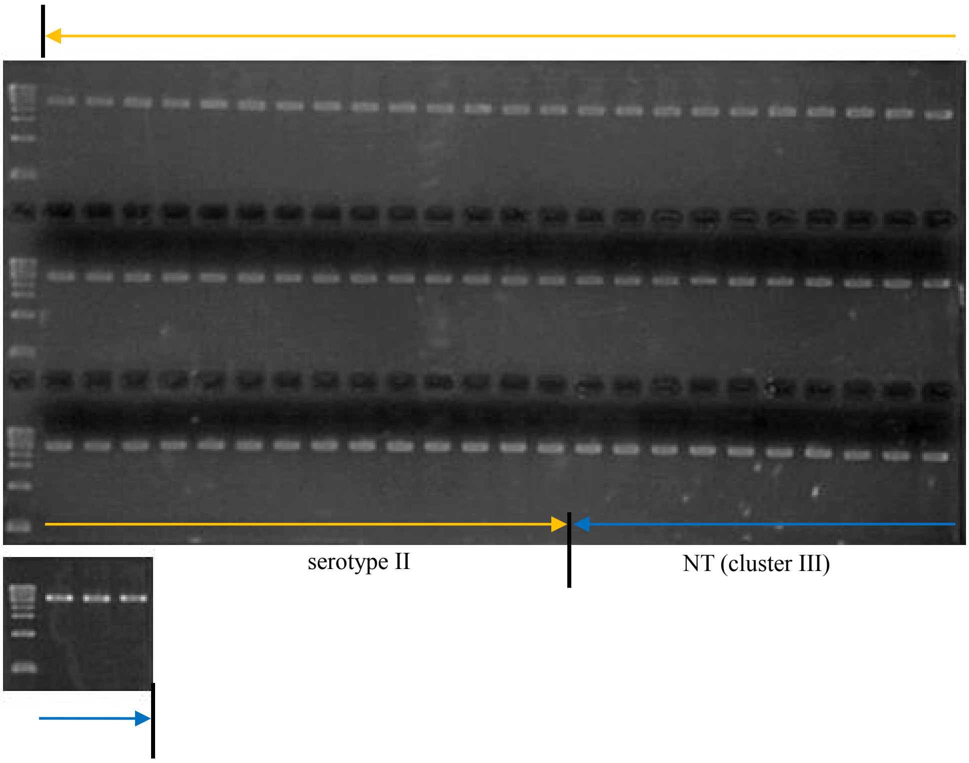


cps1bO –cpsQ 2,700 (bp)

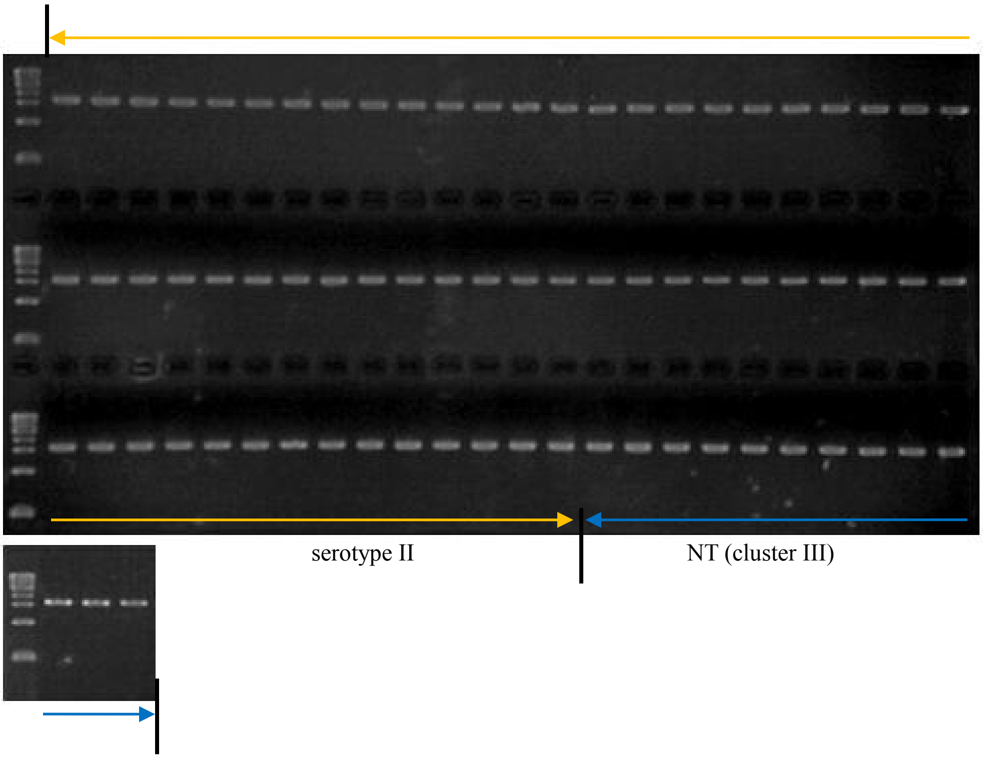


P3. PCR scanning of *cps* locus of serotype II and NT (cluster III). (62+13 strains)

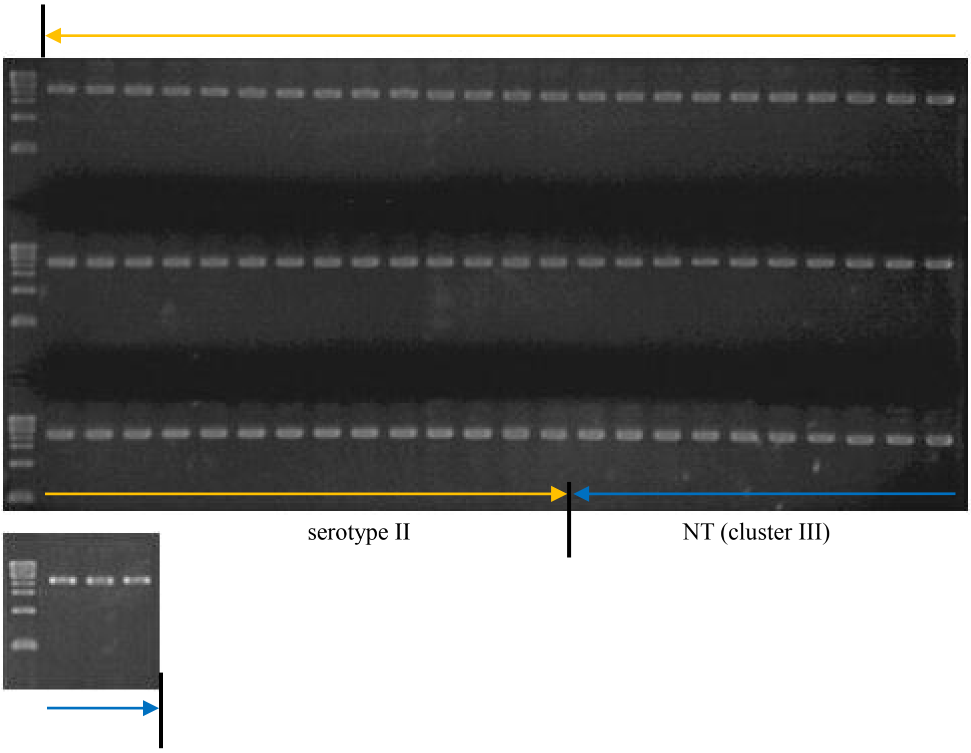
region: *lysR-cpsB* 2,600 (bp)



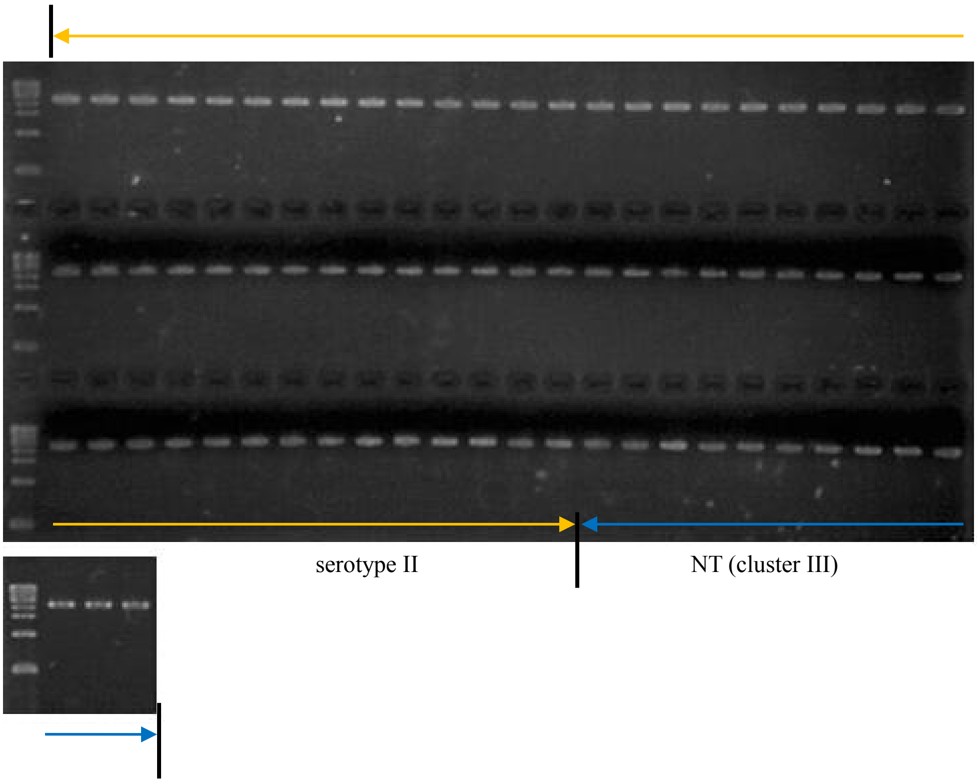
cpsA-cpsD 1,600 (bp)



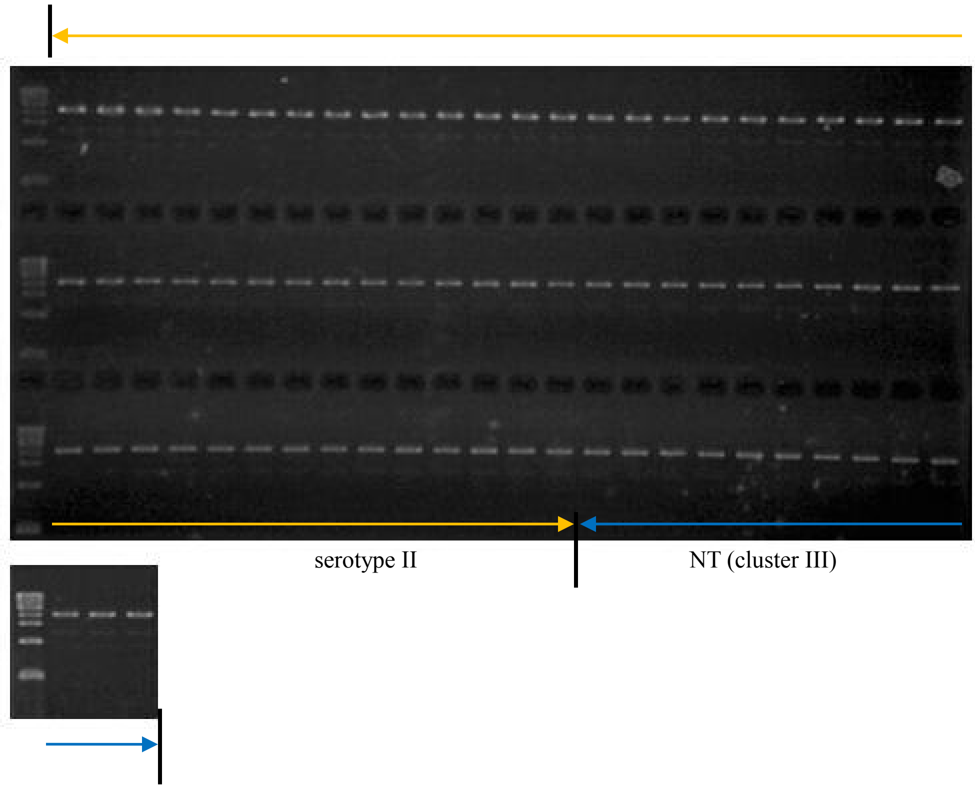
cpsC-cpsE 2,300 (bp)



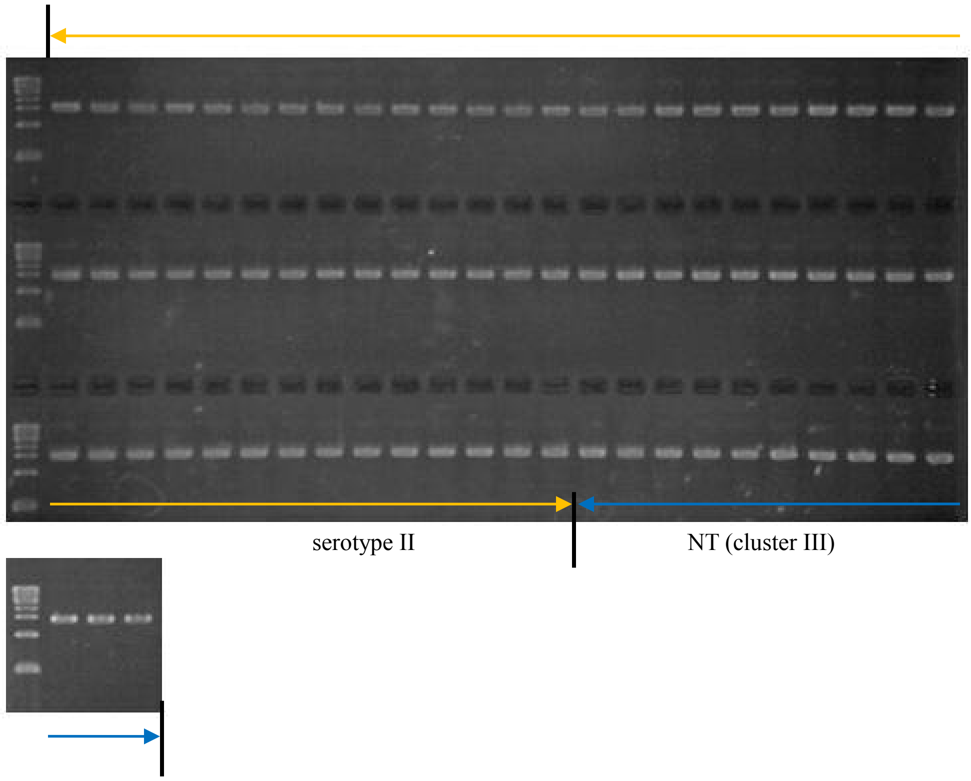
cpsE-cps2F 2,200 (bp)



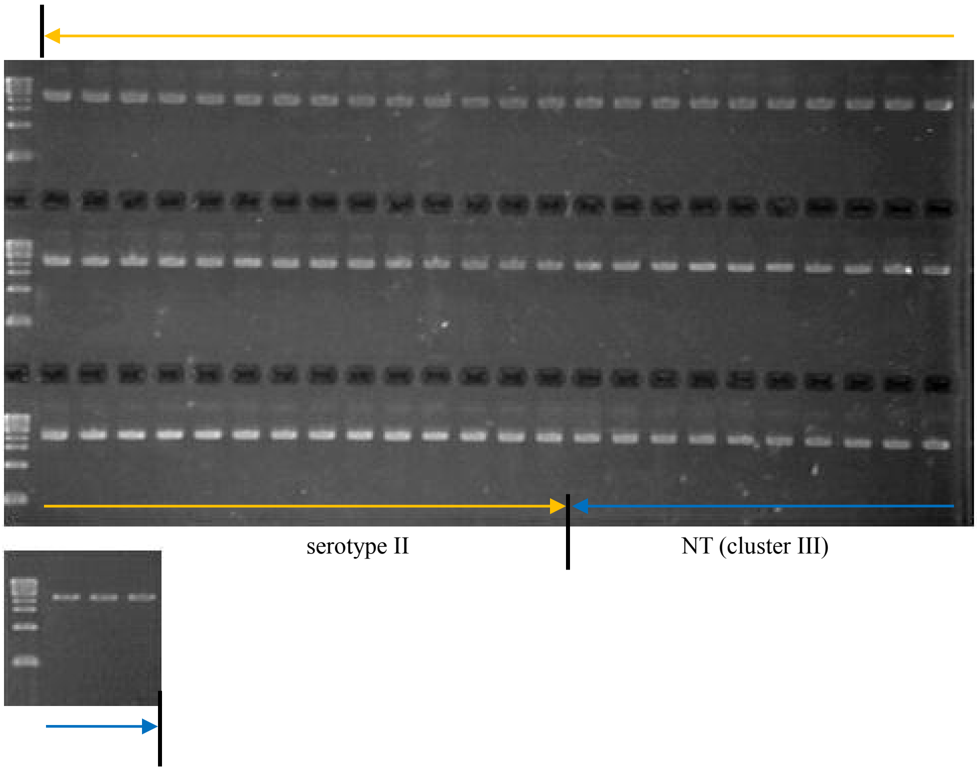
cpsE-cps2G 1,900 (bp)



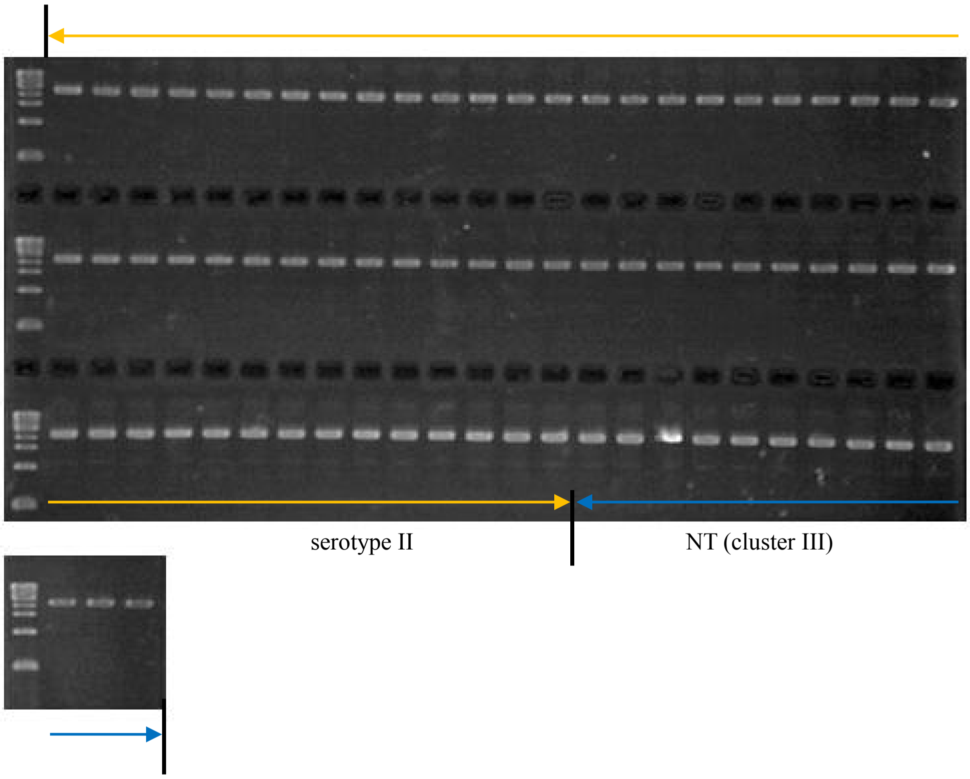
cps2G-cps2I 1,500 (bp)



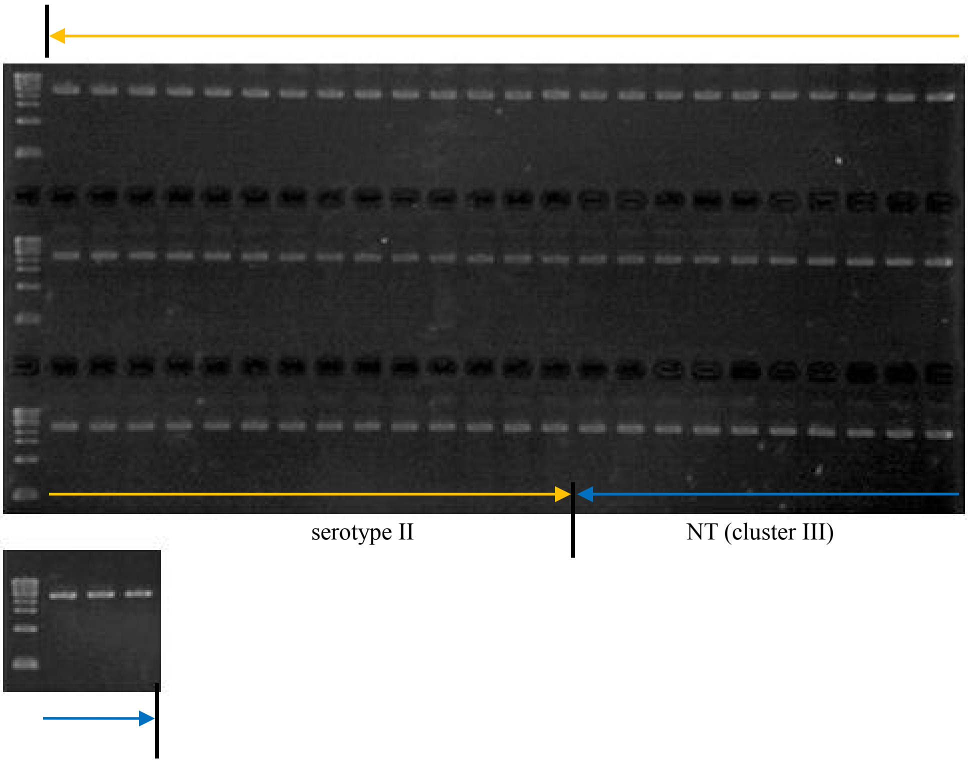
cps2H-cps2J 2,100 (bp)



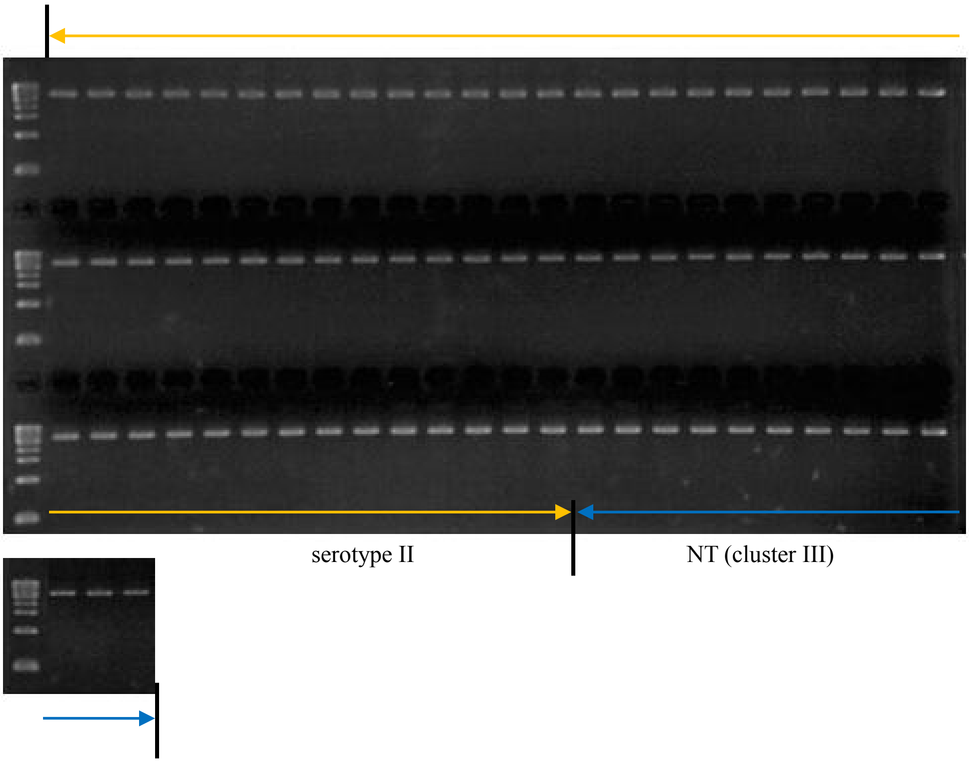
cps2J-cps2L 2,200 (bp)



cps2L-cps2N 2,300 (bp)



cps2N-cps2P 2,900 (bp)



cps2P-cpsR 2,000 (bp)