The evolution of silent β-glucoside systems in *Escherichia coli*

Inaugural-Dissertation

zur

Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Sabari Sankar Thirupathy
aus Tamilnadu, Indien

Köln, Dez 2007

Berichterstatter/in: Prof. Dr. Karin Schnetz
PD. Dr. Röbbe Wünschiers

Tag der mündlichen Prüfung: 11 Feb 2008

Acknowledgements

First and foremost, I wholeheartedly thank Karin, for being a wonderful advisor, for her guidance, suggestions, criticisms, imparting organizational capabilities in me, showing me the computer-tricks and so on. I am indebted to her for all her encouragement and support.

I thank the Schnetz Group, it has been very nice to work in a friendly atmosphere. My special thanks to Vel, Madhu, Girish, Andreas, Kathleen, Frant and Tinka. Coffee time with Raja was quite refreshing. Thanks to Maria for all the help with translation.

I am grateful to Mark, for giving me the opportunity to carry out the MLST work in his lab and for all his scientific discussions, suggestions and criticisms. He has been a great source of inspiration.

I express my sincere thankfulness to Thomas Wiehe, for helping me with the computational analysis of Z locus, discussions with him has always been very fruitful.

I thank Röbbe very much, for reading my thesis, being my referee and offering helpful suggestions at the beginning of my thesis work.

A special mention to Prof. RJ and Prof. Munavar, whose thought provoking lectures drove my fascination towards *E. coli* genetics.

I thank Dr. Ludwig Eichinger, who helped me to spot my arrays.

I am pleased with Vartul and Chris for extending their genuine support in performing MLST and microarray.

My special gratitude to Brigitte, who was there always to favor me, from housing to administration. I also thank Inge for her kind administrative help. I thank the Graduate School for the fellowship.

All my friends have been very supportive, Ras, Bala, Senthil, Palani and Palani. Special thanks to Sam, for proofreading my thesis. Mensa time with Jayan has been very nice to discuss anything under the sun.

I am thankful forever to Amma, Naina, Andal, Mari and Priya. Neil and Daya are little inspirations.

Kayal, thank you!

Contents

I Zusammenfassung **1**

I Summary **3**

II. Introduction **5**

1. Diversity of bacterial genomes 6
2. Bacterial genome evolution: the three facets 7
3. *Escherichia coli*: Phylogeny and population structure 9
4. Cryptic genes 12
5. Objectives of the current study 14

III. Results **15**

1. Evolutionary genetic analysis of the bgl operon and Z locus in the E. coli population *15*

1.1 Population structure of the *E. coli* collection 15
1.2 Genetic diversity of the *bgl* operon/ Z5211-5214 locus in *E. coli* natural isolates 20
1.3 Genetic variation in the core genome flanking the *bgl* operon/Z5211-5214 locus 23
1.4 The evolution of the *bgl*/Z5211-5214 locus is coupled to species evolution 26
1.5 Clonal evolution of the *bgl* operon/Z5211-5214 locus 28
1.6 Phylogenetic analysis of the complete *bgl* operon sequence 29
1.7 Functional analysis of *bgl* operon 32
1.8 Insertions/deletions in *bgl* /Z5211-5214 loci 34
1.9 The *bgl* operon is vertically inherited in Enterobactericiae 34
1.10 The *bgl* operon does not affect fitness in LB medium 38
1.11 Sequence evolution of the *bgl* locus 40

2. Population genetic analysis of the E. coli bgc locus *42*

2.1 The second cryptic β-glucoside operon, *bgc* 42
2.2 Typing of the *bgc* operon in the *E. coli* 42
2.3 Correlation of the prevalence of the *bgc* operon with the *E. coli* phylogeny 45

3. Comparative genomics of E. coli *47*

3.1 Evolution of bacterial genomes 47
3.2 *E. coli* oligonucleotide microarray 49
3.3 Microarray fabrication 50
3.4 Probe preparation and Hybridization 51
3.5 Data acquisition and analysis 53
3.6 Trial experiments with *E. coli* K12 MG1655 strain 55

IV Discussion **57**

1. Origin and evolution of the *bgl* operon 58
2. Phylogeny and clonality of *bgl* 59
3. String of β-glucoside systems 60
4. Function and Selection 61
5. Conclusions 62

V Materials and Methods **64**

1. Chemical, enzymes and other materials 64
2. Media and agar plates 64
3. Antibiotics 65
4. General methods 65
5. *E. coli* and other strains 65
6. Multilocus sequence typing (MLST) 67
7. Typing of *bgl* operon/Z5211-5214 locus and *bgc* operon 68
8. DNA sequencing 68
9. Phylogenetic analysis 68
10. BLAST survey 68
11. Gene deletion according to Datsenko and Wanner, 2000 69
12. Electrocompetant cells and electroporation 69
13. Transduction with phage T4GT7 70
14. Microarray-CGH protocols 71

VI Appendix **76**

VII References **85**

Erklärung **91**

Lebenslauf **92**

Curriculum vitae **93**

I Zusammenfassung

Der Aufbau bakterieller Genome ist sehr dynamisch. Beispielsweise ist bei *Escherichia coli* nur etwa 60 bis 70% des Genoms allen individuellen Isolaten gemeinsam. Der Rest des Genoms besteht aus einem flexiblen Pool von Genen, die nur in einigen Isolaten vorkommen. Diese Genom-Diversität basiert auf Genaufnahme durch horizontalen Transfer sowie auf Genverlust und der Mutation von Genen. In dieser Arbeit wurde die Evolution zweier kryptischer β-Glukosid Loci, die Teil des flexiblen Genpools sind, analysiert. Das *bgl*-Operon ist in etwa 80% aller *E. coli* Isolate vorhanden, während der Rest der Isolate stattdessen einen Locus (Z5211-5214) mit Genen unbekannter Funktion trägt. Der *bgc* Locus ist in etwa 50% aller *E. coli* Isolate vorhanden. Zur Analyse der Evolution dieser Loci, wurde die Phylogenie einer repräsentativen Kollektion von 175 *E. coli* Isolaten mittels der Methode des „multilocus sequence typing (MLST)" etabliert. Parallel dazu wurden die *bgl* und *bgc* Loci in diesen Stämmen per PCR und Sequenzierung typisiert. Dies zeigte, dass beim *bgl* / Z Locus vier Gruppen unterschieden werden können, wobei drei davon *bgl* Varianten sind. Die Korrelation dieser *bgl / Z* Gruppen mit der Spezies-Phylogenie zeigte eine erstaunliche Deckung: unter den 4 phylogenetischen Gruppen von *E. coli,* ist *bgl* in den Gruppen A, B1, and B2 vorhanden, während der Z Locus (fast) ausschließlich in D Stämmen vorkommt. Diese strikte Korrelation belegt, dass die Evolution des *bgl* / Z Locus an die Evolution der Spezies *E. coli* gekoppelt sein muss. Weiterhin zeigte diese phylogenetische Analyse, dass die *bgl* und Z Loci nicht in drei *E. coli* Isolaten vorhanden sind, die vermutlich Reste einer früheren *E.coli* Population darstellen. Auch fehlt der *bgl* / Z locus in der nah verwandten Art *Escherichia albertii* und in der Gattung *Salmonella*. Dies deutet auf einen horizontalen Transfer der *bgl* und Z Loci in Vorläufer der modernen *E.coli* Population hin. Im Widerspruch dazu zeigte eine BLAST-Suche, dass *bgl* Homologe in *Erwinia*, *Klebsiella* und *Photorhabdus* sp. vorhanden sind, wobei deren Phylogenie mit der 16S rRNA Phylogenie der Spezies übereinstimmt. Dies ist ein guter Beleg für eine vertikale Vererbung des *bgl* Locus, wobei *bgl* vermutlich in einigen enterobakteriellen Linien verloren ging. In den *E. coli*

Isolaten der phylogenetischen Gruppe D, wurde das *bgl*-Operon vermutlich mit Entstehung dieser Subgruppe durch den Z-Locus ersetzt, möglicherweise durch horizontalen gentransfer. Weiterhin zeigte die Korrelation einer funktionalen Analyse des *bgl*-Operon mit der *E.coli* Phylogenie, dass das *bgl*-Operon in den meisten Stämmen der Gruppen A und B1 intakt aber stillgelegt ist. Interessanterweise, wird das *bgl*-Operon in ~50% der Stämme der Gruppe B2 schwach exprimiert. Diese Daten kombiniert mit dem Ergebnis eines "nonsynonymous-to-synonymous substitution ratio test" (K_A/K_S Test) spricht dafür, dass das *bgl*-Operon einen unbekannten ökoligischen Selektionsvorteil bewirkt. Das zweite stumme β-Glukosid-System, *bgc*, kommt hauptsächlich in Isolaten der phylogenetischen Gruppen B1 and B2 vor. Es ist in Isolaten der Gruppe D zum Teil vorhanden und nur in einigen A-Isolaten nachweisbar. Dieses Verteilungsmuster des *bgc*-Locus kann sowohl durch Genaufnahme als auch durch Genverlust erklärt werden. Zusammengefasst zeigt die vorliegende Arbeit, dass die Flexibilität des Genoms zusätzlich zur Aufnahme von Genen auch wesentlich durch Genverlust bestimmt wird, und dass eine sorgfältige Analyse einzelner Loci notwendig ist, um zwischen diesen beiden Mechanismen unterscheiden zu können.

I Summary

The genomes of bacterial species are very dynamic. For example in *Escherichia coli,* individual isolates may share as little as 60 to 70% of their genome with other isolates. The remainder of the genome consists of a flexible pool of genes, which are present only in some isolates. This genome diversity is manifested through gain of genes by horizontal transfer as well as by loss or mutations of genes. In this study the evolution of two silent β-glucoside loci belonging to the flexible gene pool of *E. coli* was traced. The *bgl* operon is present in ~80% of all *E. coli* isolates, while in the rest it is replaced by a locus (named Z5211-5214) of unknown function. The *bgc* locus is present in roughly 50% of *E. coli* isolates. To trace the evolution of these loci, the phylogeny of a representative collection of 175 *E. coli* isolates was established by multilocus sequence typing (MLST). In parallel, the *bgl* and *bgc* loci were typed by PCR and sequencing. This revealed four groups of the *bgl* / Z locus, including 3 *bgl* variants. Mapping of these groups demonstrated a striking correlation to the species phylogeny and population structure: among the four phylogenetic groups of *E. coli, bgl* is present in the A, B1, and B2 groups, while the Z locus is present in D strains, which suggests a coupled evolution of the *bgl* / Z locus with the host. Further, three ancestral *E. coli* isolates and strains of the closely related species *Escherichia albertii* as well as the closely related genus *Salmonella enterica*, lack the *bgl* / Z locus, indicating horizontal transfer of the *bgl* and Z loci into the root of the modern *E. coli*. However, BLAST surveys revealed the presence of *bgl* homologs in *Erwinia*, *Klebsiella* and *Photorhabdus* species. The phylogeny of *E. coli bgl* and these homologs is concordant with the 16S rRNA phylogeny contradicting horizontal transfer. In conclusion, these results implicate vertical inheritance of *bgl* and its loss in some enterobacterial lineages. In *E. coli* isolates belonging to the phylogenetic group D, the *bgl* operon presumably was replaced by the Z locus, which may have been horizontally acquired. Further, correlating the data of a functional analysis of *bgl* with the species phylogeny demonstrated that *bgl* is functional although silent in the majority of strains in groups A and B1, while, interestingly, in more than 50% of B2 strains, *bgl* was not silent but

weakly expressed. These data together with the results of nonsynonymous-to-synonymous substitution ratio test (the K_A/K_S test), suggest that *bgl* may confer an unknown ecological advantage. The second silent β-glucoside system *bgc* analyzed here, is predominant in the phylogenetic groups B1 and B2, it is present in D group isolates and rarely found in A strains. The widespread occurrence could be due to either gain or loss of *bgc* in evolution. Cumulatively, the study suggests that in addition to gene gain, also gene loss may significantly contribute to the flexibility of the genome, and that a careful analysis is required for individual loci belonging to the flexible gene pool.

II. Introduction

Bacterial evolution is very dynamic. Bacterial genomes are mosaic in nature consisting of a core pool of genes, which are shared by all individuals of a species, and a flexible pool of genes, which are present only in a subset of individuals of the species. This duality of the genome allows maintaining essential function and provides the flexibility to explore new niches (Feil, 2004). Evolution of bacterial genomes is brought about by three major mechanisms; the gain of genes through horizontal transfer, gene loss, and the modification of existing genes (Lawrence, 2005). The amount of genetic diversity seen within a species is remarkable. For example, the species *Escherichia coli* includes commensals and diverse pathogens. Their mosaic genomes can vary in size by up to one megabase (Bergthorsson and Ochman, 1998), and the core genes make up only 60 to 70% of individual genomes (Welch et al., 2002). The diversity of *E. coli* is due to genome rearrangements that occurred on a microevolutionary scale, as suggested by comparative genomic studies (Fukiya et al., 2004; Perna et al., 2001; Wei et al., 2003). Among the three major mechanisms, gene modification, gene loss, and gene gain, which work behind the observed diversity of bacteria, the latter has been extensively studied for more than a decade. Horizontal transfer of genes is considered a major force in shaping bacterial genome evolution (Gogarten and Townsend, 2005; Lawrence and Hendrickson, 2003). Gene loss is evident in the evolution of obligate parasites and symbionts (Mira et al., 2001). However, the relative role of gene gain and loss in the evolution of a species is not well known. For *E. coli*, this lack of knowledge is mainly due to the focus of research on horizontally acquired pathogenicity islands (Groisman and Ochman, 1996; Hacker and Carniel, 2001; Hacker and Kaper, 2000). In contrast, the focus of the current study was on understanding the mechanisms of genome evolution by tracing the evolutionary history of cryptic genes in the population of *E. coli*.

1. Diversity of bacterial genomes

The textbook definition of 'species' is that individuals differ from others by minor but identifiable differences. However, in bacterial species the genomes display such a wide range of diversity that the definition of the bacterial species was questioned (Gevers et al., 2005). The genomes of individual isolates of bacterial 'species' can differ up to 50% in the case of *Streptococcus* (Marri et al., 2006), and 60-70% in *E. coli* as revealed by genome sequence comparison (Welch et al., 2002). Moreover, microarray based comparative genomic hybridization studies on 23 natural isolates of *E. coli* showed that ~3000 genes belong to the genomic core and ~1000-1500 genes are variable (Dobrindt et al., 2003b). However, this diversity of the bacterial genome is based on the flexible gene content rather than on sequence variation throughout the genome. The sequence of the core genome, which is the part present in all strains of a given species, is highly conserved. Further, the comparison of core genome genes between close and more distantly related 'species' can be used to build robust phylogenetic trees, which reflect the evolution of the bacterial lineages. The core genome of each species differs significantly from the core genome of closely related species, and these differences reflect the phylogeny of the species. Furthermore, the analysis of core genome genes revealed that about 200 genes are common to the gamma-proteobacteria. Only 60 genes are shared by all cellular organisms; these genes are mainly important for translation (Koonin, 2003).

In contrast to the core genome, which is assumed to encode the essential functions for the species, the flexible gene pool is considered to confer a selective advantage under specific conditions. Genes that belong to the flexible gene pool include virulence factors, antibiotic resistance genes, genes for symbiosis among others. These genes are often part of genomic and pathogenicity islands, which are horizontally transferred into the genome (Hacker and Carniel, 2001). Considering the variability of the bacterial genome, recently, the term "pan-genome" was introduced in bacterial genomics to accomplish a broader definition of bacterial species. The "pan-genome" includes the core genes as well as all the genes of the flexible pool

found in different strains of one species (Medini et al., 2005). The size of the pan genome of a given bacterial species is anticipated to increase with the availability of genome sequences of individual strains. The diversity of bacterial genomes makes it an attractive case for the analysis of bacterial evolution.

2. Bacterial genome evolution: the three facets

As mentioned before, the principal driving forces that shape bacterial genomes are i) the modification of vertically transmitted genes, ii) gene loss, and iii) gene gain (Fig. 1). In eukaryotes, evolution occurs by the modification of existing genes (McDonald and Kreitman, 1991), whereas in prokaryotes there are countable instances showing such gene modifications. One such case is the increase of pathogenicity in *Salmonella* by the alteration of the *pmrD* gene encoding polymyxin B resistance to become regulated by the PhoPQ two-component regulatory system. Another example is that in *Bordetella* the expression of a toxin gene *ptxA* is enhanced by mutations (Parkhill et al., 2003a; Winfield and Groisman, 2004). Gene gain by horizontal gene-transfer (HGT) is considered a hallmark of bacterial evolution, especially of pathogens (Hacker and Kaper, 2000; Ochman et al., 2000). Horizontal gene transfer is mediated by three mechanisms: transformation (of plasmids or naked DNA), transduction (of genomic and pathogenicity islands), and conjugation. Horizontally transferred genes generally have different GC content and Codon usage when compared to the host genomes (Ochman et al., 2000). Examination of genomes based on DNA composition of commensals and pathogens for detecting foreign genes showed abundant signs of recent acquisitions ranging from 0% in *Mycoplasma genitalium* to 17% in *Synechocystis*. HGT has been well studied in *E. coli*. Based on GC content and Codon usage analyses the *E. coli* K12 MG1655 strain is estimated to contain about 18% of foreign DNA with a transfer rate of 16 kb/Myr since speciation (Lawrence and Ochman, 1998; Ochman et al., 2000). Even in the presence of extensive transfer of DNA in *E. coli*, the chromosome size of different strains remains relatively constant around 5 megabases and in general prokaryotic genome sizes tend to remain constant. Thus, obviously gene acquisition must be balanced by loss of genes in order to reflect the

observed constant size of the genomes. Gene loss has been estimated to occur at a rate of two-three times higher as horizontal gene-transfer, when ~12,000 protein families were analyzed, but this balance need not necessarily operate on individual species (Kunin and Ouzounis, 2003). Gene loss has been the hallmark of evolution of pathogens and symbionts. Massive genome reduction is noticed in *Mycobacterium leprae*, *Buchnera*, and *Rickettsia* (Cole et al., 2001; Parkhill et al., 2003b; van Ham et al., 2003).

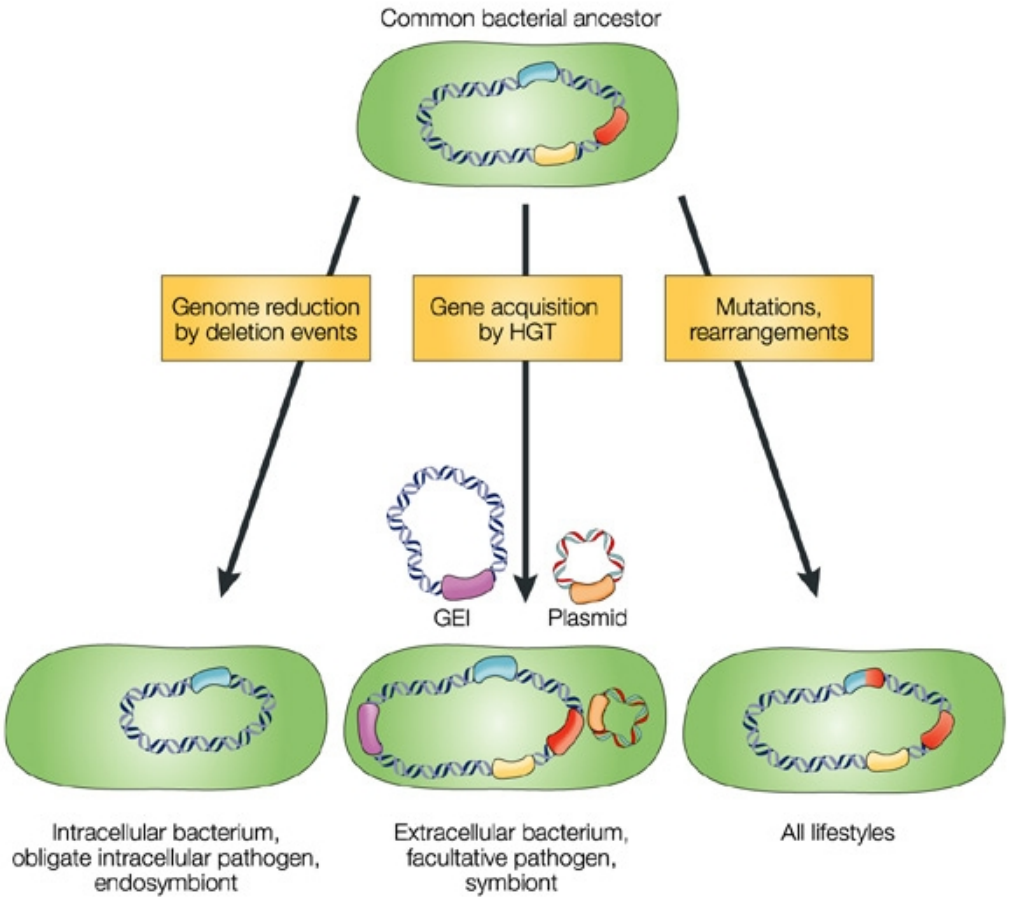

Fig. 1: Mechanisms that drive bacterial genome evolution. Three mechanisms are considered responsible for evolution of genome structures, which may reflect different bacterial lifestyles. These are, firstly the modifications of existing genes by mutations and rearrangements, which suit all lifestyles. 2. Gene loss is the major force in genome reduction by deletion events, seen in host-dependent bacteria. 3. Gene gain by horizontal gene transfer (HGT) increases the adaptability of commensal and pathogenic bacteria by introducing genomic islands (GEI), pathogenic and symbiotic islands. Figure adopted and modified from (Dobrindt, 2004).

Further, gene loss occurs on special functions that may be detrimental to pathogenic lifestyle, creating “black holes” (deletions) in the genome. For example in *Shigella,* Cadaverine produced by the decarboxylation of lysine inhibits *Shigella* enterotoxic activity, and deletion of *cadA* encoding the lysine decarboxylase was demonstrated to enhance virulence (Maurelli et al., 1998). *Shigella* are bacteria that belong to the species *E. coli*, which can be distinguished from other *E. coli* strains by specific markers (as Shiga-toxin expression) and whose name is maintained for medical historical reasons (Maurelli, 2007). Therefore, evolution of Shigella from *E. coli* is marked by gene gain of virulence traits and loss of biochemical functions that are adaptive to pathogenic lifestyle.

Detection of horizontally acquired genes is primarily performed by compositional analysis of sequences and by BLAST searches on related genomes. But it has been pointed out that gene loss can be interpreted as gene gain if one relies on BLAST searches for orthologs due to limitation of availability of genome sequences (Zhaxybayeva et al., 2007). Moreover, the quantification of gene gain and loss events has been performed on protein families rather than on individual genes. One example of a study on individual genes is the *lac* operon of *E. coli*. The *lac* operon was thought to be horizontally acquired in *E. coli* and therefore *E. coli* metabolizes lactose (Ochman et al., 2000). In contrast to this opinion, D.M. Stoebel (2006) showed that the *lac* operon is vertically transmitted in enterobactericiae and that the Lactose negative phenotype of some members including *Salmonella*, *Shigella* are due to loss of the operon. Hence, for a detailed study of loss and gain of genes in bacteria, rigorous phylogenetic methods need to be performed at the population level for individual genes.

3. *Escherichia coli*: Phylogeny and population structure

The versatile *E. coli* represents an excellent model to understand bacterial genome evolution owing to its well-established phylogenetic groups and population structure. Classic multilocus enzyme electrophoresis (MLEE) typing of 72 reference strains of *E. coli*, the ECOR collection, (Ochman and Selander, 1984) indicated the existence of four phylogenetic groups of *E. coli*,

which are designated A, B1, B2 and D. A minor group E has been neglected later, because of inconsistent clustering in subsequent analyses. Within this decade, the new molecular technique, multilocus sequence typing (MLST) was introduced for bacterial strain typing. MLST is conceptually similar to MLEE (multilocus enzyme electrophoresis) but characterizes each strain of a bacterial species by assigning alleles for seven housekeeping genes directly from the nucleotide sequence of internal fragments of genes, rather than indirectly from the electrophoretic mobilities of their gene products (Maiden, 2006). The genotype of strains characterized by MLST is defined by their allelic profiles. MLST has several advantages over MLEE. MLST is highly discriminative as it detects all the nucleotide polymorphisms within a gene rather than just those mutations that alter the electrophoretic mobility of the protein product. Bacterial strains harbor sufficient variation within the housekeeping loci that many different alleles can be resolved and by using seven genes, billions of genotypes can be obtained. A second advantage of MLST is the accuracy and portability of DNA sequence data, which can be rapidly and unambiguously compared with previously characterized strains by interrogation through a common web server (http://www.mlst.net/). MLST therefore provides a precise and unambiguous method for characterizing bacterial strains.

For *E. coli* three MLST schemes (Le et al., 2007; Reid et al., 2000) are established, one of them was designed by Wirth et al., (2006) who put forth a broader picture on the evolutionary history of *E. coli*. Briefly, MLST was used to assess the genetic relatedness of 406 natural isolates of *E. coli*, by analyzing the allelic profile of seven housekeeping genes distributed around the chromosome (Fig. 2a). Fragments of these seven genes are PCR amplified and sequenced on both strands using the PCR primers. Sequences are manually curated and each unique sequence of a gene is assigned an allele number. Thus, seven allele numbers are obtained for a strain at seven housekeeping genes. Combination of the seven numbers for a strain constitutes its allelic profile or Sequence Type (ST). In their study, Wirth et al., (2006) presented a star-like phylogeny depicting the rapid population expansion that resulted in the diversity of *E. coli* species (Fig. 2b). The four

groups A, B1, B2 and D comprise the modern *E. coli* strains and two divergent isolates, which are *E. coli*, are considered as remnants of ancestral diversity. Moreover, they identified 278 sequence types (STs) at that time and currently 721 STs (as of 10.12.2007) are deposited at the web-based server for *E. coli* MLST (http://web.mpiib-berlin.mpg.de/mlst/dbs/Ecoli). Further, they suggested

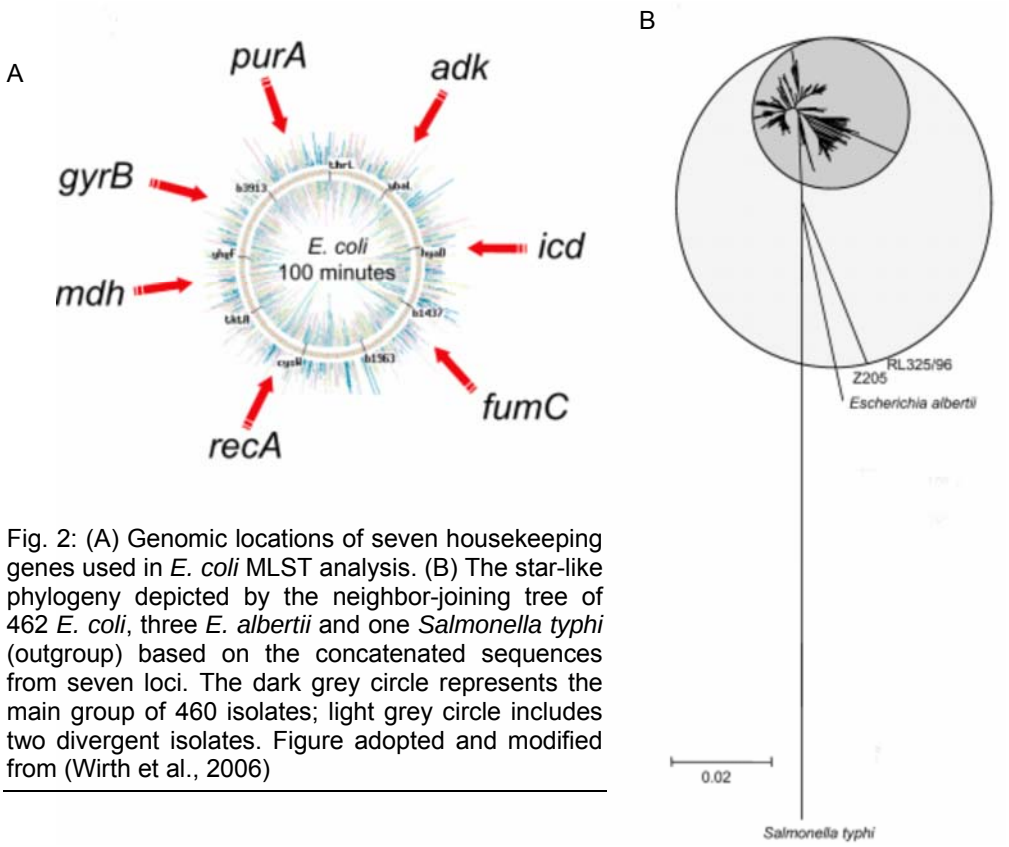

Fig. 2: (A) Genomic locations of seven housekeeping genes used in *E. coli* MLST analysis. (B) The star-like phylogeny depicted by the neighbor-joining tree of 462 *E. coli*, three *E. albertii* and one *Salmonella typhi* (outgroup) based on the concatenated sequences from seven loci. The dark grey circle represents the main group of 460 isolates; light grey circle includes two divergent isolates. Figure adopted and modified from (Wirth et al., 2006)

that the clonal structure established by MLST would provide a better framework for studying the evolution of strains, compared to the use of classic phylogenetic groups whose boundaries are fluid. Recently, Weissman et al. (2006) used MLST analysis on *E. coli* pathogens to deduce a clonal framework. In that work they identified their strains to belong to a sequence type complex ST95 and studied the evolution of fimbrial genes at the clonal complex level. From their work, Weissman et al., (2006) showed the horizontal transfer of fimbrial operons into ST95 complex strains and divergence of the genes after entry into the complex. This suggests that clonal level analysis gives a finer evolutionary framework, which when combined

with phylogenetics, can be promising to trace the history of individual genes at a deeper resolution.

4. Cryptic genes

Genes that are not expressed under any tested condition are considered cryptic or silent. Silent genes are found in bacteria in many species including *Lactobacillus*, *Bacillus*, *Escherichia*, and *Salmonella* (Birge EA, 2006). The well-studied examples exist in *E. coli*. These are the *bgl* and *asc* operons, which are involved in β-glucosides metabolism (Fig. 3). Among these two, the *bgl* operon is well characterized, known to encode proteins for uptake and hydrolysis of aryl-β-D-glucosides such as salicin, arbutin (Schnetz et al., 1987). The *asc* (arbutin, salicin, cellobiose) locus of *E. coli* encodes a regulator, a permease and a β-glucosidase necessary for transport and hydrolysis of the β-glucosides. Previously another silent β-glucoside system called *bgc* (Fig. 3) was discovered in the laboratory (Neelakanta, 2005). *bgc* (β-glucoside and cellobiose) locus comprises an operon and a divergent regulator gene, needed for utilization of β-glucosides and cellobiose at low temperature. The *bgl* operon of *E. coli* is a paradigm of crypticity, as it is not expressed under any laboratory-tested conditions. Silencing of *bgl* operon is mediated by the histone-like nucleoid structuring protein (H-NS) (Dole et al., 2004a; Nagarajavel et al., 2007). Intuitively, silent genes should be undesirable, as selection will not favor their function, ultimately leading to their erosion. On the contrary, the *bgl* operon is present in the *E. coli* laboratory strain and surprisingly, in the uropathogenic CFT073 strain as well (Welch et al., 2002). Previous work in the laboratory by G. Neelakanta (2005) showed that the *bgl* operon is predominant in natural isolates of *E. coli* including commensals and pathogens (Neelakanta, 2005). In addition, it was found that in a subset of strains it was replaced by Z5211-5214 locus of unknown function, similar to the published genome sequences of *E. coli* O157 strains. Noticeably, several types of the *bgl* locus were recognized in *E. coli*. Downstream of the operon, two hypothetical ORFs *yieJ* and *yieI* are present in one group of strains while the *yieI* gene alone is present in another group. The strains in which *bgl* is replaced by the Z211-5214 locus, both *yieJ* and *yieI*

genes are absent. The prevalence of *bgl* and its variability is intriguing in the context of its evolution.

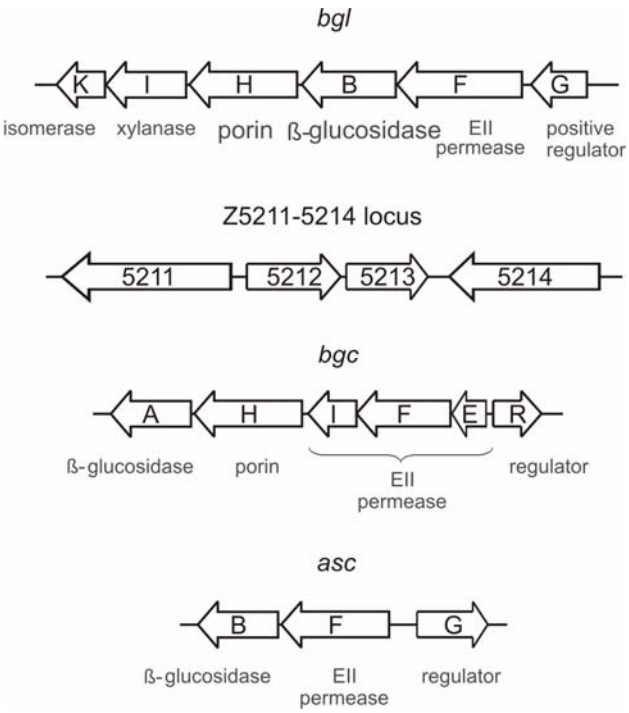

Fig. 3: Cryptic β-glucoside operons in *E. coli.* Genomic structure of three β-glucoside operons, *bgl, bgc*, *asc* and the Z locus replacing *bgl* in *E. coli*. Gene names and protein encoded are indicated. The *bgl* operon is the well-studied example of a cryptic operon, repressed by histone-like nucleoid structuring protein (H-NS) (Dole et al., 2004a). The *bgc* is another cryptic operon involved in utilization of β-glucosides and cellobiose at low termperature (Neelakanta and Schnetz, unpublished data). The *asc* operon is yet another cryptic system in *E. coli* encoding protein for metabolism of arbutin, salicin and cellobiose (Hall and Xu, 1992).

H-NS selectively represses horizontally transferred genes in *E. coli* and *Salmonella* (Lucchini et al., 2006; Navarre et al., 2007). Since *bgl* operon is a model system for the analysis of repression by H-NS, would it be a horizontally transferred operon is a question. Further, the presence of several silent systems leads to the question of the role of cryptic genes in bacterial evolution. Hall et al., (1983) proposed that under one set of conditions members with a cryptic gene are more fit than those members who express it, while under alternative conditions those members expressing the gene are at a selective advantage. This argues for the retention of silent genes, whose

evolution is rather poorly understood. Therefore, a systematic approach to trace the evolution of *bgl* and *bgc* system will shed light on the evolution of bacterial genes, which are retained but with no known advantage.

5. Objectives of the current study

The objective of this study is to contribute to the understanding of the evolution of bacterial genomes by considering the evolution of silent β-glucoside systems as a model in *E. coli*. In the present study, a diverse population of *E. coli* was typed by MLST to establish their phylogenetic and clonal structure. This laid the framework to trace the evolution of the cryptic *bgl* operon and the *bgc* operon. The results revealed the dynamic evolution of the *bgl* operon. It is vertically inherited in enterobactericiae and deleted in some lineages. In *E. coli*, the vertical history of the operon is coupled to evolutionary history of the species, indicating a strong purpose for its retention. The prevalence of the second silent system, *bgc* was analyzed in the population, which revealed the possibilities of gain or loss of *bgc* in evolution. The implications of gene gain and loss on the evolution of individual operons from a genomic perspective as well as the fate of the paradoxically silent operons is discussed.

III. Results

1. Evolutionary genetic analysis of the bgl operon and Z locus in the E. coli population

To investigate the evolution of the *bgl* operon, the prevalence of the operon within a collection of *E. coli* strains was analyzed before (Neelakanta, 2005). However, based on this previous analysis no data were available, which allowed to correlate the evolution of the *bgl* operon and Z5211-5214 locus with the species phylogeny. To achieve such a correlation it is imperative to have a collection of *E. coli* strains which is representative and for which the phylogenetic and population structure is established. In addition, the phylogeny of the *bgl* locus needs to be analyzed at the sequence level.

1.1 Population structure of the *E. coli* collection

The *E. coli* collection used in this study includes strains from diverse sources. These are 98 clinical human isolates, of which 52 are commensals and 46 are pathogens (Dr. G. Plum, Institut für Medizinische Mikrobiologie, Immunologie und Hygiene, Universität zu Köln). In addition, a septicemic *E. coli* strain i484 (Khan and Isaacson, 1998), two uropathogenic *E. coli* (UPEC) strains, J96 and 536 (Brzuszkiewicz et al., 2006), and the 72 strains of the ECOR reference collection (Ochman and Selander, 1984) were analyzed. Furthermore, in the course of the analysis two divergent *E. coli* strains RL325/96 and Z205 which are presumably of an ancestral *E. coli* type (Wirth et al., 2006) and three strains of the closely related species *Escherichia albertii* were included. Thus, the entire collection includes 178 strains.

To determine the phylogeny and the population structure of the *E. coli* collection multilocus sequence typing (MLST) was performed, using the scheme established for *E. coli* by Wirth et al., (2006). Of the collection, 99 strains were of unknown phylogeny. These were subjected to MLST by sequencing fragments of the seven housekeeping genes *adk*, *fumC*, *gyrB*, *icd*,

mdh, *purA* and *recA* for each strain (Wirth et al., 2006). Sequences were curated manually using Bionumerics software (Version 4.0), which was used in the laboratory of M. Achtman, Max Planck Institute for Infection Biology, Berlin. Any ambiguities were resolved by re-sequencing a newly generated PCR fragment and with additional internal primers in case of *adk* and *fumC* genes.

The population structure of bacterial species can be studied by allele-based population genetic analysis. For each housekeeping gene, the different sequences present within a species are assigned distinct alleles (specified by numbers). Further, for each strain, the alleles corresponding to seven loci define the allelic profile or Sequence Type (ST). STs sharing 6 or more than 6 alleles define a clonal complex referred as Sequence Type complex (ST complex). For the allele-based analyses, the MLST data for *E. coli* established in the MLST database (Wirth et al., 2006) (http://web.mpiib-berlin.mpg.de/mlst/dbs/Ecoli) was used as reference. With the help of computational algorithms the following was determined for each strain: a) the allelic profile of 7 genes, i.e. the sequence type (ST), b) the clonal relationship of strains and the sequence type complexes (ST complex) and c) the ancestral group (Wirth et al., 2006). New allele numbers and STs were assigned to strains with novel allelic sequences. Following this 98 out of the 99 strains were entered into the public *E. coli* MLST database (http://web.mpiib-berlin.mpg.de/mlst/dbs/Ecoli). One strain, E466, was identical to E464 and hence omitted. The strain collection typed in this study represents 49 different sequence types (STs). Out of 98 strains 77 occurred in 25 different ST complexes and the remaining are not assigned to any ST complex and simply designated by their STs (Refer to supplementary Table S1 in appendix). Sequences and MLST information for the 72 ECOR strains, UPEC strains J96 and 536, *E. coli* RL325/96, Z205 and three *E. albertii* strains were downloaded from the publicly available *E. coli* MLST database. Two ECOR strains 23 and 32 from the lab collection had ST different from that in the *E. coli* MLST database and hence these two strains were omitted from

this study. Taken together, the entire strain collection represented 92 STs and 115 out of 175 strains appeared in 25 different ST complexes. The MLST data including the individual alleles, STs and ST complexes are listed in Table S1. The result of these population genetic based analyses is visualized on a minimal spanning tree, referred to as MS_{TREE} (Fig. 4). Cumulatively, the analysis established the population structure of the collection of strains used, and it demonstrated that the collection is representative.

To establish the phylogenetic relationships, sequences of the housekeeping genes for each strain were concatenated (3423bp), and the concatenated sequence was used for phylogenetic reconstruction using the neighbor-joining (NJ) method included in the MEGA software V3.1 (Kumar et al., 2004). For convenience, ECOR strains were analyzed separately. The neighbor-joining tree from the sequence data of all strains resulted in four clades concordant with the classical ECOR groups A, B1, B2 and D (Fig. 5). As shown before, the two strains RL325/96 and Z205 are very divergent from the rest of the strains in four clades, and are closely related to three *E. albertii* strains (Wirth et al., 2006). Interestingly, strain E10083 isolated as a human commensal closely clustered with the two ancestral strains RL325/96 and Z205 (Fig. 5). The latter were isolated from dog and parrot respectively (Wirth et al., 2006). This suggests that the human isolate E10083 is probably another remnant of the ancestry of *E. coli*. These three divergent strains are referred to in this study as ancestral *E. coli* and the rest of the strains appearing in four phylogenetic clades, as modern group of *E. coli* as reported before (Wirth et al., 2006). The neighbor-joining tree of the concatenated sequences of housekeeping genes for the ECOR strains was also constructed. This tree was consistent with previous reports (Escobar-Paramo et al., 2004; Lecointre et al., 1998). Thus, the neighbor-joining tree (Fig. 5) represents the whole genome phylogeny of the strain collection used in this study and the strain collection is representative of all phylogenetic groups.

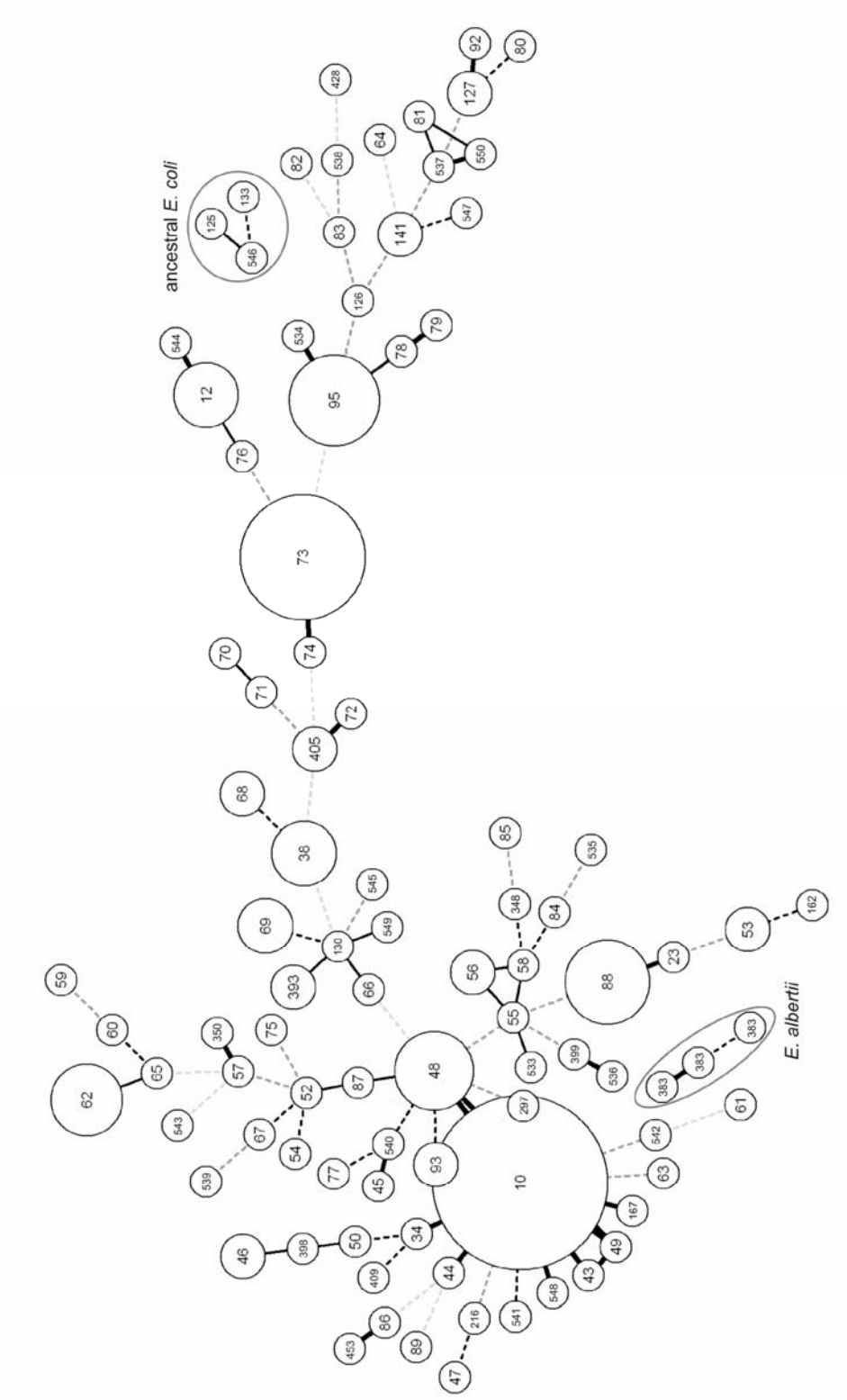

Fig. 4: Minimal spanning tree (MSTREE) depicting the Sequence Types (STs) of strain collection based on MLST analysis. Each circle represents one ST, denoted by its number on the circle. Size of the circle corresponds to number of strains, the smallest represents one strain. Black lines connecting pairs of STs indicate sharing of six alleles (thick lines), five (thin) or four (dotted) between them. Grey dotted lines of increasing length indicate sharing of three to one alleles respectively.

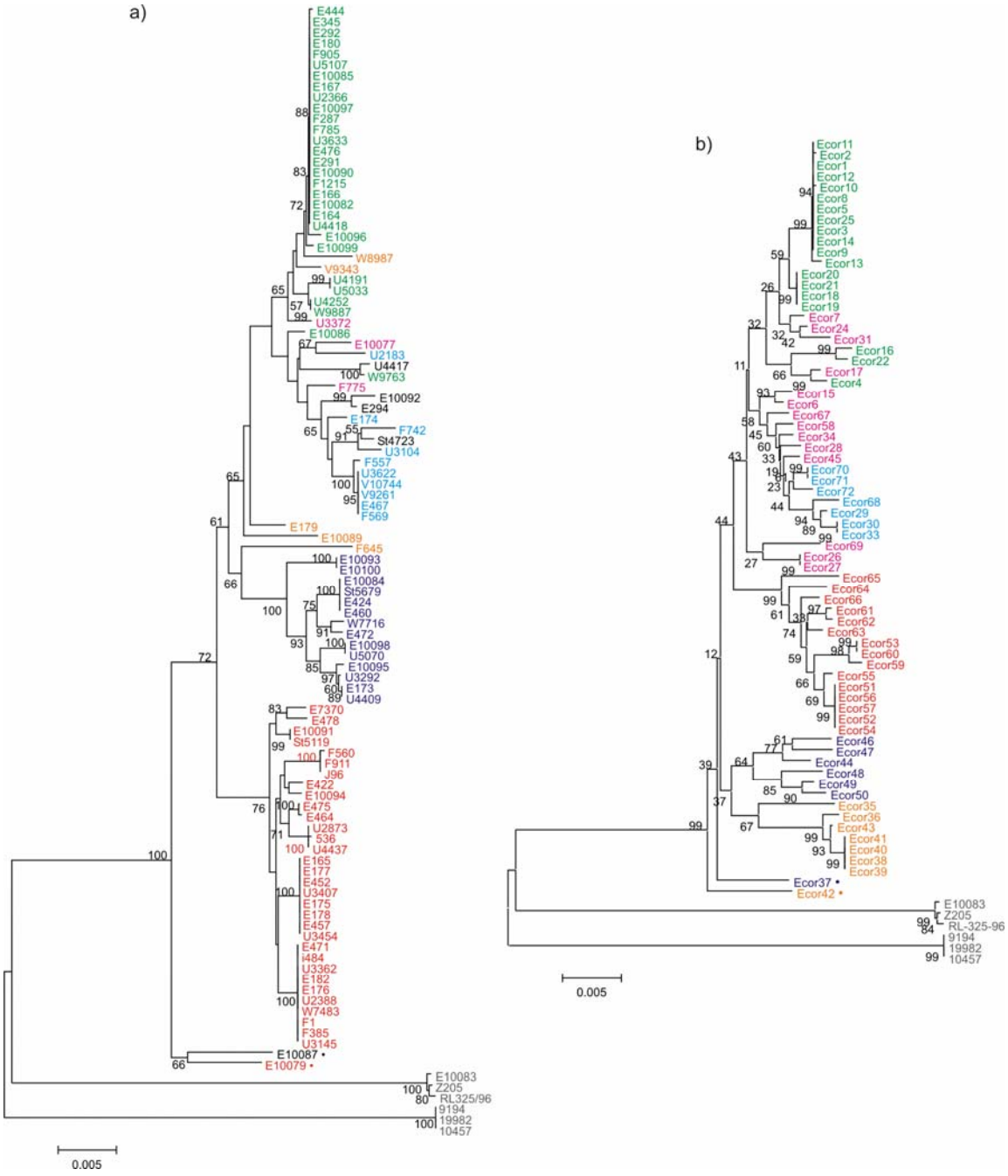

Fig. 5: Whole genome phylogeny. Neighbor-joining trees of concatenated seven MLST housekeeping genes from (a) Isolates (b) ECOR collection. Strains of different phylogenetic groups are color-coded: green-A, cyan-B1, red-B2, blue-D, magenta-AxB1, orange-ABD and grey-ancestral/*E. albertii* strains. Strains displayed in black are not assigned to any group and those indicated (•) are considered odd strains. Numbers on the nodes are bootstrap scores from 1000 replicates and scores above 50% are indicated.

Further, to discern the phylogenetic group of each strain, the information for strains of known ST was extracted from the MLST database. For strains with a new ST, the program STRUCTURE was used with the help of Vartul Sangal in M. Achtman's group as described (Wirth et al., 2006). Wirth et al., (2006) had established a scheme to correlate the ST determined by MLST to the phylogenetic groups A, B1, B2 and D established for the ECOR collection by MLEE (Ochman and Selander, 1984). Further, it was shown that some strains represent hybrids created by recombination. The hybrid group AxB1 represents strains, which derive their ancestry largely from A and B1, and the hybrid group ABD derives its ancestry from all phylogenetic groups. The ancestral group of each modern isolate is listed in (Table S1, appendix). In total 48 strains belong to the phylogenetic group A, 17 strains belong to B1, 48 to B2, 21 to D, 17 strains to AxB1, and 13 strains to ABD respectively. For 5 strains the groups have not been assigned. The ancestry of the isolates as determined by STRUCTURE and graphically displayed using the program DISTRUCT (Rosenberg NA, 2004) (Fig. 6).

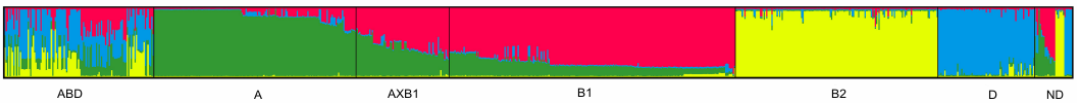

Fig. 6: Ancestry of *E. coli* isolates. Proportions of ancestry from groups A, B1, B2 and D as inferred by STRUCTURE and their assignments to six groups as displayed with DISTRUCT. The plot shows one vertical line for each isolate indicating the proportions of ancestry from the four groups, which are color-coded as green (A); red (B1); yellow (B2) and blue (D). The groups are indicated at the bottom and ND refers to strains for which groups are not assigned.

1.2 Genetic diversity of the *bgl* operon/ Z5211-5214 locus in *E. coli* natural isolates

The *bgl* operon of *E. coli* is not expressed under laboratory conditions due to effective silencing by the histone-like nucleoid structuring protein H-NS (Defez and De, 1981; Schaefler and Maas, 1967 Higgins et al., 1988, Mahadevan and Wright, 1987, Schnetz, 1995, Dole et al., 2004b; Nagarajavel et al., 2007). Previous analysis in the laboratory revealed that *bgl* operon is highly prevalent in the population (Neelakanta, 2005). In order to study the genetic diversity of the *bgl* and Z5211-5214 loci, phylogenetic analysis was carried

out. Earlier analysis in typing *bgl* and Z5211-5214 also showed that sequences upstream and downstream of *bgl* are polymorphic. To analyze the genetic variation more systematically, fragments of DNA from the *bgl* and Z5211-5214 loci were sequenced (Fig. 7). Fragments were amplified by PCR and sequenced on both the strands. The PCR analysis was consistent with earlier results (Neelakanta, 2005), which revealed that 78% of isolates (136 of 175) carry *bgl* operon and 19% lacked *bgl* operon, but carried a different locus containing four open reading frames as annotated in the *E. coli* 0157:H7 EDL933 genome sequence. The prevalence of *bgl* operon and the phylogenetic groups showed a correlation, when both were related to each other. Noticeably *bgl* operon was present in strains of A, B1, B2 groups and totally absent in D group strains. Noticeably the A and B1 strains have the genes *yieJ* and *yieI* present downstream of *bgl* operon and the B2 strains lack the *yieJ* gene. The D strains have neither *yieJ* nor *yieI* gene (Fig. 8A).

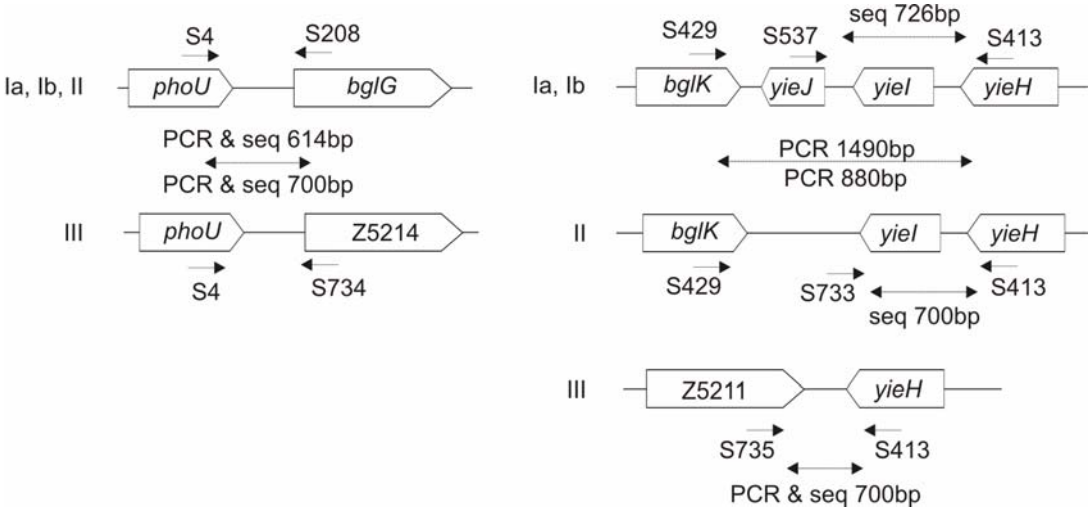

Fig. 7: PCR and sequence typing of *bgl*/Z loci. Primers and their mapping positions are indicated. Groups of *bgl*/Z loci are given at the left of the genomic structure of the loci. Sizes of the PCR product sequenced on both the strands are given. Same primers were used for both PCR and sequencing, otherwise indicated in case of sequencing fragments downstream of *bgl* operon. Multiplex PCR reaction consisting of primers S4, S208 and S734 was performed to distinguish between strains having *bgl* (group Ia, Ib and II) and Z5211-5214 locus (group III). For sequencing fragments downstream of *bgl*, PCR reaction with S429 and S413 was performed and the products were1490bp (group Ia, Ib) and 880bp (group II), which were sequenced with primers S537 and S733 respectively on the forward strand and S413 on the reverse strand. Z5211-5214 was PCR amplified and sequenced independently.

To phylogenetically reconstruct evolution of the *bgl* locus, fragments of sequences following the stop codon of the flanking genes, which are internal to a breakpoint (point from where polymorphism was observed in *bgl* locus), were concatenated. 811bp (537+277bp) of sequence of *bgl* locus for each isolate was obtained (Fig. 7B). Isolates in which the analyzed region of the *bgl* operon or the Z5211-5214 locus was disrupted by insertions and/or deletion

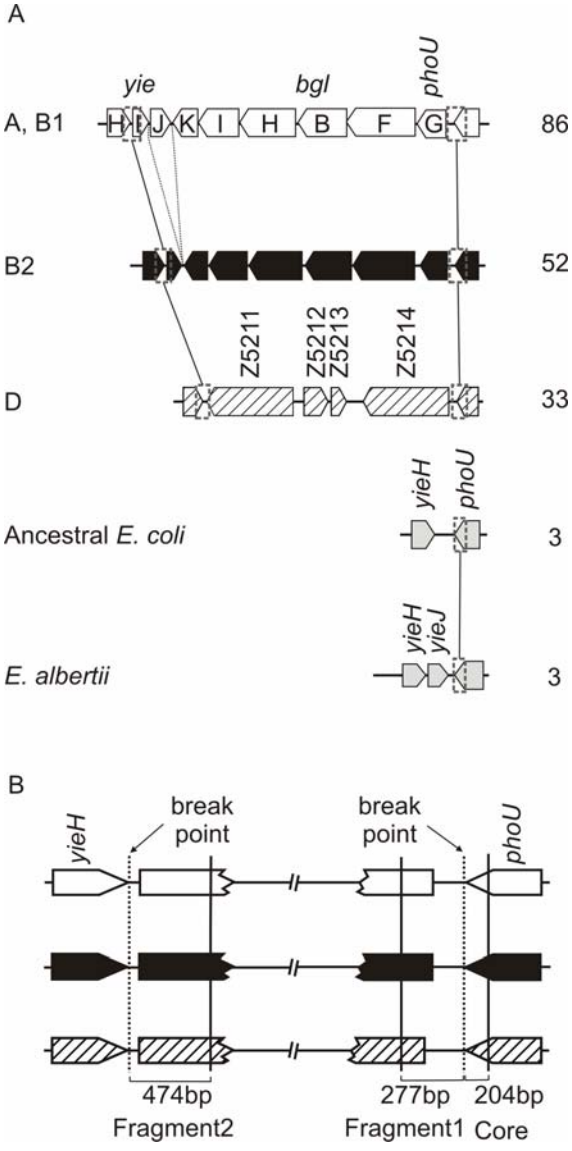

Fig. 8: A) Genomic organization of the *E. coli bgl*/Z5211-5214 loci. Gene names are indicated within genes and on the top for flanking genes and Z5211-5214 locus. *E. coli* phylogenetic groups are indicated on the left and the number of strains on the right. B2 strains lack the *yieJ* gene. Ancestral *E. coli* and *E. albertii* strains lack *bgl* operon/Z5211-5214 locus, while *E. albertii* strains carry *yieJ* gene. (B) Schematic illustration of sequencing strategy. Fragments of the *bgl* operon/Z5211-5214 locus (1& 2) and the flanking gene phoU of the core genome (core) were sequenced from the indicated breakpoints on the left and right. Sequence lengths are given at the bottom.

(see later) were omitted. A multiple alignment was generated using the concatenated 811bp sequence of *bgl* from each isolate, and this was used for phylogenetic reconstruction according to the neighbor-joining method (Fig. 9A). Again, a separate tree was built for the sequences derived from the

ECOR strains (Fig. 10). For further comparison, the respective *bgl* sequences from the published *E. coli* genome sequences were also included in the phylogenetic analysis. The tree established 3 clusters, which were named *bgl* group Ia, Ib, and II (Fig. 9A). This demonstrates the presence of three groups of the *bgl* locus in modern *E. coli* isolates.

The sequences of the Z5211-5214 locus was also phylogenetically analyzed (Fig. 26 appendix), but no groups were assigned as the tree showed several clusters, probably denoting rapid evolution of the locus (Fig. 26, Appendix). Hence, the strains harboring this locus were arbitrarily assigned to group III (Fig. 9A). The ancestral *E. coli* isolates and the *E. albertii* strains lack both the *bgl* operon and the Z5211-5214 locus (Fig. 8A). The identified structure of the genome in these strains was assigned to groups IV and V respectively. Strikingly similar results were obtained with ECOR strains (Fig. 10). These analyses demonstrate that three phylogenetic groups of the *bgl* operon exist among the modern group of *E. coli* and Z5211-5214 locus can be considered as the fourth group in which *bgl* is replaced. Intriguingly, the absence of *bgl* operon/ Z5211-5214 locus in the ancestral *E. coli* and the related *E. albertii* might indicate a probability of horizontal transfer of these two loci into modern *E. coli* isolates.

1.3 Genetic variation in the core genome flanking the *bgl* operon/Z5211-5214 locus

The sequences obtained from the regions flanking the *bgl* and Z5211-5214 loci, respectively were also phylogenetically analyzed. The *phoU* gene flanking the two loci is present in all *E. coli* isolates, including the ancestral strains and *E. albertii*. The *phoU* gene is essential for survival of *E. coli* when phosphate is limiting, a condition that is frequent in the natural habitats of *E. coli* (Buckles et al., 2006; Steed and Wanner, 1993). Therefore, *phoU* belongs to the core genome of *E. coli*. Partial sequences of the *phoU* gene were obtained from all the isolates and 204bp of fragments of these sequences were used to construct a tree by the neighbor-joining method. The tree

resulted in four clusters (Fig. 9B) comparable to the four phylogenetic groups

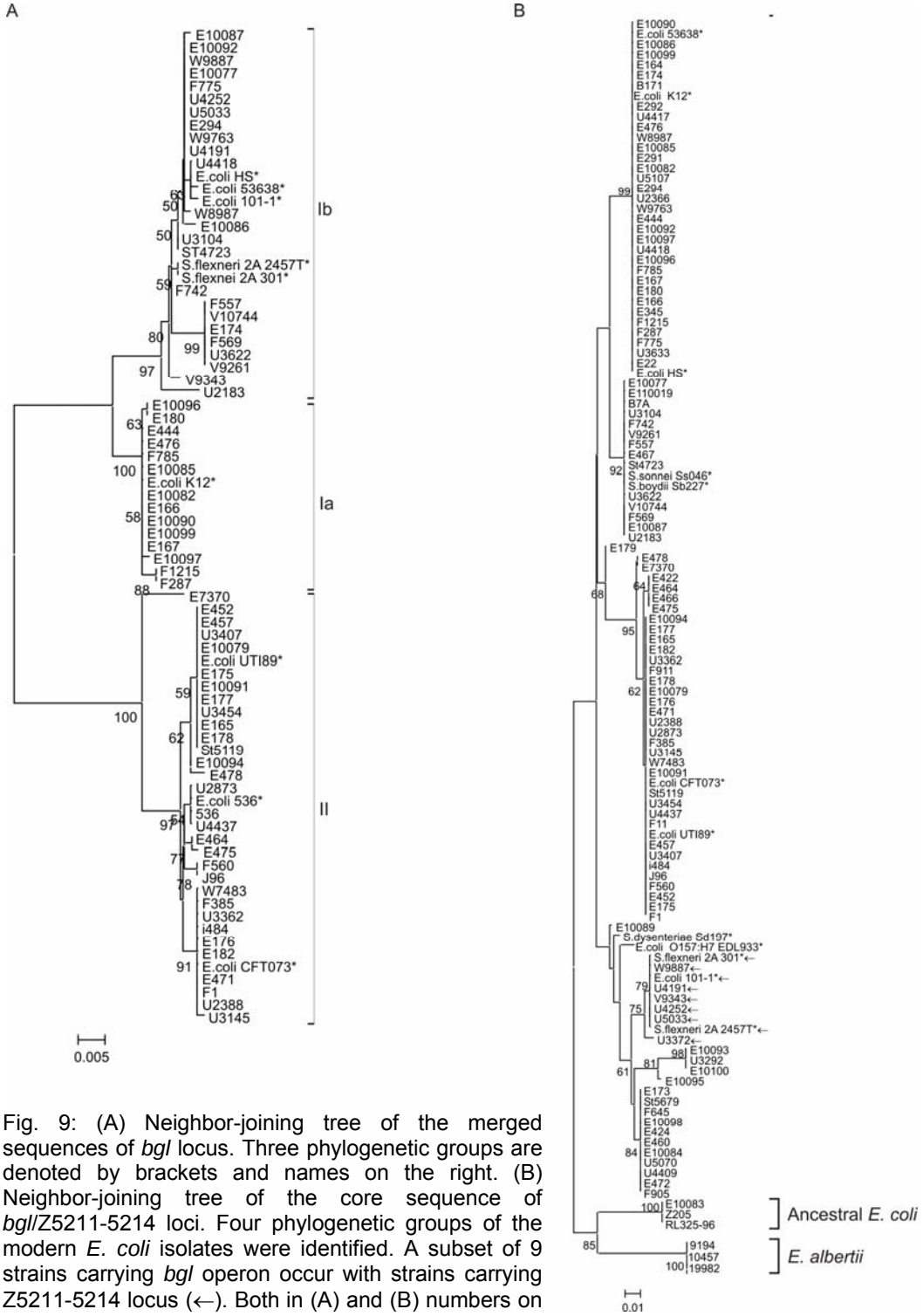

Fig. 9: (A) Neighbor-joining tree of the merged sequences of *bgl* locus. Three phylogenetic groups are denoted by brackets and names on the right. (B) Neighbor-joining tree of the core sequence of *bgl*/Z5211-5214 loci. Four phylogenetic groups of the modern *E. coli* isolates were identified. A subset of 9 strains carrying *bgl* operon occur with strains carrying Z5211-5214 locus (←). Both in (A) and (B) numbers on the nodes are bootstrap scores (above 50%) from 1000 replicates. (*) denotes strains for which sequences were obtained from NCBI.

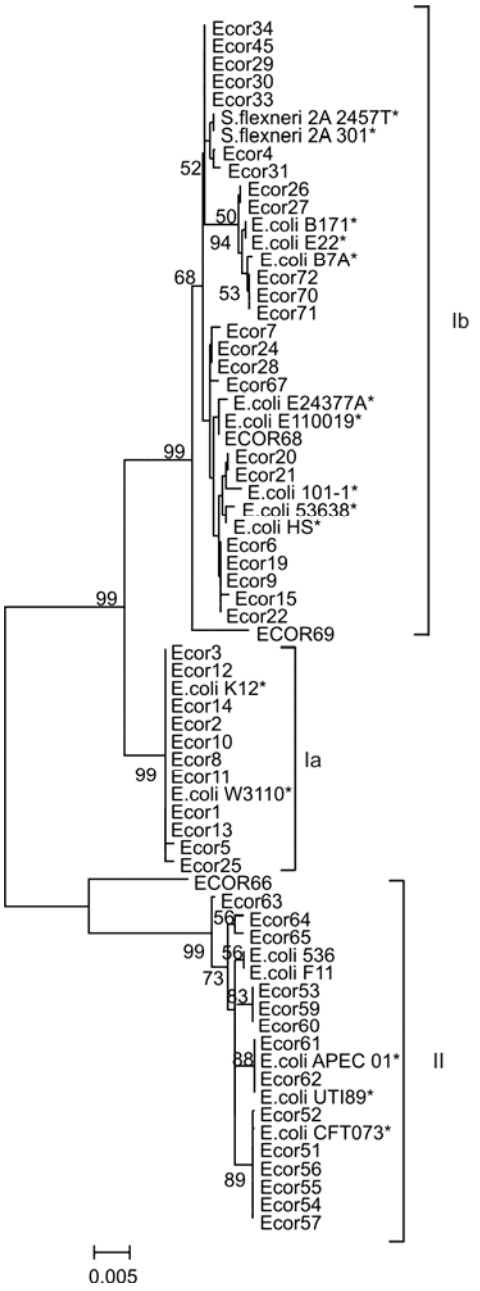

Fig. 10: Neighbor-joining tree of the merged sequences of *bgl* locus from ECOR strains (see Fig.6B). Three phylogenetic groups are denoted by brackets and names on the right. Numbers on the nodes are bootstrap values from 1000 replicates and above 50% are denoted. (*) indicates strains for which sequences were obtained from NCBI sequence databank.

of *E. coli* and comparable to the tree obtained by concatenating seven housekeeping genes. The ancestral and *E. albertii* isolates were distant from the modern isolates in the *phoU* tree, which is similar to the whole genome tree based on housekeeping genes. Strains in clades 1, 2 and 3 of the *phoU* tree possess the *bgl* operon. Strains clustering in clades 1, 2 and 3 are *bgl* Ia, Ib and II respectively. All strains that carry the Z5211-5214 locus were present in clade 4. Intriguingly, nine strains in clade 4 harbored the *bgl* operon (Ib).

Thus, for these strains the presence of *bgl* did not correlate to the *phoU* phylogenetic clades, which may indicate recombination. Thus, the diversity of *bgl* operon/ Z5211-5214 locus is reflected on the core genome flanking these loci and provides striking evidence for parallel evolution. The result may indicate a more recent recombination event for strains that carry the *bgl* group Ib operon.

1.4 The evolution of the *bgl*/Z5211-5214 locus is coupled to species evolution

The phylogeny of the *bgl* operon derived from the sequenced fragments is comparable to the phylogeny of the core gene *phoU* flanking the operon. Further, to deduce an evolutionary relationship of the *bgl* operon/Z5211-5214 locus with the species, a one-to-one phylogenetic comparison was performed between their phylogenies. The *bgl* groups were marked on the whole genome phylogenetic tree (Fig. 11). The clustering of strains in the *bgl* operon phylogeny was consistent with the major clades or phylogenetic groups seen in the whole genome phylogeny generated from the concatenation of seven housekeeping genes. This reveals that the *bgl* operon shares the same evolutionary history as the species. The Z5211-5214 locus is exclusively present in D group strains. Owing to the absence of the Z5211-5214 locus in the other three phylogenetic groups (A, B1, B2) of *E. coli,* it is possible that the Z locus was horizontally transferred into the ancestor of D group strains. An interesting exception is strain F905, which carries the Z5211-5214 locus, but clusters with the A group strains. This incongruence suggests a recent transfer of Z5211-5214 locus into strain F905. Leaving out the single exception, the strong congruence implies that the loci have a shared evolutionary history with the species. Further, this evolutionary congruence is augmented by similar results from the ECOR strains (Fig. 11).

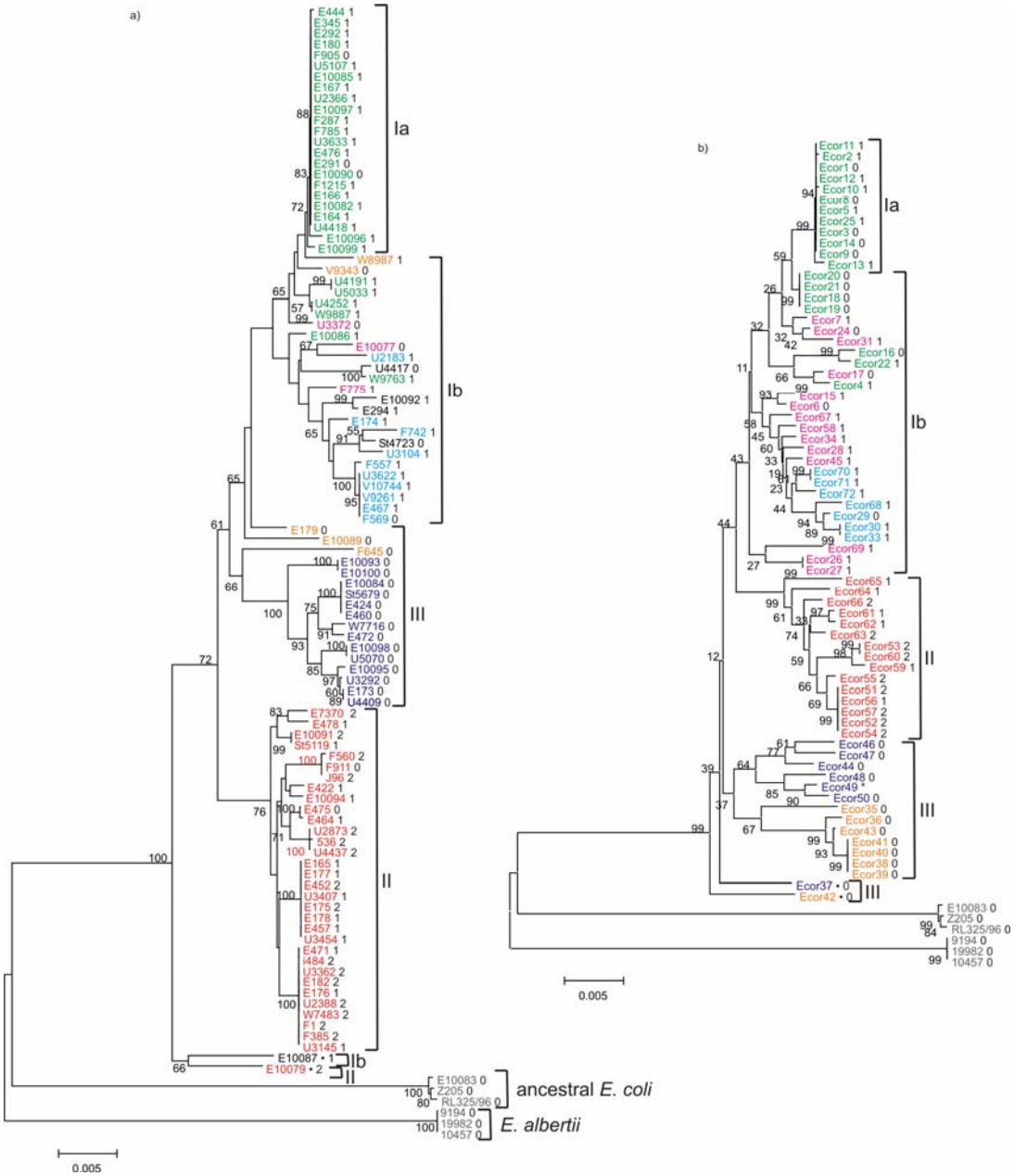

Fig. 11: *bgl* /Z5211-5214 loci groups and *Bgl* phenotype marked on the phylogenetic tree obtained from concatenation of seven housekeeping genes (represented earlier in Fig.5). Groups are indicated by brackets on the right. *Bgl* phenotype next to strain name. “1”-*Bgl*- with papillae formation indicative of presence of a functional but repressed *bgl* operon; “0“ refers to *Bgl*- without papillation indicating absence of a functional *bgl* and “2” refers to a weak positive (or relaxed) phenotype after 3-5 days of incubation at 37°C. Refer to Fig.5 for the color codes of the phylogenetic groups of the strains.

1.5 Clonal evolution of the *bgl* operon/Z5211-5214 locus

It is a controversial view that phylogenetic approaches are largely unsuitable for most modern *E. coli* and hence allele-based population genetic analyses of *E. coli* strains were considered more informative to discern the deep evolutionary relationships (Wirth et al., 2006). MLST data from the strains are used to identify clonal groups based on the sharing of the allelic profiles (see section 1.1). To deduce the clonal evolution of the *bgl* operon/Z5211-5214 locus, their groups obtained by sequencing and phylogenetic analysis were mapped on the MS_{TREE} depicting the clonal structure of the strains used in this study. The groups of the *bgl* operon/Z5211-5214 were color coded and represented within the MS_{TREE} (Fig. 12). In the MS_{TREE} every circle representing an ST acquired a uniform color indicating that all the strains were of the same *bgl* operon or Z5211-5214 group. The *bgl* Ia group mapped exclusive to the ST10 complex which contains ST10 and related STs. *bgl* Ib appeared exclusively in several ST complexes such as ST23, ST10, ST86, ST155, totally in 13 different complexes. *bgl* II was largely restricted to ST73, ST95 and ST12 complexes. Z5211-5214 strains occurred in different ST complexes, like ST31, ST38, and ST59. Importantly, there is almost no intermingling of *bgl*-Z groups in a single ST or ST complex. Two exceptions were found contradicting this strong congruence. As previously noted in the phylogenetic analysis, strain F905 with ST10 (ST10 complex, A group) is the only strain lacking the *bgl* operon but harboring the Z5211-5214 locus. This indicates the possibility of horizontal transfer of Z5211-5214 locus into ST10 complex. Another likelihood of horizontal transfer of Z5211-5214 locus was noted in ST350 complex with two strains Ecor31 and E179 (STs, ST57 and 350 respectively). Ecor31 has *bgl* operon group Ia, whereas E179 has Z5211-5214 locus, suggesting the introduction of Z locus in E179 by horizontal transfer. These results show that the prevalence of *bgl* and Z groups strongly fits the clonal structure of the species, which suggests that the *bgl* operon and Z5211-5214 locus clonally descended with *E. coli.*

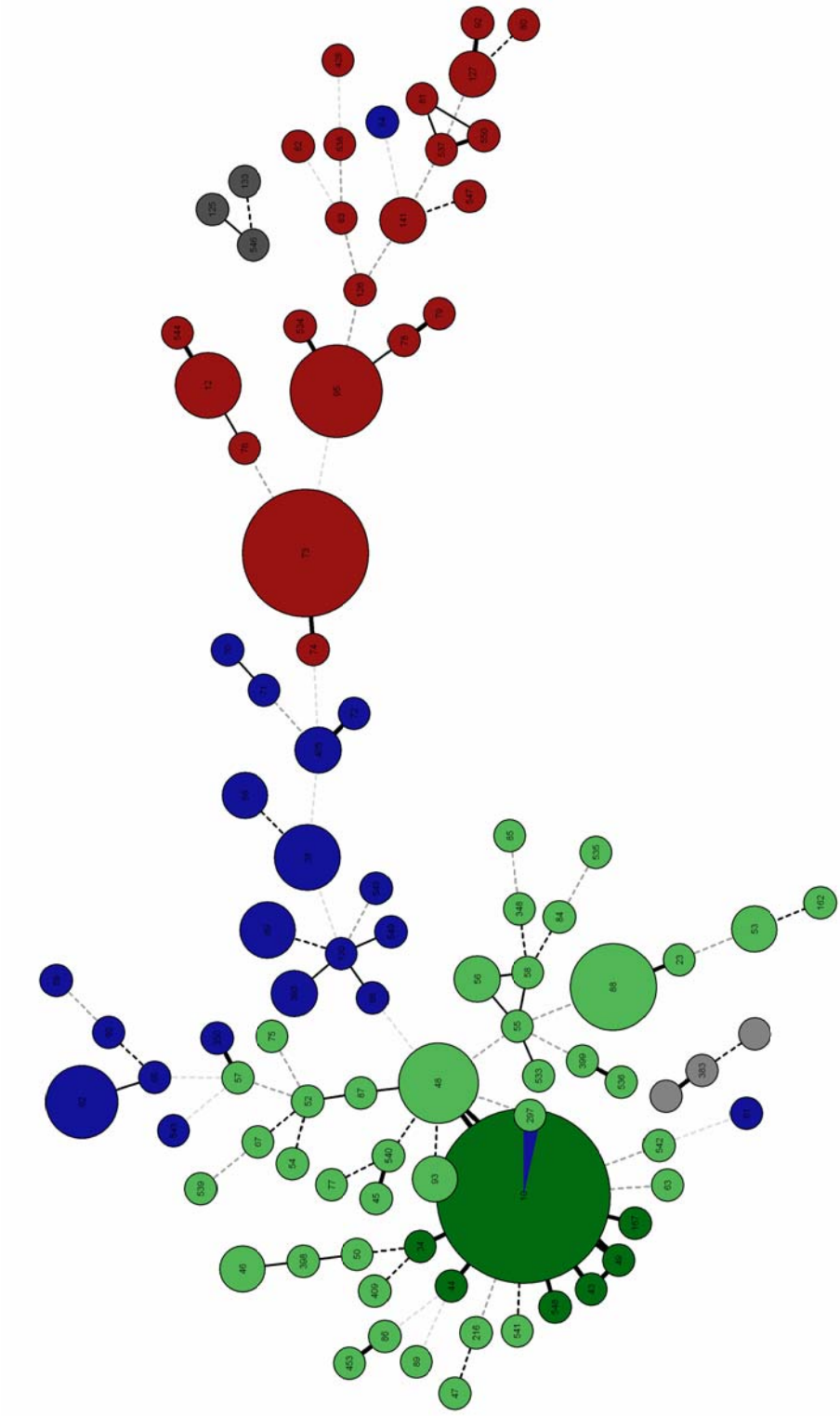

Fig. 12: Groups of *bgl*Z5211-5214 loci on MSTREE represented in Fig. 4. The groups are color codes: Ia (dark green); Ib (light green); II (dark red) and III (blue).

1.6 Phylogenetic analysis of the complete *bgl* operon sequence

The genetic diversity of the *bgl* operon, analyzed from fragments within the locus gave rise to three phylogenetic groups, which are strongly congruent with the phylogenetic groups and the clonal complexes of the species. Further, to analyze the genetic diversity of the complete *bgl* operon, the sequence of the entire locus, which includes six genes of *bgl* operon (*bglGFBHIK*) and downstream *yieJ*/*yieI* genes, were obtained from the whole genome sequences of *E. coli* and *Shigella* strains available at NCBI Microbial genomes. Sequences were extracted from 17 strains (Fig. 13) and a multiple alignment was generated using ClustalW implemented in MEGA software V3.1. Insertion sequences distorting the multiple alignment were removed from strains having IS elements (Strains *S. flexneri* 2A-301, 2457T, *E. coli* E22, E110019, 53638). Neighbor-joining method was used to build the phylogeny of the *bgl* locus including *yieJ*/*yieI* genes. The tree likewise revealed three phylogenetic groups (Ia, Ib and II) very similar to the phylogenetic tree obtained from partial sequences of the locus (Fig. 13). Thus, indeed the *bgl* locus diverged into three groups within *E. coli*.

To see, if the phylogenetic relation obtained with the complete *bgl* operon together with downstream *yieI* gene also correlates to the species evolution, of the 17 *E. coli* and *Shigella* genome sequences, the sequences of the seven housekeeping genes used in MLST analysis were extracted. The sequences were concatenated as before (see section 1.1) and used for phylogenetic analysis. The resulting neighbor-joining tree of the housekeeping genes identified three phylogenetic groups of the strains. The tree of the housekeeping genes and the *bgl* operon tree showed strong congruence (Fig. 13). Incongruence was noted with strains HS and 53658, which are closely related to strains MG1655 and W3110 (group Ia) in the *bgl* locus tree, but distantly related in the housekeeping genes tree. The incongruence could indicate a putative recombination at the *bgl* operon in strains HS and 53638 with the *bgl* Ia strains. The phylogeny of the complete *bgl* locus is comparable to the phylogenetic relations obtained by the analysis of *bgl* sequence

fragments presented above. Yet again, these results suggest that the core genome and the *bgl* locus have a shared evolutionary history.

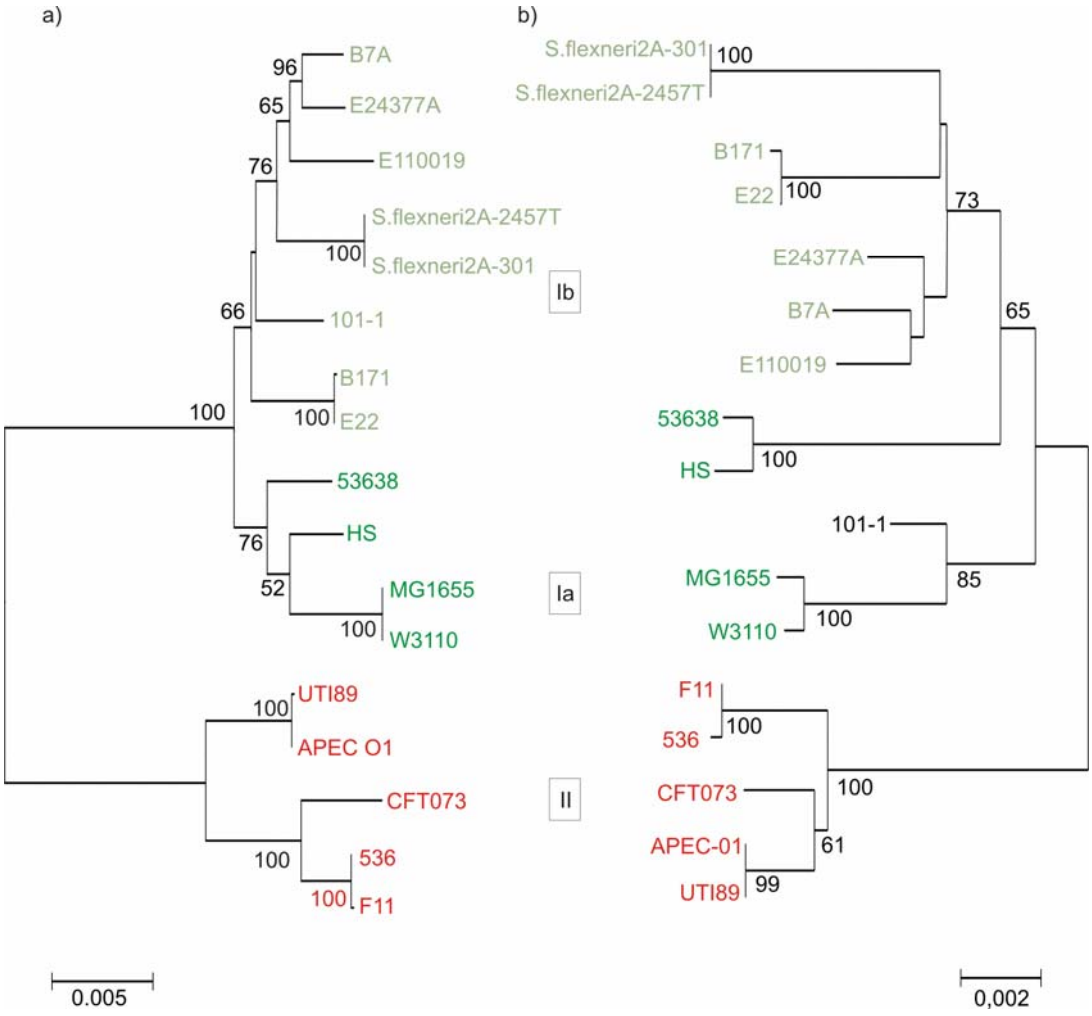

Fig. 13: Phylogenetic comparison of *bgl* locus and species trees. (a) Neighbor-joining tree of complete *bgl* locus from 17 strains obtained from NCBI sequence databank. (b) Strain phylogeny based on 7 MLST house-keeping genes. Boxed numbers in the middle refer to phylogenetic groups of *bgl* locus. Numbers on the nodes are bootstrap values from 1000 replicates (scores above 50 are indicated). Scale at the bottom depicts evolutionary distance.

1.7 Functional analysis of *bgl* operon

The *bgl* operon encodes the proteins for utilization of β-glucosidic sugars arbutin and salicin. Due to repression of *bgl* by H-NS wild type *E. coli* K12 cells are phenotypically *Bgl*$^-$. However, in *E. coli* K12 spontaneous *Bgl*$^+$ mutants arise as papillae (Schaefler and Malamy, 1969). In previous work performed in the laboratory by G. Neelakanta (2005), the *Bgl*$^-$ phenotype of all strains of the *E. coli* collection was tested on BTB salicin indicator plates at 28°C and 37°C (Neelakanta, 2005) and three phenotypes were distinguished. The phenotypes identified, were (i) *Bgl*$^-$, without papillation indicating that no functional operon is present, (ii) *Bgl*$^-$ with papillation indicative of the presence of a functional but repressed operon, (iii) a weak positive (or 'relaxed') phenotype after 3 to 5 days of incubation at 37°C, and (iv) one strain (Ecor49) showed a *Bgl*$^+$ phenotype.

In order to analyze a relationship between the *bgl* genotype, phenotype and the phylogenetic groups of *E. coli*, the phenotypes were mapped on the housekeeping genes tree (Fig. 11). To this end, the phenotypes were classified into types 0, 1 and 2, where 'type 0' was assigned to non papillating *Bgl*$^-$ strains, type 1 to papillating *Bgl*$^-$ strains, and 'type 2' to strains with a relaxed phenotype. These phenotypic types were marked on the housekeeping genes tree in which the *bgl*/Z groups were marked previously (Fig. 11). The marking of phenotypic groups on the tree revealed that majority of A, B1, hybrid AxB1 strains (*bgl* Ia or Ib group) showed *Bgl*$^+$ papillation phenotype and a minority showed *Bgl*$^-$ phenotype. All but one D group strain, which carried Z5211-5214 locus exhibited *Bgl*$^-$ phenotype, as expected. Ecor49 was the only D strain showing a weak *Bgl*$^+$ phenotype on day2. Noticeably more than half of the B2 strains corresponding to *bgl* II group exhibited the relaxed phenotype. The rest of the B2 strains showed *Bgl*$^-$ papillation phenotype except two B2 strains. Further, the phenotypic types 0, 1 and 2 were visualized on the MS_{TREE} depicting the clonal complexes (Fig.14). This visualization resulted in similar correlation seen

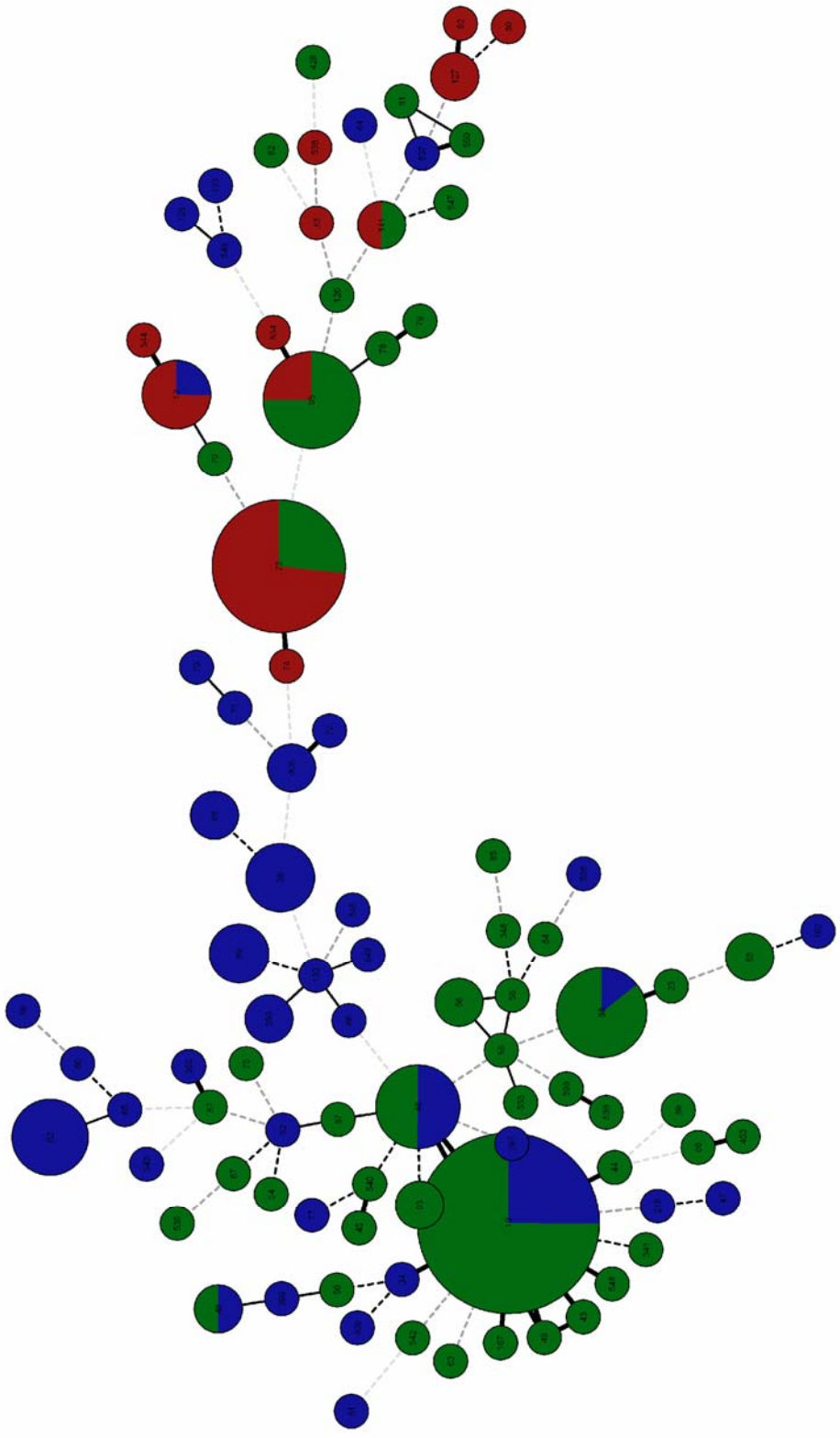

Fig. 14: Groups assigned on *Bgl* phenotype mapped on the MS_{TREE} presented earlier in Fig.4. The groups are color-coded: type “0” - *Bgl*$^{-}$ non papillating (blue); “1” - *Bgl*$^{-}$ but papillating (dark green) and “II” - weakly *Bgl*$^{+}$ (dark red)**.**

above with the phylogenetic groups. Strains in ST10, ST23 complexes and of multiple STs corresponding to *bgl* Ia or Ib displayed 20% non-papillating *Bgl*$^-$ and 80% papillating *Bgl*$^-$ phenotypes. Strains in ST73, ST95, ST12 complexes exhibited 3% *Bgl*$^-$, 40% papillating *Bgl*$^-$ and 57% relaxed phenotypes. This demonstrates that only strains that belong to the B2 phylogenetic group and concomitantly carrying a *bgl* of the group II may have a relaxed phenotype. The mapping of the phenotypes on the MS_{TREE} revealed that strains of the two clonal groups of ST73 and 12 show a high frequency of the relaxed phenotype. This relaxed phenotype and weak expression of *bgl* might be selected in these strains.

1.8 Insertions/deletions in *bgl* /Z5211-5214 loci

Typing of the *bgl* operon/Z5211-5214 locus not only revealed the presence/absence of the loci but also indicated the presence of insertion sequences or insertion associated deletions within the loci (Neelakanta, 2005). In the earlier work of G. Neelakanta (2005), there were discrepancies in the sequencing of the insertions identified in the strains. Those discrepancies were resolved in the current study by re-sequencing in 12 strains. In perceiving the correlations between occurrence of insertion/deletions and groups of *bgl*, it was notable that predominantly disruption of *bgl* operon was seen in *bgl* Ia strains and relatively less in *bgl* Ib and II strains.

1.9 The *bgl* operon is vertically inherited in Enterobactericiae

The typing of the *bgl* operon/Z5211-5214 locus revealed that both loci are absent in the ancestral *E. coli* strains and in the closest species *E. albertii*. This poses several possible scenarios. Firstly, the *bgl* operon and the Z5211-5214 locus have been horizontally introduced into the modern group of *E. coli*. Secondly, vertical inheritance of the *bgl* operon and horizontal transfer of Z5211-5214 into the ancestor of D strains and loss in ancestral and *E. albertii* strains is possible. To test the above possibilities, proteobacterial genomic sequences were searched using tblastn program of BLAST (Altschul et al.,

1990) for orthologs of genes of the *bgl* and Z5211-5214 loci. The *bgl* sequence of *E. coli* K12 strain MG1655 was used as the query, and BLAST was performed for individual genes *bgl*GFBHIK and *yieJ*, *yieI* genes. Similarly, the individual genes of the Z5211-5214 locus were used for BLAST to obtain orthologs using the *E. coli* 0157:H7 EDL933 sequence as query. Searching for *bgl* operon orthologs identified BLAST hits among members of enterobactericiae family, *Klebsiella* sp, *Erwinia* sp and *Photorhabdus* sp, with multiple hits above 30% identity in *Erwinia* and *Klebsiella*. In addition, BLAST yielded very week hits in *Yersinia* and other gamma-proteobacteria members like *Vibrio*. Protein sequences of the hits were obtained from NCBI sequence databank and used for phylogenetic analysis.

In order to determine whether the *bgl* operon is vertically transmitted, individual genealogies were constructed from protein sequences and compared to the phylogeny of the strains. Neighbor-joining trees were generated from individual protein sequences (Fig. 15). The species phylogeny for representative members of enterobactericiae was reconstructed with 16S rDNA sequences obtained from Ribosomal Database Project (RDP) hosted by Michigan State University. Independent phylogenetic analyses of *bgl* orthologs revealed a high level of congruence to that of 16S rDNA of the strains. This indicates that the *bgl* operon is vertically inherited in enterobactericiae.

Surprisingly BLAST results for *yieJ*/*yieI* genes yielded strong hits only within *E. coli/Shigella* and weak identity hits in few other bacteria. Neighbor-joining trees from the protein sequence were constructed (Fig. 15). In the *yieJ* phylogeny, the *E. albertii yieJ* gene, sequenced in this study was included. The phylogenies of *yieJ*/*yieI* are highly inconsistent with that of 16S rDNA arguing against vertical inheritance of these genes. Therefore, *yieJ* and *yieI* are potentially, horizontally transferred genes into *E. coli,* consistent with the previous report (Lawrence and Ochman, 1998). Collectively, these results suggest that genes of the *bgl* operon are vertically inherited from a common ancestor of enterobactericiae.

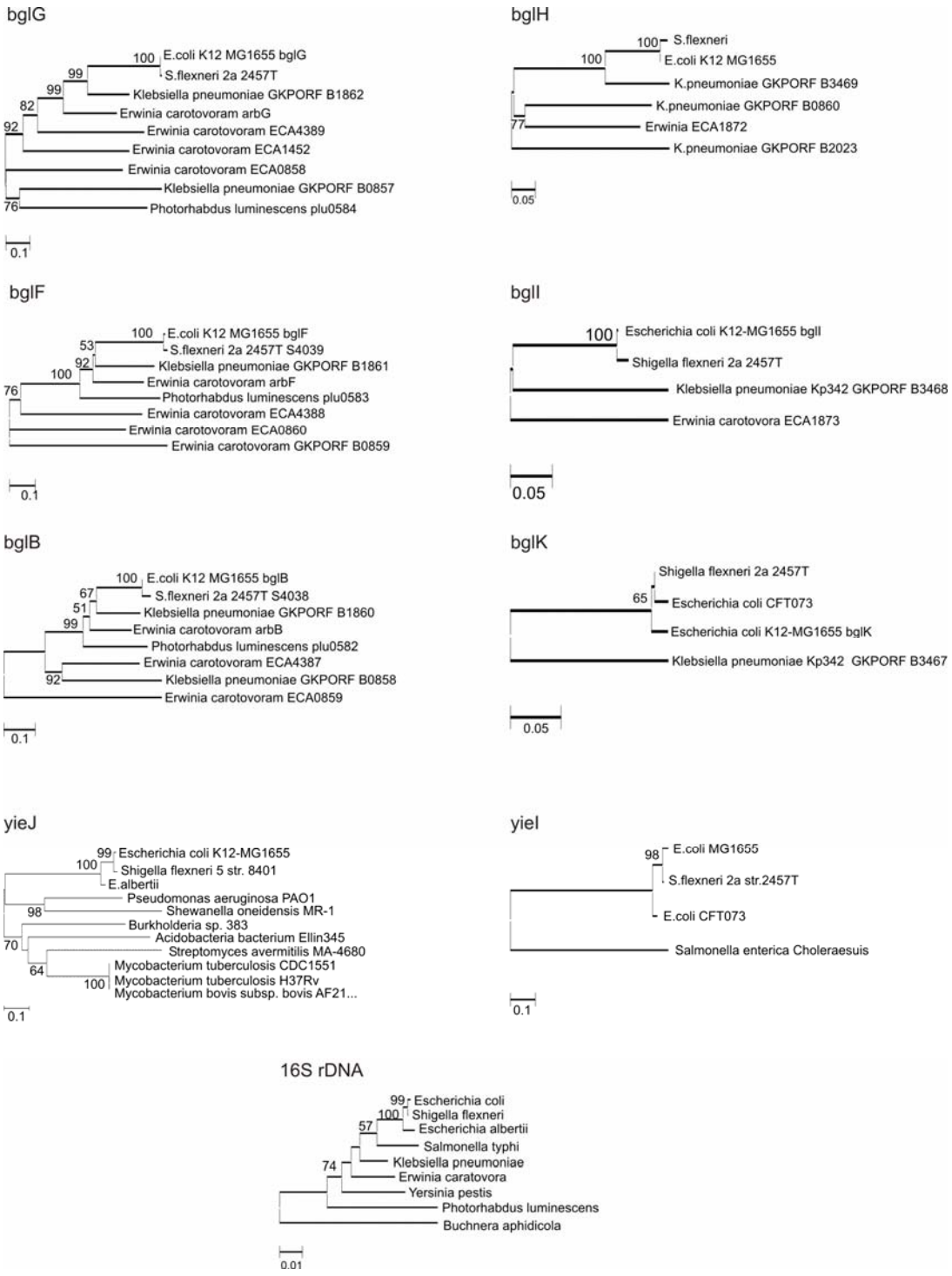

Fig. 15: Neighbor-joining trees of genes of *bgl* locus constructed from the protein sequence of *E. coli* K12 MG1655 and its orthologs obtained by BLAST; 16S rDNA sequences of representative members of enterobactericiae. Individual gene names are indicated on top of the tree. Numbers on the nodes are bootstrap values from 1000 replicates (scores above 50 are indicated). Scale at the bottom depicts evolutionary distance.

These genes were retained in *E. coli* (including *Shigella)*, *Erwinia*, and *Klebsiella* species and lost in other enterobacteriaceae members analyzed in this study. Furthermore, it was interesting to observe the structure of *bgl* orthologs in the other bacteria. In *Klebsiella sp*, *Erwinia sp*, *bgl* genes are organized in a similar fashion at least the first three genes of the operon *bglGFB* as in *E. coli* but in a different chromosomal location (Fig. 16). In *Klebsiella*, the orthologs of *bglHIK* are present in the same chromosomal position as in *E. coli*.

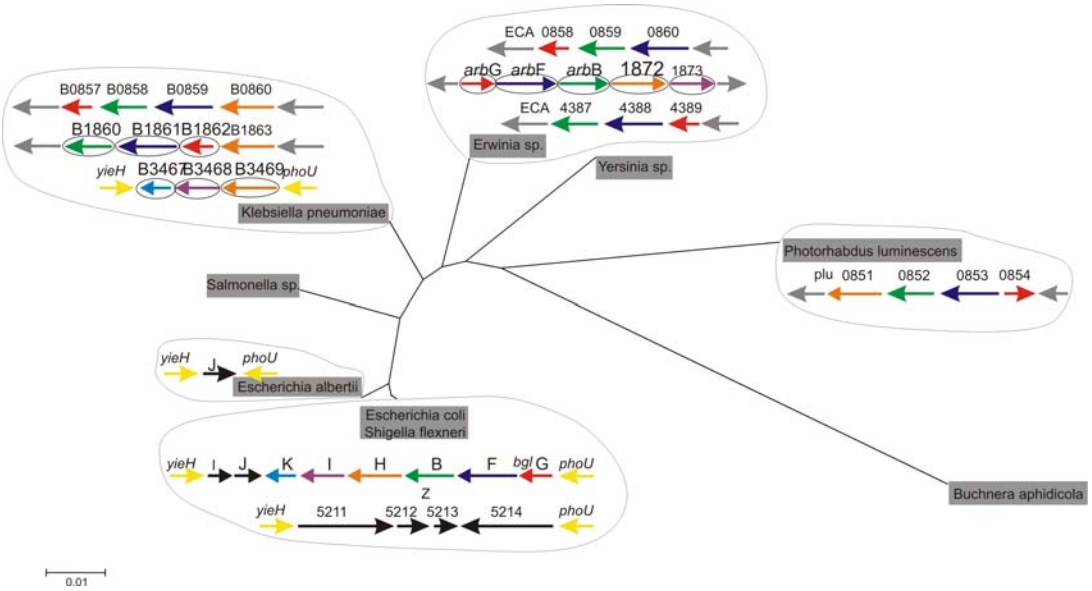

Fig. 16: Structure of *bgl* locus genes of *E. coli* K12 MG1655 and its orthologs obtained from BLAST in Enterobactericiae displayed on a neighbor-joining tree of 16S rDNA sequences of the indicated members. Gene names are indicated on top and colors refer to different genes of *bgl* locus keeping *E. coli* as reference.

In contrast, no orthologs were obtained for Z5211-5214 locus, even when the search was extended to the entire non-redundant database (nr) at the NCBI. Furthermore, the compositional analysis showed that the GC content of Z5211-5214 (31%) locus was significantly lower than the *E. coli* genome average (50.4%) (data not shown), which suggests that Z5211 to Z5214 are horizontally acquired genes.

1.10 The *bgl* operon does not affect fitness in LB medium

The data presented above demonstrate that the *bgl* operon was vertically inherited and is conserved in a silent or relaxed silent state in three of the four phylogenetic groups of *E. coli*. Until date, no conditions are known under which the operon is expressed. However, the operon can be activated by point mutations and by insertion sequences (Lopilato and Wright, 1990), (Reynolds et al., 1986), and it was speculated that activation is an evolved mechanism that allows *bgl* expression under specific physiological conditions. Further, it has been shown that the activated *bgl* operon in the presence of *rpoS819* allele confers a fitness advantage during stationary phase (Madan et al., 2005). Another possibility is that silencing of the *bgl* operon is relieved under certain conditions, for example when resident in the host as a commensal or a pathogen. This assumption is supported by the result that the *bgl* operon is weakly expressed in *E. coli* strains belonging to the phylogenetic group B2, specifically in the ST73 and ST12, 95 complex strains. This may suggest that *bgl* has a function in this group of *E. coli* under certain conditions. In order to analyze whether the presence, absence or activation of *bgl* provides a selective advantage, growth competition experiments were performed. To this end, wild-type *E. coli* K12, a *bgl*$^+$ derivative, and a *bgl*-deletion derivative were analyzed in pairwise competition experiments. In these experiments, one of the two strains was tagged with an antibiotic resistance and a *lac* deletion as phenotype marker. The *lac* marker and the antibiotic resistance were swapped between the competition pair to check if the marker has a neutral effect on growth. The competition experiment was performed with 5 replicates for three days. Every day the culture was diluted 1:100 and an aliquot of the culture was plated on X-gal indicator plates (Fig.17A). In all pairs, the competing strains grew similarly demonstrating that the presence, absence or activation of *bgl* does not affect the fitness of *E. coli* K12. Since this strain belongs to phylogenetic group A, ST10 complex and carries *bgl* operon of the group Ia, in addition a similar experiment was performed with *E. coli* strain i484, which is a phylogenetic group B2, ST73 complex and *bgl* group II strain. Wild type i484, an activated *bgl*$^+$ mutant of

this strain, as well as a *bgl* deletion mutant were tested in competition experiments in liquid LB cultures. Similar to the results for *E. coli* K12, no growth advantage was seen for the wild-type, the activated *bgl* mutant, and the *bgl* deletion mutant. (Fig. 18). Hence, from these experiments, no *in vitro* growth advantage conferred by *bgl* operon could be observed.

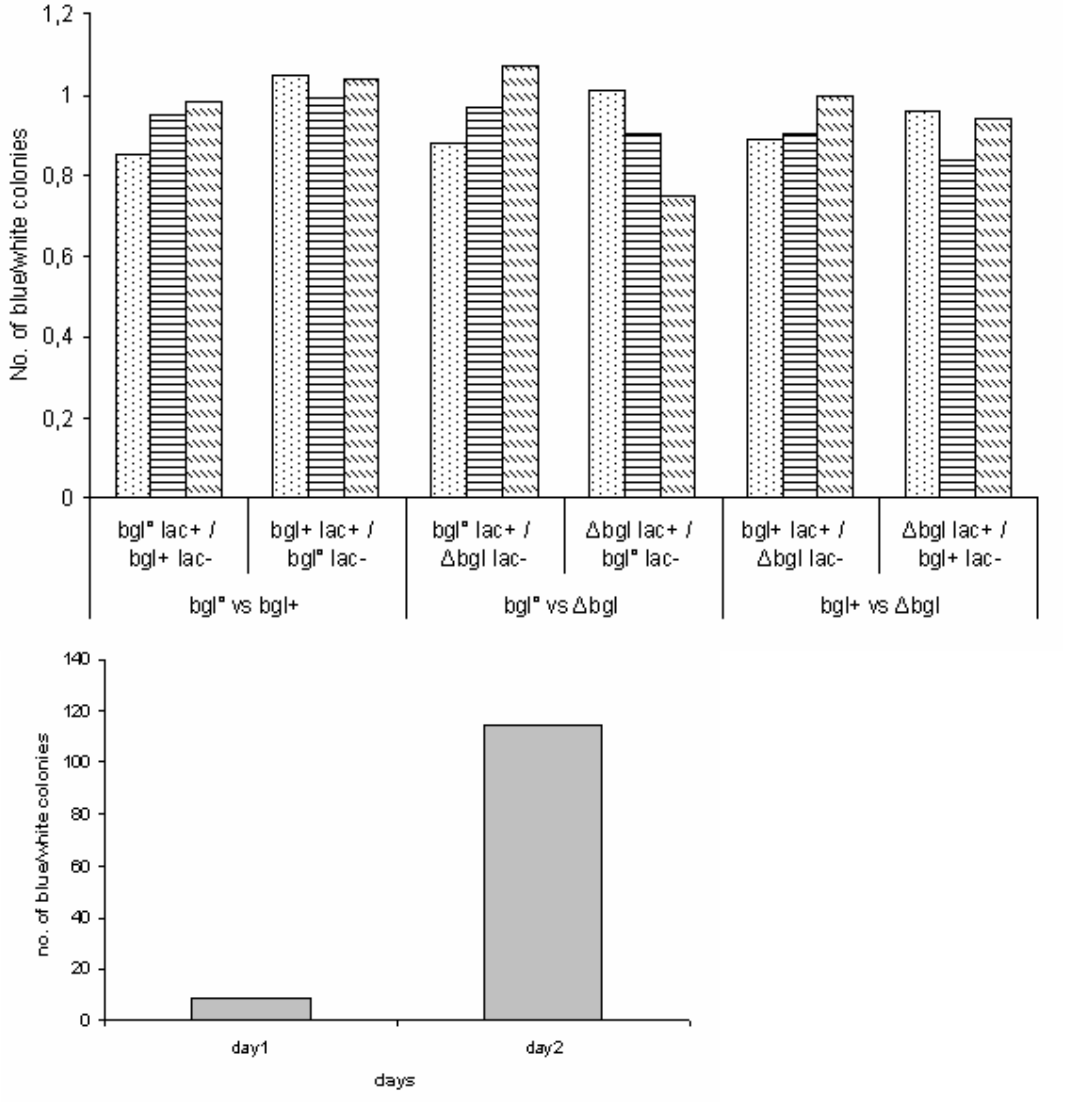

Fig. 17: a) *E. coli* K12 competition growth experiments. Results of the competition growth between wildtype *bgl* strain (*bgl*°) and activated mutant (*bgl*$^{+}$) competed against a deletion mutant (Δ*bgl*). On the X-axis, the competitive pairs are indicated and the relevant strain genotypes are given. Y-axis represents the average of the ratio of blue versus white colonies counted from 5 independent cultures. Bars with dots represent measurements from day 1, dashed lines – day 2 and diagonal lines – day 3. b) Competition growth experiment results for K12 strain (wildtype) competed against its *hns* mutant.

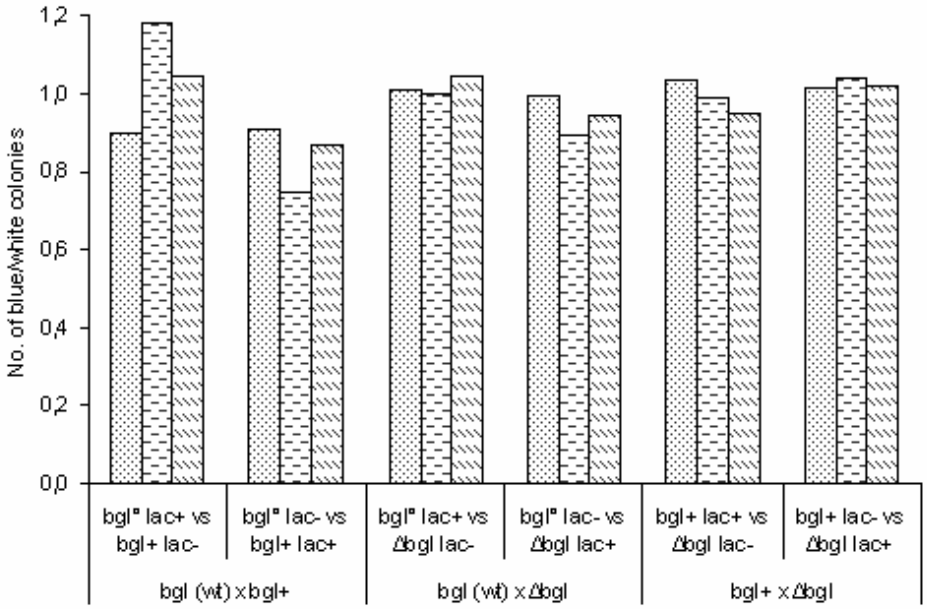

Fig. 18: *E. coli* i484 competition growth experiment. Results of the competition growth between wildtype *bgl* strain and an activated mutant competed against a deletion mutant. X-axis represents the competitive pairs with relevant genotypes. Y-axis represents the average of the ratio of blue versus white colonies counted from 5 independent cultures. Bars with dots represent measurements from day1, dashes - day2, and diagonal lines - day 3.

As a control experiment, the fitness of a pair of wild-type *E. coli* K12 and its isogenic *hns* mutant was also tested. In *Salmonella, hns* mutants have a severely reduced fitness (Navarre et al., 2006). Strikingly, the *E. coli hns* mutant was out competed by the wild-type on day 2 (Fig. 17B) and on day 3 *hns* mutant could not be detected anymore. Since H-NS is a global regulator (Dorman, 2007), the growth incompetence is attributable to a pleiotropic effect of *hns* mutation.

1.11 Sequence evolution of the *bgl* locus

In the attempt to better understand the evolution of the silent *bgl* locus, the type of selection occurring at the locus was analyzed. One of the powerful approaches to detect selection is the K_A/K_S test (Hurst, 2002). This test estimates the number of nonsynonymous (K_A) and synonymous substitutions (K_S) per non-synonymous and synonymous site, respectively. A K_A/K_S ratio above 1 is indicative of positive selection and values below 1 indicate purifying selection, which prevents the accumulation of non-synonymous mutations.

For the *bgl* locus, the average Ka/Ks ratio was determined for 7 individual genes using the genome sequences of 17 *E. coli* and *Shigella* strains (NCBI microbial genomes). Since sequences of strain W3110 and *S. flexneri* 2a 301 were identical to MG1655 and 2457T respectively, they were omitted from the Ka/Ks test. The Ka/Ks test was performed using the Nei-Gojobori method (Nei and Gojobori, 1986) as implemented in the DNAsp package (Rozas et al., 2003). The optimally aligned DNA sequence used previously in the

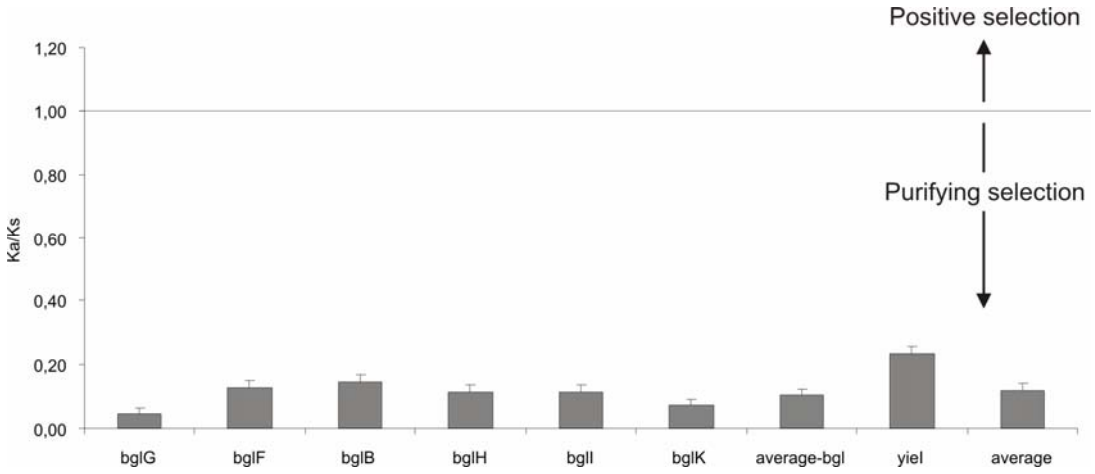

Fig. 19: Ka/Ks ratio for the *bgl* locus. On the X-axis are the individual genes of the *bgl* locus. For each gene the Ka/Ks ratio determined from pairwise comparison and average values are plotted on the Y-axis. The averages for the *bgl* operon genes and the same including *yieI* gene are given.

phylogenetic analysis (section. 1.7) was utilized for the selection test. Average Ka and Ks values and their ratio for each gene were calculated from individual pairwise comparison, and the Ka/Ks ratio for each gene was calculated from these average values. Depending on the gene, the Ka/Ks ratios ranged from the lowest 0.045 for *bgl*G to the highest 0,232 for *yieI*. The average for all the 7 genes was 0.12, which is characteristic of purifying selection (Fig. 19). Thus, purifying selection operates to maintain the operon.

2. Population genetic analysis of the E. coli bgc locus

2.1 The second cryptic β-glucoside operon, *bgc*

The *bgc* (β-glucoside and cellobiose) operon was discovered in the septicemic *E. coli* strain i484 (Neelakanta, 2005). The *bgc* locus consists of six genes, of which five form the *bgcEFIHA* operon and one gene, *bgcR*, maps upstream in divergent orientation. Gene *bgcR* encodes a regulator, while genes *bgcEFIHA* encode the subunits of the permease, a phopho-β-glucoside hydrolase and a putative β-glucoside specific porin. The *bgc* operon resembles an additional cryptic β-glucoside system that can be mutationally activated. Expression of the operon results in β-glucoside utilization at low temperature (28°C) but not at 37°C. The *bgc* operon is present in the sequenced UPEC strains CFT073 (Welch et al., 2002) and UTI89 and absent in the non pathogenic laboratory *E.coli* K12 strain MG1655 and the enterohemorrhagic strain O157 (Neelakanta, 2005). Here it was shown that strain i484, in which the operon was discovered, belongs to the same phylogenetic group B2 and sequence type complex ST73 as the UPEC strains. In this study, the prevalence of this additional β-glucoside system was analyzed and its evolution was traced in *E. coli*. To this end, the strain collection was typed for the presence/absence of the *bgc* operon and these *bgc* types were mapped onto the MS_{TREE} of the collection.

2.2 Typing of the *bgc* operon in the *E. coli*

Most of the strains of the laboratory *E. coli* strain collection were typed previously by PCR for the presence or absence of the *bgc* locus (Neelakanta, 2005). In order, to obtain additional sequence information, a slightly different PCR scheme (Fig. 20) was established for typing of *bgc*, which allowed obtaining sequence information that could also be phylogenetically analyzed. This PCR/sequencing scheme was used to type the representative collection of 175 *E. coli* strains. PCR primers were designed, which map to the flanking genes *marB* and *ydeD* as well as the *bgcR* gene using the published genome sequences of *E. coli* K12 MG1655, uropathogen CFT073 and

enterohemorrhagic 0157 EDL933 strains. In this PCR, strategy amplicons of different sizes are expected for the presence or absence of *bgc* operon, which allowed to type *bgc* in a single PCR reaction. In addition, the PCR products were sequenced on both strands with the primers used for PCR. The typing revealed that 57% (99 of 175 strains) carry the *bgc* operon and that 42% (72 of 175 strains) lack the *bgc* operon. Strain Ecor9 and the three *E. albertii* strains could not be typed because no PCR products were obtained. Possibly, the locus is organized differently in these strains. The typing confirmed that the *bgc* operon is widespread among the *E. coli* population. It is present even among the strains lacking the *bgl* operon and it is absent in the three ancestral *E. coli* strains. In addition, it has been shown before that the structure of *bgc* operon is altered in six strains by insertions and deletions (Neelakanta, 2005). The analysis of the sequence by alignments and neighbor joining trees revealed that at this locus the obtained sequences were of insufficient length for phylogenetic analyses.

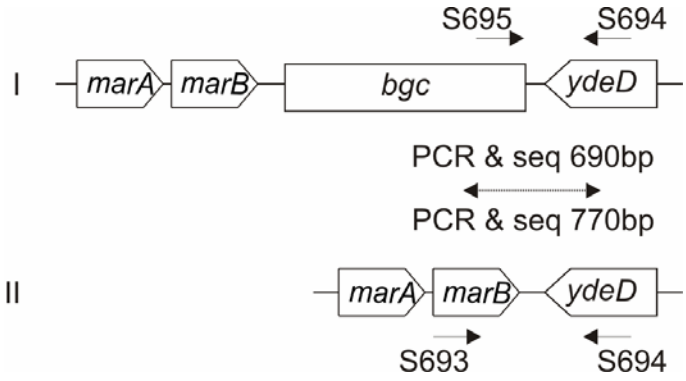

Fig. 20 Schematic showing the PCR and sequencing of the *bgc* loci. Same primers used for both PCR and sequencing are indicated. Multiplex PCR reaction with primers S693, S694 and S695 was performed to distinguish strains with (690bp) and without (770bp) *bgc* operon and the products were sequenced on both the strands. Gene structures are not drawn to scale.

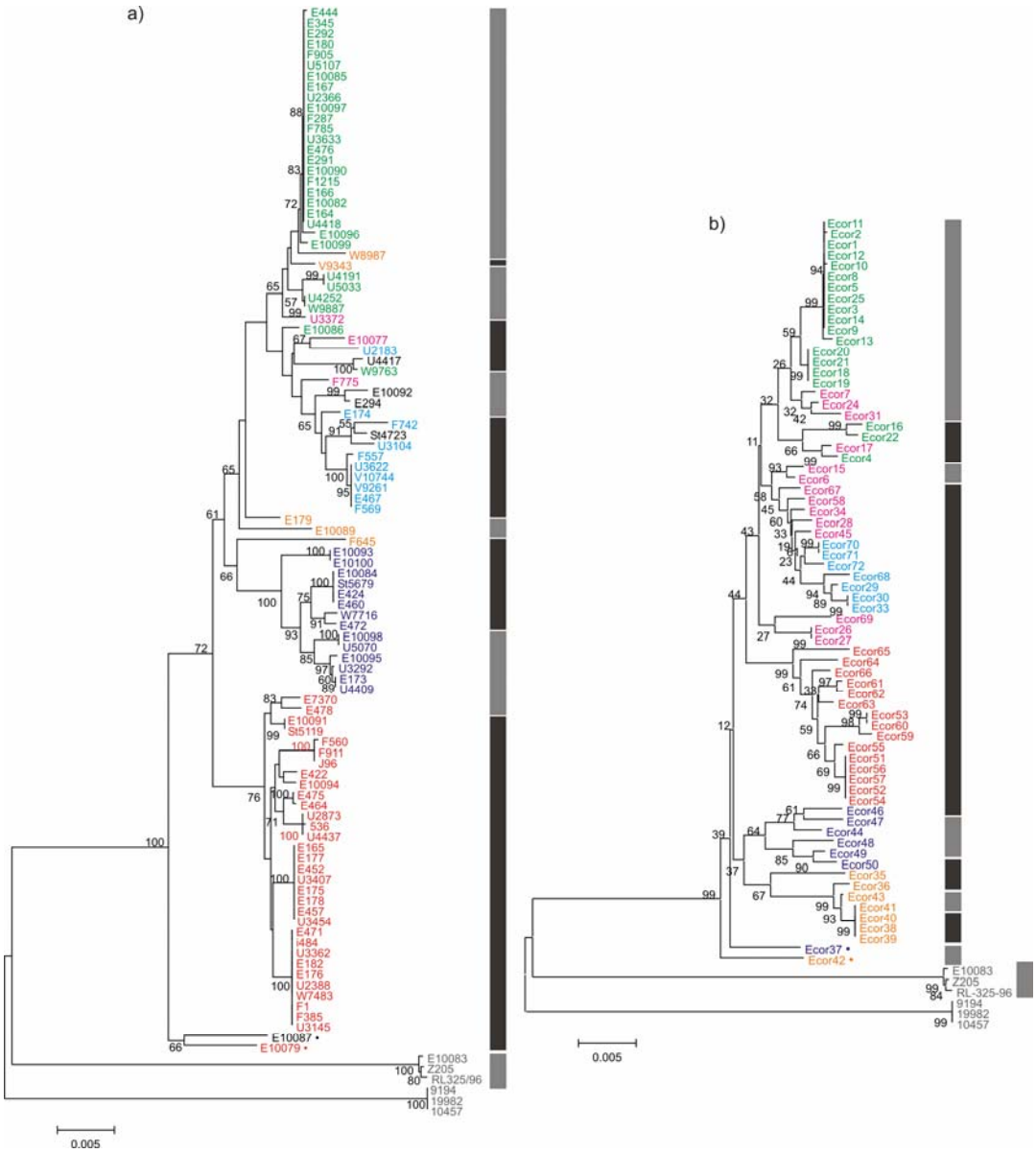

Fig. 21: *bgc* groups are marked on the tree based on concatenation of seven housekeeping gene represented earlier in Fig.5. Groups are indicated by black (I, present) and grey (II, absent) on the right. Strains of different phylogenetic groups are color-coded as in Fig.5: green-A, cyan-B1, red-B2, blue-D, magenta-AxB1, orange-ABD and grey-ancestral/*E. albertii* strains. Strains displayed in black are not assigned to any group and those indicated (•) are considered odd strains.

2.3 Correlation of the prevalence of the *bgc* operon with the *E. coli* phylogeny

To analyze the occurrence of the *bgc* operon with respect to the phylogeny of *E. coli*, the presence and absence, respectively, of *bgc* was associated with the phylogenetic groups of *E. coli* on the neighbor joining tree based on the housekeeping genes, which was presented earlier in section 1.1 (Fig. 21). In addition, the *bgc* types (presence or absence) were mapped onto the MS_{TREE} (Fig. 22). The analysis revealed that all but 4 A group strains lack the *bgc* operon and that all B1 strains possess the *bgc* operon. Further, all but three B2 strains carry *bgc*, and surprisingly 50% of D group strains (all of which lack the *bgl* operon) possess *bgc*. Among the hybrid strains 65% of hybrid AxB1 strains and 50% of ABD strains carried *bgc*, respectively. Hence, *bgc* is not restricted to any particular phylogenetic group. Notably, it is over-represented in the B1 and B2 phylogenetic groups and underrepresented in the A group. Furthermore, to discern the prevalence of *bgc* in the population, the *bgc* groups (presence/absence) were color coded and marked on the MS_{TREE} representing the population structure (section 1.1, Fig. 4). By these means, a more intuitive visualization of the presence/absence polymorphism of *bgc* was achieved. It became evident, that *bgc* is totally absent in the ST10 complex. The *bgc* locus is present in all ST12, 14, 23, 38, 73, 95, and 155 complex strains. It is absent in ST31, 69, 350 and 399 complex strains. All the results are summarized in Table S1 in appendix. Thus prevalence of *bgc* is widespread, which may indicate either horizontal transfer or deletion in different clonal complexes.

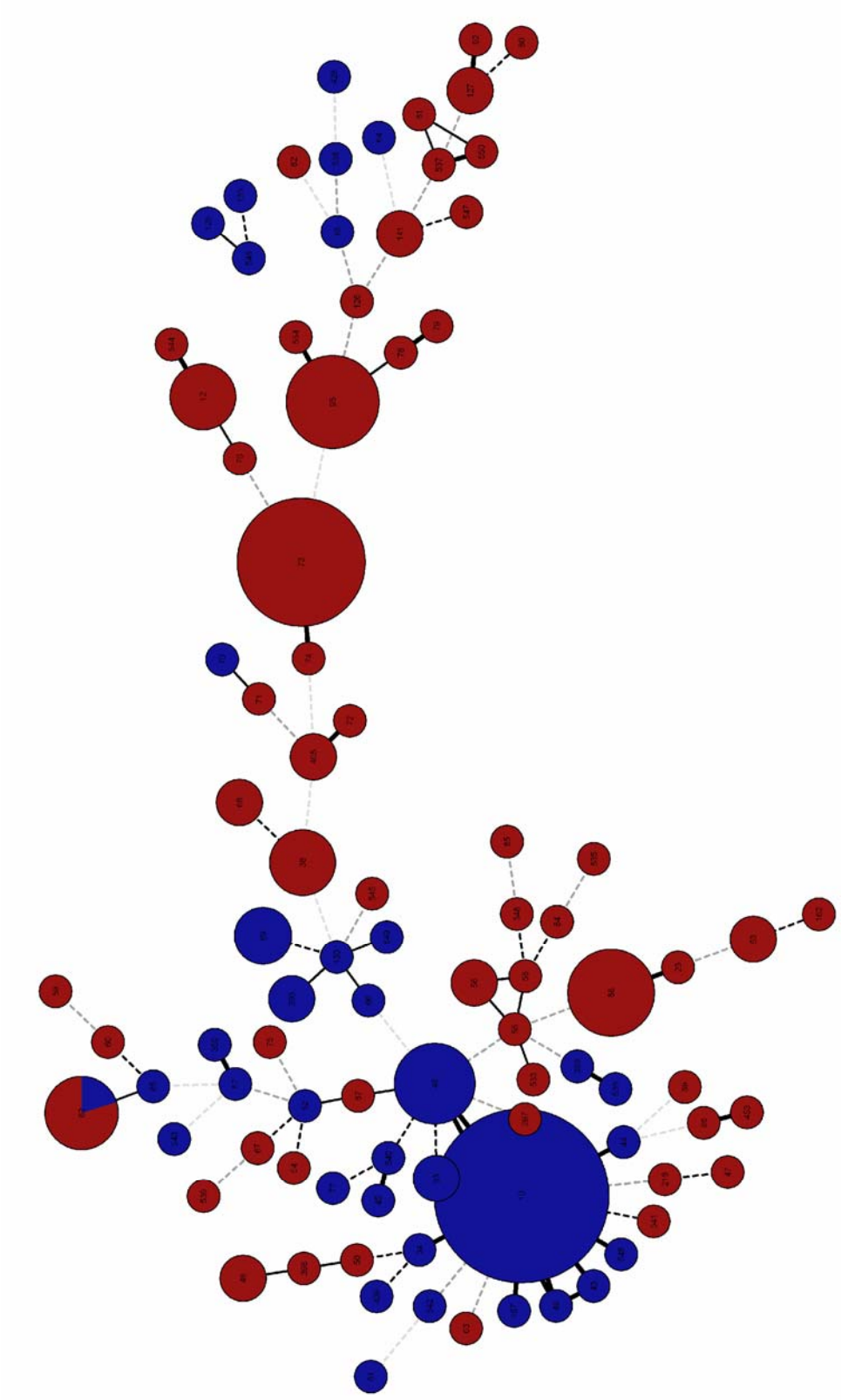

Fig. 22: Groups of *bgc* operon on MS_{TREE} represented in Fig.4. The groups are I (red, *bgc* present) and II (blue, *bgc* absent)

3. Comparative genomics of E. coli

3.1 Evolution of bacterial genomes

Bacterial genomes are remarkably stable and extremely fluid at the same time. Genomes acquire mutations and undergo gene deletions, acquisitions, and rearrangements. They evolve over time, constantly adapting to the environment. *E. coli* is an ecologically very diverse species, which includes both non-pathogens and pathogens, and the *E. coli* genomes are mosaic and diverse owing to the ecological and pathogenic diversity of the species (Welch et al., 2002). Therefore, *E. coli* serves as a fascinating model to study bacterial genome evolution. The analysis of the *bgl* and *bgc* loci in this study provided two examples for variations between *E. coli* strains of different phylogenetic groups, and for *bgl* the results suggests that this variability is based on gene loss. If this type of study were extended to the whole genome level, it would identify events of gene loss and gain in the evolution of modern *E. coli* strains, and indeed comparative genomics can help to identify gene loss and gain events in bacterial genome evolution. DNA array based Comparative Genomic Hybridization (microarray-CGH) has become a powerful method to do whole genome comparisons of bacterial strains (Behr et al., 1999) (Dobrindt et al., 2003a). The test strains are compared to available reference genomes spotted on the arrays and the presence or absence of genes is detected, with the limitation that genes absent in the reference strains, cannot be detected. Thus, complete genome comparisons give a snapshot of the genome content of strains of interest, which facilitates to track the evolution of strains. A microarray-based approach was initiated here by establishing the experimental setup and methodology (Fig. 23).

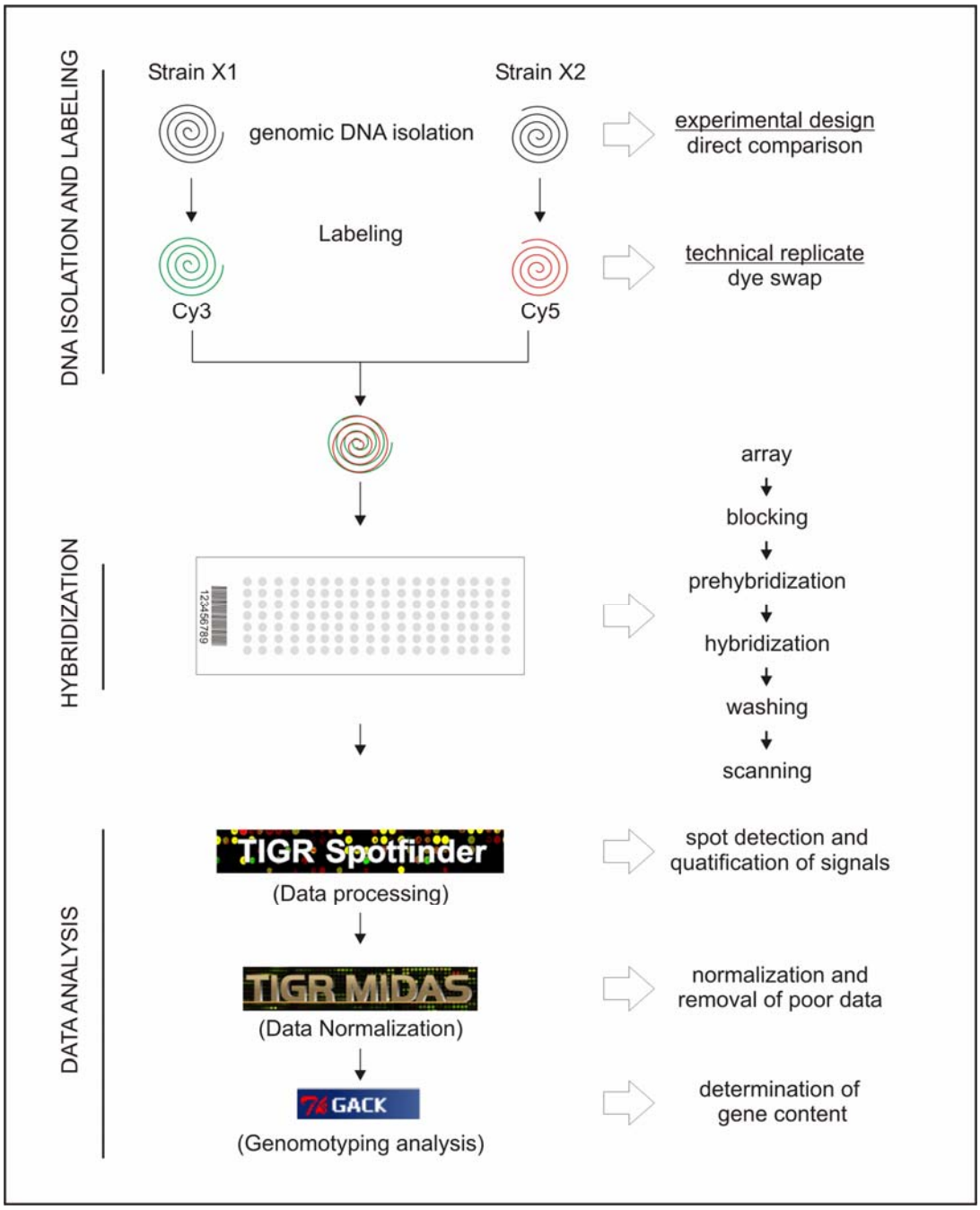

Fig. 23: Overview of the Microarray based Comparative genomic hybridization experiment. Major steps involved in the experiment are depicted and elaborated in the text.

3.2 *E. coli* oligonucleotide microarray

For performing the proposed microarray experiment, a whole genome microarray oligonucleotide set, *E. coli* Genome AROS™ was used (Qiagen Operon, *E. coli* Genome AROS™ Version 2.0). The oligo set represents four sequenced *E. coli* strains and three plasmids. These are the *E. coli* laboratory strains K12 MG1655, enterohemorrhagic strains O157:H7 EDL933, Sakai and uropathogenic strain CFTO73. The plasmids are pO157_EDL933 of O157:47 EDL933 strain and two plasmids pOSAK1, pO157_Sakai of the O157:H7 Sakai strain. The oligo set consists of 70mer probes and the number of ORFs represented by the oligonucleotides corresponding to each genome is given in table 1. In addition, 12 positive and randomized negative control oligonucleotides are included in the array. The oligonucleotides are designed to have a T_m of 75°C within a range of $\pm$ 5°C and were supplied at 600pmol concentration. Each oligo carries an amino linker at the 5' end for efficient attachment to amine reactive glass slides.

Table 1: Oligonucleotide set for microarray analysis of *E. coli*

E. coli strain (No of ORFs)	No. of Oligos	No. of ORFs represented by Oligos corresponding to each genome			
		K12	O157:H7 (EDL933)	O157:H7 (Sakai)	CFT073
K12 MG1655 (4289)	4289	4289	3643	3632	2046
O157:H7 EDL933 (5349)	1416	0	1416	931	0
O157:H7 Sakai (5360)	273	0	0	273	0
CFT073 (5366)	3320	0	0	0	3320
Total	9298	4289	5059	4836	5366

3.3 Microarray fabrication

The *E. coli* genome oligonucleotide array can be designed to one's own needs from the commercial oligonucleotide kit (*E. coli* Genome AROS™ Version 2.0). To manufacture an array, the following criteria are taken into consideration: a) choosing appropriate glass slides for spotting the oligonucleotides; b) printing buffer; c) physical conditions for printing and d) spotting pattern. In order to meet these criteria a sample oligonucleotide set called E.coli AROS Sample set V 1.0 was purchased (Qiagen, Operon). The sample version consisted of 96 70mer probes at 600pmol concentration. For standardizing the fabrication process, this trial oligonucleotide set was utilized. The lyophilized oligonucleotides were resuspended in water to a final concentration of 40 µM and separate aliquots were prepared from resuspended master set for printing. During array fabrication, the following criteria were taken into consideration for desirable results: a) morphology, b) size homogeneity c) signal reproducibility, d) overall intensity, and e) background. Four different buffers and commercial glass slides were used in trial printing to optimize manufacturing of arrays. From the trial experiments, the buffer Sodium phosphate 200mM pH9.0 and Nexterion slide H (Schott GmBH) gave the best results in printing. On to the microtiter plate with aliquots of oligonucleotides, 200mM Sodium Phosphate pH 9.0 was added to achieve a final concentration of 20µM of oligonucleotide and 100mM of the buffer. The Nexterion slide H has a hydrogel coating containing activated functional groups on the glass surface for efficient attachment of amino linked oligonucleotides. Sodium phosphate buffer offers low evaporation, so that coupling can be efficient. Environmental factors such as humidity and temperature play a role in determining the quality of the slides. All the spotting runs were performed in a controlled situation, 18°-22°C and 30%-50% humidity. A 4x4 grid was designed for spots in one sub grid and total, 4x12 grids were spotted. Printed slides were incubated in the arrayer at 75% humidity overnight for effective coupling and drying. Later slides were stored in desiccators until use. The quality of printing was checked by post printing staining of slides with SYBR® green dye (Molecular Probes). The sample

array was then used for optimizing the microarray experiments. For optimization, trial experiments were conducted with the sample array using labeled genomic DNA from *E. coli* K12 MG1655 strain and a Δ*lac* strain CSH50. Since the array consists of oligo for *lac* operon, CSH50 Δ*lac* strain is expected to give no signal. This trial experiment was repeated until hybridization was optimized.

Once the microarray experiment was optimized including sample preparation, labeling with fluorescent dyes and hybridization, whole genome array was fabricated. The protocol for each step in the experiment is described in the Materials and Methods section. The whole genome array was developed in two versions. In version 1.0 the oligonucleotide set was divided into two sets, one containing *E. coli* K12 MG1655 strain specific oligos called K12 array, another called pathoarray containing oligos representing the three pathogenic strains. The version 1.0 was a result of limitation of the spotting robot, which possibly achieved a maximum of 5808 spots on a single slide and hence the total of 9692 oligos were split on to two slides. The spotting pattern was 4x12 grids with 11x11 spots in each grid. With the version 1.0 of the array, it was required to hybridize each strain pair onto two slides, which causes increase in noise of the experiment. Therefore, in addition the whole genome oligonucleotide set was spotted onto a single slide with a high density arrayer. This resulted in *E. coli* microarray slides version 2.0. The spotting pattern of version 2.0 was 4x12 grids with 21x22 spots in each grid. The quality of the array was determined by staining with Cy3 labeled nonamers. The final version has 19384 spots spanning the entire oligonucleotide set for each ORF representing four *E. coli* genome sequences.

3.4 Probe preparation and Hybridization

Genomic DNA was isolated from test strains K12 MG1655 and *E. coli* i484. 500ng of DNA was labeled via an indirect method of labeling genomic DNA. In the indirect method, DNA is labeled with a modified base, aminoallyl dUTP using random primers. Then the modified DNA is chemically coupled to

fluorescent Cy dyes. The individual labeling reactions were purified and pooled before hybridization. Initially the DNA was sheared approximately to a size range of 1-3Kb and was used for labeling. Better labeling was obtainable with unsheared DNA and therefore shearing was avoided. Several concentrations of starting sample DNA were tried and 500ng was found to be satisfactory in labeling. Labeled DNA was run on an agarose gel and scanned in a microarray scanner to check the incorporation of Cy dyes and the same gels were stained with Ethidium bromide to visually approximate the DNA concentration.

The pooled Cy labeled samples were hybridized to the slides. Prior to hybridization, the array was blocked in Ethanolamine buffer to inactivate the reactive surface on the unspotted area and in addition prehybridized in SSC buffer supplemented with SDS and BSA. An important parameter to be taken care during hybridization is the hybridization temperature. Therefore a series of temperature ranging from 42°C to 65°C was tried for standardization. Hybridization at low temperatures yielded higher non-specific hybridization leading to higher false positive rates. On the other hand, hybridization at higher temperature (65°C) gave lower false positive rates, though signal intensity was weak relative to hybridization at low temperatures but quantifiable. In order to reduce the non-specific hybridization several different hybridization buffers were tested. SSC buffer supplemented with formamide and/or Salmon sperm DNA was used. Addition of formamide or salmon DNA did not greatly influence the non specific signals and therefore additives to SSC buffer were omitted. Hybridizations were performed overnight ~16h in a commercial hybridization station (Slidebooster^{TM}, Implen GmbH), which provides a controlled condition for hybridization. In addition, the hybridization station has the facility to keep the probes in circulation flooding the whole slide uniformly, which gives better signals.

Post hybridization washing is one of the most important steps in a microarray experiment. Washing determines the stringency, the signal strength and

influences the background. So it is very crucial to standardize the washing. Stringency is defined by the sum total of the external factors that affect hybridization efficiency, including temperature, salt and pH. Washing was optimized from trials of different stringencies from low to high. A higher stringency washing was required for reducing the false positives due to non-specific hybridization. Initially low stringency washing conditions was (high salt, moderate detergent concentration and low temperature) used during the trial experiments with the version 1.0 array. Later when the same condition was applied on the Version 2.0 array, non-specific hybridization was observed. Therefore, washing conditions were optimized by changing the salt concentration and temperature. After trials, it was found that low salt concentration, detergent concentration and low temperature were ideal in reducing false positives giving good signal intensities. Washing was performed in a very controlled condition using a commercial slide washing station (Advawash^{TM}, Implen GmbH).

3.5 Data acquisition and analysis

The processed slides were scanned with a Genepix scanner (Genomic solutions) on both Cy3 and Cy5 channels and individual TIFF images of 10µM resolution were generated. Raw image files were analyzed using Spotfinder of TM4 microarray software suite (Saeed et al., 2006). Image analysis includes finding the spots, measuring the signal intensity along with background subtraction. Scanned images of both channels (Cy3 and Cy5) are loaded and overlaid for quantification. Manually a grid was constructed to define the area of spots on the scanned images. The Otsu-threshold method (Saeed et al., 2006) implemented in the program, which requires assigning of minimum and maximum spot diameter, was used for spot detection. A quality control (QC) filter with default parameters is applied. Annotation file containing the information pertaining to each spot is loaded before image processing. Then the program uses the grid template to scan the spots and estimates the signal intensity from both Cy3 and Cy5 channels. The exported raw data from processed images consisted of signal intensities of each spot along with the

background values. Bad spots are flagged, given zero value and included in the raw data. Exported data is in the form of tab-delimited text file called tav, which is suitable for data analysis with Midas software of the TM4 suite.

Prior to extracting the real biological information from the data obtained, it is important to process the raw data. Data are normalized using Midas software (Saeed et al., 2006). Normalization helps to compensate for variability between slides and fluorescent dyes, as well as other systematic sources of error, by appropriately adjusting the measured array intensities. Various normalization procedures are available in Midas and locfit (LOWESS) method (Quackenbush, 2002) was employed for analysis. Locfit (LOWESS) normalization function normalizes Cy3 or Cy5 intensities of all spots in the data file by applying LOWESS algorithm and adjusting either Cy3 or Cy5 intensities of each spot by the computed LOWESS factors. Further data filtering refines the dataset by removing poor or questionable data, in addition to data that are considered irrelevant to the analysis. Thus, the normalized data with signal intensities corresponding to each gene in the array are exported as tab-delimited text file.

The final and the most interesting step in data analysis is to find out the genome content of the strains hybridized. Classically presence or absence of genes is determined by comparing a test against a reference or pooled reference, typically the strain(s) whose genome sequences have been used for synthesizing the array. In this study, a different approach was planned to have a simple and direct comparison between two test strains. In the classical approach to detect the presence/absence of genes, a constant ratio value as a cut-off is used. This cut-off is empirically determined from the comparison of the reference strain to the test strain. Using a constant cut-offs may lead to falsely assigning presence of genes (Kim et al., 2002). Moreover constant cut-off demand high levels of reproducibility, which is practically difficult. For the simple experimental design applied in this study, an alternative and more suitable method of assigning presence/absence of genes was planned. A

program called GACK (Kim et al., 2002) was intended to use. The methodology of GACK depends on the shape of the signal-ratio distribution. The cutoff to assign presence/absence is determined on individual hybridization data and hence accounts for array-to-array variation. Moreover, the GACK algorithm assigns a probability score for each gene if present. This score is called estimated probability of presence (EPP). Percentage of EPP is calculated by dividing the expected normal value from the observed value. Theoretically, an EPP of 0% means absent, 100% for presence of genes and in between values are meant to be divergent. Binary and tertiary classification of the data can be done. Thus, this methodology can be used to determine the genome content of strains of interest.

3.6 Trial experiments with *E. coli* K12 MG1655 strain

In order to apply the established experimental methodology, trial experiments were performed with sequenced laboratory strain K12 MG1655, based on which the oligos were designed. A self-hybridization experiment was set up, that is K12 was compared with K12. This hybridization would prove to be a control, since signals are expected from all the oligos representing K12 genome and no signals from the other pathogen specific genes. The experiments were performed on V 1.0 array, which includes K12 and pathoarray separately. In these trial hybridizations was performed under various conditions (temperature 42° to 65°C; different washing protocols) which were described earlier. This allowed testing the efficiency of different protocols. Data analysis was carried out as described. The results are summarized in Fig. 24. Hybridization at low temperatures with low stringent washing resulted in higher percentage of true positives (91%) and at the same time greater false positives (43%). True positives refer to signals expected from the self-hybridization on the K12 array (4481 spots) and 109 spots on the pathoarray. False positives refer to signals, which are not expected from the K12 self-hybridization on 273 and 4846 spots in K12 and pathoarray respectively. The hybridization at 55°C with moderate stringent washing resulted in similar percentage of true positives, but reduced the percent of

false positives (33%). The final hybridization at 65°C with higher stringent washing detected less percentage of true positives (83%) whereas greatly reduced the false positives to 22%. Comparable values were obtained on the pathoarray. The spots that were scored as false positives are due to cross hybridization from K12. The cross hybridization data obtained from Qiagen Operon GmbH, revealed that indeed intergenic region of K12 genome could cross hybridizes with the pathogen specific genes. Hence, in this study the false positives are neglected from data analysis. To conclude, hybridization at 65°C followed by highest stringent washing protocol is best suited for comparative genomic hybridization experiment.

(A)

	low stringent washing 42°C	moderate stringent washing 55°C	high stringent washing 65°C
true positives (%)	91 4079/4481	91 4102/4482	83 3721/4481
false positives (%)	43 118/273	33 90/273	22 61/273

(B)

	low stringent washing 42°C	moderate stringent washing 55°C	high stringent washing 65°C
true positives (%)	63 61/109	59 64/109	48 52/109
false positives (%)	31 1496/4846	26 1290/4846	23 1131/4846

Fig. 24: Results from the standardization experiments. *E. coli* K12 MG1655 strain was self-hybridized under different conditions indicated on the top and the results for hybridization on K12 array (A) and pathoarray (B) are presented. The number of spots scored as present over total number of spots present on the array and its percentage is given.

IV Discussion

In this work, the evolution of two loci of *Escherichia coli* encoding proteins for the utilization of aryl-β-glucosides was analyzed and correlated to the species evolution. The results suggest that the *bgl* operon was vertically inherited by *E. coli*. Homologs are present in other enteric bacteria, while in some enterobacterial species it presumably was lost. The operon was also lost in one of the four phylogenetic groups of *E. coli*, where it is replaced by a cluster of 4 genes (Z5211-5214) of unknown function. Three phylogenetic groups of the *bgl* operon exist in the modern *E. coli* population. Strikingly, these groups are congruent with the phylogenetic groups of the species and perfectly match the clonal population structure, which suggests coupled evolution. Intriguingly, the estimation of K_A/K_S ratio in identifying selection revealed that *bgl* is under purifying selection, implying high functional conservation. Further, the *Bgl* phenotype also correlates with the phylogenetic groups and the clonal population structure of the species. The *bgl* operon is functional but silent in the majority of *E. coli* isolates of the A and B1 phylogenetic groups. In isolates of the B2 phylogenetic group it is intact in almost all isolates, and furthermore is not silent but weakly expressed in approximately 50% of the B2 isolates. This may indicate that *bgl* provides a selective advantage under certain conditions in B2 strains although *bgl* does not affect the fitness of the strains in laboratory growth conditions. The analyses on the second silent β-glucoside system, *bgc*, demonstrated its occurrence in all the phylogenetic groups, with the *bgc* locus being highly prevalent in strains of the B1 and B2 phylogenetic groups and being under-represented in strains of the group A. This diversity could indicate loss or gain of *bgc* during evolution of *E. coli*; it may also be based on intra-species recombination. Taken together, the study points out that in order to understand the role of gain and loss of genes in evolution, it is important to consider phylogeny of genes in a population. In case of the *bgl* operon, such an analysis provides compelling evidence that the *bgl* operon is indeed maintained in *E. coli* for an unknown purpose.

1. Origin and evolution of the *bgl* operon

The phylogenetic analysis suggests that the *bgl* operon is vertically inherited in enterobacteriaceae and that it has been vertically transmitted from a hypothetical ancestor of enterobacteriaceae to *E. coli* (Fig. 14 and 15). This result excludes the possibility that the *bgl* operon was horizontally transferred into *E. coli*. Horizontal transfer of *bgl* could be assumed based on the absence of *bgl* in ancestral *E. coli* and *E. albertii* strains and in the *E. coli* strains of the phylogenetic group D. Furthermore, by applying a Bayesian method on nucleotide composition of the microbial genomes Nakamura et al (2004) predicted *bgl*G gene to be a putative horizontally acquired gene. Based on GC content and codon usage analysis Lawrence and Ochman (1998) predicted that *yieJ*/*yieI* genes present downstream of *bgl* operon as candidates of horizontal transfer in *E. coli*. Hence, the PCR analysis in this study (Fig. 6) together with the computational predictions favors the notion of horizontal transfer of *bgl*. This is also supported by the finding that H-NS silences foreign genes and the fact that it exceptionally represses the *bgl* operon in *E. coli* (Dole et al., 2004b; Lucchini et al., 2006; Nagarajavel et al., 2007; Navarre et al., 2007). Nevertheless, the phylogenetic analysis of *bgl* and its orthologs in Proteobacteria confirm that *bgl* has been a part of the ancestral genome of Enterobactericiae. Therefore, it has to be assumed that during evolution, it was lost in *Salmonella*, *Yersinia*, and in a group of *E. coli*, whereas it was retained in three other groups of *E. coli*, in *Erwinia* sp. and in *Photorhabdus* sp. In the phylogenetic group D of *E. coli*, the *bgl* operon was replaced by the Z5211-5214 locus, which possibly was acquired by horizontal transfer. Replacement by a foreign gene is one of the mechanisms of operon death in bacteria (Price et al., 2006). Briefly, a common ancestral population of *E. coli* presumably had the *bgl* operon and when the species expanded into four lineages A, B1, B2 and D, *bgl* was eliminated in D and retained in the others. PCR and sequence based typing of the ancestral and *E. albertii* strains revealed that these strains have neither *bgl* nor Z5211-5214 locus at the same chromosomal map position (Fig. 6). Assuming vertical inheritance of *bgl*, it

could be expected that *bgl* is present in the three ancestral strains. However, the PCR and sequencing analyses were performed with primers specific for the chromosomal position at which *bgl* maps in the modern population of *E. coli*. Therefore, these analyses do not exclude the presence of *bgl* genes elsewhere in the ancestral genome. *E. albertii* is considered as a distinct species of the *Escherichia* genus, which split from an *E. coli* like ancestor 28 million years ago (Hyma et al., 2005). Hence, it is possible that *bgl* was eliminated upon speciation of *E. albertii*. Further analyses, possibly using Southern or ST-PCR (Semi-random PCR) (Chun et al., 1997) may resolve these interesting questions.

2. Phylogeny and clonality of *bgl*

Three phylogenetic groups of *bgl* (Ia, Ib and II) were defined based on partial sequences of the *bgl* locus, and the prevalence of these groups show a striking congruence with the phylogeny of the species (see results section 1.3 and 1.5; Fig. 8A and 10). In addition, this congruence is strongly supported by phylogenetic reconstruction of the complete operon taken from 17 *E. coli* and *Shigella* genome sequences, and comparison of the *bgl* phylogenetic tree with the tree based on the sequences of the MLST genes (Fig. 12). It is interesting that a locus like *bgl* belonging to the flexible part of genome parallels the phylogeny of the core genome, which is exceptional in *E. coli*. Earlier, in *E. coli* the phylogeny of the locus of the *mutS* gene involved in Methyl directed Mismatch repair (MMR) was analyzed. This locus is known to be polymorphic and part of the flexible genome. Phylogenetic analysis of *mutS* gene in natural isolates of *E. coli* showed higher incongruence with the species phylogeny, indicative of extensive horizontal transfer (Brown et al., 2001; Kotewicz et al., 2003). Additionally, the *fim* operon encoding type1 pili, whose evolution was studied at clonal level, specifically in ST95 clonal complex exhibit less degree of clonality, indicative of frequent horizontal transfer of fimbrial genes and/or intraspecies recombination (Weissman et al., 2006).

Recombination is a well-known phenomenon that influences the evolution of strains and genes (Feil, 2004). Core genes are expected to display clonality in bacterial species without recombination (Maiden, 2006), whereas flexible genes do not necessarily need to show clonality (Gogarten and Townsend, 2005). Surprisingly, *bgl* exhibits strict clonality in *E. coli*. The three groups of *bgl* and the fourth group, where *bgl* is replaced by the Z5211-5214 locus, perfectly match the clonal structure of strains tested and characterized by MLST (Fig. 11). This result is even more striking as previous reports indicated that even housekeeping genes in *E. coli* do not exhibit such a clonal behavior (Wirth et al., 2006). Strict clonality of *bgl* reflects the almost complete absence of recombination. Some recombination events can still be witnessed in the evolution of *bgl*. These are firstly, the elimination of *bgl* in D and ABD strains, in agreement with results that strains of these groups have a recombinogenic nature (Wirth et al., 2006). Secondly, in strains F905 and E179 recombination has obscured the phylogenetic congruence and clonality of *bgl*. Strain F905 is the only strain in A group in ST10 complex, in which the *bgl* operon is replaced by the Z locus. Similarly, strain E179 in the ST350 complex (ABD group) may have lost *bgl* by replacement with the Z locus. Further, recombination can be inferred from the tree based on the core gene *phoU*, which maps next to *bgl*. In this tree, nine strains having the *bgl* operon cluster together with strains possessing the Z5211-5214 locus (Fig. 8B). It might also be that the *phoU* tree is not as reliable as the tree based on seven or more housekeeping genes (Lecointre et al., 1998).

3. String of β-glucoside systems

It is intriguing that all the species in which *bgl* homologs were identified by BLAST carry multiple genes involved in β-glucoside and cellobiose metabolism (Fig. 15). Only in *Erwinia*, the *arb* operon required for metabolism of arbutin and salicin is expressed (An et al., 2004; El et al., 1990; Hong et al., 2006). In *Klebsiella* and *Photorhabdus*, beta-glucosidic genes are annotated as cryptic (Duchaud et al., 2003). Therefore, the functional status of these genes has to be examined in different species.

The second cryptic operon identified and characterized earlier in the laboratory, the *bgc* locus is prevalent in more than 50% of the population. Prevalence of *bgc* is widespread and not strictly confined to any phylogenetic group or clonal complex. However, it is almost absent in the A group, and completely absent in the ST10 complex. This diversity makes it hard to conclude with the limited analysis performed, whether rampant gain or loss characterizes the evolution of *bgc*. Further BLAST surveys as performed for *bgl* combined with phylogenetic analyses will give insights into evolution of *bgc*. It thus appears more inquisitive why several systems are maintained and in a silent state.

4. Function and Selection

Earlier in the laboratory, a weak expression of *bgl* operon was observed in a set of strains and from the current analysis, it turned out that those are B2 strains carrying *bgl* II (Fig. 10 and 13). Khan et al (1998) published that the *bgl* operon is expressed in vivo upon infection of the mouse liver by *E. coli*. The septicemic strain i484 used in that study belongs to the ST73 complex of group *bgl* II and exhibits a weakly positive phenotype on salicin indicator plates. Therefore, it is reasonable to speculate that *bgl* might confer a selective advantage in B2 strains in their niche. Apart from the relaxed phenotype, the majority of strains in the phylogenetic groups A, B1 and B2 displayed a *Bgl* negative but papillating phenotype, indicative of the presence of a functional operon, which can be mutationally activated. In some strains, which are *Bgl*$^-$ and in which *bgl* is not mutationally activated, this phenotype is associated with disruption of the operon due to insertions and deletions, while in other strains of this phenotype, structure of the locus based on PCR and Southern analyses (Neelakanta, 2005) is unchanged. Insertions and deletions as well as point mutation might indicate the conversion of *bgl* genes into pseudogenes, which may lead to erosion (Mira et al., 2001). This is indeed seen in Shigella genomes including *S. sonnei*, *S. dysenteriae* and *S. boydii* in which *bgl* is disrupted. (Yang et al., 2005).

An interesting report by Madan and co workers (2005) showed that activated *bgl* confers a growth advantage in stationary phase (GASP) phenotype when competed against wildtype *bgl* strain in *rpoS* mutant. In the present study, wildtype or activated *bgl* did not confer a competitive growth advantage over its deletion, neither in the laboratory strain K12 nor in the B2 strain i484, which exhibits a weakly positive phenotype (Fig. 16 and 17). Since the environment conducive to *bgl* expression is unknown, these experiments might not reflect the real benefits of maintaining *bgl*.

The weakly positive *Bgl* phenotype in B2 strains prompted to check for any signs of positive selection on *bgl* sequences. However, the Ka/Ks ratio analysis performed rather showed evidence for purifying selection on *bgl* genes (Fig.18). When selection is not strong enough to maintain a function, then genes are lost by accumulation of mutations (Lawrence and Roth, 1996),(Mira et al., 2001). Therefore, one would expect a cryptic locus that is under weak or no selection to be lost in evolution. Recently in *Shigella* it is shown that weaker selection led to accelerated gene loss (Hershberg et al., 2007). In contrast, *bgl* is under purifying selection, which reflects on the significance of its biological function. Thus, there might be a strong unknown ecological reason behind the existence of *bgl* in *E. coli* and probably in other enterobactericiae members as well.

5. Conclusions

A simple model for evolution of the *bgl* operon is postulated in Fig. 25. It is assumed that the silent *bgl* operon is vertically inherited in Enterobacteriaceae, following loss and retention in different species. In *E. coli, bgl* descended clonally in the three lineages A, B1, B2 and got deleted in lineage D, cumulatively reflecting the evolutionary history of the species. The second cryptic system *bgc* is prevalent in the major groups of *E. coli*. Further phylogenetic analysis is required to trace the evolution of *bgc*. The preservation of a functional *bgl* operon and the weakly positive phenotype together with constrained protein evolution argue for a specific purpose in

conserving *bgl*. With respect to genome evolution, *bgl* portrays itself as an exciting model involving gene loss and gain, which are the major driving forces behind genome evolution. Over a decade, importance of gene gain by horizontal transfer has been emphasized much overshadowing the significant role of gene loss. Gene loss has been well appreciated only in the evolution of parasitic bacteria. From this study, it is hypothesized that loss of *bgl* has significance in the evolution of the D lineage and retention of *bgl* in other lineages might have an ecologically equal importance. In conclusion, for a better understanding of the evolution of bacterial genomes, loss and gain of genes should be carefully investigated, especially from a population perspective.

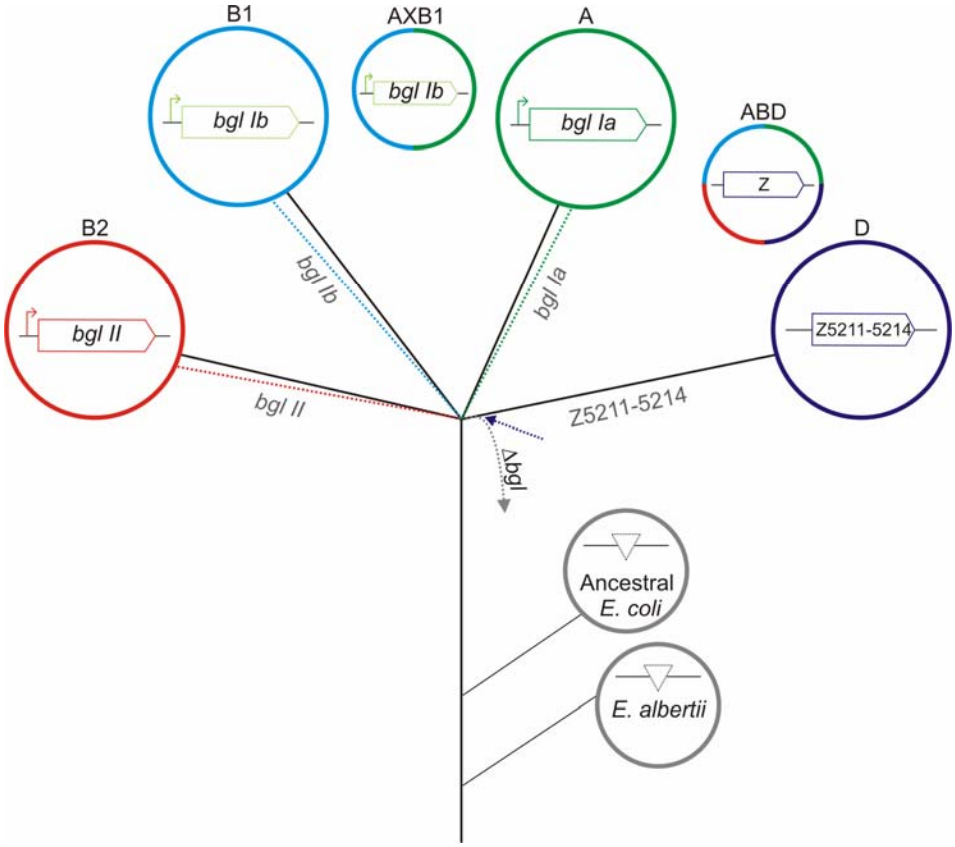

Fig. 25: Evolution of *bgl* and Z loci. *bgl* operon in vertically inherited in enterobactericiae. In *E. coli*, it is maintained in the three phylogenetic groups A, B1 and B2. In the D group, *bgl* is lost by horizontal transfer of Z5211-5214 (Z) locus. Groups of *bgl* locus Ia, Ib and II are concordant with species groups. The phylogenetic groups are represented by circles and the *bgl* or Z locus is indicated within each group. In hybrid groups, AXB1 and ABD *bgl* Ib and Z5211-5214 locus are present respectively. Thus, the evolutionary history of the *bgl* locus argues for its retention in *E. coli*

V Materials and Methods

1. Chemical, enzymes and other materials

Chemicals and enzymes were purchased from commercial sources. Oligonucleotides were purchased from Invitrogen life technologies or Sigma Aldrich.

2. Media and agar plates

LB (1L)
10 gr Bacto Trypton
5 gr Yeast Extract
5 gr NaCl
for plates add 15 gr Bacto Agar

SOB (1L)
20 gr Bacto Tryptone
5g Bacto Yeast Extract
0.5g NaCl
1.25ml 2M KCl
adjust pH to 7.0 with NaOH,
after autoclaving just before use add 10ml 1M $MgCl_2$ per liter

SOC per liter SOB add 19.8ml 20% Glucose.

Bromthymol blue plates (BTB-plates)

15g Bacto Agar
1g Yeast-Extrakt
1g Trypton
5g NaCl
add 900 ml H_2O, autoclave

add sterile:
1 ml 1 M $MgSO_4$
1 ml 0,1 M $CaCl_2$
1 ml Vitamin B1 (stock solution 1mg/ml, filter sterilize)
0,5 ml $FeCl_3$ 1mM
20 ml 10% (w/v) Casaminoacids
50 ml sugar (e.g. 10 % Salicin, 20% Lactose, etc.)
10 ml BTB stock solution (2% bromthymol blue in 50% EtOH, 0,1N NaOH)

Antibiotics if required. The medium should be turquoise, if medium is green add 1N NaOH, if it is blue add 37% HCl

3. Antibiotics

Antibiotic	stock solution	storage temp.	final conc.
ampicillin	50mg/ml in 50 % EtOH	-20°C	50 µg/ml
chloramphenicol	30 mg/ml in Ethanol	-20°C	15 µg/ml
kanamycin	10 mg/ml in H_2O	4°C	25 µg/ml

4. General methods

Molecular biology methods like PCR, other enzymatic reactions were performed as described (Sambrook and Russel, 2001) or according to manufacturer's instructions, unless otherwise stated.

5. *E. coli* and other strains

The *E. coli* K12 strains used in this study is listed in Table.2. 98 clinical *E. coli* isolates comprising of 52 commensals, 22 uropathogenic and 24 septicemic were obtained from Dr. Goerg Plum, Institut für Medizinische Mikrobiologie, Immunologie und Hygiene, Universität zu Köln, Germany. The speticemic *E. coli* strain i484 was obtained from Dr. Richard E. Isaacson, Department of Veterinary Pathobiology, University of Illinois, Urbana, Illinois, USA. The 72 ECOR strains (*E. coli* reference collection) were obtained from STEC (Shiga Toxin-producing *E. coli*) Center, Michigan State University, MI, USA. *E. coli* strains RL325/96 and Z205 were obtained from Dr. Mark Achtman, Max Planck Institute for Infections Biology, Berlin, Germany. *Escherichia albertii* strains were obtained from Dr. Lothar H. Wieler, Department of Veterinary Medicine, Freie University, Berlin, Germany. Strains RL325/96 and Z205 are reported divergent from the rest of *E. coli* isolates representing deepest known evolutionary lineage in this species (Wirth T, 2006). MLST sequence-based phylogenetic analysis showed that strain E10083 differed markedly from the remaining strains, but clustered with RL325/96 and Z205. In the current study,

those 3 strains are referred to as ancestral strains. The *E. coli* i484 strain used in the competition experiments are listed in Table.3.

Table.2 List of *E. coli* K12 strains used in this study

Strain	Relevant genotype[a]	Source/Reference[b]
S527	K12 MG1655, CGSC # 7740	(Blattner et al., 1997)
S541	CSH50 ∆*bgl*-AC11 ∆*lac*Z	(Dole et al., 2004b)
S638	MG1655 (=S527) *bgl*-CAP -66 C→T (=C234)	lab collection
S642	MG1655 (=S527) Δ*bgl*-AC11	lab collection
S3010	S541 Δhns::KD4-KanR	(Nagarajavel et al., 2007)
S3710	S527 *lac*A::KD4kan	PCR fragment (S911/912) from pKD4 transformed into S527/pKD46
S3716	S638 *lac*A::KD4kan	PCR fragment (S915/916) from pACYC184 transformed into S638/pKD46
S3722	S642 *lac*A::KD4kan	PCR fragment (S911/912) from pKD4 transformed into S642/pKD46
S3748	S527 Δ*lac*::KD3cm	PCR fragment (S911/S937 from pKD3) transformed into S527/pKD46.
S3750	S638 Δ*lac*::KD3cm	PCR fragment (S911/S937 from pKD3) transformed into S638/pKD46.
S3752	S642 Δ*lac*::KD3cm	PCR fragment (S911/S937 from pKD3) transformed into S642/pKD46.
S3754	S527 Δhns::KD4kan	xT7 (S3010), selected on LB Kan

a. The relevant genotype of the strains (which are all derivatives of *E. coli* K12 MG1655 strains except S541, a CSH50 derivative) refers to *bgl*, *lac*, and *hns* loci. Mutation resulting in the activation of the silent *bgl* operon includes *bgl*-CRP (a C to T exchange in the CRP binding site at position -66 relative to transcription start)

b. Construction of strains by transduction using T4GT7 is explained in Material and Methods. Δhns::KD4, *lac*A::KD4kan, *lac*A::KD3cm, Δ*lac*::KD4kan, and Δ*lac*::KD3cm refers to replacement of relevant locus on the chromosome by Kanamycin (KD4) or Chloramphenicol (KD3) cassettes constructed according to (Datsenko and Wanner, 2000) and described in detail in Material and Methods.

Table.3 List of *E. coli* i484 strains used in this study

Strain	Relevant genotype[a]	Source/Reference[b]
EC1	*E. coli* i484	(Khan and Isaacson, 1998)
EC2	i484 *bgl*$^{+}$ *bgl*:313-360 (Δ47bp)	lab collection
Ec396	Ec1 Δ*lac*IZYA::KD3-Cm^{R}	PCR fragment (S937/T021 from pKD3) transformed into Ec1/pKD46
Ec398	Ec2 Δ*lac*IZYA::KD3-Cm^{R}	PCR fragment (S937/T021 from pKD3) transformed into Ec2/pKD46
Ec400	Ec1 Δ*bgl*GFBHIK::KD3-Cm^{R}	PCR fragment (T019/T020 from pKD3) transformed into Ec1/pKD46
Ec404	Ec396 Δ*lac*IZYA::FRT	Ec396 transformed with pCP20
Ec406	Ec398 Δ*lac*IZYA::FRT	Ec398 transformed with pCP20
Ec408	Ec400 Δ*bgl*GFBHIK::FRT	Ec400 transformed with pCP20
Ec412	Ec408 Δ*lac*::KD3	PCR fragment (S937/T021 from pKD3) transformed into Ec408/pKD46
Ec416	Ec412 Δ*lac*::FRT	Ec412 transformed with pCP20

a. The relevant genotype of the strains (which are all derivatives of *E. coli* i484 strain) refers to *bgl*, and *lac hns* loci.

b. Construction of Δ*lac*IZYA::KD3 and Δ*bgl*GFBHIK::KD3 refers to replacement of relevant locus on the chromosome by Chloramphenicol (KD3) cassette according to (Datsenko and Wanner, 2000) and described in detail in Material and Methods. Construction of Δ*lac*IZYA::FRT, ΔbglGFBHIK::FRT by transformation with pCP20 expressing the FLP recombinase, and selected on LB Amp at 28°C and then shifted to 42°C for loss of the plasmid and Amp and Cm sensitivity is tested and confirmed by PCR.

6. Multilocus sequence typing (MLST)

Multilocus sequence typing of *E. coli* strains was performed as described (Wirth T, 2006). PCR primers and conditions were adopted from MLST protocol for *E. coli* available at the *E. coli* MLST database (http://web.mpiib-berlin.mpg.de/mlst/dbs/Ecoli). MLST analyses including assigning alleles, STs, ST complexes, generating MS_{TREE} were performed using Bionumerics software (Applied Maths NV). For sequences of *adk* and *fumC* genes, additional internal primers (*adk* - S776, S777; *fumC* - S778) were used.

7. Typing of *bgl* operon/Z5211-5214 locus and *bgc* operon

PCR based typing of the *bgl* operon/Z211-5214 locus was performed with gene specific primers and primers from the flanking genes and the PCR fragments were viszualized on 1% agarose gel. Then they were sequenced on both the strands with the same primers. Ambiguity in sequences was resolved by re-sequencing from fresh PCR fragments. Sequences were manually curated using Contig Express program of Vector NTI Suite (Invitrogen) and Bionumerics software (Applied Maths NV). Similarly, the *bgc* operon was typed using PCR and sequencing with primers from flanking genes and within the operon.

8. DNA sequencing

DNA sequencing was performed using Big dye terminator cycle sequencing kit (version 1.1 and 3.1, Applied Biosystems) according to manufacture's instruction and using an automated DNA sequencer.

9. Phylogenetic analysis

Sequences of the seven housekeeping locus were concatenated into a single sequence for each strain prior to subsequent analysis. Sequences of fragments of *bgl*/Z5211-5214 were separated into core (flanking gene *phoU*) and two locus specific sequences. Locus sequences were merged into one prior to subsequent analysis for each strain. MEGA package V3.1 (http://www.megasoftware.net/) was used for phylogenetic analyses. Sequences were aligned using ClustalW and manually checked for optimal alignment. Phylogenetic trees were constructed using the Neighbor-joining algorithm with default parameters with 1000 bootstrap replicates.

10. BLAST survey

For identifying *bgl* operon orthologs, the sequence from *E. coli* K12 MG1655 strains was used as query. tblastn program of NCBI microbial genomes BLAST were used and searched against all available proteobacterial genomes. Similar BLAST was carried out in Comprehensive microbial

resources of TIGR and results from independent searches were same. Protein sequences of the BLAST hits were downloaded from above two resources and phylogenetically analyzed. 16S rDNA sequences were obtained from Ribosomal Data project hosted by Michigan State University (http://rdp.cme.msu.edu/index.jsp).

11. Gene deletion according to Datsenko and Wanner, 2000

Deletion of *lac* operon was done as described (Datsenko and Wanner, 2000). This method is based on the λ Red based recombination between linear DNA fragment and the chromosomal locus. The strategy is to replace the chromosomal locus with a selectable antibiotic resistance gene that is generated by PCR by using primers with 30-50 bp homology extensions of the locus to be deleted. Briefly, the strains were transformed with temperature sensitive plasmid (pKD46) which expresses the λ Red recombinase under an inducible arabinose promoter. The PCR product for deletion of *lac* operon was generated using primers S911/S937 and plasmid pKD3 or pKD4 as template. This PCR generates a fragment with either chloramphenicol resistance (pKD3) or kanamycin resistance gene (pKD4) flanked by 40bp homology to upstream and downstream sequences of *lac* operon. Then 100ng of gel purified PCR fragment was used to electro-transform strains harboring the helper plasmid (pKD46). The recombinants were selected at 42°C on LB chloramphenicol or kanamycin plates. The loss of helper plasmid was confirmed by sensitivity to ampicillin and deletion of the *lac* operon was confirmed by PCR using primers S920/921. Two independent colonies were stored and used in competition experiments.

12. Electrocompetant cells and electroporation

Cells were grown overnight in 3ml SOB medium with appropriate antibiotics and at appropriate temperature. Of this culture 200µl were inoculated to 50ml of SOB media with appropriate antibiotics and grown to an OD_{600} of 0.7. The culture was transferred to prechilled tubes and centrifuged at 4°C for 15 minutes. The pellet was resuspended in 50 ml of ice-cold H_2O and spun at

4°C for 15 minutes at 3000rpm. The pellet was again resuspended in 25 ml of prechilled H_2O and centrifuged at 4°C for 15 minutes at 3000rpm. Then the cells were resuspended in 2ml of ice-cold 10% glycerol and pelleted by centrifugation (3000 rpm for 15 minutes). Finally, cells were resuspended in 200µl of ice-cold 10% glycerol. The cells were either used immediately for electroporation or for long term storage, further incubated for 1 hour on ice and stored in 40µl aliquots at -80°C. For transformation 40µl of competent cells were mixed with plasmid DNA or a DNA fragment and incubated for 10 minutes on ice. The mix was transferred to prechilled electroporation cuvette (Biorad). The cuvettes were placed in the electroporator and the electric shock was given for 3 seconds at 1.8kV. 1ml of SOC medium was immediately added to the cuvettes. Then the cells were transferred to glass tubes and incubated at 37°C for 1 hour. 100µl of the culture was plated on appropriate selection plates.

13. Transduction with phage T4GT7

T4-Topagar
6g Bacto-Agar (Difco)
10g Bacto-Tryptone (Difco)
8g NaCl
2g Tri-Natriumcitrate-Dihydrate
3g Glucose
add 1l H_2O

The technique is based on generalized transduction, which makes use of the bacteriophage T4*GT7* to transfer DNA between bacteria. Briefly, 100µl of the overnight culture to be transduced was incubated with 10µl, 5µl, and 2µl of T4GT7 lysate prepared from the cells which had the DNA of interest (Donor strain). The incubation was carried out for 20 minutes at room temperature and 100µl was plated on respective selection plates. The tranductants were restreaked at least three-four times to get rid of the contaminating phages and the transfer of the gene was verified by PCR.

14. Microarray-CGH protocols

14.1 Spotting of Oligos on the glass slide

Oligos are spotted on the Nexterion slide H (Schott GmBH). Each slide comes with a bar code for reference and side with the bar code is used for spotting. Before spotting slides are stored at -20°C. Prior to spotting slides should be allowed to equilibirate to room temperature. Handling the slides is very important to obtain strong signals and reduce background. Slides should always be handled wearing gloves holding at the edge and the printed area should be untouched. Oligos are prepared as aliquots in 384-well microtiter plates from the master plates to final concentration of 20µM with the spotting buffer 200mM Sodium Phosphate pH 9.0. Oligos on plates sealed with adhesive covers should be shaken overnight at 4°C in a rotary shaker. Slides are spotted at the Microarray facility, Biochemistry, Uni-Klinik, Universität Zu Köln. Spotting aliquots can be stored at -20°C in between different spotting runs. Finally, after printing, the oligo plates should be dried in the spotting chamber overnight and later stored at -20°C. For future printing the aliquots should be resuspended with ultra pure H_20 (Biochrom AG) with 0.5µL less volume compensating for loss during previous spotting. Spotted slides should be left overnight at 75% humidity for drying and efficient coupling of oligos. Spotted slides should be stored in slides-box in vacuum desiccators until use.

14.2 Quality checking of the spotted slides

SYBR green (Molecular Probes) is used to stain the slides, to check the quality of the spotting. SYBR green is diluted 10,000 fold in TBE buffer (45mM Tris-borate, 1mM EDTA pH 8.0). Flood the spotted slide with the diluted stain and incubate at room temperature for 2-3 minutes. Wash the slides 3-4 times with TBE buffer and spin-dry slides at 1000 rpm for 2 minutes. Scan the slides at 1000x3000resolution in Cy3 channel. To remove the stain off the slide, incubate at room temperature for 1 hour in a solution of 0.1%

SDS, 10mM Tris, 1mM EDTA pH 7.5. Slides are then dried by centrifugation and stored.

14.3 Isolation of genomic DNA from E. coli strains

Genomic DNA was isolated from 2-4mL of overnight cultures in LB medium. For isolation, Presto Spin D universal kit (Molzym GmbH) was used according to manufacture's instructions. The final elution volume was 100µL and quality was checked on 0.7% agarose gel. Isolated DNA was stored at 4°C.

14.4 Labeling of genomic DNA for microarray hybridization

For hybridization on *E. coli* array, genomic DNA from test strains was labeled with Alexa Flour® 555 (green) and 647 (red) dyes (Molecular Probes). 500ng of genomic DNA was used for one labeling reaction. For labeling, BioPrime Plus Array CGH Genomic Labeling System (Invitrogen) was used according to manufacture's instructions. During labeling, light exposure was avoided. The quality of labeling was checked with 2µL of sample on a 1% agarose gel on a microscopic slide. The gels are poured as thin as it can hold 2µL of sample. The gels were scanned in green (Cy3) and red (Cy5) channels at a resolution of 1000x3000. Quantity was estimated by measuring the absorbance at A_{260}. Labeled DNA samples should be away from light, stored in dark-colored tubes or covered with aluminum foils.

14.5 Prehybridization of the array

In order to avoid unspecific binding unspotted areas must be blocked; hence, slides were blocked in 50mM Ethanolamine borate buffer pH9.0 (50mM of Ethonolamine and Boric acid in water) for 1h at room temperature in a glass chamber for slides. Then slides were washed in sterile distilled H_20 3-4 times and dried by centrifugation at 1000 rpm for 5min. Blocked slides should be used immediately (within 2-3h) for hybridization.

Later slides were prehybridized in Prehybridization buffer (3.5X SSC, 0.1% SDS, 10mg/mL BSA) at 65°C for 20min. Slides were washed thoroughly in

sterile distilled H_20 3-4 times and briefly dipped in Isopropanol in a glass chamber for slides. Immediately slides were dried by centrifugation at 1000rpm for 5 min. Prehybridized slides should be used immediately (less than an hour) for hybridization. Slides can be cleaned very briefly using compressed air spray to remove the dust from the surface before hybridization.

14.6 Hybridization

Labeled genomic DNA from two test strains, one with Alexa Flour® 555 (green) and another with Alexa Flour® 647 (red) are pooled in equal amounts and vaccum-dried down completely in Speedvac (Thermo Electron Corporation). Then the probes are resuspended in 60µL of hybridization buffer (4X SSC, 0.1% SDS). Probes are heated to 95°C for 5 min and briefly centrifuged. After labeling the probes, the tubes are kept in dark until applying onto the slides. Probes should be applied on the slides for hybridization within 5-10 minutes. Slides are mounted on the hybridization station (Slidebooster™, Implen GmbH). Before mounting the slides, 50µL of coupling liquid (Implen GmbH) should be applied on the chamber in 3-4 drops. This enables circulation of the probes on the slide during hybridization. Lifter glass cover slip is laid over the slides and probes are added from one end, without creating air bubbles. Probes should be added slowly to avoid air bubbles. In case if air bubbles are formed, they should be removed out of the slide by gently taping the cover slip. 250µL of the Humidifying buffer (Implen GmbH) should be applied on the chamber in the wells at the end of the chamber. This will maintain a humid condition for hybridization. Hybridization was performed at 65°C overnight (~16h).

14.7 Post hybridization washing

After hybridization, slides were washed in a washing station (Advawash™, Implen GmbH). First the slides were washed in a moderately stringent buffer I and then in high stringent buffer II. Each washing step was performed twice for 10min at room temperature. Slides are inserted into the washing chamber

and flooded with the first solution, immediately the cover slips are taken off gently using fine forceps. Not more than four slides can be washed at a time. After washing is done, slides should be dried immediately (with 5 seconds) by centrifugation at 100rpm for 5 minutes. Slides should be exposed in air for longer time before drying.

Washing buffer I: 1X SSC+0.05% SDS

Washing buffer II: 0.06X SSC

14.8 Scanning the array

Hybridized slides are scanned in microarray scanner (Genomic solutions) in both green and red channels at 10μM resolution. First a preview scan is run at 1000x3000 resolution, during which the parameter gain of the scanner can adjusted to get optimal image. In addition, another parameter like Black can be adjusted to get a optimal image of the hybridization. Generally, a good hybridization would not require higher gains and higher gains would lead to saturation of pixels from the spots. Accordingly, gain should be fixed not to get saturated pixel intensities. Once the parameters are fixed, final scan at 10μM resolution can be performed. Individual TIFF files are generated, which are stores and used for data analyses. Slides can be stored in slide box and can be rescanned when necessary. However, the intensity of signals is weakened by time.

14.9 Buffers and reagents for Microarray

Spotting Buffer 2x (200mM Phosphate-Buffer (ph=9,0) /10% Na2SO4) 1L

$Na_2HPO_4*2H_2O$ = 3.56g

$Na_2SO_4*10H_2O$ = 10.0g

adjust pH to 9.0± 0.1 by adding 1N NaoH.

20X SSC 1L

Nacl = 175.3g

Sodium Citrate = 88.2g

adjust to pH 7.0 with Hcl

5X TBE 1L

Tris base = 54g

Boric acid = 27.5g

0.5M EDTA pH 8.0 = 20mL

50mM Ethanolamine borate buffer 1L

Ethanolamine = 1.22g

Boric acid = 3.09g

adjust pH to 9.0± 0.1 by adding 1N NaoH.

VI Appendix

Table.4 List of Oligonucleotides used in this study

Name[a]	Length[b]	Oligo Sequence [c]	Description [d]
S4	19	GGATGGACATTGACGAAGC	*bgl*:13-31
S208	23	CACCACAACATTATTGTTGAGAA	*bgl* +175 to +154 in *E. coli* K12 and CFT073
S352	22	GAGCGGCATAACCTGAATCTGA	IS1-primer
S375	26	GCAGGAAGTCAGATTATGAAATTTGA	E.coliCFT073-C1955-C1960region-6894-6919bp
S401	18	GGGCGTTGCGGAACAAAC	E.coliCFT073-C1955-C1960region-924-941
S402	29	GTATTTGGTTTGCGGTGGCGTAAAGCGGT	E.coliCFT073-C1955-C1960region-7411-7383
S413	25	CCGATCGTTCACCCGAAAGTCACCA	yieH-14579-14555
S429	24	GGCGAAAAACTTGCTGATAATTGT	*bgl*K-13090-13113
S537	23	CAGTGGCTTGGGATGATATTTGA	yieJ: 13854-13876
S548	27	GTCGATTGTGATGATAAAATACGTTCT	Z5211: 2248-2274
S693	25	AACGTGACAACGTCACTGAGGCAAT	marB for MLST
S694	21	AACGGTCAGCATGTGGCGATG	MLST ydeD
S695	27	TGAAATCGCCAGTATTTTACGGATCAG	MLST c1960
S712	20	ATTCTGCTTGGCGCTCCGGG	MLST adk forward
S713	20	CCGTCAACTTTCGCGTATTT	MLST adk reverse
S715	20	GTACGCAGCGAAAAAGATTC	MLST fumC reverse
S716	20	TCGGCGACACGGATGACGGC	MLST gyrB forward
S717	20	GTCCATGTAGGCGTTCAGGG	MLST gyrB reverse
S718	29	ATGGAAAGTAAAGTAGTTGTTCCGGCACA	MLST icd forward
S719	19	GGACGCAGCAGGATCTGTT	MLST icd reverse
S720	32	ATGAAAGTCGCAGTCCTCGGCGCTGCTGGCGG	MLST mdh forward
S721	35	TTAACGAACTCCTGCCCCAGAGCGATATCTTTCTT	MLST mdh reverse
S722	20	TCGGTAACGGTGTTGTGCTG	MLST purA forward
S723	20	CATACGGTAAGCCACGCAGA	MLST purA reverse
S724	21	AGCGTGAAGGTAAAACCTGTG	MLST recA forward
S725	20	ACCTTTGTAGCTGTACCACG	MLST recA reverse
S727	20	TCACAGGTCGCCAGCGCTTC	MLST fumC forward
S733	23	CGGATGTGTGAATTACGCTCCGG	*bgl* downstream seq - intergenic of *bgl*K -yieJ/I of K12 & CFT073
S734	27	CTCCTGAACACAATATTTATTCGCCCG	seq Z5214

S735	34	AAAGCACTATCAAACCAACTGGAACATATAAAAT	seq Z5211
S766	21	CGCATTCGCTTTACCCTGACC	MLST-recA-forward1
S767	22	TCGTCGAAATCTACGGACCGGA	MLST-recA-reverse1
S776	23	GCCTTTCTTGAGGCAATCGCCTG	MLST adk F1
S777	30	CAACTTGTTGATAATTGTAGCGGAAAAGTG	MLST adk R1
S778	26	CAGGTAATGACTGCCAGTTCATCTGC	MLST fumC F1
S782	21	TCGAACCAATCCAGAATATTA	seq Z5211 reverse
S785	20	TACGGACTGCCCGTTGACGG	MLST IS629 forward
S786	25	CCAGGTAATGATTTACAGCGGCAAG	MLST IS1397 forward
S787	23	TCCGGTGCATTTGCAATTAACTG	MLST Z5211 forward
S788	20	GCATCCGGCAATGTGTCCAG	MLST E. albertii phoU-yieJ forward
S789	22	TTCCACGAGCAGACAGGACGTT	E.albertii igs yieJ-yieI - MLST Rev
S911	61	GTTCCTGCGCTTTGTTCATGCCGGATGCGGCTAATGTAGAGTGTAGGCTGGAGCTGCTTCG	*lac* operon Datsenko
S912	70	TACAAGTTCAGCGATCTACATTAGCCGCATCCGGCATGAACATATGAATATCCTCCTTAGTTCCTATTCC	*lac* operon Datsenko
S915	66	TTTGTTCATGCCGGATGCGGCTAATGTAGATCGCTGAACTTTCTCATGTTTGACAGCTTATCATCG	*lac* operon Datsenko
S916	63	AAATTGCCTGATACGCTGCGCTTATCAGGCCTACAAGTTCGATTGGCTCCAATTCTTGGAGTG	*lac* operon Datsenko
S921	26	CGTAGTATCAGCGGCAATTACCTGAT	cynX - to check Datsenko insertions downstream *lac*A
S922	20	TGGCGCGGGTAGTATCGTCA	*lac*A - to check Datsenko insertions downstream *lac*A
S937	70	ATGATAGCGCCCGGAAGAGAGTCAATTCAGGGTGGTGAATCATATGAATATCCTCCTTAGTTCCTATTCC	*lac*I for Datsenko deletion
T-19	71	AGCTCGATAAACTGCTGGCAGAAAAAGATAGCGATAAATAATTCACCAAGGTGTAGGCTGGAGCTGCTTCG	Ec1 *bgl* operon Dasenko
T-20	75	CCCGGATTGGATATTTCATGTCCTGAAACAGACTCTTTAAGCTAACATATGAATATCCTCCTTAGTTCCTATTCC	Ec1 *bgl* operon Datsenko
T-21	59	TTTATGCCGGATGCGGCGTGAACACCTTATCCGGCCTAGTGTAGGCTGGAGCTGCTTCG	Ec1 *lac* operon Datsenko

a. name of the oligonucleotide as in the laboratory collection

b. sequence represented from 5'-3'

c. gene names for which oligonucleotide were designed are indicated. For some oligos the strain names are given. Detailed descriptions are maintained in laboratory records

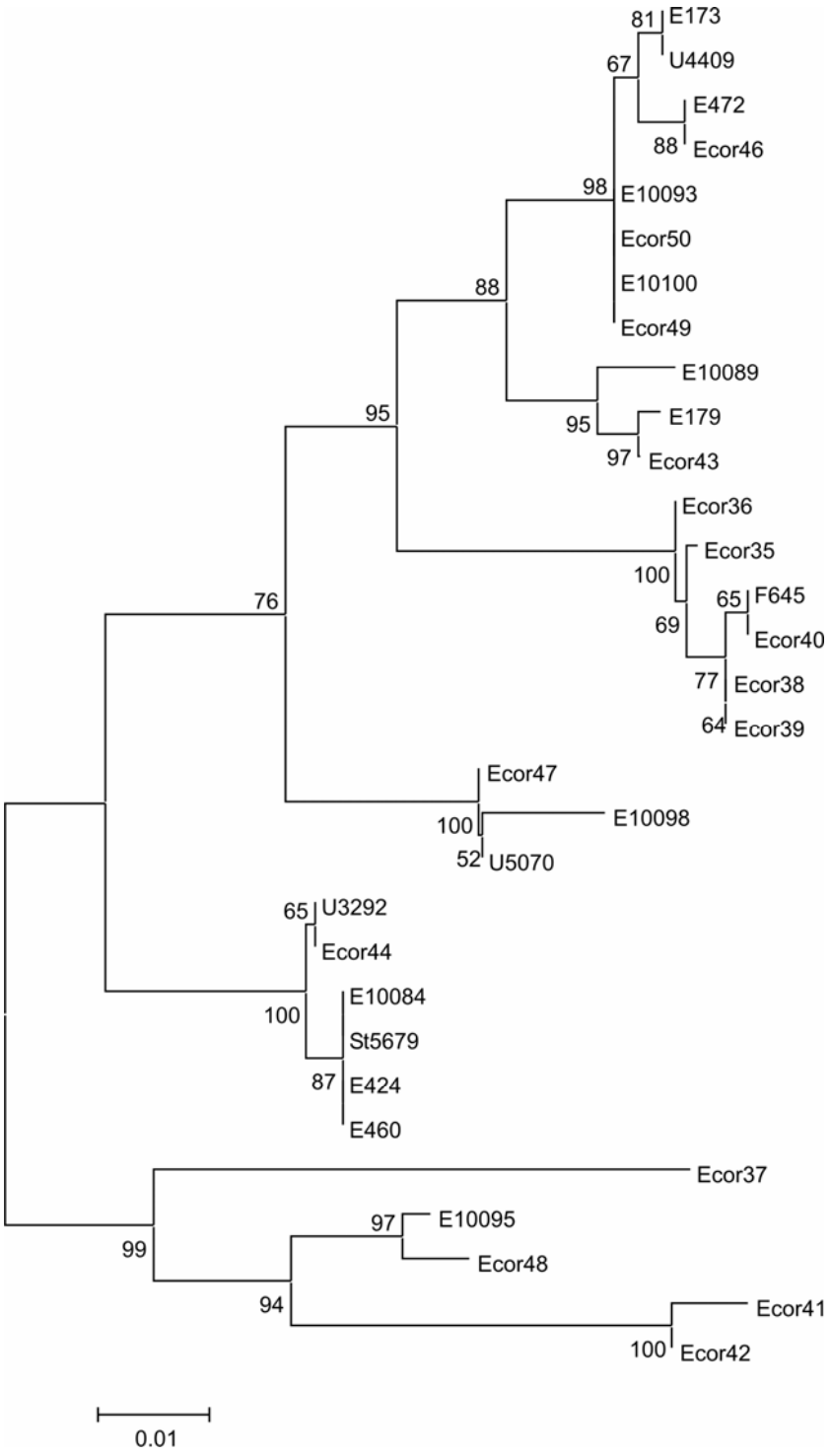

Fig. 26: Neighbor-joining tree of the Z5211-5214 locus obtained by phylogenetically reconstruction of 474bp sequence of the locus. Numbers on the nodes are bootstrap scores from 1000 replicates.

Table S1: List of the *E. coli* and *E. albertii* isolates used in the present study

Ec. No.	Strain Name	Ancestral Group	ST	ST Complex	bgl/Z groups	bgc groups	Phenotype Groups
Ec293	536	B2	92	None	II	I	2
Ec347	*E. albertii* 10457	nil	383	None	V	ND	0
Ec343	*E. albertii* 19982	nil	383	None	V	ND	0
Ec345	*E. albertii* 9194	nil	383	None	V	ND	0
Ec62	E10077	AxB1	162	ST469	Ib	I	0
Ec59	E10079	B2	534	ST95	II	I	2
Ec69	E10082	A	10	ST10	Ia	II	1
Ec67	E10083	nil	546	None	IV	II	0
Ec68	E10084	D	38	ST38	III	I	0
Ec72	E10085	A	10	ST10	Ia	II	1
Ec61	E10086	A	541	ST522	Ib	I	1
Ec60	E10087	ND	535	None	Ib	I	1
Ec56	E10089	ABD	543	None	III	II	0
Ec70	E10090	A	10	ST10	Ia	II	0
Ec73	E10091	B2	141	None	II	I	2
Ec57	E10092	ND	536	ST399	Ib	II	1
Ec55	E10093	D	405	ST405	III	I	0
Ec66	E10094	B2	126	None	II	I	1
Ec71	E10095	D	549	None	III	II	0
Ec63	E10096	A	167	ST10	Ia	II	1
Ec58	E10097	A	10	ST10	Ia	II	1
Ec74	E10098	D	69	ST69	III	II	0
Ec64	E10099	A	34	ST10	Ia	II	1
Ec65	E10100	D	405	ST405	III	I	0
Ec103	E164	A	10	ST10	Ia	II	1
Ec104	E165	B2	95	ST95	II	I	1
Ec106	E166	A	10	ST10	Ia	II	1
Ec105	E167	A	10	ST10	Ia	II	1
Ec123	E173	D	393	ST31	III	II	0
Ec124	E174	B1	348	ST156	Ib	I	1
Ec126	E175	B2	95	ST95	II	I	2
Ec125	E176	B2	73	ST73	II	I	1
Ec119	E177	B2	95	ST95	II	I	1
Ec120	E178	B2	95	ST95	II	I	1
Ec122	E179	ABD	350	ST350	III	II	0

Table S1: List of the *E. coli* and *E. albertii* isolates used in the present study

Ec. No.	Strain Name	Ancestral Group	ST	ST Complex	bgl/Z groups	bgc groups	Phenotype Groups
Ec121	E180	A	10	ST10	Ia	II	1
Ec111	E182	B2	73	ST73	II	I	2
Ec107	E291	A	10	ST10	Ia	II	0
Ec108	E292	A	10	ST10	Ia	II	1
Ec110	E294	ND	399	ST399	Ib	II	1
Ec109	E345	A	10	ST10	Ia	II	1
Ec114	E422	B2	547	None	II	I	1
Ec113	E424	D	38	ST38	III	I	0
Ec115	E444	A	548	ST10	Ia	II	1
Ec116	E452	B2	95	ST95	II	I	2
Ec118	E457	B2	95	ST95	II	I	1
Ec101	E460	D	38	ST38	III	I	0
Ec127	E464	B2	550	ST14	II	I	1
Ec130	E467	B1	88	ST23	Ib	I	1
Ec129	E471	B2	73	ST73	II	I	1
Ec99	E472	D	68	None	III	I	0
Ec100	E475	B2	537	ST14	II	I	0
Ec102	E476	A	10	ST10	Ia	II	1
Ec117	E478	B2	428	None	II	II	1
Ec112	E7370	B2	538	ST538	II	II	2
Ec151	Ecor1	A	10	ST10	Ia	II	0
Ec160	Ecor10	A	43	ST10	Ia	II	1
Ec161	Ecor11	A	10	ST10	Ia	II	1
Ec162	Ecor12	A	10	ST10	Ia	II	1
Ec163	Ecor13	A	44	ST10	Ia	II	1
Ec164	Ecor14	A	10	ST10	Ia	II	0
Ec165	Ecor15	AxB1	45	None	Ib	II	1
Ec166	Ecor16	A	46	ST46	Ib	I	0
Ec167	Ecor17	AxB1	47	None	Ib	I	0
Ec168	Ecor18	A	48	ST10	I	II	0
Ec169	Ecor19	A	48	ST10	Ib	II	1
Ec152	Ecor2	A	49	ST10	Ia	II	1
Ec170	Ecor20	A	48	ST10	Ib	II	0
Ec171	Ecor21	A	48	ST10	Ib	II	0
Ec172	Ecor22	A	50	None	Ib	I	1

Table S1: List of the *E. coli* and *E. albertii* isolates used in the present study

Ec. No.	Strain Name	Ancestral Group	ST	ST Complex	bgl/Z groups	bgc groups	Phenotype Groups
Ec173	Ecor23	ND	73			ND	2
Ec174	Ecor24	AxB1	52	None	Ib	II	0
Ec175	Ecor25	A	10	ST10	Ia	II	1
Ec176	Ecor26	AxB1	53	None	Ib	I	1
Ec177	Ecor27	AxB1	53	None	Ib	I	1
Ec178	Ecor28	AxB1	54	None	Ib	I	1
Ec179	Ecor29	B1	55	ST155	Ib	I	0
Ec153	Ecor3	A	10	ST10	Ia	II	0
Ec180	Ecor30	B1	56	ST155	Ib	I	1
Ec181	Ecor31	AxB1	57	ST350	Ib	II	1
Ec182	Ecor32	ND	73			ND	2
Ec183	Ecor33	B1	56	ST155	Ib	I	1
Ec184	Ecor34	AxB1	58	ST155	Ib	I	1
Ec185	Ecor35	ABD	59	ST59	III	I	0
Ec186	Ecor36	ABD	60	None	III	I	0
Ec187	Ecor37	D	61	ST11	III	II	0
Ec188	Ecor38	ABD	62	None	III	I	0
Ec189	Ecor39	ABD	62	None	III	I	0
Ec154	Ecor4	A	63	None	Ib	I	1
Ec190	Ecor40	ABD	62	None	III	I	0
Ec191	Ecor41	ABD	62	None	III	II	0
Ec192	Ecor42	ABD	64	None	III	II	0
Ec193	Ecor43	ABD	65	None	III	II	0
Ec194	Ecor44	D	66	None	III	II	0
Ec195	Ecor45	AxB1	67	None	Ib	I	1
Ec196	Ecor46	D	68	None	III	I	0
Ec197	Ecor47	D	69	ST69	III	II	0
Ec198	Ecor48	D	70	None	III	II	0
Ec199	Ecor49	D	71	None	III	I	3
Ec155	Ecor5	A	10	ST10	Ia	II	1
Ec200	Ecor50	D	72	ST405	Iv	I	0
Ec201	Ecor51	B2	73	ST73	II	I	2
Ec202	Ecor52	B2	73	ST73	II	I	2
Ec203	Ecor53	B2	12	ST12	II	I	2
Ec204	Ecor54	B2	73	ST73	II	I	2

Table S1: List of the *E. coli* and *E. albertii* isolates used in the present study

Ec. No.	Strain Name	Ancestral Group	ST	ST Complex	bgl/Z groups	bgc groups	Phenotype Groups
Ec205	Ecor55	B2	74	ST73	II	I	2
Ec206	Ecor56	B2	73	ST73	II	I	1
Ec207	Ecor57	B2	73	ST73	II	I	2
Ec208	Ecor58	AxB1	75	None	Ib	I	1
Ec209	Ecor59	B2	76	None	II	I	1
Ec156	Ecor6	AxB1	77	ST206	Ib	II	0
Ec210	Ecor60	B2	12	ST12	II	I	2
Ec211	Ecor61	B2	78	None	II	I	1
Ec212	Ecor62	B2	79	None	II	I	1
Ec213	Ecor63	B2	80	ST568	II	I	2
Ec214	Ecor64	B2	81	ST14	II	I	1
Ec215	Ecor65	B2	82	None	II	I	1
Ec216	Ecor66	B2	83	None	II	II	2
Ec217	Ecor67	AxB1	84	None	Ib	I	1
Ec218	Ecor68	B1	85	None	Ib	I	1
Ec219	Ecor69	AxB1	86	ST86	Ib	I	1
Ec157	Ecor7	AxB1	87	None	Ib	I	1
Ec220	Ecor70	B1	88	ST23	Ib	I	1
Ec221	Ecor71	B1	88	ST23	Ib	I	1
Ec222	Ecor72	B1	89	None	Ib	I	1
Ec158	Ecor8	A	10	ST10	Ia	II	0
Ec159	Ecor9	A	10	ST10	Ib	ND	0
Ec9	F1	B2	73	ST73	II	I	2
Ec21	F1215	A	10	ST10	Ia	II	1
Ec10	F287	A	10	ST10	Ia	II	1
Ec11	F385	B2	73	ST73	II	I	2
Ec12	F557	B1	23	ST23	Ib	I	1
Ec13	F560	B2	544	ST12	II	I	2
Ec14	F569	B1	88	ST23	Ib	I	0
Ec15	F645	ABD	62	None	III	I	0
Ec16	F742	B1	539	None	Ib	I	1
Ec17	F775	AxB1	540	None	Ib	II	1
Ec18	F785	A	10	ST10	Ia	II	1
Ec19	F905	A	10	ST10	III	II	0
Ec20	F911	B2	12	ST12	II	I	0

Table S1: List of the *E. coli* and *E. albertii* isolates used in the present study

Ec. No.	Strain Name	Ancestral Group	ST	ST Complex	bgl/Z groups	bgc groups	Phenotype Groups
Ec1	i484	B2	73	ST73	II	I	2
Ec292	J96	B2	12	ST12	II	I	2
Ec332	RL325/96	nil	133	None	IV	II	0
Ec22	St4723	ND	297	None	Ib	I	0
Ec23	St5119	B2	141	None	II	I	1
Ec24	St5679	D	38	ST38	III	I	0
Ec33	U2183	B1	453	ST86	Ib	I	1
Ec34	U2366	A	10	ST10	Ia	II	0
Ec35	U2388	B2	73	ST73	II	I	2
Ec36	U2873	B2	127	None	II	I	2
Ec37	U3104	B1	533	None	Ib	I	1
Ec38	U3145	B2	73	ST73	II	I	1
Ec41	U3292	D	130	ST31	III	II	0
Ec42	U3362	B2	73	ST73	II	I	2
Ec39	U3372	AxB1	409	None	ND	II	0
Ec40	U3407	B2	95	ST95	II	I	1
Ec43	U3454	B2	95	ST95	II	I	1
Ec45	U3622	B1	88	ST23	Ib	I	1
Ec44	U3633	A	10	ST10	Ia	II	1
Ec46	U4191	A	93	ST168	Ib	II	1
Ec47	U4252	A	48	ST10	Ib	II	1
Ec51	U4409	D	393	ST31	III	II	0
Ec50	U4417	ND	398	ST398	Ib	I	0
Ec48	U4418	A	10	ST10	Ia	II	1
Ec49	U4437	B2	127	None	II	I	2
Ec53	U5033	A	93	ST168	Ib	II	1
Ec54	U5070	D	69	ST69	III	II	0
Ec52	U5107	A	10	ST10	Ia	II	1
Ec27	V10744	B1	88	ST23	Ib	I	1
Ec25	V9261	B1	88	ST23	Ib	I	1
Ec26	V9343	ABD	216	None	Ib	I	0
Ec28	W7483	B2	73	ST73	II	I	2
Ec29	W7716	D	545	None	III	I	0
Ec30	W8987	ABD	542	None	Ib	II	1
Ec31	W9763	A	46	ST46	Ib	I	1

Table S1: List of the *E. coli* and *E. albertii* isolates used in the present study

Ec. No.	Strain Name	Ancestral Group	ST	ST Complex	bgl/Z groups	bgc groups	Phenotype Groups
Ec32	W9887	A	48	ST10	Ib	II	1
Ec336	Z205	nil	125	None	IV	II	0

TableS1. List of *E. coli* and *E. albertii* strains used in this study. Each strain is referred by its name and an Ec.No. as documented in the lab strain collection. The Sequence Type (ST), Sequence Type complex (ST Complex), Ancestral groups as determined from the MLST analyses are depicted. The allele numbers of seven MLST genes are available at the E. coli MLST web server (http://web.mpiib-berlin.mpg.de/mlst/dbs/Ecoli/). The groups of *bgl*/Z (Ia, Ib, II and III) and *bgc* (I, II) loci and the Bgl phenotype groups (0-Bgl^-; 1- Bgl^-, papillating and 3- weak Bgl^+) are listed.

VII References

Altschul,S.F., Gish,W., Miller,W., Myers,E.W., and Lipman,D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403-410.

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.

Behr,M.A., Wilson,M.A., Gill,W.P., Salamon,H., Schoolnik,G.K., Rane,S., and Small,P.M. (1999). Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science *284*, 1520-1523.

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., III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B., and Shao,Y. (1997). The complete genome sequence of Escherichia coli K-12, Science *277*, 1453-1474.

Brown,E.W., LeClerc,J.E., Li,B., Payne,W.L., and Cebula,T.A. (2001). Phylogenetic evidence for horizontal transfer of mutS alleles among naturally occurring Escherichia coli strains
J. Bacteriol. *183*, 1631-1644.

Brzuszkiewicz,E., Bruggemann,H., Liesegang,H., Emmerth,M., Olschlager,T., Nagy,G., Albermann,K., Wagner,C., Buchrieser,C., Emody,L., Gottschalk,G., Hacker,J., and Dobrindt,U. (2006). How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic Escherichia coli strains. Proc. Natl. Acad. Sci. U. S. A *103*, 12879-12884.

Buckles,E.L., Wang,X., Lockatell,C.V., Johnson,D.E., and Donnenberg,M.S. (2006). PhoU enhances the ability of extraintestinal pathogenic Escherichia coli strain CFT073 to colonize the murine urinary tract. Microbiology *152*, 153-160.

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.

Cole,S.T., Eiglmeier,K., Parkhill,J., James,K.D., Thomson,N.R., Wheeler,P.R., Honore,N., Garnier,T., Churcher,C., Harris,D., Mungall,K., Basham,D., Brown,D., Chillingworth,T., Connor,R., Davies,R.M., Devlin,K., Duthoy,S., Feltwell,T., Fraser,A., Hamlin,N., Holroyd,S., Hornsby,T., Jagels,K., Lacroix,C., Maclean,J., Moule,S., Murphy,L., Oliver,K., Quail,M.A., Rajandream,M.A., Rutherford,K.M., Rutter,S., Seeger,K., Simon,S., Simmonds,M., Skelton,J., Squares,R., Squares,S., Stevens,K., Taylor,K., Whitehead,S., Woodward,J.R., and Barrell,B.G. (2001). Massive gene decay in the leprosy bacillus. Nature *409*, 1007-1011.

Datsenko,K.A. and Wanner,B.L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc. Natl. Acad. Sci. U. S. A *97*, 6640-6645.

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.

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

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

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

Dorman,C.J. (2007). H-NS, the genome sentinel. Nat. Rev. Microbiol. *5*, 157-161.

Duchaud,E., Rusniok,C., Frangeul,L., Buchrieser,C., Givaudan,A., Taourit,S., Bocs,S., Boursaux-Eude,C., Chandler,M., Charles,J.F., Dassa,E., Derose,R., Derzelle,S., Freyssinet,G., Gaudriault,S., Medigue,C., Lanois,A., Powell,K., Siguier,P., Vincent,R., Wingate,V., Zouine,M., Glaser,P., Boemare,N., Danchin,A., and Kunst,F. (2003). The genome sequence of the entomopathogenic bacterium Photorhabdus luminescens. Nat. Biotechnol. *21*, 1307-1313.

EA Birge (2006). Bacterial and Bacteriophage Genetics., 4th edition, Springer-Verlag.

El,H.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.

Escobar-Paramo,P., Sabbagh,A., Darlu,P., Pradillon,O., Vaury,C., Denamur,E., and Lecointre,G. (2004). Decreasing the effects of horizontal gene transfer on bacterial phylogeny: the Escherichia coli case study. Mol. Phylogenet. Evol. *30*, 243-250.

Feil,E.J. (2004). Small change: keeping pace with microevolution. Nat. Rev. Microbiol. *2*, 483-495.

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.

Gevers,D., Cohan,F.M., Lawrence,J.G., Spratt,B.G., Coenye,T., Feil,E.J., Stackebrandt,E., Van de,P.Y., Vandamme,P., Thompson,F.L., and Swings,J. (2005). Opinion: Re-evaluating prokaryotic species. Nat. Rev. Microbiol. *3*, 733-739.

Gogarten,J.P. and Townsend,J.P. (2005). Horizontal gene transfer, genome innovation and evolution. Nat. Rev. Microbiol. *3*, 679-687.

Groisman,E.A. and Ochman,H. (1996). Pathogenicity islands: bacterial evolution in quantum leaps. Cell *87*, 791-794.

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.

Hacker,J. and Kaper,J.B. (2000). Pathogenicity islands and the evolution of microbes. Annu. Rev. Microbiol. *54*, 641-679.

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.

Hershberg,R., Tang,H., and Petrov,D.A. (2007). Reduced selection leads to accelerated gene loss in Shigella. Genome Biol. *8*, R164.

Hong,S.Y., An,C.L., Cho,K.M., Lee,S.M., Kim,Y.H., Kim,M.K., Cho,S.J., Lim,Y.P., Kim,H., and Yun,H.D. (2006). Cloning and comparison of third beta-glucoside utilization (bglEFIA) operon

with two operons of Pectobacterium carotovorum subsp. carotovorum LY34. Biosci. Biotechnol. Biochem. *70*, 798-807.

Hurst,L.D. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet. *18*, 486.

Hyma,K.E., Lacher,D.W., Nelson,A.M., Bumbaugh,A.C., Janda,J.M., Strockbine,N.A., Young,V.B., and Whittam,T.S. (2005). Evolutionary genetics of a new pathogenic Escherichia species: Escherichia albertii and related Shigella boydii strains. J. Bacteriol. *187*, 619-628.

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

Kim,C.C., Joyce,E.A., Chan,K., and Falkow,S. (2002). Improved analytical methods for microarray-based genome-composition analysis. Genome Biol. *3*, RESEARCH0065.

Koonin,E.V. (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat. Rev. Microbiol. *1*, 127-136.

Kotewicz,M.L., Brown,E.W., Eugene,L.J., and Cebula,T.A. (2003). Genomic variability among enteric pathogens: the case of the mutS-rpoS intergenic region. Trends Microbiol. *11*, 2-6.

Kumar,S., Tamura,K., and Nei,M. (2004). MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief. Bioinform. *5*, 150-163.

Kunin,V. and Ouzounis,C.A. (2003). The balance of driving forces during genome evolution in prokaryotes. Genome Res. *13*, 1589-1594.

Lawrence,J.G. (2005). Common themes in the genome strategies of pathogens. Curr. Opin. Genet. Dev. *15*, 584-588.

Lawrence,J.G. and Hendrickson,H. (2003). Lateral gene transfer: when will adolescence end? Mol. Microbiol. *50*, 739-749.

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,G.T., Clermont,O., Gouriou,S., Picard,B., Nassif,X., Denamur,E., and Tenaillon,O. (2007). Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group Escherichia coli strains. Mol. Biol. Evol. *24*, 2373-2384.

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.

Lopilato,J. and Wright,A. (1990). Mechanisms of activation of the cryptic *bgl* operon of *Escherichia coli* K-12. In The bacterial chromosome, K.Drlica and M.Riley, eds. (Washington, D.C.: American Society for Microbiology), pp. 435-444.

Lucchini,S., Rowley,G., Goldberg,M.D., Hurd,D., Harrison,M., and Hinton,J.C. (2006). H-NS mediates the silencing of laterally acquired genes in bacteria. PLoS. Pathog. *2*, e81.

Madan,R., Kolter,R., and Mahadevan,S. (2005). Mutations that activate the silent bgl operon of Escherichia coli confer a growth advantage in stationary phase. J. Bacteriol. *187*, 7912-7917.

Mahadevan,S. and Wright,A. (1987). A bacterial gene involved in transcription antitermination: regulation at a rho-independent terminator in the bgl operon of E. coli. Cell *50*, 485-494.

Maiden,M.C. (2006). Multilocus sequence typing of bacteria. Annu. Rev. Microbiol. *60*, 561-588.

Marri,P.R., Hao,W., and Golding,G.B. (2006). Gene gain and gene loss in streptococcus: is it driven by habitat?. Mol. Biol. Evol. *23*, 2379-2391.

Maurelli,A.T. (2007). Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens. FEMS Microbiol. Lett. *267*, 1-8.

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.

McDonald,J.H. and Kreitman,M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature *351*, 652-654.

Medini,D., Donati,C., Tettelin,H., Masignani,V., and Rappuoli,R. (2005). The microbial pan-genome. Curr. Opin. Genet. Dev. *15*, 589-594.

Mira,A., Ochman,H., and Moran,N.A. (2001). Deletional bias and the evolution of bacterial genomes. Trends Genet. *17*, 589-596.

Nagarajavel,V., Madhusudan,S., Dole,S., Rahmouni,A.R., and Schnetz,K. (2007). Repression by binding of H-NS within the transcription unit. J. Biol. Chem. *282*, 23622-23630.

Navarre,W.W., McClelland,M., Libby,S.J., and Fang,F.C. (2007). Silencing of xenogeneic DNA by H-NS-facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA. Genes Dev. *21*, 1456-1471.

Neelakanta, Girish. Genome variations in commensal and pathogenic *E.coli*. 2005.
Ref Type: Thesis/Dissertation

Nei,M. and Gojobori,T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. *3*, 418-426.

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.

Parkhill,J., Sebaihia,M., Preston,A., Murphy,L.D., Thomson,N., Harris,D.E., Holden,M.T., Churcher,C.M., Bentley,S.D., Mungall,K.L., Cerdeno-Tarraga,A.M., Temple,L., James,K., Harris,B., Quail,M.A., Achtman,M., Atkin,R., Baker,S., Basham,D., Bason,N., Cherevach,I., Chillingworth,T., Collins,M., Cronin,A., Davis,P., Doggett,J., Feltwell,T., Goble,A., Hamlin,N., Hauser,H., Holroyd,S., Jagels,K., Leather,S., Moule,S., Norberczak,H., O'Neil,S., Ormond,D., Price,C., Rabbinowitsch,E., Rutter,S., Sanders,M., Saunders,D., Seeger,K., Sharp,S., Simmonds,M., Skelton,J., Squares,R., Squares,S., Stevens,K., Unwin,L., Whitehead,S., Barrell,B.G., and Maskell,D.J. (2003a). Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. Nat. Genet. *35*, 32-40.

Perna,N.T., Plunkett,G., III, Burland,V., Mau,B., Glasner,J.D., Rose,D.J., Mayhew,G.F., Evans,P.S., Gregor,J., Kirkpatrick,H.A., Posfai,G., Hackett,J., Klink,S., Boutin,A., Shao,Y., Miller,L., Grotbeck,E.J., Davis,N.W., Lim,A., Dimalanta,E.T., Potamousis,K.D., Apodaca,J.,

Anantharaman,T.S., Lin,J., Yen,G., Schwartz,D.C., Welch,R.A., and Blattner,F.R. (2001). Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature *409*, 529-533.

Price,M.N., Arkin,A.P., and Alm,E.J. (2006). The life-cycle of operons. PLoS. Genet. *2*, e96.

Quackenbush,J. (2002). Microarray data normalization and transformation. Nat. Genet. *32 Suppl*, 496-501.

Reid,S.D., Herbelin,C.J., Bumbaugh,A.C., Selander,R.K., and Whittam,T.S. (2000). Parallel evolution of virulence in pathogenic Escherichia coli. Nature *406*, 64-67.

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.

Rosenberg NA (2004). DISTRUCT: a program for the graphical display of population structure. Molecular Ecology Notes *4*, 137-138.

Rozas,J., Sanchez-DelBarrio,J.C., Messeguer,X., and Rozas,R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. *19*, 2496-2497.

Saeed,A.I., Bhagabati,N.K., Braisted,J.C., Liang,W., Sharov,V., Howe,E.A., Li,J., Thiagarajan,M., White,J.A., and Quackenbush,J. (2006). TM4 microarray software suite. Methods Enzymol. *411*, 134-193.

Schaefler,S. and Maas,W.K. (1967). Inducible System for the Utilization of β-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.

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

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.

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.

van Ham,R.C., Kamerbeek,J., Palacios,C., Rausell,C., Abascal,F., Bastolla,U., Fernandez,J.M., Jimenez,L., Postigo,M., Silva,F.J., Tamames,J., Viguera,E., Latorre,A., Valencia,A., Moran,F., and Moya,A. (2003). Reductive genome evolution in Buchnera aphidicola. Proc. Natl. Acad. Sci. U. S. A *100*, 581-586.

Wei,J., Goldberg,M.B., Burland,V., Venkatesan,M.M., Deng,W., Fournier,G., Mayhew,G.F., Plunkett,G., III, Rose,D.J., Darling,A., Mau,B., Perna,N.T., Payne,S.M., Runyen-Janecky,L.J., Zhou,S., Schwartz,D.C., and Blattner,F.R. (2003). Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T. Infect. Immun. *71*, 2775-2786.

Weissman,S.J., Chattopadhyay,S., Aprikian,P., Obata-Yasuoka,M., Yarova-Yarovaya,Y., Stapleton,A., Ba-Thein,W., Dykhuizen,D., Johnson,J.R., and Sokurenko,E.V. (2006). Clonal analysis reveals high rate of structural mutations in fimbrial adhesins of extraintestinal pathogenic Escherichia coli. Mol. Microbiol. *59*, 975-988.

Welch,R.A., Burland,V., Plunkett,G., III, Redford,P., Roesch,P., Rasko,D., Buckles,E.L., Liou,S.R., Boutin,A., Hackett,J., Stroud,D., Mayhew,G.F., Rose,D.J., Zhou,S., Schwartz,D.C., Perna,N.T., Mobley,H.L., Donnenberg,M.S., and Blattner,F.R. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc. Natl. Acad. Sci. U. S. A *99*, 17020-17024.

Winfield,M.D. and Groisman,E.A. (2004). Phenotypic differences between Salmonella and Escherichia coli resulting from the disparate regulation of homologous genes. Proc. Natl. Acad. Sci. U. S. A *101*, 17162-17167.

Wirth,T., Falush,D., Lan,R., Colles,F., Mensa,P., Wieler,L.H., Karch,H., Reeves,P.R., Maiden,M.C., Ochman,H., and Achtman,M. (2006). Sex and virulence in Escherichia coli: an evolutionary perspective. Mol. Microbiol. *60*, 1136-1151.

Yang,F., Yang,J., Zhang,X., Chen,L., Jiang,Y., Yan,Y., Tang,X., Wang,J., Xiong,Z., Dong,J., Xue,Y., Zhu,Y., Xu,X., Sun,L., Chen,S., Nie,H., Peng,J., Xu,J., Wang,Y., Yuan,Z., Wen,Y., Yao,Z., Shen,Y., Qiang,B., Hou,Y., Yu,J., and Jin,Q. (2005). Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. Nucleic Acids Res. *33*, 6445-6458.

Zhaxybayeva,O., Nesbo,C.L., and Doolittle,W.F. (2007). Systematic overestimation of gene gain through false diagnosis of gene absence. Genome Biol. *8*, 402.

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 Prof. Dr. Karin Schnetz betreut worden.

Köln
Dezember 2007 Sabari Sankar Thirupathy

Lebenslauf

Name:	Sabari Sankar Thirupathy
Geburtsdatum:	02.12.1979
Geburtsort:	Srivilliputhur, Tamilnadu, Indien
Staatsangehörigkeit:	Indisch
1994-1995	Gymanasium (Matriculation), Tamilnadu, Indien
1996-1997	Abitur (Higher Secondary), Tamilnadu, Indien
1998-2001	Bachelor of Science (B. Sc), Madurai Kamaraj University, Indien
2001-2003	(Diplom) Master of Science (M. Sc), Madurai Kamaraj University, Indien
2004-2008	Doktorarbeit bei Prof. Dr. Karin Schnetz am Institüt für Genetik, Universität zu köln Title: The evolution of silent β-glucoside systems in *Escherichia coli*

Köln,
Dezember 2007 Unterschrift

Curriculum vitae

Name:	Sabari Sankar Thirupathy
Date of Birth:	02.12.1979
Place of Brith:	Srivilliputhur, Tamilnadu, India
Nationality:	Indian

1994-1995	Matriculation, Tamilnadu, India
1996-1997	Higher Secondary, Tamilnadu, India
1998-2001	Bachelor of Science (B. Sc), Madurai Kamaraj University, India
2001-2003	Master of Science (M. Sc), Madurai Kamaraj University, India
2004-2008	Doctoral studies under the guidance of Prof. Dr. Karin Schnetz Institute for Genetics University of Cologne Title: The evolution of silent β-glucoside systems in *Escherichia coli*

Cologne
December 2007 Signature


FOR NO REASON
WHAT'S THAT GUY DOING ON THE MOUNTAIN OVER THERE?
LET'S GO ASK HIM.
ARE YOU STANDING HERE WAITING FOR A FRIEND?
NO.
THEN YOU'RE HERE TO BREATHE THE FRESH AIR, RIGHT?
NO.
ARE YOU HERE TO TAKE IN THE BEAUTIFUL SCENERY?
NO.
THEN WHY ARE YOU STANDING HERE?
I'M JUST HERE, STANDING.
BGL