Genome variations in commensal and pathogenic *E.coli*

INAUGURAL-DISSERTATION

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Girish Neelakanta

aus Bangalore, Indien

2005

Referees/Berichterstatter : Prof. Dr. Karin Schnetz
Prof. Dr. Diethard Tautz

Date of oral examination : 01.02.2005
Tag der mündlichen Prüfung

The present research work was carried out under the supervision and the direction of Prof. Dr. Karin Schnetz in the Institute for Genetics, University of Cologne, Cologne, Germany, from June 2001 to February 2005.

Diese Arbeit wurde von Juni 2001 bis February 2005 am Institut fur Genetik der Universität zu Köln unter der Leitung und der Betreuung von Prof. Dr. Karin Schnetz durchgeführt.

Amma-Nanna..
(to my parents)

ACKNOWLEDGEMENTS

First and foremost I would like to thank my advisor, Prof. Dr. Karin Schnetz for providing her guidance and support during my graduate studies. Her constructive help during the course of my study is acknowledged.

I acknowledge my special thanks to Prof. Diethard Tautz for his constant co-operation and support all through my studies. His immense help and inspirations are not only unforgettable but also praiseworthy. I sincerely thank Dr. Georg Plum for providing the *E.coli* isolates and also for his constant help and encouragement. I owe special thanks to Prof. Angelika Noegel for her help and encouragements during my graduate studies.

I owe a special gratitude to Inge Götz-Krichi, who made all the necessary official things easy and fast. Her encouragement and friendly support is admirable and unforgettable. My special thanks are due to Eva for her necessary help and support in the official formalities.

I thank all the past and present lab colleagues for providing a friendly atmosphere and a special one to Yvonne and Sandra. I record my special thanks to the Professors and members of the "Graduiertenkolleg" for the friendly scientific discussions during the graduate programme meetings.

My everloving parents Amma-Nanna were the source of inspiration and motivation throughout my life. I am deeply indebted for their love and affection which stood by me as a strong support and without their blessings it would have been a difficult task to complete this work.

I owe a huge indebtedness to my everloving brother for his constant encouragement and incredible love which bolstered my days to reach my goals. I thank my friends Satish and Madhu for their inspirational support.

I am greatly indebted to my affectionate wife for all her support throughout this study. Without her, everything would have been impossible and meaningless. Also the support and encouragements from her family is highly acknowledged.

Finally, the financial assistance received from the Graduiertenkolleg "Genetik zellulärer Systeme" University of Cologne, Germany, in the form of Stipend is highly recognized. Also the financial assistance provided from the DFG is acknowledged.

Cologne
10/12/2004 Girish Neelakanta

Contents

Abbreviations

bp	base pairs
BTB	bromothymol blue
bgl/Z	*bgl*/Z5211-Z5214 genomic island
c1955-c1960	c1955-c1960 genomic island
cAMP	cyclic adenosine monophosphate
cpm	counts per minute
CRP (CAP)	catabolite regulator protein
dNTP	deoxyribonucleotide triphosphate
DMSO	dimethylsulphoxide
DNA	deoxyribonucleic acid
EDTA	ethylenediaminetetraacetic acid
H-NS	histone-like nucleotide-structuring protein
IPTG	isopropyl-β-D-thiogalactopyranoside
kb	kilo base pairs
kDa	kilo Dalton
OD	optical density
ONPG	o-nitrophenyl-β, D-galactopyranoside
ORF	open reading frame
PCR	polymerase chain reaction
PNPG	p- nitrophenyl-β, D-galactopyranoside
rpm	rotations per minute
U	unit
v/v	volume by volume
wt	wild type
w/v	weight by volume
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside

Zusammenfassung

Die vergleichende Analyse des Genoms der vier *E. coli* Stämme MG1655, CFT073, O157-EDL933 und Sakai hat einen detaillierten Einblick in das Verständnis von Expansion und Verkleinerung von *E. coli* Genomen geliefert. In der vorliegenden Arbeit wurden die DNA Polymorphismen analysiert die in den Gen Regionen von *bgl*/Z5211-Z5214, c1955-c1960 und *lac* vorkommen und zwar in 25 septischen, 32 uropathogenen, 1 asymptomatischen Bakteriuria, 81 Mensch kommensalen und 32 Tier kommensalen *E. coli* Stämmen, im Vergleich zu den vier sequenzierten Genomen.

Auf der Basis der Ergebnisse an *bgl*/Z5211-Z5214 konnten die typisierten *E.coli* Stämme in fünf Haupttypen und einen Subtypen gruppiert werden: MG1655 Typ, CFT073 Typ, O157 Typ, vierter Typ, fünfter Typ und gemixter Typ. Ungefähr 20 % der Stämme haben eine *bgl* Region ähnlich zu MG1655, 26% ähnlich zu CFT073, 20% haben eine Z5211-Z5214 ähnlich zu O157, 20% haben eine upstream Sequenzen ähnlich zu O157, gefolgt von einer *bgl* und downstream-artigen Region ähnlich zu MG1655 (mit Mischling *yieI* Gens). 11% der Stämme, mit der Ausnahme eines *yieI* Gens, haben MG1655 Sequenzen in der upstream Region, *bgl* und auch downstream Region. Mix Typ Stämme haben eine Mixtur von MG155, CFT073 und O157 in der *bgl*/Z5211-Z5214 Region.

Weiterhin wurden drei unterschiedliche ß-Glukosid Nutzungstypen gefunden. 35% der Stämme papillieren wie MG1655, 16% der Stämme öfter als MG1655 und 15% zeigen schwache Bgl^+ (*relaxed*) Phänotypen. Alle Stämme mit dem *relaxed* Phänotyp zeigten ein CFT073 artiges *bgl* operon, was andeutet daß die CFT073 *bgl* Sequenz eine Vorraussetzung für den schwachen Phänotyp ist und nicht umgekehrt. Mutationen in Genen die für den generellen zellulären Metabolismus benötigt werden, wie etwa Aminosäure Synthese oder Nukleotid Biosynthese, führten zu einer Veränderung des *relaxed* Phänotyps. Die Ergebnisse zeigen auch, daß Sequenz Variation in der *bgl* Promotor Region in den CFT073 *bgl*/Z Typ Stämmen keinen signifikanten Einfluß auf die *bgl* Expression im *E.coli* K-12 Hintergrund hat.

Es konnte auch ein zusätzliches ß-Glukosid System identifiziert werden. Dieses System entspricht der c1955-c1960 Region des CFT073 Chromosoms. Die Analyse der c1955-c1960 Region zeigte, daß 97 von 171 Stämmen das c1955-c1960 System besitzen. In den Stämmen mit dem CFT073 *bgl* Typ kam das c1955-c1960 System vorzugsweise vor. Es konnte gezeigt werden, daß das c1955-c1960 System Gene für die ß-Glukosid Nutzung kodiert. Es hat einen CAP abhängigen Promotor und ist durch Glukose katabolisch reprimiert.

Um zu untersuchen, ob der *bgl*/Z5211-Z5214 Lokus Korrelationen mit anderen Zucker Nutzungs Systemen zeigt, wurden Lactose Nutzungs Phänotypen analysiert. Neun der 171 Stämme zeigte einen Lac^- Phänotyp, wobei sechs dem O157 Typ (*bgl*/Z5211-Z5214) entsprechen. Diese Untersuchungen zeigen die genetische Diversität von *E.coli* Stämmen. Die Ergebnisse ergeben einen Einblick in die Frage, ob sich die *bgl*/Z5211-Z52124 Region als Marker für eine neue Typisierungsstrategie von *E.coli* Isolaten eignet.

Summary

Comparative genomics of the four *E.coli* strains MG1655, CFT073 and O157-EDL933 and Sakai has provided a wealth of information in understanding the continual expansion and retraction of *E.coli* genomes in detail. In this study, a systematic analysis was performed to assess the DNA polymorphisms at the region of *bgl*/Z5211-Z5214 island encoded systems, c1955-c1960 island encoded system and *lac* region in 25 septicemic, 32 uropathogenic, 1 asysmptomatic bacteriuria, 81 human commensals and 32 animal commensal *E.coli* strains and were compared to that seen in the four sequenced *E.coli* strains.

Based on the observations *E.coli* strains were typed at the *bgl*/Z5211-Z5214 locus into five main types and one sub type: MG1655 type, CFT073 type, O157 type, fourth type, fifth type and mixed type. Approximately, 20% of the strains have *bgl* region like MG1655, 26% have *bgl* region like CFT073, 20% have Z5211-Z5214 region like O157, 20% have upstream sequence like O157 followed by *bgl* and downstream like MG1655 (with hybrid *yieI* gene) and 11% of the strains with the exception of downstream *yieI* gene have MG1655 sequence in the upstream, *bgl* and in the downstream region. Mixed type strains have mixture of sequences from MG1655, CFT073 and O157 in the *bgl*/Z5211-Z5214 region.

In addition, three different types of β-glucoside utilization phenotypes were seen. 35% of the strains papillate like MG1655, 16% of the strains papillate more frequently than MG1655 and 15% showed weak Bgl^+ (relaxed) phenotype. All the strains that showed relaxed phenotype carried CFT073 type *bgl* operon, indicating CFT073 *bgl* sequence is important for relaxed phenotype and not *vice versa.* Mutations in the genes that are necessary for general cellular metabolism like amino acid biosynthesis and nucleotide biosynthesis abolished the relaxed phenotype. The analysis also demonstrated that the sequence variations seen at the *bgl* promoter region in the CFT073 *bgl*/Z type strains does not significantly influence the *bgl* expression in *E.coli* K-12 background.

Furthermore, an additional β-glucoside system was identified. This system corresponds to c1955-c1960 region of the CFT073 chromosome. The analysis of c1955-c1960 region revealed that 97 out of 171 strains carried c1955-c1960 system. The presence of c1955-c1960 was observed to be predominant in the strains that carry CFT073 type *bgl*. In a second line of investigation, it was demonstrated that c1955-c1960 system encodes genes for β-glucoside utilization. It carries a CAP dependent promoter and is catabolically repressed in the presence of glucose.

In order to analyze whether the typing at the *bgl*/Z5211-Z5214 locus has any correlations with the other sugar utilizing systems, the lactose utilization phenotypes were determined. Nine out of 171 strains showed Lac^- phenotype, in which six of them belong to O157 type (at *bgl*/Z5211-Z5214). Taken together, the analysis demonstrates the genetic diversity among the *E.coli* strains. Moreover, it may provide an insight in considering *bgl*/Z5211-Z5214 island region as a marker for devising a novel strain-typing method for *E.coli* isolates.

II. Introduction

Strains of the *Escherichia coli* species display a wide range of genome variations with some strains differing by more than 1 Mb (Bergthorsson and Ochman, 1998). These differences could be attributed to: i) the acquisition of foreign DNA, including integration of genetic entities like pathogenicity islands and genomic islands; ii) deletions and duplications of the existing genes; and iii) accumulation of repetitive DNA, such as insertion sequences and transposons into the chromosome. Horizontal gene transfer of DNA is acknowledged to be a key player for the generation of genome variations in *E.coli* (Lawrence and Roth, 1996; Ochman et al., 2000). The availability of the genome sequences of four *E.coli* strains has provided a wealth of information in understanding the genetic diversity among the natural isolates (Blattner et al., 1997; Welch et al., 2002; Perna et al., 2001; Hayashi et al., 2001). However, the mechanisms involved in the genetic differences between the strains of *E.coli* are not completely understood. Analyzing variations in carbohydrate fermentation systems may provide a useful starting point for investigation of genome variations. Hence, in this study a systematic approach was carried out to analyze the genetic variations in 171 commensal and pathogenic strains at three loci that comprise of β-glucoside operon (*bgl*) or Z5211-Z5214 genomic islands, c1955-c1960 genomic island and *lac* region.

1. Pathogenicity islands, genomic islands and bacterial evolution

Over the past decade, considerable insight has been gained into the role of accessory genetic elements to the process of bacterial evolution. A type of genetic element called as ‘pathogenicity island’ (PAIs) has been shown to contribute to the molecular evolution of bacterial pathogens (Falkow, 1996). PAIs are chromosomal clusters of pathogen-specific virulence genes that are characteristically found in pathogenic strains and also rarely in nonpathogenic variants. Some of the features of pathogenicity islands include GC content that differs from the rest of the bacterial genomes, insertion at the 3’ end of the tRNA genes, the presence of flanking direct repeats and insertion elements, ability to encode mobility genes such as integrases, transposases and origins of plasmid replication (Hacker et al., 1997; Blum et al., 1994).

Pathogenicity islands belong to the group of ‘Genomic islands’ (Dobrindt et al., 2004). Genomic islands are called as symbiosis, fitness, metabolic or resistance islands depending on the functions they encode and the advantages they confer relative to the specific lifestyle of a bacterium (Hentschel and Hacker, 2001; Hacker and Carniel, 2001). It may also carry genes of unknown function. An evolutionary advantage of genomic islands is that large number of genes -

for example entire operons that confer new traits - can be horizontally transferred into the genome of the recipient that allows the recipient for more successful adaptation and increased fitness in a specific ecological niche (Fig. 1). Thus, the occurrence and genetic organization of genomic islands reflect the importance of gene acquisition and genome reduction events for the evolution of bacterial pathogens, as well as of non-pathogenic microorganisms. Recent studies have elucidated that genomic islands and other horizontally transferable structures are more commonly found in those bacteria that are present in niches colonized by diverse bacterial species rather than in isolated or sparsely populated environments (Dobrindt et al., 2004).

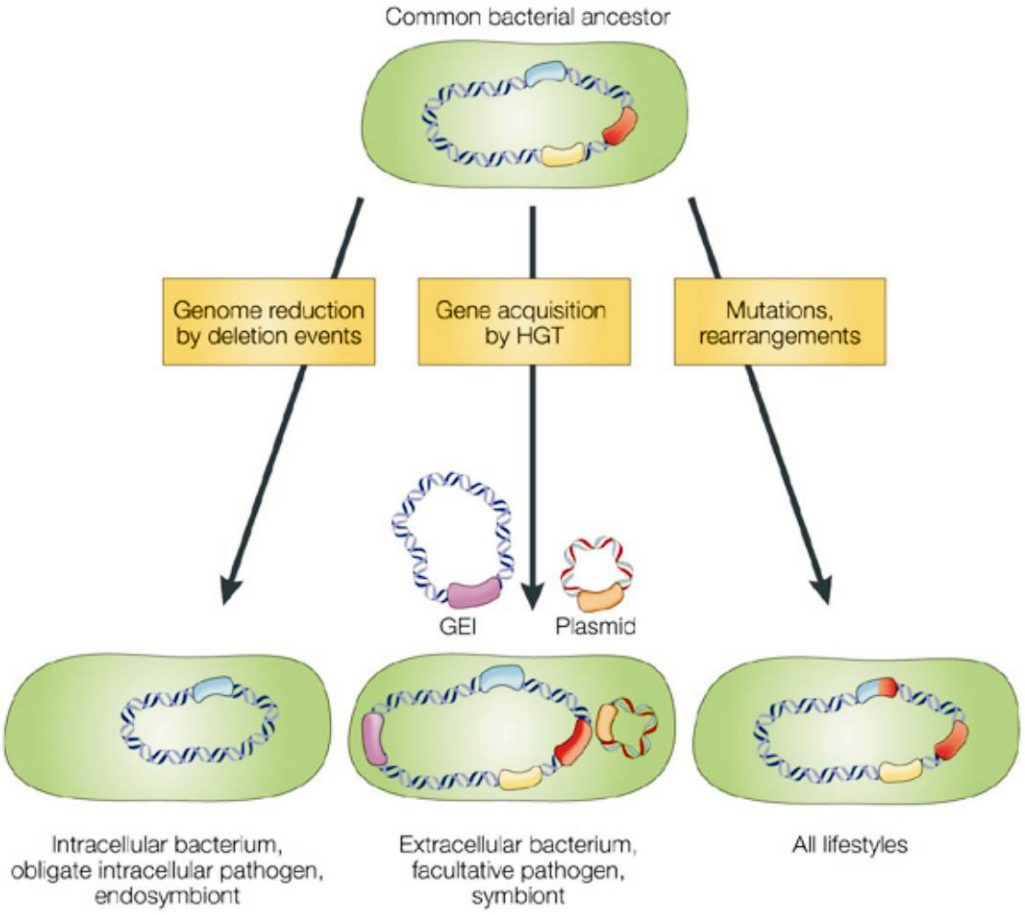


Figure 1: Evolution of bacterial variants by acquisition and loss of genetic information. Genome structure reflects the bacterial life style. Genome reduction is common in intracellular bacteria such as obligate intracellular pathogens and endosymbionts, and contributes to the evolution of strictly host-dependent bacterial variants-as bacteria rely on the host cell to compensate for the gene functions that are lost. Gene acquisition by horizontal transfer between different species, involves mobile genetic elements, such as plasmids, genomic islands (GEIs) and bacteriophages (not shown), and increases the versatility and adaptability of the recipient. This is common in extracellular bacteria, such as facultative pathogens and symbionts, and the acquisition of genes in this way allows bacteria to adapt to a new or changing environment. In addition to these processes, point mutations and genetic rearrangements constantly contribute to evolution of new gene variants in all types of bacteria. HGT, horizontal gene transfer. Figure is taken from Dobrindt et al., 2004.

2. *E.coli*, a model to study bacterial genome evolution

The species *Escherichia coli* represents an exceptional model to study the contribution of genomic islands to the evolution of bacterial genomes in detail (Lawrence and Ochman, 1998). *E.coli* is remarkably a diverse species as numerous ecotypes that live as harmless commensals in intestines of humans and animals exist (Ochman and Selander, 1984). In addition, other distinct genotypes including the enteropathogenic, enterohemorrhagic, enteroinvasive, enterotoxigenic and enteroaggregative *E.coli* cause significant morbidity and mortality as human intestinal pathogens. Extra-intestinal *E.coli* is another varied group of life-threatening pathogens that include distinct clonal groups responsible for neonatal meningitis/sepsis and urinary tract infections. Todate, genomes of four *E.coli* strains: the laboratory strain *E.coli* K-12 MG1655 (Blattner et al., 1997) with a genome size of 4.6 Mb, uropathogenic strain CFT073 (Welch et al., 2002) with genome size of 5.2 Mb and two variants of enterohemorrhagic strain O157-EDL933 (Perna et al., 2001) and Sakai (Hayashi et al., 2001) with genome sizes of 5.4 Mb have been sequenced.

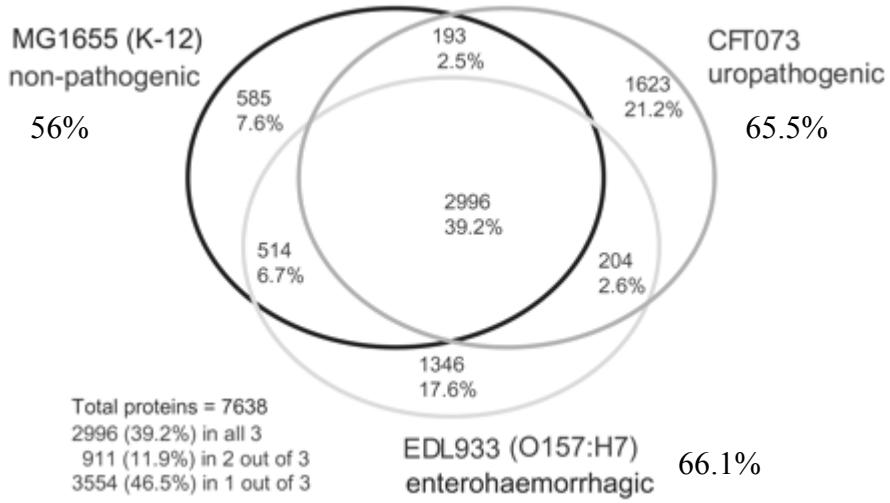


Figure 2: Shared *E.coli* proteins. Comparison of the predicted proteins of the three *E.coli* strains shows the number of orthologs in each shared category and numbers of strain-specific proteins. Hypervariable proteins and proteins spanning island-backbone junctions were excluded from the analysis. Number of proteins counted: K-12, 4,288; CFT073, 5,016; EDL933, 5, 063. In the totals for the three strains, orthologous proteins are counted only once. Orthologous proteins meet the same match criteria used for designation of backbone. Total numbers of proteins analyzed in each strain are indicated: MG1655 (56%), CFT073 (65.5%) and EDL933 (66.1%). Figure modified from Welch et al., 2002.

Comparative genomics of the four sequenced *E.coli* genomes has revealed that approximately 60 to 70% of the genome is composed of a conserved 'core genome' which contains the genetic information that is required for essential cellular functions. The remaining 30 to 40% is composed of 'flexible' gene pool, which encodes additional traits consisting of pathogenicity

islands and genomic islands that can be beneficial under certain circumstances (Welch et al., 2002; Fig. 2). Moreover, the genome size variations (up to 1Mb) seen in the natural *E.coli* isolates (Bergthorsson and Ochman, 1998) has revealed an extensive evidence for the horizontal transfer events in the genetic variability among *E.coli* isolates. Thus, *E.coli* serves as an excellent model to understand the genetic basis for pathogenicity and the evolutionary diversity and its impact on bacterial evolution.

3. Phylogeny and strain typing of *E.coli*

A set of 72 reference strains of *Escherichia coli* (ECOR strains) isolated from a variety of hosts and geographical locations has been established for use in studies of variation and genetic structure in natural populations. These strains are representative of the range of genotypic variation in the species as a whole (Ochman and Selander, 1984). The phylogenetic relationships among the 72 ECOR strains was studied by the neighbor-joining (NJ) method applied to a genetic distance matrix based on electrophoretically detected allelic variation at 38 enzyme-coding loci (for e.g. Alchohol dehydrogenase, Malate dehydrogenase etc.,). The principle underlying the analysis is that any allelic difference in electrophoretic mobility results from at least one codon difference at the nucleotide level. Under the assumption that codon changes occur independently, standard genetic distance is an estimate of the mean number of net codon differences per genetic locus (Selander et al., 1986, Herzer et al., 1990 and references therein). Thus, based on the genetic distance matrix ECOR strains have been phylogenetically classified into 5 groups: A, B1, B2, D and E (Herzer et al., 1990). Group A is predominant with K-12 and K-12 like strains (25 strains) isolated from humans. Group B1 is predominant with strains isolated from non primate mammals (16 strains). Group B2 is predominant with strains isolated from humans and other primates (15 strains). Group D is a heterogenous group with mixture of strains from humans, non primates and other primates (12 strains). The fifth group E is a variant from the other four types that consists of strains from humans and non primate mammals (4 strains).

In the past decade or so, several studies have used genome comparison techniques like macro-restriction analysis, PFGE (Pulsed Field Gel Electrophoresis), genomic subtraction, RFLP (Restriction fragment length polymorphism), analysis of DNA sequences of housekeeping genes and DNA microarray analysis to assess the genetic variability and phylogenetic relatedness among *E.coli* isolates (Bonacorsi et al., 2000; Melkerson-Watson et al., 2000; Rode et al., 1999; Lecointre et al., 1998; Milkman and Bridges, 1993; Pupo et al., 1997; Akman and Aksoy, 2001;

Dobrindt et al., 2003; Fukiya et al., 2004). Using RFLP and nucleotide sequencing, Herbelin and coworkers (2000) have assessed the DNA polymorphism in the *mutS-rpoS* region and inferred an evolutionary history of divergence of this region among the isolates of *E.coli*. The study was focused on a collection of *E.coli* strains comprising of enteropathogenic, enterohemorrhagic and ECOR strains. Their results showed that the length of the genomic sequence between the *mutS* and *rpoS* genes is variable in the natural *E.coli* isolates (Herbelin et al., 2000). Based on the variations seen, the authors have proposed an evolutionary model that categorizes *E.coli* strains into four main types and one sub type. The grouping from their analysis correlated with the phylogenetic classification of ECOR strains.

4. Impact of genome variations on the utilization of carbon source in *E.coli*

Carbohydrates are excellent carbon sources for all bacteria. In *E.coli*, a vast amount of information is available on the components of the pathways necessary for the utilization of various carbohydrates. The pathways of carbohydrate utilization follow the same theme in all the natural *E.*coli isolates. However, the impact of genetic variability within the genome may have an influence on the metabolic properties of the carbohydrate utilization systems. Horizontal gene transfer even at very low levels produces a mosaic chromosome. As a result, species-specific traits such as those encoded by horizontally transferred genes (e.g., lactose utilization, indole production) attributes to the phenotypic characterization of *E.coli* (Lawrence and Ochman, 1998). Analyzing the genetic variations in the carbohydrate systems may provide an insight in understanding the role of horizontal gene transfer in shaping the ecological and pathogenic character of *E.coli.*

Three of the loci that encode genes required for the utilization of carbohydrates that show variations in the four published *E.coli* sequences are ; i) *bgl*/Z5211-Z5214 island encoded region at ~84 min on the MG1655 chromosome (Blattner et al., 1997) includes β-glucoside utilization, *bgl* (in MG1655 and CFT073) and Z5211-Z5214 (in O157-EDL933 and Sakai) islands. ii) The region between *marB* and *ydeD* at 34.8 min on the MG1655 chromosome includes c1955-c1960 island encoded system (characterized in this work) that is present in CFT073 and absent in the other three strains. iii) The *lac* operon at 7.8 min on the MG1655 chromosome is conserved in all the four sequenced strains (Fig. 3).

Figure 3: Schematic representation of the three loci analyzed in the current study. Horizontal lines represent the linear chromosomes of MG1655, CFT073 and two variants of O157-EDL933 and Sakai. The two variants of O157 are identical to one another at the three loci. Positions of the three loci are relative to the MG1655 chromosome and are indicated as 7.8 min (7.8'), 34.8 min (34.8') and 84 min (84'). The β-glucoside (*bgl*) genomic island is present in MG1655 and CFT073. O157 carries Z5211-Z5214 genomic island in place of *bgl*. The c1955-c1960 genomic island is present in CFT073 and absent in the other three strains. The *lac* operon (*lac*) is conserved in all the four sequenced *E.coli* strains.

5. The *bgl*/*Z5211-Z5214* locus in *E.coli*

The *bgl* operon in MG1655 and CFT073 contains six genes (*bglG*, F, B, H, I and K) (Schnetz et al., 1987; Welch et al., 2002) (Fig. 4) in which the first three genes are necessary and sufficient for the utilization of aryl β-glucosides like arbutin and salicin (Prasad and Schaefler, 1974; Schnetz et al., 1987; Mahadevan et al., 1987). The gene products of the *bgl* operon are the positive regulator and antiterminator BglG, the β-glucoside specific permease EII^{Bgl} (or BglF), the phospho-β-D-glucosidase BglB, Porin like protein BglH, endo-1-4-xylanase homology protein BglI and glucosamine-6-phosphate-isomerase homology protein BglK (Fig. 4). In contrast to MG1655 and CFT073, strains O157-EDL933 and Sakai have Z5211-Z5214 region in place of *bgl*. The Z5211-Z5214 region contains four ORF's that encode proteins of unknown functions (Perna et al., 2001; Hayashi et al., 2001) (Fig. 4).

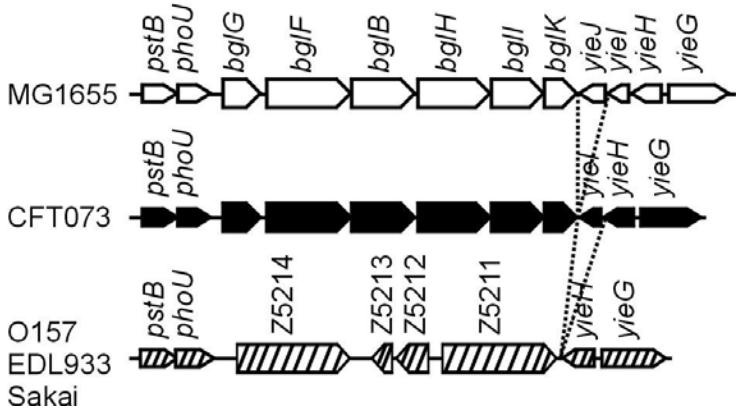


Figure 4: Structure of the *bgl*/*Z5211-Z5214* locus in the four sequenced *E.coli* strains. The genes in MG1655 *bgl* region are presented by open black arrows (Blattner et al., 1997), genes of the CFT073 *bgl* region are presented in black (Welch et al., 2002), and genes of the O157-EDL933 (Perna et al., 2001) and Sakai (Hayashi et al., 2001) Z5211-Z5214 region are presented by hatched black arrows. CFT073 lacks the *yieJ* gene at the 3' end of the *bgl* operon (indicated by dotted lines). Strains O157-EDL933 and Sakai are identical to one another at this locus, and lack the *yieJ* and *yieI* genes. Genes present in upstream and downstream of the *bgl*/Z5214-Z5211 locus are shown with different arrows (open, black, hatched) indicating sequence variations between the strains.

The alignment of *bgl*/Z5211-Z5214 region including the upstream and downstream regions in the four sequenced *E.coli* strains shows that CFT073 lacks the *yieJ* gene. Strains O157-EDL933 and Sakai lacks the *yieJ* and *yieI* genes and are identical to one another at this locus (Fig. 4). Furthermore, the nucleotide sequence alignment including the upstream and downstream regions in the four sequenced *E.coli* strains show that sequence variations are seen in and around the *bgl*/Z5211-Z5214 region in the four sequenced strains. To this end, the nucleotide sequence alignment of the *bgl*/Z5211-Z5214 island encoded systems with its flanking regions distinguishes the four sequenced *E.coli* strains.

6. Crypticity of the *bgl* operon

An interesting feature of the *E.coli bgl* operon is its crypticity (silent) (Schaefler and Maas, 1967; Reynolds et al., 1981) i.e. it is neither expressed nor induced under all laboratory conditions. However, Khan and Isaacson (1998) have reported that the expression of *bgl* operon is seen when septicemic *E.coli* strain i484 infects mouse liver. Why the operon is silent in the laboratory conditions and what may cause it to be expressed in the host, remains to be an open question.

The silencing of the *bgl* operon is determined at the *bgl* promoter and within the region of *bglG* gene. The abundant nucleoid-associated protein H-NS, that affects the expression of many genes (Ussery et al., 1994), is essential for silencing of the *bgl* operon (Dole et al., 2002; Dole et al., 2004; Defez and de Felice, 1981) (Fig. 5). It represses the CRP/cAMP dependent *bgl* promoter (Schnetz, 1995, Schnetz and Wang, 1996; Mukerji and Mahadevan, 1997; Caramel and Schnetz, 1998) as well as region downstream to the promoter, where it causes a strong polarity of the *bglG* gene leading to low expression of *bglG* and further downstream genes (Dole et al., 2002, Dole et al., 2004; Fig. 5). In addition to H-NS, Fis a pleiotropic DNA bending protein, RpoS a stationary phase sigma factor, Crl, transcriptional-regulator-like proteins LeuO and BglJ, the protease Lon, RNA binding protein Hfq and H-NS homologue StpA are also necessary for silencing of the *bgl* operon (Finkel and Johnson, 1992; Caramel and Schnetz, 2000; Schnetz, 2002; Tsui et al., 1994; Giel et al., 1996; Free et al., 1998; Ueguchi et al., 1998; Ohta et al., 1999; Dole et al., 2004). Spontaneous mutations that activate the *bgl* operon map close to the CRP-dependent promoter that include deletion of an AT-rich silencer upstream of the promoter, integration of insertion elements, and point mutations that improve the CRP-binding site (Reynolds et al., 1986; Reynolds et al., 1981; Schnetz and Rak, 1992; Schnetz, 1995; Lopilato and Wright, 1990; Mukerji and Mahadevan, 1997).

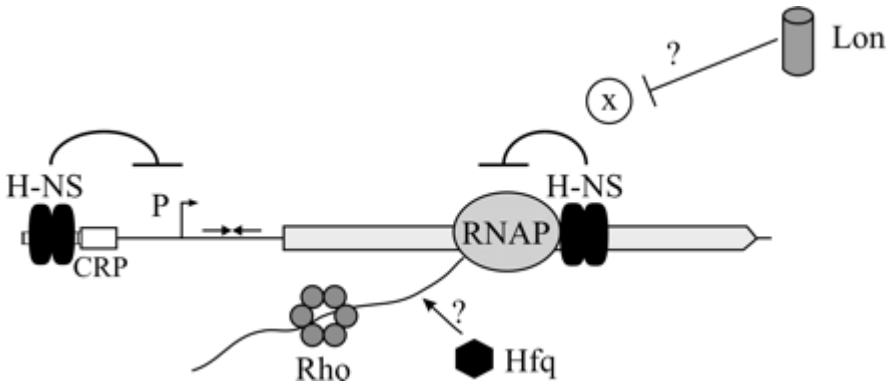


Figure 5: Model of the H-NS-mediated repression of the *bgl* operon at two levels (Dole et al., 2004). H-NS binds upstream of the promoter and represses the transcription initiation. In addition, H-NS binds within the coding region of the first gene, approximately 600 to 700 bp downstream of the transcription initiation site, where it induces a Rho-dependent polarity (Dole et al., 2004). Hfq and Lon reduce the H-NS-induced polarity.

Todate, no information is available at the genetic and molecular level for the organization and expression of *bgl* genes in the naturally occurring *E.coli* isolates. Analyzing the impact of genetic variations at the *bgl*/Z5211-Z5214 locus on the organization and expression of *bgl* genes might give an insight on the reasons for the maintenance of the operon in a cryptic state. Moreover, it may help in elucidating the complex pleiotropic regulation of the *bgl* operon.

7. β-glucoside utilization system in other organisms

The β-glucosides such as Salicin, Cellobiose, Arbutin and Esculin are abundantly found in nature. They can be found in foods containing plant extracts. The general structure of these compounds is a glucose moiety with various groups attached at the C-1 hydroxyl of the glucose core. The β-glucosides are used as a carbon source by many bacteria. In *Shigella* (a close relative of *E.coli*) the genes encoding the proteins for utilization of aryl-β-glucosides are organized as in *E.coli*. However, in *Shigella* the *bglB* gene is inactivated by an insertion element and thus two step mutations are necessary to allow the utilization of salicin (Kharat and Mahadevan, 2000). In *Erwinia chrysanthemi* a plant pathogen, the *arb* genes encoding the gene products required for the utilization of β-glucosides (arbutin) are homologous to *E.coli bgl* operon. However, the *arb* genes are not cryptic as *bgl.* This difference in *Erwinia* could be due to the divergence of the promoter region in comparison to *E.coli* (el Hassouni et al., 1990). The divergence in the utilization of β-glucosides in *Erwinia chrysanthemi* and *E.coli* could also be because of their different natural habitats (Fig. 6). *Klebsiella aerogenes* carry *bgl* genes as in *E.coli*. However, the *bgl* operon of *K. aerogenes* is not cryptic. This is due to the differences at the *bgl* promoter region in the *K.aerogenes* (Raghunand and Mahadevan, 2003). Gram positive bacteria have shown to

contain β-glucoside systems similar to those of the Gram-negative bacteria. In *Bacillus subtilis*, the *bglPH* operon is responsible for the utilization of β-glucosides. Expression of this operon is regulated by LicT a BglG homolog encoded at a separate locus (Kruger and Hecker, 1995; Schnetz et al., 1996).

Bacteria	habitat	ß-glucoside utilization system	Reference
Shigella	water, intestinal pathogen	*bgl* (ß-glucoside) operon	Kharat and Mahadevan, 2000
Erwinia chrysanthemi	soil, plant pathogen	*arb* (arbutin)operon	el Hassouni et al., 1990
Klebsiella aerogenes	intestine, respiratory tract	*bgl* operon	Raghunand and Mahadevan, 2003
Bacillus subtilis	soil	*bglPH* operon	Kruger and Hecker, 1995
Clostridium longisporum	rumen, intestine	*abg* (aryl-ß-glucoside)	Brown and Thompson, 1998
Azospirillum irakense	soil	*salCAB* (salicin) operon	Faure et al., 2001
Thermoanaerobacter brokii	soil, volcanic habitat	*cglT* and *xglS* genes	Breves et al., 1997
Listeria monocytogenes	soil	*bvr* locus	Brehm et al., 1999
Streptococcus gordonii	human teeth	*bgl* regulon, esc locus *bfb* locus, *gom* locus	Kilic et al., 2004
Streptococcus mutans	soil, plant pathogen	*bglPCA* regulon	Cote and Honeyman, 2002
Pectobacterium carotovorum	intestine or pathogen	*bgl* operon	An et al., 2004
E.coli	intestine or pathogen	*bgl* operon (silent)	Schnetz et al., 1987

Figure 6: Overview of β-glucoside utilization systems present in diverse bacteria. Shown is the different β-glucoside systems present in diverse groups of bacteria. The names of the genes or the systems involved in β-glucoside utilization are indicated.

Clostridium longisporum, a ruminal Gram-positive bacterium carries an aryl-β-glucoside uptake and utilization system that is composed of several *abg* (aryl-β-glucoside) genes (Brown and Thomson, 1998). The other systems involved in the utilization of the β-glucosides are the *salCAB* operon in *Azospirillum irakense* (Faure et al., 2001), the *bglPCA* regulon in *Streptococcus mutans* (Cote and Honeyman, 2002), the *cglT* and *xglS* genes in *Thermoanaerobacter brokii* (Breves et al., 1997), *bgl* operon of *Pectobacterium carotovorum* (An et al., 2004) and the *bvr* locus in *Listeria monocytogenes* (Brehm et al., 1999). None of the β-glucoside systems mentioned above are cryptic. Recent report from Kilic and co-workers (2004) have reported that *Streptococcus gordonii* have four separate genetic loci that encodes genes for the utilization of β-glucosides. In addition, the authors have also reported that the genes required for β-glucoside utilization are associated with adhesion, biofilm formation and *in vivo* gene

expression. This data suggest a unique role for β-glucoside utilization systems in the environment inside the host.

8. Aim of the thesis

There is growing evidence that genomic diversity is high among the natural *E.coli isolates*. Genetic entities like genomic islands play a profound role in these processes. An enhanced understanding of the role of genomic islands may provide an insight into how genome dynamics can contribute to bacterial evolution in general. In the present study we took advantage of the readily available genome sequences of MG1655, CFT073 and O157-EDL933 and Sakai and we have systematically analyzed the two island encoded regions: the *bgl*/Z5211-Z5214 locus and c1955-c1960 locus in a repertoire of 171 naturally occurring *E.coli* isolates.

The *E.coli* isolates used in the current study comprises of 99 clinical (25 septicemic, 22 uropathogenic, 52 human commensals) and the 72 strains of the ECOR collection (10 uropathogenic, 1 asysmptomatic bacteriuria, 29 human commensals and 32 animal commensal). A combination of PCR, Southern hybridization and nucleotide sequencing was used to characterize the genetic variations at *bgl*/Z5211-Z5214 locus and c1955-c1960 locus and compared the variations seen to that in the four sequenced *E.coli* strains. In addition, β-glucoside utilization phenotypes of all the strains were analyzed. In order to know whether the *bgl*/Z5211-Z5214 locus analysis has any correlations with the other carbohydrate utilizing systems, all the strains were analyzed for their lactose utilization phenotypes. With the approaches undertaken the current study addresses a method of typing *E.coli* strains at *bgl*/Z5211-Z5214 genomic island.

III. Results

1. Analysis of the *bgl*/Z5211-Z5214 genomic island in naturally occurring *E.coli*

(This section, in part, is in preparation for a publication)

1.1 Variations at the *bgl*/Z5211-Z5214 locus in the four sequenced *E.coli* strains

Comparative genomics of four sequenced *E.coli* strains K-12 MG1655, CFT073 and two variants of O157 (EDL933 and Sakai) shows that the *bgl* and the Z5211-Z5214 locus are alternative genomic islands in *E.coli*. Strains MG1655 and CFT073 carry the *bgl* genomic island while the O157 strain, carries the Z5211-Z5214 genomic island at the same chromosomal map position (Fig. 4, Introduction). Alignment of the *bgl*/Z5211-Z5214 region (including upstream and downstream regions) from the four sequenced strains showed variations between the strains (Fig. 7).

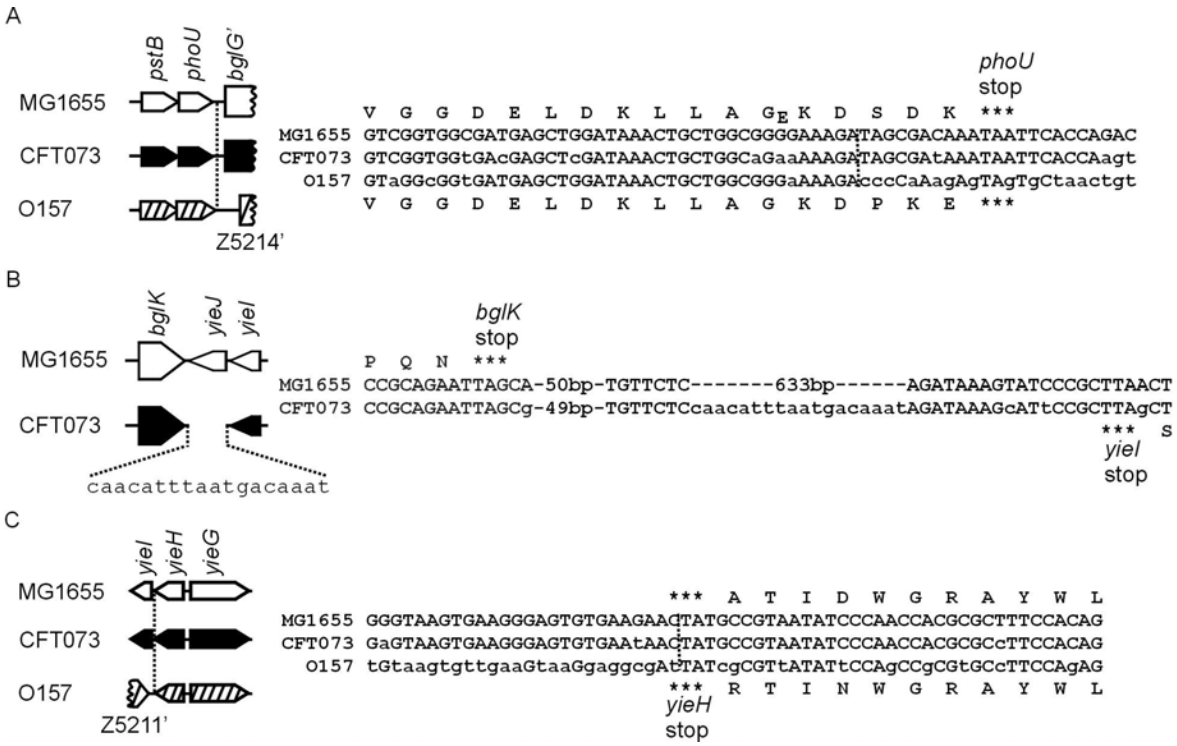


Figure 7: Nucleotide sequence alignment of *bgl*/Z5211-Z5214 regions which are different in the sequenced *E.coli* strains. Structures are represented as described in Figure 4 (Introduction). Nucleotide changes are shown as non-capitals and amino acid variations are shown in bold. A) Comparison of the *phoU* region of MG1655 (AE000449: 8290-8227), CFT073 (AE016769: 175747-175684) and O157-EDL933 (AE005603: 8244-8181). Deduced amino acid sequences of MG1655 and CFT073 are shown at the top and O157 at the bottom. The dotted line indicates the 5' end of Z5211-Z5214 region. In the schematic presentation shown to the left *bglG*' and Z5214' indicate the 5' end of the *bglG* and Z5214 genes respectively. B) Comparison of the *bglK-yieI* region of MG1655 (AE000449: 78-1 + AE000448: 10332-9685) and CFT073 (AE016769: 167538-167429). CFT073 lacks the *yieJ* gene and carries an additional 18 bp sequence in comparison to MG1655. Deduced amino acid sequences of BglK (at the top) and YieI (at the bottom) are also shown. C) Comparison of the *yieH* region of MG1655 (AE000448: 9181-9121), CFT073 (AE016769: 166919-166859) and O157-EDL933 (AE005603: 804-744). Deduced amino acid sequence is shown as in A. Dotted lines indicate the 3' end of Z5211-Z5214 region. Z5211' indicates 3' end of Z5211. O157-EDL933 and Sakai possess identical sequences at the regions shown in A and C.

The upstream *phoU* gene is conserved in all the four sequenced strains. However, nucleotide sequence alignment of the *phoU* region showed variations between the strains (Fig. 7A). The nucleotide sequence alignment of the downstream region of *bgl*/Z5211-Z5214 locus also showed variations between the strains (Fig. 7B and C). Strain CFT073 lacks the *yieJ* gene and carries an additional 18 bp in comparison to MG1655 (Fig. 7B). Both the O157 strains lack *yieJ* and *yieI* genes and are identical to each other at the *bgl* Z5211-Z5214 locus (Fig. 7C).

1.2 Typing of 171 *E.coli* isolates at the region of *bgl*/Z5211-Z5214 genomic islands

In order to analyze whether the variations at the *bgl*/Z5211-Z5214 locus seen in the four sequenced strains are also present in the *E.coli* isolates a total of 171 *E.coli* strains encompassing 99 clinical (Table 6a, Appendix) and the 72 strains of the ECOR collection (Table 6b, Appendix) were analyzed at this locus. Strains were analyzed by PCR with *bgl* or Z5211-Z5214 specific oligos (Fig. 37, materials and methods and Table 4, Appendix), nucleotide sequencing and Southern hybridization using *bgl* specific probes (see materials and methods).

Based on the analysis the strains were typed into five main types and one sub type at *bgl*/Z5211-Z5214 region (Fig. 8). Approximately, 20% of the strains that have MG1655 like nucleotide sequences in the upstream, *bgl*, and downstream regions are grouped as MG1655 type (Fig. 8A to K). 26% of the strains that have CFT073 like sequences in the upstream, *bgl* and downstream regions are grouped as CFT073 type (Fig. 8L to N). 20% of the strains that have Z5211-Z5214 locus are grouped as O157 type (Fig. 8O and P). 20% of the strains that have the upstream sequence like O157 followed by *bgl* and downstream sequences like MG1655 are grouped as fourth type (Fig. 8Q to W). 11% of the strains that have upstream, *bgl* like MG1655 and downstream like MG1655 with 5'end of the *yieI* gene like CFT073 are grouped as fifth type (Fig. 8X to Z). In addition to these five main types, 3% of the strains have mixed sequences of MG1655, CFT073 and O157 in the *bgl*/Z5211-Z5214 region and are grouped as mixed type in this study (Fig. 8AA to AD). The *phoU* gene that is upstream of *bgl*/Z5211-Z5214 is conserved in all the isolates analyzed; however, the downstream region is variable.

Eleven out of 33 MG1655 *bgl/Z* type strains show alterations in the *bgl* region by insertions of IS1, IS2, IS629, IS186 and ISEc8 (insertion associated fragment from *pheV* locus of CFT073) and deletions within the *bgl*/Z5211-Z5214 locus (Fig. 8B to K). Strains U3633, E10096 and U4418 have same insertion site for ISEc8 fragment (Fig. 8B and C). Likewise, strains E291 and E292 have same insertion site for IS629 (Fig. 8H and I) suggesting that the strains could be derivatives of each other. Two strains out of 46 CFT073-type show alterations by insertion of IS1397 and IS1294 associated deletion (Fig. 8M and N respectively).

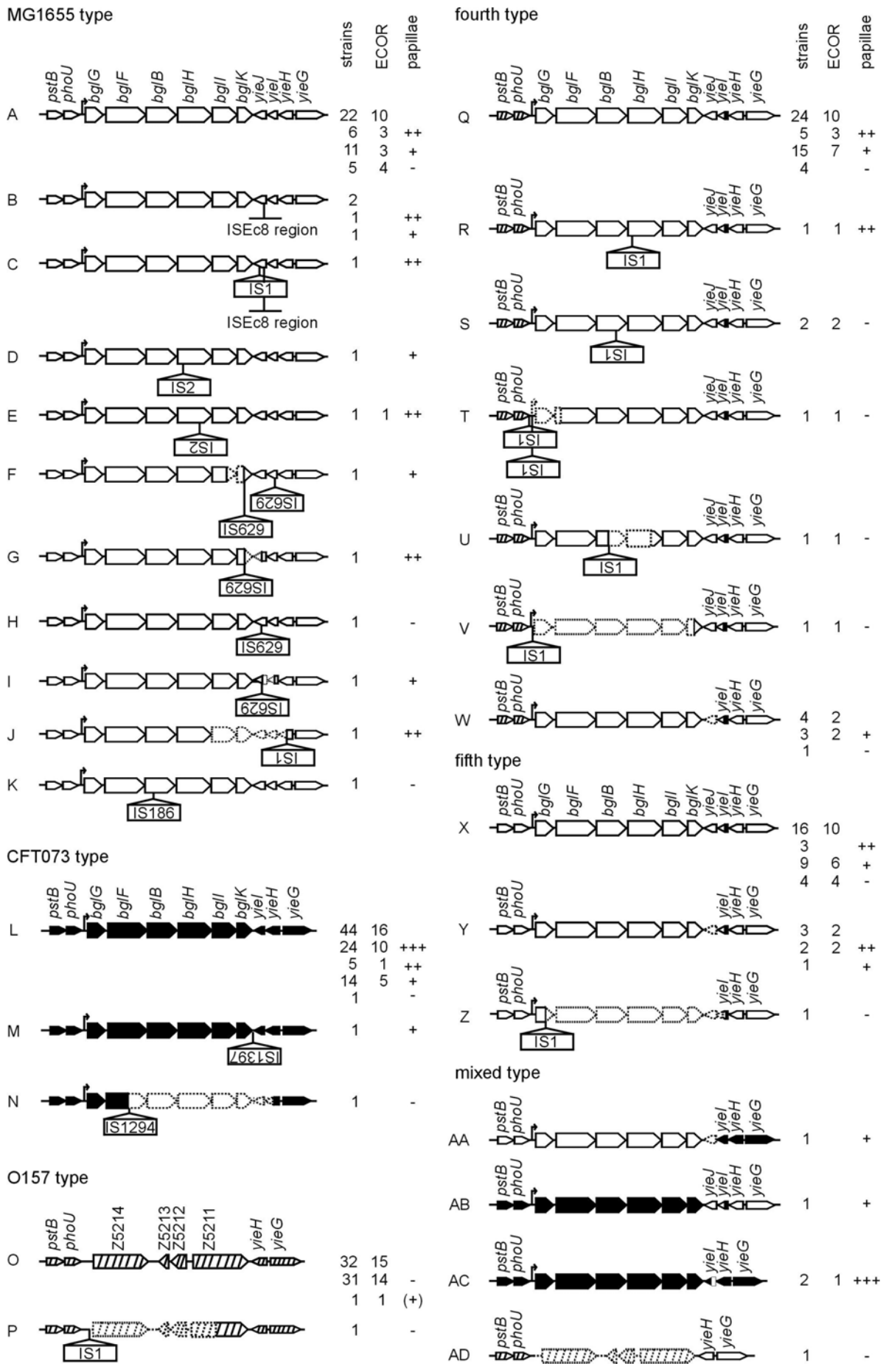

MG1655 type
strains
ECOR
papillae
pstB phoU bglG bglF bglB bglH bglI bglK yieJ yieI yieH yieG
A
22 10
6 3 ++
11 3 +
5 4 -
B
2
1 ++
1 +
ISEc8 region
C
1 ++
IS1
ISEc8 region
D
1 +
IS2
E
1 1 ++
IS2
F
1 +
IS629
IS629
G
1 ++
IS629
H
1 -
IS629
I
1 +
IS629
J
1 ++
IS1
K
1 -
IS186
CFT073 type
pstB phoU bglG bglF bglB bglH bglI bglK yieI yieH yieG
L
44 16
24 10 +++
5 1 ++
14 5 +
1 -
M
1 +
IS1397
N
1 -
IS1294
O157 type
pstB phoU Z5214 Z5213 Z5212 Z5211 yieH yieG
O
32 15
31 14 -
1 1 (+)
P
1 -
IS1
fourth type
pstB phoU bglG bglF bglB bglH bglI bglK yieJ yieI yieH yieG
Q
24 10
5 3 ++
15 7 +
4 -
pstB phoU yieJ yieI yieH yieG
R
1 1 ++
IS1
S
2 2 -
IS1
T
1 1 -
IS1
IS1
U
1 1 -
IS1
V
1 1 -
IS1
pstB phoU yieI yieH yieG
W
4 2
3 2 +
1 -
fifth type
pstB phoU bglG bglF bglB bglH bglI bglK yieJ yieI yieH yieG
X
16 10
3 ++
9 6 +
4 4 -
Y
3 2
2 2 ++
1 +
Z
1 -
IS1
mixed type
AA
1 +
AB
1 +
AC
2 1 +++
AD
1 -

Figure 8: *bgl*/Z5211-Z5214 locus and β-glucoside utilization phenotypes of 171 commensal and pathogenic *E.coli* strains. Strains were analyzed by PCR; sequencing and Southern hybridization using *bgl* specific probes. Structures are represented as shown in Figure 4 (see introduction). Based on the structural analysis strains were grouped into five main types and one sub type. Total number of strains and number of ECOR strains in each type are shown. The β-glucoside utilization phenotypes on BTB salicin plates at 37°C are also shown. +++ indicates weakly Bgl^+ at day 3 incubation (relaxed phenotype), ++ more papillae than MG1655, + papillae like MG1655, (+) weak Bgl^+ at day 2 incubation and – indicates no/late papillae. **MG1655 type 33 strains:** A) 22 strains have intact *bgl* locus like MG1655, in which 6 strains (F287, F1215 (t1 +105ta), E476, ECOR2, ECOR5, ECOR13) show more papillae, 11 strains (F785, E10097, E10099, E10082, E10085, E166, E444, E180, ECOR10, ECOR11, ECOR25) papillate like MG1655 and the remaining 5 strains (E10090, ECOR1, ECOR3, ECOR8, ECOR14) does not papillate. B) 2 strains (U3633, (++)) and (E10096, (+)) have insertion of 2456 bp ISEc8 (insertion associated) fragment (AE016766: 83067 to 85522) with 6 bp target site duplication (*yieJ*: 256-261 (numbering relative to the translational start)). C) Strain U4418 (++) carries an IS1 with a 9 bp target site duplication (*yieJ*: 304-312) and also carries the ISEc8 fragment as in strains shown in B. D) Strain E167 (+) carries an IS2 insertion with 5bp target site duplication (*bglH*: 33-37). E) Strain ECOR12 (++) carries an IS2 insertion with 4bp target site duplication (*bglH*: 934-937). F) Strain E345 (+) carries an IS629 and associated deletion from *bglI*: 917 to *bglK*: 105 (AE000449: 1078-604) and carry a second IS629 in inverted orientation with a 3 bp target site duplication (*yieI*: 235-237). G) Strain U5107 (++) carries an IS629 and associated deletion from *bglK*: 106 to *yieJ*: 573 (AE000449: 603-1 + AE000448: 10332-10311). H) Strain E291 (-) carries an IS629 insertion with a target site duplication of 3 bp (*yieJ*: 547-549). I) Strain E292 (+) carries an IS629 and associated deletion from *yieJ*: 546 to *yieI*: 355 (AE000448: 10284-9577). J) Strain E164 (++) carries an IS1 and associated deletion from *bglI*: 33 to *yieH*: 238 (AE000449: 1962-1 + AE000448: 10332-8728). K) Strain U2366 (-) carries an IS186 with 8bp target site duplication (*bglB*: 352-359). **CFT073 type 46 strains:** L) 44 strains have intact *bgl* locus like CFT073, in which 26 strains (i484, F1, F385, F560, W7483, U2388, U2873, U3362, U4437, E10079, E10091, E182, E175, E452, ECOR23, ECOR32, ECOR51, ECOR52, ECOR53, ECOR54, ECOR55, ECOR57, ECOR60, ECOR63) show a relaxed phenotype, 5 strains (St5119, U3145, E176, E471, ECOR56) show more papillae, 14 strains (U3454, U3407, E10094, E478, E457, E177, E178, E464, E466, ECOR59, ECOR61, ECOR62, ECOR64, ECOR65) papillate like MG1655 and 1 strain (E475) does not papillate. M) Strain E422 (+) carries an IS1397 with 3 bp target site duplication (AE016769: 167486-167488). N) Strain F911 (-) carries an IS1294 and associated deletion from *bglF*: 931 to *yieH*: 256 (AE016769: 173583-166484). **O157 type 33 strains:** O) 32 strains have intact Z5211-Z5214 locus like O157, in which one strain (ECOR49) show weak Bgl^+ phenotype (day 2) of the remaining 31 strains, 22 strains (F645, F905, St5679, U3292, U4409, U5070, E10093, E10098, E460, E173, E10100, E472, E10084, E424, E10089, E179, E10095, ECOR37, ECOR38, ECOR39, ECOR40, ECOR43, ECOR44, ECOR47, ECOR48, ECOR50) does not papillate and 9 strains (E10100, E472, E10084, E424, ECOR35, ECOR36, ECOR41, ECOR42, ECOR46) show late papillae. P) Strain W7716 (-) carries an IS1 insertion and associated deletion (AE005603: 8093-1600). **Fourth type 34 strains** have upstream region like O157, followed by MG1655 type *bgl* and downstream with 5' end of the *yieI* gene like CFT073. Q) 24 strains have intact bgl like MG1655, 5 of these (U4191, U3104, ECOR19, ECOR45, ECOR67) show more papillae, 15 strains (F557, F742, V9261, V10744, U3622, U4252, U5033, E10087, ECOR7, ECOR26, ECOR27, ECOR28, ECOR70, ECOR71, ECOR72) papillate like MG1655 and in the remaining four strains, 3 strains (E10077, F569, U3372) show late papillae and one strain (St4723) does not papillate. R) Strain W9887 (++) carries an IS1 insertion with 9bp target site duplication (*bglH*: 97-105). S) Two strains ECOR20 (t1 +102gt) (-), ECOR21 (-) carries an IS1 insertion with 9bp target site duplication (*bglB*: 1060-1068). T) Strain ECOR18 carries an IS1 insertion with 9 bp target site duplication (AE000449: 8196-8204) and also carries IS1 associated deletion (AE000449: 8165-6293). U) Strain ECOR9 carries an IS1 associated deletion from *bglB*: 862 to *bglH*: 1602 (AE000449: 4224-2003). V) Strain ECOR17 carries an IS1 insertion and associated deletion from *bgl*t1 to *bglK*: 597 (AE000449: 8016-112). W) 4 strains (U2183, (+), ECOR58 (+), ECOR69 (+), and V9343, (late papillae)) have MG1655 sequence and lacks *yieJ* as CFT073. **Fifth type 20 strains** have mixture of sequences from MG1655 and CFT073. The upstream, *bgl* and downstream (with the exception of 5' yieI gene like CFT073) are like MG1655. X) 16 strains have intact *bgl* locus and downstream like MG1655 with 5' end of *yieI* gene like CFT073. In which 3 strains (F775, W8987 (t1 +102ga), E294) show more papillae, 9 strains (W9763 (t1 +102ga), E10092, E174, ECOR15, ECOR22, ECOR33, ECOR34, ECOR68, ECOR30) papillate like MG1655 and in the remaining 4 strains, 3 strains ECOR16, ECOR24 and ECOR29 show late papillae and strain ECOR6 does not papillate. ? indicates sequence information is not known in that region. Y) 3 strains (E10086 (+), ECOR4 (+), ECOR31 (+)) have MG1655 sequence and lacks *yieJ* gene as in CFT073 and have 5' of the *yieI* gene like CFT073. Z) Strain U4417 (-) carries an IS1 and associated deletion from *bglG*: 297 to *yieI*: 153 (AE000449: 7655-1 + AE000448: 10332-9375) and carries 5' end of *yieI* gene like CFT073. **Mixed type 4 strains** have mixture of sequences from MG1655, CFT073 and O157. AA) Strain E165 (+) has an upstream region and *bgl* like MG1655 followed by the downstream sequence like CFT073 with no *yieJ* gene (sequence of CFT073 starts at AE016769: 167468). AB) Strain E467 (+) have upstream and *bgl* like CFT073 followed by downstream sequences like MG1655 (sequence of MG1655 starts at AE000448: 10339) with 5' *yieI* gene like CFT073. AC) 2 strains E7370

(+++) and ECOR66 (+++) have upstream, *bgl* and downstream like CFT073 with 5' end of the *yieI* gene like MG1655. AD) Strain E10083, (-) has neither *bgl* nor Z5211-Z5214. However, it has some O157 sequence beyond the upstream breakpoint and downstream sequence in the *yieH* gene is like MG1655. The deletion in E10083 (AE005603: 8120 to 780) is associated with the insertion of 9 additional bases (5' TTTCTTTAT) in between the deletion endpoints. Text direction in the insertion elements represents the relative orientations of the insertion elements according to their defined left and right ends. Strain names are shown as indicated in Table 6 (Appendix).

Out of 33 O157 *bgl*/Z type strains, one strain has an IS1 insertion and associated deletion within Z5211-Z5214 region (Fig. 8P). Nucleotide sequencing at the downstream region of the fourth *bgl*/Z type strains revealed that all the 34 strains have *yieJ* and *yieH* sequences like MG1655. However, the *yieI* gene is with hybrid sequences i.e., the 5' end of the *yieI* gene is like CFT073 (with 6 additional base pairs in comparison to MG1655 *yieI* sequence) and the 3'end is like MG1655. Out of these 34 fourth *bgl*/Z type strains, 10 strains show alterations by insertions of IS1 and deletions within the *bgl*/Z5211-Z514 locus (Fig. 8Q to W). 16 Strains in the fifth *bgl*/Z type have upstream and intact *bgl* operon like MG1655 followed with MG1655 downstream structure and 5' end of *yieI* gene like CFT073 (Fig. 8X). Strains E10086, ECOR4 and ECOR31 that have upstream and *bgl* like MG1655 lacks the *yieJ* gene like CFT073 and possess the 5' part of the *yieI* gene like CFT073 (Fig. 8Y). Strain U4417 carries an IS1 associated deletion and has upstream, *bgl*, and downstream like MG1655 with 5' end of *yieI* gene like CFT073 (Fig. 8Z).

Strains in the mixed *bgl*/Z type differ from the other types in having a mixture of sequences from MG1655, CFT073 and O157. Strains E7370 and ECOR66 have upstream, *bgl* and downstream like CFT073 but carries 5' end of the *yieI* gene like MG1655 (Fig. 8AC). One of the mixed type strain E10083 does not have neither the *bgl* nor the Z5211-Z5214 genes however, the upstream *phoU* sequence is like O157 followed by the 5' part of Z5211-Z5214 region and the downstream *yieH* sequence is like MG1655 (Fig. 8AD). The details of the structural and phenotype analysis are presented below.

1.3 β-glucoside (salicin) utilization phenotypes of *E.coli* isolates

Aryl-β-glucosides are used as a carbon source by many bacteria (Kharat and Mahadevan, 2000; El Hassouni et al., 1992; Schnetz et al., 1996; Faure et al., 2001; Cote et al., 2000; Brown and Thomson, 1998; Breves et al., 1997; Brehm et al., 1999). In *E.coli* K-12 utilization of β-glucosides like arbutin and salicin requires the expression of *bgl* operon. Wild type *E.coli* K-12 cells are phenotypically Bgl^-. However, upon spontaneous activation Bgl^+ mutants arise as papillae (Schaefler and Malamy, 1969). To understand whether the variations seen at the *bgl*/Z5211-Z5214 locus in the 171 *E.coli* isolates have an impact on the β-glucoside utilization, the strains were analyzed for their phenotypes on BTB salicin plates at 37°C. Phenotypes were

noted daily up to 5 days (see materials and methods). Three different types of papillation phenotypes were seen (Fig. 9).

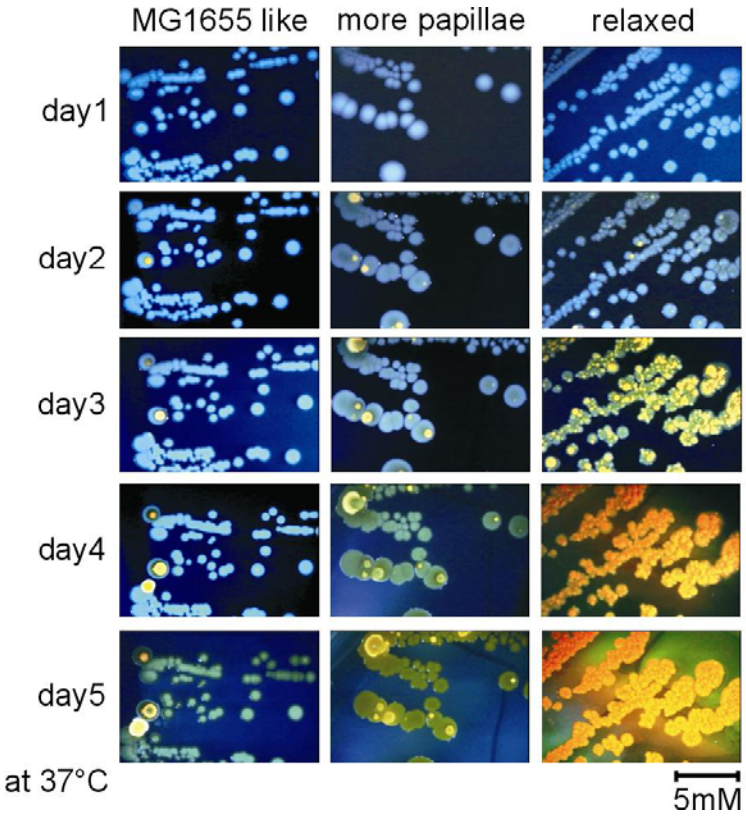


Figure 9: Phenotypes on BTB salicin plates at 37°C. All the 171 *E.coli* strains were streaked on Bromthymol blue (BTB) salicin indicator plates and incubated at 37°C. Following day 1 incubation images were taken using Zeiss stemi 2000-C Microscope and the observations were recorded up to day 5. Shown are the representative images of the three different papillation phenotypes seen. *E.coli* K-12 MG1655 strains are phenotypically Bgl^- at day 1 (seen as blue colonies). Upon further incubation (from day 2 onwards) spontaneous Bgl^+ mutants arise (seen as orange papillae). 60 strains show K-12 MG1655 like phenotype where the number of papillaes seen were comparable to that of MG1655 (MG1655 phenotype). 27 strains papillate more frequently than MG1655 (more papillae), 26 strains show weakly Bgl^+ phenotype at day 3 (relaxed phenotype) where the entire colony surface is covered with numerous tiny orange papillae. 1 strain show weak Bgl^+ on day 2 incubation and in the remaining 57 strains, 16 strains papillate late after day 5 incubation and 41 strains do not papillate even after prolonged incubation. (Table 6, Appendix).

35% of the strains papillate like K-12 MG1655 with 10-30 papillae's (in the conditions we have analyzed) are grouped under MG1655 like phenotype. 16% of the strains that papillate more frequently than MG1655 with 100-200 papillae's are grouped under more papillae phenotype. In contrast to the other two types, 15% of the strains showed a weak Bgl^+ phenotype on day 3 (Fig. 9) which is assigned as relaxed phenotype in the current study. Strains that show relaxed phenotype papillate when incubated at 28°C (data not shown). In addition to these three types, one strain (ECOR49) that has Z5211-Z5214 locus showed a weak Bgl^+ phenotype on day 2 of incubation at 37°C. Among the other remaining 33% some of the strains either papillate late (after day 5), or do not papillate even after prolonged incubation (up to day 10) on BTB salicin plates.

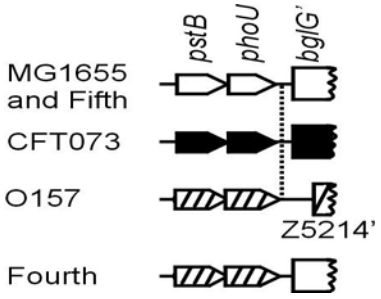
1.4 Nucleotide polymorphism at the upstream region of *bgl*/Z5211-Z5214

Nucleotide sequence alignment of *bgl*/Z5211-Z5214 region (including upstream and downstream regions) from the four sequenced strains showed sequence variations between the strains (Fig. 7). Based on the sequence variations at the upstream and the downstream regions, the four sequenced strains can be distinguished from each other (Fig. 7). In addition, alignment of the

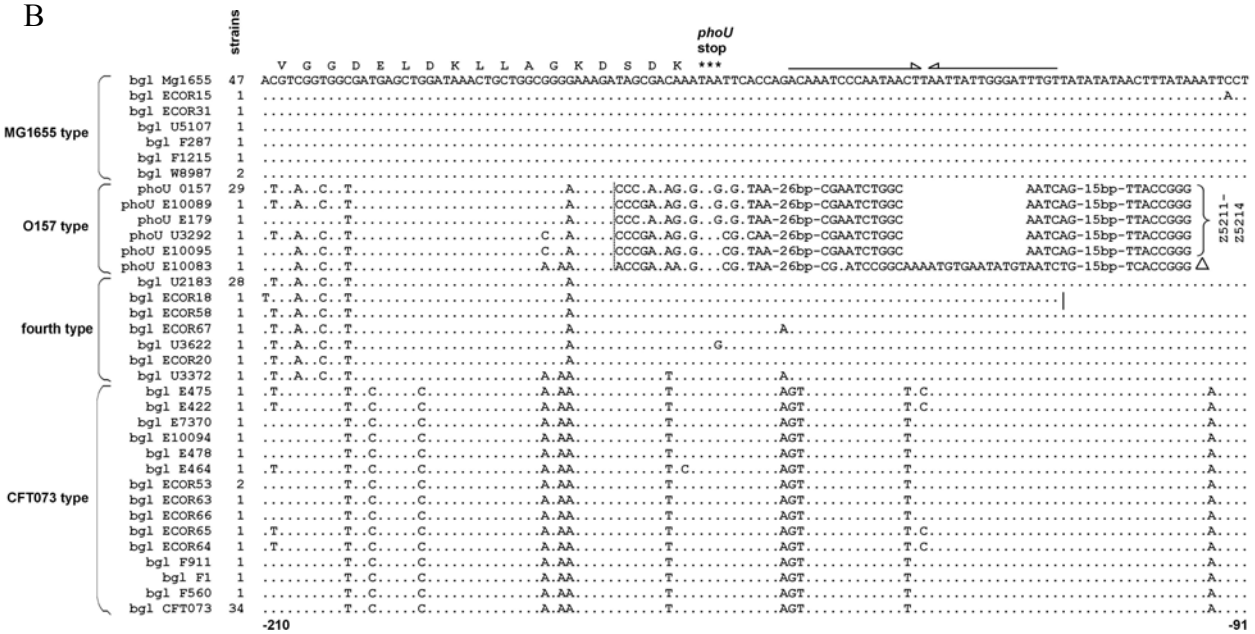
nucleotide sequence of the *bgl* operon from MG1655 and CFT073 also do show variations. The strain CFT073 carries around 12 bp changes in the *bgl* promoter region in comparison to MG1655 (Fig. 10). To analyze whether these polymorphisms are present in the *E.coli* isolates, the nucleotide sequence at the upstream and *bgl* promoter region was determined. The strains were analyzed by PCR with *bgl* specific primers (Fig. 37 materials and methods, Table 4, Appendix) to see whether they carry the *bgl* operon. In the strains that carry the *bgl* operon the sequence of the upstream region (*phoU* gene) and the *bgl* regulatory region were amplified by PCR with *bgl* specific primers and the PCR fragments were sequenced. The strains that showed no PCR product with *bgl* specific primers and carry no *bgl* (negative PCR) were analyzed by ST-PCR (for the presence of Z5211-Z5214) and the obtained PCR products were sequenced.

Based on the nucleotide sequence alignment with the published sequences of MG1655, CFT073 and O157, the strains were grouped into four types at the upstream and 5'end of the *bgl*/Z5211-Z5214 locus: MG1655 type, CFT073 type, O157 type and fourth type.

A



B



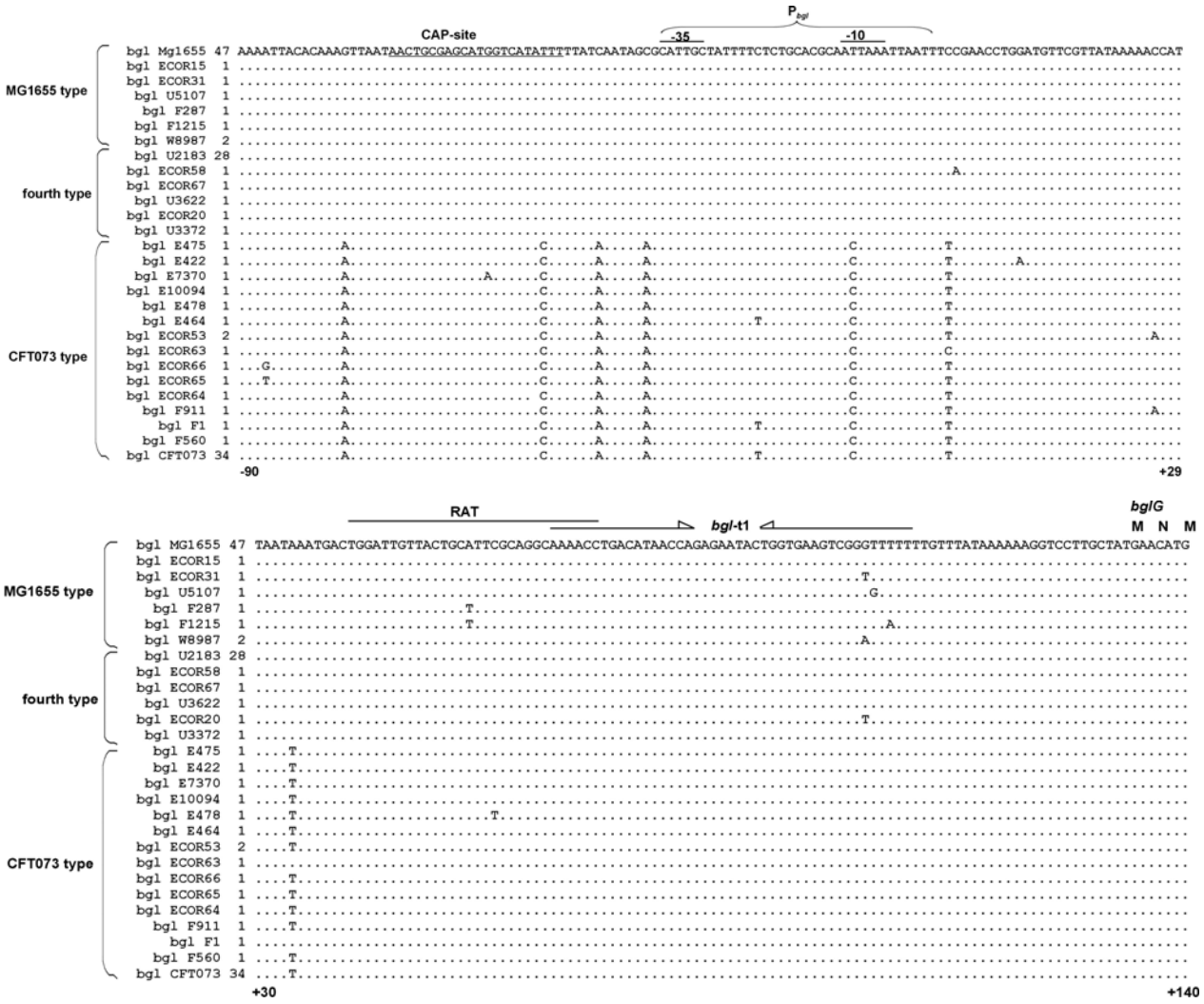


Figure 10: Nucleotide sequence alignment of the upstream region of the *bgl* and Z5211-Z5214 islands. **A)** Schematic presentation of the upstream region of *bgl* island in MG1655 and CFT073, Z5211-Z5214 island in O157 (EDL933 and Sakai). Fourth types have upstream sequence like O157 followed by MG1655 *bgl* sequence. Structures are represented as described in Figure 4 (introduction). The dotted line indicates the 5' end of Z5211-Z5214 region. The *bglG*' and Z5214' indicate the 5' end of the *bglG* and Z5214 genes respectively. **B)** Nucleotide sequences of 171 strains were aligned and compared with the published sequences of MG1655, CFT073 and O157. Representative strains in each type along with the strains that have specific base-pair changes are shown (refer Table 6, Appendix). Conserved nucleotides in comparison to MG1655 are shown as dots. Base pair changes in each strain are shown. Deduced amino acid sequence of MG1655 PhoU and BglG is shown at the top of the alignment. Horizontal arrows above the sequence represent the inverted repeats in the *phoU* and *bgl* terminator t_1. Promoter P_{bgl} is marked by a brace. The CAP binding site is underlined. BglG binding site (RAT) is shown as a line above the sequence. Typing of the strains based on the nucleotide sequence alignment is shown at the left. Strains of MG1655 type have sequence like MG1655; Strains of fourth type possess *phoU* gene sequence like O157 followed by MG1655 *bgl* sequence. Strains of CFT073 type have 11 to 12 base pair changes in the *bgl* promoter region. The sequence variations seen in the strains of O157 type at the end of *phoU* and the vertical line mark the start of Z5211-Z5214 region. Strain E10083 neither carries *bgl* nor Z5211-Z5214; however, some of the Z5211-Z5214 sequence (with additional 15 bp, see alignment) is present till the breakpoint (Δ). Numbering below each alignment is relative to the transcription start site of the *bgl* operon. Number of strains in each sequence types is shown. Strain names are listed as in Table 6 (Appendix).

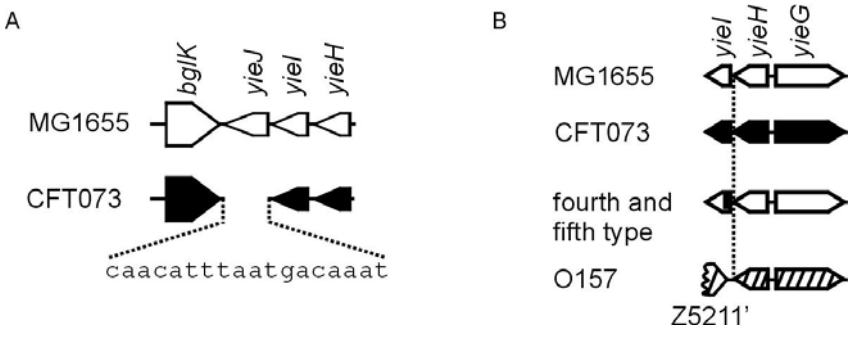
Strains that have MG1655 *bgl* sequence are grouped as MG1655 type. Some of the MG1655 *bgl*/Z type strains have single nucleotide change in the *bgl* terminator t1 (Fig. 10B). Strains of the MG1655 and CFT073 *bgl*/Z type differ at 11 to 12 bp in the *bgl* promoter region (Fig. 10B).

Strains of the O157 *bgl*/Z type have Z5211-Z5214 region as in O157. In addition to these three types, a group of strains (fourth *bgl*/Z type) that have sequence variations in the *phoU* gene (as in O157) followed by MG1655 *bgl* sequence was observed. The strains that have specific variations within each type are shown in the alignment (Fig. 10B and Table 6, Appendix).

1.5 Nucleotide polymorphism at the downstream region of *bgl*/Z5211-Z5214

The grouping of the *E.coli* strains based on the alignment of the nucleotide sequences of the upstream region aggravated an interest to analyze the nucleotide sequences at the downstream region. Hence, for the strains in the fourth *bgl*/Z type the nucleotide sequence for the downstream region of *bgl*/Z5211-Z5214 was determined. In addition, ST-PCR was performed for the strains of O157 *bgl*/Z type (Table 6a, Appendix) and the PCR products were sequenced. Based on the nucleotide sequence alignment with the published sequences of MG1655, CFT073 and O157, the strains were grouped into five types at the downstream and 3'end of the *bgl*/Z5211-Z5214 locus: MG1655 type, CFT073 type, O157 type, fourth type and fifth type (Fig. 11). Most of the strains that were typed based on the nucleotide sequence alignment of the upstream region were found to be clustered into the same type in the alignment of the downstream sequences (Table 6, Appendix).

The downstream sequences in the *yieJ* and *yieH* genes of the fourth *bgl*/Z type strains are like MG1655 (with few sequence variations). However, the *yieI* gene is with a mixture of sequences from CFT073 and MG1655. The strain CFT073 has additional 6 bp AGTACC in the *yieI* gene in comparison to MG1655 (Fig. 11C). The *yieI* gene observed in the fourth type strains has these additional six bases and a stretch of CFT073 sequence at the 5' end of the gene. However, the 3'end of the *yieI* gene is like MG1655 (Fig. 11C and 8Q to W). In addition, nucleotide sequencing results revealed that twenty strains have downstream sequences like MG1655 with the exception of carrying hybrid *yieI* gene. This group is named as fifth *bgl*/Z type in the current study (Fig. 8X to Z). Fifth *bgl*/Z type strains have upstream and *bgl* like MG1655 followed by downstream like MG1655 with hybrid *yieI* gene.



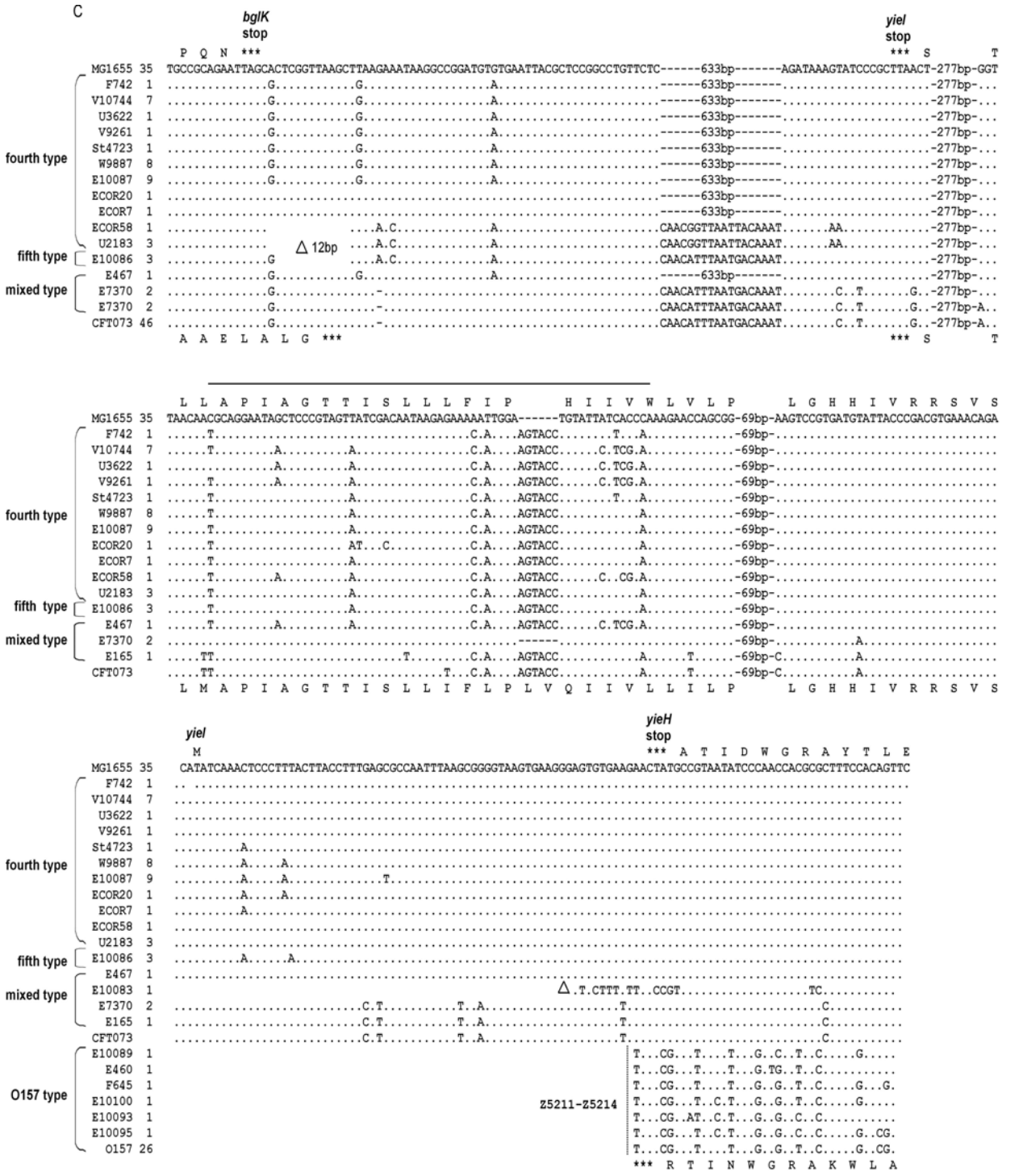


Figure 11: Nucleotide sequence alignment of the downstream region of *bgl*/Z5211-Z5214. **A)** Comparison of the *bglK-yieI* region of MG1655 and CFT073 is shown. CFT073 lacks the *yieJ* gene and carries an additional 18 bp sequence in comparison to MG1655. **B)** Comparison of the *yieH* region of MG1655, CFT073, O157 and fourth type. Fourth type and fifth type are identical to one another at the downstream region. Dotted lines indicate the 3' end of Z5211-Z5214 region. Z5211' indicates 3' end of Z5211. O157-EDL933 and Sakai possess identical sequences at the regions shown in A and B. **C)** Nucleotide sequences of the strains were aligned and compared with the published sequences of MG1655 (AE000449: 80-1+ AE000448: 10332-9118), CFT073 (AE016769: 167540-166856) and O157 (AE005603: 779-741). Base pair changes and conserved nucleotides are represented as in Fig. 10. Deduced amino acid sequence of MG1655 BglK, YieI, YieH (top of the alignments); CFT073 BglK, YieI (bottom) and YieH (top); O157 YieH (bottom) is shown. Horizontal line represents the hybrid *yieI* part. In fourth and fifth type apart from the sequence shown with the horizontal line the rest of the sequences are like MG1655 (with few sequence variations). Typing of the strains based on the nucleotide sequence alignment is shown at the left. Representative

strains in each type along with the strains that have specific base-pair changes are shown. Strains of fourth type possess *bglK* sequence like MG1655 followed by MG1655 downstream sequence (except the 5'end of *yieI* gene). O157 type strains have sequence variations in the *yieH* gene as in O157. The vertical line marks the 3'end of Z5211-Z5214 region. Strain E10083 neither carries *bgl* nor Z5211-Z5214; Δ represents the downstream breakpoint till the upstream sequences (shown in Fig. 10). Deletion of 12bp in strains U2183, ECOR58 and E10086 are shown as Δ12bp. Number of strains in each sequence types is shown. Refer Table 6 for more details.

Furthermore, apart from the five main types seen, a sub-type of strains (Table 6, Appendix) that have mixture of sequences from MG1655, CFT073 and O157 in the *bgl*/Z5211-Z5214 locus was observed. This sub-group is designated as a mixed *bgl*/Z type in this study (Fig. 8AA to AD) for e.g. strain E165 that has upstream and *bgl* sequence like MG1655 has downstream sequences like CFT073 (Fig. 11C and Fig. 8AA). In addition, hybrid *yieI* gene was also observed in mixed *bgl*/Z type strains (Fig. 8AB and AC). The strains of O157 *bgl*/Z type have *yieH* gene sequence as in O157 (Fig. 11).

1.6 Southern hybridization analysis for the strains that did not papillated on BTB salicin plates.

Out of the 99 clinical strains, 24 strains did not papillate on BTB salicin plates at 37°C (Table 6a, Appendix). All these strains were analyzed by Southern hybridization with various *bgl* region specific probes (Figure 12 and Fig. 41, Appendix). 18 out of 24 strains are of O157 *bgl*/Z type (determined by ST-PCR, Fig. 8O and P). The templates used for probe preparation were isolated either as PCR products from MG1655 cells or by restriction analysis of pFDX733 or pFDY52 (see materials and methods). Hybridization to the genomic DNA from O157 *bgl*/Z type strains with *bgl* specific probes (probe C, D, E, F, G, and H, see Fig.41IV to X , Appendix) did not give any signals as seen in MG1655 (used as control) indicating the lack of *bgl* operon in this group of strains. However, with probes A, B, I and J signals were seen as expected in case of O157 (Fig. 12 and Fig.41I to III, XI to XV Appendix), indicating the presence of intact Z5211-Z5214 region (Fig. 8O). The presence of Z5211-Z5214 region correlates to upstream and downstream sequencing analysis of O157 *bgl*/Z type strains (Fig. 10 and 11). Strains W7716 and E10083 have Z5211-Z5214 sequences, however, unexpected signals were seen in both the strains indicating either insertions or deletions within the Z5211-Z5214 region (Fig. 12, Fig. 8P, 8AD and Fig.41I to III, XII to XV, Appendix). In addition to O157 *bgl*/Z type, 6 strains that have *bgl* operon (determined by upstream and downstream nucleotide sequencing) and do not papillate were also analyzed by Southern hybridization with *bgl* specific probes (Fig. 8A, 8H, 8K, 8Q, 8Z, and Fig. 12). Strain U4417 that has upstream *bgl* sequences like MG1655, gave unexpected signals with probes A, C and I (Fig. 8Z, Fig. 12 and Fig.41II, V, XII). No signals were seen with probe D, F, H indicating a possible deletion within the *bgl* operon (Fig.41VI, VII, Appendix).

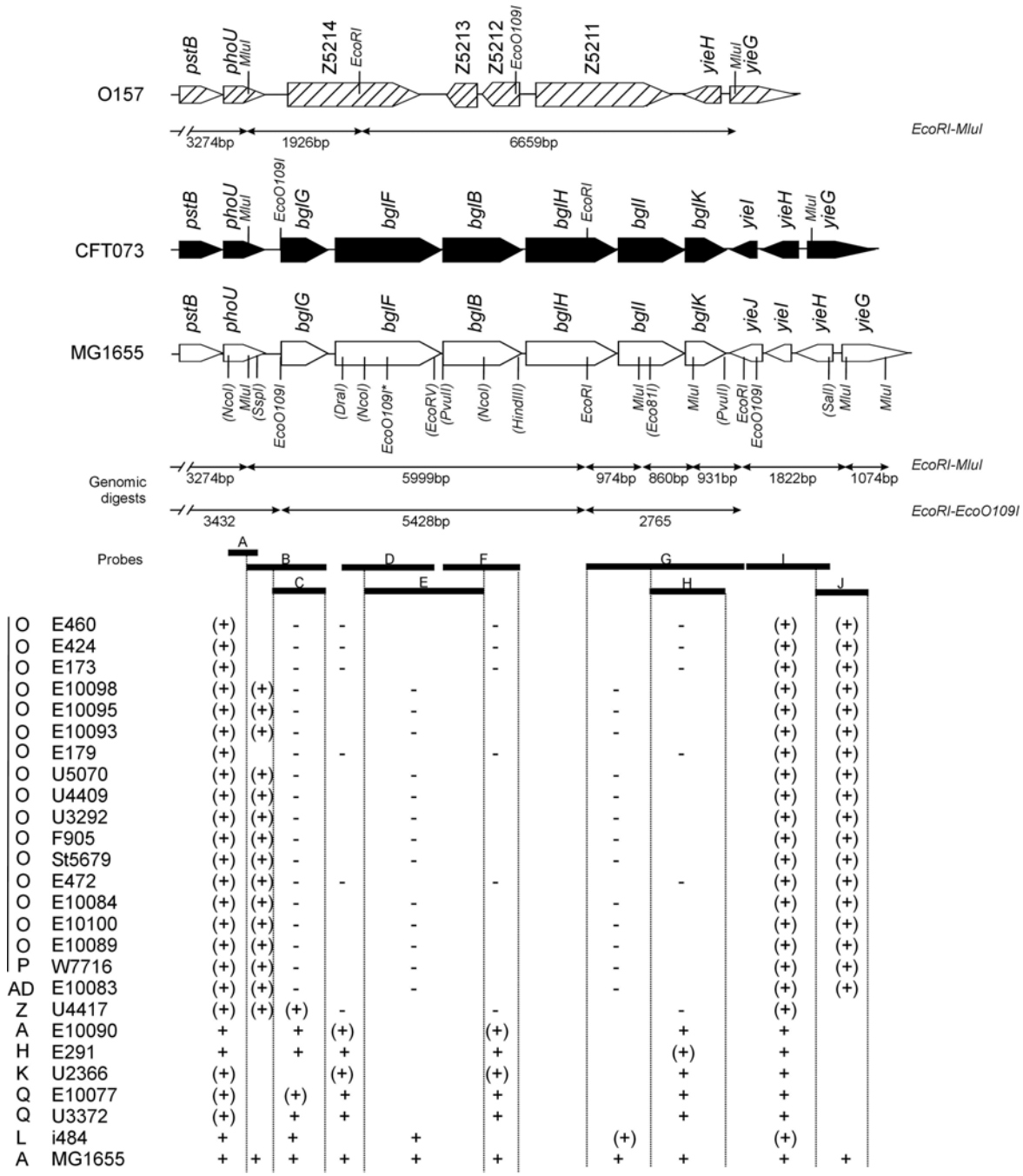


Figure 12: Southern analysis of the *bgl*/Z5211-Z5214 region in *E.coli* isolates. Strain names are indicated as listed in Table 6 (Appendix). Letters corresponds to the structures shown in Fig. 8. Restriction sites are indicated as black lines, the regions used as probes are shown as black bars: probe A (NcoI-SspI), B (S145/S201), C (S157/S201), D (DraI-EcoRV), E (NcoI-NcoI), F (PvuII-HindIII), G (EcoRI-EcoRI), H (Eco81I-PvuII), I (EcoRI-SalI), J (S335/S336). + indicates presence, - indicates absence, (+) indicates signal of unexpected size in comparison to MG1655. Strains that have O157 sequences are indicated by vertical black line at the left. Primers used for probe preparation are named as listed in Table 4 (Appendix). Gel images are shown in Figure 41 (Appendix).

To map the deletion point in strain U4417, PCR was carried out with the primer matching in the upstream region (*bglG* gene) and the other in downstream region (*yieI* gene). The obtained product was sequenced from both the ends. The sequencing result showed that U4417 carries an IS1 associated deletion from *bglG* gene to *yieI* gene (Fig. 8Z). Strains E10090, E10077, U3372

carry the intact *bgl* operon with no structural alterations, indicating that a point mutation(s) might have resulted in the non-utilization of β-glucosides in these three strains (Fig. 8A and 8Q).

Furthermore, Southern hybridization was performed with 6 different genomic digests of strain i484 along with MG1655 genomic digests as control (Fig. 13A & Fig.41XVIII, XIX). The genomic digests were probed with two probes: Probe K that maps in the *bglH* gene and Probe L that map in the *yieIHG* genes. The results revealed that i484 have *bgl* region like CFT073 that lacks *yieJ* gene (Fig.8L, 13A and Fig.41XVIII, XIX, Appendix).

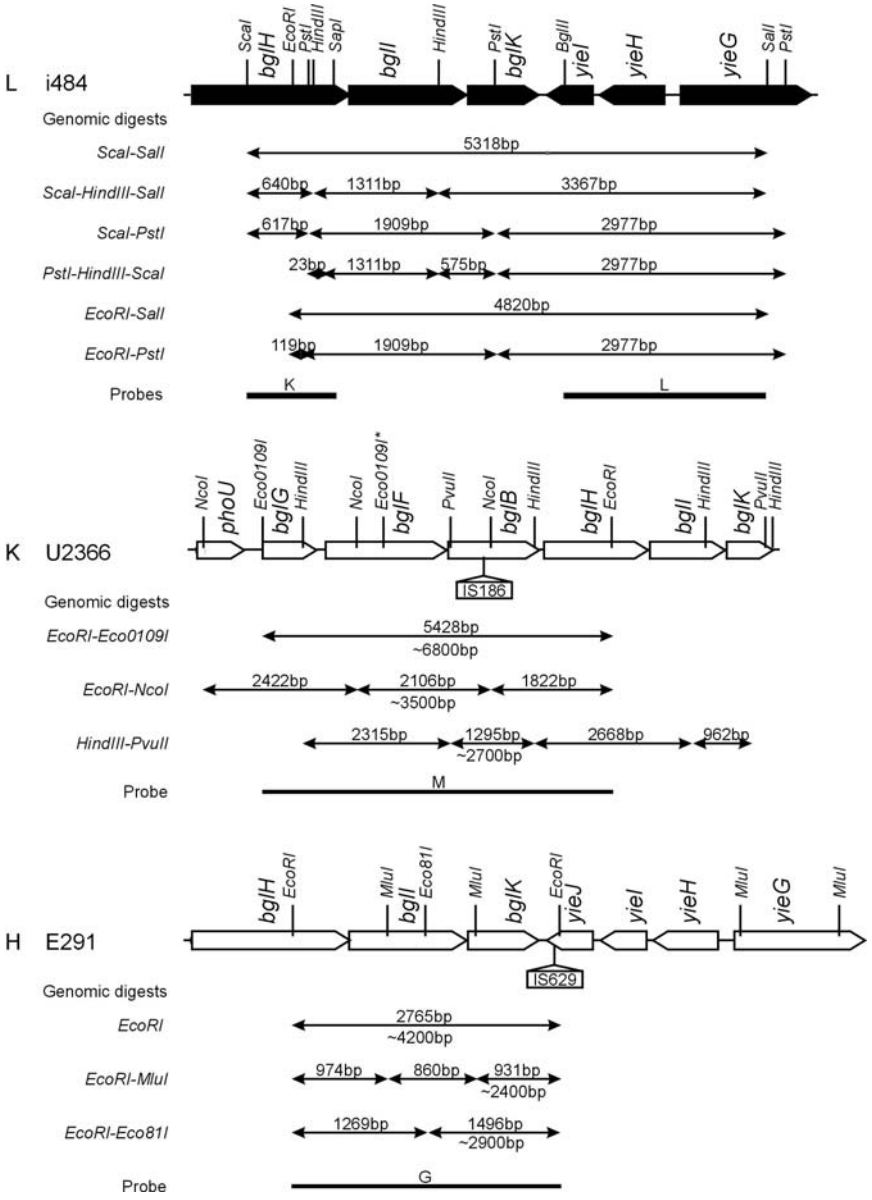
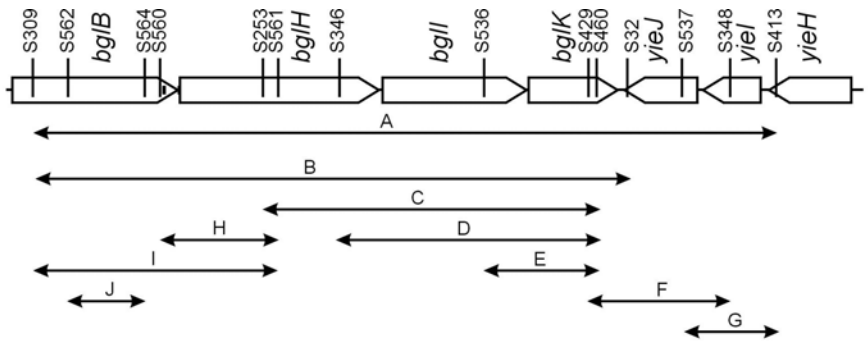


Figure 13: Southern analysis to fine map the insertions or deletions in the *bgl* region. Restriction sites are indicated as black lines. The regions used as probes are shown as black bars: probe K (ScaI-SapI), L (BglII-SalI), M (Eco01091-EcoRI),G (EcoRI-SalI). Sizes expected for MG1655 genomic DNA digests are shown at the top of horizontal arrows and the variation in the signal size is shown at the bottom **A)** Strain i484 genomic digests are shown, signals were seen as expected in CFT073. **B)** Strain U2366 genomic digests probed with M showed increased signal of ~1.4kb **C)** Strain E291 genomic DNA was digested in 3 different sets and probed with G. An increase of ~1.5kb size was seen. Gel images are shown in Fig. 41 (Appendix). Letters corresponds to the structures shown in Fig. 8.

Strain E291 showed an increase in size (~1.4 kb) with probe H, indicating an insertion in the *bglHIK-yieJ* region (Fig.12H and Fig.41X, Appendix). Further fine mapping (Fig. 13C & Fig.41XVI), PCR and nucleotide sequencing revealed that E291 carries IS629 insertion in *yieJ* gene (Fig. 13C and Fig. 8H). The increased in size corresponds to the insertion element shown in Figure 8H. Insertion in *yieJ* gene cannot explain why the strain cannot utilize β-glucoside, as the first three genes are necessary and sufficient for the utilization of β-glucosides. It is possible that in strain E291 a point mutation (s) in the structural genes might have resulted in the non-utilization of β-glucosides. Strain U2366 showed an increase in size (~1.5 kb) with probe A. D, F indicating a possible insertion within the first four genes of the *bgl* operon (Fig. 12K and Fig.41II, VI, VIII). Further mapping of the insertion by using probe M and three different genomic digests showed insertion in *bglB* (Fig.13B & Fig.41XVII). The insertion size corresponds to the insertion element shown in Figure 8K. PCR and nucleotide sequencing revealed that U2366 carry IS186 insertion in *bglB* (Fig. 8K). The ability to utilize β-glucosides in strain U2366 might have been lost due to the inactivation of *bglB* by IS186 insertion.

1.7 Long PCR analysis to analyze the alterations within the *bgl*/Z5211-Z5214 locus

Results from the Southern hybridization and nucleotide sequencing analysis revealed that strains show alterations within the *bgl*/Z5211-Z5214 locus by deletions and insertions. In order to analyze the genomic alterations within the *bgl*/Z5211-Z5214 locus in the strains that were not analyzed by Southern hybridization (Table 6, Appendix), a long PCR strategy was employed. This strategy involves the use of Elongase enzyme (Invitrogen) that can amplify DNA up to 30 kb. In order to analyze for the presence of intact *bgl* operon or Z5211-Z5214, long PCR's were carried out with the primers that map in the upstream region (*phoU*) and in the downstream region (*yieH*) and the PCR products were compared to the MG1655 and i484 (like CFT073) sizes (controls). Strains that have intact *bgl* region like MG1655 gave expected band size like MG1655 and strains that have CFT073 like *bgl* region gave expected band size like i484 (like CFT073). Strains of O157 *bgl*/Z type gave expected PCR product size as expected in O157. The strains that gave expected sizes like MG1655, CFT073 and O157 were further analyzed by PCR's with primers specific for *bgl* or Z5211-Z5214 (Table 6a and 6b, Appendix). Furthermore, out of 21 strains that showed differences in the PCR product sizes in comparison to the expected sizes of MG1655, CFT073 and O157, 19 strains (remaining 2 strains were analyzed in Southern analysis, see section 1.6, Results) were analyzed by several internal PCR's to map the possible insertion and/or deletion points within the *bgl*/Z5211-Z5214 region (Fig. 14 and Table 6a and 6b, Appendix). A principle strategy of such an analysis is shown in Figure 14.



	Strains	PCR A	B	C	D	E	F	G	H	I	J
D	E167	(+)	(+)	+	+	+	+	+	(+)	(+)	+
R	W9887	(+)	(+)	+	+	+	+	+	(+)	(+)	+
E	ECOR12	(+)	(+)	+	+	+	+	+	(+)	(+)	+
S	ECOR20	(+)	(+)	+	+	+	+	+	+	(+)	(+)
S	ECOR21	(+)	(+)	+	+	+	+	+	+	(+)	(+)
A	MG1655	+	+	+	+	+	+	+	+	+	+

Figure 14: Schematic representation of PCR analysis to map the insertions or deletions within *bgl*/Z5211-Z5214 region. Strain names are shown as in Table 6 (Appendix) and oligos are shown as in Table 4 (Appendix). The MG1655 *bgl* region is shown as open black arrows. Oligos are shown with vertical lines above the genes. Horizontal arrows represent the PCR reactions performed with the respective oligos. A: S309/S413, B: S309/S32, C: S253/S460, D: S346/S460, E: S536/S460, F: S429/S348, G: S537/S413, H: S560/S561, I: S309/S561, J: S562/S564. + indicates expected size seen as in MG1655, (+) indicates increased size seen in comparison to MG1655. Refer Table 6 for the strain names and the mutations they carry in the *bgl*/Z5211-Z5214 locus. Out of 21 strains that show insertions or deletions, 19 strains were mapped using long PCR strategy. The 19 strains (letters) that corresponds to the structures shown in Fig. 8 are : E167 (D), W9887 (R), ECOR12 (E), ECOR20 (S), ECOR21 (S), U3633 and E10096 (B), U4418 (C), E345 (F), U5107 (G), E292 (I), E164 (J), E422 (M), F911 (N), ECOR18 (T), ECOR9 (U), ECOR17 (V), U4417 (Z) and W7716 (P).

In the analysis shown above initial long PCR was carried out with oligos S309 and S413. The results showed that all the strains shown in Figure 14 (Table 6, Appendix) showed an increase in the PCR product size in comparison to MG1655. To further fine map the insertion points, several internal PCR's were carried out (Fig. 14).The relative PCR product was later sequenced from both the ends for e.g. the PCR product (S560/S561) of strain E167 was sequenced with oligos S560 and S561. The nucleotide sequencing results revealed that strains E167 and ECOR12 carry IS2 insertions in the *bglH* gene (Fig. 8D and 8E, respectively), strain W9887 carries IS1 insertion in *bglH* gene (Fig. 8R). Strains ECOR20 and ECOR21 carry IS1 insertions at the same position in *bglB* (Fig. 8S). Both these strains do not papillate on BTB salicin indicator plates. Since, *bglB* is essential for β-glucoside utilization the insertion of IS1 in the *bglB* may perhaps explain the no papillae phenotype seen in these two strains. To map the deletion points in the strains W7716 and E10083 (detected by Southern hybridization) PCR was carried out using a primer that maps in the

upstream region (*phoU* gene) and the primer that map in the downstream gene (*yieH*). The obtained PCR product was sequenced from both the ends. PCR and nucleotide sequencing results showed that strain W7716 carries an IS1 associated deletion from Z5212 to Z5214 (Fig. 8P) and strain E10083 carries complete deletion of Z5211-Z5214 region (Fig. 8AD).

From the PCR and sequencing analysis an insertion associated fragment encoded by ISEc8 was mapped in strains U3633, E10096 and U4418 (Fig. 8B and C). It was observed that ISEc8 generates 6 bp target site duplication at its insertion site. The insertion element ISEc8 belongs to IS66 family and carries a 22bp imperfect terminal inverted repeats. It shows 51.1% identity at the nucleotide level to the insertion element ISRm14 from *Sinorihizobium meliloti* (Schneiker, 1999). The mechanism of ISEc8 transposition is currently not understood.

1.8 A refined PCR strategy to analyze the downstream region and the presence of hybrid *yieI* gene

Nucleotide sequencing data of the upstream and downstream *bgl* and Z5211-Z5214 islands revealed the existence of *E.coli* strains that carry a hybrid *yieI* gene. This hybrid *yieI* gene is identical to the *yieI* gene of K12-MG1655 with a patch of sequence within its 5' end that is identical to the sequence of the *yieI* gene in CFT073. In addition some strains have a mixed structure (section 1.4 and 1.5). In order to analyze whether strains which according to PCR analysis have a downstream structure as K12-MG1655, i.e. carry the *yieJ* gene, or which lack *yieJ* as strain CFT073 carry a hybrid *yieI* gene or another mixed sequence structure a refined PCR strategy was designed. This strategy involves a two step PCR reactions using two sets of three primers (Fig. 15). Both sets of primers include a MG1655 specific *bglK* primer (S588) and a CFT073 specific *bglK* primer (S589).These two primers match at different positions and thus a PCR product generated with the MG1655 specific primer has a different size than a PCR product generated with the CFT073 specific primer. The reverse primer in the first set matches to the 5' end of the CFT073 type *yieI* gene (S587), while in the 2nd set it matches to the 5' end of the MG1655 type *yieI* (S586) (Fig. 15). In brief, by the first PCR reaction strains carrying a *yieI* gene with the 5' end of the CFT073 type can be characterized (see Figure 15B, C, D and F). Those strains that showed no PCR product in this first PCR were analyzed using the second set of three primers. This allowed to characterize strains in which the 5' end of the *yieI* gene is identical to MG1655 (see Fig. 15A and E). To distinguish whether strains carry a hybrid *yieI* gene or not, the PCR fragments of some strains were sequenced (Fig. 15B, C, E and F). All together, the data of the current study shows that out of 135 isolates that have *yieI* gene 43% have a hybrid *yieI* gene. The significance of the presence of the hybrid *yieI* gene is yet to be understood.

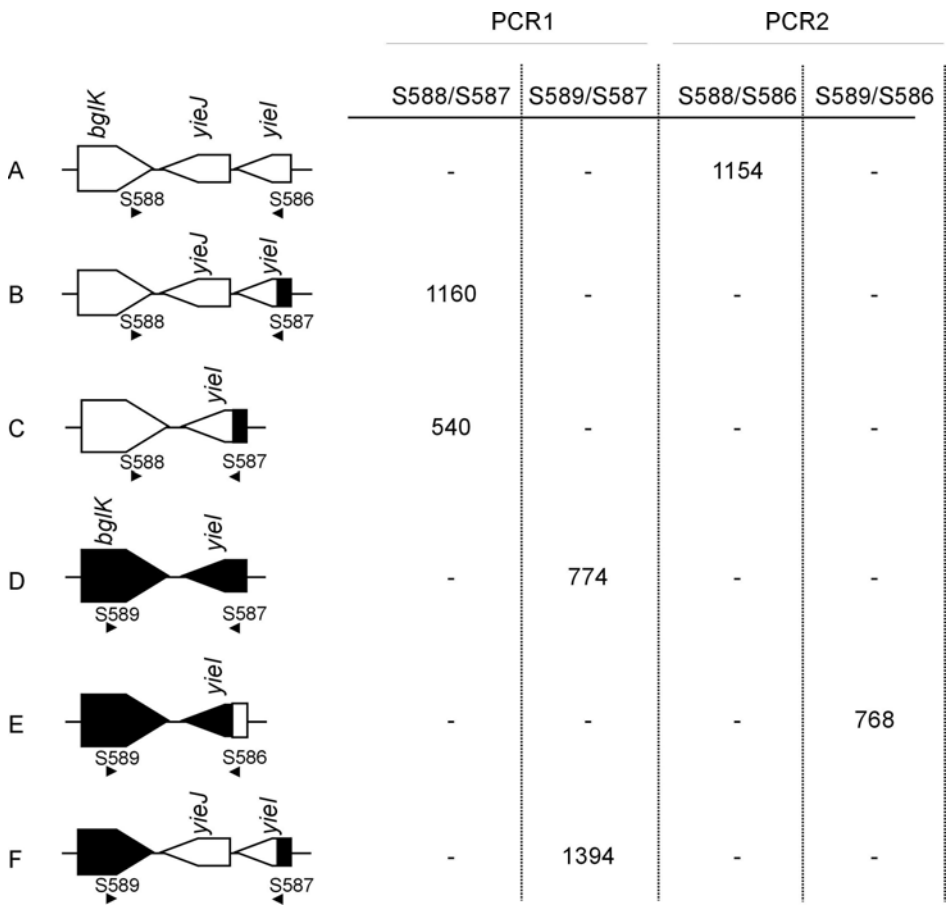


Figure 15: Schematic presentation of a PCR strategy used for typing *E.coli* strains at the downstream region of the *bgl* operon. Structures of *bglK-yieH* region are shown. Open black arrows and blocks represent MG1655 sequence, closed arrows and blocks represents CFT073 sequence. Structures seen in the current analysis at the downstream region are shown in A to F. Primer mapping positions are shown beneath each structure by arrow heads. This strategy involves two PCR reactions. The first PCR is carried out with the primer combination S588/S89/S587. This PCR results in the size of one of the PCR products that corresponds to the structures shown in B, C, D, F. Strains that showed no PCR product in the first PCR were analyzed with a second PCR using primers S588/S589/S586. All strains of which no PCR product was obtained with the first set of primers yielded a PCR product in the second PCR corresponding to the structures shown in A, E. The expected size of the PCR product for each structure is shown in base pairs. – indicates no PCR product.

1.9 Correlations of *bgl*/Z5211-Z5214 on β-glucoside utilization phenotypes.

A strong correlation was seen for the presence of *bgl*/Z5211-Z5214 and the β-glucoside utilization phenotypes of the strains. Out of 57 strains that show no papillae or late papillae, 58% of them are of O157 *bgl*/Z type with no *bgl* genes (Fig. 8O, P and AD) (Table 1a, b, and c), 14% have functional inactivation of *bgl* genes either by insertions or deletions (Fig. 8K, N, S, T, U, V, Z) within the *bgl* operon. The remaining 28% of the strains have no structural alterations in the *bgl* genes (Fig. 8A, H, L, Q, W and X). These 28% of the strains may possess point mutations inactivating β-glucoside utilization. Our analysis indicates that the *bgl* operon is a primary β-glucoside utilization system in *E.coli* present in approximately 70% of the isolates, and in all of the strains in which it is present, its silent state is conserved.

Table 1a: *bgl*/Z5211-Z5214 locus and β-glucoside utilization phenotypes in 171 *E.coli* isolates

	171 strains[a]	Phenotypes on BTB salicin plates at 37°C[C]				
	human com-81 animal com-32 uro-32 sep-25 ABU-1	K-12- 60 (35%) human com-31 (38%) animal com-12 (37%) uro-10 (31%) sep-6 (24%) ABU-1 (100%)	more papillae-27 (16%) human com-11 (13%) animal com-4 (12%) uro-6 (19%) sep-6 (24%) ABU-	relaxed-26(15%) human com-10 (13%) animal com-5 (16%) uro-6 (19%) sep-5 (20%) ABU-	no papillae-57 (33%) human com-28 (34%) animal com-11 (34%) uro-10 (31%) sep-8 (32%) ABU-	weak Sal$^+$-1(1%) human com-1 (1%) animal com- uro- sep- ABU-
bgl/Z5211-Z5214 locus[b]	MG1655 *bgl*-88 (51%) human com-36 (44%) animal com-22 (69%) uro-15 (47%) sep-14 (56%) ABU-1 (100%)	44 (26%) 20 (25%)[d] 11 (34%) 6 (19%) 6 (24%)[e] 1 (100%)	22 (13%) 8 (10%)[f] 4 (12%) 5 (16%)[g] 5 (20%)[h] -	-	22 (13%) 8 (10%)[i] 7 (22%)[j] 4 (12%)[k] 3 (12%)[o] -	-
	CFT073 *bgl*-49 (29%) human com-25 (31%) animal com- 6 (19%) uro-11 (34%) sep-7 (28%) ABU-	16 (9%) 11 (13%)[l] 1 (3%) 4 (12%) - -	5 (3%) 3 (4%) - 1 (3%) 1 (4%) -	26 (15%) 10 (13%) 5 (16%) 6 (19%) 5 (20%) -	2 (1%) 1 (1%) - 1 (4%)[m] -	-
	Z5211-Z5214-34 (20%) human com-20 (25%) animal com-4 (12%) uro-6 (19%) sep-4 (16%) ABU-	-	-	-	33 (19%) 19 (23%)[n] 4 (12%)[p] 6 (19%) 4 (16%) -	1 (1%) 1 (1%) - - - -

a: Total number of strains analyzed, human com-commensal strains from stool samples of healthy humans, animal com-commensals from animals, uro-uropathogenic strains, sep-septicemic isolates, ABU-asymptomatic bacteriuria causing strain **b**: *bgl*/Z5211-Z5214 locus was analyzed by PCR, sequencing, and southern hybridization. Numbers of strains are indicated, percentages are within brackets and the total numbers of strains in each category are also shown **c**: phenotypes on BTB salicin plates at 37°C. MG1655: papillates like MG1655, more papillae: papillates more frequently than MG1655, relaxed-weak Bgl$^+$ on day3 of incubation, none-no papillation or late papilallation, weak Sal$^+$-weakly salicin positive phenotype (day2) **d**: Strain E167 (*bglH*::IS2), E10096 (*yieJ*::ISEc8),E345 Δ(*bglH-bglI*)::IS629, *yieI*::IS629, E292 Δ(*yieJ-yieI*)::IS629 **e**: W9763 t1 +102ga **f**: ECOR12 (*bglH*::IS2), E164 Δ(*bglI-yieH*)::IS1 **g**: U3633 *yieJ*::ISEc8, U4418 *yieJ*::IS1, *yieJ*::ISEc8, U5107 Δ(*bglK-yieJ*)::IS629 **h**: F1215 t1 +105ta, W8987 t1 +102 ga, W9887 (*bglH::IS1*) **i**: E291 *yieJ*::IS629, ECOR9 Δ(*bglB-bglH*)::IS1. Strains E10077 and ECOR24 show late papillation **j**: ECOR17 Δ(*bgl*t1-*bglK*), ECOR20 *bglB*::IS1, ECOR21 *bglB*::IS1, ECOR18 Δ (P_{bgl}-*bglF*)::IS1. Strains ECOR16, ECOR29 show late papillation **k**: U4417 Δ*bglG-yieI*, U2366 bglB::IS186 and strain U3372 show late papillation, **l**: E422 *bglK*-::IS1397-*yieJ*, **m**: F911 Δ(*bgl*F-*yieI*)::IS1294 **n**: E10083 ΔZ5211-Z5214. Strains ECOR35, ECOR36, ECOR41, ECOR42, E10100, E10084, E472, E424 show late papillation **o**: F569, V9343 late papillation **p**: ECOR46 show late papillation

Table1b: *bgl*/Z5211-Z5214 locus and β-glucoside utilization phenotypes of 99 *E.coli* strains

	99 strains[a]	Phenotypes on BTB salicin plates at 37°C[c]			
	human com-52	K-12-37(37%)	more papillae-17(17%)	relaxed-15(15%)	no papillae-30 (30%)
	uro-22	human com-25 (48%)	human com-5 (10%)	human com-6 (11%)	human com-15 (29%)
	sep-25	uro-6 (27%)	uro-6 (27%)	uro-4 (18%)	uro-6 (27%)
		sep-6 (24%)	sep-6 (24%)	sep-5 (20%)	sep-8 (32%)
bgl/Z5211-Z5214 locus[b]	MG1655 *bgl*-48 (48%)	26 (26%)	13 (13%)	-	9 (9%)
	human com-22 (42%)	16 (36%)[d]	3 (6%)[f]		3 (6%)[i]
	uro-12 (54%)	4 (41%)	5 (23%)[g]		3 (14%)[k]
	sep-14 (56%)	6(44%)[e]	5 (20%)[h]		3 (12%)[o]
	CFT073 *bgl*-32 (32%)	11 (15%)	4 (4%)	15 (15%)	2 (2%)
	human com-18 (35%)	9 (21%)[l]	2 (4%)	6 (11%)	1 (2%)
	uro-7 (32%)	2 (14%)	1 (4%)	4 (18%)	-
	sep-7 (28%)	-	1 (4%)	5 (20%)	1 (4%)[m]
	Z5211-Z5214-19 (19%)	-	-	-	19 (19%)
	human com-12 (23%)				12 (23%)[n]
	uro-3 (14%)				3 (14%)
	sep-4 (16%)				4 (16%)

a to **p** as in Table 1a

Table 1c: *bgl*/Z5211-Z5214 locus and β-glucoside utilization phenotypes of 72 ECOR strains

	72 strains[a]	Phenotypes on BTB salicin plates at 37°C[C]				
	human com-29	K-12 -23 (32%)	more papillae-10(14%)	relaxed-11 (15%)	no papilale-27 (37%)	Weak Sal$^+$-1 (1%)
	animal com-32	human com-6 (21%)	human com-6 (21%)	human com-4 (14%)	human com-12 (41%)	human com-1 (3%)
	uro-10	animal com-12 (37%)	animal com-4 (12%)	animal com-5 (16%)	animal com-11 (34%)	animal com-
	ABU-1	uro-4 (40%)	uro -	uro -2 (20%)	uro -4 (40%)	uro -
		ABU-1(100%)	ABU-	ABU-	ABU-	ABU-
bgl/Z5211-Z5214 locus[b]	MG1655 *bgl*-40 (55%)	18 (25%)	9 (13%)	-	13 (18%)	-
	human com-14 (65%)	4 (14%)	5 (17%)[f]		5 (17%)[i]	
	animal com-22 (68%)	11 (34%)	4 (12%)		7 (29%)[j]	
	uro -3 (30%)	2 (20%)	-		1 (10%)	
	ABU-1 (100%)	1 (100%)	-		-	
	CFT073 *bgl*-17 (24%)	5 (7%)	1 (1%)	11(15%)	-	-
	human com-7 (24%)	2 (7%)	1 (3%)	4 (14%)		
	animal com-6 (19%)	1 (3%)	-	5 (16%)		
	uro -4 (40%)	2 (30%)	-	2 (20%)		
	ABU-	-	-	-		
	Z5211-Z5214-15 (21%)	-	-	-	14 (19%)	1 (1%)
	human com-8 (27%)				7 (24%)[n]	1 (3%)
	animal com-4 (12%)				4 (12%)[p]	-
	uro -3 (30%)				3 (30%)	-
	ABU-				-	-

a to **p** as in Table 1a

A second significant correlation was seen for the presence of the CFT073 type *bgl* operon to the relaxed phenotype. All the strains that showed a relaxed phenotype carry CFT073 *bgl* (Fig. 8L, AC, Table 1a, b and c), indicating that CFT073 *bgl* is important for the strains to show relaxed phenotype and not *vice versa*.

1.10 Correlations of the *bgl*/Z5211-Z5214 typing with phylogenetic distribution of ECOR strains.

The 72 ECOR strains are phylogenetically classified into 5 different groups: A, B1, B2, D, and E based on the genetic distance matrix on electrophoretically detected allelic variation at 38 enzyme-encoding loci (Herzer et al., 1990 and references therein; Whittam et al., 1983). It was reported that Group A is a distinct lineage comprising of K-12 and K-12-like strains isolated for the most part from humans. B1 group is predominant with the strains isolated from non-primate mammals, whereas the group B2 strains are mostly from humans and other primates. The group D is a heterogeneous group that consists of strains from humans, non-primate mammals and other primates and group E is a variant group from the other four with the strains isolated from humans and non-primate mammals (Herzer et al., 1990).

Comparison of the *E.coli* types seen at the *bgl*/Z5211-Z5214 locus in the present study to the phylogenetic distribution of ECOR strains revealed significant correlations. 14 out of 15 ECOR strains that belong to group B2 are of CFT073 *bgl*/Z type (Table 2). All the MG1655 *bgl*/Z type strains belong to group A. Likewise; all the strains that belong to group D are of O157 *bgl*/Z type (Table 2). Strains in the fourth *bgl*/Z type are distributed in group A and B1. Fifth *bgl*/Z type strains are distributed in A, B1 and E and mixed *bgl*/Z type strain is distributed in group B2.

Table 2: *bgl*/Z5211-Z5214 type with ECOR phylogeny

	72 strains[a]	Phylogenetic groups[c]				
		A	B1	B2	D	E
bgl/Z5211-Z5214[b]	MG1655 type-11 (15%)[d]	11(15%)	-	-	-	-
	O157 type-15 (21%)	-	-	-	12 (17%)	3 (4%)
	CFT073 type-16 (22%)	1 (1%)	1 (1%)	14 (20%)	-	-
	Fourth type-17 (24%)	7 (10%)	10 (14%)	-	-	-
	Fifth type-12 (16%)	6 (8%)	5 (7%)	-	-	1 (1%)
	Mixed type	-	-	1 (1%)	-	-

a: 72 ECOR strains used in the study
b: types based on the *bgl*/Z5211-Z5214 region analysis
c: phylogenetic groups are based on Herzer et al., 1990
d: number of strains are shown and percentages are calculated to the total number of ECOR strains analyzed in the study.

1.11 Spontaneous activation of the *bgl* operon in natural *E.coli* isolates.

In *E.coli* K-12 95-98% of the mutations that activate *bgl* operon are found in the vicinity of the promoter that includes integration of mobile DNA elements, point mutation within the binding site of the catabolite gene activator protein (CAP) and deletions encompassing an upstream AT-rich sequence element (Reynolds et al., 1981; Reynolds et al., 1986; Schnetz and Rak, 1988). Only 2-5% of the activating mutations are unlinked, involving loci such as *hns*, *gyr*, *bglJ* and *leuO* (Defez and de Felice, 1981; DiNardo et al., 1982; Giel et al., 1996; Ueguchi et al., 1998). To know whether the pattern of activation is identical in the strains that show different papillation phenotypes on BTB salicin plates at 37°C, the spontaneous mutants of a selection of strains that show the MG1655 like phenotype, the more papillae phenotype or relaxed phenotype or those that papillate late (after day 5) were analyzed (described in materials and methods). Expect from the strains that show the relaxed phenotype all the spontaneous mutants were isolated at 37°C. Spontaneous mutants from the strains that show relaxed phenotype were isolated at 28°C.

The seven Bgl^+ mutants of the uropathogenic strain U3454 (MG1655 like phenotype) and two mutants of U3372 (late papillae) show activation of the *bgl* operon by insertions in the upstream AT-rich silencer sequences (Fig. 16). Out of seven spontaneous mutants of the septicemic isolate i484 (relaxed phenotype) one has a deletion of 47 bp, one mutant has a point mutation in CRP binding site and the other five mutants do not carry any sequence change in the *bgl* promoter vicinity (Fig. 16). Similarly, out of 6 spontaneous mutants of a septicemic isolate W7483 (relaxed phenotype), 2 of them have a deletion of 72 bp in the promoter vicinity and the remaining 4 mutants do not have any sequence change. Of two commensal strains E10091 (relaxed phenotype) and E176 (more papillae), all of the 3 Bgl^+ mutants of E10091 and 1 mutant of E176 do not have sequence change in the promoter vicinity. The Bgl^+ mutants that do not have any change in the promoter vicinity may carry mutations elsewhere in the genome which is yet to be determined. The pattern of activation of the *bgl* operon seen in the analyzed strains resemble to that of the activation seen in *E.coli* K-12. However, it varies from strain to strain and may depend on its cellular pleiotropic control of the *bgl* operon or of transposition.

A

	strain name	type	typing at *bgl*/Z5211-Z5214	phenotype	number of mutants analyzed	number of mutants in *cis*
L	U3454	uro	CFT073	MG1655	7	7
L	i484	sep	CFT073	relaxed	7	2
L	W7483	sep	CFT073	relaxed	6	2
L	E10091	com	CFT073	relaxed	3	0
L	E176	com	CFT073	more papillae	1	0
Q	U3372	uro	Fourth	late papillae	2	2

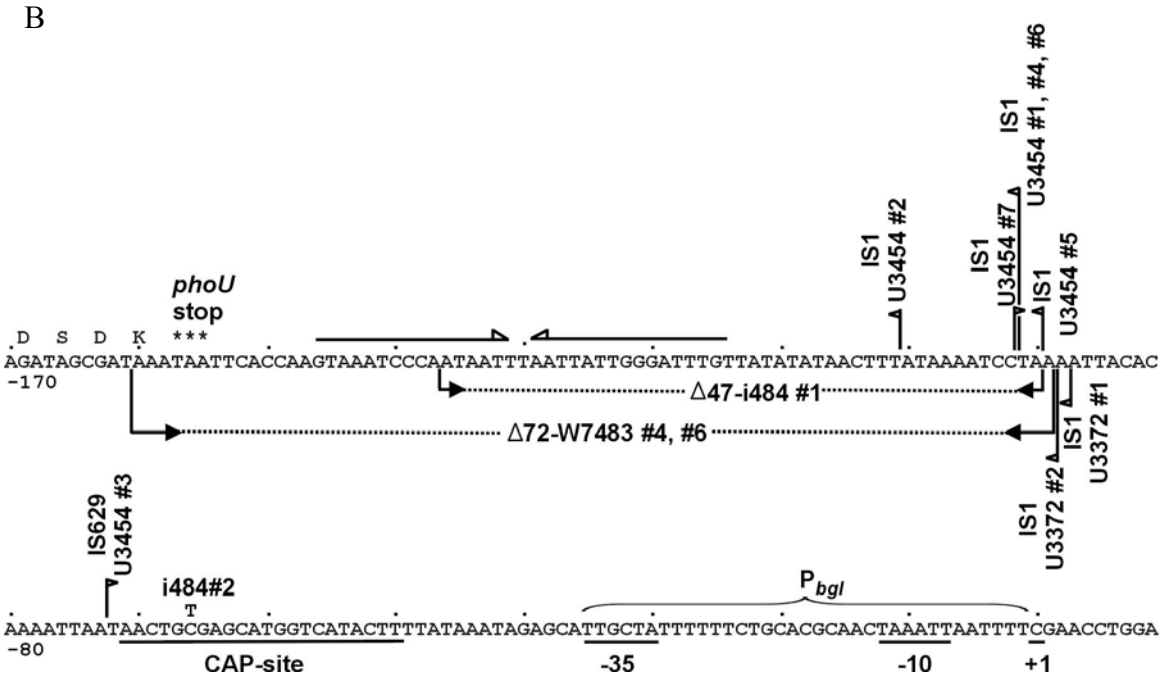


Figure 16: Spontaneous activation of the *bgl* operon. A) **S**train names are shown as listed in Table 6. Uro indicates uropathogenic, sep-septicemic, com-commensal. Letters corresponds to the structures shown in Fig. 8. Typing at the *bgl*/Z5211-Z5214 locus is shown. Phenotypes are on BTB salicin plates at 37°C. No. of mutants analyzed are shown. B) DNA sequence of CFT073 *bgl* promoter region is shown. Sequenced Bgl^+ alleles with insertions of IS1, IS629 and deletions in the promoter region are indicated. Positions of the IS elements are represented by straight lines with the arrowhead ends showing the relative orientations according to their defined left and right ends. The proximal deletion endpoints are indicated by right angled arrows marked with Δ. Numbering is relative to the transcription start of the *bgl* operon. The CAP binding site is underlined and base exchange within the CAP site is given above the sequence. Promoter P_{bgl} is marked by a brace and the transcriptional start site is indicated. Horizontal arrows above the sequence represent the inverted repeats. Two uropathogenic strains (U3454 and U3372) were analyzed for the spontaneous activation of the *bgl* operon. Out of seven spontaneous mutants of U3454, 6 of them carry IS1 insertion and one carry IS629. Two of the U3372 Bgl^+ mutants that arise after day 7 incubation at 37°C carry IS1 insertions. Two septicemic isolates (i484 and W7483) were analyzed. Out of Seven Bgl^+ mutants of i484, 1 had a deletion of 47 bp, and other had a mutation from C to T in CAP binding site, the rest 5 mutants have no changes in the *bgl* promoter region. Deletion of 72 bp is seen in 2 out of 6 W7483 Bgl^+ mutants. Two commensal strains (E10091 and E176) were analyzed. All the 3 mutants of E10091 and 1 mutant of E176 did not have any changes in the *bgl* promoter region. Bgl^+ mutants from relaxed phenotype strains were isolated at 28°C.

1.12 Deduced amino-acid sequence alignment of BglG

Of some strains (7 MG1655 *bgl/*Z type, 8 fourth *bgl/*Z type and 12 CFT073 *bgl/*Z type, Table 6, Appendix) in addition to the *bgl* promoter-leader region, the first gene *bglG* was sequenced and the deduced amino acid sequences of BglG were aligned with the deduced amino acid sequences of MG1655 BglG and CFT073 BglG (Fig. 17). The alignment results showed that the strains of MG1655 *bgl/*Z type do not have sequence variations. Some of the strains of the fourth *bgl/*Z type (*phoU* sequence like O157 and *bgl* sequence like MG1655, Table 6 Appendix) and all of the CFT073 *bgl/*Z type carry few variations in comparison to MG1655 (Fig. 17).

The antiterminator BglG, belongs to SacY/BglG family of related proteins (Stulke et al., 1998). The proteins belonging to this family have an N-terminal RNA-binding domain which is involved in the anti-termination activity (Manival et al., 1997) and two conserved PRD (PTS

regulatory domain) domains which are targets for both positive and negative control by phosphorylation. In each PRD domain there are one to two conserved histidines which are the target residues for phosphorylations (Stulke et al., 1998).

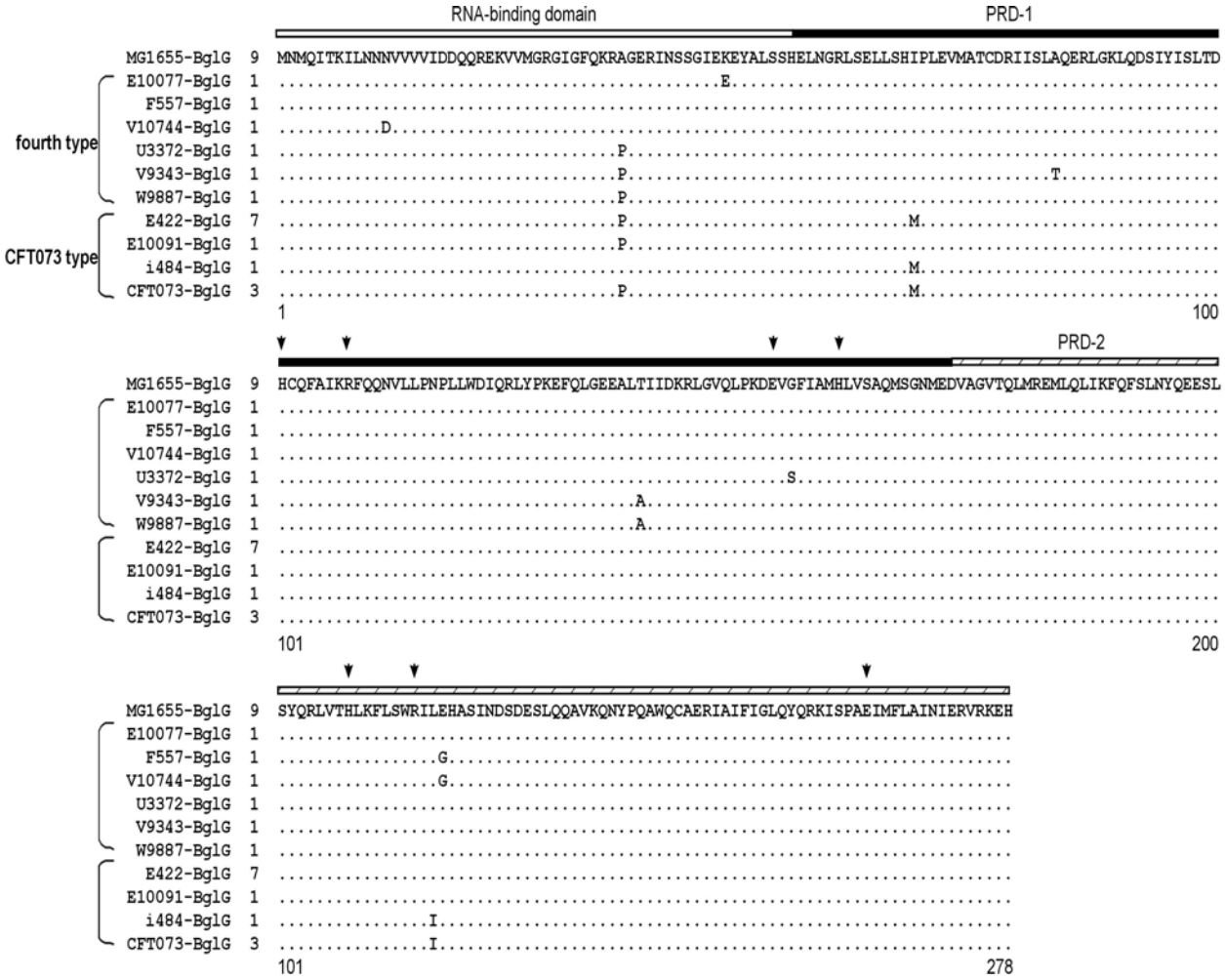


Figure 17: Deduced amino acid sequence alignment of BglG. Deduced amino acid sequences of the 27 strains (7 MG1655 type, 8 fourth types and 12 CFT073 type) were aligned and compared with the deduced amino acid sequences of MG1655 and CFT073 BglG. Amino acid changes in comparison to MG1655 BglG are shown. Conserved amino acids in comparison to MG1655 BglG are represented as dots. Typing of the strains based on the nucleotide sequence alignment is shown at the left. Representative strains in each type along with the strains that have specific amino acid changes are shown. Strains of MG1655 type do not have any change. Two strains of fourth type have no sequence variations are like MG1655; remaining 5 strains have few amino acid changes. CFT073 *bgl* type strains carry 2 to 3 amino acid changes. Inverted arrows represent the conserved amino acid residues in the SacY/BglG family of proteins. Open bar represents RNA binding domain, Closed black bar represents PRD-1 domain and hatched bar represents PRD-2 domain.

In the alignment of the deduced amino acid sequences, strains of the fourth *bgl/Z* type and CFT073 *bgl/Z* type have amino acid changes in the RNA binding domain as well as in the PRD domains. With the exception of septicemic isolate i484, all the CFT073 *bgl/Z* type strains carry A37P. However, this may not affect BglG activity since in the antiterminator ArbG (from the gram negative bacterium *Erwinia chrysanthemi*) also carries a proline residue in place of alanine at this position (el Hassouni et al., 1992). In addition, the structure of SacY (another member of

the SacY/BglG family) RNA binding domain (Manival et al., 1997) indicates that this residue is located in the loop region and thus may not have a role in the function of BglG. Strain E10077 (a fourth *bgl/Z* type strain) carries a K48E change in BglG. This could affect the BglG activity, since Lysine is a conserved residue in the SacY/BglG group of anti-terminators (Declerck et al., 2002). The conserved histidine residues in the PRD-domains of BglG was found to be unaffected in all the strains which were analyzed (Fig. 17).

1.13 Do the sequence variations in the CFT073 *bgl* type strains influence *bgl* expression?

Khan and Isaacson (1998) have reported that the expression of the *bgl* operon occurs in infected mouse liver and suggested a unique role for this operon *in vivo*. The septicemic isolate i484 which was used in their studies was also sequenced at the *bgl* region in our present work. Strain i484 have 12 bp sequence variations in the *bgl* promoter region like CFT073 in comparison to MG1655. However, it has BglG-P37A (number refers to the translational start) in the deduced amino acid sequence which is in contrast to CFT073. To analyze whether the sequence variations seen at the *bgl* regulatory region (includes *bgl* promoter, t1 and the first gene *bglG*) in the CFT073 *bgl/Z* type strains have any influence on the *bgl* expression, plasmidic construct that carry *lacZ* gene fused at the end of *bgl* promoter, terminator t1 and the first gene *bglG* was used (Fig. 18). The P_{bgl}-t1-*bglG* fragment of the strains shown in Figure 18 was isolated by PCR with oligos S145 and S201 and fused to the *lacZ* gene in the plasmid vector pKES15 that contains pACYC origin (refer Table 4 and Table 8 for primer position and plasmid construction, respectively). The expression of these constructs were determined in K-12 Δ*bgl*, Δ*lacZ* background from cultures grown to the exponential phase (OD_{600} of 0.5) in M9 medium with Casamino acids, Vitamin B1, 1% glycerol and antibiotics (see materials and methods).

No significant difference was seen in the β-galactosidase activity in the constructs carrying fourth *bgl/Z* type (*phoU* sequence like O157 and *bgl* sequence like MG1655) or CFT073 *bgl/Z* type sequence in comparison to MG1655. However, constructs that carry specific mutations as seen in the case of U5107 (t1 +103tg) show a ~2 fold increase in the β-galactosidase activity compared to MG1655 (1090 and 550 U, respectively Fig. 18C and A). Strain E10077 (a fourth *bgl/Z* type) that carries a K48E change in BglG (from the deduced amino acid sequence) showed ~ 2 fold lesser activity (275 U, Fig. 18F) than that of MG1655. The Lysine residue in the RNA binding domain of BglG is a conserved residue among the BglG-SacY group of anti-terminators (Declerck et al., 2002). Thus, change from Lysine to Glutamine might have a role in the function of BglG in strain E10077.

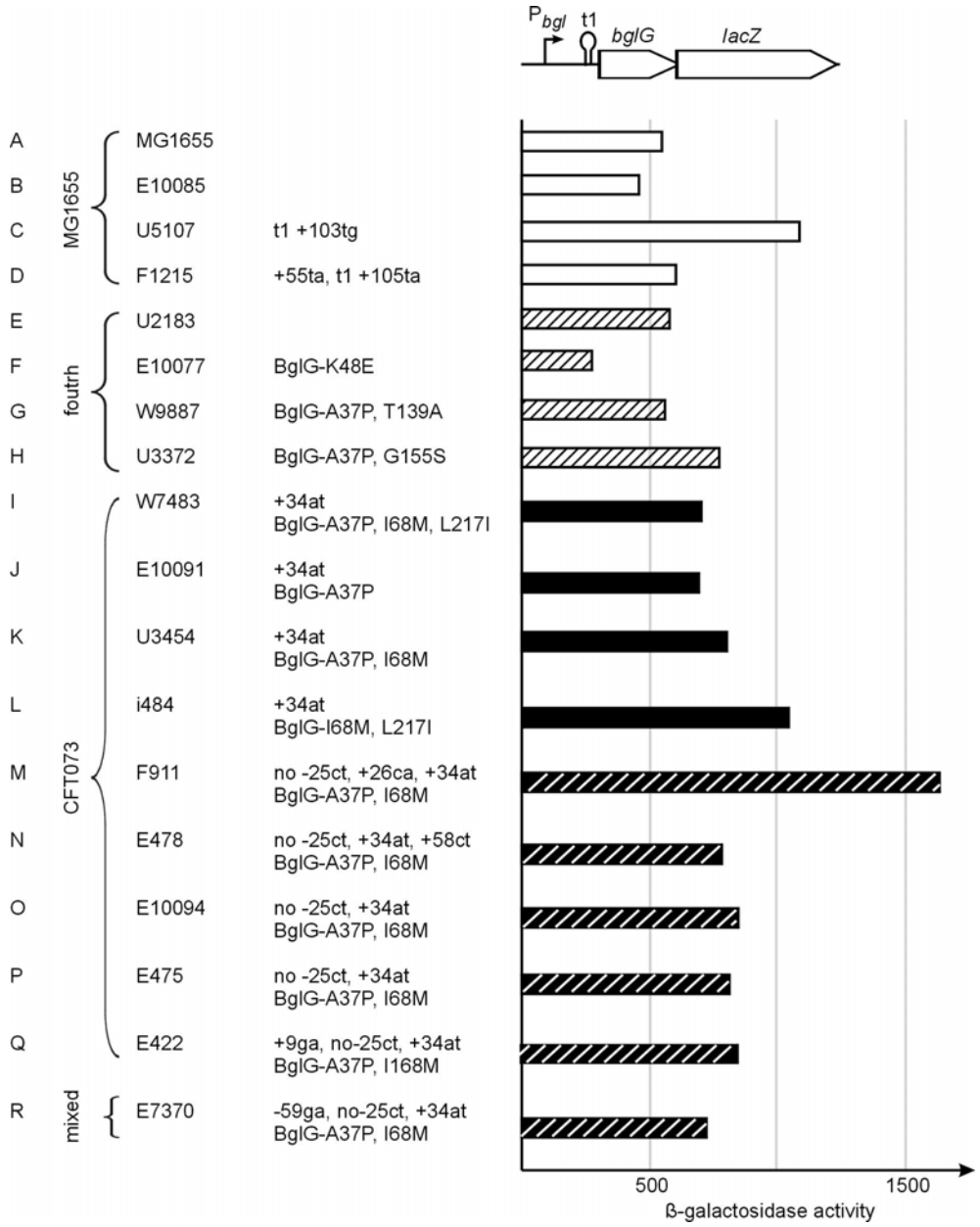


Figure 18: β-galactosidase activity of the P_{bgl} reporter constructs directed in K-12. Shown are the results of plasmidic construct P_{bgl}-t1-*bglG-lacZ* whose expression is directed in K-12. Strain names and specific base pair changes are shown. Nucleotide sequence changes in comparison to MG1655 *bgl* sequence are shown as non-capitals and the numbering indicated is relative to the transcription start of the *bgl* operon. Amino acid changes in comparison to MG1655-BglG are shown as capitals and numbering is relative to the translational start of BglG. Plasmids were transformed into S541 and β-galactosidase activity at OD_{600}=0.5 was determined. β-galactosidase activity is expressed in miller units (Miller, 1972). Shown are the average results from at least 3 independent experiments and from at least two independent tranformants. Standard deviations errors are less than 10%.Bars represent the β-galactosidase activity. The P_{bgl}-t1-*bglG-lacZ* readings for each plasmidic constructs (units) are as follows: A) pKES83 (550U) B) pKEGN1 (460U) C) pKEGN2 (1090U) D) pKEGN3 (505U) E) pKEGN5 (580) F) pKEGN4 (275) G) pKEGN6 (560) H) pKEGN33 (775) I) pKEGN9 (705) J) pKEGN7 (695) K) pKEGN8 (805) L) pKEGN84 (1040) M) pKEGN34 (1640) N) pKEGN35 (785) O) pKEGN36 (850) P) pKEGN37 (815) Q) pKEGN39 (850) R) pKEGN38 (725).

The variants of CFT073 *bgl/Z* type shows ~ 1.5 to 2 fold enhancement in the β-galactosidase activity in comparison to the MG1655 construct (Fig. 18I to R). Strain F911 (CFT073 *bgl/Z* type)

that does not have -25ct change (position relative to transcription start of *bgl* operon) carries an additional base pair change in the leader region (+26ca), shows 3 fold higher activity (1640 U) than MG1655. Taken together the results demonstrate that with the exceptions of the specific sequence variants (Fig. 18B, C, F, G, H, J to R), the sequence variations seen in the fourth *bgl/Z* type and CFT073 *bgl/Z* type at the *bgl* promoter and in the first gene *bglG* do not influence the *bgl* expression in K-12 background. However, at this stage, it cannot be ruled out that the sequence variations seen in the CFT073 *bgl/Z* type may have an impact on the *bgl* expression in an untested background (could be a different strain background or in an environment inside the host).

1.14 Sequence variations in the CFT073 *bgl* type strains do not have significant influence on the *bgl* promoter activity

To analyze whether the sequence variations seen at the vicinity of the *bgl* promoter in the CFT073 *bgl/Z* type strains have any influence on the *bgl* promoter activity, strains (same as in Fig. 18) representing unique sequence variations as compared to MG1655 *bgl* sequence were selected for constructing reporter constructs that carry *bgl* promoter +25bp (numbering relative to transcription start of *bgl* operon) fused to *lacZ* (Fig 19).

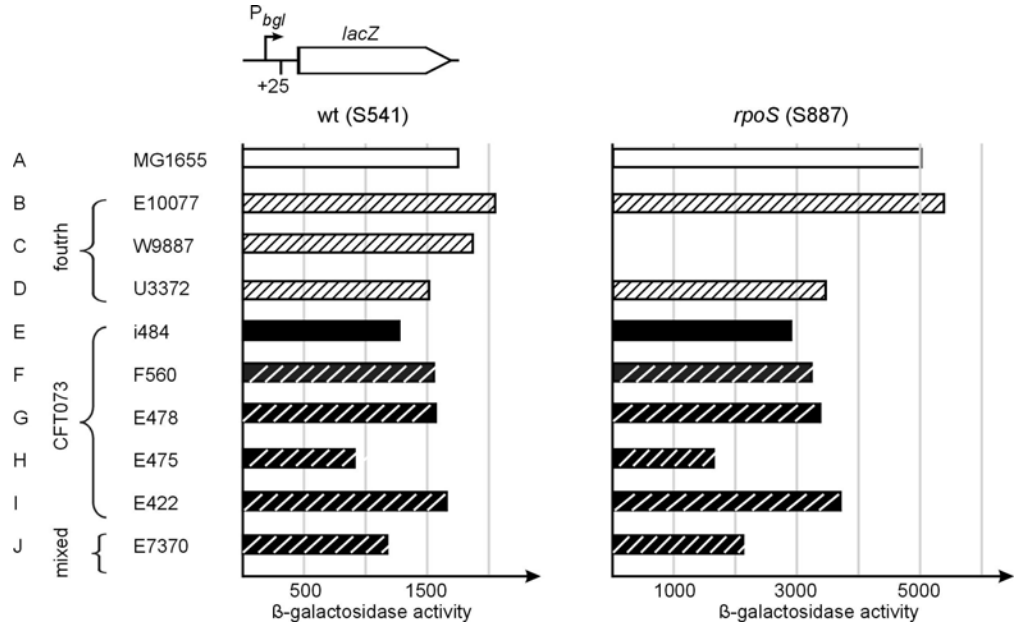


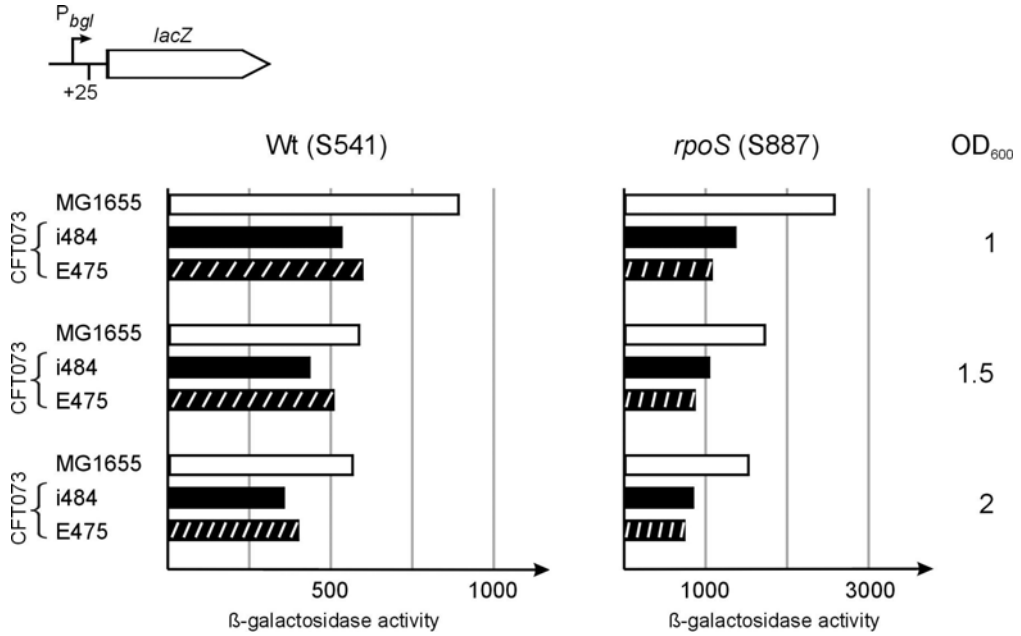
Figure 19: β-galactosidase activity of the P_{bgl} +25-*lacZ* reporter constructs in wt (S541) and in rpoS⁻ background. Shown are the results of plasmidic constructed of P_{bgl}+25-*lacZ* whose expressions are directed in K-12. Strain names and typing at the *bgl*/Z5211-Z5214 locus is shown as in Table 6 (Appendix). Bars represents the β-galactosidase activity expressed in miller units (Miller, 1972). Shown are the average results from at least 3 independent experiments and from at least two independent tranformants. Standard deviation errors are less than 10%. The plasmidic constructs (units in wt, *rpos*) are as follows: A) pKEKB30 (1775, 5080) B) pKEGN10 (2080, 5455) C) pKEGN40 (1895, not determined) D) pKEGN15 (1535, 3505) E) pKEGN17 (1290, 2935) F) pKEGN16 (1565, 3325) G) pKEGN14 (1590, 3415) H) pKEGN12 (920, 1740) I) pKEGN11 (1680, 3750) J) pKEGN13 (1190, 2325).

RpoS is the alternative sigma factor of *E.coli* RNA polymerase and is required for transcription of many genes expressed during the onset of stationary growth phase (Loewen and Hengge-Aronis, 1994). And also, RpoS was reported to be necessary for silencing of the *bgl* operon. Hence, the expression levels of the constructs carrying various sequence variations were analyzed in both wt (S541) and in *rpoS* (S887) background in the exponential growth phase of the cells (Fig. 19).

Constructs carrying fragments from strains i484 and E7370 showed a decreased activity when compared to the MG1655 construct. In addition, construct carrying fragment from E475 showed ~ 2 fold decreased in the promoter activity in comparison to MG1655 construct (920U to 1775U, Fig. 19A and 19H). Strain E475 carries specific sequence variations (no -25ct) in comparison to CFT073 sequence. Thus, the decreased in promoter activity could be due to specific sequence changes seen in E475 construct. The strain E7370 (mixed *bgl/Z* type) have upstream and *bgl* sequence like CFT073 followed by downstream sequence like CFT073 with exception of 5' end of *yieI* gene like MG1655. With the exceptions of the constructs carrying fragments from i484, E475 and E7370, the activity of all the other P_{bgl}+25-*lacZ* constructs was more or less similar to the MG1655 construct.

In order to analyze whether the sequence variations has any role in the RpoS mediated *bgl* regulation, activity of all the P_{bgl}+25-*lacZ* constructs were analyzed in *rpoS* background. The results showed a two fold increase in the *rpoS* background in all the constructs similar to that of MG1655 construct, suggesting that the sequence variations seen in the CFT073 *bgl/Z* type does not influence the RpoS mediated *bgl* regulation.

Furthermore, the decreased in the promoter activity seen in the constructs of i484 and E475 (in comparison to MG1655 construct) at the exponential growth phase of the cells in both wt and *rpoS* initiated further studies in analyzing the expression levels of the constructs following the growth curve (Fig. 20). The growth curve of the cultures was determined measuring the OD_{600} at different time points along with MG1655 construct. Bacterial cells were harvested at the onset and early stationary phase (OD_{600}=1, 1.5, 2) and the expressed LacZ activity was determined (Fig. 20). All three constructs (i484, E475, MG1655) were repressed to two fold at the onset of the stationary phase. The expression levels of the i484 and E475 constructs decreased further while the expression directed by the MG1655 constructs was constant.



.

Figure 20: β-galactosidase activity of the P_{bgl}+25-*lacZ* reporter constructs in wt (S541) and in *rpoS* background following the growth curve of the strains. Shown are the results of plasmidic construct of P_{bgl}+25-*lacZ* whose expression is directed in K-12. Strain names and typing at the *bgl*/Z5211-Z5214 locus is shown. Bars represents the β-galactosidase activity expressed in miller units (Miller, 1972). Shown are the average results from at least 3 independent experiments and from at least two independent tranformants. Standard deviation errors are less than 10%. β-galactosidase assay results of the three constructs repressed two fold at the onset of the stationary phase. The expression of the MG1655 construct decreased further while the expression of i484 and E475 construct was constant. The plasmidic constructs (units in wt, *rpos*) at OD_{600}=1 are as follows: MG1655 construct pKEKB30 (900, 5080), i484 construct pKEGN17 (535, 1715) and E475 construct pKEGN12 (600, 1345). OD_{600}=1.5 readings in wt, *rpos* are as follows: pKEKB30 (590, 2175), pKEGN17 (435, 1305), pKEGN12 (510, 1080). OD_{600}=2 readings in wt and rpos are as follows: pKEKB30 (570, 1920), pKEGN17 (355, 1055), pKEGN12 (400, 920).

1.15 A mutagenesis screen to identify factors that are involved in the relaxed phenotype in *E.coli*

In *E.coli* K-12 the wt-*bgl* operon is silent in the laboratory conditions i.e. it is not expressed in all tested laboratory conditions. However, it was observed that 15% of the natural isolates do exhibit a weak Bgl^+ phenotype (relaxed phenotype) on BTB salicin plates at 37°C (on day 3, Fig. 9). This made us to speculate that in these 15% of the strains there could be additional factors which are important for the strains to show relaxed phenotype. Hence, in order to identify those factors we chose one strain i484 and a transposon mutagenesis screen were carried out using a rep^{ts} plasmid that carries miniTn10-cm^R cassette (pKESK18, see materials and methods). In brief, wt i484 was mutagenised with pKESK18 and the mutants that show a phenotype like MG1655, more papillae or no papillae were screened (Fig. 21A). In total, the screen yielded 16 mutants that met the above criterion. Of these mutants, 5 showed a MG1655 like phenotype, 8 mutants showed more papillae and 3 mutants showed no papillation. In addition to these 16 mutants, one mutant was isolated that showed a strong Bgl^+ phenotype (Fig. 21A). In order to eliminate the

mutants that carry mTn10- cm^R insertions in the *bgl* operon, two long PCR's were performed for all 17 mutants with oligos: a) S157 (maps upstream to *bgl* promoter) and S312 (maps in *bglB*) b) S309 (maps in *bglB*) and S348 (maps in *yieI*). Those mTn10-mutants that yielded larger PCR products (expected sizes with miniTn10-cm^R insertions) were eliminated from further analysis.

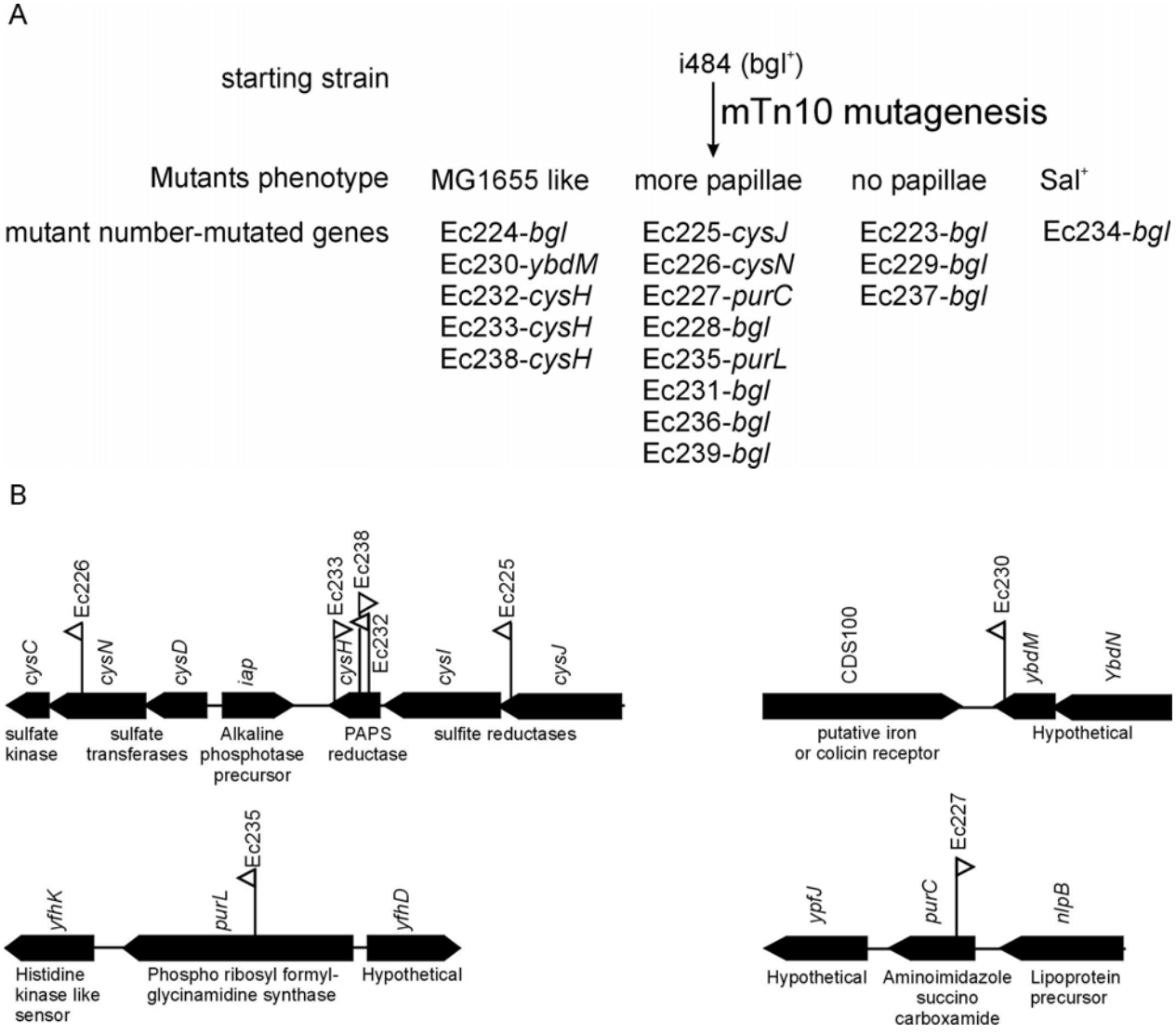


Figure 21: Mutagenesis screen to identify the factors that are essential for relaxed phenotype. A) Schematic representation of the mTn10-*cat* mutagenesis strategy is shown. The starting strain i484 has the natural *bgl* operon. This strain is mutagenised with pKESK18 and mutants were screened on BTB salicin plates at 37°C. Wild type i484 strains show relaxed phenotype on day 3 incubation. In this screen mutants were selected that papillate like MG1655, more papillae or no papillae. Mutant numbers and the mutated genes/system is shown. B) This strategy yielded mutants carrying transposon insertions in *cysN*, *cysH*, *cysJ*, *purL*, *ybdM*, and *purC* genes. The insertions are shown with respect to the published *E.coli* CFT073 chromosome. The arrowhead indicates the direction of *cat* gene. The positions of mTn10 insertions are as follows: Ec225 (AE016765: 141948), Ec226 (AE016765: 135774), Ec227 (AE016764: 146222), Ec230 (AE016762: 215048), Ec232 (AE016765: 139681), Ec233 (AE016765: 139499), Ec235 (AE016764: 235798), Ec238 (AE016765: 139673). The encoded gene products are shown beneath the respective genes.

The mutant that showed the strong Bgl^+ phenotype carried miniTn10-cm^R insertion upstream of *bglF*. The Bgl^+ phenotype in this mutant could be attributed to the constitutive expression of *bgl*F and *bglB* genes. Those mutants that gave PCR products like wt-i484 were further processed with ST-PCR to identify the miniTn10-cm^R insertions (see materials and methods). From the obtained ST-PCR products nucleotide sequencing was determined. Blast searches were carried out with the obtained sequence information. To this end, the screen yielded the following mutants that have insertions in the following genes: *cysN* (mutant #Ec226), *cysH* (#Ec233, Ec238, Ec232), *cysJ* (#Ec225), *purL* (#Ec235), *ybdM* (#Ec230), *purC* (#Ec227) (Fig. 21B). The genes *cysN*, *cysH* and *cysJ* are involved in the sulfate assimilation pathway. This pathway is involved in the conversion of sulfate to sulfide, in preparation for incorporation into cysteine and methionine. Mutations in *cysN*, *cysH* and *cysJ* may lead to the alterations in the sulfate assimilation which may in-turn affect the amino acid synthesis. On the other-hand the gene *cysH* is involved in conversion of PAPS (phosphoadenosine phosphosulfate) to adenosine 3', 5'-phosphate which in-turn is required for the synthesis of phosphate that can be ultilized for the synthesis of phosphoenol pyruvate. Thus, mutation in *cysH* may have direct or indirect effects on the PTS system, which may consecutively, affect the transport of β-glucosides. However, this hypothesis remains to be thoroughly explored. The genes *purL* and *purC* are involved in the *de novo* biosynthesis of purine nucleotides. The mutations in these genes could be attributed for the alterations in the purine nucleotide pool which may affect the cellular physiology of the strain.

The gene product of *ybdM* is not known at present. However, homology searches with the deduced amino acid sequence showed significant identities that matches with the complete coding sequences of the following proteins: Z1644 (*E.coli* O157-ParB like nuclease, AAQ19140)-91%, IbrB (*E.coli* ECOR9-DNA binding, arginine biosynthesis, AF460182)-88%, *YbdM* (*S. typhimurium*-putative transcriptional regulator, AE008724)-60%, BF2122 (*B. fragilis*-putative transcriptional regulator, AP006841)-46%, EF0120 (*E. faecalis*-ParB like nuclease, AF454824)-45%. Most of the matches seen were with the hypothetical proteins from other bacterial species. To this end, the effect of the mutation in *ybdM* gene on β-glucoside utilization is not understood.

1.16 A mutagenesis screen in the mixed Sal^+ mutants isolated from strains that show relaxed phenotype at 37°C.

In order to know whether the relaxed phenotype is caused by the mutations or is due to the leaky expression of the *bgl* operon, strains i484 and E7370 (both show relaxed phenotype at 37°C) were selected for the analysis. Wild type strains of i484 and E7370 were streaked on BTB

salicin plates and incubated at 37°C. After 3 days of incubation both the strains showed relaxed phenotype. The Sal^+ part of the streak was purified to isolate independent mutants. However, the subsequent purification steps resulted in a mixture of colonies, where some of the colonies showed weak Sal^+ phenotype and some colonies showed Sal^- phenotypes. Re-streaking of the Sal^+ or Sal^- colonies also resulted in the colonies with mixture of phenotypes. Thus, in the current study we were unable to purify complete Sal^+ mutants from the strains that show relaxed phenotype at 37°C. Two of such mutants from each strain that show mixed phenotype were selected for further analysis (Fig. 22A).

A miniTn10-cm^R mutagenesis strategy is shown in Figure 22A. The two mixed mutants of i484 (Ec240 and Ec241) and two mixed mutants of strain E7370 (Ec254 and Ec255) were mutagenised with plasmid pKESK18 and the mutants that show $Sal^–$ were screened. The screen yielded 3 mutants from Ec240 (i484 mixed Sal^+ #1), 9 mutants from Ec241 (i484 mixed Sal^+ #2), 5 mutants from Ec254 (E7370 mixed Sal^+ #1) and 8 mutants from Ec255 (E7370 mixed Sal^+ #2) (Fig. 22A). The screen yielded mutations in *bgl* operon, *pgi*, *cysG*, *serC*, *ybdM*, *carB* and *rfaC*. Mutations in the *bgl* genes hinder the β glucoside utilization which explains the Sal^- phenotype in those mutants. The *pgi* gene codes for the enzyme phosphoglucoisomerase of the glycolysis pathway. It was shown that in a *pgi* mutant glycolysis is blocked and this accelerates RNaseE mediated degradation of the *ptsG* gene transcript. The *ptsG* gene codes for a transport protein for Glucose uptake and is a part of the PTS transport system (Kimata et al., 2001). In addition, the work from Dole et al., (2004) has reported that mutation in *pgi* results in a decreased expression of the *bgl* operon. The gene *cysG* codes for seroheme synthase that act as a co-factor for sulfite reductases (gene product from *cysJ*, see Fig. 21B). Mutations in *serC* may have an effect on serine biosynthesis. The gene c*arB* codes for carbamoylphosphate synthase that is involved in the *de novo* synthesis of pyrimidine ribonucleotides and arginine biosynthesis. The gene *rfaC* codes for product that is involved in lipopolysaccharide core biosynthesis. Lipopolysaccharide is a major component of outer membrane of gram-negative bacteria, such as *E.coli*. Mutations in *rfaC* may have pleiotropic effect within the cell (Chen and Coleman, 1993; Bauer and Welch, 1997). In addition to the above mentioned mutations, the screen also yielded a mutant that carries miniTn10-cm^R insertion in *ybdM* gene (which was obtained from the earlier screen as well, Fig. 22B). Taken together, the results from the miniTn10-cm^R mutagenesis of the wt-i484 and the screen performed with the two mixed mutants of relaxed phenotype strains primarily picked up mutations in the genes that are required for nucleotide and/or amino acid biosynthesis (Fig. 21 and 22). The effect of the mutation in these genes may lead to the alterations in the nucleotide

pool and/or amino acid pool of the bacterial cell which in-turn may have pleiotropic effect on the expression of various genes (that may include regulators of *bgl* operon silencing such as H-NS). Thus, the alterations in the nucleotide or amino acid pool may have direct or indirect effect on H-NS activity which in turn may affect the β-glucoside utilization. The other possibility is that it could be a direct effect on the repression of *bgl* by H-NS.

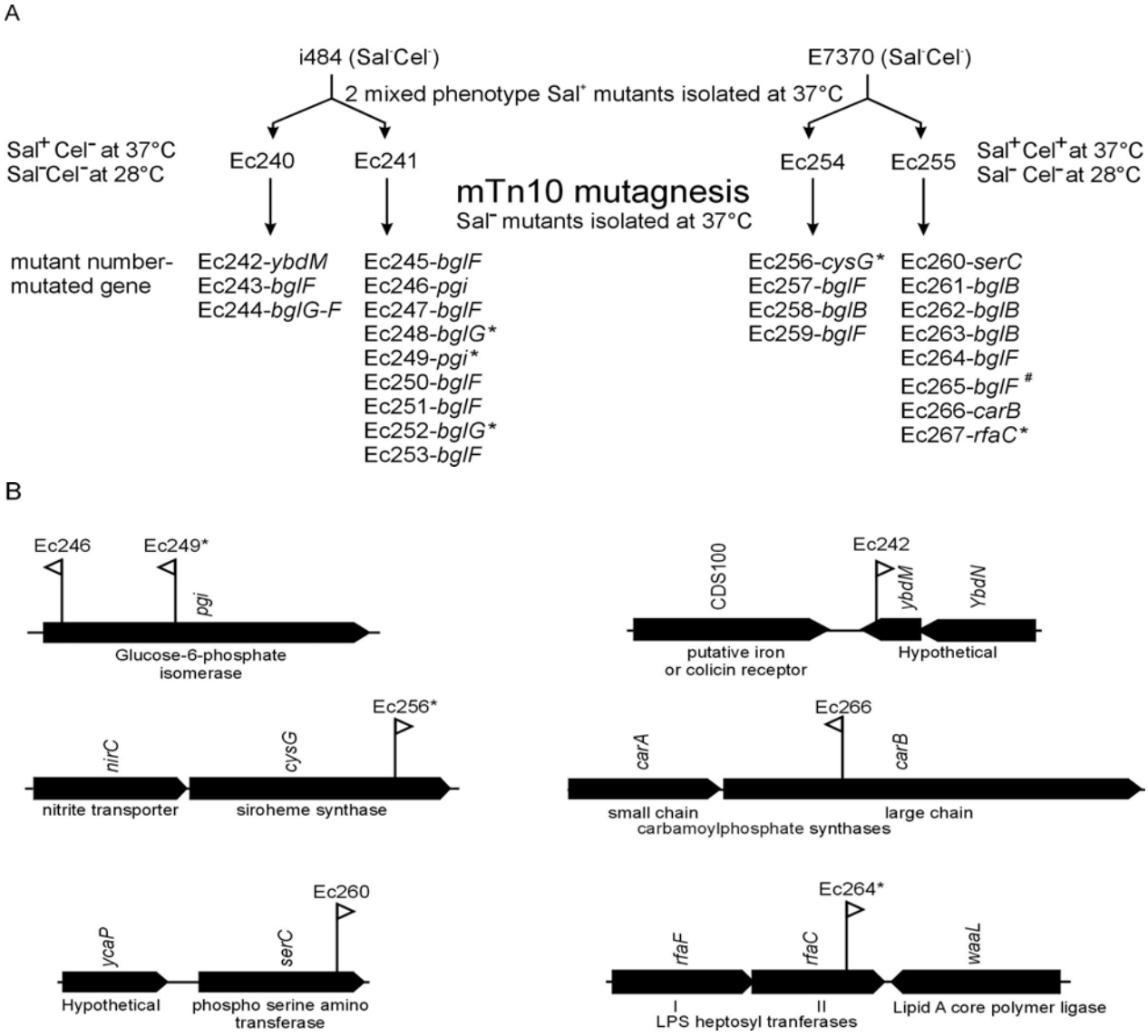


Figure 22: Mutagenesis screen to identify the factors that are essential for relaxed phenotype. A) Schematic representation of the mTn10-*cat* mutagenesis strategy is shown. The starting strain i484 and E7370 has the natural *bgl* operon. Both strains were streaked on BTB salicin plates at 37°C and mixed Sal^+ mutants were isolated. Two such mutants from each strain were mutagenised with pKESK18 and Sal^- mutants were screened on BTB salicin plates at 37°C. B) This strategy yielded mutants carrying transposon insertions in *pgi, ybdM, cysG, serC, carB, rfaC* and in *bgl* operon. The insertions are shown with respect to the published *E.coli* CFT073 chromosome. The arrowhead indicates the direction of *cat* gene. The positions of mTn10 insertion with respect to the Genbank primary accession numbers are as follows: Ec246 (AE016770: 235679), Ec249 (AE016770: 236173), Ec242 (AE016762: 215064), Ec256 (AE016767: 299877), Ec260 (AE016758: 108192), Ec266 (AE016755: 35300), Ec 264 (AE016768: 298893). The encoded gene products are shown beneath the respective genes. * indicates papillation seen after day 7. # indicates Sal^+

2. Identification and analysis of an additional β-glucoside system in *E.coli*

(This section, in part, is in preparation for a publication)

2.1 Strain i484 Δ*bgl* and O157 type (at *bgl*/*Z*5211-*Z*5214 locus) strains papillates on BTB salicin plates

In the *bgl*/*Z*5211-*Z*5214 analysis it was observed that out of 33 O157 *bgl/Z* type strains, 9 strains papillate late on BTB salicin plates at 37°C (Table 6, Appendix). This suggested an existence of an additional system required for the β-glucoside utilization which can be activated.

In addition, the spontaneous Bgl^+ mutant of i484 (Ec2, Δ47 Bgl^+) that was isolated at 28°C was transformed with plasmid pFMAC11 (rep^{ts}-Tet^R, Δ*bgl* in *bgl* region) and the *bgl* operon was deleted as described in materials and methods (also see Fig. 23A). The resultant strain i484 Δ*bgl* (Ec93) was tested for its phenotype on BTB Salicin plates at 37°C and at 28°C. At both the temperatures i484 Δ*bgl* showed Sal^- phenotype. Phenotypes were also determined for the utilization of three other β-glucosides (Arbutin, Cellobiose and Esculin). The results revealed that i484 Δ*bgl* showed Arb^- Cel^- and Esc^- phenotypes at both 37°C and at 28°C. Interestingly, it was observed that after 5 days of incubation at 28°C, i484 Δ*bgl* showed papillation on BTB salicin plates. This observation supported the earlier speculation from the 9 O157 *bgl/Z* type strains that in *E.coli* additional β glucoside system (s) may exist and can be activated. Hence, to identify the system that is responsible for the papillation phenotype in the absence of *bgl* genes the strain i484 Δ*bgl* was selected for further studies.

Four of the Sal^+ spontaneous mutants from i484 Δ*bgl* were purified at 28°C as mentioned in materials and methods (also see Fig. 23A). All the four mutants (Ec131, Ec132, Ec133, and Ec134) showed Sal^+ Arb^+ Cel^+ and Esc^+ phenotypes on respective indicators plates at 28°C and were Sal^- Arb^- Cel^- and Esc^- at 37°C (Fig. 23A). Phenotypes on Salicin and Esculin plates at 28°C were stronger compared to the other two sugars.

2.2 A miniTn10-cm^R mutagenesis screen to identify the additional β-glucoside system

In an approach to identify the system that is involved in the utilization of the β-glucosides, a miniTn10-cm^R mutagenesis was performed. The four Sal^+ mutants isolated from i484 Δ*bgl* at 28°C (Ec131, Ec132, Ec133, Ec134) were mutagenised with plasmid pKESK18 (as described in the previous sections, Fig. 21 and 22) and the mutants that show Sal^- phenotypes at 28°C and 37°C were screened (Fig. 23).

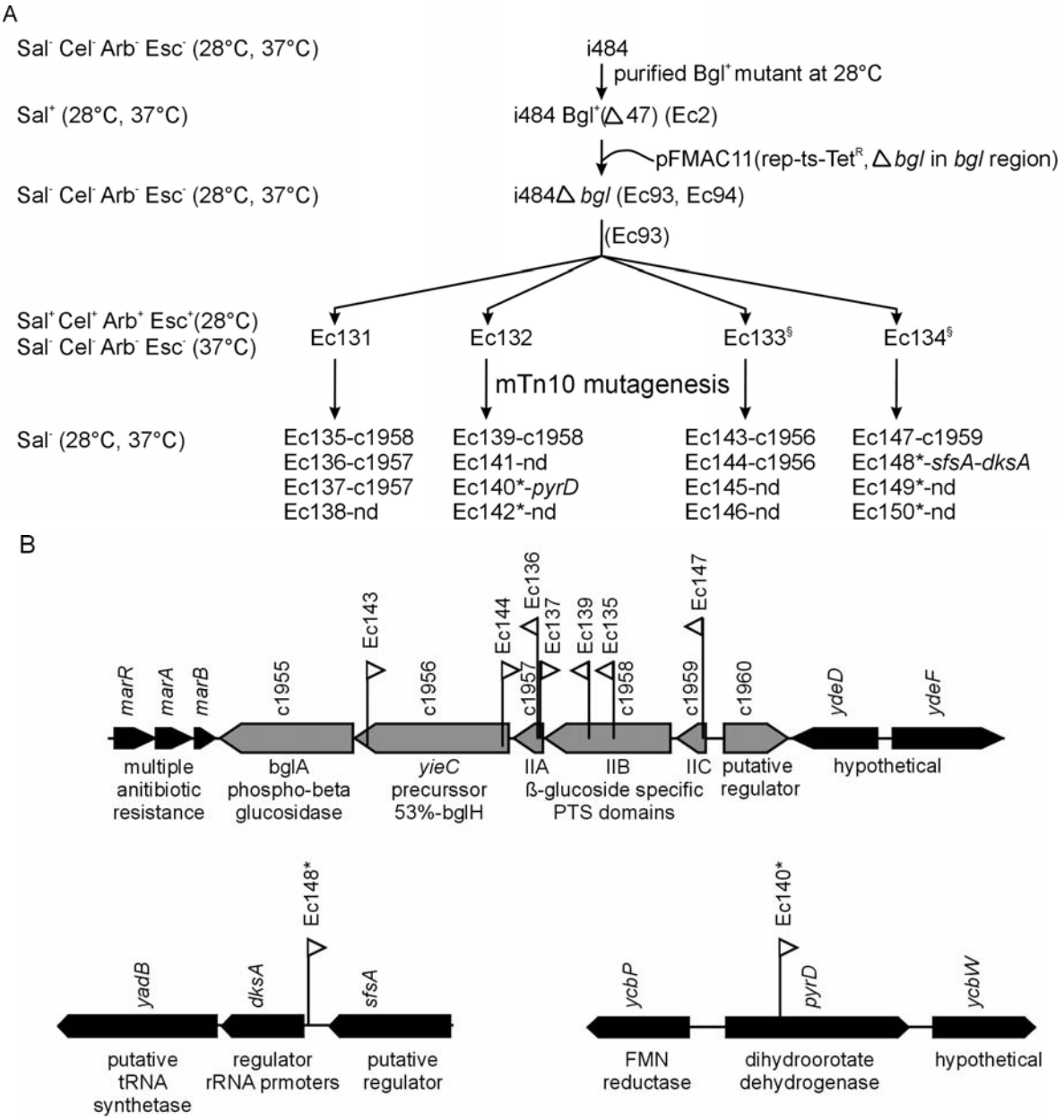


Figure 23: Summary of the identification of additional β-glucoside system in *E.coli*. A) Summary of the identification of additional β glucoside system. In brief, spontaneous Bgl^+ mutant of wt-i484 was isolated at 28°C and *bgl* operon was deleted using plasmid pFMAC11. The resulting strain Ec93 (i484 Δ*bgl*) papillae's at 28°C. Four such spontaneous mutants were purified and mutagenised using plasmid pKESK18. The mTn10-insertion mutants that showed Sal^- phenotype at 28°C were screened. Shown are the respective strain numbers used in the strategy. The obtained mTn10-insertion mutants were later analyzed by ST-PCR and the obtained PCR products were sequenced from both ends. § indicates that strain papillates after day 10 at 37°C. * indicates weak Sal^+ phenotype at 28°C. Phenotypes at 28°C and at 37°C are also shown. Mutants-mutated genes are indicated. nd indicates not determined. B) Based on the homology searches from the obtained nucleotide sequence information, the mTn10-insertions were mapped. This strategy yielded mutants carrying transposon insertions in c1956-c1960 locus, *dksA* region and in *pyrD* region. The insertions are shown with respect to the published *E.coli* CFT073 chromosome. The arrowhead indicates the direction of *cat* gene. The positions of mTn10 insertion with 9bp target site duplications are relative to the Genbank primary accession numbers: Ec135 (AE016760: 291472-291480), Ec136 and Ec137 (AE016760: 291646-291654), Ec139 (AE016760: 291205-291213), Ec140 (AE016758: 147545-147553), Ec143 (AE016760: 288797-288805), Ec144 (AE016760: 290312-290320), Ec147 (AE016760: 292417-292425), Ec148 (AE016755: 171175-171183). The mutants Ec138, Ec141, Ec142, Ec145, Ec146, Ec149, and Ec150 were not analyzed further. The encoded gene products are shown beneath the respective genes. * indicates as in A.

In total the screen yielded 16 miniTn10-cm^R mutants with four mutants from each starting strain (Fig. 23A). The miniTn10-cm^R insertion mutants Ec135 to Ec139, Ec141, and Ec143 to Ec147 showed Sal$^-$ phenotype at both 28°C and 37°C. However, mutants Ec140, Ec142, Ec148, Ec149 and Ec150 showed weak Sal$^+$ phenotype at 28°C. ST-PCR and nucleotide sequencing were carried out to map the insertion site in the chromosome (materials and methods).

Out of nine mutants that were analyzed by nucleotide sequencing, 7 mutants carried miniTn10-cm^R insertions in the c1955-c1960 region of the sequenced strain CFT073 (Fig. 23). The c1955-c1960 region comprises of six putative ORF's: c1955 to c1960. Five ORF's c1955 to c1959 reads in the same orientation, whereas c1960 reads in the opposite orientation. All the 7 mutants that carry miniTn10-cm^R insertions in c1955-c1960 showed Sal$^-$ phenotype at both 28°C and 37°C, indicating a possible second locus in *E.coli* for β-glucoside utilization. Furthermore, comparative genomic analysis of the c1955-c1960 region with the other three sequenced strains indicated that this locus is uniquely present in CFT073 and is absent in MG1655 and two variants of O157 (EDL933 and Sakai). Thus, this locus comprises of putative genes that are encoded in a genomic island.

In addition to the mutants with insertions in c1955-c1960 region, the screen also yielded two mutants that carried miniTn10-cm^R in two other loci. Both the mutants showed weak Sal$^+$ phenotype at 28°C (Fig. 23). Mutant Ec148 carried miniTn10-cm^R insertion close to the gene *dksA* (Fig. 23). DksA was originally identified in *E. coli* as a multi-copy suppressor of the temperature sensitivity of *dnaKJ* mutant (Kang and Craig, 1990). Deletion and/or over-expression of *dksA* have been shown to have pleiotropic effects that includes defects in the chaperonin function, gene expression, cell division, amino acid biosynthesis, quorum sensing, and virulence (Kang and Craig 1990; Bass et al., 1996; Turner et al., 1998; Webb et al., 1999; Ishii et al., 2000; Branny et al., 2001; Hirsch and Elliott, 2002 and Brown et al., 2002). In addition, recent report has shown that DksA is absolutely required for rRNA regulation (Paul, et al., 2004).

Mutant Ec140 carried miniTn10-cm^R insertion in the gene *pyrD* (Fig. 23). The gene pyrD encodes for dihydroorotate dehydrogenase enzyme that is involved in pyrimidine ribonuclotide biosynthesis. Mutations in *dskA* and *pyrD* might have pleiotropic effects in the cell which consecutively may affect the β-glucoside utilization. However, both these mutants were not analyzed in more detail in this study.

2.3 Homology searches for the deduced amino acid sequences of c1955-c1960 genes

The c1955-c1960 region carries six open reading frames and is thought to encode putative proteins (Welch et al., 2002). Based upon analysis of the deduced amino acid sequences, five of these proteins have similarity with proteins associated with the utilization of β-glucosides as carbon sources in *E.coli* and other bacteria (Fig. 24).

c1955: The last ORF spans 1446 nucleotides and potentially encodes a deduced protein of 482 amino acids. The deduced protein from this ORF displayed homology with the complete coding sequences of 6-phospho-β-glucosidases (BglA) from *Enterobacteriaceae* members such as *E.coli, Yersinia, Shigella, Salmonella, Erwinia* and *Klebsiella*. In addition, C1955 also showed significant similarity to 6-phospho-β-glucosidases from Gram-positive bacteria (Fig. 24). With the high degree of similarity to the phospho-β-glucosidases it suggests that c1955 may encode a putative phospho-β-glucosidase enzyme for the utilization of β-glucosides.

c1956: The fifth ORF spans 1674 nucleotides and could encode a protein of 558 amino acids. The deduced protein from this ORF displayed similarity to the complete coding sequences of YieC precursors (BglH) from *E.coli* (68%)*, Shigella* (67%) (Fig. 24). BglH is a product of the *bgl* operon and encodes for putative membrane protein (porin) specific for β-glucosides (Andersen et al., 1999). Sequence similarities were also seen with LamB proteins from *Aeromonas, Yersinia, Erwinia, Shigella* and *Klebsiella*. LamB proteins are known to act as substrate-specific porins which is involved in the guided diffusion of maltose and maltodextrins into the *E.coli* cells (Boos and Shuman, 1998).

c1957: The fourth ORF spans 315 nucleotides and could encode a protein of 105 amino acids. The deduced protein from this ORF displayed high percentage similarities with the complete coding sequences of β-glucoside specific PTS IIA domains from *Erwinia* and *Photorhabdus* (Fig. 24). It also showed significant similarities with *Listeria, E.coli, Shigella, Salmonella, Bacillus* β-glucoside specific PTS IIA domains. Therefore, this ORF may possibly encode the enzyme IIA domain of PTS system for transport of sugars like β-glucosides.

c1958: The third ORF spans 1362 nucleotides and could encode a protein of 454 amino acids. The deduced protein from this ORF displayed similarity to several GenBank entries that showed homology with the complete coding sequences of enzyme IIC domain of PTS system specific for β-glucoside transport. *L.monocytogenes* CelB, *B.lichenformis* Blo1162 and *L.plantaurum* PtsBC showed a very high similarity followed by PTS IIC domains from *Enterobacteriaceae* members (Fig. 24). Thus, this ORF may encode the enzyme IIC domain of PTS system and may be involved in the transport of β-glucosides.

c1959: The second ORF spans 312 nucleotides and could encode a protein of 104 amino acids. The deduced protein shows greater similarities to the complete coding sequences of enzyme IIB domains of *L.monocytogenes B.lichenformis* followed by *Streptococcus*, *Pseudomonas* and *Enterococcus*. *E.coli* ChbB proteins show 41% identities and 56% similarities to C1959 (Fig. 24). Therefore, c1959 may encode a putative PTS IIC domain for the transport of β-glucosides in the c1955-c1960 system in *E.coli*.

c1960: The first ORF spans 699 nucleotides and would encode deduced protein of 233 amino acids. This ORF reads in the opposite orientation in comparison to the other five ORF's. The deduced protein from this ORF displayed similarity to several GenBank entries that show homology to Bacillus GntR family of transcriptional regulators. Highest similarity was seen with *Erwinia* ECA1849 (putative transcriptional regulator, GntR family), *E.coli* and *Shigella* YidP (putative transcriptional regulator) (Fig. 25). Similarities were also seen with DNA binding protein IolR from *Bacillus*, Fatty acyl transcriptional regulator FuxR from *E.coli*, GntR family proteins from *Bacillus*, *Erwinia*, *Listeria*, *Lactobacillus* and *Corynebacterium*. In addition, weak similarities were seen with hypothetical proteins YegW from CFT073 and YydK from *Bacillus subtilis*. From the blast searches it suggested that c1960 could possibly act as a regulator in the c1955-c1960 system. The role of C1960 in the c1955-c1960 system is discussed later in the next sections.

Protein	Function	Identity	Similarity	Reference
		complete coding sequences		
	(a) similarity with the deduced C1955 protein			
E.coli CFT073 B*glA*	6-P-ß-glucosidase	60	74	Welch et al., 2002
E.coli MG1655 B*glA*	6-P-ß-glucosidase	60	74	Blattner et al., 1997
E.coli Sakai B*glA*	6-P-ß-glucosidase	60	74	Hayashi et al., 2001
E.coli EDL933 BglA	6-P-ß-glucosidase	59	73	Perna et al., 2001
Yersinia pestis BglA	6-P-ß-glucosidase	60	75	Deng et al., 2002
Salmonella typhimurium BglA	6-P-ß-glucosidase	59	73	McClelland et al., 2001
Shigella flexneri BglA	6-P-ß-glucosidase	59	73	Wei et al., 2003
Bacillus subtilis BglA	6-P-ß-glucosidase	61	74	Zhang & Aronson, 1994
B.subtilis BglH	6-P-ß-glucosidase	55	68	Le Coq et al., 1995
Clostridium longisporum AbgA	6-P-ß-glucosidase	52	65	Genbank AAC05714
E.coli MG1655 AscB	6-P-ß-glucosidase	48	64	Hall & Zu, 1992
B.subtilis YckE	probable ß-glucosidase	39	51	Fujishima & Yamane,1995
MG1655 BglB	6-P-ß-glucosidase	50	64	Schnetz et al., 1987
CFT073 BglB	6-P-ß-glucosidase	49	63	Welch et al., 2002
S.flexneri BglB	6-P-ß-glucosidase	48	61	Wei et al., 2003
Erwinia chrysanthemi ArbB	6-P-ß-glucosidase	50	55	el Hassouni et al., 1992
Klebsiella oxytoca CasB	6-P-ß-glucosidase	51	54	Lai et al., 1997

Protein	Function	Identity complete coding sequences	Similarity	Reference
	(b) similarity with the deduced C1956 protein			
MG1655 BglH	Porin for ß-glucosides	52	68	Blattner et al., 1997
CFT073 BglH	Porin for ß-glucosides	51	67	Welch et al., 2002
S.flexneri BglH	Porin for ß-glucosides	51	67	Wei et al., 2003
Aeromonas Hydrophila LamB	Maltoporin	31	49	GenBank CAD43291
A.salmonicida LamB	Maltoporin	30	47	Dordsworth et al., 1993
Y.pestis LamB	Maltoporin	25	44	Deng et al., 2002
MG1655 LamB	Maltoporin	25	44	Blattner et al., 1997
V.cholerae LamB	Maltoporin	27	45	Heidelberg et al., 2002
S.typhimurium LamB	Maltoporin	25	44	McClelland et al., 2001
S.typhi LamB	Maltoporin	24	41	Deng et al., 2003
Klebsiella pneumoniae LamB	Maltoporin	25	43	Werts et al., 1992
	(c) similarity with the deduced C1957 protein			
Erwinia carotovora	ß-glucoside specific PTS IIA	59	78	GenBank CAG76546
Photorhabdus luminescens CelC	ß-glucoside specific PTS IIA	53	71	Duchaud et al., 2003
Y.pestis CelC	ß-glucoside specific PTS IIA	51	70	Deng et al., 2002
Listeria monocytogenes	ß-glucoside specific PTS IIA	49	72	Nelson et al., 2004
S.typhimurium CelC	ß-glucoside specific PTS IIA	48	71	McClelland et al., 2001
S.typhi CelC	ß-glucoside specific PTS IIA	48	71	Deng et al., 2003
S.flexneri	ß-glucoside specific PTS IIA	48	69	Wei et al., 2003
MG1655 ChbA	ß-glucoside specific PTS IIA	48	69	Blattner et al., 1997
CFT073 CelC	ß-glucoside specific PTS IIA	48	69	Welch et al., 2002
EDL933 CelC	ß-glucoside specific PTS IIA	48	69	Perna et al., 2001
Sakai CelC	ß-glucoside specific PTS IIA	48	69	Hayashi et al., 2001
B.subtilis CelC	ß-glucoside specific PTS IIA	40	62	Tobisch, 1997
Streptococcus pyogenes	ß-glucoside specific PTS IIA	39	65	Banks et al., 2004
Streptococcus mutans LacF	lactose specific PTS IIA	37	55	Rosey & Stewart, 1992
Lactobacillus casei LacE	lactose specific PTS IIA	33	35	Alpert & Chassy, 1988
	(d) similarity with the deduced C1958 protein			
L.monocytogenes CelB	ß-glucoside specific PTS IIC	69	85	Nelson et al., 2004
B.licheniformis Blo1162	ß-glucoside specific PTS IIC	64	82	Rey et al., 2004
Lactobacillus plantarum Pts8C	ß-glucoside specific PTS IIC	55	74	Kleerebezem et al., 2004
MG1655 ChbC	ß-glucoside specific PTS IIC	36	56	Blattner et al.,1997
CFT073 CelB	ß-glucoside specific PTS IIC	36	55	Welch et al., 2002
Sakai Ecs2443	ß-glucoside specific PTS IIC	36	56	Hayashi et al., 2001
EDL933 CelB	ß-glucoside specific PTS IIC	36	55	Perna et al., 2001
S.typhimurium CelB	ß-glucoside specific PTS IIC	36	56	McClelland et al., 2001
S.flexneri CelB	ß-glucoside specific PTS IIC	36	55	Jin et al., 2002
S.typhi CelB	ß-glucoside specific PTS IIC	36	56	Deng et al., 2003
Enterococcus faecalis Ef1019	ß-glucoside specific PTS IIC	36	54	Paulsen et al., 2003
Y.pestis CelB	ß-glucoside specific PTS IIC	35	55	Deng et al., 2002
S.pyogenes PTSIIC	putative PTS IIC	34	54	Beres et al., 2002
Serratia marcescens ChbC	ß-glucoside specific PTS IIC	33	54	GenBank BAB92991
	(e) similarity with the deduced C1959 protein			
L.monocytogenes CelA	ß-glucoside specific PTS IIB	58	81	Nelson et al., 2004
B.licheniformis Bli02506	ß-glucoside specific PTS IIB	60	78	Rey et al., 2004
S.pneumoniae PtcB	ß-glucoside specific PTS IIB	50	70	Tettelin et al., 2001
P. luminescens CelA	ß-glucoside specific PTS IIB	50	66	Duchaud et al., 2003
E.faecalis Ef1769	ß-glucoside specific PTS IIB	49	66	Paulsen et al., 2003
B.halodurans BH3921	ß-glucoside specific PTS IIB	45	64	Takami et al., 2000
C.acetobutylicum licB	ß-glucoside specific PTS IIB	44	62	Nolling et al., 2001
S.typhi CelA	ß-glucoside specific PTS IIB	42	57	Deng et al., 2003
CFT073 CelA	ß-glucoside specific PTS IIB	41	56	Welch et al., 2002
MG1655 CelA	ß-glucoside specific PTS IIB	41	56	Blattner et al., 1997
EDL933 CelA	ß-glucoside specific PTS IIB	41	56	Perna et al., 2001
Sakai CelA	ß-glucoside specific PTS IIB	41	56	Hayashi et al., 2001
Y.pestis CelA	ß-glucoside specific PTS IIB	41	59	Deng et al., 2002
S.flexneri CelA	ß-glucoside specific PTS IIB	40	55	Jin et al., 2002

Figure 24: Identities and similarities of deduced proteins from the c1955-c1959 region of *E.coli* CFT073. Identities and similarities are shown are percentages. The homology searches using the deduced amino acid sequences were carried out in NCBI BLAST and in EMBL BLAST. Shown are the identities and similarities relative to the complete amino acid sequence match. P indicates phospho.

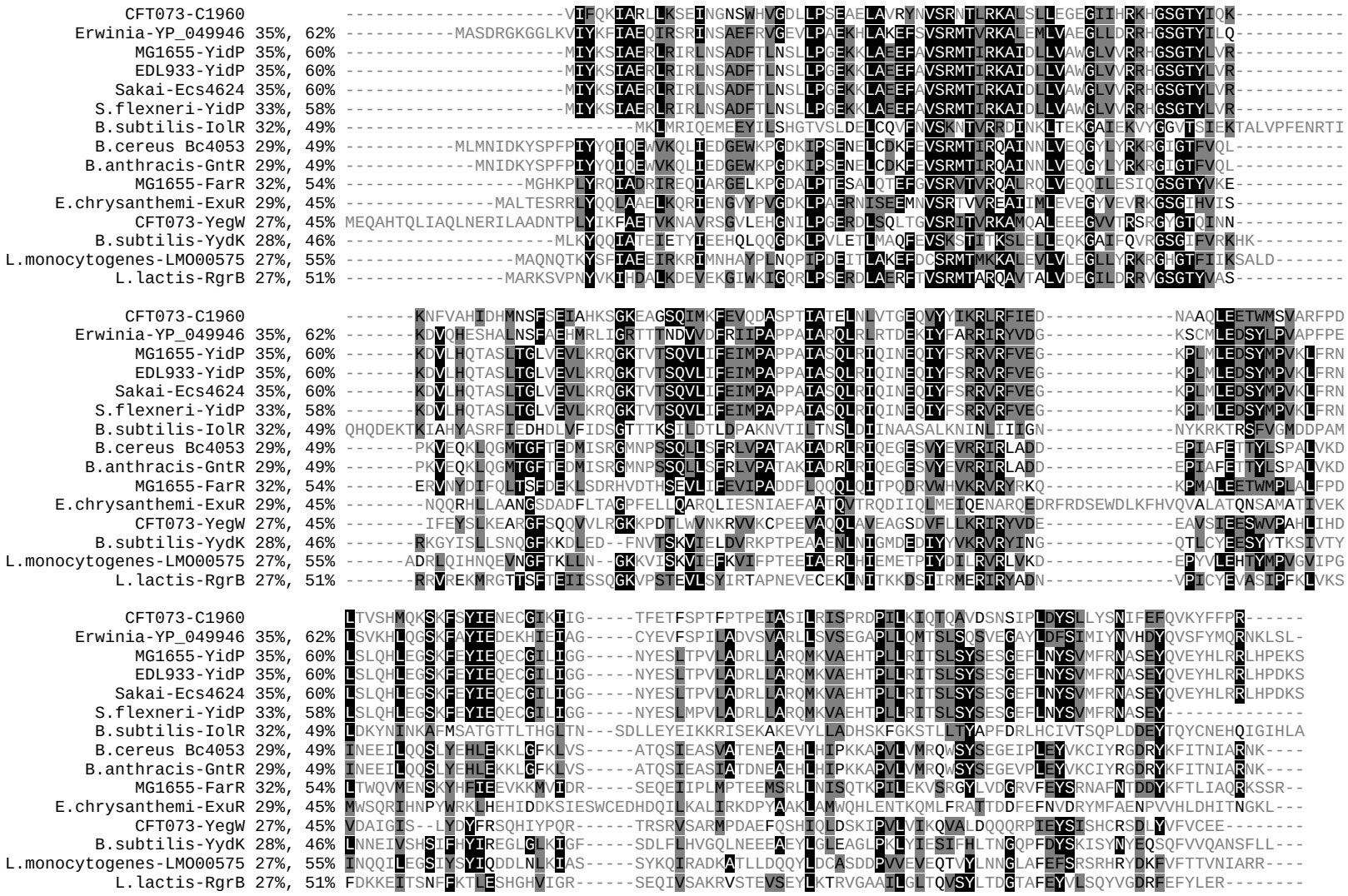


Figure 25: Deduced amino acid sequence alignment of c1960. Identical residues are blocked in black boxes. Similar residues are blocked in grey boxes. Residues that are weakly similar are shown in bold letters and residues that are non-similar are shown in grey. Percentages of – (Identities, similarities)- of C1960 with the follwing proteins are as follows: *Erwinia carotovara* ECA1840 (35%, 62%, GenBank accession number YP_049946), MG1655 YidP (35%, 60%, Blattner et a., 1997), EDL933 YidP (35%, 60%, Perna et al., 2001), Sakai Ecs4624 (35%, 60%, Hayashi et al., 2001), *S.flexneri* YidP (33%, 58%, Wei et al., 2003) *B.subtilis* IolR (32%, 49%, Yoshida et al., 1995), *B.cereus* Bc4053 (29%, 49%, Ivanova et al., 2003), *B.anthracis* GntR (29%, 49%, Read et al., 2002), MG1655 FarR (32%, 54%, Blattner et a., 1997), *E.chrysanthemi* ExuR (29%, 45%, GenBank Q9X930), CFT073 YegW (27%, 45%, Welch et al., 2002), B.subtilis YydK (28%, 46%, Kunst et al., 1997), *L.monocytogenes* GntR (27%, 55%, Nelson et al., 2004), *L.lactis* RgrB (27%, 51%, Bolotin et al., 2001),. Percentages are relative to full length protein.

2.4 Analysis of additional β-glucoside system locus in 171 *E.coli* isolates.

Comparative genomic analysis of the c1955-c1960 region (additional β-glucoside system) of *E.coli* CFT073 with the other three sequenced strains indicate that this locus is uniquely present in CFT073 and is absent in MG1655 and two variants of O157 (EDL933 and Sakai). Thus, this locus comprises of genomic island encoded region in *E.coli* (Fig. 26). The genes upstream to c1955-c1960 region comprise of *marR*, *marA* and *marB* encodes gene products for multiple antibiotic resistances and the genes in the downstream region of c1955-c1960 region comprises of *ydeD* and *ydeF* encodes for proteins of unknown function (hypothetical proteins).

To analyze whether the natural population of *E.coli* carries c1955-c1960 system, all the 171 isolates were analyzed by PCR with c1955-c1960 region specific primers. The primers were designed based on the published nucleotide sequence of strain CFT073 (Welch et al., 2002). All the strains were analyzed by long PCR with a primer mapping in *marB* gene (S401) and the other in *ydeD* gene (S402) (Table 6a and b, Appendix). Those strains that gave expected PCR products like that of CFT073 (in the current study wt-i484 strain was used as control because it carries intact c1955-c1960 system) were further analyzed with specific primers that maps in the internal regions of c1955-c1960 genes and the PCR product sizes were compared to the expected PCR product sizes of CFT073 c1955-c1960 region (see Table 4 and Table 6a and b, Appendix). Strains that do not carry c1955-c1960 system gave expected PCR product size as in MG1655 (used as control) with primers S401 and S402 (Table 4, Appendix). In addition, PCR with primers that maps specifically in the internal regions of c1955-c1960 genes resulted in no PCR products in this group of strains. Thus, confirming the absence of c1955-c1960 region.

Furthermore, it was observed that out of 171 strains, 7 of them resulted in PCR products of unexpected size with primers S401 and S402 (due to insertions or deletions) in comparison to the expected sizes of CFT073. This group of strains was further analyzed with primers that map in the internal region of c1955-1960 and the PCR products that correspond to the possible insertion or deletion point was sequenced from both the ends (Fig. 26). The detailed analysis is shown in Figure 26. The analysis revealed that out of 171 strains: 74 strains do not have c1955-c1960 region and are like MG1655 or O157 (Fig. 26A), 90 strains have c1955-c1960 region intact (Fig. 27B), 3 strains carry IS1 insertions within the c1955-c1960 system (Fig. 26C and D), 1 strain carry IS1 associated deletion from c1957 to c1956 (Fig. 26E) and 3 strains show alterations within the c1955-c1960 region by deletions (Fig. 26F to H).

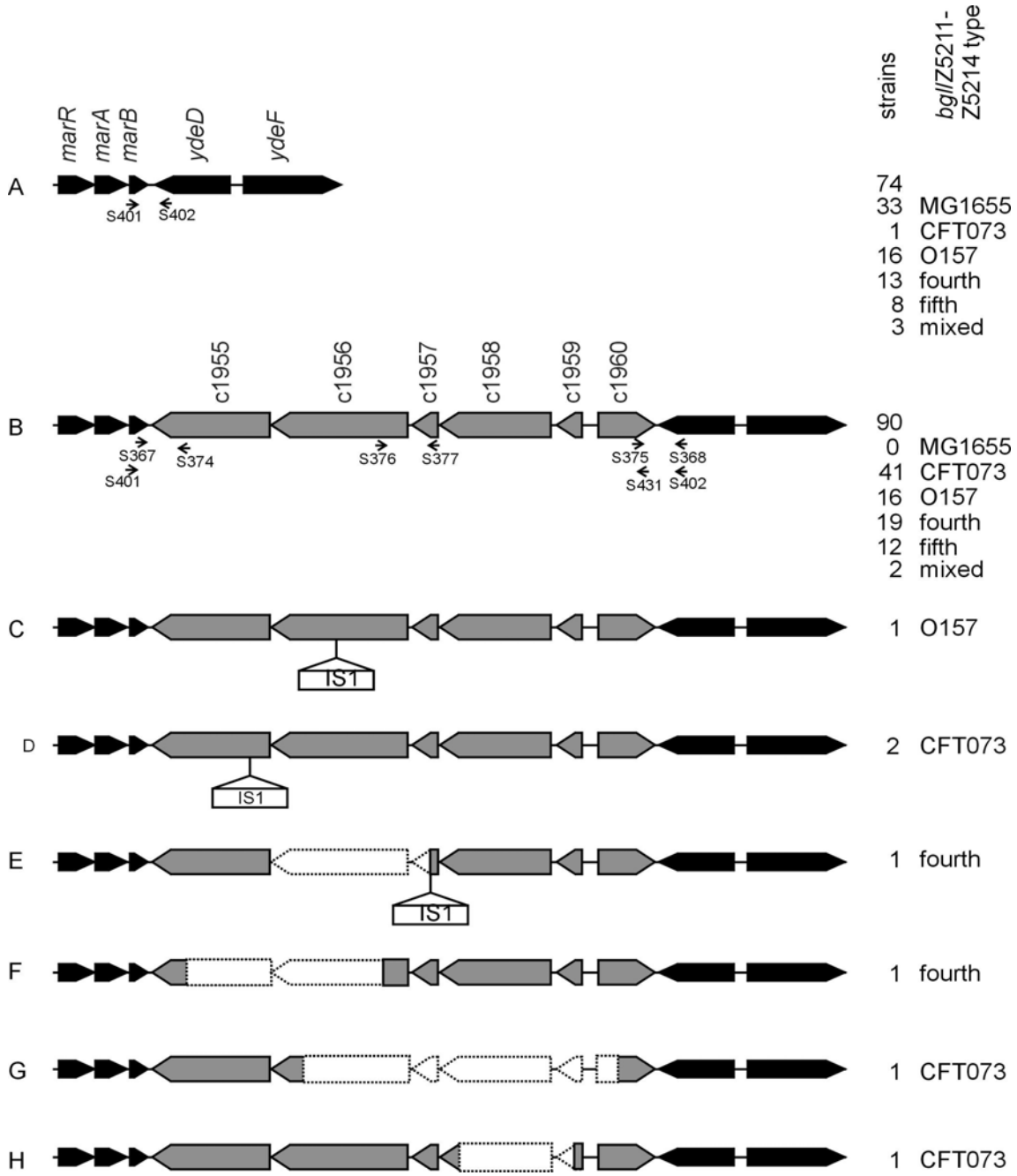


Figure 26: c1955-c1960 region in 171 *E.coli* isolates. Strains were analyzed by PCR. PCR was performed with the primers mapping in *marB* and *ydeD* genes. The c1955-c1960 region is shown in grey and upstream and downstream regions are in black. Out of 171 strains: A) 74 strains gave expected products like MG1655 indicating the absence of c1955-c1960 region. B) 90 strains gave expected products like CFT073 indicating the presence of intact c1955-c1960 region. C) Strain F645 carries IS1 insertion in c1956 with 8 bp target site duplication (AE016760: 289424-298431) D) 2 strains E464 and E466 carries IS insertion in c1955 with 9bp target site duplication (AE016760: 288391-288399). E) Strain ECOR17 carries IS1 associated deletion from AE016760: 288646 to 290509. F) Strain St4723 carries deletion from c1955 to c1956 (AE016760: 287503-290170) G) Strain U2873 carries deletion from c1956 to c1960 (AE016760: 288957 to 292879) H) Strain ECOR63 carries deletion from c1958 to c1959 (AE016760: 290809 to 292442). The positions mentioned are relative to the Genebank primary accession number of *E.coli* CFT073 genome. Text direction in the insertion elements represents the relative orientations of the insertion elements according to their defined left and right ends. Typing at *bgl*/Z5211-Z5214 locus is also shown and numbers in each type are also shown. Primer mapping positions are shown.

The strains E464 and E466 carry IS1 insertion in c1955 at the same insertion point suggesting that the strains could be derivatives of each other (Fig. 26D). Taken together, the analysis at the c1955-c1960 region also revealed evidence that in nature deletions and insertions are common features of bacterial genome evolution.

2.5 Correlations of c1955-c1960 analysis with the phylogenetic distribution of ECOR strains

Analysis of c1955-c1960 region in the 72 ECOR strains significantly correlated to its phylogenetic distribution (Table 3). 19 out of 25 strains that belong to Group A do not carry c1955-c1960 system (like MG1655). In contrast, 13 out of 16 strains that belong to Group B1 and 14 out of 15 strains that belong to Group B2 carry c1955-c1960 system (like CFT073). 8 out of 12 strains that belong to Group D carry c1955-c1960 system and none of the Group E strains carry c1955-c1960 system. Overall, the data shows that the presence of c1955-c1960 system is predominant in the strains that belong to Group B2 and B1.

Table 3: c1955-c1960 system and ECOR phylogeny

	72 strains[a]	c1955-c1960 system[b]	
		+	-
Phylogenetic types[c]	A-25 (35%)[d]	6 (8%)	19 (27%)
	B1-16 (22%)	13 (18%)	3 (4%)
	B2-15 (21%)	14 (20%)	1 (1%)
	D-12 (17%)	8 (12%)	4 (5%)
	E-4 (5%)	0	4 (5%)

a: 72 ECOR strains used in the study
b: Presence (+) or absence (-) of c1955-c1960 system
c: phylogenetic groups are based on Herzer et al., 1990
d: number of strains are shown and percentages are calculated to the total number of ECOR strains analyzed in the study.

2.6 The four spontaneous mutants carry identical point mutation in the putative regulatory region

In order to characterize the spontaneous activation of the c1955-c1960 system, the putative regulatory region between c1959 (β-glucoside specific PTS IIB domain) and c1960 (putative regulator) was sequenced from the starter strain i484 Δ*bgl* (Ec93) and from the four spontaneous mutants (Ec131 to Ec134). Nucleotide sequence alignment of the i484 Δ*bgl* and the four mutants with published sequence of CFT073 revealed that strain i484 Δ*bgl* has identical sequence as in CFT073. While the four spontaneous mutant carry an identical point mutation from G to A at

position AE016760- 292542 (position relative to GenBank accession number) (Figure 27). The sequence around the mutation matches with the consensus sequence of CAP binding site. In addition, the mutation was seen to be in the non-conserved nucleotide of the putative CAP-binding site. Thus, the sequence in this region looks like a typical Class II CAP dependent promoter region. In case of Class II CAP dependent promoters the CAP binding site is centered at the -41.5 nucleotide position (with respect to the transcription start) and the -35 box is replaced with CAP binding site. To this end, the c1955-c1960 system may possess a putative Class II CAP dependent promoter which needs to be further substantiated.

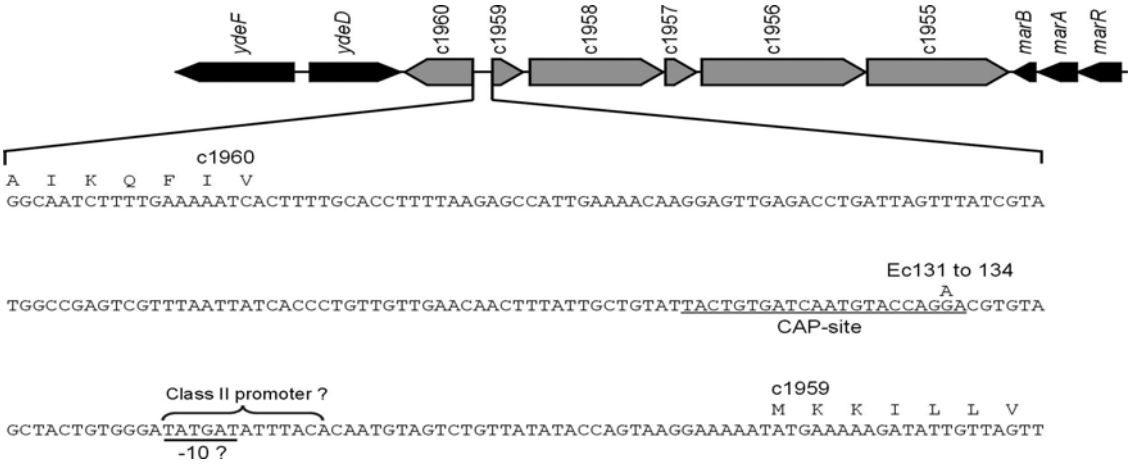


Figure 27: Nucleotide sequence of the putative regulatory region of the c1955-c1960 system. The c1955-c1960 system is shown in inverted orientation with respect to Fig. 19. Nucleotide sequence is shown from AE016760: 292694 to 292455 of the *E.coli* CFT073 genome. The nucleotide sequence of Ec93 (i484 Δ*bgl*) is identical to CFT073 with no sequence changes. The four spontaneous mutants (Ec131 to Ec134) carry identical point mutations from G to A. The nucleotide change is shown at the top of the sequence. Putative CAP binding site is underlined. Putative Class II CAP dependent promoter is shown by a brace and putative -10 regions are shown by horizontal line below the sequence. Deduced amino acid sequences of c1959 and c1960 are shown at the top of the nucleotide sequence.

2.7 c1955-c1960 system encodes genes for β-glucoside utilization

In order to test whether the constitutive expression of the c1955-c1960 system allows utilization of β-glucosides, the putative promoter of the c1955-c1960 system was replaced with *tac* promoter under the control of *lacI*. The expression of these constructs can be regulated with IPTG. The putative structural genes c1955 to c1959 from i484 Δ*bgl* and from two spontaneous mutants Ec132 and Ec134 were cloned and the phenotypes of the resulting plasmids in Ec93 (i484 Δ*bgl*) and S541 (K-12 Δ*bgl*) were determined on BTB indicator plates in the presence of four β-glucosides (Salicin, Cellobiose, Arbutin and Esculin, Fig. 28). The results revealed that in i484 Δ*bgl* background, wt-construct (that carries c1955-c1959 genes from i484 Δ*bgl*) is sufficient for the utilization of β-glucosides (Salicin, Cellobiose, and Arbutin). The results are also similar

with other two constructs carrying activated fragments (pKEGN53 and 54). However, in K-12 Δ*bgl* background the wt-construct did not complemented for the β-glucoside utilization at both 28°C and 37°C. In the two constructs carrying activated fragments, construct pKEGN53 was able to complement whereas, the other construct did not complemented. Complementation in K-12 Δ*bgl* for the β-glucoside utilization with one of the activated derivative (pKEGN53) could be due to PCR generated errors in the structural genes. Interestingly, it was observed that an Arbutin positive (Arb^+) phenotype was seen with both wt-constructs and activated constructs in i484 Δ*bgl* and in K-12 Δ*bgl* at both 28°C and 37°C (Fig. 28). This is possibly due to the presence of the gene *bglA*. The gene *bglA* is conserved in all the four sequenced *E.coli* strains, encodes an enzyme phospho β-glucosidase that is necessary for the hydrolysis of β-glucosides like Arbutin. The enzyme BglA can cleave β-glucosides only when the sugars are transported into the cell (Prasad et al., 1973). Thus, the constitutive expression of c1955 to c1959 (putative transporters) in the constructs analyzed could function for the transport of arbutin and the Arb^+ phenotype seen in the i484 Δ*bgl* and K-12 Δ*bgl* transformants could be attributed to the BglA activity. Taken together, the data suggests that c1955-c1960 system encodes genes necessary for β-glucoside utilization system.

2.8 c1955-c1960 system is ON in septicemic isolate background but OFF in K-12 background

The constitutive expression of the structural genes (c1955-c1959) showed that c1955-c1960 system encodes genes for β-glucoside utilization. In order to analyze whether the c1955-c1960 system allows for the utilization of β-glucosides in i484 Δ*bgl* and K-12 Δ*bgl* background, the genes c1955 to c1960 from i484 Δ*bgl* and from the four spontaneous mutants were cloned into a plasmid vector containing pACYC origin. The plasmids were transformed into i484 Δ*bgl* background (Ec93) and in K-12 Δ*bgl* (S541) background. The phenotypes of the tranformants were determined on BTB indicator plates in the presence of four β-glucosides (Fig. 28). The results revealed that Ec93 tranformants with the plasmids carrying fragments from spontaneous mutants (pKEGN42 to 45) showed Sal^+ Cel^+ Arb^+ Esc^- phenotype at 28°C. However, at 37°C only Arb^+ phenotype was seen. This suggested that the enzyme activity of the phospho β-glucosidase (encoded by c1959) is temperature sensitive. In the K-12 Δ*bgl* transformants the results revealed that the plasmids carrying fragments from spontaneous mutants (pKEN42 to 45) did not allow for the utilization of Salicin, Cellobiose and Esculin at 28°C and 37°C. However, a strong Arb^+ phenotype was seen at both the temperatures (Fig. 28). As mentioned earlier (in

section 2.7) the Arb$^+$ phenotype is possible due to the presence of *bglA* gene. At an outlook, it also suggests that additional factors might be necessary for the expression of c1955-c1960 system and these factors might be absent in K-12.

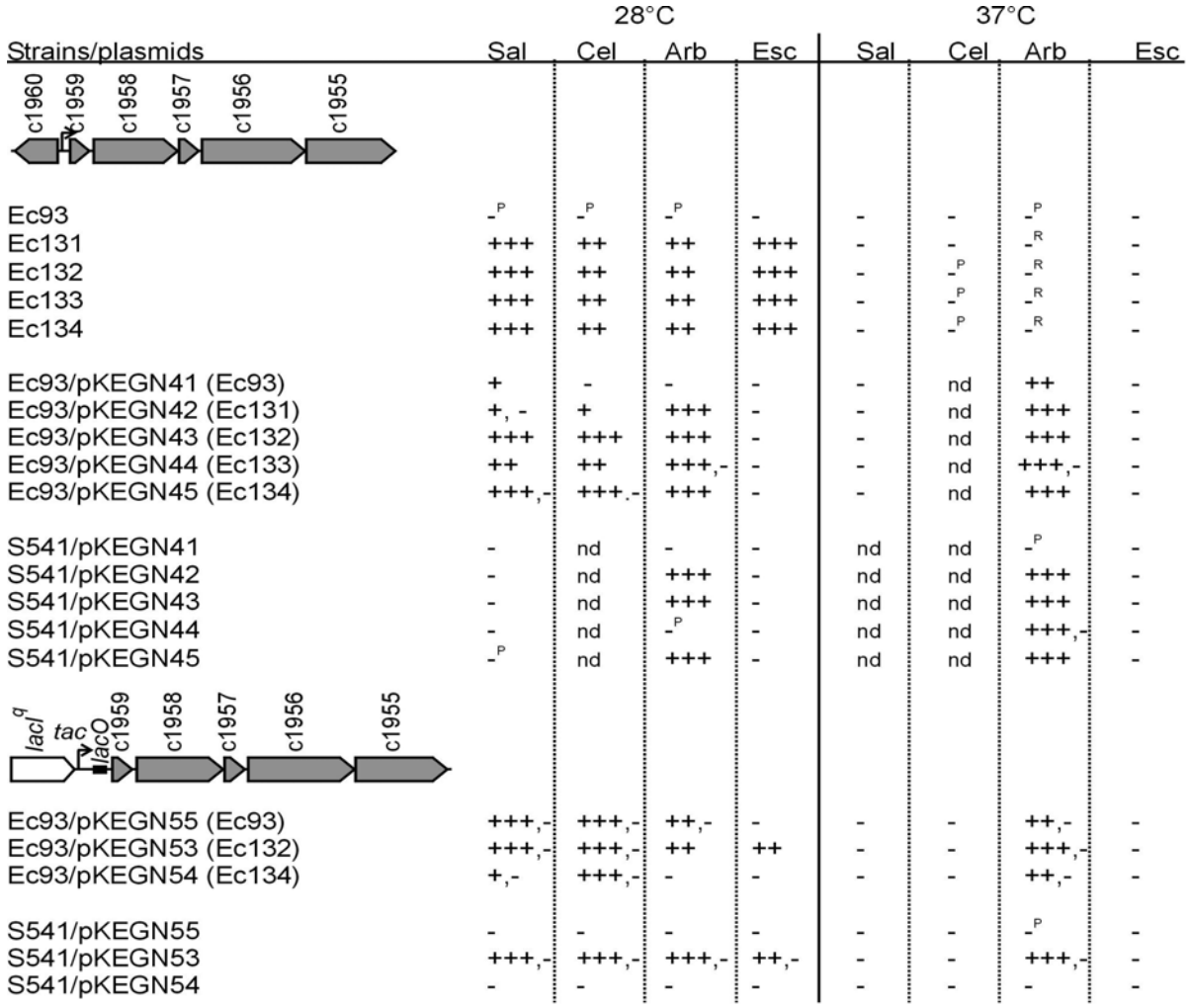


Figure 28: Phenotypes of the c1955-c1960 system constructs in i484 Δ*bgl* (Ec93) and K-12 Δ*bgl* (S541). Shown are the phenotypes on BTB indicator plates with 0.5% Sal-Salicin, Arb-Arbutin, Cel-Cellobiose, Esc-Esculin at 28°C and 37°C. The structures of the two plasmidic constructs used in the complementation analysis are also shown. Strain Ec93 carries wt-c1955-c1960 system and strain Ec131 to Ec134 carries activated derivatives of c1955-c1960 system. The phenotypes of the un-transformed strains are also shown at the top panel. Followed by phenotypes of the Ec93 transformants complemented with plasmids (pKEGN41 to 45). Plasmids and strains from which the fragments are taken are shown as follows: pKEGN41 (Ec93), pKEGN42 (Ec131), pKEGN43 (Ec132), pKEGN44 (Ec133), pKEGN45 (Ec134). Constructs pKEGN42 to 45 can complement the utilization of β-glucosides in Ec93 (at 28°C) but not in S541. The constructs pKEGN53 to 55 carry fragments from Ec132, Ec133 and Ec93 respectively and the complementation was seen only in Ec93 background and not in S541. Phenotypes of the transformants with plasmids pKEGN53 to 55 were determined in the presence of 1mM IPTG. Phenotypes were determined from two independent experiments and from two independent plasmidic clones. The results are reproducible in each case. +++ Sal$^+$, ++ medium Sal$^+$, + weak Sal$^+$, - Sal$^-$, –^{P} Sal$^-$ but papillates and –^{R} relaxed phenotype.

2.9 The promoter of c1955-c1960 system is CAP dependent and is catabolically repressed in the presence of glucose

To test whether the point mutation seen in the CAP-binding site has influence on the promoter activity, two different *lacZ* fusions were made from wt (i484 Δ*bgl*) and from the activated derivative that carry point mutation in the CAP-binding site of c1955-c1960 system (Ec134). The first construct comprises of the promoter ($P_{c1955-c1960}$) fused to *lacZ* and the second construct with the gene c1960 with $P_{c1955-c1960}$ fused to the *lacZ* (Fig. 29). The plasmidic constructs were integrated into S541 (K-12 Δ*bgl* Δ*lacZ*) chromosomal *attB* site (see materials and methods). The expression of the integrated derivates were determined from cultures grown to the exponential phase (OD_{600} of 0.5) in M9 medium with Casamino acids, Vitamin B1, 0.5% glycerol and antibiotics. It was seen that in the wt construct, the promoter activity was reduced 2 fold in the presence of putative regulator c1960 (200 to 100U, Fig. 29I A and B) suggesting c1960 as a putative repressor in the c1955-c1960 system. The promoter activity in the constructs carrying activated promoter, activity was increased to ~10 fold in comparison to the wt-$P_{c1955-c1960}$-lacZ construct (200 to 1905U, Fig.29I A and C). In addition, the repression of the promoter activity was not affected by c1960 in the activated derivatives suggesting that the point mutation might have hindered the binding of putative repressor in the promoter region (Fig. 29).

In order to test whether the $P_{c1955-c1960}$ activity is catabolically repressed in the presence of glucose, expression of the constructs were determined in the presence of 0.5% glucose. A 2 to 3 fold decrease in the promoter activity was seen in all the constructs suggesting that the promoter activity is catabolically repressed in the presence of Glucose. To test whether the additional β-glucoside system promoter is CAP dependent, *cyaA* gene that codes for the adenylate cyclase enzyme catalyzing production of cyclic AMP (cAMP) was deleted resulting in K-12 Δ*bgl* Δ*lacZ* Δ*cyaA* (strains S2272 to S2279). Due to a lack of cAMP in these strains the active CAP-cAMP complex cannot form and bind to the CAP binding site of the c1955-c1960 promoter region thereby expected to cause low or no expression. The results revealed that the promoter activity was very low in all the four reporter constructs in the absence of cAMP (see –cAMP values, Fig. 29). However, in the presence of 1mM cAMP the activity of the constructs with activated promoter showed ~10 fold increase in comparison to the wt-constructs (Fig. 29I compare values in +cAMP of A, B to C, D). To further support that the promoter is CAP dependent, expression of the plasmidic constructs were determined in *crp* background (this strain lacks CAP protein)

and the results revealed that activity in all the constructs were drastically reduced (Fig. 29II, see *crp* readings). Expression of the plasmidic constructs in wt in the presence of Glucose or Glycerol (Fig. 29II) supported the data with the chromosomal constructs (Fig. 29I), where it was seen that promoter activity of the constructs that carry activated derivates was enhanced to ~ 20 fold in comparison to the wild type constructs (Fig. 29II compare A, B to C, D).

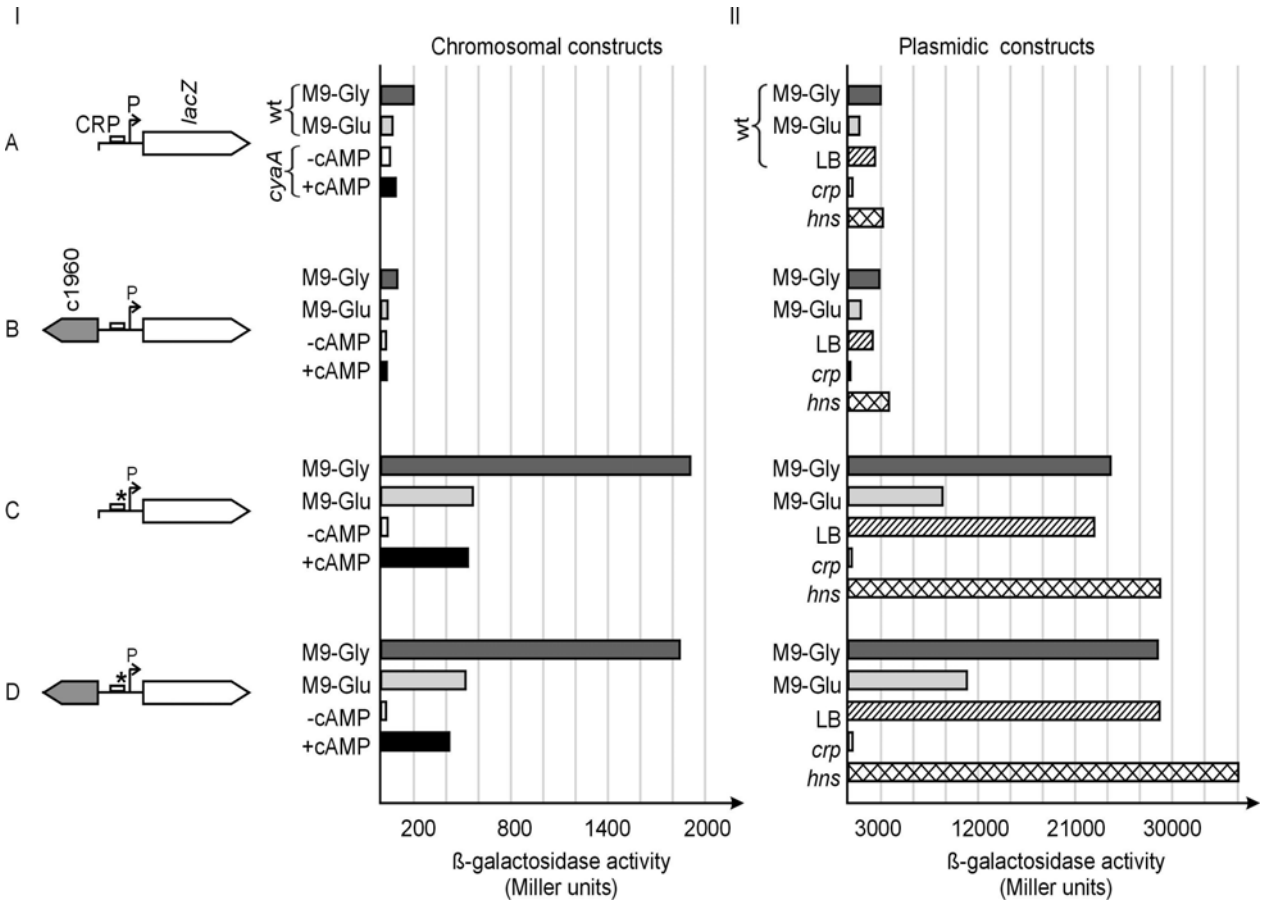


Figure 29: β-galactosidase acitivity of $P_{c1955\text{-}c1960}$-*lacZ* reporter constructs. β-galactosidase activity is expressed in miller units (Miller, 1972). Bars represent the β-galactosidase activity. Shown are the average results from atleast 3 independent experiments and from at least two independent clones. **I)** Assays of the wt (S541) cultures were carried out in M9 medium with either Glycerol or Glucose as carbon source (materials and methods) and assays for c*yaA* were carried out in LB medium with or without 1mMcAMP. Plasmids used for chromosomal integrations are as follows: A) pKEGN46 B) pKEGN51 C) pKEGN48 D) pKEGN52. Strains carrying the reporter constructs in *attB* (units in M9-Gly, M9-Glu, cyaA –cAMP and cyaA +cAMP) are as follows: A) S2180/81 (200, 70, 55, 90) B) S2184/5 (100, 40, 30, 35) C) S2182/3 (1905, 565, 40, 535) D) S2186/7 (1840, 520, 30, 420). Standard deviation errors are less than 10%. The strains with Δ*cyaA* carrying identical reporter constructs are as follows: A) S2272/3 B) S2276/7 C) S2274/5D) S2278/9. * indicates mutation in CAP binding site **II)** Plasmidic assays of the reporter constructs are shown. Plasmids used for the assays are shown as in I. Assays in wt (S541) in M9 and LB medium are indicated. Assays of *crp* (S996) and *hns* (S614) were carried out in LB medium. Units in Wt-M9-Glu, M9-Gly, LB, *crp* and *hns* are as follows: A) 2995, 1020, 2465, 380, 3200 B) 2875, 1150, 2250, 195, 3760 C) 24245, 8720, 22738, 318, 25840 D) 28605, 10960, 28858, 390, 36050.

It's been known that the histone like protein (H-NS) is involved in the silencing of the *bgl* operon (Dole et. al., 2004). To test whether H-NS has any role in the regulation of the c1955-c1960 system, the expression of the plasmidic $P_{c1955\text{-}c1960}$-*lacZ* reporter fusions were determined in *hns* background. The results showed that no significant difference was observed in the *hns* in comparison to the wt (Fig. 29II, compare values of wt (LB) and *hns*). Taken together, the data suggest that the promoter of the c1955-c1960 system is CAP dependent and is catabolically repressed in the presence of glucose. In addition, it also suggests that H-NS has no role on the promoter activity.

2.10 Expression of $P_{c1955\text{-}c1960}$-*lacZ* reporter constructs are induced by salicin in septicemic isolate background (i484 Δ*bgl)* that carries activated c1955-c1960 system.

To test whether the expression of the $P_{c1955\text{-}c1960}$-*lacZ* reporter constructs can be induced with the β-glucosides (like salicin, cellobiose). The strain i484 Δ*bgl* and two mutants that carry activated derivatives of c1955-c1960 system (Ec132 and Ec134) were transformed with plasmidic constructs carrying wt-$P_{c1955\text{-}c1960}$-*lacZ* (pKEGN46), wt-c1960-$P_{c1955\text{-}c1960}$-*lacZ* (pKEGN52) and respective $P_{c1955\text{-}c1960}$ reporter constructs carrying fragments with point mutation (activated derivatives, pKEGN48, 52) and the expression was determined in exponential growth phase of the cells grown at 28°C and 37°C (Fig. 30).

In strain i484 Δ*bgl* no difference was seen in all the four constructs in the presence of Salicin or cellobiose at both temperatures. However, in the c1955-c1960 spontaneous mutants (Ec132 and Ec134), the wt-c1960-$P_{c1955\text{-}c1960}$-*lacZ* construct (pKEGN51) showed ~5 to 8 fold increase in the activity in the presence of Salicin in comparison to the absence (8220U, 7800U in comparison to 920U at 28°C and 14615U, 14140U in comparison to 2905U at 37°C, Fig. 30B). The construct that carried activated c1960-$P_{c1955\text{-}c1960}$-*lacZ* fragment (pKEGN52) showed a ~1.5 fold enhancement in the promoter activity in the presence of Salicin at 28°C (13575U and 12200U in comparison to 8310U, Fig. 30D). No effect of Salicin was seen in the constructs carrying just the promoter alone fused to the *lacZ* gene (Fig. 30A and C). In addition, no induction with cellobiose was seen at both temperatures in all the four constructs. To this end, the data suggests that upon transport of the β-glucosides into the cell the putative regulator c1960 can also act as an activator of the c1955-c1960 system.

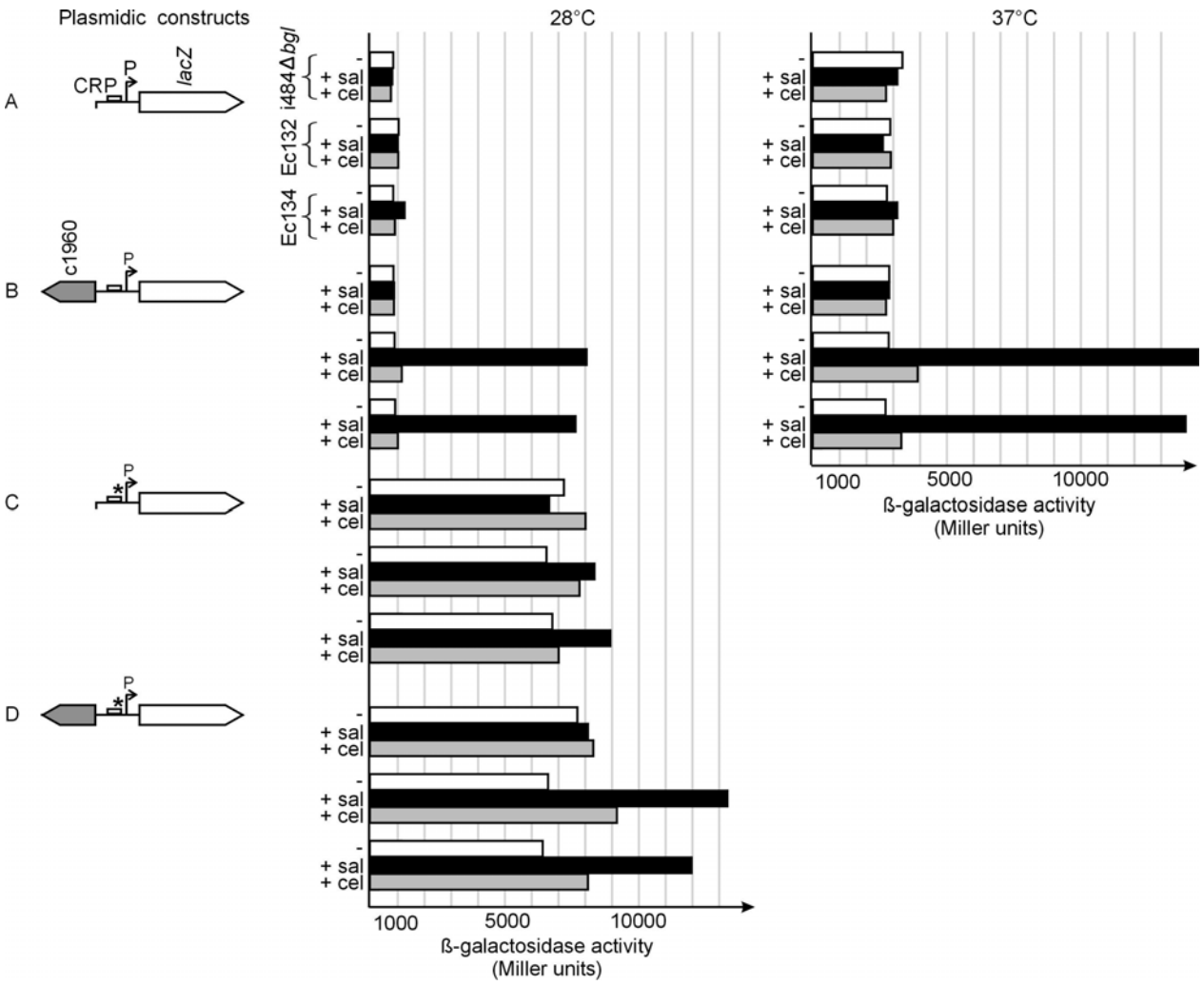


Figure 30: β-galactosidase acitivity of $P_{c1955\text{-}c1960}$-*lacZ* reporter constructs in i484 Δ*bgl* and two spontaneous mutants of c1955-c1960 system (Ec132 and Ec134). β-galactosidase activity is expressed in miller units (Miller, 1972). Bars represent the β-galactosidase activity. Shown are the average results from at-least 3 independent experiments and from at least two independent clones. All assays were carried out in M9 medium with 0.5% glycerol, vitamin B1, casa amino acids, antibiotics and with or without β-glucosides like salicin, cellobiose. – indicates no β-glucosides in the medium, +sal indicates-with 0.2% salicin, +cel-0.2% cellobiose. Plasmids used in the analysis are A) pKEGN46 B) pKEGN51 C) pKEGN48 D) pKEGN52. Strain name, units (-sal, +sal, +cel) are as follows: A) i484 Δ*bgl* (885, 850, 800 at 28°C) (3400, 3210, 2780 at 37°C). Ec132 (1090, 1060, 1075 at 28°C) (2940, 2660, 2965 at 37°C), Ec134 (885, 1320, 945 at 28°C) (2810, 3246, 3050 at 37°C). B) i484 Δ*bgl* (895, 920, 915 at 28°C) (2900, 2905, 2780 at 37°C). Ec132 (935, 8220, 1210 at 28°C) (2880, 14615, 3980 at 37°C). Ec134 (960, 7800, 1060 at 28°C) (2765, 14140, 3360 at 37°C). C) i484 Δ*bgl* (7360, 6995, 8180). Ec132 (6700, 8520, 7950). Ec134 (6920, 9020, 7160). D) i484 Δ*bgl* (7870, 8310, 8475).Ec132 (6775, 13575, 9370). Ec134 (6585, 12200, 8270). * indicates mutation in the CAP binding site.

2.11 The β-glucosides salicin, cellobiose, chitobiose, arbutin and esculin are not inducers of c1955-c1960 in K-12 background.

The phenotypes of the S541 (K-12 Δ*bgl* Δ*lacZ*) complemented with two different types of constructs carrying c1955-c1960 system revealed that c1955-c1960 system cannot complement for the utilization of β-glucosides in K-12 (Fig. 28). In addition, the induction of wt-c1960-$P_{c1955\text{-}c1960}$-*lacZ* activity with β-glucoside (Salicin) in the septicemic isolate background (i484 Δ*bgl*) (Fig.30) initiated further studies to carry out in K-12 Δ*bgl* Δ*lacZ* background. Thus, to test

whether the c1955-c1960 promoter activity is induced with any of the β-glucoside in K-12 background, the expression of the chromosomal constructs carrying wt-$P_{c1955\text{-}c1960}$ was determined in the presence of two different constructs provided *in trans* (Fig. 31). One construct carries structural genes (c1955-c1959) from the activated derivatives driven by inducible *tac* promoter (pKEGN53) and the other construct carries activated c1955-c1960 system driven by its own promoter ($P_{c1955\text{-}c1960}$) (pKEGN43).

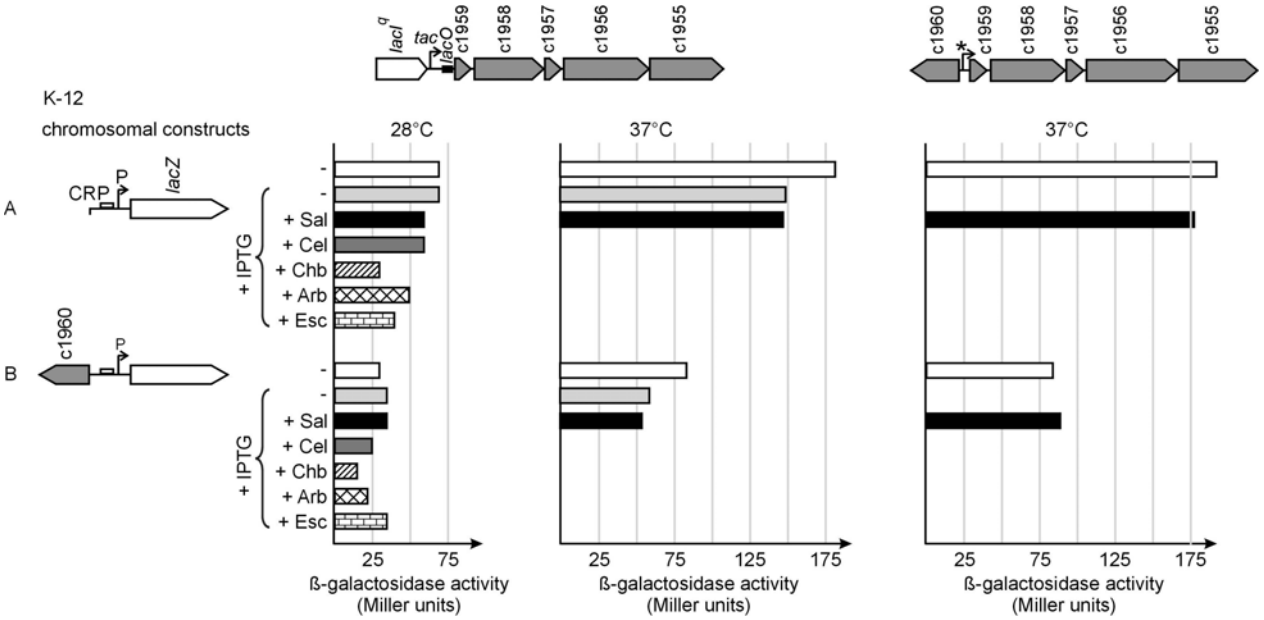


Figure 31: β-galactosidase acitivity $P_{c1955\text{-}c1960}$-*lacZ* reporter constructs in K-12 Δ*bgl*. β-galactosidase activity is expressed in miller units (Miller, 1972). Shown are the average results from at-least 3 independent experiments and from at least two independent clones. * indicates activated c1955-c1960 system. Bars represent the β-galactosidase activity. Strains carrying the chromosomal constructs are listed as in Fig. 29. Shown are the values with two constructs provided in trans: pKEGN53-carries structural genes of c1955-c1959 (from the activated derivative) driven by inducible *tac* promoter, pKEGN43-carries the intact c1955-c1960 system (from the activated derivative). Identical results were seen with two other respective independent clone pKEGN52 and pKEGN44. Assays with pKEGN53 *in trans* were performed at both 28°C and 37°C in the presence or absence of β-glucosides like salicin (sal), cellobiose (cel), chitobiose (chb), arbutin (arb) and esculin (esc) with or without 1mM IPTG. Assays with pKEGN43 *in trans* were performed with or without salicin at 37°C. All assays were carried out in M9 medium with 0.5% glycerol, vitamin B1, casa amino acids and with relevant antibiotics. – indicates no β-glucosides. β-galactosidase activity in units with pKEGN53 in trans are as follows: at 28°C absence of β-glucosides, +IPTG, +Sal, +Cel, +Chb, +Arb, +Esc (construct A: 70, 70, 60, 60, 30, 50, 60; construct B: 30, 35, 35, 25, 15, 22, 35) at 37°C absence of β-glucosides , +IPTG, +Sal (A: 185, 150, 150 B: 85, 60, 55). β-galactosidase activity in units with pKEGN54 in trans: absence of β-glucosides, +Sal (A: 195, 180 B: 85, 90).

The promoter activity was determined from exponential growth phase of the cells (Fig. 31). The results revealed that in the strains that carry construct with the structural genes (c1955-c1959) under the control of inducible *tac* promoter, none of the tested β-glucosides (salicin, cellobiose, chitobiose, arbutin and esculin) induced the $P_{c1955\text{-}c1960}$ activity. Also, no effect of Salicin was seen on the $P_{c1955\text{-}c1960}$ activity in the presence of intact c1955-c1960 system. Taken together, these results demonstrate that the tested β-glucosides are not inducers of c1955-c1960 system in K-12 background.

2.12 Expression of $P_{c1955\text{-}c1960}$-*lacZ* reporter construct is induced by salicin and arbutin in K-12 background that carries activated copy of *bgl* operon.

The expression of $P_{c1955\text{-}c1960}$-*lacZ* constructs in septicemic isolate background carrying activated copy of c1955-c1960 system suggested that upon transport of the β-glucosides into the cell c1955-c1960 system is induced. To test this idea in the K-12 background, constructs carrying wt-$P_{c1955\text{-}c1960}$-*lacZ* reporter systems were analyzed in wt (S49), K-12 Δ*bgl* (S162) and spontaneous mutant of *bgl* operon that carry *bglR*::IS1 (S157) (Fig. 32).

In the wt strain and in K-12 Δ*bgl* no difference was seen in the presence or absence of Salicin at both temperatures. However, in the strains that carries activated copy of the *bgl* operon promoter activity was significantly enhanced in the presence of salicin at 37°C. In the spontaneous mutant that carry *bglR*::IS1 the promoter activity of the wt-c1960-$P_{c1955\text{-}c1960}$-*lacZ* construct was enhanced to 4 fold at 37°C in the presence of salicin and arbutin in comparison to the absence (compare 14370U and 11320U to 3505U Fig. 32B). However, no such difference was seen at 28°C.

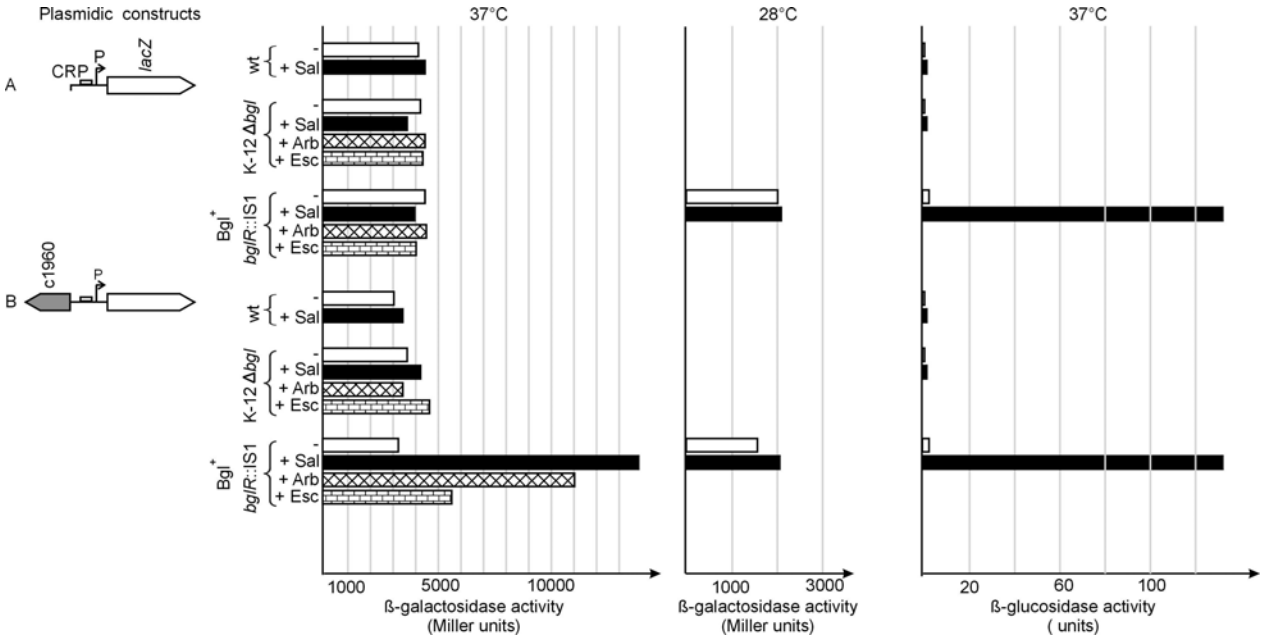


Figure 32: β-galactosidase and β-glucosidase activity of the wt-$P_{c1955\text{-}c1960}$-*lacZ* reporter constructs in K-12 background. β-galactosidase activity is expressed in miller units (Miller, 1972). Shown are the average results from at-least 3 independent experiments and from at least two independent clones. Plasmids used in the analysis are A) pKEGN46 B) pKEGN51. Plasmids were transformed in to strains wt (S49), K-12 Δ*bgl*, and Bgl^+ mutant bglR::IS1 (S157). All assays were carried out in M9 medium with 0.5% glycerol, vitamin B1, casa amino acids, and antibiotics and with or without β-glucosides like Salicin (Sal) Arbutin (Arb), and Esculin (Esc). β-glucosidase activity was determined for strains S49, S162, S157, S432 to quantitate the level of β-glucosidase expression (see materials and methods). β-galactosidase activity in Units at 37°C (no β-glucosides, +Sal, +Arb, +Esc): Construct A wt (4315, 4890, nd, nd) K-12 Δ*bgl* (4392, 3905, 4635, 4520), Bgl^+ (4690, 4155, 4665, 4200) Construct B wt (3360, 3680, nd, nd) K-12 Δ*bgl* (3905, 4505, 3670, 4990), Bgl^+ (3505, 14370, 11320, 5900). At 28°C construct A Bgl^+ (2055, 2135, nd, nd) construct B (1680, 2120). β-glucosidase activity in units (-Sal, +Sal): construct A wt (1, 2), K-12 Δ*bgl* (1, 2), Bgl^+ (3, 135) construct B wt (1, 2), K-12 Δ*bgl* (1, 2), Bgl^+ (3, 135).

In order to analyze the level of *bgl* operon expression in the spontaneous mutant of *bgl* operon carrying *bglR*::IS1 (S157), β-glucosidase assays were performed (materials and methods). The results revealed that the mutant with *bglR*::IS1 showed ~70 fold enhancement in the β-glucosidase activity in comparison to the wt (135U to 2U, see β-glucosidase activity results Fig. 32). This suggests that in the spontaneous mutant of the *bgl* operon that carry *bglR*::IS1, due to the high expression of *bgl* genes the intake of phosphorylated sugars might have resulted in higher promoter activity of the wt-$P_{c1955\text{-}c1960}$-c1960-*lacZ* reporter construct. To this end, the results suggest that the amount of phosphorylated β-glucosides transported into the cells may influence the promoter activity of c1955-c1960 system in K-12 background. However, whether the intake of phosphorylated β-glucosides or the degradation products of β-glucosides is responsible for this activation remains to be answered.

3. Correlations of the genome variations at *bgl*/Z5211-Z5214 locus to the other carbohydrate utilization systems

(This section, in part, is in collaboration with Prof. Diethard Tautz, Institute for Genetics, University of Cologne, Germany)

3.1 Correlations of *bgl*/Z5211-Z5214 locus typing with c1955-c1960 locus analysis and lactose utilization phenotypes

Comparison of the typing at the *bgl*/Z5211-Z5214 genomic island to that of the presence of c1955-c1960 genomic island revealed interesting correlations. Out of 46 strains that have CFT073 type *bgl* region, 45 strains carry c1955-c1960 system. None of the strains that are MG1655 *bgl/*Z type carry the c1955-c1960 island. In the remaining 4 types, 17 out of 33 O157 *bgl/*Z type, 21 out of 34 fourth *bgl/*Z type strains, 12 out of 20 fifth *bgl/*Z type strains and 2 out of 5 mixed *bgl/*Z type strains have c1955-c1960 system (Fig. 33). The predominance of c1955-c1960 system in the strains that have CFT073 type *bgl* region is statistically significant.

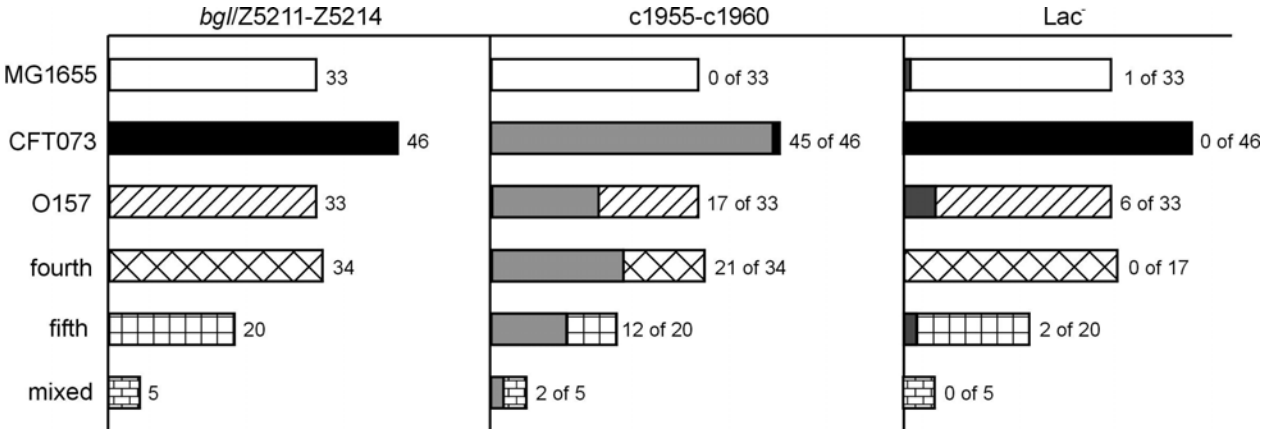


Figure 33: Comparison of *bgl*/Z5211-Z5214 typing with c1955-c1960 locus analysis and lactose utilization phenotypes. Bars represent number of strains in each category. Open bars represent number of strains that are MG1655 type at *bgl*/Z5211-Z5214 locus, closed black bar-CFT073 type, hatched bar-O157 type, cross hatched-fourth type bricked bar-fifth type and squared bar-mixed type. No. of strains that have c1955-c1960 region are represented as light grey bars. The c1955-c1960 region is predominant in CFT073 type, and none of the MG1655 type strains carry c1955-c1960. Out of 171 strains, 99 strains were analyzed for lactose utilization phenotypes. Shown are the numbers of Lac$^-$ strains in each type represented by dark grey bars. Out of nine Lac$^-$ strains, six of them belong to O157 type. The correlations for the predominance of c1955-c1960 region in CFT073 type and predominance of Lac$^-$ in O157 type are found to be statistically significant.

Furthermore, in order to analyze whether the typing of the *E.coli* isolates at the *bgl*/Z5211-Z5214 locus has any correlations with the utilization of other carbohydrates, the Lac phenotype of all the *E.coli* isolates and the *lac* locus of Lac negative strains were analyzed. Statistically significant correlation was seen in the comparison of *lac* operon analysis with the *bgl*/Z5211-Z5214 typing.

Out of nine strains that showed Lac$^-$ phenotype, 6 of them have functional LacI and LacZ and are LacY negative. All these six strains are of O175 *bgl/Z* type (Fig. 33).

3.2 Analysis of *lac* operon in 171 *E.coli* isolates

The *E.coli lac* operon encodes the gene products that are required for the utilization of lactose. The products of *lac* operon are beta-galactosidase LacZ, permease LacY and galactosidase acetyl-transferase LacA. The *lacI* gene that is not the part of *lac* operon encodes for a repressor. The comparative genomics of the four sequenced *E.coli* strains revealed that the *lac* operon is conserved in all the four strains (Fig. 34). However, variations were seen around the *lac* operon region as shown in Figure 34.

All the isolates were analyzed for their phenotype on MacConkey lactose indicator plates at 37°C. The analysis revealed that out of 171 strains analyzed, 162 of them are Lac positive indicating that these strains have intact *lac* operon (Fig. 34A to C). The remaining nine strains that showed Lac$^-$ phenotype were analyzed by PCR with various primers specific for *lac* operon region. Strain W8987 did not gave any PCR product with various primer combinations, indicating the lack of *lac* operon in this strain (Fig. 34H). Strain ECOR6 gave expected PCR products with S435/S546 and no PCR product were seen with S545/S436, S95/S434 and S545/S434 indicating a possible deletion within the *lacZ* gene (Fig. 34G).

The remaining seven strains gave expected PCR products like that of MG1655 indicating the presence of intact *lac* operon. However, strain E10084 gave an increased PCR product with primers S95/S165 (see Table 4, materials and methods) and the obtained PCR product was sequenced from both the ends. It was seen that E10084 carries IS1 insertion in *lacY* gene (Fig. 34E). This could explain why the strain cannot utilize lactose. Due to the functional inactivation of the *lacY* gene, lactose cannot be transported into the cell. To analyze whether LacI and LacZ are functional in these seven strains, β-galactosidase assays in the presence or absence of IPTG were performed. Induction of the β-galactosidase acitivity was seen in the presence of IPTG in strains F645, U4409, E179 E173 and ECOR43 indicating that both LacI and LacZ are functional and the Lac$^-$ phenotypes of these strains could be due to the mutations in the *lacY* gene (Fig. 34D and E). However, in case of strain E10085 β-galactosidase activity was not induced in the presence of IPTG. In an approach to address why the strain E10085 is Lac negative, the *lacI* and *lacZ* genes from E10085 were PCR amplified and cloned into a low copy vector resulting in plasmid pKEGN50. The pKEGN50 was transformed into S527 (MG1655, *lac*$^+$) and phenotypes

were determined. The results revealed that S527 transformed with pKEGN50 showed Lac$^+$ phenotype. This answers that E10085 LacI does not carry mutation that is dominant negative. Furthermore, to see if the promoter of the E10085 *lac* operon is affected, nucleotide sequencing was performed from the PCR product that contain promoter region. The nucleotide sequencing revealed that the promoter region of E10085 is unaffected.

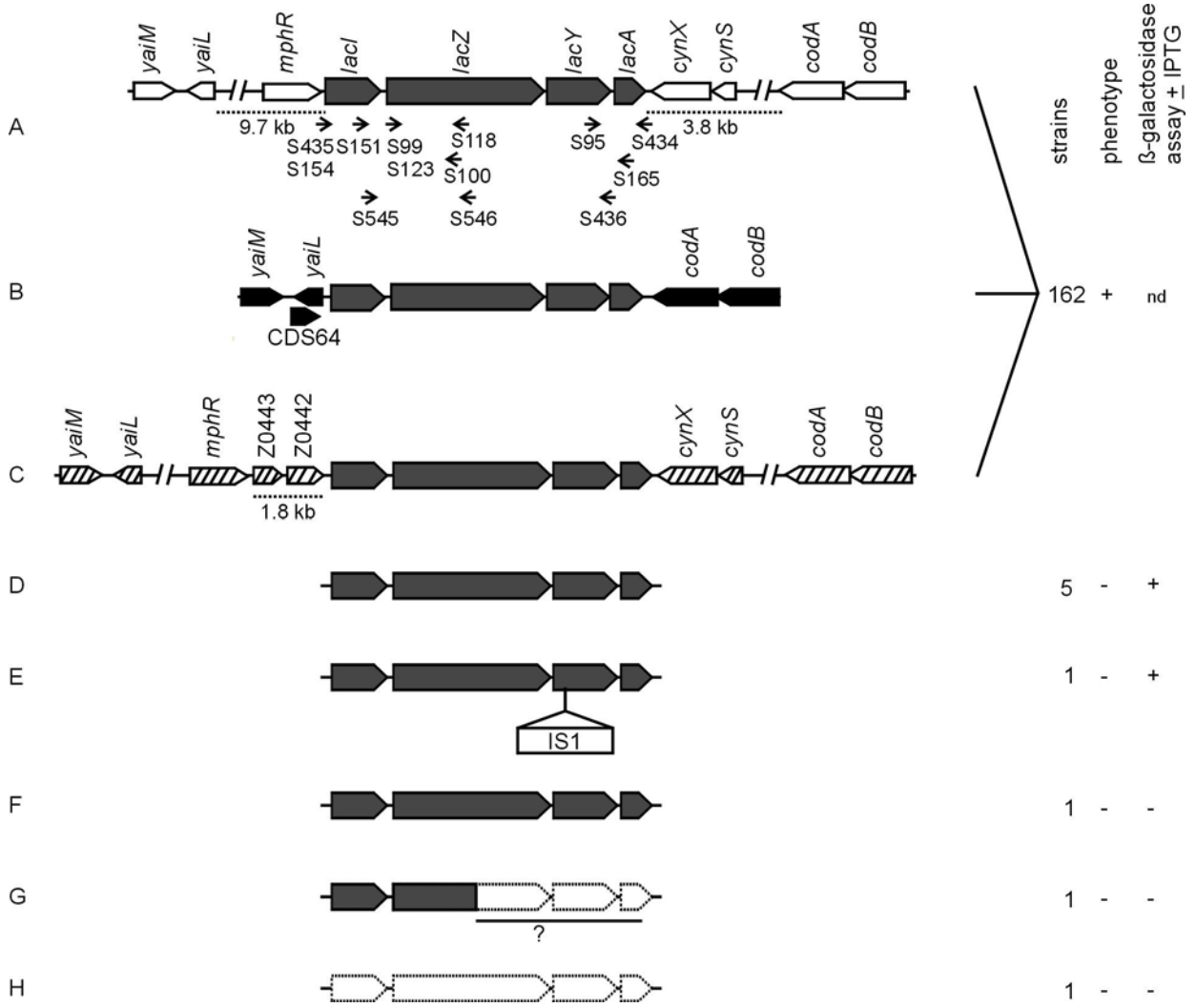


Figure 34: *lac* operon in *E.coli* isolates. The structures of the *lac* operon region in A) MG1655 B) CFT073 C) O157-EDL933 and Sakai are shown. Arrow head represents direction of the oligos. Strain CFT073 lacks a ~9.7kb region at the upstream and a ~3.8kb region at the downstream of the *lac* operon (shown by dotted lines in A). Strain O157 carries an additional 1.8Kb fragment with two open reading frames Z0443 and Z0442 at the upstream region of *lac* operon. However, the downstream region is like MG1655. Phenotypes on MacConkey lactose plates are shown. + indicates Lac$^+$ and – indicates Lac$^-$Results of β-galactosidase assay are shown. + indicates induction of β-galactosidase acitivity in the presence of IPTG and –indicates no induction. D) five strains F645, U4409, E179, E173 and ECOR93 show Lac$^-$ phenotype and show induction in β-galactosidase assay. E) Strain E10084 carries IS1 insertion in *lacY* with target site duplication of 9bp from +3367 to +3375 (positions with respect to transcription start of *lac* operon) shows Lac$^-$ phenotype. F) Strain E10085 and ECOR6 shows Lac$^-$ phenotype and no induction in β-galactosidase assay. G) Strain ECOR6 did not gave PCR product with S435/S434, S545/S436, S545/S434, S95/S434, however it gave expected PCR product size as in MG1655 with S435/S546, S545/S546 and S151/S118. H) Strain W8987 did not result PCR product with S154/S100, S123/S165, S95/S165 and S123/S9 (refer Table 4, materials and methods for primer positions).

3.3 Nucleotide polymorphisms at the *lac* promoter region

The nucleotide sequence alignment of the *lac* promoter region from the four published sequenced strains shows that each strain differs from each other with sequence variations. The Nucleotide sequence alignment of the four sequenced strains is shown in Figure 35.

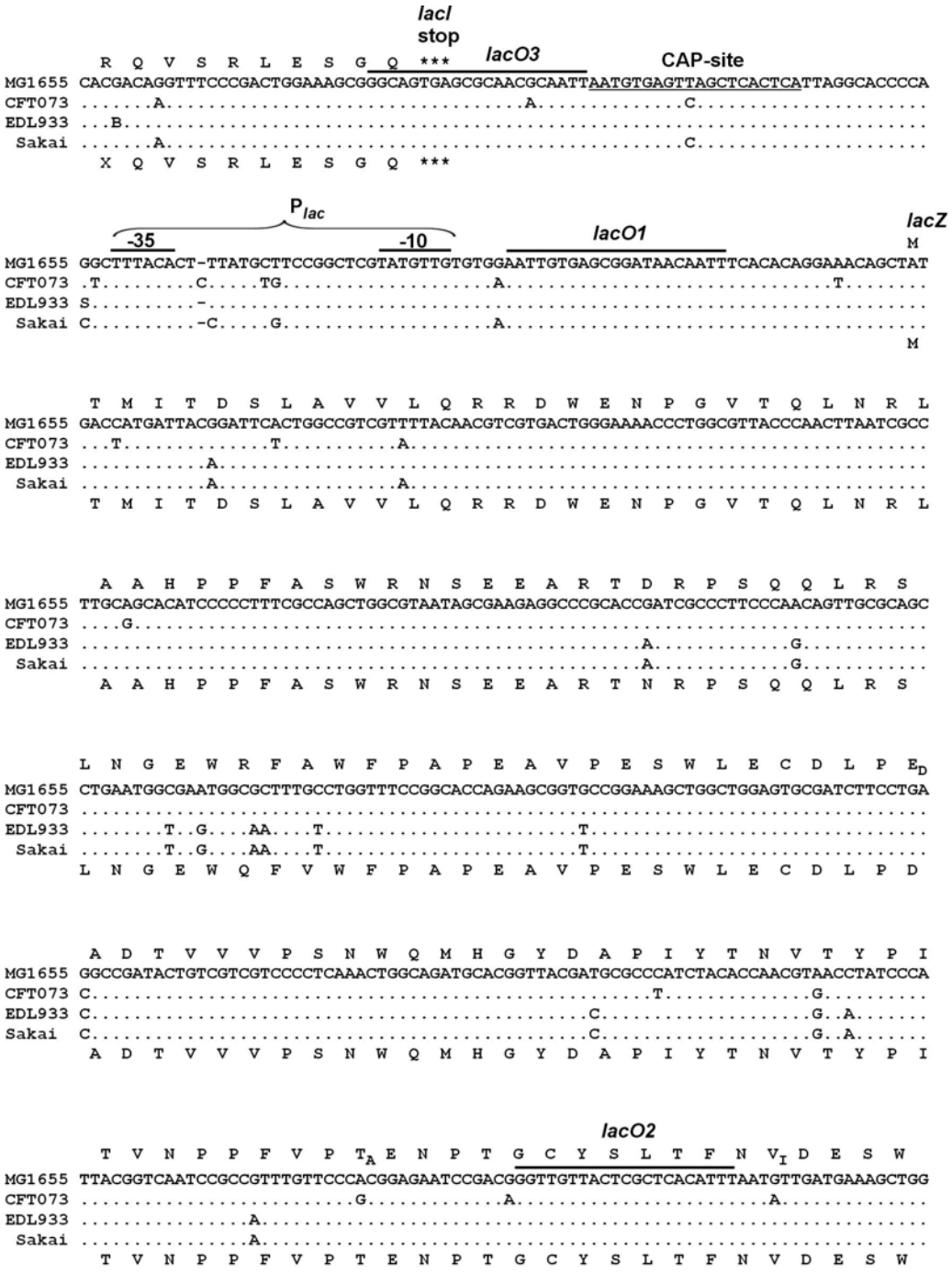


Figure 35: Nucleotide sequence alignment of the *lac* promoter region in the four published sequenced strains. Shown is the nucleotide sequence alignment of MG1655, CFT073, O157-EDL933 and Sakai. Nucleotide sequence changes in comparison to MG1655 are indicated. Conserved nucleotides are shown as dots. Deduced amino acid sequence of MG1655 and CFT073 are shown at the top and O157 variants EDL933 and Sakai are shown at the bottom of the alignments. Amino acid changes in CFT073 are shown as subscripts. *lac* operators are indicated by horizontal lines above the sequence. *lac* promoter is shown by brace and -10 and -35 regions are shown by horizontal lines. CAP binding site is underlined.

Strain CFT073 carries around 16bp sequence variations in comparison to MG1655 sequence. It carries an additional base pair in the promoter region in comparison to MG1655 and O157. Strain O157 EDL933 carries around 15bp variations and strain Sakai carries around 19bp sequence variations in comparison to MG1655. No sequence changes were seen in the -10 and -35 box (Fig. 35). In order to analyze whether the polymorphism seen in the four sequenced strains are also present in the natural isolates of *E.coli*, the nucleotide sequencing of the *lac* promoter region (shown in Fig. 35) in all the isolates is currently under progress (the sequencing part of the work is in collaboration with Prof. Diethard Tautz).

IV. Discussion

Strains of *Escherichia coli* differ significantly in their genome sizes. Considerable insight has been gained into the role of 'pathogenicity islands' and 'genomic islands' for these differences. In this study, the DNA polymorphisms at two of the regions that show variations between the four sequenced *E.coli* strains have been assessed in 171 commensal and pathogenic *E.coli* isolates: *bgl*/Z5211-Z5214 region and c1955-c1960 region. The *bgl*/Z5211-Z5214 region and c1955-c1960 region are genomic islands in *E.coli*. In addition to the analysis of the two loci, lactose utilization phenotypes were assessed in all the isolates and the *lac* locus of the Lac negative strains were analyzed. Based on the observations at the *bgl*/Z5211-Z5214 region strains were typed into five main types and one sub type. It was observed that the typing of the *E.coli* strains at the *bgl*/Z5211-Z5214 locus correlated as well with the analysis of the other two loci. Taken together, this study addresses a possible method of typing *E.coli* strains at *bgl*/Z5211-Z5214 locus.

1. Genome variations at three loci in *E.coli* isolates

In the present work correlations for the five following features were observed: *bgl*/Z5211-Z5214 typing, c1955-c1960 system analysis, *lac* operon analysis, β-glucoside utilization phenotypes and phylogenetic distribution of ECOR strains (Fig. 36).

Out of 171 strains analyzed 33 strains have MG1655 *bgl/Z* locus, *lac* operon and none of them have c1955-c1960 system. All of the ECOR strains that have this structure belong to phylogenetic group A. 45 of the strains have CFT073 *bgl*/Z locus, c1955-c1960 system and *lac* operon. 8 out of 11 ECOR strains that show relaxed phenotype have this structure and are predominant in phylogenetic group B2.

Strains that have O157 *bgl/Z* locus are predominant in group D. Six out of 9 strains that are Lac^- have O157 *bgl*/Z locus and all of them lack c1955-c1960 system indicating a high rate of genetic variability at the region of carbohydrate utilization systems in this group of strains. Strains that have fourth *bgl*/Z locus are predominant in group B1, where 21 of them have both *lac* operon and c1955-c1960 system and 13 have *lac* operon but lacks c1955-c1960 system. Out of 20 strains that have fifth *bgl*/Z locus, 12 of them have both *lac* operon and c1955-c1960 system and 8 of them lacks c1955-c1960 system. Out of 5 strains that have mixed *bgl*/Z locus, 2 strains have both *lac* operon and c1955-c1960 system and 3 strains have *lac* operon but lacks c1955-c1960 system. The occurrence of the mixed *bgl*/Z type strains in the natural population of *E.coli*

may provide an insight in understanding the bacterial genome dynamics. In addition, it provides an evidence for the role of horizontal gene transfer in shaping the bacterial genome evolution.

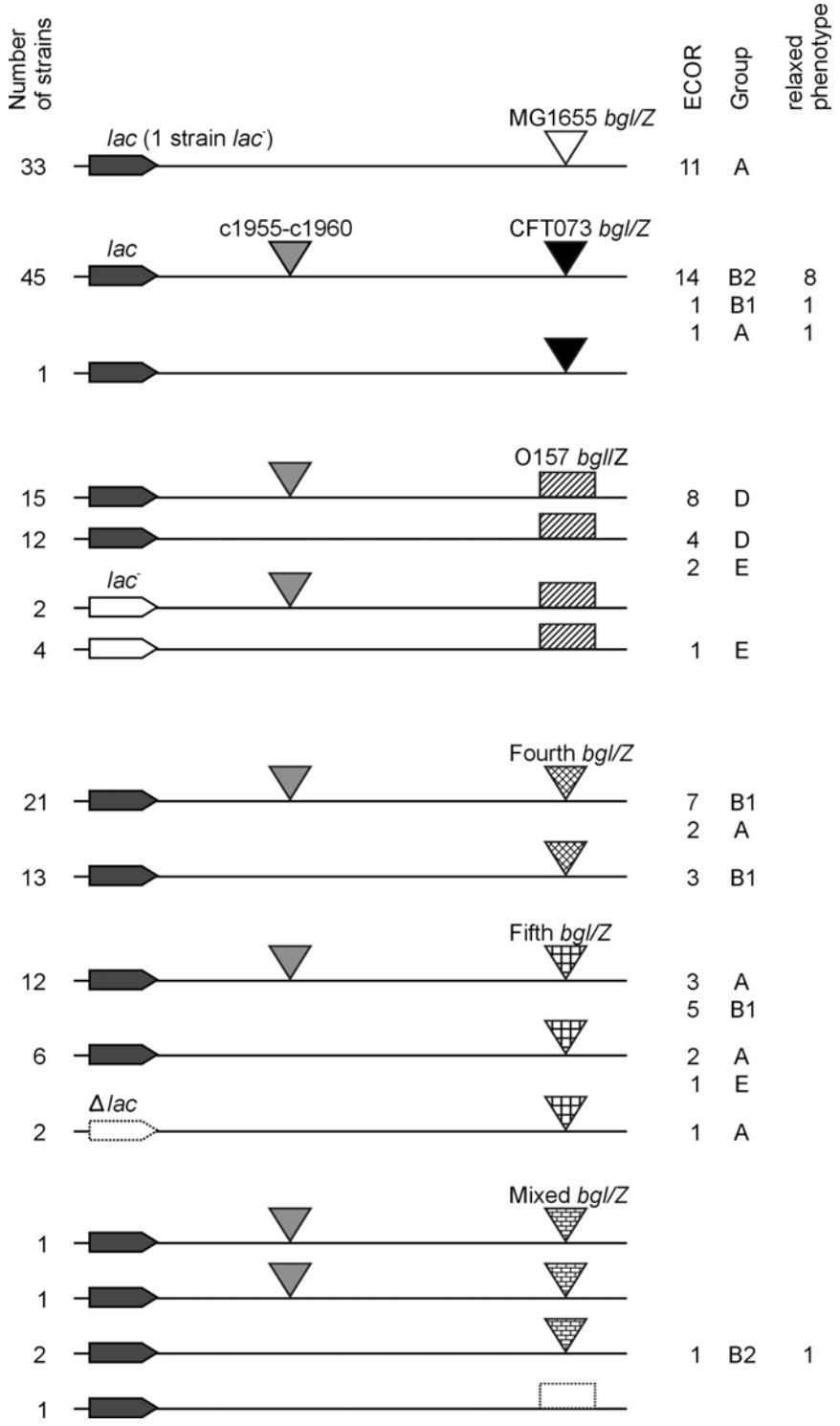


Figure 36: Schematic representation of the three loci analyzed in 171 strains. Horizontal lines represent the linear chromosomes of the 171 strains analyzed in this study. Positions of the three loci are shown not to scale. Typing at the *bgl*/Z5211-Z5214 locus (MG1655 *bgl*/Z, CFT073 *bgl*/Z, O157 *bgl*/Z, Fourth *bgl*/Z, Fifth *bgl*/Z and Mixed *bgl*/Z) is indicated. c1955-c1960 indicates c1955-c1960 system. Total numbers of strains in each type are shown. Number of ECOR strains and its phylogenetic grouping (Herzer et al., 1990) are indicated. Numbers of ECOR strains that show relaxed phenotype on BTB salicin plates at 37°C are also shown.

2. Structure of the *bgl*/Z5211-Z5214 locus in *E.coli*

The *bgl*/Z5211-Z5214 region may represent a good model to study the process of genomic island evolution in *E.coli* (Fig. 4, see introduction). The results from the current analysis reveal that five main and one sub *bgl*/Z types of *E.coli* strains may exists in the natural population. It was observed that, out of 171 strains approximately 20% of the strains have *bgl* locus like MG1655, 26% have *bgl* locus like CFT073, 20% of the strains have Z5211-Z5214 locus like O157, 20% have upstream sequence like O157 followed by the *bgl* and downstream like MG1655, 11% have upstream, *bgl* and downstream like MG1655 with hybrid *yieI* gene. In addition, 3% of the strains have a mixture of sequences from MG1655, CFT073 and O157 in *bgl*/Z5211-Z5214 locus which is assigned as mixed type in the current study (Fig. 8, Results). The grouping of the *E.coli* strains in the current study supports the evolutionary model proposed by Herbelin and co-workers (2000) and reflects the hypothesis that in the natural population four to five types of *E.coli* strains may exist. However, existence of the variants from these types cannot be ruled out (for e.g., mixed type strains in the current study). Furthermore, the random mutagenesis screen in a septicemic strain i484 that has CFT073 type *bgl* revealed that strain i484 is similar to CFT073 in the other loci as well, supporting the hypothesis in considering *bgl*/Z5211-Z5214 locus as a possible marker for typing *E.coli* isolates.

It was observed that the upstream *phoU* gene is conserved in all the isolates analyzed. However, the downstream region is variable as seen in the four sequenced *E.coli* strains (Fig. 8, Results). *phoU* is an essential gene in *E.coli* and mutations in *phoU* results in severe growth defects and alterations in phosphate levels (Steed & Wanner, 1993). The current study also revealed that in the natural population of *E.coli* approximately 43% of the strains carry hybrid *yieI* gene. The implication of this observation is yet to be understood. Furthermore, the alterations seen within the *bgl*/Z5211-Z5214 locus by insertions and deletions provides a model for how genes are lost from a genome once they no longer provide any selective advantage. In addition, it was observed that all of the CFT073 *bgl*/Z type strains and some strains in fourth *bgl*/Z type and mixed *bgl*/Z type do not carry *yieJ* gene. The absence of *yieJ* gene may possibly suggest the formation of "black holes" i.e., deletions of genes that may enable commensal bacteria to evolve toward a pathogenic life style. The gene product of the *yieJ* gene is currently unknown. The formation of "black holes" is observed in the deletion of *cadA* gene that encodes for Lysine decarboxylase that catalyzes the reaction for the formation of cadaverine which acts as an inhibitor of *Shigella* enterotoxin activity (Maurelli et al., 1998). Thus, it suggests that the creation

of black holes is a pathway that complements gene acquisition in the evolution of bacterial pathogens.

The sequence analysis and PCR-based studies used in this study could be further explored in designing an optimal PCR based strategy that would lead to a more convenient and easy approach to type *E.coli* at *bgl*/Z5211-Z5214 locus. The availability of genome sequences of four *E.coli* strains has provided wealth of information in designing strain specific primers for the optimal PCR strategies. Which could be further explored in analyzing the variations at other loci that might give an insight into a novel PCR based strategy for typing *E.coli*.

3. Silencing of the *bgl* operon is conserved

The current analysis has provided extensive evidence for the β-glucoside utilization phenotypes of the natural *E.coli* isolates from diverse sources. The observations revealed that silencing of the *bgl* operon is conserved in *E.coli*. Three different kinds of papillation phenotypes were seen, where majority of the strains papillated like MG1655 or more frequently than MG1655. However, around 15% of the strains showed a weak Bgl^+ phenotype (relaxed) (Fig. 9). Strains that showed relaxed phenotype can papillate only at 28°C, suggesting a temperature sensitive regulation of β-glucoside utilization in this group of strains.

Furthermore, the pattern of activation of the *bgl* operon seen in the representative strains of MG1655 like phenotype, more papillae phenotype and relaxed phenotype demonstrated that the *bgl* operon is spontaneously activated in the natural *E.coli* isolates as in MG1655 (Fig. 16, Results). However, difference in the activation varies from strain to strain. The basis for these differences between the strains is not known at this stage; it could be speculated that this could be because of the strain to strain background and may depend on the cellular pleiotropic control of the *bgl* operon or of transposition.

4. Sequence variations in the *bgl* operon have no significant influence on the *bgl* expression in K-12 background

The observation of the weak Bgl^+ phenotype (relaxed phenotype) in 15% of the strains in the laboratory condition suggests that the basal expression of the *bgl* operon could be high in this group of strains. In addition, it was observed that all the strains that showed relaxed phenotype carried CFT073 type *bgl* (Table 1, Results). The analysis of the expression levels of P_{bgl}-*lacZ* and P_{bgl}-t1-*bglG-lacZ* fusions in K-12 background showed that the sequence variations in the *bgl* promoter region seen in CFT073 *bgl*/Z type do not have significant influence on the promoter

activity. However, constructs carrying fragments with additional base pair changes show significant differences in comparison to MG1655 indicating the importance of the sequence change(s) on *bgl* expression (Fig. 18 and 19, Results). At this stage, it cannot be concluded that the sequence variations seen in the CFT073 *bgl* type may not have an impact on the *bgl* expression. It might be possible that in an environment inside the host or in an untested strain background the sequence variations might play a role in the *bgl* expression.

5. Positive regulatory factors necessary for relaxed phenotype

To ensure whether the relaxed phenotype is not due to the multiple papillation, we tried purifying the papillae from day 3 incubated BTB salicin plates at 37°C. However, all the attempts in purifying complete Bgl^+ mutants were unsuccessful. A mTn10 mutagenesis screen to unravel the factors that are essential for relaxed phenotype in wt-i484 yielded mutations in *bgl* operon, *cysN*, *cysH*, *cysJ*, *purL*, *purC* and *ybdM* (Fig. 21). Mutations in *bgl* operon hinder the utilization of β-glucosides. Mutations in *cysN*, *cysH*, and *cysJ* affect sulfate assimilation and consecutively affect the biosynthesis of cysteine and methionine. Mutations in *purC* and *purL* affect the *de novo* biosynthesis of purine nucleotides. The second round of mTn10 mutagenesis yielded mutations in 5 new genes: *pgi*, *cysG*, *serC*, *carB* and *rfaC* (Fig. 22). It was shown that *bgl* operon could be specifically downregulated by *pgi* mutation (Dole et al., 2004). Mutations in *cysG*, *serC*, *carB* and *rfaC* have general response on the cellular metabolism by affecting the sulfate assimilation, amino acid biosynthesis, ribonucleotides synthesis and lipopolysaccharide biosynthesis respectively.

The two rounds of screen basically yielded mutations in the genes that are involved in amino acid biosynthesis or nucleotide biosynthesis. Mutations in these genes may lead to the alterations in the nucleotide pool and/or amino acid pool of the bacterial cell. This could have a potential influence on the expression of various genes that may include regulators of the *bgl* operon silencing such as H-NS. Thus, the pleiotropic response during these events may have effect (through H-NS) on the β-glucoside utilization. The direct effect on the repression of *bgl* by H-NS also cannot be ruled out. The precise mechanism to the effect of mutations in these genes on the β-glucoside utilization remains to be determined.

On the other-hand, in both the independent miniTn10-mutagenesis screen mentioned above, mutation in the gene *ybdM* was obtained. The effect of the mutation in the hypothetical gene *ybdM* on β-glucoside utilization cannot be speculated at this stage. However, homology searches revealed that the *ybdM* may encode protein that is homologous to putative transcriptional

regulators from *E.coli, Salmonella, Bacillus,* and *Enterococcus* (see results section 1.15). To this end, it would be interesting to further characterize the role of hypothetical protein YbdM in the β-glucoside utilization.

6. c1955-c1960 locus in *E.coli*

The studies on natural isolates of *E.coli* have revealed that *E.coli* K-12 may possess additional systems for the utilization of β-glucoside sugars (Hall, 1988; Parker and Hall, 1988). Due to the remarkable ecotype diversity of *E.coli*, the presence of various β-glucoside systems may play a potential role that permits its adaptation to a novel environment in which β-glucosides are the sole carbon source. In this study, the isolation and characterization of an additional β-glucoside specific locus from *E.coli* is reported.

Some of the strains that carry Z5211-Z5214 locus in place of *bgl* showed late papillation on BTB salicin plates at 37°C. This papillation phenotype suggested an existence of additional β-system which is activated upon prolonged incubation in the presence of β-glucosides. A miniTn10-cm^R mutagenesis screen to unravel the presence of additional β glucoside-system primarily yielded mutations in c1955-c1960 locus of the published genome sequence of *E.coli* CFT073 (Fig. 23).

The c1955-c1960 represents an island encoded system that is present uniquely in the CFT073 genome and absent in the other three sequenced *E.coli* strains. The analysis revealed that 57% of the strains have c1955-c1960 region like CFT073 and the remaining 43% do not carry this locus (Fig. 26). The data also showed that c1955-c1960 system is predominant in the strains that have CFT073 type *bgl* region. None of the MG1655 *bgl*/Z type strains carried c1955-c1960 system (Fig. 26). Furthermore, the nucleotide sequence alignment of the upstream (*marB*) and downstream (*ydeD*) regions of c1955-c1960 in the four sequenced *E.coli* strains showed sequence variations between the strains (data not shown). Typing of *E.coli* isolates at c1955-c1960 locus might give correlations with the typing at *bgl*/Z5211-Z5214 locus which may lead in elucidating novel typing methods for *E.coli* strains.

7. c1955-c1960 system encodes genes for β-glucosides utilization

Homology searches of the deduced amino acid sequences of the genes in the c1955-c1959 region showed significant similarity to the proteins that are involved in the β-glucoside utilization in diverse bacteria (Fig. 24, Results). The deduced amino acid sequences of c1960 showed

significant similarity with GntR family of transcriptional regulators from *Bacillus*, *Listeria*, *Erwinia* and *E.coli*.

In an attempt to analyze the potential functions of the proteins encoded by this locus, phenotypes were determined in the presence of various β-glucosides. The analysis revealed that the c1955-c1960 encode genes that can hydrolyze four β-glucosides (Salicin, Arbutin, Cellobiose and Esculin) at 28°C, supporting the assignment of the putative gene products traced from the blast searches (Fig. 24). The results were further strengthened by the complementation of the c1955-c1960 region in i484 Δ*bgl* background (Fig. 28). Moreover, the constitutive expression of the putative structural genes (c1955-c1960) from wt-strain (i484 Δ*bgl*) complemented the utilization of three of the tested sugars. Furthermore, it was observed that c1955-c1960 genes cannot complement in K-12 background. This data suggested that in addition to c1955-c1960, additional factors that are absent in K-12 are required for the utilization of β-glucosides.

8. Regulation of c1955-c1960 system

Nucleotide sequencing of the putative promoter region in between the genes c1959-c1960 revealed that all the four spontaneous mutants carried identical point mutations (Fig. 27). The mutation was seen to be in the non-conserved nucleotide of a putative CAP binding site. Along with, a putative -10 box that consists of conserved nucleotides essential for RNA polymerase binding was mapped. Thus, this region looks like a typical Class II CAP dependent promoter, where the -35 box is replaced with a CAP-binding site.

Expression of $P_{c1955\text{-}c1960}$-*lacZ* fusions showed that the mutation seems to be involved in the activation of c1955-c1960 system (Fig. 29). The two fold repression in the promoter activity noticed in the wt-construct suggested a repressor activity for c1960. This repression was abolished in the constructs that carry mutation in the putative CAP binding site. The promoter activity was two to three fold repressed in the presence of glucose, suggesting a carbon catabolite repression by primary sugars. In addition, the results from Δ*cyaA* and *crp* backgrounds reveal that the promoter acitivity is CAP dependent. The protein H-NS is involved in the silencing of *E.coli bgl* operon at multiple levels (Dole et al., 2002; Dole et al., 2004) however, no such influence of H-NS was seen in the tested c1955-c1960 constructs. To this end, the data suggests that the mutation in the putative regulatory region might be responsible for the activation of the c1955-c1960 system.

Furthermore, it was observed that salicin, cellobiose, arbutin and esculin are not inducers of c1955-c1960 system in the wt-K-12 background (Fig. 31). However, it was observed that the promoter activity of the wt-construct carrying c1960 was induced with salicin in K-12 background that carry activated copy of the *bgl* operon (Fig. 32). Similar observations were seen from i484 Δ*bgl* background that carry activated copy of c1955-c1960 system (Fig. 30). Taken together, the data implies that transport and/or phosphorylation of β-glucosides is necessary for induction of c1955-c1960 system. It also suggests that upon transport of the β-glucosides c1960 could possibly act as an activator of the c1955-c1960 system. However, whether the intake of phosphorylated β-glucosides or the degradation products of the β-glucosides is responsible for this activation remains to be answered.

9. Correlations of *bgl*/Z5211-Z5214 typing with other carbohydrate utilizing systems

Most studies on the utilization of carbon sources by *E.coli* date back before the unraveling of its complex ecotype diversity. A reassessment of some characteristics within the updated genomic context appeared warranted. Lactose is abundantly found in nature and assimilation of lactose is an important feature of enteric bacteria. *E.coli* and *Klebsiella pneumoniae* can utilize lactose as a carbon source. Where as, the other enteric bacteria like *Shigella*, *Salmonella*, *Serratia* and *Yersinia* cannot utilize lactose. Thus, lactose utilization can be attributed as a phenotypic characteristic that distinguishes *E.coli* and *Klebsiella* from other enteric bacteria (Ochman et al., 2000). Lawrence and Ochman (1998) have pointed out that the species specific traits such as lactose utilization are derived from the functions encoded by horizontally transferred genes. To test whether the heterogeneity seen at the *bgl*/Z5211-Z5214 locus and c1955-c1960 locus have correlations with the other carbon utilizing systems, studies on *lac* operon were performed. The data showed that 162 out of 171 strains can utilize lactose. The higher percentage of Lac^+ strains in the natural population of *E.coli* suggests that *lac* operon is highly conserved in its functional state (Fig. 34).

Furthermore, a strong correlation of the lactose utilization phenotypes with the *bgl*/Z5211-Z5214 typing was observed. Out of nine strains that are Lac^-, six of them possibly carry mutations in *lacY* gene that result in functional inactivation of the permease (Fig. 34). Moreover, all these six strains belong to O157 type (at *bgl*/Z5211-Z5214 locus). This suggests that O157 type (at *bgl*/Z5211-Z5214) strains might be with a high level of genetic variability that is required for the higher probability of its survival in a constantly changing environment.

10. Outlook

The current analysis might give an insight in answering the rapid bacterial evolution which is concerned with some of the relevant issues of current studies. Firstly, what types and how many types of *E.coli* strains are present in the natural population? Secondly, is there any correlation for the presence of different loci in the genome of *E.coli* to its habitat? Thirdly, where do the genomic islands come from and by what mechanisms are they transferred? Finally, how is it possible to identify cases of horizontal gene transfer?

The report from Kilic and coworkers (2004) suggested a unique role for β-glucoside utilization systems in the maintenance of pathogenic life style of a bacterium. Further studies on c1955-c1960 system or *bgl* on these aspects would give an insight in understanding whether the β-glucoside utilization systems such as *bgl* or c1955-c1960 system contributes to the virulence of *E.coli*.

In summary, the present study highlights the genetic diversity at the *bgl*/Z5211-Z5214 region and c1955-c1960 region among pathogenic and commensal *E.coli* isolates. The correlations seen with the *bgl*/Z5211-Z5214 locus typing, c1955-c1960 system analysis, *lac* operon analysis, β-glucoside utilization phenotypes and phylogenetic distribution of ECOR strains might implicate *bgl*/Z5211-Z5214 locus as a possible marker for typing *E.coli* strains. In addition, the study supports the hypothesis that in nature five main types of *E.coli* strains may exist. Furthermore, the genetic variations seen at the three loci provide evidence for frequent acquisition of foreign DNA through horizontal gene transfer events, as well as for the deletions within the genomes during the course of bacterial evolution.

V Materials and Methods

1. Chemicals, enzymes and other materials

Chemicals and enzymes were purchased from commercial sources. Oligonuleotides were purchased from Invitrogen life technologies (Karlsruhe, Germany) or Eurogentec.

2. Media and agar plates

LB medium (1 l)	10g	Bacto Tryptone (Difco)
	5g	Yeast-extract (Difco)
	5g	NaCl
		(for plates 15g Bacto Agar, Difco)
NB medium (1 l)	8g	Bacto NB broth, dehydrated (Difco)
		(3g Bacto Beef extract, 5g Bacto peptone)
SOC medium (1 l)	prepare SOB in 970ml H_2O:	
		prepare SOB in 970ml H_2O
	20g	Bacto Tryptone (Difco)
	5g	Yeast-extract (Difco)
	0.5g	NaCl
	1.25ml	2M KCl
		adjust pH 7.0 with NaOH
		autoclave, add 10ml 1M $MgCl_2$
		for SOC add 19.8ml 20% glucose
MacConkey lactose indicator plates (1 l)	50g	MacConkey agar with lactose (Difco)
BTB indicator plates (Schaefler, 1967)	15g	Bacto Agar (Difco)
	1g	Yeast-extract (Difco)
	1g	Tryptone (Difco)
	5g	NaCl
		add 900ml H_2O, autoclave
		add sterile:
	1ml	1M $MgSO_4$
	1ml	0.1M $CaCl_2$
	1ml	Vitamin B1 (stock solution 1mg/ml, filter sterilized)
	0.5ml	1mM $FeCl_3$
	20ml	10% (w/v) Casaminoacids
	50ml	10% (w/v) (Salicin or Cellobiose or Arbutin or Esculin, filter sterilized)
	10ml	Bromthymol blue stock solution (2% BTB in 50% ethanol, 0.1N NaOH)
		adjust colour with NaOH

M9 Medium
(Miller, 1972)

prepare 20X M9 (stock solution):

140g	Na_2HPO_4 x 2 H_2O
60g	KH_2PO_4
20g	NH_4Cl
	H_2O to 1 l

M9 medium (prepare from sterile solutions):

50ml	20x M9
1ml	0.1M $CaCl_2$
1ml	1M $MgSO_4$
0.5ml	1mM $FeCl_3$
1ml	Vitamin B1 (stock solution 1mg/ml)
66ml	10% Casamino acids

carbon source 0.5% final concentration:

25ml	20% Glucose
or 6.25ml	80% Glycerol
	H_2O to 1 l

3. Antibiotics

Antibiotics	stock	storage	final conc.
Ampicillin	50mg/ml in 50% ethanol	-20°C	50µg/ml
Chloramphenicol	30mg/ml in ethanol	-20°C	15µg/ml
Kanamycin	10mg/ml in H_2O	+4°C	25µg/ml
Spectinomycin	50mg/ml in 30% ethanol	-20°C	50µg/ml
Tetracyclin	5mg/ml in 70% ethanol	-20°C	12µg/ml

4. General Methods

The molecular biology methods like restriction enzyme digestions, ligations and other enzyme reactions, PCR, plasmid DNA purification, auto-radiography were performed as described (Sambrook et al., 2001; Sambrook et al., 1989) or according to the manufacturer instructions, unless otherwise stated. Large scale preparations of plasmid DNA were performed using the plasmid maxi kit (Qiagen or Promega) according to manufacturer instructions.

5. *E.coli* isolates and growth conditions

The *E.coli* K-12 strains used in this study are listed in Table 5 (Appendix). 98 clinical *E.coli* isolates comprising of 52 commensals, 22 uropathogenic and 24 septicemic were obtained from Dr. Georg Plum, Institute for Medical Microbiology, University of Cologne, Germany (Table 6a, Appendix). All these 98 isolates are isolated from human sources and were obtained from the local hospitals. Septicemic strains are isolates from the blood of the patients with septicemia. Uropathogenic strains are isolates from urine samples of patients with urinary tract infections and commensal strains are isolates from the stool samples of healthy individuals. The septicemic strain *E.coli* i484 was a kind gift from Dr. Richard E. Isaacson, Department of Veterinary Pathobiology, University of Illinois, Urbana, Illinois, USA (Table 6a, Appendix). The 72 ECOR (*E.coli* reference collection) strains that comprises of 29 commensals from humans, 32 commensals from animals (Zoo and domesticated), 10 uropathogenic and 1 asysmptomatic bacteriuria strain were obtained from STEC (Shiga Toxin-producing *Escherichia coli*) Center, Michigan State University, MI, USA (Table 6b, Appendix). Strains were grown in Luria-Bertani (LB) liquid/agar medium (Miller, 1972) and preserved as DMSO stocks (1.5ml of overnight grown culture in LB medium + 50µl DMSO) at -80°C. Where necessary, antibiotics were added to final concentrations as mentioned (see materials and methods 3).

6. PCR analysis of the *bgl*/Z5211-Z5214 locus in *E.coli* isolates

Oligonucleotides used in this study are listed in Table 4 (Appendix). To analyze the genetic diversity of the *E.coli* isolates at the *bgl*/Z5211-Z5214 locus, a PCR strategy was designed based upon the use of the published MG1655, CFT073, O157-EDL933 and Sakai genome sequences. The mapping positions of the PCR primers used in the *bgl*/Z5211-Z5214 locus analysis are schematically shown in Figure 37. PCR reactions were performed either with single colony suspension or with the plasmid constructs of the respective strains as templates (Table 8,

Appendix). The strains were analyzed by PCR with various *bgl* specific primers to see whether they carry the *bgl* operon. In the strains that carry the *bgl* operon the sequence of the upstream region (*phoU* gene), *bgl* promoter region and the downstream region (*yieJIH*) was determined. To analyze whether the strains carry an intact *bgl* operon long PCR (using Elongase, Invitrogen, Germany) with oligos mapping in the *phoU* and *yieH* genes was performed (Table 4, Appendix). PCR products were analyzed in comparison to the expected PCR product sizes in MG1655 and i484 (like CFT073). Strains that do not carry the *bgl* operon were analyzed by PCR with Z5211-Z5214 specific primers (strains in Table 6b, Appendix) or with ST-PCR (strains in Table 6a, Appendix). ST-PCR products were sequenced. The presence of intact Z5211-Z5214 region in all the strains was analyzed by long PCR.

The strains that showed differences in the long PCRs in comparison to the expected PCR product sizes of MG1655, CFT073 and O157 were analyzed further with several internal PCRs to map the possible deletion or insertion points. An example of such an analysis is shown in Figure 14 (Results). The relative PCR products (e.g., as described in Fig. 14, Results) were sequenced from both the ends. Details of the PCR analysis are shown in Table 6 (Appendix).

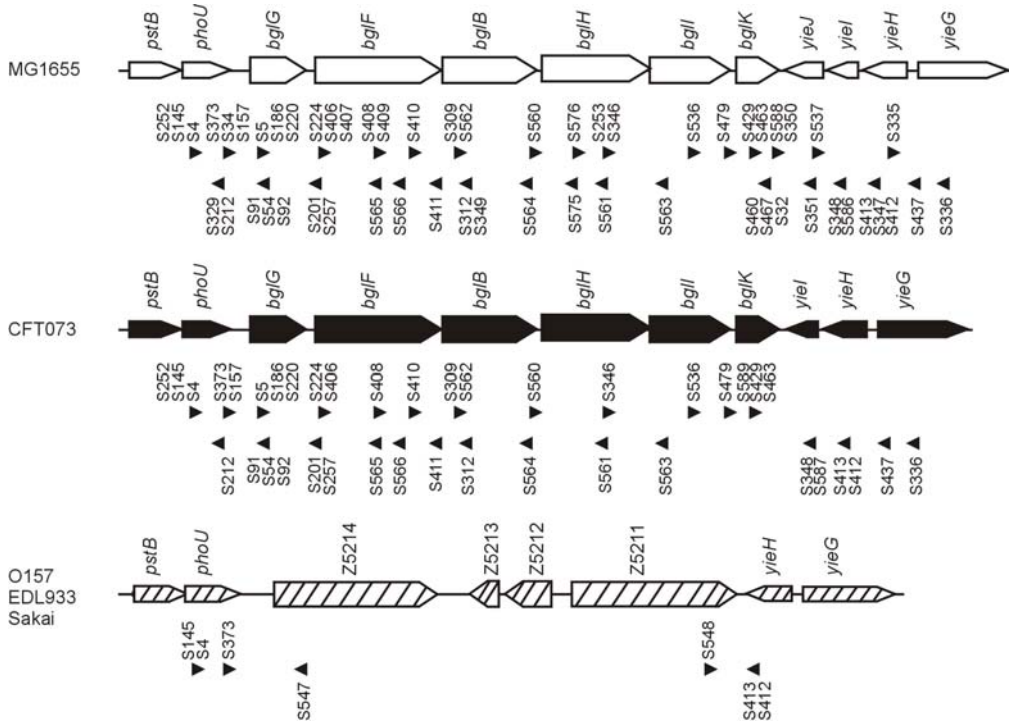


Figure 37: Schematic presentation of the oligos mapping positions in MG1655, CFT073 *bgl* region and in O157 Z5211-Z5214 region. Oligos are shown with black arrows and arrowhead indicates the direction of the oligo from 5' to 3'. MG1655, CFT073 and O157 *bgl*/Z5211-Z5214 region is shown as in Figure 4 (Introduction).

7. ST-PCR (Semi-Random PCR)

ST-PCR is a semi-random two step PCR method that involves two successive PCR reactions and two pairs of PCR primers. This method is simple and very specific to identify the novel DNA sequences next to previously known sequences (Chun et al., 1997). Chun and co-workers (1997) have used ST-PCR to identify genes that are mutated in miniTn3 mutagenesis. In brief, in the first reaction, one primer anneals to the end of the transposable element, while the other contains a specific 20-nucleotide sequence followed by ten bases of degenerate sequence and a specific five-nucleotide sequence. A subset of these degenerate primers anneals to an unknown DNA sequence near the transposable element and allows initial amplification. The second pair of primers anneals to specific DNA sequences, one from the transposable element and the other from the 20-nucleotide sequence in the semi-random primer resulting in the amplification of a specific PCR product.

In the current study ST-PCR was employed for two reasons: a) to identify the genes that are mutated in the miniTn10-cm^R mutagenesis (also see 8, materials and methods) and b) to identify the presence of Z5214-Z5211 in the *bgl*/Z5214-Z5211 locus. Oligos used for ST-PCR are listed in Table 4 (Appendix). The ST-PCR approach to map the miniTn10-cm^R insertions is schematically shown in Figure 38. In brief, the first round of PCR reaction (PCR1) was performed in a 20µl reaction volume with a combination of degenerate primer S360 composed of 20 bases of defined sequence, followed by ten random bases, followed by the bases GATC (four-nucleotide sequence GATC was used because of its higher frequency in the *E.coli* genome) (modified from (Chun et al., 1997) and with primer S357 (or S358) that maps in the miniTn10-cm^R cassette. PCR1 reaction mix was then diluted to 1:5 times with water to 100µl. 1µl from the diluted mix was used as template in the second round of PCR reaction (PCR2) performed in a 20µl reaction volume with S361 (anneals to the complement of the 20 bases of defined sequence at the 5' end of primer S360) and S359 (nested primer in the miniTn10-cm^R cassette). The obtained PCR products were later sequenced with S359.

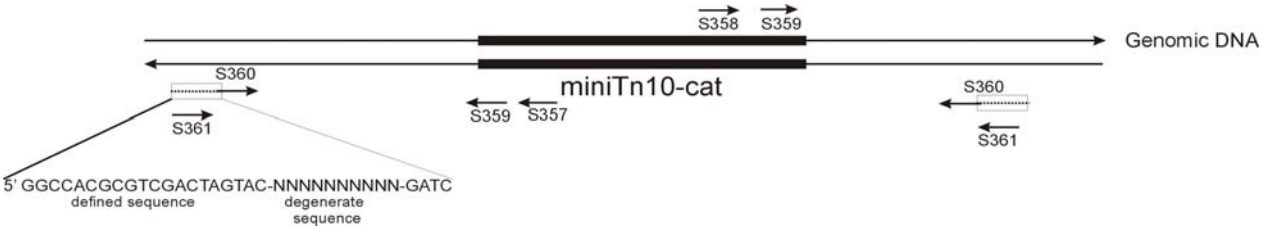


Figure 38: Summary of ST-PCR (modified from Chun et al., 1997). The miniTn10-cm^R insertions in the chromosome can be screened by ST-PCR. Briefly, PCR 1 is performed with S357/S360 or S358/S360 followed by PCR2 reaction with S359/S361. The obtained PCR products were sequenced with S359.

The PCR conditions for PCR1 and PCR2 are as described in (Chun et al., 1997). To identify the presence of Z5211-Z5214 region, briefly, the first round of PCR reaction (PCR1) was performed with a combination of degenerate primer S360 and the specific primer that map in *phoU* (S145) or *yieH* (S412) gene. The second round of PCR reaction (PCR2) was performed with S361 and nested primers in *phoU* (S373) or *yieH* (S413). The obtained PCR products were later sequenced with primers mapping in *phoU* or *yieH* gene.

8. miniTn10-cm^R mutagenesis

The miniTn10-cm^R mutagenesis was performed with plasmid pKESK18 (pSC101 derivative). Plasmid pKESK18 is rep^{ts}, carries a miniTn10-cm^R transposon, and the Tn10 transposase gene driven by the phage λ pR promoter (Fig. 39). The transposase expression is repressed at 28°C by the temperature sensitive cI-857. However, at 42°C the cI-857 repressor is inactive. Hence, at 42°C the transposase is expressed, leading to the transposition of the miniTn10-cm^R transposon. Concomitantly, the plasmid is lost as it does not replicate owing to the Rep^{ts} protein.

day 1:

* transform the strain (to be mutagenised) with the plasmid pKESK18, select transformants at 28°C on LB kanamycin + chloramphenicol plates

day 2:

* pick single colony from the transformants and inoculate into 3ml LB with kanamycin and chloramphenicol. Grow overnight at 28°C

day 3:

* prepare 10^{-4}, 10^{-6} and 10^{-7} dilutions of the overnight culture to a final volume of 2 ml in sterile Mg-saline
* from 10^{-4} dilution, plate 100µl of the diluted cells on LB-chloramphenicol plates and incubate at 42°C (plates are pre warmed to 42°C before plating the cells). In parallel, for determining the cell titer, plate 200µl from 10^{-6} and 10^{-7} dilutions on LB kanamycin + chloramphenicol plates and incubate at 28°C
* replica plate the colonies on BTB Salicin-chloramphenicol plates and in parallel check for kan^s
* characterize the miniTn10-cm^R insertions by ST-PCR (Fig. 38, see materials and methods 7)

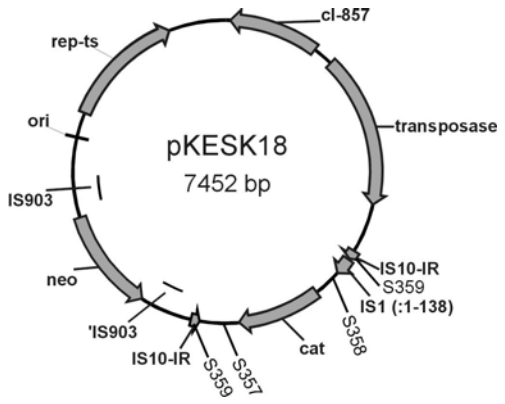


Figure 39: Schematic presentation of plasmid pKESK18. Refer to text for more details. Primers used in ST-PCR are shown as listed in Table 4 (Appendix).

9. DNA sequencing and sequence data analysis

DNA sequencing was performed with the Big dye terminator cycle sequencing kit (version 1.0 or 3.0, ABI Prism) according to the manufacturer's instructions and using an automated DNA sequencer. The details of the sequence information's are documented in the lab records. Homologues searches were performed with BLASTN (National Center for Biotechnology Information) and with Fasta3 (European Molecular Biology Laboratory). Nucleotide Sequence alignments and amino acid sequence alignments were performed using Vector NTI programme.

10. Statistical tests

Statistical correlations were calculated by Chi square tests and Fishers Exact test using the online resources available at http://www.georgetown.edu/faculty/ballc/webtools/web_chi.html and http://www.unc.edu/~preacher/fisher/fisher.htm respectively.

11. Preparation of competent cells and transformation ($CaCl_2$ method)

TEN buffer: 20mM Tris-HCl pH 7.5, 1mM EDTA, 50mM NaCl

* grow cells in 25ml LB medium till OD_{600}=0.3
* centrifuge at 3000rpm and re-suspend the cell pellet in 12.5 ml ice cold 0.1M $CaCl_2$
* incubate on ice for 20 minutes
* centrifuge again and re-suspend pellet in 1ml ice cold 0.1M $CaCl_2$
* use 100µl of these cells for transformation
* add 1-10ng of the plasmid or 10µl (=1/2) of the ligation mix to be transformed into the pre-cooled eppendorf tube and make up the volume to 50µl with TEN buffer. Cool the mix on ice.
* add 100µl of competent cells and incubate on ice for 20 minutes
* heat-shock at 42°C for exactly 2 minutes
* incubate on ice for 10 minutes
* add 1ml LB medium and shake at 37°C for 1 hour
* plate 100µl on suitable selective plates

12. Preparation of electrocompetent cells and electroporation

* inoculate 3ml LB with a single (fresh O/N) colony. Shake at 37°C overnight
* from overnight grown culture inoculate 200µl to 50ml fresh LB medium in 250ml conical flask
* grow cells (37°C in a shaker) to an OD_{600} of 0.6-0.7
* place the flask with the culture on ice for 1hr
* transfer culture to sterile, prechilled centrifuge tubes (50ml Blue capped tubes)
* centrifuge the cells at 3000rpm for 15 minutes at 4°C. Decant the supernatant.
* re-suspend the pellet in approximately 50ml of cold (0-4°C) sterile water

* centrifuge cells at 3000rpm for 15min at 4°C and decant the water carefully without disturbing the pellet. Place tubes back on ice
* re-suspend the pellet in approximately 25ml of cold (0-4°C) sterile water
* centrifuge cells at 3000rpm at 4°C 15min. Decant the water and place tubes back on ice.
* using glass pipette re-suspend cells in 2ml cold sterile 10% glycerol.
* centrifuge cells at 6000rpm for 15min at 0-4°C. Decant the supernatant and place tubes back on ice.
* re-suspend pellet in 200µl cold sterile 10%glycerol.
* take 40µl from this for single transformation, alternatively cells can be frozen and stored at -80°C (see below)
* incubate cells for at least 1 additional hour on ice
* make aliquots: 40µl in ice-cold eppendorf tubes
* freeze in liquid N_2, store at -80°C

Electroporation:

* prechill the electrocuvettes on ice
* if case of frozen electrocompetent cells: thaw cells slowly on ice
* prepare Eppendorf-tube on ice with plasmid DNA (free from salts)
* add 40µl of competent cells
* incubate for 10minutes
* transfer the mixture (DNA + Competent cells) into the prechilled Electrocuvettes near Electroporator
* electroshock at 1.8kv for 3seconds
* remove the Electrocuvette immediately and add 1ml of SOC medium as quickly as possible into the cuvette
* transfer the medium from the cuvette into the test tube and incubate for 1hr at 37°C
* plate 100µl on suitable selective plates

13. Plasmids and DNA fragments

A brief description of plasmid constructions can be found in Table 8 (Appendix). Details of plasmid constructions are documented in the lab records and sequences compiled in Vector NTi.

A series of plasmids starting with pKEGN1 were constructed which have a pACYC (p15A) replication origin (Fig.40). The starting material was a similar plasmid with MG1655 *bgl* (pKES15). They have λ phage attachment site (*attP*) cloned into them that allows the λ integrase mediated recombinational insertion into the chromosomal *attB* site of *E.coli* (Diederich et al., 1992). These plasmids also have a Ω cassette which contains the Spectinomycin resistance gene, *aadA* and strong transcriptional terminators at its 3'end. As seen in Figure 40, plasmid pKEGN1 has convenient restriction enzyme sites which could be used to replace the *bgl* construct with a different construct. For integrations into the chromosomal *attB* site, these plasmids were cut with *BamHI* (in some cases *BglII*) and the originless fragment containing the Spectinomycin resistance gene and the gene(s) of interest was used for insertion into the *attB* site as described (see materials and methods, 14).

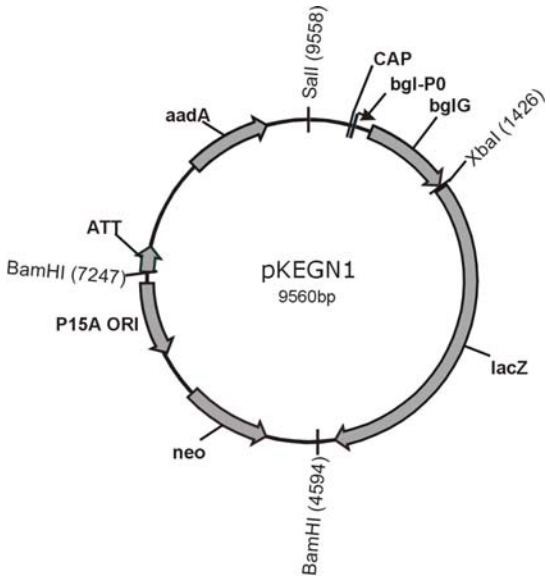


Figure 40: Schematic representation of plasmid pKEGN1. The starting material for this plasmid was a similar plasmid with MG1655 *bgl*. This plasmid and later plasmids based on the same principle carry a λ phage attachment site attP for integration into the E.coli genome. Resistance markers for Spectinomycin (aadA) and Kanamycin (neo) are shown. Plasmid pKEGN1 has E10085-P_{bgl}-t1-*bglG-lacZ* construct. This plasmid can be conveniently replaced with a different constructs to generate a series of plasmids used in the study. Restriction sites used for cloning (SalI-XbaI) are indicated.

14. Integration of plasmids in the *attB* site of *E.coli* chromosome (Diederich et al., 1992; Dole et al., 2002)

* transform the recipient strain with helper plasmid pLDR8 and select on LB Kanamycin plates at 28°C
* inoculate 3-4ml LB Kanamycin medium in glass tubes from a fresh single colony of the strain/pLDR8-plates
* shake culture over night at 28°C
* inoculate 25ml LB Kanamycin medium in 100ml Erlenmeyer flask with 1.25ml (1:20 dilution) of the fresh over night culture
* shake culture at 37°C for 90 minutes
* prepare competent cells using the $CaCl_2$ method
* cut 5µg of the plasmid containing the construct to be integrated with *BamHI* or *BglII* (10U per 100µl) incubate overnight at 37°C
* run on agarose gel and extract the originless fragment using Qiagen gel extraction kit
* use 10ng *BamHI* (or *BglII*) fragment for re-ligation in 20µl total volume
* transform the competent cells prepared as above using 10µl (=1/2) of *BamHI* (or *BglII*) re-ligation reaction
* plate: 2 x 0.2ml on LB Spectinomycin plates and if appropriate 1x 0.2ml on MacConkey lactose Spectinomycin plates
* incubate plates at 42°C over night

testing of clones via PCR:

* pick a colony and resuspend the cells in 100µl H_2O
* use the PCR-primer: S93/S164: to test the attB/P`-side
  S95/S96: to test the attP/B`-side
  S95/S164: to see integrations of dimers
  suitable primers to the test the cloned fragment

15. β-galactosidase assay (Miller, 1972)

Z-buffer	60mM Na_2HPO_4, 40mM NaH_2PO_4, 10mM KCl, 1mM $MgSO_4$, 100µg/ml Chloramphenicol, pH 7.0

Na_2CO_3	1M
ONPG	4 mg/ml in 0.1M phosphate buffer pH 7.0
SDS	0,1 % (w/v)
chloroform	

day 1:

* prepare overnight cultures (3-4 ml) in minimal media M9 with necessary supplements as required e.g. glycerol or other carbon sources, casamino acids, thiamine, antibiotics etc. or LB medium with necessary supplements as required

day2:

* measure the OD_{600} of the overnight cultures diluted 1:5
* inoculate 8 ml cultures to an OD_{600} of 0.1 in 100 ml conical flasks
* in case of induction add 1mM IPTG and/or 0.2% (w/v) Salicin (or Cellobiose or Arbutin or Esculin)
* grow cultures to OD_{600}=0.5 at 37°C
* harvest the cultures on ice
* to perform enzyme assay prepare appropriate 1ml dilutions from the cooled cultures in ice cold Z-buffer (at least 3 dilutions in duplicates were prepared)
* add 10µl 0.1% (w/v) SDS and 20µl chloroform to all dilutions
* set up blanks with 1 ml Z-buffer in four separate 2 ml Eppendorf-tubes
* vortex the probes for 15 seconds and then immediately pre-incubate them for 10 minutes at 28°C
* start the reaction by adding 0.2 ml ONPG with multi-pipette and mix
* stop the reaction after 30 minutes by adding 0.5 ml 1M Na_2CO_3 (or later, when the color turns to a strong yellow, note exact time)
* centrifuge the probes for 5 to 10 minutes at room-temperature and measure the OD_{420} calculate the enzyme activity in Miller units as:

$$1 \text{ unit} = \frac{OD_{420} \text{ x dilution factor x } 1000}{OD_{600} \text{ x time (minutes)}}$$

16. β-glucosidase assay

Z-buffer	60mM Na_2HPO_4, 40mM NaH_2PO_4, 10mM KCl, 1mM $MgSO_4$, 100µg/ml Chloramphenicol, pH 7.0
Na_2CO_3	1M
PNPG	8mg/ml in 0.1M phosphate buffer pH 7.0

day 1:

* grow bacteria in (3-4ml) in minimal media M9 medium with necessary supplements as required e.g.: glycerol, casamino acids, thiamine, antibiotics etc.

day 2:

* measure the OD_{600} of the overnight cultures diluted 1:5
* inoculate 8 ml cultures to an OD_{600} of 0.1 in 100 ml conical flasks
* for induction add 0.2% (w/v) Salicin
* grow cultures to OD_{600}=0.5 at 37°C
* harvest the cultures on ice
* to perform enzyme assay prepare appropriate 1ml dilutions from the cooled cultures in ice cold Z-buffer (at least 3 dilutions in duplicates were prepared) and set up four blanks with 1 ml Z-buffer

* pre-incubate them for 5 minutes at 37°C
* start the reaction by adding 0.2 ml PNPG (with multipipette), mix and incubate at 37°C
* stop the reaction after 30 minutes by adding 0.5 ml Na_2CO_3 after 30 minutes (or earlier, when the color turns to a strong yellow, note exact time).
* centrifuge the samples for 5 to 10 min at room-temperature and then measure the OD_{410}
* calculate the enzyme activity in as:

$$1 \text{ unit} = \frac{OD_{410} \text{ x dilution factor x } 1000}{OD_{600} \text{ x time (minutes)}}$$

17. Construction of *ΔcyaA* strains by T4GT7-transduction

* inoculate strains S2180, S2182, S2184 and S2186 into LB medium containing Spectinomycin
* mix 100µl of a fresh overnight cultures with 10µl, 5µl and 2µl T4GT7 lysate (prepared from strain S632 (Δ*cyaA*::kan), lab collection)
* incubate for 20' at room temperature
* plate entire mix onto LB plates containing Kanamycin for selection of the transductants
* re-streak colonies several times to get rid of contaminating T4GT7 phages
* test the clones via PCR with S502/S503 (expected PCR product size: ~1.9kb).

18. Isolation of Genomic DNA

Mg-Saline	0.85% NaCl, 10mM $MgSO_4$
Solution I	50mM Glucose, 25mM Tris-Cl pH-8.0, 10mM EDTA, 1mg/ml Lysozyme
Lysis buffer	1M NaCl, 25mM Tris-Cl, pH 7.5, 10mM EDTA, 4% SDS
TE buffer	10mM Tris-Cl pH 8.0, 1mM EDTA

* spin down 10 ml fresh overnight grown culture (5000 rpm; 10 min)
* wash 1x with Mg-Saline
* resuspend in 1 ml Solution I and add lysozyme. Incubate for 10-15min at room temperature
* add 2ml Lysis buffer and mix by inversion and incubate for 5min at room temperature
* add 10µl RNase and mix by inversion and incubate at 65°C for 15min
* add 200µl Proteinase K and mix by inversion and incubate at 37°C for 15min
* add 1 volume of Phenol/Choloroform (1:1) and mix by inversion, transfer to new tube and spin to separate aqueous and organic phases
* transfer the upper phase to a fresh tube using cut off end (blunt) pipetting tips and add 2-3 volumes of Ethanol
* wind out DNA carefully
* resuspend in 500µl TE
* add 20µl RNase and incubate at 37°C for 10 min
* add 5µl Proteinase K and incubate at 37°C for 10 min
* extract with 1 volume of Phenol/Chloroform, spin and transfer upper phase to a fresh Eppendorf tube with cut off end (blunt) tips and add 2 volumes of Ethanol.
* wind out DNA as above.
* dry under vaccum for few minutes
* resuspend in 250µl TE.

19. Southern hybridization (modified from Sambrook and Russell/Current protocols)

depurination solution		0.125M HCl
denaturation solution		1.5M NaCl, 0.5M NaOH
neutralisation solution		1M Tris, 1.5M NaCl
20x SSPE (1 l)		2.4M NaCl, 0.16M NaH_2PO_4. H_2O, 16mM EDTA in 800ml of H_2O, adjust pH to 7.4 with NaOH (~6.5 ml of a 10N solution), make up the final volume to 1 l with H_2O, sterilize by autoclaving
50x Denhardt's reagent		1% (w/v) Ficoll (Type 400), 1% (w/v) polyvinylpyrrolidone, 1% (w/v) bovine serum albumin
pre-hybridization buffer		5x Denhardt's reagent, 5x SSPE, 0.5% (w/v) SDS, 100µg/ml Salmon sperm DNA (ssDNA) (100µl from 100mg/ml stock solution, boiled for 10min and chilled on ice for 5min before adding)
washing solutions:	Buffer A	2x SSPE + 0.1% SDS
	Buffer B	2x SSPE + 0.5% SDS
	Buffer C	1x SSPE + 0.1% SDS
	Buffer D	2x SSPE

10µg of the genomic DNA was digested with appropriate restriction enzymes (20U) in a volume of 100µl. Incubate overnight at 37°C. The samples were loaded onto 0.7% agarose gel and electrophoresis was carried out by running the gels in low voltage (40-60volts) in TAE buffer. Following the electrophoresis, DNA samples in the gel were visualized in UV light along with the fluorescent rulers placed adjacent to the gel and photographed. The gel was then processed for depurination by submerging the gel in depurination solution for 10-20 min with gentle agitation, followed by denaturation for 15-30 min and neutralization for 15-30 min with gentle agitation at room temperature. The fragments were transferred to uncharged nylon membrane (Amersham Biosciences, Germany) by ascending capillary method in 20x SSPE transfer buffer (Sambrook et al., 2001). Briefly, the uncharged nylon membrane was placed on top of the gel and both the gel and membrane were sandwiched in between 4-5 Whatman filter papers (type 3MM). This setup was then placed on the wick (long Whatman filter paper) that runs on a solid support and in contact to the buffer 20x SSPE. A 500ml water bottle was placed above the complete apparatus for efficient transfer of DNA. The transfer of DNA was allowed for 8-24 hours. To fix the DNA, the membrane was incubated for 30 min to 2 hours at 80°C in a vaccum oven. The membrane was pre-hybridized with pre-hybridization solution at 68°C for 2-8hrs in hybridization bottles. Pre-hybridization was followed by hybridization to α^{32}-P dCTP specific probes (prepared using Ready To Go DNA labeling kit, Amersham) at 68°C for 8-12hrs. Following hybridization the membrane was washed in Buffer A for 3 min at room temperature, Buffer B for 15 min at room temperature, Buffer C for 30 min at 65°C, and Buffer D for 10 min at room temperature with gentle agitation. The membrane was then dried and exposed to X-ray film with an intensifying screen for 16-24 hours at room temperature to obtain an auto-radiographic image. The *bgl* region specific probes were generated either by PCR with MG1655 cells using *bgl* region specific primers or by restriction digestion of the plasmids pFDY52 and pFDX733 (Schnetz et

al., 1987). Detailed description of the positions of probes and the genomic digests used for Southern hybridization are shown in Figure 11 (Results).

20. Construction of i484Δ*bgl* strain (Ec93)

The strain Ec2 (i484 Bgl^+ mutant) that carries an activated *bgl* operon with a Δ47 bp in the silencer-*bgl* promoter region was transformed with plasmid pFMAC11 (rep^{ts}-tet^R, Δ*bgl* in *bgl* region, Caramel and Schnetz, 1998). The tranformants were selected on LB plate with Tetracycline at 28°C. Single colony was inoculated into LB medium with Tetracycline and was grown overnight at 28°C. The grown culture was plated on LB plate with Tetracycline at 42°C (at this step the plasmidic construct is integrated into the chromosome). Single colonies from this plate were inoculated into LB medium and grown at 28°C over night (this step allows recombination event to occur). The grown culture is plated on BTB salicin plate at 37°C and screened for Bgl^-. The obtained Bgl^- colonies were tested for Tetracycline sensitivity and PCR with oligos S4/S32.

21. Isolation of Bgl^+ and/or Sal^+ mutants

To isolate Bgl-positive mutants, cells carrying the wild-type *bgl* operon were plated on BTB-salicin plates and incubated at 37°C or 28°C. Bgl^+ papillae were picked after 2-4 days of incubation at 37°C or 28°C and re-streaked on BTB-salicin plates for purification. Expect from the strains that show relaxed phenotype all the spontaneous mutants were isolated at 37°C. Spontaneous mutants from the strains that show relaxed phenotype were isolated at 28°C. To characterize the mutation that caused activation, the *bgl* promoter region was amplified by PCR and the fragments were sequenced.

The c1955-c1960 salicin-positive mutants (Ec131 to Ec134), the strain Ec93 (i484 Δ*bgl*) carrying wild-type c1955-c1960 system was plated on BTB-salicin plates and incubated at 28°C. Sal^+ papillae were picked after 5 days of incubation at 28°C and re-streaked on BTB-salicin plates for purification. The four isolated Sal^+ mutants also show Cel^+ (cellobiose-positive), Arb^+ (arbutin-positive) and Esc^+ (esculin-positive) phenotype on respective BTB indicator plates at 28°C. Schematic presentation of the isolation of c1955-c1960 mutants are shown in Figure 23 (Results).

22. Microscopy for imaging β-glucoside utilization phenotypes

For imaging the phenotypes of the *E.coli* strains, strains were streaked on Bromthymol blue (BTB) salicin (β-glucoside) indicator plates and phenotypes were documented up to 5 days of incubation using a Stereomicroscope Zeiss stemi 2000-C Microscope. For processing images Adobe photoshop and/or CorelDRAW was used.

Table 4: Synthetic oligonucleotides used in the present study

name[a]	sequence[b]	position[c] or reference
S4	GGATGGACATTGACGAAGC	*phoU*: 443 to 461
S5	GGATTGTTACTGCATTCGC	*bgl*: +42 to +60
S9	TGAGGGGACGACGACAGT	*lac:* +305 to +288
S32	CCACTGCGGCAAGCTGAG	MG1655 *yieJ*: 549 to 566
S34	GGATAAACTGCTGGCGGG	MG1655, O157 *phoU*: 690 to 707
S54	*CCTCTAGA*AATTCCGCGCCCCATGACGA	*bgl*: +224 to +205
S91	GTGATTTGCATGTTCATAGCAAGGAC	*bgl*: +148 to +123
S92	CAAGAGGAATATGACTTAAGAGTTCG	*bgl*: +342 to +317
S94	GCTTTACTAAGCTGATCCGGTGGA	pKES15: 8479 to 8502
S95	CATATGGGGATTGGTGGCGA	*lac*: +3009 to +3028
S100	CATCGTAACCGTGCATCTGCCA	*lac*: +330 to +309
S118	TGCGGGCCTCTTCGCTATTA	*lac*: +171 to +152
S123	TGTGGAATTGTGAGCGGATA	*lac:* -6 to +15
S145	*CCGGTCGAC*GCGTTCGCGCGGATGGACATTGACGAAGC	*phoU*: 431 to 461
S154	pGTGAAACCAGTAACGTTATACGATGTCG	*lac*: -1168 to -1141
S157	*CCCGTCGAC*TTATAACTGCGAGCATGGTCA	*bgl*: -76 to -55
S165	GTAACAGTGGCCCGAAGATA	MG1655, O157 *lac*: +3462 to +3443
S186	ATAACCAGAGAATACTGGTGAAGTCGGGT	*bgl*: +75 to +103
S201	*GCGTCTCTAGA*AATATTTCAGTGTTCTTTGCGCACG	*bgl*: +974 to +950
S212	*CCTCTAGA*TTTTTATAACGAACATCCAGGTTCG	*bgl*: +25 to +1
S220	*GCGGATCC*ATGAACATGCAAATCACCAAAATTCTCA	*bgl*: +132 to +159
S224	*GGGGATCC*ACCCGCAAGCATGGCAATGT	*bgl*: +841 to +860
S252	AGATGCTGCACGACGTGCTGG	*phoU*: 410 to 430
S253	GTCACGGCAAACGAAAGCGC	MG1655 *bgl*: +5427 to +5446
S257	*GCTCTAGA*TGCCCTCTACCGCTTTGCG	*bgl*: +1097 to +1079
S309	*CCGTCTAGA*CATCGATTTTTATCACCGT	*bgl*: +3171 to +3189
S312	TGATCCCCGCCTGCGC	*bgl*: +3352 to +3337
S335	CTTCAGGATCGAGCGTAATACCA	MG1655 *yieH*: 121 to 99
S336	CGCGGAAAATCGTCAGTAACA	*yieG*: 394 to 374
S346	CCACTCTGATTTCGAATCTATTCGT	*bgl*: +5777 to +5801
S347	GCACAATAAGCCGATCGTTCA	*yieH*: 567 to 587
S348	CGGGTCACAGAAACGTTATCGT	MG1655, CFT073 *yieI*: 219 to 240
S349	GTAACGACATGTTGATTTCATTAAACGT	MG1655 *bgl*: +3529 to +3502
S350	GCTGCCGCAGAATTAGCACT	MG1655 *bgl*: +8002 to +8021
S351	GGATGATGCCAGCATAGAAGGT	MG1655 *yieJ*: 222 to 243
S352	GAGCGGCATAACCTGAATCTGA	IS1: 625-646
S357	GGCAGGGTCGTTAAATAGCCGCTTATGT	miniTn10-Cm^R oligo: 1180 to 1214
S358	CGGTATCAACAGGGACACCAGGATTTATTTATTCT	miniTn10-Cm^R oligo: 263 to 236
S359	*GCTCTAGA*GATCATATGACAAGATGTGTATCCACCTTAACT	miniTn10-Cm^R oligo: 70 to 38 and 1411 to 1443
S360	GGCCACGCGTCGACTAGTACNNNNNNNNNNGATC	ST-PCR oligo, modified from (Chun et al., 1997)
S361	*GCTCTAGA*GGCCACGCGTCGACTAGTAC	ST-PCR oligo, modified from (Chun et al., 1997)
S367	*CGGAATTCGGGCCC*GCCCTATTATAATCAACACGCTATGTAGT	MG1655, CFT073 *marB*: 192 to 220

Table 4: Synthetic oligonucleotides used in the present study

name[a]	sequence[b]	position[c] or reference
S368	*CCGCTCGAGCGGCCGCGTCGAC*GGCGTAAAGCGGTAAAGGTCA	CFT073 *ydeD*: 893 to 913
S373	*CCGGTCGAC*GCTGCCAGAATATTTGTGAGTTTATCT	*phoU*: 613 to 640
S374	GGTCAGCAACCCTTATATCGATACA	CFT073 c1955: 1020 to 1044
S375	GCAGGAAGTCAGATTATGAAATTTGA	CFT073 c1960: 271 to 296
S376	CACCGTTTTCATTTTTGATGTTATCA	CFT073 c1956: 295 to 270
S377	GACCGCGTTAATTAGTCAGGATGA	CFT073 c1957:168 to 191
S401	GGGCGTTGCGGAACAAAC	*marB*: 54 to 71
S402	GTATTTGGTTTGCGGTGGCGTAAAGCGGT	*ydeD:* 877-905
S406	*GCGGATCC*GCAAAGCGGTAGAGGGC	*bgl*: +1078 to +1096
S407	TGATGATAAAGGTAATCTGCTAAACCG	MG1655 *bgl*: +1371 to +1397
S408	GCCTGGTTGTGCAGCATTCTG	*bgl*: +1771 to +1791
S409	GATCAGCGGCGTTGACGAG	MG1655 *bgl*: +2171 to +2189
S410	CATTCACGTCGCTGATACCACG	*bgl*: +2571 to +2592
S411	*GCGGA*TCCTTTTATCGTTAGCGAATGATGG	*bgl*: +2990 to +2966
S412	CCATGCGGCAAAAGCGATGAATGT	*yieH*: 447 to 470
S413	CCGATCGTTCACCCGAAAGTCACCA	*yieH*: 557 to 601
S429	GGCGAAAAACTTGCTGATAATTGT	*bgl*: +7860 to +7883
S430	GCATGATCAATCATACGTTTAACCA	CFT073 c1959: 74 to 50
S431	GTCGCAATAGTAGGAGAAGCATCCT	CFT073 c1960: 326 to 302
S432	*GCTCTAGA*TTTTTCCTTACTGGTATATAACAGACTACATT	CFT073 AE016760: 292477 to 292508
S433	*CCGGTCGACGAATTC*TTTTGCACCTTTTAAGAGCCATT	CFT073 AE016760: 292673 to 292651
S434	*CGGGATCC*GCTTAAGCGACTTCATTCACCTGA	*lac*: +4428 to +4397
S435	*CCGGTCGAC*GAATGGCGCAAAACCTTTC	*lac*: -1238 to -1219
S436	*CGGGATCC*CGGTTATTATTATTTTTGACACCA	*lac*: +3124 to +3099
S437	CGACGGTACGCTGGTCGA	*yieH*: 33 to 50
S460	CTTCGGTAACCGGACCTTGC	MG1655 *bgl*: +7942 to +7923
S463	AGTGCCTGACAGCTACGTGACG	*bgl*: +7815 to +7836
S467	*CAATCCTTTACTCAG*TAAGCTTAACCGAGTGCTAATTCTGC	MG1655 *bgl*: +8033 to +8008
S479	*CCGTCGACCC*ACCAGCAAATGAGCCGTGTCGCGG	*bgl*: +7337 to +7354
S490	*CGGAATTC*GTAGTCTGTTATATACCAGTAAGGAAAAATATGA	CFT073 AE016760: 292505 to 292472
S491	*GCTCTAGA*CCGTTTGATTTAAATCACATGGGTA	*marB*: 132 to 160
S502	CCGTGGTCCATCCTAACATCCT	*cyaA* AE000456: 7930-7951
S503	CATTATTCCGGTGACGGATGAATC	*cyaA* AE000456: 11315-11292
S536	CTGAATGCTAAAGCGGCAGATC	*bgl*: +6848 to +6869
S537	CAGTGGCTTGGGATGATATTTGA	MG1655 *yieJ*: 32 to 54
S545	CTGTTGCCCGTCTCACTGGT	*lac*: -217 to -198
S546	CTGTCCTGGCCGTAACCGA	*lac*: +535 to +517
S547	CGCTTAGTTTTTCATTATCATTAGGGA	O157 Z5214: 481-455
S548	GTCGATTGTGATGATAAAATACGTTCT	O157 Z5211: 2248-2274
S560	GCGAAGGAAGCCTCACAAGA	*bgl*: +4304 to 4323
S561	CCCGTTATTATCCTGATTATCTTTTTC	*bgl*: +5477 to +5451
S562	GTCGTTACACGCGCCATTCAC	*bgl*: +3522 to 3542
S563	CGGGTGAATATTGTCCGGAAC	*bgl*: +6346 to +6326
S564	CCGTAGCGCTTAGACATTTGTGA	*bgl*: +4273 to 4252
S565	GCGTTCCAGAATGCTGCACA	*bgl*: +1797 to 1778
S566	GCTGGTTCGGTGATACCAAACA	*bgl*: +2213 to 2192

Table 4: Synthetic oligonucleotides used in the present study

name[a]	sequence[b]	position[c] or reference
S579	GACCGCATTTGCCGTTCTAC	CFT073 c1956: 603 to 584
S580	CGCCAGCCACCATATTTTGT	CFT073 c1955: 440-421
S581	GCCGTAAAACAGAAGTCAGCATC	CFT073 c1956:1102-1080
S586	GGTTCTTTGGGTGATAATACATCCA	MG1655 *yieI*: 111-135
S587	CTTTTGGTAATAATACAGGTACTTCCATTGT	CFT073 *yieI*: 115-145
S588	CCGGTTACCGAAGATGTTCCA	MG1655 *bgl*: +7930 to +7950
S589	CATTTTTGTGGCAATCTGCCA	CFT073 *bgl*: +7692 to +7712

a: name of the oligo in the lab collection

b: sequence is shown from 5' to 3' end. P indicates 5' phosphate. Restriction enzyme sites are underlined; overhangs are in italics.

c: Oligo mapping positions are indicated. Numbering in *bgl* and *lac* are relative to the transcription start of the *bgl* operon and *lac* operon respectively. Wherever relevant, oligos mapping specifically to K-12 MG1655 sequence are prefixed with MG1655. Likewise oligos specific for the CFT073 sequence are prefixed with CFT073 and those specific for O157 sequence are prefixed with O157. Oligos that are not prefixed maps to all the four published sequences. Un-prefixed *bgl* oligos maps to both MG1655 and CFT073 *bgl* operon. Numbering in *phoU*, *yieJIHG*, c1955-c1960, *marB*, *ydeD* are relative to their respective translational start. Positions of the oligos S250, S302, S432, S433 and S490 are relative to the NCBI accession numbers. Numbering in IS1 is relative to the defined left and right ends of the insertion element (Grindley, 1978). Additional descriptions and strategies when used for cloning are documented in lab records.

Table 5: *E.coli* K-12 strains used in the study

strain[a]	relevant genotype or structure[b]	construction[c]/ reference
S49	CSH50 *bgl*° Δ (*lac-pro*) *ara thi*	(Miller, 1972)
S157	CSH50 *bglR*::IS1	(Schnetz, 1992)
S162	CSH50 Δ *bgl*-AC11	(caramel and Schnetz, 1998)
S432	CSH50 (=S49) Bgl^+ *bgl*-CAP-C234 (C-66→T)	(Schnetz, 2002)
S484	CSH50 *bgl*° Δ (*argF-lac*) U169 Pro^+	(Dole et al., 2002)
S486	CSH50 (*gpt-lac*)-positive	(Dole et al., 2002)
S524	= S486 Δ lacZ-Y217	(Dole et al., 2002)
S527	MG1655 laboratory strain K-12	(Blattner et al., 1997)
S541	CSH50 Δ*bgl*-AC11 Δ*lacZ*	(Dole et al., 2002)
S614	= S541 *hns*::Ap^R	lab collection
S632	= lambda$^-$, e14$^-$, *relA1*, *spoT1*, Δ(*cyaA*1400)::kan thi-1	(Shah and Peterkofsky, 1991)
S887	=S541 rpoS::Tn10	(Dole et al., 2002)
S996	= S524 Δ*crp zhd*::Tn10	lab collection
S2180/1	=S541 attB:: $spec^R$ Ec93-$P_{c1955\text{-}c1960}$-*lacZ*	x pKEGN46
S2182/3	=S541attB:: $spec^R$ Ec134-$P_{c1955\text{-}c1960}$-*lacZ*	x pKEGN48
S2184/5	=S541 attB:: $spec^R$ Ec93-c1960-$P_{c1955\text{-}c1960}$-*lacZ*	x pKEGN51
S2186/7	=S541attB:: $spec^R$ Ec134-c1960-$P_{c1955\text{-}c1960}$-*lacZ*	x pKEGN52
S2272/3	=S2180 Δ*cyaA*::kan	x T4GT7 S632 kan^R
S2274/5	=S2182 Δ*cyaA*::kan	x T4GT7 S632 kan^R
S2276/7	=S2184 Δ*cyaA*::kan	x T4GT7 S632 kan^R
S2278/9	=S2186 Δ*cyaA*::kan	x T4GT7 S632 kan^R

a: Strain names in the lab collection. Strain numbers of duplicates of the respective strains are indicated with /.

b: The relevant genotype of the constructed CSH50 derivatives refers to the *bgl*, *lac*, c1955-c1960, *cyaA*, *crp*, *hns* and *rpoS* loci. Mutations causing activation of the silent *bgl* operon include *bgl*-CRP (a C to T exchange in the CRP binding site at position -66, relative to the transcription start), *bglR*::IS1 (integration of IS1 in orientation II generating a target site duplication from -88 to -80). $P_{c1955\text{-}c1960}$ indicates the putative promoter region from c1955-c1960 system (position relative to CFT073 AE016760: 292674-292476).

c**:** Transductants (using T4GT7) of *cyaA*::kan alleles were selected on LB plates containing kanamycin. Integrations into *attB* were performed as described (see materials and methods) (Diederich et al., 1992).

Table 6a: clinical *E.coli* isolates used in the study

Strain[a]	original strain name[b]	strain type[c]	typing[d]	bgl/Z5211-Z5214 upstream sequence[e]	*bgl/Z5211-Z5214 locus*[f]	*yieJ* PCR[g]	5'-*yieI* [h]	*bgl*/Z5211-Z5214 downstream sequence[i]	mutations[j]	Sal Phenotype 37°C[k]	28°C	c1955-c1960[l]	Lac Phenotype 37°C[m]	lac genotype[n]
Ec1	i484	sept	CFT073	CFT073	CFT073	CFT073	CFT073	CFT073		+++	++	+	+	
Ec9	F1	sept	CFT073	CFT073	CFT073	CFT073	CFT073			+++		+	+	
Ec11	F385	sept	CFT073	CFT073	CFT073	CFT073	CFT073			+++		+	+	
Ec13	F560	sept	CFT073	CFT073	CFT073	CFT073	CFT073			+++		+	+	
Ec28	W7483	sept	CFT073	CFT073	CFT073	CFT073	CFT073			+++		+	+	
Ec35	U2388	uro	CFT073	CFT073	CFT073	CFT073	CFT073			+++		+	+	
Ec36	U2873	uro	CFT073	CFT073	CFT073	CFT073	CFT073			+++		Δ c1956-c1960	+	
Ec42	U3362	uro	CFT073	CFT073	CFT073	CFT073	CFT073			+++		+	+	
Ec49	U4437	uro	CFT073	CFT073	CFT073	CFT073	CFT073			+++		+	+	
Ec59	E10079	com	CFT073	CFT073	CFT073	CFT073	CFT073			+++		+	+	
Ec73	E10091	com	CFT073	CFT073	CFT073	CFT073	CFT073			+++		+	+	
Ec111	E182	com	CFT073	CFT073	CFT073	CFT073	CFT073			+++		+	+	
Ec126	E175	com	CFT073	CFT073	CFT073	CFT073	CFT073			+++		+	+	
Ec116	E452	com	CFT073	CFT073	CFT073	CFT073	CFT073			+++		+	+	
Ec23	St5119	sept	CFT073	CFT073	CFT073	CFT073	CFT073			++		+	+	
Ec38	U3145	uro	CFT073	CFT073	CFT073	CFT073	CFT073			++		+	+	
Ec125	E176	com	CFT073	CFT073	CFT073	CFT073	CFT073			++		+	+	
Ec129	E471	com	CFT073	CFT073	CFT073	CFT073	CFT073			++		+	+	
Ec43	U3454	uro	CFT073	CFT073	CFT073	CFT073	CFT073			+		+	+	
Ec40	U3407	uro	CFT073	CFT073	CFT073	CFT073	CFT073			+		+	+	
Ec66	E10094	com	CFT073	CFT073	CFT073	CFT073	CFT073			+		+	+	
Ec117	E478	com	CFT073	CFT073	CFT073	CFT073	CFT073		+58ct	+		-	+	
Ec118	E457	com	CFT073	CFT073	CFT073	CFT073	CFT073			+		+	+	
Ec119	E177	com	CFT073	CFT073	CFT073	CFT073	CFT073			+		+	+	
Ec120	E178	com	CFT073	CFT073	CFT073	CFT073	CFT073			+		+	+	
Ec127	E464	com	CFT073	CFT073	CFT073	CFT073	CFT073			+		c1955::IS1	+	
Ec128	E466	com	CFT073	CFT073	CFT073	CFT073	CFT073			+		c1955::IS1	+	
Ec100	E475	com	CFT073	CFT073	CFT073	CFT073	CFT073			none	++	+	+	
Ec114	E422	com	CFT073	CFT073	CFT073	CFT073	CFT073	CFT073	+9ga, *bglK*-::IS1397-*yieI*	+		+	+	
Ec20	F911	sept	CFT073	CFT073	CFT073	CFT073	CFT073	CFT073	+26ca, Δ (bglF-yieH) ::IS1294	none	+	+	+	
Ec18	F785	sept	MG1655	MG1655	MG1655	MG1655	MG1655			+		-	+	
Ec58	E10097	com	MG1655	MG1655	MG1655	MG1655	MG1655			+		-	+	
Ec64	E10099	com	MG1655	MG1655	MG1655	MG1655	MG1655			+		-	+	

Table 6a: clinical *E.coli* isolates used in the study

Strain[a]	original strain name[b]	strain type[c]	typing[d]	bgl/Z5211-Z5214 upstream sequence[e]	*bgl/Z5211-Z5214 locus*[f]	*yieJ* PCR[g]	5'-yieI [h]	bgl/Z5211-Z5214 downstream sequence[i]	mutations[j]	Sal Phenotype 37°C[k]	28°C	c1955-c1960[l]	Lac Phenotype 37°C[m]	lac genotype[n]
Ec69	E10082	com	MG1655	MG1655	MG1655	MG1655	MG1655			+		-	+	
Ec72	E10085	com	MG1655	MG1655	MG1655	MG1655	MG1655			+		-	-	
Ec105	E167	com	MG1655	MG1655	MG1655	MG1655	MG1655		bglH::IS2	+		-	+	
Ec106	E166	com	MG1655	MG1655	MG1655	MG1655	MG1655			+		-	+	
Ec115	E444	com	MG1655	MG1655	MG1655	MG1655	MG1655			+		-	+	
Ec121	E180	com	MG1655	MG1655	MG1655	MG1655	MG1655			+		-	+	
Ec10	F287	sept	MG1655	MG1655	MG1655	MG1655	MG1655		+55at	++		-	+	
Ec21	F1215	sept	MG1655	MG1655	MG1655	MG1655	MG1655		+55at, t1 (t105a)	++		-	+	
Ec102	E476	com	MG1655	MG1655	MG1655	MG1655	MG1655			++		-	+	
Ec70	E10090	com	MG1655	MG1655	MG1655	MG1655	MG1655			none	none	-	+	
Ec63	E10096	com	MG1655	MG1655	MG1655	MG1655	MG1655	MG1655	yieJ::ISEc8 region	+		-	+	
Ec109	E345	com	MG1655	MG1655	MG1655	MG1655	MG1655	MG1655	Δ(bglI-bglK)::IS629, yieJ::IS629	+		-	+	
Ec108	E292	com	MG1655	MG1655	MG1655		MG1655	MG1655	Δ(yieJ-yieI)::IS629	+		-	+	
Ec44	U3633	uro	MG1655	MG1655	MG1655	MG1655	MG1655	MG1655	yieJ::ISEc8 region	++		-	+	
Ec48	U4418	uro	MG1655	MG1655	MG1655	MG1655	MG1655	MG1655	yieJ::IS1, yieJ::ISEc8 region	++		-	+	
Ec52	U5107	uro	MG1655	MG1655	MG1655		MG1655	MG1655	t1 ('+103tg), Δ(bglK-yieJ)::IS629	++		-	+	
Ec103	E164	com	MG1655	MG1655	MG1655		MG1655	MG1655	Δ(bglI-yieH)::IS1	++		-	+	
Ec34	U2366	uro	MG1655	MG1655	MG1655	MG1655	MG1655		bglB::IS186	none	none	-	+	
Ec107	E291	com	MG1655	MG1655	MG1655	MG1655	MG1655	MG1655	yieJ::IS629	none	none	-	+	
Ec12	F557	sept	fourth	O157	MG1655	MG1655	CFT073	MG1655	+784ag (BglG-E218G)	+		+	+	
Ec16	F742	sept	fourth	O157	MG1655	MG1655	CFT073	MG1655		+		+	+	
Ec25	V9261	sept	fourth	O157	MG1655	MG1655	CFT073	MG1655		+		+	+	
Ec27	V10744	sept	fourth	O157	MG1655	MG1655	CFT073	MG1655	+165ag (BglG-N12D), '+784ag (BglG-E218G)	+		+	+	
Ec45	U3622	uro	fourth	O157	MG1655	MG1655	CFT073	MG1655		+		+	+	
Ec47	U4252	uro	fourth	O157	MG1655	MG1655	CFT073	MG1655		+		-	+	
Ec53	U5033	uro	fourth	O157	MG1655	MG1655	CFT073	MG1655		+		-	+	
Ec60	E10087	com	fourth	O157	MG1655	MG1655	CFT073	MG1655		+		+	+	

Table 6a: clinical *E.coli* isolates used in the study

Strain[a]	original strain name[b]	strain type[c]	typing[d]	bgl/Z5211-Z5214 upstream sequence[e]	*bgl/Z5211-Z5214 locus*[f]	*yieJ* PCR[g]	5'-yiel [h]	*bgl*/Z5211-Z5214 downstream sequence[i]	mutations[j]	Sal Phenotype 37°C[k]	28°C	c1955-c1960[l]	Lac Phenotype 37°C[m]	lac genotype[n]
Ec46	U4191	uro	fourth	O157	MG1655	MG1655	CFT073	MG1655		++		-	+	
Ec37	U3104	uro	fourth	O157	MG1655	MG1655	CFT073	MG1655		++		+	+	
Ec32	W9887	sept	fourth	O157	MG1655	MG1655	CFT073	MG1655	+546ag (BglG-T139A), bglH::IS1	++		-	+	
Ec62	E10077	com	fourth	O157	MG1655	MG1655	CFT073	MG1655	+273ag (BglG-K48E)	late	late	+	+	
Ec14	F569	sept	fourth	O157	MG1655	MG1655	CFT073	MG1655		late		+	+	
Ec39	U3372	uro	fourth	O157	MG1655	MG1655	CFT073	MG1655	+594ga (BglG-G155S)	late	+	-	+	
Ec22	St4723	sept	fourth	O157	MG1655	MG1655	CFT073	MG1655		none	none	Δ c1955-c1956	+	
Ec26	V9343	sept	fourth	O157	MG1655	CFT073	CFT073	MG1655	+378ga (BglG-A83T), +546ag (BglG-T139A)	late		+	+	
Ec33	U2183	uro	fourth	O157	MG1655	CFT073	CFT073	MG1655		+		+	+	
Ec31	W9763	sept	fifth	MG1655	MG1655	MG1655	CFT073		t1(+102ga)	+		+	+	
Ec124	E174	com	fifth	MG1655	MG1655	MG1655	CFT073			+		+	+	
Ec57	E10092	com	fifth	MG1655	MG1655	MG1655	CFT073			+		-	+	
Ec17	F775	sept	fifth	MG1655	MG1655	MG1655	CFT073			++		-	+	
Ec30	W8987	sept	fifth	MG1655	MG1655	MG1655	CFT073		t1(+102ga)	++		-	-	Δlac
Ec110	E294	com	fifth	MG1655	MG1655	MG1655	CFT073			++		-	+	
Ec50	U4417	uro	fifth	MG1655	MG1655		CFT073	MG1655	Δ(bglG-yiel)::IS1	none	none	+	+	
Ec61	E10086	com	fifth	MG1655	MG1655	CFT073	CFT073	MG1655		+		+	+	
Ec104	E165	com	mixed	MG1655	MG1655	CFT073	CFT073	CFT073		+		+	+	
Ec112	E7370	com	mixed	CFT073	CFT073	CFT073	MG1655	CFT073		+++		-	+	
Ec130	E467	com	mixed	CFT073	CFT073	MG1655	CFT073	MG1655		+		+	+	
Ec67	E10083	com	mixed	O157	O157	O157	O157	MG1655	ΔZ5211-Z5214	none	none	-	+	
Ec15	F645	sept	O157	O157	O157	O157	O157	O157		none	none	c1956::IS1	-	lacY$^-$
Ec19	F905	sept	O157	O157	O157	O157	O157	O157		none	none	-	+	
Ec24	St5679	sept	O157	O157	O157	O157	O157	O157		none	none	+	+	
Ec41	U3292	uro	O157	O157	O157	O157	O157	O157		none	none	-	+	
Ec51	U4409	uro	O157	O157	O157	O157	O157	O157		none	none	-	-	lacY$^-$
Ec54	U5070	uro	O157	O157	O157	O157	O157	O157		none	none	-	+	
Ec55	E10093	com	O157	O157	O157	O157	O157	O157		none	none	+	+	
Ec74	E10098	com	O157	O157	O157	O157	O157	O157		none	none	-	+	
Ec101	E460	com	O157	O157	O157	O157	O157	O157		none	none	+	+	
Ec123	E173	com	O157	O157	O157	O157	O157	O157		none	none	-	-	lacY$^-$

Table 6a: clinical *E.coli* isolates used in the study

Strain[a]	original strain name[b]	strain type[c]	typing[d]	bgl/Z5211-Z5214 upstream sequence[e]	*bgl/Z5211-Z5214 locus*[f]	*yieJ* PCR[g]	5'-yiel [h]	bgl/Z5211-Z5214 downstream sequence[i]	mutations[j]	Sal Phenotype 37°C[k]	28°C	c1955-c1960[l]	Lac Phenotype 37°C[m]	lac genotype[n]
Ec65	E10100	com	O157	O157	O157	O157	O157	O157		late	none	+	+	
Ec99	E472	com	O157	O157	O157	O157	O157	O157		late	none	+	+	
Ec68	E10084	com	O157	O157	O157	O157	O157	O157		late	+	+	-	lacY::IS1
Ec113	E424	com	O157	O157	O157	O157	O157	O157		late	late	+	+	
Ec56	E10089	com	O157	O157	O157	O157	O157	O157		none	none	-	+	
Ec71	E10095	com	O157	O157	O157	O157	O157	O157		none	none	-	+	
Ec122	E179	com	O157	O157	O157	O157	O157	O157		none	none	-	-	lacY$^-$
Ec29	W7716	sept	O157	O157	O157	O157	O157	O157	Δ Z5211-Z5214::IS1	none	late	+	+	

Table 6b: ECOR strains analyzed in the study

Strain[a]	original strain name[b]	strain source	strain type[c]	Phylo-genetic groups Herzer et.al., 1990[o]	bgl-Hall, 1988[p]	typing[d]	bgl/Z5211-Z5214 upstream sequence[e]	*bgl/Z5211 - Z5214 locus*[f]	*yieJ* PCR[g]	5'-yiel[h]	bgl/Z5211-Z5214 downstream sequence[i]	mutations[j]	Sal Phenotype 37°C[k]	c1955-c1960[l]	Lac Phenotype 37°C[m]
Ec173	ECOR23	Elephant	healthy	A	+	CFT073	CFT073	CFT073	CFT073	CFT073			+++	+	+
Ec182	ECOR32	Giraffe	healthy	B1	+	CFT073	CFT073	CFT073	CFT073	CFT073			+++	+	+
Ec201	ECOR51	Human	healthy	B2	+	CFT073	CFT073	CFT073	CFT073	CFT073			+++	+	+
Ec202	ECOR52	Orangutan	healthy	B2	+	CFT073	CFT073	CFT073	CFT073	CFT073			+++	+	+
Ec203	ECOR53	Human	healthy	B2	+	CFT073	CFT073	CFT073	CFT073	CFT073			+++	+	+
Ec204	ECOR54	Human	healthy	B2	+	CFT073	CFT073	CFT073	CFT073	CFT073			+++	+	+
Ec205	ECOR55	Human	UTI (P)	B2	+	CFT073	CFT073	CFT073	CFT073	CFT073			+++	+	+
Ec207	ECOR57	Gorilla	healthy	B2	+	CFT073	CFT073	CFT073	CFT073	CFT073			+++	+	+
Ec210	ECOR60	Human	UTI(C)	B2	+	CFT073	CFT073	CFT073	CFT073	CFT073			+++	+	+
Ec213	ECOR63	Human	healthy	B2	+	CFT073	CFT073	CFT073	CFT073	CFT073			+++	Δ (c1958-c1959)	+
Ec206	ECOR56	Human	healthy	B2	+	CFT073	CFT073	CFT073	CFT073	CFT073			++	+	+
Ec209	ECOR59	Human	healthy	B2	+	CFT073	CFT073	CFT073	CFT073	CFT073			+	+	+
Ec211	ECOR61	Human	healthy	B2	+	CFT073	CFT073	CFT073	CFT073	CFT073			+	+	+
Ec212	ECOR62	Human	UTI (P)	B2	+	CFT073	CFT073	CFT073	CFT073	CFT073			+	+	+
Ec214	ECOR64	Human	UTI (C)	B2	+	CFT073	CFT073	CFT073	CFT073	CFT073			+	+	+
Ec215	ECOR65	Ape	healthy	B2	+	CFT073	CFT073	CFT073	CFT073	CFT073	CFT073		+	+	+
Ec152	ECOR2	Human	healthy	A	+	MG1655	MG1655	MG1655	MG1655	MG1655			++	-	+
Ec155	ECOR5	Human	healthy	A	+	MG1655	MG1655	MG1655	MG1655	MG1655			++	-	+
Ec162	ECOR12	Human	healthy	A	+	MG1655	MG1655	MG1655	MG1655	MG1655		bglH::IS2	++	-	+
Ec163	ECOR13	Human	healthy	A	+	MG1655	MG1655	MG1655	MG1655	MG1655			++	-	+
Ec160	ECOR10	Human	healthy	A	+	MG1655	MG1655	MG1655	MG1655	MG1655			+	-	+
Ec161	ECOR11	Human	UTI(C)	A	+	MG1655	MG1655	MG1655	MG1655	MG1655			+	-	+
Ec175	ECOR25	Dog	healthy	A	+	MG1655	MG1655	MG1655	MG1655	MG1655			+	-	+
Ec151	ECOR1	Human	healthy	A	+	MG1655	MG1655	MG1655	MG1655	MG1655			none	-	+
Ec153	ECOR3	Dog	healthy	A	+	MG1655	MG1655	MG1655	MG1655	MG1655			none	-	+
Ec158	ECOR8	Human	healthy	A	+	MG1655	MG1655	MG1655	MG1655	MG1655			none	-	+
Ec164	ECOR14	Human	UTI (P)	A	+	MG1655	MG1655	MG1655	MG1655	MG1655			none	-	+
EC169	ECOR19	Ape	healthy	A	+	fourth	O157	MG1655	MG1655	CFT073	MG1655		++	-	+
Ec195	ECOR45	Pig	healthy	B1	+	fourth	O157	MG1655	MG1655	CFT073	MG1655		++	+	+
Ec217	ECOR67	Goat	healthy	B1	+	fourth	O157	MG1655	MG1655	CFT073	MG1655		++	+	+
Ec157	ECOR7	Orangutan	healthy	A	+	fourth	O157	MG1655	MG1655	CFT073	MG1655		+	+	+
Ec176	ECOR26	Human	healthy	B1	+	fourth	O157	MG1655	MG1655	CFT073	MG1655		+	-	+
Ec177	ECOR27	Giraffe	healthy	B1	+	fourth	O157	MG1655	MG1655	CFT073	MG1655		+	+	+
Ec178	ECOR28	Human	healthy	B1	+	fourth	O157	MG1655	MG1655	CFT073	MG1655		+	+	+

Table 6b: ECOR strains analyzed in the study

Strain[a]	original strain name[b]	strain source	strain type[c]	Phylo-genetic groups Herzer et.al., 1990[o]	bgl-Hall, 1988[p]	typing[d]	bgl/Z5211-Z5214 upstream sequence[e]	*bgl/Z5211 - Z5214 locus*[f]	*yieJ* PCR[g]	5'-yieI[h]	bgl/Z5211-Z5214 downstream sequence[i]	mutations[j]	Sal Phenotype 37°C[k]	c1955-c1960[l]	Lac Phenotype 37°C[m]
Ec220	ECOR70	Gorilla	healthy	B1	+	fourth	O157	MG1655	MG1655	CFT073	MG1655		+	-	+
Ec221	ECOR71	Human	ABU	B1	+	fourth	O157	MG1655	MG1655	CFT073	MG1655		+	-	+
Ec222	ECOR72	Human	UTI (P)	B1	+	fourth	O157	MG1655	MG1655	CFT073	MG1655		+	+	+
Ec171	ECOR21	Cow	healthy	A	+	fourth	O157	MG1655	MG1655	CFT073	MG1655	bglB::IS1	none	-	+
Ec170	ECOR20	Cow	healthy	A	+	fourth	O157	MG1655	MG1655	CFT073	MG1655	t1 (+102gt), bglB::IS1	none	-	+
Ec167	ECOR17	Pig	healthy	A	Δ	fourth	O157	MG1655	MG1655	CFT073	MG1655	Δ (*bgl*t1-*bglK*)::IS1	none	Δ (c1956-c1957)::IS1	+
Ec168	ECOR18	Ape	healthy	A	+	fourth	O157	MG1655	MG1655	CFT073	MG1655	Δ (P_{bgl}-*bglF*)::IS1	none	-	+
Ec159	ECOR9	Human	healthy	A	+	fourth	O157	MG1655	MG1655	CFT073	MG1655	Δ (bglB-bglH)::IS1	none	-	+
Ec219	ECOR69	Ape	healthy	B1	+	fourth	O157	MG1655	CFT073	CFT073	MG1655		+	+	+
Ec208	ECOR58	Lion	healthy	B1	+	fourth	O157	MG1655	CFT073	CFT073	MG1655		+	+	+
Ec218	ECOR68	Giraffe	healthy	B1	+	fifth	MG1655	MG1655	MG1655	CFT073			+	+	+
Ec165	ECOR15	Human	healthy	A	+	fifth	MG1655	MG1655	MG1655	CFT073			+	-	+
Ec172	ECOR22	Cow	healthy	A	+	fifth	MG1655	MG1655	MG1655	CFT073			+	+	+
Ec180	ECOR30	Bison	healthy	B1	+	fifth	MG1655	MG1655	MG1655	CFT073	MG1655		+	+	+
Ec183	ECOR33	Sheep	healthy	B1	+	fifth	MG1655	MG1655	MG1655	CFT073			+	+	+
Ec184	ECOR34	Dog	healthy	B1	+	fifth	MG1655	MG1655	MG1655	CFT073			+	+	+
Ec174	ECOR24	Human	healthy	A	+	fifth	MG1655	MG1655	MG1655	CFT073	MG1655		late	-	+
Ec166	ECOR16	Leopard	healthy	A	+	fifth	MG1655	MG1655	MG1655	CFT073			late	+	+
Ec179	ECOR29	Kangaroo rat	healthy	B1	+	fifth	MG1655	MG1655	MG1655	CFT073			late	+	+
Ec156	ECOR6	Human	healthy	A	+	fifth	MG1655	MG1655	MG1655	CFT073			none	-	-
Ec181	ECOR31	Leopard	healthy	E	+	fifth	MG1655	MG1655	CFT073	CFT073	MG1655	t1 (+102gt)	++	-	+
Ec154	ECOR4	Human	healthy	A	+	fifth	MG1655	MG1655	CFT073	CFT073	MG1655		++	+	+
Ec216	ECOR66	Ape	healthy	B2	+	mixed	CFT073	CFT073	CFT073	MG1655	CFT073		+++	-	+
Ec185	ECOR35	Human	healthy	D	Δ, R	O157	O157	O157	O157	O157	O157		late	+	+
Ec186	ECOR36	Human	healthy	D	Δ, R	O157	O157	O157	O157	O157	O157		late	+	+
Ec192	ECOR42	Human	healthy	E	Δ, R	O157	O157	O157	O157	O157	O157		late	-	+
Ec196	ECOR46	Ape	healthy	D	Δ, R	O157	O157	O157	O157	O157	O157		late	+	+
Ec191	ECOR41	Human	healthy	D	Δ, R	O157	O157	O157	O157	O157	O157		late	-	+
Ec187	ECOR37	Marmoset	healthy	E	Δ, R	O157	O157	O157	O157	O157	O157		none	-	+

Table 6b: ECOR strains analyzed in the study

Strain[a]	original strain name[b]	strain source	strain type[c]	Phylo-genetic groups Herzer et.al., 1990[o]	bgl-Hall, 1988[p]	typing[d]	bgl/Z5211-Z5214 upstream sequence[e]	*bgl/Z5211 - Z5214 locus*[f]	*yieJ* PCR[g]	5'-yieI [h]	bgl/Z5211-Z5214 downstream sequence[i]	mutations[j]	Sal Phenotype 37°C[k]	c1955-c1960[l]	Lac Phenotype 37°C[m]
Ec188	ECOR38	Human	healthy	D	Δ, R	O157	O157	O157	O157	O157	O157		none	+	+
Ec189	ECOR39	Human	healthy	D	Δ, R	O157	O157	O157	O157	O157	O157		none	+	+
Ec190	ECOR40	Human	UTI (P)	D	Δ, R	O157	O157	O157	O157	O157	O157		none	+	+
Ec193	ECOR43	Human	healthy	E	Δ	O157	O157	O157	O157	O157	O157		none	-	-
Ec194	ECOR44	Cougar	healthy	D	+	O157	O157	O157	O157	O157	O157		none	-	+
Ec197	ECOR47	Sheep	healthy	D	Δ, R	O157	O157	O157	O157	O157	O157		none	-	+
Ec198	ECOR48	Human	UTI(C)	D	Δ, R	O157	O157	O157	O157	O157	O157		none	-	+
Ec200	ECOR50	Human	UTI (P)	D	Δ, R	O157	O157	O157	O157	O157	O157		none	+	+
Ec199	ECOR49	Human	healthy	D	Δ, R	O157	O157	O157	O157	O157	O157		weak Sal+	+	+

Table 6a and **Table 6b**: Isolates in Table 6a are obtained from Dr. Georg Plum, Institute for Medical Microbiology, University of Cologne, Cologne, Germany and isolates in Table 6b are obtained from STEC Center, Michigan State University, MI, USA. **a**: name of the strain in lab collection. **b**: original strain name **c**: All the isolates in Table 6a are isolated from human source. septicemic strains (sept) are isolates from the blood of the patients with septicaemia. Uropathogenic strains (uro) are isolates from urine samples of patients with urinary tract infections and commensal (com) strains are isolates from the stool samples of healthy individuals. Commensal *E.coli* isolates in Table 6b are from human/animal sources and pathogenic isolates are from human sources. UTI (C)- symptomatic urinary tract infection with acute cystitis, UTI (P)-urinary tract infection with acute pyelonephritis, ABU-asymptomatic bacteriuria**,** healthy-isolates from healthy humans/animas. **d**: Typing of the *E.coli* isolates based on the nucleotide sequences at the upstream, *bgl*/Z5211-Z5214 and downstream regions. **e:** sequence group with respect to the nucleotide sequence alignment of upstream sequences (*phoU* region). **f**: types based on the nucleotide sequence alignment of *bgl* operon or Z5211-Z5214. **g***: yieJ* PCR-MG1655 refers to the presence of *yieJ* gene as like in strain MG1655, CFT073 refers to the absence of *yieJ* gene as in strain CFT073, O157 refers to absence of *yieJ* gene as in strain O157. **h:** nucleotide sequence at the 5' end of the *yieI* gene determined by PCR and sequencing (for fourth and fifth type strains) **i:** sequence groups with respect to the nucleotide sequence alignment of downstream sequences (*yieJIH* region). **j:** mutations seen in the *bgl*/Z5211-Z5214 region. **k:** phenotypes on BTB salicin indicator plates at 37°C and at 28°C. +++ indicates relaxed phenotype, ++ more papillae, + MG1655 like papillae, none-no papillae and late-late papillation seen after day5 incubation, weak sal^+-Bgl^+ phenotype at day 2 incubation. **l:** c1955-c1960 locus analysis-+ indicates presence of c1955-c1960 region, - indicates absence. Alterations/mutations within the c1955-c1960 region are shown. **m:** phenotypes on MacConkey lactose plates at 37°C. + indicates Lac^+, - indicates Lac^-. **n:** genotypes relative to the *lac* operon is shown. **o:** Phylogenetic groups of ECOR strains based on Herzer et.al., 1990. **p:** *bgl* region analysis of the ECOR strains reported earlier by Hall., 1988. + indicates presence of *bgl* operon, - absence, Δ-carries *bgl* deletion, R-carries replacement in the *bgl* region. For more details see results section.

Table 7: miniTn10-Cm^R mutants and Sal$^+$ mutants analyzed in this study

strain[a]	relevant genotype or structure[b]	construction[c]/ reference
Ec2	i484 Bgl$^+$#1, Δ47 *bgl*	results section 1.11, Fig. 16
Ec3	i484 Bgl$^+$#2, *bgl*-CRP-C234 (C-66→T)	results section 1.11, Fig. 16
Ec4	i484 Bgl$^+$#3, no change in *bgl* promoter region	results section 1.11, Fig. 16
Ec5	i484 Bgl$^+$#4, no change in *bgl* promoter region	results section 1.11, Fig. 16
Ec6	i484 Bgl$^+$#5, no change in *bgl* promoter region	results section 1.11, Fig. 16
Ec7	i484 Bgl$^+$#6, no change in *bgl* promoter region	results section 1.11, Fig. 16
Ec8	i484 Bgl$^+$#7, no change in *bgl* promoter region	results section 1.11, Fig. 16
Ec93	i484 Δ*bgl*	results section 2.2, Fig. 23
Ec131	=Ec93 Sal$^+$, Cel$^+$, Arb$^+$, Esc$^+$ at 28˚C and Sal$^-$, Cel$^-$, Arb$^-$, Esc$^-$ at 37˚C	results section 2.2, Fig. 23
Ec132	=Ec93 Sal$^+$, Cel$^+$, Arb$^+$, Esc$^+$ at 28˚C and Sal$^-$, Cel$^-$, Arb$^-$, Esc$^-$ at 37˚C	results section 2.2, Fig. 23
Ec133	=Ec93 Sal$^+$, Cel$^+$, Arb$^+$, Esc$^+$ at 28˚C and Sal$^-$, Cel$^-$, Arb$^-$, Esc$^-$ at 37˚C	results section 2.2, Fig. 23
Ec134	=Ec93 Sal$^+$, Cel$^+$, Arb$^+$, Esc$^+$ at 28˚C and Sal$^-$, Cel$^-$, Arb$^-$, Esc$^-$ at 37˚C	results section 2.2, Fig. 23
Ec135	Ec131 c1958::mTn10-cm^R, AE016760: 291472-291480	results section 2.2, Fig. 23
Ec136	Ec131 c1957::mTn10-cm^R, AE016760: 291646-291654	results section 2.2, Fig. 23
Ec137	Ec131 c1957::mTn10-cm^R, AE016760: 291646-291654	results section 2.2, Fig. 23
Ec138	Ec131 mTn10-cm^R	results section 2.2, Fig. 23
Ec139	Ec132 c1958::mTn10-cm^R, AE016760: 291205-291213	results section 2.2, Fig. 23
Ec140	Ec132 *pyrD*::mTn10-cm^R, AE016758: 147545-147553	results section 2.2, Fig. 23
Ec141	Ec132 mTn10-cm^R	results section 2.2, Fig. 23
Ec142	Ec132 mTn10-cm^R	results section 2.2, Fig. 23
Ec143	Ec133 c1956::mTn10-cm^R, AE016760:288797-288805	results section 2.2, Fig. 23
Ec144	Ec133 c1956::mTn10-cm^R, AE016760:290312-290320	results section 2.2, Fig. 23
Ec145	Ec133 mTn10-cm^R	results section 2.2, Fig. 23
Ec146	Ec133 mTn10-cm^R	results section 2.2, Fig. 23
Ec147	Ec134 c1959::mTn10-cm^R, AE016760: 292417-292425	results section 2.2, Fig. 23
Ec148	Ec134 (dksA-sfsA)::mTn10-cm^R, AE016760: 171175-171183	results section 2.2, Fig. 23
Ec149	Ec133 mTn10-cm^R	results section 2.2, Fig. 23
Ec150	Ec133 mTn10-cm^R	results section 2.2, Fig. 23
Ec223	i484 *bgl*::mTn10-cm^R	results section 1.15, Fig. 21
Ec224	i484 *bgl*::mTn10-cm^R	results section 1.15, Fig. 21
Ec225	i484 *cysJ*::mTn10-cm^R, AE016765: 141948	results section 1.15, Fig. 21
Ec226	i484 *cysN*::mTn10-cm^R, AE016765: 135774	results section 1.15, Fig. 21
Ec227	i484 *purC*::mTn10-cm^R, AE016764: 146222	results section 1.15, Fig. 21
Ec228	i484 *bgl*::mTn10-cm^R	results section 1.15, Fig. 21
Ec229	i484 *bgl*::mTn10-cm^R	results section 1.15, Fig. 21
Ec230	i484 *ybdM*::mTn10-cm^R, AE016762: 215048	results section 1.15, Fig. 21
Ec231	i484 *bgl*::mTn10-cm^R	results section 1.15, Fig. 21
Ec232	i484 *cysH*::mTn10-cm^R, AE016765: 139681	results section 1.15, Fig. 21
Ec233	i484 *cysH*::mTn10-cm^R, AE016765: 139499	results section 1.15, Fig. 21
Ec234	i484 *bgl*::mTn10-cm^R	results section 1.15, Fig. 21
Ec235	i484 *purL*::mTn10-cm^R, AE016764: 235798	results section 1.15, Fig. 21
Ec236	i484 *bgl*::mTn10-cm^R	results section 1.15, Fig. 21
Ec237	i484 *bgl*::mTn10-cm^R	results section 1.15, Fig. 21
Ec238	i484 *cysH*::mTn10-cm^R, AE016765: 139681	results section 1.15, Fig. 21
Ec239	i484 *bgl*::mTn10-cm^R	results section 1.15, Fig. 21
Ec240	i484 mixed phenotype Sal$^+$Cel$^-$ #1	results section 1.16, Fig. 22
Ec241	i484 mixed phenotype Sal$^+$Cel$^-$ #2	results section 1.16, Fig. 22
Ec242	Ec240 *ybdM*::mTn10-cm^R, AE016762: 215064	results section 1.16, Fig. 22
Ec243	Ec240 *bglF*::mTn10-cm^R	results section 1.16, Fig. 22

Table 7: miniTn10-Cm^R mutants and Sal$^+$ mutants analyzed in this study

strain[a]	relevant genotype or structure[b]	construction[c]/ reference
Ec244	Ec240 (*bglG-bglF*):: mTn10-cm^R	results section 1.16, Fig. 22
Ec245	Ec241 *bglF*::mTn10-cm^R	results section 1.16, Fig. 22
Ec246	Ec241 *pgi*::mTn10-cm^R, AE016770: 235679	results section 1.16, Fig. 22
Ec247	Ec241 *bglF*::mTn10-cm^R	results section 1.16, Fig. 22
Ec248	Ec241 *bglF*::mTn10-cm^R	results section 1.16, Fig. 22
Ec249	Ec241 *pgi*::mTn10-cm^R, AE016770: 236173	results section 1.16, Fig. 22
Ec250	Ec241 *bglF*::mTn10-cm^R	results section 1.16, Fig. 22
Ec251	Ec241 *bglF*::mTn10-cm^R	results section 1.16, Fig. 22
Ec252	Ec241 *bglG*::mTn10-cm^R	results section 1.16, Fig. 22
Ec253	Ec241 *bglF*::mTn10-cm^R	results section 1.16, Fig. 22
Ec254	E7370 mixed phenotype Sal$^+$Cel$^+$ #1	results section 1.16, Fig. 22
Ec255	E7370 mixed phenotype Sal$^+$Cel$^+$ #2	results section 1.16, Fig. 22
Ec256	Ec254 *cysG*:: mTn10-cm^R, AE016767: 299877	results section 1.16, Fig. 22
Ec257	Ec254 *bglF*::mTn10-cm^R	results section 1.16, Fig. 22
Ec258	Ec254 *bglB*::mTn10-cm^R	results section 1.16, Fig. 22
Ec259	Ec254 *bglF*::mTn10-cm^R	results section 1.16, Fig. 22
Ec260	Ec255 *serC*:: mTn10-cm^R, AE016758:108192	results section 1.16, Fig. 22
Ec261	Ec255 *bglB*::mTn10-cm^R	results section 1.16, Fig. 22
Ec262	Ec255 *bglB*::mTn10-cm^R	results section 1.16, Fig. 22
Ec263	Ec255 *bglB*::mTn10-cm^R	results section 1.16, Fig. 22
Ec264	Ec255 *bglF*::mTn10-cm^R	results section 1.16, Fig. 22
Ec265	Ec255 *bglF*::mTn10-cm^R	results section 1.16, Fig. 22
Ec266	Ec255 *carB*:: mTn10-cm^R, AE016755: 35300	results section 1.16, Fig. 22
Ec267	Ec255 *bglB*::mTn10-cm^R	results section 1.16, Fig. 22
Ec268	U3454 Bgl$^+$#1, *bglR*::IS1	results section 1.11, Fig. 16
Ec269	U3454 Bgl$^+$#2, *bglR*::IS1	results section 1.11, Fig. 16
Ec270	U3454 Bgl$^+$#3, *bglR*::IS629	results section 1.11, Fig. 16
Ec271	U3454 Bgl$^+$#4, *bglR*::IS1	results section 1.11, Fig. 16
Ec272	U3454 Bgl$^+$#5, *bglR*::IS1	results section 1.11, Fig. 16
Ec273	U3454 Bgl$^+$#6, *bglR*::IS1	results section 1.11, Fig. 16
Ec274	U3454 Bgl$^+$#7, *bglR*::IS1	results section 1.11, Fig. 16
Ec275	W7483 Bgl$^+$#1, no change in the promoter region	results section 1.11, Fig. 16
Ec276	W7483 Bgl$^+$#2, no change in the promoter region	results section 1.11, Fig. 16
Ec277	W7483 Bgl$^+$#3, no change in the promoter region	results section 1.11, Fig. 16
Ec278	W7483 Bgl$^+$#4, Δ72 *bgl*	results section 1.11, Fig. 16
Ec279	W7483 Bgl$^+$#5, no change in the promoter region	results section 1.11, Fig. 16
Ec280	W7483 Bgl$^+$#4, Δ72 *bgl*	results section 1.11, Fig. 16
Ec281	E10091 Bgl$^+$#1, no change in the promoter region	results section 1.11, Fig. 16
Ec282	E10091 Bgl$^+$#2, no change in the promoter region	results section 1.11, Fig. 16
Ec283	E10091 Bgl$^+$#3, no change in the promoter region	results section 1.11, Fig. 16
Ec284	E176 Bgl$^+$#1, no change in the promoter region	results section 1.11, Fig. 16
Ec285	U3372 Bgl$^+$#1, *bglR*::IS1	results section 1.11, Fig. 16
Ec286	U3372 Bgl$^+$#2, *bglR*::IS1	results section 1.11, Fig. 16

a: strain names in the lab collection.

b: The relevant genotype of the constructed derivatives refers to the *bgl*, c1955-c1960, *cysH, cysN, cysJ, cysG, serC, pgi, purC, purL, carB,* and *ybdM* loci. Mutations causing activation of the silent *bgl* operon include *bgl*-CRP (a C to T exchange in the CRP binding site at position -66, relative to the transcription start), *bglR*::IS1 (integration of IS1), bglR::IS629 (integration of IS629). Position of mn10-cm^R insertions are shown in GenBank Accession numbers.

c: details of the construction and isolation of mutants are schematically shown in the respective figures indicated.

Table 8: Plasmids used in the present work

name[a]	relevant structure/description and replicon/resistance[b]	Source/reference[c]
pLDR8	λ-repressor, ts-cI-857, *int* under-λ-P_R, pSC101-rep ts-kan^R	(Diederich et al., 1992)
pKES15	MG1655-P_{bgl} +54-*lacZ*	Lab collection
pKES65	F742-P_{bgl}-t_1-*bglG-lacZ*	Lab collection
pKES66	F775-P_{bgl}-t_1-*bglG-lacZ*	Lab collection
pKES67	F785-P_{bgl}-t_1-*bglG-lacZ*	Lab collection
pKES71	W7483-P_{bgl}-t_1-*bglG-lacZ*	Lab collection
pKES73	W9763-P_{bgl}-t_1-*bglG-lacZ*	Lab collection
pKES75	F1-P_{bgl}-t_1-*bglG-lacZ*	Lab collection
pKES76	F287-P_{bgl}-t_1-*bglG-lacZ*	Lab collection
pKES77	F385-P_{bgl}-t_1-*bglG-lacZ*	Lab collection
pKES78	F557-P_{bgl}-t_1-*bglG-lacZ*	Lab collection
pKES79	F911-P_{bgl}-t_1-*bglG-lacZ*	Lab collection
pKES80	F1215-P_{bgl}-t_1-*bglG-lacZ*	Lab collection
pKES81	V9343-P_{bgl}-t_1-*bglG-lacZ*	Lab collection
pKES82	V10744-P_{bgl}-t_1-*bglG-lacZ*	Lab collection
pKES83	MG1655-P_{bgl}-t_1-*bglG-lacZ*	Lab collection
pKES84	i484--P_{bgl}-t_1-*bglG-lacZ*	Lab collection
pKEKB30	MG1655-P_{bgl} +25-*lacZ*	(Dole et al., 2004)
pKESK18	cI-857-Tn10 transposase-miniTn10-cm^R in rep-ts-kan^R	Lab collection
pKESK22	*lacI*-q-*lacO3*-P_{tac}-*lacO1*-MCS-in pACYC-kan^R	Lab collection
pKESK24	*attB* integration vector (single BamHI and BglII site)	Lab collection
pKESK25	*attB* integration vector (single BamHI and BglII site)	Lab collection
pKESD8	wt-P_{bgl}-*bglG-lacZ*	(Dole et al., 2002)
pKESD9	CRP^+-P_{bgl}-*bglG-lacZ*	(Dole et al., 2002)
pKESD20	*lacUV5*-P_{bgl}-*bglG-lacZ*	(Dole et al., 2002)
pFDY52	wt-*bgl* region in pUC12	lab collection
pFDY284	pACYC kan^R vector	lab collection
pFDY217	*lacI-lacOP-lacY* in rep-ts tet^R	(Dole et al., 2002)
pFMAC11	rep^{ts}-tet^R, Δ*bgl* in *bgl* region	(Caramel and Schnetz, 1998)
pFDX733	wt-*bgl* operon, kan	(Schnetz et al., 1987)
pBR322	cloning vector	(Bolivar, 1978)
pHPΩ-Tc	cloning vector carrying-Ω-tet cassette	(Fellay et al., 1987)
pKEGN1	E10085-P_{bgl}-t_1-*bglG-lacZ*	V: pKES15, SalI-XbaI, phosphatase F: E10085, PCR S145/S201, SalI-XbaI
pKEGN2	U5107-P_{bgl}-t_1-*bglG-lacZ*	V: pKES15, SalI-XbaI, phosphatase F: U5107, PCR S145/S201, SalI-XbaI
pKEGN3	F1215-P_{bgl}-t_1-*bglG-lacZ*	V: pKES15, SalI-XbaI, phosphatase F: F1215, PCR S145/S201, SalI-XbaI
pKEGN4	E10077-P_{bgl}-t_1-*bglG-lacZ*	V: pKES15, SalI-XbaI, phosphatase F: E10077, PCR S145/S201, SalI-XbaI
pKEGN5	U2183-P_{bgl}-t_1-*bglG-lacZ*	V: pKES15, SalI-XbaI, phosphatase F: U2183, PCR S145/S201, SalI-XbaI
pKEGN6	W9887-P_{bgl}-t_1-*bglG-lacZ*	V: pKES15, SalI-XbaI, phosphatase F: W9887, PCR S145/S201, SalI-XbaI
pKEGN7	E10091-P_{bgl}-t_1-*bglG-lacZ*	V: pKES15, SalI-XbaI, phosphatase F: E10091, PCR S145/S201, SalI-XbaI
pKEGN8	U3454-P_{bgl}-t_1-*bglG-lacZ*	V: pKES15, SalI-XbaI, phosphatase F: U3454, PCR S145/S201, SalI-XbaI

Table 8: Plasmids used in the present work

name[a]	relevant structure/description and replicon/resistance[b]	Source/reference[c]
pKEGN9	W7483-P_{bgl}-t_1-*bglG-lacZ*	V: pKES15, SalI-XbaI, phosphatase F: W7483, PCR S145/S201, SalI-XbaI
pKEGN10	E10077-P_{bgl} +25-*lacZ*	V: pKES15, SalI-XbaI, phosphatase F: pKEGN4, PCR S145/S212, SalI-XbaI
pKEGN11	E422-P_{bgl} +25-*lacZ*	V: pKES15, SalI-XbaI, phosphatase F: E422, PCR S145/S212, SalI-XbaI
pKEGN12	E475-P_{bgl} +25-*lacZ*	V: pKES15, SalI-XbaI, phosphatase F: E475, PCR S145/S212, SalI-XbaI
pKEGN13	E7370-P_{bgl} +25-*lacZ*	V: pKES15, SalI-XbaI, phosphatase F: E7370, PCR S145/S212, SalI-XbaI
pKEGN14	E478-P_{bgl} +25-*lacZ*	V: pKES15, SalI-XbaI, phosphatase F: E478, PCR S145/S212, SalI-XbaI
pKEGN15	U3372-P_{bgl} +25-*lacZ*	V: pKES15, SalI-XbaI, phosphatase F: U3372, PCR S145/S212, SalI-XbaI
pKEGN16	F560-P_{bgl} +25-*lacZ*	V: pKES15, SalI-XbaI, phosphatase F: F560, PCR S145/S212, SalI-XbaI
pKEGN17	i484-P_{bgl} +25-*lacZ*	V: pKES15, SalI-XbaI, phosphatase F: pKES84, PCR S145/S212, SalI-XbaI
pKEGN32	i484-*bglH-yieH*-pBR-amp^R	V: pBR322, EcoRI-EcoRV, phosphatase F: i484, PCR S253/S336, EcoRI-Bst1107I
pKEGN33	U3372-P_{bgl}-t_1-*bglG-lacZ*	V: pKES15, SalI-XbaI, phosphatase F: U3372, PCR S145/S201, SalI-XbaI
pKEGN34	F911-P_{bgl}-t_1-*bglG-lacZ*	V: pKES15, SalI-XbaI, phosphatase F: F911, PCR S145/S201, SalI-XbaI
pKEGN35	E478-P_{bgl}-t_1-*bglG-lacZ*	V: pKES15, SalI-XbaI, phosphatase F: E478, PCR S145/S201, SalI-XbaI
pKEGN36	E10094-P_{bgl}-t_1-*bglG-lacZ*	V: pKES15, SalI-XbaI, phosphatase F: E10094, PCR S145/S201, SalI-XbaI
pKEGN37	E475-P_{bgl}-t_1-*bglG-lacZ*	V: pKES15, SalI-XbaI, phosphatase F: E475, PCR S145/S201, SalI-XbaI
pKEGN38	E7370-P_{bgl}-t_1-*bglG-lacZ*	V: pKES15, SalI-XbaI, phosphatase F: E7370, PCR S145/S201, SalI-XbaI
pKEGN39	E422-P_{bgl}-t_1-*bglG-lacZ*	V: pKES15, SalI-XbaI, phosphatase F: E422, PCR S145/S201, SalI-XbaI
pKEGN40	W9887-P_{bgl}-t_1-*bglG-lacZ*	V: pKES15, SalI-XbaI, phosphatase F: W9887, PCR S145/S212, SalI-XbaI
pKEGN41	Ec93 (i484 Δ*bgl*)-c1955-c1960-kan^R	V: pFDY284, SalI-EcoRI F: Ec93, PCR S367/S368, 6370bp, XhoI-EcoRI
pKEGN42	Ec131 (Ec93 Sal$^+$ Cel$^+$ Arb$^+$ Esc$^+$)-c1955-c1960-kan^R	V: pFDY284, SalI-EcoRI F: Ec131, PCR S367/S368, 6370bp, XhoI-EcoRI
pKEGN43	Ec132 (Ec93 Sal$^+$ Cel$^+$ Arb$^+$ Esc$^+$)-c1955-c1960-kan^R	V: pFDY284, SalI-EcoRI F: Ec132, PCR S367/S368, 6370bp, XhoI-EcoRI
pKEGN44	Ec133 (Ec93 Sal$^+$ Cel$^+$ Arb$^+$ Esc$^+$) c1955-c1960-kan^R	V: pFDY284, SalI-EcoRI F: Ec133, PCR S367/S368, 6370bp, XhoI-EcoRI
pKEGN45	Ec134 (Ec93 Sal$^+$ Cel$^+$ Arb$^+$ Esc$^+$) c1955-c1960-kan^R	V: pFDY284, SalI-EcoRI F: Ec134, PCR S367/S368, 6370bp, XhoI-EcoRI
pKEGN46	Ec93-$P_{c1955\text{-}c1960}$-*lacZ*	V: pKES15, SalI-XbaI F: Ec93, PCR S432/S433, SalI-XbaI
pKEGN47	Ec93-c1960-$P_{c1955\text{-}c1960}$-*lacZ*	V: pKES15, SalI-XbaI F: Ec93, PCR S432/S368, SalI-XbaI
pKEGN48	Ec134-$P_{c1955\text{-}c1960}$-*lacZ*	V: pKES15, SalI-XbaI F: Ec134, PCR S432/S433, SalI-XbaI

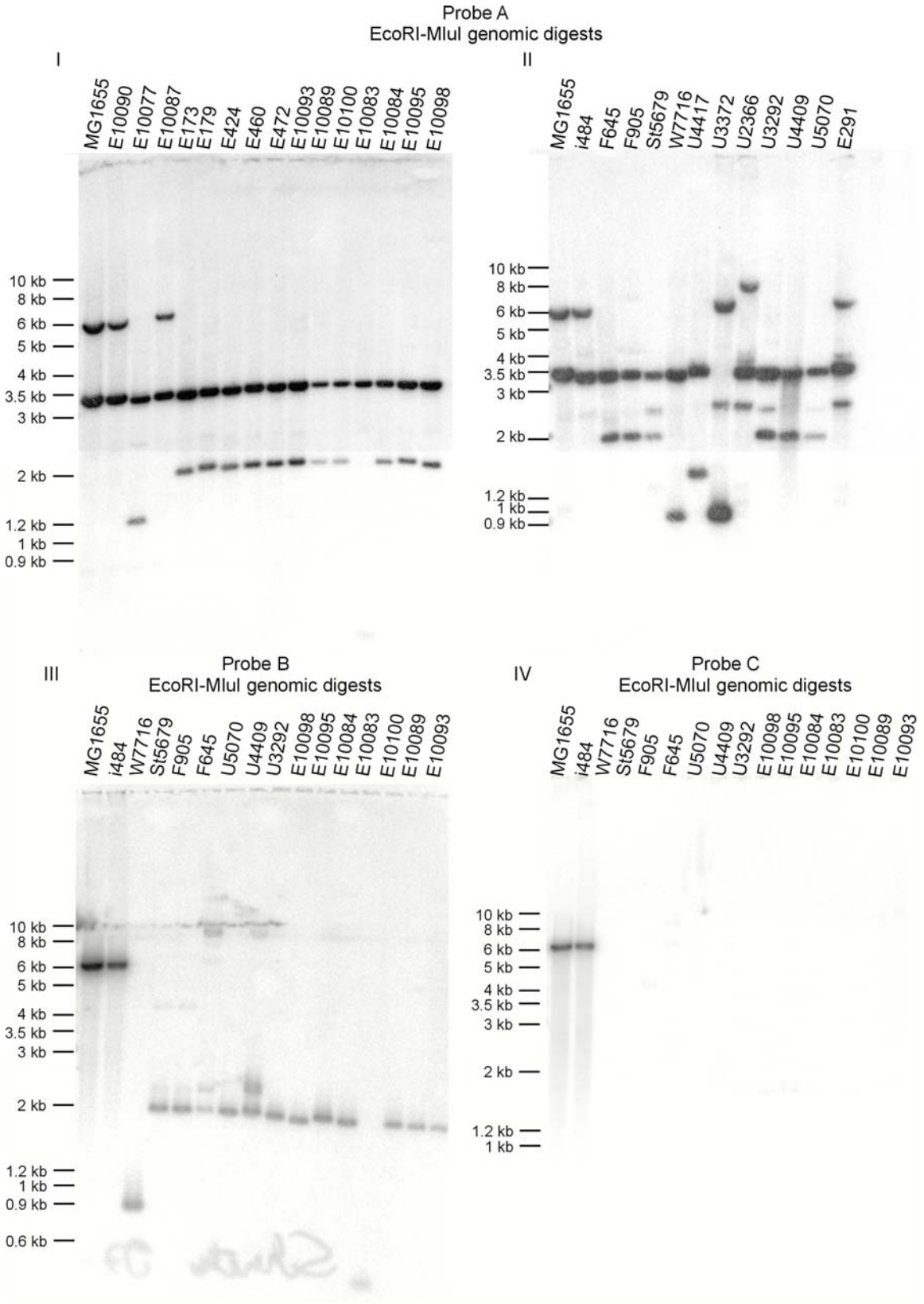
Table 8: Plasmids used in the present work

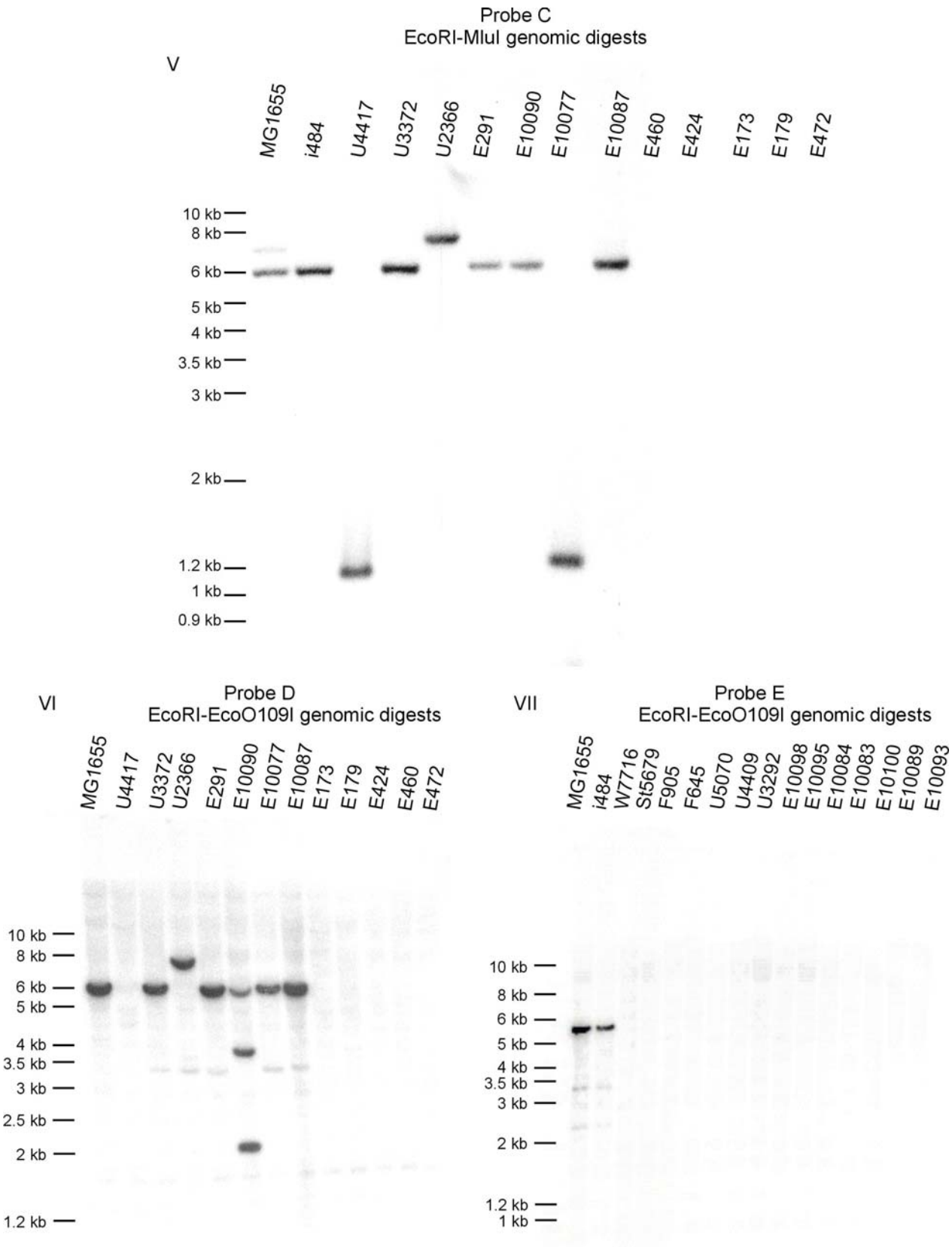
name[a]	relevant structure/description and replicon/resistance[b]	Source/reference[c]
pKEGN49	Ec134-c1960-$P_{c1955\text{-}c1960}$-*lacZ*	V: pKES15, SalI-XbaI F: Ec134, PCR S432/S368, SalI-XbaI
pKEGN50	E10085-*lacI*, *lacZ*-kan^R	V: pFDY284, SalI-BamHI F: PCR E10085 with S435/S436, SalI-BamHI,
pKEGN51	i484 Δ*bgl*-c1960-$P_{c1955\text{-}c1960}$-*lacZ*	V: pKEGN47, SalI-XbaI F: pKEGN29, SalI-XbaI
pKEGN52	Ec134-c1960-$P_{c1955\text{-}c1960}$-*lacZ*	V: pKEGN49, SalI-XbaI F: pKEGN29, SalI-XbaI
pKEGN53	Ec132-$lacI^q$-P_{tac}-c1959-c1955	V: pKESK22, EcoRI-XbaI F: Ec132, PCR S490/S491, EcoRI-XbaI
pKEGN54	Ec132- $lacI^q$-P_{tac}-c1959-c1955	V: pKESK22, EcoRI-XbaI F: Ec134, PCR S490/S491, EcoRI-XbaI
pKEGN55	Ec93- $lacI^q$-P_{tac}-c1959-c1955	V: pKESK22, EcoRI-XbaI F: i484 Δ*bgl*, PCR S490/S491, EcoRI-XbaI

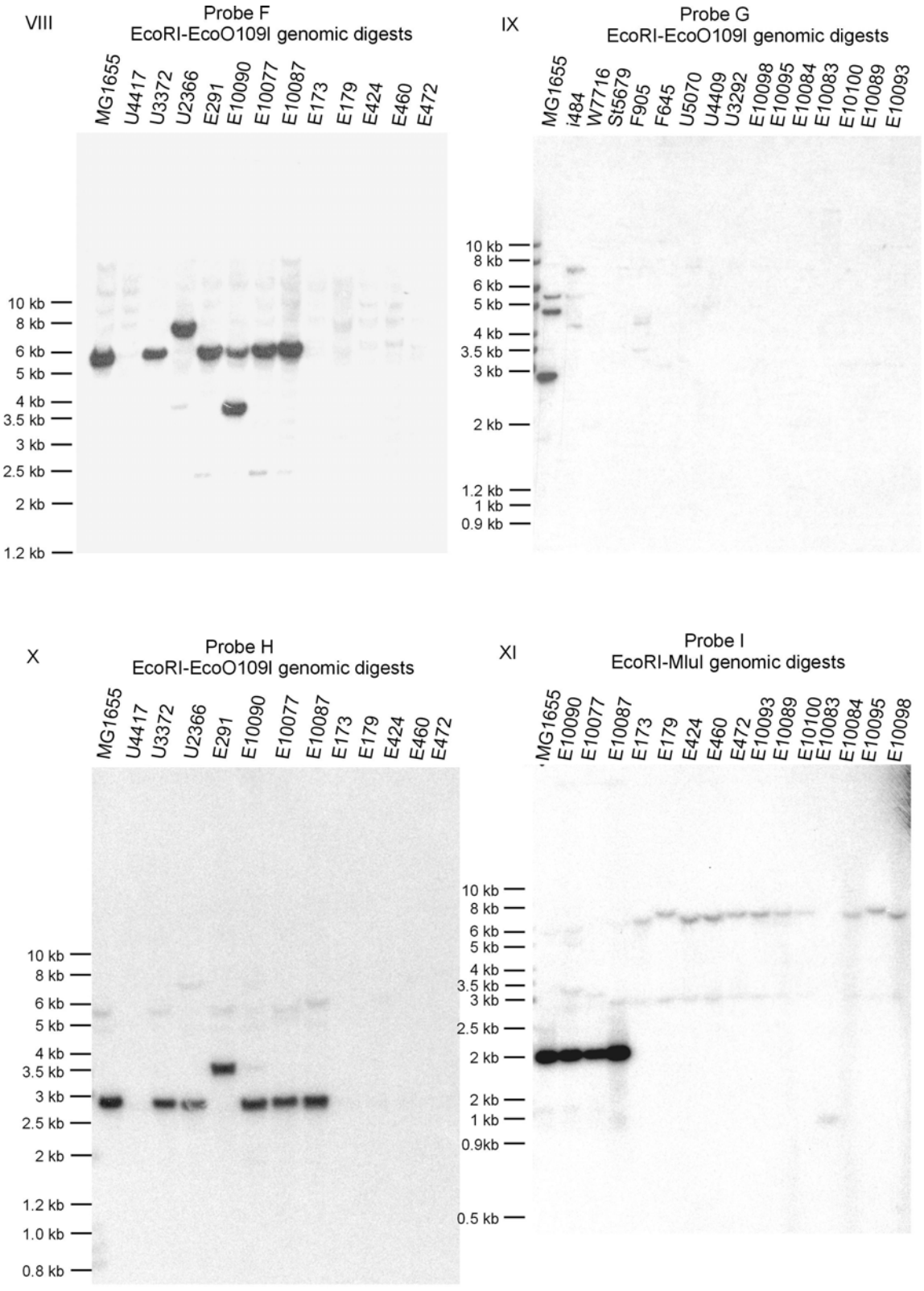
a: name of the plasmid in the lab collection

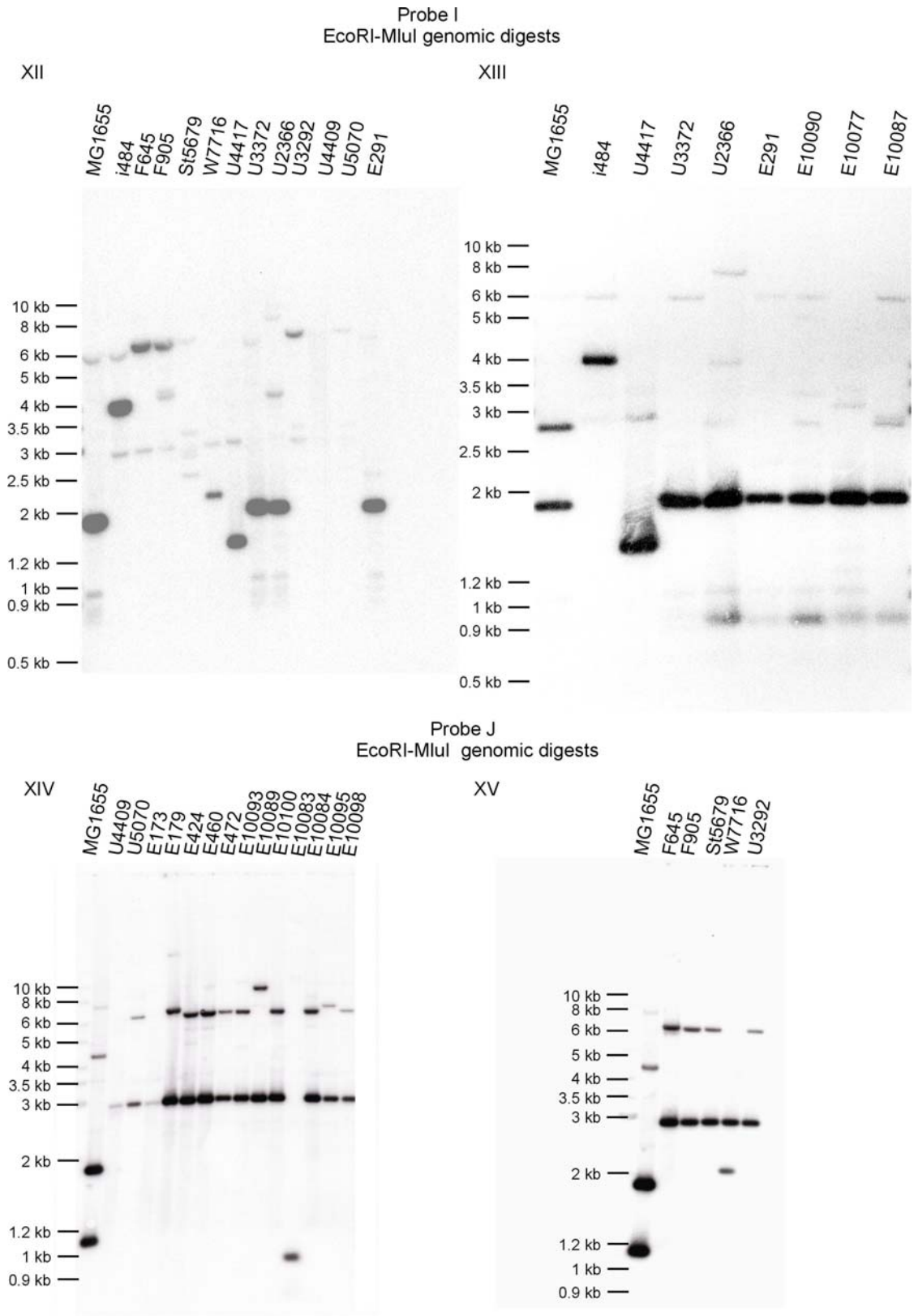
b: Number stated in the P_{bgl}-*lacZ* fusions is relative to the transcription start site of the *bgl* operon. t_1 indicates *bgl* terminator t_1. Mutations causing activation of the silent *bgl* operon include *bgl*-CRP^+ (a C to T exchange in the CRP binding site at position -66, relative to the transcription start) and *bgl*-Δ (a deletion of the upstream silencer, extending from position -77). Strain names shown are listed as in Table 5 and 6 (for e.g. pKEGN1 carries P_{bgl} fragment from strain E10085). Unless otherwise indicated plasmids are pACYC derivatives (Chang and Cohen, 1978) carrying a neo (kan^R), attP for integration into the K-12 chromosome and Ω $spec^R$ cassette. $P_{c1955\text{-}c1960}$ indicates the putative promoter region from c1955-c1960 system (position relative to CFT073 AE016760: 292674-292476). Expression of the Constructs carrying the constitutive promoter P_{tac} can be induced with IPTG. The inducible promoter P_{tac} is in between *lac* operator's *lacO3* and *lacO1* and is under the control of *lacI* gene. Plasmids carrying pSC101 replication origin (Hashimotoh-Gotoh et al., 1981) or pBR replication origin (Bolivar, 1978) and chloramphenicol (cm) or tetracycline (tet), or ampicillin (amp) resistance markers are indicted.

c: A short description of plasmid construction. V indicates the vector fragments. F indicates the insert fragment preparation. PCR reactions are indicated in order, PCR, template and primers used. Detailed plasmid construction descriptions are documented in lab records and all sequences are compiled in Vector NTi. EtOH ppt indicates ethanol precipitation.


Probe A
EcoRI-MluI genomic digests
I
MG1655
E10090
E10077
E10087
E173
E179
E424
E460
E472
E10093
E10089
E10100
E10083
E10084
E10095
E10098
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2 kb
1.2 kb
1 kb
0.9 kb
II
MG1655
i484
F645
F905
St5679
W7716
U4417
U3372
U2366
U3292
U4409
U5070
E291
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2 kb
1.2 kb
1 kb
0.9 kb
III
Probe B
EcoRI-MluI genomic digests
MG1655
i484
W7716
St5679
F905
F645
U5070
U4409
U3292
E10098
E10095
E10084
E10083
E10100
E10089
E10093
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2 kb
1.2 kb
1 kb
0.9 kb
0.6 kb
IV
Probe C
EcoRI-MluI genomic digests
MG1655
i484
W7716
St5679
F905
F645
U5070
U4409
U3292
E10098
E10095
E10084
E10083
E10100
E10089
E10093
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2 kb
1.2 kb
1 kb


Probe C
EcoRI-MluI genomic digests
V
MG1655
i484
U4417
U3372
U2366
E291
E10090
E10077
E10087
E460
E424
E173
E179
E472
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2 kb
1.2 kb
1 kb
0.9 kb
VI
Probe D
EcoRI-EcoO109I genomic digests
MG1655
U4417
U3372
U2366
E291
E10090
E10077
E10087
E173
E179
E424
E460
E472
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2.5 kb
2 kb
1.2 kb
VII
Probe E
EcoRI-EcoO109I genomic digests
MG1655
i484
W7716
St5679
F905
F645
U5070
U4409
U3292
E10098
E10095
E10084
E10083
E10100
E10089
E10093
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2 kb
1.2 kb
1 kb


VIII
Probe F
EcoRI-EcoO109I genomic digests
MG1655
U4417
U3372
U2366
E291
E10090
E10077
E10087
E173
E179
E424
E460
E472
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2.5 kb
2 kb
1.2 kb
IX
Probe G
EcoRI-EcoO109I genomic digests
MG1655
i484
W7716
St5679
F905
F645
U5070
U4409
U3292
E10098
E10095
E10084
E10083
E10100
E10089
E10093
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2 kb
1.2 kb
1 kb
0.9 kb
X
Probe H
EcoRI-EcoO109I genomic digests
MG1655
U4417
U3372
U2366
E291
E10090
E10077
E10087
E173
E179
E424
E460
E472
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2.5 kb
2 kb
1.2 kb
1.0 kb
0.8 kb
XI
Probe I
EcoRI-MluI genomic digests
MG1655
E10090
E10077
E10087
E173
E179
E424
E460
E472
E10093
E10089
E10100
E10083
E10084
E10095
E10098
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2.5 kb
2 kb
2 kb
1 kb
0.9kb
0.5 kb


Probe I
EcoRI-MluI genomic digests
XII
MG1655
i484
F645
F905
St5679
W7716
U4417
U3372
U2366
U3292
U4409
U5070
E291
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2.5 kb
2 kb
1.2 kb
1 kb
0.9 kb
0.5 kb
XIII
MG1655
i484
U4417
U3372
U2366
E291
E10090
E10077
E10087
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2.5 kb
2 kb
1.2 kb
1 kb
0.9 kb
0.5 kb
Probe J
EcoRI-MluI genomic digests
XIV
MG1655
U4409
U5070
E173
E179
E424
E460
E472
E10093
E10089
E10100
E10083
E10084
E10095
E10098
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2 kb
1.2 kb
1 kb
0.9 kb
XV
MG1655
F645
F905
St5679
W7716
U3292
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2 kb
1.2 kb
1 kb
0.9 kb

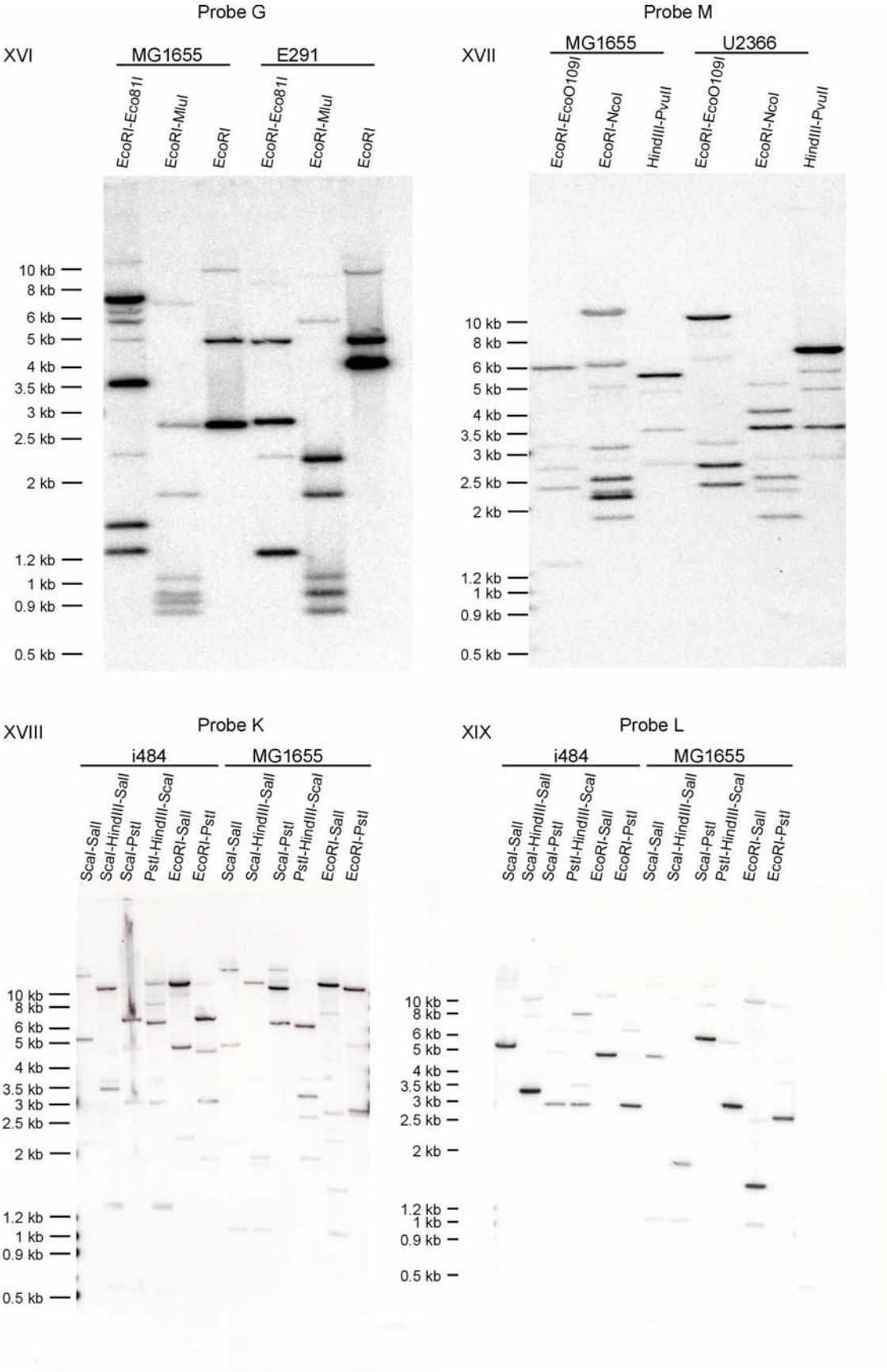

Probe G
XVI
MG1655
E291
EcoRI-Eco81I
EcoRI-MluI
EcoRI
EcoRI-Eco81I
EcoRI-MluI
EcoRI
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2.5 kb
2 kb
1.2 kb
1 kb
0.9 kb
0.5 kb
Probe M
XVII
MG1655
U2366
EcoRI-EcoO109I
EcoRI-NcoI
HindIII-PvuII
EcoRI-EcoO109I
EcoRI-NcoI
HindIII-PvuII
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2.5 kb
2 kb
1.2 kb
1 kb
0.9 kb
0.5 kb
XVIII
Probe K
i484
MG1655
ScaI-SalI
ScaI-HindIII-SalI
ScaI-PstI
PstI-HindIII-ScaI
EcoRI-SalI
EcoRI-PstI
ScaI-SalI
ScaI-HindIII-SalI
ScaI-PstI
PstI-HindIII-ScaI
EcoRI-SalI
EcoRI-PstI
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2.5 kb
2 kb
1.2 kb
1 kb
0.9 kb
0.5 kb
XIX
Probe L
i484
MG1655
ScaI-SalI
ScaI-HindIII-SalI
ScaI-PstI
PstI-HindIII-ScaI
EcoRI-SalI
EcoRI-PstI
ScaI-SalI
ScaI-HindIII-SalI
ScaI-PstI
PstI-HindIII-ScaI
EcoRI-SalI
EcoRI-PstI
10 kb
8 kb
6 kb
5 kb
4 kb
3.5 kb
3 kb
2.5 kb
2 kb
1.2 kb
1 kb
0.9 kb
0.5 kb

Figure 41: Southern hybridization images. Enzymes used in the genomic digestions and probes are indicated. Strain names are as listed in Table 6 (Appendix). I and II) Genomic DNA digested with EcoRI and MluI and probed with probe A (*phoU* gene, *NcoI-SspI* fragment from pFDY52). III) EcoRI and MluI genomic digests and probed with probe B (*phoU*-P_{bgl}-*bglG*, S145/S201 PCR fragment from MG1655). IV and V) EcoRI and MluI digests probed with probe C (P_{bgl}-*bglG*, S157/S201 PCR fragment from MG1655). VI) EcoRI-EcoO109I genomic digests probed with probe D (*bglF*, DraI-EcoRV fragment from pFDX733) VII) EcoRI-EcoO109I genomic digests probed with probe E (*bglF-bglB*, NcoI-NcoI fragment from pFDX733) VIII) EcoRI-EcoO109I genomic digests probed with probe F (*bglB*, PvuII-HindIII fragment from pFDX733) IX) EcoRI-EcoO109I genomic digests probed with probe G (bglHIK, EcoRI-EcoRI fragment from pFDY52) X) EcoRI-EcoO109I genomic digests probed with probe H (bglIK, Eco81I-PvuII fragment from pFDY52). XI, XII and XIII) EcoRI-MluI genomic digests probed with probe I (yieJIH, EcoRI-SalI fragment from pFDY52). XIV and XV) EcoRI-MluI genomic digests probed with probe J (yieHG, PCR S335/S336 fragment from MG1655) XVI) EcoRI-Eco81I, EcoRI-MluI and EcoRI genomic digests of MG1655 (as control) and strain E291, probed with probe G. XVII) EcoRI-EcoO109I, EcoRI-NcoI and HindIII-PvuII genomic digests probed with probe M (*bglGFBH*, Eco01091-EcoRI fragment from pFDY52). XVIII) ScaI-SalI, ScaI-HindIII-SalI, ScaI-PstI, PstI-HindIII-SalI, EcoRI-SalI, EcoRI-PstI genomic digests from MG1655 (control) and i484 probed with probe K (*bglH*, ScaI-SapI fragment from pFDY52). XIX) ScaI-SalI, ScaI-HindIII-SalI, ScaI-PstI, PstI-HindIII-SalI, EcoRI-SalI, EcoRI-PstI genomic digests from MG1655 (control) and i484 probed with probe L (*yieIHG*, BglII-SalI fragment from pFDY52). Refer Figure 11 and 12 (Results) for detailed descriptions.

VII. Bibliography

Akman, L., and Aksoy, S. (2001). A novel application of gene arrays: Escherichia coli array provides insight into the biology of the obligate endosymbiont of tsetse flies. Proc Natl Acad Sci U S A *98*, 7546-7551.

An, C. L., Lim, W. J., Hong, S. Y., Kim, E. J., Shin, E. C., Kim, M. K., Lee, J. R., Park, S. R., Woo, J. G., Lim, Y. P., and Yun, H. D. (2004). Analysis of bgl Operon Structure and Characterization of beta-Glucosidase from Pectobacterium carotovorum subsp. carotovorum LY34. Biosci Biotechnol Biochem *68*, 2270-2278.

Andersen, C., Rak, B., and Benz, R. (1999). The gene bglH present in the bgl operon of Escherichia coli, responsible for uptake and fermentation of beta-glucosides encodes for a carbohydrate-specific outer membrane porin. Mol Microbiol *31*, 499-510.

Banks, D. J., Porcella, S. F., Barbian, K. D., Beres, S. B., Philips, L. E., Voyich, J. M., DeLeo, F. R., Martin, J. M., Somerville, G. A., and Musser, J. M. (2004). Progress toward characterization of the group A Streptococcus metagenome: complete genome sequence of a macrolide-resistant serotype M6 strain. J Infect Dis *190*, 727-738.

Bass, S., Gu, Q., and Christen, A. (1996). Multicopy suppressors of prc mutant Escherichia coli include two HtrA (DegP) protease homologs (HhoAB), DksA, and a truncated R1pA. J Bacteriol *178*, 1154-1161.

Bauer, M. E., and Welch, R. A. (1997). Pleiotropic effects of a mutation in rfaC on Escherichia coli hemolysin. Infect Immun *65*, 2218-2224.

Beres, S. B., Sylva, G. L., Barbian, K. D., Lei, B., Hoff, J. S., Mammarella, N. D., Liu, M. Y., Smoot, J. C., Porcella, S. F., Parkins, L. D.*, et al.* (2002). Genome sequence of a serotype M3 strain of group A Streptococcus: phage-encoded toxins, the high-virulence phenotype, and clone emergence. Proc Natl Acad Sci U S A *99*, 10078-10083.

Bergthorsson, U., and Ochman, H. (1998). Distribution of chromosome length variation in natural isolates of Escherichia coli. Mol Biol Evol *15*, 6-16.

Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F.*, et al.* (1997). The complete genome sequence of Escherichia coli K-12. Science *277*, 1453-1474.

Blum, G., Ott, M., Lischewski, A., Ritter, A., Imrich, H., Tschape, H., and Hacker, J. (1994). Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an Escherichia coli wild-type pathogen. Infect Immun *62*, 606-614.

Bolivar, F. (1978). Construction and characterization of new cloning vehicles. III. Derivatives of plasmid pBR322 carrying unique Eco RI sites for selection of Eco RI generated recombinant DNA molecules. Gene *4*, 121-136.

Bolotin, A., Wincker, P., Mauger, S., Jaillon, O., Malarme, K., Weissenbach, J., Ehrlich, S. D., and Sorokin, A. (2001). The complete genome sequence of the lactic acid bacterium Lactococcus lactis ssp. lactis IL1403. Genome Res *11*, 731-753.

Bonacorsi, S. P., Clermont, O., Tinsley, C., Le Gall, I., Beaudoin, J. C., Elion, J., Nassif, X., and Bingen, E. (2000). Identification of regions of the Escherichia coli chromosome specific for neonatal meningitis-associated strains. Infect Immun *68*, 2096-2101.

Boos, W., and Shuman, H. (1998). Maltose/maltodextrin system of Escherichia coli: transport, metabolism, and regulation. Microbiol Mol Biol Rev *62*, 204-229.

Branny, P., Pearson, J. P., Pesci, E. C., Kohler, T., Iglewski, B. H., and Van Delden, C. (2001). Inhibition of quorum sensing by a Pseudomonas aeruginosa dksA homologue. J Bacteriol *183*, 1531-1539.

Brehm, K., Ripio, M. T., Kreft, J., and Vazquez-Boland, J. A. (1999). The bvr locus of Listeria monocytogenes mediates virulence gene repression by beta-glucosides. J Bacteriol *181*, 5024-5032.

Breves, R., Bronnenmeier, K., Wild, N., Lottspeich, F., Staudenbauer, W. L., and Hofemeister, J. (1997). Genes encoding two different beta-glucosidases of Thermoanaerobacter brockii are clustered in a common operon. Appl Environ Microbiol *63*, 3902-3910.

Brown, G. D., and Thomson, J. A. (1998). Isolation and characterisation of an aryl-beta-D-glucoside uptake and utilisation system (abg) from the gram-positive ruminal Clostridium species C. longisporum. Mol Gen Genet *257*, 213-218.

Brown, L., Gentry, D., Elliott, T., and Cashel, M. (2002). DksA affects ppGpp induction of RpoS at a translational level. J Bacteriol *184*, 4455-4465.

Caramel, A., and Schnetz, K. (1998). Lac and lambda repressors relieve silencing of the Escherichia coli bgl promoter. Activation by alteration of a repressing nucleoprotein complex. J Mol Biol *284*, 875-883.

Caramel, A., and Schnetz, K. (2000). Antagonistic control of the Escherichia coli bgl promoter by FIS and CAP in vitro. Mol Microbiol *36*, 85-92.

Cerdeno-Tarraga, A. M., Efstratiou, A., Dover, L. G., Holden, M. T., Pallen, M., Bentley, S. D., Besra, G. S., Churcher, C., James, K. D., De Zoysa, A.*, et al.* (2003). The complete genome sequence and analysis of Corynebacterium diphtheriae NCTC13129. Nucleic Acids Res *31*, 6516-6523.

Chang, A. C., and Cohen, S. N. (1978). Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid. J Bacteriol *134*, 1141-1156.

Chen, L., and Coleman, W. G., Jr. (1993). Cloning and characterization of the Escherichia coli K-12 rfa-2 (rfaC) gene, a gene required for lipopolysaccharide inner core synthesis. J Bacteriol *175*, 2534-2540.

Chun, K. T., Edenberg, H. J., Kelley, M. R., and Goebl, M. G. (1997). Rapid amplification of uncharacterized transposon-tagged DNA sequences from genomic DNA. Yeast *13*, 233-240.

Cote, C. K., Cvitkovitch, D., Bleiweis, A. S., and Honeyman, A. L. (2000). A novel beta-glucoside-specific PTS locus from Streptococcus mutans that is not inhibited by glucose. Microbiology *146 (Pt 7)*, 1555-1563.

Cote, C. K., and Honeyman, A. L. (2002). The transcriptional regulation of the Streptococcus mutans bgl regulon. Oral Microbiol Immunol *17*, 1-8.

Declerck, N., Minh, N. L., Yang, Y., Bloch, V., Kochoyan, M., and Aymerich, S. (2002). RNA recognition by transcriptional antiterminators of the BglG/SacY family: mapping of SacY RNA binding site. J Mol Biol *319*, 1035-1048.

Defez, R., and De Felice, M. (1981). Cryptic operon for beta-glucoside metabolism in Escherichia coli K12: genetic evidence for a regulatory protein. Genetics *97*, 11-25.

Deng, W., Burland, V., Plunkett, G., 3rd, Boutin, A., Mayhew, G. F., Liss, P., Perna, N. T., Rose, D. J., Mau, B., Zhou, S.*, et al.* (2002). Genome sequence of Yersinia pestis KIM. J Bacteriol *184*, 4601-4611.

Deng, W., Liou, S. R., Plunkett, G., 3rd, Mayhew, G. F., Rose, D. J., Burland, V., Kodoyianni, V., Schwartz, D. C., and Blattner, F. R. (2003). Comparative genomics of Salmonella enterica serovar Typhi strains Ty2 and CT18. J Bacteriol *185*, 2330-2337.

Diederich, L., Rasmussen, L. J., and Messer, W. (1992). New cloning vectors for integration in the lambda attachment site attB of the Escherichia coli chromosome. Plasmid *28*, 14-24.

DiNardo, S., Voelkel, K. A., Sternglanz, R., Reynolds, A. E., and Wright, A. (1982). Escherichia coli DNA topoisomerase I mutants have compensatory mutations in DNA gyrase genes. Cell *31*, 43-51.

Dobrindt, U., Agerer, F., Michaelis, K., Janka, A., Buchrieser, C., Samuelson, M., Svanborg, C., Gottschalk, G., Karch, H., and Hacker, J. (2003). Analysis of genome plasticity in pathogenic and commensal Escherichia coli isolates by use of DNA arrays. J Bacteriol *185*, 1831-1840.

Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol *2*, 414-424.

Dodsworth, S. J., Bennett, A. J., and Coleman, G. (1993). Molecular cloning and nucleotide sequence analysis of the maltose-inducible porin gene of Aeromonas salmonicida. FEMS Microbiol Lett *112*, 191-197.

Dole, S., Klingen, Y., Nagarajavel, V., and Schnetz, K. (2004). The protease Lon and the RNA-binding protein Hfq reduce silencing of the Escherichia coli bgl operon by H-NS. J Bacteriol *186*, 2708-2716.

Dole, S., Kuhn, S., and Schnetz, K. (2002). Post-transcriptional enhancement of Escherichia coli bgl operon silencing by limitation of BglG-mediated antitermination at low transcription rates. Mol Microbiol *43*, 217-226.

Dole, S., Nagarajavel, V., and Schnetz, K. (2004). The histone-like nucleoid structuring protein H-NS represses the Escherichia coli bgl operon downstream of the promoter. Mol Microbiol *52*, 589-600.

Duchaud, E., Rusniok, C., Frangeul, L., Buchrieser, C., Givaudan, A., Taourit, S., Bocs, S., Boursaux-Eude, C., Chandler, M., Charles, J. F.*, et al.* (2003). The genome sequence of the entomopathogenic bacterium Photorhabdus luminescens. Nat Biotechnol *21*, 1307-1313.

el Hassouni, M., Chippaux, M., and Barras, F. (1990). Analysis of the Erwinia chrysanthemi arb genes, which mediate metabolism of aromatic beta-glucosides. J Bacteriol *172*, 6261-6267.

Faure, D., Saier, M. H., Jr., and Vanderleyden, J. (2001). An evolutionary alternative system for aryl beta-glucosides assimilation in bacteria. J Mol Microbiol Biotechnol *3*, 467-470.

Falkow, S. (1996). Escherichia and Salmonella: cellular and molecular biology. 2nd ed, 2723-2729.

Fellay, R., Frey, J., and Krisch, H. (1987). Interposon mutagenesis of soil and water bacteria: a family of DNA fragments designed for in vitro insertional mutagenesis of gram-negative bacteria. Gene *52*, 147-154.

Finkel, S. E., and Johnson, R. C. (1992). The Fis protein: it's not just for DNA inversion anymore. Mol Microbiol *6*, 3257-3265.

Free, A., Williams, R. M., and Dorman, C. J. (1998). The StpA protein functions as a molecular adapter to mediate repression of the bgl operon by truncated H-NS in Escherichia coli. J Bacteriol *180*, 994-997.

Fujishima, Y., and Yamane, K. (1995). A 10 kb nucleotide sequence at the 5' flanking region (32 degrees) of srfAA of the Bacillus subtilis chromosome. Microbiology *141 (Pt 2)*, 277-279.

Fukiya, S., Mizoguchi, H., Tobe, T., and Mori, H. (2004). Extensive genomic diversity in pathogenic Escherichia coli and Shigella Strains revealed by comparative genomic hybridization microarray. J Bacteriol *186*, 3911-3921.

Giel, M., Desnoyer, M., and Lopilato, J. (1996). A mutation in a new gene, bglJ, activates the bgl operon in Escherichia coli K-12. Genetics *143*, 627-635.

Grindley, N. D. (1978). IS1 insertion generates duplication of a nine base pair sequence at its target site. Cell *13*, 419-426.

Hacker, J., Blum-Oehler, G., Muhldorfer, I., and Tschape, H. (1997). Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol *23*, 1089-1097.

Hacker, J., and Carniel, E. (2001). Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. EMBO Rep *2*, 376-381.

Hall, B. G. (1988). Widespread distribution of deletions of the bgl operon in natural isolates of Escherichia coli. Mol Biol Evol *5*, 456-467.

Hall, B. G., and Xu, L. (1992). Nucleotide sequence, function, activation, and evolution of the cryptic asc operon of Escherichia coli K12. Mol Biol Evol *9*, 688-706.

Hashimoto-Gotoh, T., Franklin, F. C., Nordheim, A., and Timmis, K. N. (1981). Specific-purpose plasmid cloning vectors. I. Low copy number, temperature-sensitive, mobilization-defective pSC101-derived containment vectors. Gene *16*, 227-235.

Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C. G., Ohtsubo, E., Nakayama, K., Murata, T.*, et al.* (2001). Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res *8*, 11-22.

Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Clayton, R. A., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Umayam, L.*, et al.* (2000). DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae. Nature *406*, 477-483.

Hentschel, U., and Hacker, J. (2001). Pathogenicity islands: the tip of the iceberg. Microbes Infect *3*, 545-548.

Hentschel, U., Steinert, M., and Hacker, J. (2000). Common molecular mechanisms of symbiosis and pathogenesis. Trends Microbiol *8*, 226-231.

Herbelin, C. J., Chirillo, S. C., Melnick, K. A., and Whittam, T. S. (2000). Gene conservation and loss in the mutS-rpoS genomic region of pathogenic Escherichia coli. J Bacteriol *182*, 5381-5390.

Herzer, P. J., Inouye, S., Inouye, M., and Whittam, T. S. (1990). Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of Escherichia coli. J Bacteriol *172*, 6175-6181.

Hirsch, M., and Elliott, T. (2002). Role of ppGpp in rpoS stationary-phase regulation in Escherichia coli. J Bacteriol *184*, 5077-5087.

Ishii, Y., Yamada, H., Yamashino, T., Ohashi, K., Katoh, E., Shindo, H., Yamazaki, T., and Mizuno, T. (2000). Deletion of the yhhP gene results in filamentous cell morphology in Escherichia coli. Biosci Biotechnol Biochem *64*, 799-807.

Ivanova, N., Sorokin, A., Anderson, I., Galleron, N., Candelon, B., Kapatral, V., Bhattacharyya, A., Reznik, G., Mikhailova, N., Lapidus, A.*, et al.* (2003). Genome sequence of Bacillus cereus and comparative analysis with Bacillus anthracis. Nature *423*, 87-91.

Jin, Q., Yuan, Z., Xu, J., Wang, Y., Shen, Y., Lu, W., Wang, J., Liu, H., Yang, J., Yang, F.*, et al.* (2002). Genome sequence of Shigella flexneri 2a: insights into pathogenicity through comparison with genomes of Escherichia coli K12 and O157. Nucleic Acids Res *30*, 4432-4441.

Kang, P. J., and Craig, E. A. (1990). Identification and characterization of a new Escherichia coli gene that is a dosage-dependent suppressor of a dnaK deletion mutation. J Bacteriol *172*, 2055-2064.

Khan, M. A., and Isaacson, R. E. (1998). In vivo expression of the beta-glucoside (bgl) operon of Escherichia coli occurs in mouse liver. J Bacteriol *180*, 4746-4749.

Kharat, A. S., and Mahadevan, S. (2000). Analysis of the beta-glucoside utilization (bgl) genes of Shigella sonnei: evolutionary implications for their maintenance in a cryptic state. Microbiology *146 (Pt 8)*, 2039-2049.

Kilic, A. O., Tao, L., Zhang, Y., Lei, Y., Khammanivong, A., and Herzberg, M. C. (2004). Involvement of Streptococcus gordonii beta-glucoside metabolism systems in adhesion, biofilm formation, and in vivo gene expression. J Bacteriol *186*, 4246-4253.

Kleerebezem, M., Boekhorst, J., van Kranenburg, R., Molenaar, D., Kuipers, O. P., Leer, R., Tarchini, R., Peters, S. A., Sandbrink, H. M., Fiers, M. W.*, et al.* (2003). Complete genome sequence of Lactobacillus plantarum WCFS1. Proc Natl Acad Sci U S A *100*, 1990-1995.

Kruger, S., and Hecker, M. (1995). Regulation of the putative bglPH operon for aryl-beta-glucoside utilization in Bacillus subtilis. J Bacteriol *177*, 5590-5597.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A., Borchert, S.*, et al.* (1997). The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature *390*, 249-256.

Lai, X., Davis, F. C., Hespell, R. B., and Ingram, L. O. (1997). Cloning of cellobiose phosphoenolpyruvate-dependent phosphotransferase genes: functional expression in recombinant Escherichia coli and identification of a putative binding region for disaccharides. Appl Environ Microbiol *63*, 355-363.

Lawrence, J. G., and Ochman, H. (1998). Molecular archaeology of the Escherichia coli genome. Proc Natl Acad Sci U S A *95*, 9413-9417.

Lawrence, J. G., and Roth, J. R. (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics *143*, 1843-1860.

Le Coq, D., Lindner, C., Kruger, S., Steinmetz, M., and Stulke, J. (1995). New beta-glucoside (bgl) genes in Bacillus subtilis: the bglP gene product has both transport and regulatory functions similar to those of BglF, its Escherichia coli homolog. J Bacteriol *177*, 1527-1535.

Lecointre, G., Rachdi, L., Darlu, P., and Denamur, E. (1998). Escherichia coli molecular phylogeny using the incongruence length difference test. Mol Biol Evol *15*, 1685-1695.

Loewen, P. C., and Hengge-Aronis, R. (1994). The role of the sigma factor sigma S (KatF) in bacterial global regulation. Annu Rev Microbiol *48*, 53-80.

Mahadevan, S., Reynolds, A. E., and Wright, A. (1987). Positive and negative regulation of the bgl operon in Escherichia coli. J Bacteriol *169*, 2570-2578.

Manival, X., Yang, Y., Strub, M. P., Kochoyan, M., Steinmetz, M., and Aymerich, S. (1997). From genetic to structural characterization of a new class of RNA-binding domain within the SacY/BglG family of antiterminator proteins. Embo J *16*, 5019-5029.

Maurelli, A. T., Fernandez, R. E., Bloch, C. A., Rode, C. K., and Fasano, A. (1998). "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of Shigella spp. and enteroinvasive Escherichia coli. Proc Natl Acad Sci U S A *95*, 3943-3948.

McClelland, M., Sanderson, K. E., Spieth, J., Clifton, S. W., Latreille, P., Courtney, L., Porwollik, S., Ali, J., Dante, M., Du, F.*, et al.* (2001). Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. Nature *413*, 852-856.

Melkerson-Watson, L. J., Rode, C. K., Zhang, L., Foxman, B., and Bloch, C. A. (2000). Integrated genomic map from uropathogenic Escherichia coli J96. Infect Immun *68*, 5933-5942.

Milkman, R., and Bridges, M. M. (1993). Molecular evolution of the Escherichia coli chromosome. IV. Sequence comparisons. Genetics *133*, 455-468.

Miller, J.H. (1972). Experiments in Molecular Genetics. (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory).

Mukerji, M., and Mahadevan, S. (1997). Characterization of the negative elements involved in silencing the bgl operon of Escherichia coli: possible roles for DNA gyrase, H-NS, and CRP-cAMP in regulation. Mol Microbiol *24*, 617-627.

Nelson, K. E., Fouts, D. E., Mongodin, E. F., Ravel, J., DeBoy, R. T., Kolonay, J. F., Rasko, D. A., Angiuoli, S. V., Gill, S. R., Paulsen, I. T.*, et al.* (2004). Whole genome comparisons of serotype 4b and 1/2a strains of the food-borne pathogen Listeria monocytogenes reveal new insights into the core genome components of this species. Nucleic Acids Res *32*, 2386-2395.

Nolling, J., Breton, G., Omelchenko, M. V., Makarova, K. S., Zeng, Q., Gibson, R., Lee, H. M., Dubois, J., Qiu, D., Hitti, J.*, et al.* (2001). Genome sequence and comparative analysis of the solvent-producing bacterium Clostridium acetobutylicum. J Bacteriol *183*, 4823-4838.

Ochman, H., and Jones, I. B. (2000). Evolutionary dynamics of full genome content in Escherichia coli. Embo J *19*, 6637-6643.

Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature *405*, 299-304.

Ochman, H., and Selander, R. K. (1984). Standard reference strains of Escherichia coli from natural populations. J Bacteriol *157*, 690-693.

Ohnishi, M., Tanaka, C., Kuhara, S., Ishii, K., Hattori, M., Kurokawa, K., Yasunaga, T., Makino, K., Shinagawa, H., Murata, T.*, et al.* (1999). Chromosome of the enterohemorrhagic Escherichia coli O157:H7; comparative analysis with K-12 MG1655 revealed the acquisition of a large amount of foreign DNAs. DNA Res *6*, 361-368.

Ohta, T., Ueguchi, C., and Mizuno, T. (1999). rpoS function is essential for bgl silencing caused by C-terminally truncated H-NS in Escherichia coli. J Bacteriol *181*, 6278-6283.

Parker, L. L., and Hall, B. G. (1988). A fourth Escherichia coli gene system with the potential to evolve beta-glucoside utilization. Genetics *119*, 485-490.

Paul, B. J., Barker, M. M., Ross, W., Schneider, D. A., Webb, C., Foster, J. W., and Gourse, R. L. (2004). DksA: a critical component of the transcription initiation machinery that potentiates the regulation of rRNA promoters by ppGpp and the initiating NTP. Cell *118*, 311-322.

Paulsen, I. T., Banerjei, L., Myers, G. S., Nelson, K. E., Seshadri, R., Read, T. D., Fouts, D. E., Eisen, J. A., Gill, S. R., Heidelberg, J. F.*, et al.* (2003). Role of mobile DNA in the evolution of vancomycin-resistant Enterococcus faecalis. Science *299*, 2071-2074.

Perna, N. T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., Kirkpatrick, H. A.*, et al.* (2001). Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature *409*, 529-533.

Prasad, I., and Schaefler, S. (1974). Regulation of the beta-glucoside system in Escherchia coli K-12. J Bacteriol *120*, 638-650.

Prasad, I., Young, B., and Schaefler, S. (1973). Genetic determination of the constitutive biosynthesis of phospho- -glucosidase A in Escherichia coli K-12. J Bacteriol *114*, 909-915.

Pupo, G. M., Karaolis, D. K., Lan, R., and Reeves, P. R. (1997). Evolutionary relationships among pathogenic and nonpathogenic Escherichia coli strains inferred from multilocus enzyme electrophoresis and mdh sequence studies. Infect Immun *65*, 2685-2692.

Raghunand, T. R., and Mahadevan, S. (2003). The beta-glucoside genes of Klebsiella aerogenes: conservation and divergence in relation to the cryptic bgl genes of Escherichia coli. FEMS Microbiol Lett *223*, 267-274.

Read, T. D., Salzberg, S. L., Pop, M., Shumway, M., Umayam, L., Jiang, L., Holtzapple, E., Busch, J. D., Smith, K. L., Schupp, J. M.*, et al.* (2002). Comparative genome sequencing for discovery of novel polymorphisms in Bacillus anthracis. Science *296*, 2028-2033.

Rey, M. W., Ramaiya, P., Nelson, B. A., Brody-Karpin, S. D., Zaretsky, E. J., Tang, M., Lopez de Leon, A., Xiang, H., Gusti, V., Clausen, I. G.*, et al.* (2004). Complete genome sequence of the industrial bacterium Bacillus licheniformis and comparisons with closely related Bacillus species. Genome Biol *5*, R77.

Reynolds, A. E., Felton, J., and Wright, A. (1981). Insertion of DNA activates the cryptic bgl operon in E. coli K12. Nature *293*, 625-629.

Reynolds, A. E., Mahadevan, S., LeGrice, S. F., and Wright, A. (1986). Enhancement of bacterial gene expression by insertion elements or by mutation in a CAP-cAMP binding site. J Mol Biol *191*, 85-95.

Rode, C. K., Melkerson-Watson, L. J., Johnson, A. T., and Bloch, C. A. (1999). Type-specific contributions to chromosome size differences in Escherichia coli. Infect Immun *67*, 230-236.

Rosey, E. L., and Stewart, G. C. (1992). Nucleotide and deduced amino acid sequences of the lacR, lacABCD, and lacFE genes encoding the repressor, tagatose 6-phosphate gene cluster, and sugar-specific phosphotransferase system components of the lactose operon of Streptococcus mutans. J Bacteriol *174*, 6159-6170.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular cloning: a laboratory manual.(Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).

Sambrook, J., Fritsch, E.F., and Maniatis, T. (2001). Molecular cloning: a laboratory manual. (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).

Schaefler, S., and Maas, W. K. (1967). Inducible system for the utilization of beta-glucosides in Escherichia coli. II. Description of mutant types and genetic analysis. J Bacteriol *93*, 264-272.

Schaefler, S., and Malamy, A. (1969). Taxonomic investigations on expressed and cryptic phospho-beta-glucosidases in Enterobacteriaceae. J Bacteriol *99*, 422-433.

Schaefler, S., Malamy, A., and Green, I. (1969). Phospho-beta-glucosidases and beta-glucoside permeases in Streptococcus, Bacillus, and Staphylococcus. J Bacteriol *99*, 434-440.

Schneiker, S., Kosier, B., Puhler, A., and Selbitschka, W. (1999). The Sinorhizobium meliloti insertion sequence (IS) element ISRm14 is related to a previously unrecognized IS element located adjacent to the Escherichia coli locus of enterocyte effacement (LEE) pathogenicity island. Curr Microbiol *39*, 274-281.

Schnetz, K. (1995). Silencing of Escherichia coli bgl promoter by flanking sequence elements. Embo J *14*, 2545-2550.

Schnetz, K. (2002). Silencing of the Escherichia coli bgl operon by RpoS requires Crl. Microbiology *148*, 2573-2578.

Schnetz, K., and Rak, B. (1988). Regulation of the bgl operon of Escherichia coli by transcriptional antitermination. Embo J *7*, 3271-3277.

Schnetz, K., and Rak, B. (1992). IS5: a mobile enhancer of transcription in Escherichia coli. Proc Natl Acad Sci U S A *89*, 1244-1248.

Schnetz, K., Stulke, J., Gertz, S., Kruger, S., Krieg, M., Hecker, M., and Rak, B. (1996). LicT, a Bacillus subtilis transcriptional antiterminator protein of the BglG family. J Bacteriol *178*, 1971-1979.

Schnetz, K., Toloczyki, C., and Rak, B. (1987). Beta-glucoside (bgl) operon of Escherichia coli K-12: nucleotide sequence, genetic organization, and possible evolutionary relationship to regulatory components of two Bacillus subtilis genes. J Bacteriol *169*, 2579-2590.

Selander, R. K., Caugant, D. A., Ochman, H., Musser, J. M., Gilmour, M. N., and Whittam, T. S. (1986). Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. Appl Environ Microbiol *51*, 873-884.

Shah, S., and Peterkofsky, A. (1991). Characterization and generation of Escherichia coli adenylate cyclase deletion mutants. J Bacteriol *173*, 3238-3242.

Steed, P. M., and Wanner, B. L. (1993). Use of the rep technique for allele replacement to construct mutants with deletions of the pstSCAB-phoU operon: evidence of a new role for the PhoU protein in the phosphate regulon. J Bacteriol *175*, 6797-6809.

Stulke, J., Arnaud, M., Rapoport, G., and Martin-Verstraete, I. (1998). PRD--a protein domain involved in PTS-dependent induction and carbon catabolite repression of catabolic operons in bacteria. Mol Microbiol *28*, 865-874.

Takami, H., Nakasone, K., Takaki, Y., Maeno, G., Sasaki, R., Masui, N., Fuji, F., Hirama, C., Nakamura, Y., Ogasawara, N.*, et al.* (2000). Complete genome sequence of the alkaliphilic bacterium Bacillus halodurans and genomic sequence comparison with Bacillus subtilis. Nucleic Acids Res *28*, 4317-4331.

Tettelin, H., Nelson, K. E., Paulsen, I. T., Eisen, J. A., Read, T. D., Peterson, S., Heidelberg, J., DeBoy, R. T., Haft, D. H., Dodson, R. J.*, et al.* (2001). Complete genome sequence of a virulent isolate of Streptococcus pneumoniae. Science *293*, 498-506.

Tobisch, S., Glaser, P., Kruger, S., and Hecker, M. (1997). Identification and characterization of a new beta-glucoside utilization system in Bacillus subtilis. J Bacteriol *179*, 496-506.

Tsui, H. C., Leung, H. C., and Winkler, M. E. (1994). Characterization of broadly pleiotropic phenotypes caused by an hfq insertion mutation in Escherichia coli K-12. Mol Microbiol *13*, 35-49.

Turner, A. K., Lovell, M. A., Hulme, S. D., Zhang-Barber, L., and Barrow, P. A. (1998). Identification of Salmonella typhimurium genes required for colonization of the chicken alimentary tract and for virulence in newly hatched chicks. Infect Immun *66*, 2099-2106.

Ueguchi, C., Ohta, T., Seto, C., Suzuki, T., and Mizuno, T. (1998). The leuO gene product has a latent ability to relieve bgl silencing in Escherichia coli. J Bacteriol *180*, 190-193.

Ussery, D. W., Hinton, J. C., Jordi, B. J., Granum, P. E., Seirafi, A., Stephen, R. J., Tupper, A. E., Berridge, G., Sidebotham, J. M., and Higgins, C. F. (1994). The chromatin-associated protein H-NS. Biochimie *76*, 968-980.

Webb, C., Moreno, M., Wilmes-Riesenberg, M., Curtiss, R., 3rd, and Foster, J. W. (1999). Effects of DksA and ClpP protease on sigma S production and virulence in Salmonella typhimurium. Mol Microbiol *34*, 112-123.

Wei, J., Goldberg, M. B., Burland, V., Venkatesan, M. M., Deng, W., Fournier, G., Mayhew, G. F., Plunkett, G., 3rd, Rose, D. J., Darling, A.*, et al.* (2003). Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T. Infect Immun *71*, 2775-2786.

Welch, R. A., Burland, V., Plunkett, G., 3rd, Redford, P., Roesch, P., Rasko, D., Buckles, E. L., Liou, S. R., Boutin, A., Hackett, J.*, et al.* (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci U S A *99*, 17020-17024.

Werts, C., Charbit, A., Bachellier, S., and Hofnung, M. (1992). DNA sequence analysis of the lamB gene from Klebsiella pneumoniae: implications for the topology and the pore functions in maltoporin. Mol Gen Genet *233*, 372-378.

Whittam, T. S., Ochman, H., and Selander, R. K. (1983). Multilocus genetic structure in natural populations of Escherichia coli. Proc Natl Acad Sci U S A *80*, 1751-1755.

Zhang, J., and Aronson, A. (1994). A Bacillus subtilis bglA gene encoding phospho-beta-glucosidase is inducible and closely linked to a NADH dehydrogenase-encoding gene. Gene *140*, 85-90.

Erklärung

Ich versichere, daß ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit- einschließlich Tabellen, Karten und Abbildungen-, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; daß diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; daß sie - abgesehen von unten angegebenen Teilpublikationen- noch nicht veröffentlicht worden ist, sowie, daß ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Frau Prof. Dr. Karin Schnetz betreut worden.

Teilpublikationen:

Girish Neelakanta and Karin Schnetz. Genome variations at the *bgl*/Z5211-Z5214 genomic island region in commensal and pathogenic *E.coli*. (Manuscript in preparation)

Girish Neelakanta and Karin Schnetz. Identification and analysis of an additional β-glucoside system in *E.coli.* (Manuscript in preparation)

Datum: 10/12/2004 Unterschrift

Girish Neelakanta

Curriculum vitae

Name **Girish Neelakanta**

Date of Birth 01.09.1976

Place of Birth Bangalore, India.

Nationality Indian

Contact details Institute for Genetics,
University of Cologne,
Weyertal 121,
50931, Cologne,
Germany
Email: girish.neelakanta@uni-koeln.de
Telephone: +49-0221-470 7887

Academic profile

University studies:

1994-1997 Bachelor of Science (BSc) in Microbiology,
Bangalore University, Bangalore, India

1997-1999 Master of Science (MSc) in Biotechnology,
Bangalore University, Bangalore, India

Dissertation title: Purification and characterization of DNA from rat brain.
Advisor: Prof. Dr. N.B. Joshi,
Head of the Department, Department of Biophysics,
National Institute of Mental Health and Neurosciences
Bangalore, India

Research Assistantship

1999-2001 Research assistant,
Department of Molecular Reproduction, Development and Genetics, Indian Institute of Science (IISc), Bangalore, India
Project advisor: Prof. Dr. S. Mahadevan

Doctoral Study

2001-2005 Thesis title: Genome variations in commensal and pathogenic *E.coli*
Advisor: Prof. Dr. Karin Schnetz,
Institute for Genetics,
University of Cologne, Weyertal 121,
50931, Cologne

Date: 10/12/2004
Place: Cologne

Girish Neelakanta

Lebenslauf

Name	**Girish Neelakanta**
Geburtsdatum	01.09.1976
Geburtsort	Bangalore, Indien
Staatsangehörigkeit	Indisch
Anschrift	Institut fur Genetik, Weyertal 121, 50931, Koln, Deutschland Email: girish.neelakanta@uni-koeln.de Telefon: +49-0221-470 7887

Voruniversität

Studium

1994-1997	Studium der Mikrobiologie (*Bachelor of Sciences*, *B.Sc*), Bangalore University, Bangalore, Indien
1997-1999	Diplomstudium der Biotechnologie (*Master of Sciences,* Biotechnology*),* Bangalore university, Bangalore, Indien Titel der Diplomarbeit: Purification and characterization of DNA from rat brain

Promotionsstudium

Juni 2001-Februar 2005	Institut fur Genetik, Weyertal 121, 50931, Koln, Deutschland Thema: 'Genome variations in commensal and pathogenic *E.coli'* Betreuerin: Frau Prof. Dr. Karin Schnetz,

Ort: Koeln
Datum: 10/12/2004

Girish Neelakanta