Detection and characterisation of integrons, gene cassettes and cassette-located antibiotic resistance genes in the human oral metagenome

This thesis is submitted by

Md. Ajijur Rahman, MPharm

For the degree of

Doctor of Philosophy (PhD)

In the

Faculty of Medical Sciences

University College London

Department of Microbial Diseases

UCL Eastman Dental Institute

256 Gray’s Inn Road

London WC1X8LD, UK

-2017-
Declaration

I, Md. Ajijur Rahman confirm that the work presented in this thesis is my own.

Where information has been derived from other sources, I confirm that this has been indicated in the thesis.
Abstract

Antibiotic resistance is a serious threat to public health. Horizontal gene transfer (HGT) plays a critical role in the dissemination of antibiotic resistance, however, knowledge about the source of the resistance determinants and their mobile vectors is still limited. Integrons are natural gene capture and expression systems and are known for their role in dissemination of antibiotic resistance genes (ARGs) in clinically important pathogens. The human oral cavity is a reservoir of ARGs many of which were found on plasmids and transposons. However, the association of these ARGs with integrons have not been investigated before.

In this study, a PCR-based metagenomic approach was used to investigate the presence of integrons carrying ARGs in the oral cavity of healthy human individuals from the UK (n=11) and Bangladesh (n=10). PCR primers targeting the mobile integrons (class 1, 2 and 3) as well as chromosomal integrons of Treponema were used to amplify integrons and associated array of gene cassettes (GCs) in the oral metagenome.

By analysing the libraries of PCR amplicons, a large pool of GCs including the cassettes located at the first position of an integron were identified and most of them were found to be novel. The cassettes were predicted to carry open-reading frames encoding proteins of a diverse range of functions including antibiotic resistance, competence, plasmid stability and adaptation to stress.

Two novel variants of D-alanine-D-alanine ligase (ddl) were identified and located in the first position within an integron for the first time. It was found that expression of ddlS increase the resistance of the surrogate host (Escherichia coli) to D-cycloserine (D-cycloserine), an antibiotic used to treat multi-drug-
resistant (MDR) and extensively drug-resistant (XDR) tuberculosis. A SNP at c.777 of the *ddl* variants was found to be responsible for alteration of minimum inhibitory concentration (MIC) of D-cycloserine. Cloning and sequencing of upstream sequence showed that the putative host of the Ddl encoding integron is likely to be a strain of *Treponema denticola*, one of the causative agents of periodontitis. The Ddl proteins encoded by the cassette genes were expressed and purified and their specific ligase activity was confirmed. The predicted 3D structures were also determined using I-TASSER tools and the generated 3D models were used to test the hypothesis that plant-based flavonoids could play an important role in the evolution *ddls* as integron GCs. It was found that the flavonoids namely quercetin and apigenin can bind to both ATP and D-alanine binding sites of the Ddls.

This study shows that a PCR-based metagenomic approach can recover novel GCs including functional genes that confer resistance to clinically important antibiotics. The evidence for HGT of integron GCs and a hypothesis for evolution of *ddls* within integrons is also presented.
Drug-resistant microbes kill 25,000 people each year in Europe alone, and an estimated number of 700,000 around the world. To tackle the problem of antibiotic resistance, more research is needed to identify the source of the antibiotic resistance genes (ARGs) and their mobile vectors as well as the drivers that select for ARGs in bacteria. The results presented in this thesis have provided some useful insights into the role of oral bacteria as a carrier of mobile ARGs as well as the factors that might contribute to selecting the resistance genes in the oral cavity. I have shown that oral cavity carries ARGs present within integrons. I have reported the discovery of two novel variants of D-alanine-D-alanine ligase genes (ddl) (named as ddl6 and ddl7) that are located within first gene cassette (GC) of a reverse integron and form a separate phylogenetic clade. The expression of ddl6 and ddl7 were shown to confer resistance to D-cycloserine, an antibiotic currently used for the treatment of MDR and XDR-TB. I have also shown for the first time that a single base pair mutation in ddl can alter the levels of D-cycloserine resistance conferred by these ligases.

I have also developed a hypothesis which could explain why ddl genes have evolved as integron GC and acquired and maintained within first position of the integron. The hypothesis states that dietary flavonoids have Ddl inhibitory activity and could act as a driver for selecting ddls within integrons. By analysing the similarity of the predicted 3D structures of GC-encoded Ddls with the vancomycin resistant proteins, we developed a hypothesis that a single mutation in omega-loop of integron-encoded Ddls could confer vancomycin resistance.
Some of the results of this thesis have been published in two scholarly journal articles which have been downloaded and viewed by more than a thousand people around the world. In addition, I have presented, discussed and shared my results in 10 microbiology conferences and focused meetings on AMR held in the UK and abroad. This PhD project also led to at least two cross-disciplinary collaborations outside UCL.

The work presented in the thesis will have implications in the field of antimicrobial resistance, discovery of inhibitors targeting Ddl as well as understanding the evolutionary and environmental drivers of antimicrobial resistance.
Acknowledgements

First and foremost I would like to offer my thanks and sincere gratitude to my supervisors, Dr. Adam Roberts and Prof Peter Mullany for their guidance, suggestions and teachings throughout my PhD work. Their thoughtful comments and criticisms were very useful.

I am also grateful to Dr. Frank Kaiser for teaching me the protein techniques including the use of FPLC to purify the proteins. A lot of thanks to Dr. K. Miraz Rahman and Dr. Jamshidi Shirin of KCL for their input on molecular docking experiment. A lot of thanks to Dr. Haitham Hussain for his technical assistance and support. Thanks a lot to Dr. Morgana Vianna for providing me the 16S profile of the metagenomic DNA samples. A big thank you to Prof Anwar Ul Islam for his help to collect the saliva samples from Bangladesh and sending them to the UK. Special thanks to Dr. Andrew Smith for allowing me to use the CLARIOstar microplate reader and to Erni Marlina teaching me how to use the CLARIOstar.

I am also grateful to all of the members of Dr. Robert’s group at UCL including Liam Reynolds, Nur Asyura Binti Nor Amdan, Zeina Al Kudmani, Deena Al Harbi, Supanan Sucharit, Sarah Al Harbi, Norashirene Mohamad Jamil, Enas Newire, Marta Monteiro, Agata Antepowicz and Supathep Tansirichaiya. The group helped me a lot for professional development and collaboration.

Thanks a lot to British Society of Antimicrobial Chemotherapy (BSAC), American Society for Microbiology (ASM), Microbiology Society and Federation of European Microbiological Societies (FEMS) for their travel grants and awards including the FEMS-ASM Makela-Cassel Award. Their support helped me to
attend and present my research at the different microbiology meetings and conferences held in the UK, Europe and the USA.

Thanks to the UCL Doctoral School for their very useful professional development trainings. Thanks to the UCL Eastman Dental Institute and the Department of Microbial Diseases for allowing me to study here and providing necessary support to carry out my studies. Also thanks to my colleagues at the Department of Pharmacy of Rajshahi University for their support and the authority of Rajshahi University for granting me the leave for my PhD study.

I would like to offer my special gratitude to the Commonwealth Scholarship Commission in the UK (CSC) and the UK Government for the Commonwealth Scholarship Award. Through this award the CSC has paid all of my tuition fees at UCL and given me stipends to support the living cost for me and my family. The CSC has also provided me with necessary guidance and trainings for professional development and excellence. Also thanks to the University Grants Commission of Bangladesh for nominating me for this award.

Lastly, I would like to thank my wife, Fahmi Ferdous Moni for her love, care, tolerance and sacrifice. Thanks to my lovely daughter Afeefa Rahman Shiza for her sacrifice. Thanks to my father, Md. Shamsul Haque and mother, Begum Shamsun Nahar and my elder brothers and their families for their love and support. Thanks a lot to my friends, colleagues and their families here in the UK including Mr. Mohammad Jahirul Hoque Shakil and his wife Mrs. Tajmina Sultana Shipu, Dr. Mukhlesur Rahman, Dr. Proma Khondkar, Mr. Rahat M Ali, Dr. Maruf Ahmed and his wife Mrs. Mubina Afroz for their care, love and support for me and my family during my PhD journey.
Table of contents

Declaration ..................................................................................................................2
Abstract .......................................................................................................................3
Impact Statement .......................................................................................................5
Acknowledgements ....................................................................................................7

1 Chapter One: Introduction ......................................................................................37

1.1 Antibiotic resistance: A global crisis .................................................................38
1.2 Genetic basis of antibiotic resistance .................................................................38
  1.2.1 Intrinsic resistance .......................................................................................39
  1.2.2 Acquired resistance .....................................................................................39

1.3 Integrons .............................................................................................................42
  1.3.1 Structure of integrons ..................................................................................43
  1.3.2 Gene cassettes (GCs) ..................................................................................44
  1.3.3 attC: the cassette associated recombination sites .......................................45

1.4 Classification of integrons ..................................................................................46
1.5 Variants and locations of Pc promoters .............................................................51
1.6 Expression of intI gene: Role of the SOS response and cAMP-CRP complex ....53
1.7 Expression of cassette-located genes and its control by cAMP-CRP complex ....55
1.8 Recombination reactions in integron-gene cassette systems .............................58
1.9 Bacteria carrying integrons ...............................................................................60
  1.9.1 Class 1 integron carrying bacteria ...............................................................61
1.9.2 Class 2 integron carrying bacteria..................................................64
1.9.3 Class 3 integron carrying bacteria..................................................66
1.9.4 Chromosomal integron (CI) carrying bacteria...............................67
1.10 Detection of integrons......................................................................69
  1.10.1 Sequence-based............................................................................70
  1.10.2 PCR-based....................................................................................70
  1.10.3 Capturing of functional GCs by IntI-mediated recombination ....75
1.11 Investigation of ARGs using metagenomics.................................76
1.12 Microbial communities in the oral cavity.......................................78
1.13 ARGs in the cultivable oral bacteria................................................81
1.14 ARGs in the oral cavity detected by different metagenomic approaches
  83
1.15 Integrons in the oral cavity...............................................................84
1.16 Aims of this study.............................................................................88

2 Chapter Two: Materials and Methods..................................................89
  2.1 Introduction.......................................................................................90
  2.2 Source of chemicals, reagents and antibiotics.................................90
  2.3 Bacterial strains and plasmids used in this study..............................90
  2.4 Ethical Approval...............................................................................93
  2.5 Collection and storage of saliva samples.........................................94
  2.6 Extraction of metagenomic DNA from saliva...................................95
  2.7 Purification of genomic DNA from bacterial culture........................97
  2.8 Synthesis of oligonucleotides used in this study............................98
2.9 General PCR protocol ................................................................. 101
2.10 Agarose gel electrophoresis ...................................................... 102
2.11 PCR purification ................................................................. 103
2.12 Extraction of DNA from agarose gel ........................................ 104
2.13 Cloning of PCR products .......................................................... 105
  2.13.1 Cloning into pGEM-T Easy .................................................. 105
  2.13.2 Cloning into pET-28a(+) vector .......................................... 106
2.14 Preparation of E. coli competent cells ...................................... 109
2.15 Preparation of Bacillus subtilis competent cells ......................... 109
2.16 Transformation of ligation reactions into E. coli α-Select .......... 110
2.17 Transformation of E. coli BL21 (DE3) ....................................... 110
2.18 Transformation of B. subtilis .................................................... 111
2.19 Colony PCR ................................................................. 111
2.20 Plasmid DNA isolation ........................................................... 112
2.21 DNA sequencing and analysis .................................................. 112
2.22 Agar dilution method for determination of minimum inhibitory concentrations (MIC) .................................................. 114
2.23 Preparation of sample for dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) .................................................. 114
2.24 Casting a discontinuous SDS-PAGE Gel ................................... 115
2.25 16S metagenomic profiling .................................................... 116
2.26 Submission of DNA sequences to GenBank ................................ 117
3 Chapter Three: Detection of integrons and associated gene cassettes in the human oral metagenome .................................................. 118
  3.1 Introduction ............................................................................. 119
3.2 Materials and Methods

3.2.1 Design of primers to target the integron gene cassettes

3.2.2 PCR amplification of GCs and cloning into pGEM-T Easy

3.2.3 Analysis of the sequences for the features of GCs

3.2.4 Nomenclature of the gene cassettes

3.2.5 Submission of the sequences to GenBank

3.3 Results

3.3.1 High molecular weight metagenomic DNAs were isolated from the saliva samples collected from UK and Bangladesh

3.3.2 16S metagenome profiling of pooled saliva samples of the UK and Bangladesh shows that they are truly metagenomic

3.3.3 Class 1 and class 2 integrons could not be recovered in the salivary metagenomic DNA of UK and Bangladeshi samples

3.3.4 PCR products having all the structural features of an integron were recovered from both UK and Bangladeshi samples

3.3.5 The partial intl sequences associated with attI and first GCs were related to different Treponema sp.

3.3.6 The cassette promoters, Pc were found to be located upstream of first GCs and within the attI sites

3.3.7 The core sites of attC were identified at the 3'-end of first GCs

3.3.8 The first GCs encoding proteins of different types of functions were recovered
3.3.9 GCs with different size and orientations were amplified using $attC$-based primers

3.3.10 The core sites of $attC$ (R' and R'') were identified on the GCs

3.3.11 Diversity of the functions of putative proteins encoded by ORFs within the GCs detected by novel $attC$ primers

3.4 Discussion

3.5 Conclusion

4 Chapter 4: Genetic features and resistance phenotypes of integron-located $ddls$

4.1 Introduction

4.1.1 The role of Ddls in peptidoglycan synthesis in bacteria

4.1.2 Mechanism of action of D-cycloserine

4.1.3 Mechanisms of D-cycloserine resistance

4.2 Materials and methods

4.2.1 Determination of minimum inhibitory concentration (MIC)

4.2.2 Construction of $ddl6$ mutants by site-directed mutagenesis

4.2.3 Cloning of $ddl6$ and $ddl7$ into pHCMC05 and pET-28a

4.2.4 Determination of the upstream region of integron carrying $ddl6$ and $ddl7$

4.2.5 Determination of functionality of putative gene cassette promoter, $Pc$ located upstream of $ddl$ cassettes

4.2.6 Growth curve analysis of $B. subtilis$ overexpressing $ddl7$
4.2.7 PCR to detect *ddl* in individual samples of oral metagenomic DNA

173

4.2.8 Isolation of the putative host of the integron carrying *ddl6* and *ddl7*

173

4.2.9 Tools and software used for bioinformatics analysis .................. 174

4.3 Results .................................................................................. 175

4.3.1 Genetic features of integron-located *ddl* genes were studied...... 175

4.3.2 The second-copy of *ddl* in *T. pedis* B683 is the closest homologue of *ddl6* and *ddl7* ................................................................................. 179

4.3.3 The *ddl*(GC) of *T. pedis* B683, the only homologue of *ddl6* and *ddl7*, is likely to be acquired by HGT ........................................ 184

4.3.4 Integron located *ddls* are phylogenetically diverged from house-keeping *ddls* of *Treponema* ........................................................................ 185

4.3.5 The putative active sites of Ddl6 and Ddl7 are conserved........... 187

4.3.6 Expression of *ddl6* and *ddl7* located within the 2,024 bp insert in pGEM-T Easy confers a different level of resistance to D-cycloserine .... 189

4.3.7 Site-directed mutagenesis confirmed that c.777 G>T substitution at *ddl6* is responsible for the alteration of MIC of D-cycloserine ............ 191

4.3.8 Expression of *ddls* does not confer cross-resistance to other antibiotics targeting cell-wall biosynthesis ................................... 193

4.3.9 The putative *Pc* promoter is located within the *attl* site and can express downstream GCs...................................................... 195
4.3.10 The upstream region of integron carrying *ddl7* was found to be similar to *T. denticola*.................................................................197

4.3.11 The *intI* of the integron carrying *ddl7* is closely associated with *intI* of different strains of *T. denticola*.........................................................202

4.3.12 Growth curve of *B. subtilis* 168 overexpressing *ddl6* or *ddl7*......203

4.3.13 The GCs encoding *ddl* are present in almost all individual samples of the UK and BD.................................................................204

4.3.14 The original host of the integron carrying *ddl* could not be recovered from the oral cavity.................................................................206

4.4 Discussion ......................................................................................................207

4.5 Conclusion.....................................................................................................211

5 Chapter Five: Functional characterization of Ddls encoded by integron gene cassettes .........................................................................................213

5.1 Introduction ....................................................................................................214

5.2 Materials and Methods................................................................................216

5.2.1 Chemicals and reagents...........................................................................216

5.2.2 Preparation of the reagents and buffers.................................................217

5.2.2.1 *Reagents for EnzCheck phosphate assay kit*.....................................217

5.2.3 Optimization of time for maximum expression of Ddl .........................217

5.2.4 Western blotting .......................................................................................218

5.2.4.3 *Electrophoretic transfer* ....................................................................218

5.2.5 Optimization of purification.....................................................................219
5.2.6 Preparation and cell lysates and purification of Ddl6, Ddl7, DdlEc and DdlTd using FPLC system

5.2.6.1 Preparation of cleared E. coli lysates under native conditions

5.2.7 FPLC to purify 6x His-tagged proteins

5.2.8 Buffer exchange to remove imidazole from the protein solutions

5.2.9 Determination of protein concentration

5.2.10 Enzymatic activity assay

5.2.10.1 Detection of formation of D-ala-D-ala dipeptide by paper chromatography

5.2.11 Inhibition of Ddl activities by D-cycloserine

5.2.12 Standard curve for Pi

5.2.13 Measurement of \( K_m \), \( V_{\text{max}} \) and \( K_{\text{cat}} \) of Ddl6 and Ddl7, DdlAEc and DdlTd

5.2.14 Statistical Analysis

5.3 Results

5.3.1 Optimization of time for expression and the purification process

5.3.2 Purification of Ddl6, Ddl7, DdlAEc and DdlTd using FPLC

5.3.3 The integron-encoded Ddl7 catalyse the formation of D-ala-D-ala dipeptide, not D-ala-D-ser or D-ala-D-lac

5.3.4 Kinetic properties of Ddl6, Ddl7, DdlAEc and DdlTd
5.3.5 The instability index, aliphatic index (AI) and grand average of hydropathicity (GRAVY) of DdI6, Ddi7 were compared with Ddls encoded by house-keeping genes ................................................................. 239

5.4 Discussion ......................................................................................................................... 242

5.5 Conclusion ....................................................................................................................... 247

6 Chapter Six: Determination of predicted 3D structures of Ddls and molecular docking ........................................................................................................................................ 249

6.1 Introduction ..................................................................................................................... 250

6.2 Materials and methods .................................................................................................... 258
  6.2.1 Prediction of 3D structures of Ddi6, Ddi7 and DdiTd .................................................. 258
  6.2.2 Molecular docking ..................................................................................................... 259

6.3 Results ............................................................................................................................ 260
  6.3.1 Predicted 3D structures of Ddi6 and Ddi7 obtained using I-TASSER could be superimposed with each other ................................................................. 260
  6.3.2 The putative omega-loop of Ddi6 and Ddi7 are located within the C-terminal domain of the proteins and contain two important residues for catalytic activity and substrate specificity ................................................................. 262
  6.3.3 Structural homologues of Ddi6 and Ddi7 as identified by I-TASEER include vancomycin resistance proteins ................................................................. 264
  6.3.4 The residues at 259 position of Ddi6 (Trp) and Ddi7 (Cys) are not part of putative ligand binding sites of Ddls, but are located very close to the omega loop .............................................................................................................. 267
Molecular docking experiments showed that W259C mutation in Ddl6 does not alter the binding affinity of D-alanine and D-cycloserine at the D-alanine binding site.  

Quercetin and apigenin bind to D-ala binding sites with higher affinity than the ATP-binding sites.

Discussion

Conclusion

Chapter Seven: Final conclusion and future directions

References

Appendix I: Multiple sequence alignment of empty gene cassettes detected in this study

Appendix II: Translated sequence of ddl6

Appendix III: Translated sequence of ddl7

Appendix IV: Translated sequence of intI gene located upstream of ddl7

Appendix V: List of poster and oral presentations

Appendix VI: List of awards, prizes and travel grants

Appendix VII: Research articles published as a result of the work presented in the thesis
List of Figures

**Figure 1-1.** Mechanisms of HGT in bacteria. a) Transformation b) Transduction, c) Conjugation. The figure was reproduced from Furuya and Lowy (2006). ................. 41

**Figure 1-2.** Structure of a typical class 1 integron where intI and the GCs in the array are oriented in the opposite direction. intI1, integrase gene for class 1 integron; Pc, GC promoter; Pint, promoter for intI; attI, recombination site; qacEΔ1, partially deleted genes that encode resistance to quaternary ammonium compound resistance; sul1, gene encoding sulphonamide resistance and attC, recombination site. ...................... 43

**Figure 1-3.** Structure and sequence of attI1 and attC. A. Sequence of attI1 (Cambray, Guerout et al. 2010). L and R, the simple integrase binding sites; DR1 and DR2, direct repeat sequences. B. Sequence of attC (Cambray, Guerout et al. 2010). R' and R" are the inverted repeats located at the boundaries and L' and L" are the inverted repeats located 6 bp upstream of R' and 5 bp downstream of R", respectively. ...................... 44

**Figure 1-4.** Structure of different class of integrons. A) Class 1 integron in Tn21 (Krin, Cambray et al. 2014) carrying aadA1 cassette encoding streptomycin/spectinomycin resistance where intI and the GCs in the array are oriented in the opposite direction. B) A class 2 integron (Hussein, Ahmed et al. 2009) carrying three resistance GCs. C) A class 3 integron (Collis, Kim et al. 2002). intI, integrase gene; Pc, gene cassette promoter; attI, recombination site; qacEΔ1, partially deleted genes that encode resistance to quaternary ammonium compound resistance; sul1, gene encoding sulphonamide resistance; orfx, gene with unknown function and attC, recombination site on the GCs recognized by the integrase. The sequence GTTRRRY (R is a purine, Y is a pyrimidine) is the integrons cross-over point where integration of GCs occur.... 48

**Figure 1-5.** Phylogenetic tree of known integron integrases (IntI). A single integron IntI is included for each bacterial species, provided that all IntI from that species cluster together in a preliminary analysis. Only species for which most of the GC array has been sequenced are included. Black boxes indicate integrons that are associated with antibiotic resistance GCs, with the particular cassette identified in the box. Class 1, 2 and 3 integrons can contain multiple ARGs. The accession number of each integron integrase is in parentheses next to the taxon name of its host and the number of GCs associated with it is in brackets. The tree and bootstrap support values were inferred by maximum likelihood using PHYML. The different colours represent different families of...
proteobacteria. The position of IntI of Treponema denticola in the tree has been shown in a red-lined box. The tree and description were adapted from Boucher, Labbate et al. (2007) (p. 304).

**Figure 1-6.** Alignment of the promoter regions of intI genes from the CI of V. cholerae (Vch), V. metschnikovii (Vme), V. parahaemolyticus (Vpa), and V. natriegens (Vna) and from class 1 (cl1-MI), class 2 (cl2-MI), class 3 (cl3-MI), and class 5 (cl5-MI) MIs (1). Putative LexA-binding sequences are boxed, whereas putative σ70 promoter elements (−35 and −10) are underlined and the translation start site of intI is boxed in red. The location of the LexA motif within a class 1 integron (on the top) is shown. The alignment section of the figure was adapted from Guerin, Cambray et al. (2009).

**Figure 1-7.** The regulatory regions of intI and cassette array are shown in the alignment of V. cholerae and V. mimicus strains. The intIA gene sequence is shown in bold italicized characters. Regulator binding site are framed. -10 and -35 promoter boxes are shown in grey. The attI recombination point is indicated by a vertical bar. The figure was taken from Krin, Cambray et al. (2014) (Supplementary figure S1).

**Figure 1-8.** Model of insertion and excision of circular GCs into integron platforms by site-specific recombination. The 5ʹ-CS contains the core integron consisting of intI, attl and Pc and the 3ʹ-CS contains the qacEΔ1 and sul1. The first event of insertion of a circular GC is an example of attIxattC recombination and the second insertion is an example of attCxattC recombination.

**Figure 1-9.** The structure of the class 1 integron detected on the plasmid pCG4 of gram-positive bacterium Corynebacterium glutamicum. The white box represents the integrated GC and black boxes the 5ʹ- and 3ʹ-conserved segments (CS). Numbers below the vertical arrows represent positions of the base pairs differing from those of integron InC. Pant (now named as Pc) indicates the integron promoter involved in expression of the streptomycin/spectinomycin resistance gene associated with the cassette. The following genes are indicated: int, site-specific integrase; aadA2a, streptomycin/spectinomycin resistance; qacEΔ1, antiseptics resistance; sul1, sulfonamide resistance; ORF5, unknown function. The figure was adapted from Nesvera, Hochmannova et al. (1998).

**Figure 1-10.** Schematic representation of arrays of class 2 integrons found in the bacterial isolates of clinical and environmental samples (n = 126). The thin vertical closed bar represents the attI2 site, and the ovals represent the attC sites of the GCs.
All intI2 genes from these arrays were sequenced, and they revealed the usual internal stop codon. The figure and description were adapted from Ramírez, Piñeiro et al. (2010) (p.703).

**Figure 1-11.** Schematic representation of a class 3 integron detected in a K. pneumoniae isolate. The figure was adapted from Jones-Dias, Manageiro et al. (2016) (p.10).

**Figure 1-12.** Agarose gel electrophoresis of the variable size PCR products obtained using 5'-CS and 3'-CS primers. Lane 1, 1-kb DNA ladder; lane 2, K. pneumoniae 154; lane 3, P. aeruginosa 702; lane 4, E. aerogenes 177; lane 5, E. cloacae 588; lane 6, S. marcescens 616; lane 7, S. marcescens 946; lane 8, S. marcescens 947; lane 9, S. marcescens 909; lane 10, S. typhimurium 101; lane 11, E. coli 801; lane 12, Proteus mirabilis 820. The figure was adapted from (Levesque, Piche et al. 1995).

**Figure 1-13.** The approximate location of the widely used PCR primers for detection of integrons.

**Figure 1-14.** Multiplex PCR amplification of the class 1, 2 and 3 integrons (Lane 1 and 2). PCR bands show that the amplification products of class 1, 2 and 3 integrons were 280, 788 and 979 bp, respectively. The figure was adapted from Ren, Zhao et al. (2013).

**Figure 1-15.** Flowchart showing the steps for different metagenomic approaches to investigate the presence of ARGs in an environmental or clinical sample. The common steps for all different approaches are shown in orange colour on the top.

**Figure 1-16.** The anatomy of the human mouth. The photo was adapted from Wikipedia on 20 June 2017.

**Figure 1-17.** Microbial diversity in the oral cavity according to the phylum. The number before the name of the phyla indicates the number of taxa in each phylum. The photo was adapted from Dewhirst (2015).

**Figure 1-18.** A subgingival biofilm with Actinomyces sp. (green), bacteria (red) and eukaryotic cells (large green cells on top). The four different layers of the biofilm are depicted with the root surface orientated to the bottom of the image. Scale bar = 10 μm. The image was adapted from Zijng, van Leeuwen et al. (2010) (p.3).
**Figure 1-19.** Structure of the unusual chromosomal integron of T. denticola ATCC35405. The intI gene and cassette ORFs (block arrows, sizes not to scale) are numbered according to GenBank accession no. NC_002967. White ORFs have no known function, light grey ORFs are related to conserved hypothetical proteins and dark grey ORFs are related to proteins of known function (blast E values<0.00001). Circles indicate putative 59-39 recombination sites. Underlined cassettes (A–G) are duplicates, defined as containing ORFs with >95% amino acid identity. The figure was taken from Coleman, Tetu et al. (2004) (p.3524).………85

**Figure 1-20.** The phylogenetic tree of the 14 integron-carrying T. denticola strains found in the oral cavity. The tree was built using the 31 marker genes. Red numerals indicate that the groupings are consistent with the phylogenetic relationship and the blue numeral highlights an unusual sharing of GCs between not-so-closely-related strains OTK and SP37. The figure was adapted from Wu, Doak et al. (2013) (p.7).…….87

**Figure 2-1** Saliva DNA collection, Preservation and Isolation Kit (Norgen, Canada) used for collecting saliva from Bangladeshi volunteers. The photo was modified from Norgen’s website. ……………………………………………………………………………………………………… 95

**Figure 2-2** Map of pGEM-T Easy vector (Promega, UK) (30 May 2016).……………… 105

**Figure 2-3** Vector map of pET-28a(+) used for cloning and expression of ddl genes in E.coli BL21 (DE3). The map was created using SnapGene software (from GSL Biotech; available at snapgene.com). ………………………………………………………………………………… 108

**Figure 3-1.** The target sites of primers used to detect different types of integrons in the oral metagenomic DNA. The primers MARS2 and MARS3 designed based on the attC of the unusual integron of T. denticola are shown aligned with their target binding sites. The core sites R'' and L'', and R' and L' have been indicated by bold fonts. ………………… 120

**Figure 3-2.** Flow chart of the methods used to detect integrons and associated gene cassettes in the oral metagenomic DNA. …………………………………………………………… 122

**Figure 3-3.** Agarose gel electrophoresis of the metagenomic DNA isolated from pooled saliva samples of UK (left) and Bangladeshi (right) volunteers. HyperLadder I (Bioline) was used as size marker (200 bp-10 kbp). The gel was imaged in Alphalnager gel documentation system (Alpha Innotech). Extraction was done in triplicate……………… 123
**Figure 3-4.** Abundance of different bacterial phyla in pooled saliva of UK and Bangladeshi individuals. ................................................................. 124

**Figure 3-5.** Abundance of bacteria in the oral cavity in the pooled saliva sample of UK and Bangladesh grouped according to their order......................................................... 125

**Figure 3-6.** PCR products generated by previously published primers using the UK metagenomic DNA as template. The name of the primers used are given under the figures. The PCR were run in triplicate................................................................. 126

**Figure 3-7.** A) Products of integrase PCR using the UK metagenomic DNA as template. B) EcoRI digestion of plasmids prepared from randomly picked clones to find the clone containing 280 bp insert which is the expected size of the intI1 PCR........ 127

**Figure 3-8.** PCR products obtained using TDIF and MARS 2 primers using the metagenomic DNA of UK saliva samples (A). Agarose gel electrophoresis of the EcoRI digest of the plasmids containing different size inserts (B-C). The numbers on the top indicate the number of the clones. The inserts were sequenced and analysed......... 129

**Figure 3-9.** PCR products obtained using TDIF and MARS2 primers using the metagenomic DNA of Bangladeshi saliva samples (A). Agarose gel electrophoresis of the EcoRI digest of the plasmids containing different size inserts (B-C). The numbers on the top indicates the number of the clones................................................................. 130

**Figure 3-10.** Phylogenetic tree of partial intI sequences detected in the oral metagenomic DNA. The evolutionary history was inferred using the Neighbour-Joining method and the evolutionary distances were computed using the Maximum Composite Likelihood method. Evolutionary analyses were conducted in MEGA6. The intI1 of P. aureginosa was used as outgroup. ................................................................. 132

**Figure 3-11.** Multiple sequence alignment of putative attI sites detected in the clones containing all the features of integrons located upstream of the first gene cassettes with attI1, attI2 and attI3 (Partridge, Recchia et al. 2000, Collis and Hall 2004). The inversely oriented integrase binding sites were labelled as S1 and S2. The recombination point as well as -35 sequence and -10 sequence of the putative Pc have also been indicated. The sequences after the recombination point are shown in lower case................................................................. 133
Figure 3-12. Detection of the core site R" with a nearly consensus RYYYYAAC sequence on the first GCs in the library of PCR amplicons. The location of the reverse primer (MARS2) which contains the core site L'' has been indicated with an arrow... 134

Figure 3-13: A) PCR products generated using attC-based forward (MARS5) and reverse (MARS2) primers using Bangladeshi metagenomic DNA. B, C and D) EcoRI digestion of the plasmids to see the size of the inserts................................. 139

Figure 3-14. A) PCR products generated using attC-based forward (MARS5) and reverse (MARS2) primers using UK metagenomic DNA. B and C) EcoRI digestion of the plasmids to see the size of the inserts........................................... 140

Figure 3-15. Core sites R' (1R) and R" (1L) abutting the forward (MARS5) and reverse attC primers (MARS2) in two representative GCs (MMU26 and MMB3/9) recovered in this study. The location of primers binding sites are marked with arrow. ...................... 142

Figure 4-1. The common biosynthetic pathway of peptidoglycan in bacteria. The role of Ddl and other enzymes in synthesis of peptidoglycan has been shown. The figure was adapted from Bugg, Braddick et al. (2011)......................................................... 161

Figure 4-2. The mechanism of the action of D-cycloserine. D-cycloserine inhibits the two key enzymes responsible for peptidoglycan synthesis in bacteria (Strominger, Ito et al. 1960, Lambert and Neuhaus 1972). The structures were obtained using ChemDraw Professional 16.0................................................................. 162

Figure 4-3. The model of mechanisms of resistance of D-cycloserine in M. tuberculosis and other bacteria including E. coli. Overproduction of targets, target modification by mutation, mutation in the transporter genes, increased efflux and mutation metabolic genes are the known mechanisms of D-cycloserine. Ddl and Alr were represented by blue and orange shapes. D-cycloserine is in blue squares.............................. 164

Figure 4-4. (A and B) Site-directed mutagenesis to change the cytosine and guanine at 490 and 777 position of the coding sequence of ddl6, respectively. The primers used are indicated with small arrows. Red cross marks indicate the location of point mutations. The PCR amplified products were ligated using the ligase provided with the Phusion site-directed mutagenesis kit...................................................... 168

Figure 4-5. Cloning of ddl6 and ddl7 into pHCMC05 vector. The vector is digested with BamHI and XbaI and purified with PCR purification kit. The coding sequence of ddl6
andddl7 are PCR amplified using TddlF and TddlR primers having BamHI and XbaI restriction sites. The PCR products are purified and digested with BamHI and XbaI and purified again. The digested PCR products were then ligated to the digested vector and transformed into E. coli α-select for manipulation of the plasmids. After confirming the correctness of the insert by sequencing, the plasmids are transformed to B. subtilis 168.

Figure 4-6. A) Approximate location of the primers (small arrow) on the genome of T. denticola ATCC35405 (NC_002967.9) (ORFs were annotated according GenBank database). Four forward primers were designed upstream of the intI of T. denticola. B) These forward primers were coupled with a reverse primer, ddlR designed from the end of ddlS located on the 2,024 bp PCR amplicon recovered from the library.

Figure 4-7. Primers used to clone ddl with their upstream attl sequence containing putative Pc promoter.

Figure 4-8. Construction of two new pET-28a(+) vectors carrying ddl6 and ddl7 with their 198 bp upstream region that contains the putative attl sites and putative Pc. Another two plasmids carrying only the coding sequence of the genes were used as negative control.

Figure 4-9. (A) Genetic organisation of 2,024 bp inserts in pGEM-T Easy vector containing a partial intI, Pc, and ddl within the first cassette. (B) Comparison of putative attl sequence preceding ddl in the pGEM-T Easy inserts (TMU6/7) with the putative attl of integron of T. denticola ATCC 35405 (Coleman, Tetu et al. 2004). Location of putative Pc with their consensus -35 and -10 sequences on attl has been shown. The integrase binding sites S1 and S2 as well as DR1 and DR2 are also shown. The recombination point G↓TT and the putative transcription start site (TSS) located at the 3'-end of attl are also shown. (C) Comparison of partial sequence of attC detected at the 3'-end of the 2,024 bp insert with a typical complete attC associated with Tde1837 of T. denticola integron (Coleman, Tetu et al. 2004).

Figure 4-10. The sequence of ddl6 (in bold font) with its upstream and downstream sequence within the 2,024 bp insert. The positions of SNPs (marked with red) by which ddl6 differed from ddl7 have been shown underneath. The non-coding sequence in between partial intI (underlined red at the 5'-end) and ddl6 is shown in lower-case letter. This region contains the putative attl site (underlined green), RBS, IntI-binding sites (S1 and S2) and the -10 and -35 sequences of the putative Pc. The recombination
point is shown by a vertical arrow. The R'' and L'' sequences within the partial attC at the downstream of the gene have also been shown (grey shaded). The location of priming sites of TDIF and MARS2 primers is shown by arrows. ................................. 178

**Figure 4-11.** The genetic organisation of ddl of T. pedis found on a wgs contig (GenBank accession number: NZ_AOTN01000179) which shows 98% nucleotide sequence identity to the integron-located ddls. This homologous ddl was found to be flanked by two attC sites. ................................................................. 179

**Figure 4-12.** The location of the ddl homologous to ddl6 and ddl7 (bolded and red underlined) within the 5699 bp contig of T. pedis B683 (GenBank accession number: NZ_AOTN01000179). The noncoding sequences are marked with small case letters. The R", L", L' and R' core-sites of the putative attC sites are bolded and marked with blue, green, purple and red coloured font, respectively. The different ORFs located upstream and downstream of ddl are underlined with different colour (orange-multidrug transporter; black-hypothetical protein; green-hypothetical protein; blue-HDIG-domain containing protein; red-ddl and purple hypothetical protein). ................................................................. 180

**Figure 4-13.** Sequence and size of the putative attC sites detected within the contig of T. pedis B683 that contains a ddl. The attC sequences of the contig were compared with attC of T. denticola (Tde1837) and partial attC located downstream of ddl6 or ddl7. .................................................................................................................. 181

**Figure 4-14.** Multiple alignment of the coding sequences of ddl6, ddl7 and ddl(GC) of T. pedis. The alignment was done by using T-Coffee server (Notredame, Higgins et al. 2000). The position of two mutations at c.490 and c.777 by which ddl6 is different from ddl7 is shown by the blue arrows. .................................................................................................................. 183

**Figure 4-15.** The identity of the flanking sequence of ddl6/ddl7 (40 bp upstream and 29bp downstream) with their closest homologue, ddl(GC) located on a 5699 bp contig of T. pedis B683 genome. The putative core sites of attC (R" and L"), the simple integrase binding site (S1) are shown. The recombination points located on S1 and core site, R' of the attC located upstream of ddl(GC) (attC_{HDIG}) are marked with a down arrow. The primer binding sites are shown by small horizontal arrows. The start codons (ATG) of the homologous ddls are bolded. The attC sites downstream of the genes are the attC for this gene. .................................................................................................................. 185
**Figure 4-16.** Phylogenetic tree of ddl6, ddl7 and ddl(GC) along with the ddls of different strains of T. denticola and some other Treponema species. The ddlA of E. coli was used as an out group. The evolutionary history was inferred using the Neighbour-Joining method. Evolutionary analyses were conducted in MEGA6 (Tamura, Stecher et al. 2013). The ddls found within GCs are shown in rectangular box. ......................... 186

**Figure 4-17.** The phylogenetic tree of deduced amino acid sequences of Ddl6 and Ddl7 (highlighted in red rectangle) along with the Ddl homologues of other species. The evolutionary history was inferred using the Neighbour-Joining method. Evolutionary analyses were conducted in MEGA6 (Tamura, Stecher et al. 2013). ......................... 187

**Figure 4-18.** Multiple sequence alignment of Ddl and Ddl7 with the Ddl(GC) of T. pedis (Ddl_Tped) (WP_029410393), DdlA (BAE7676162) and DdlB (EGT67290) of E. coli, Van A (AAA65956) of E. faecium as well as VanB (YP_009076352) of E. faecalis. The amino acid residues known to bind D-Ala and ATP have been marked on their top. The conserved residues are marked with an asterisk and grey shaded. The position of the omega loop has been indicated by a box. The changes between Ddl6 and Ddl7 are numbered on top and marked with red rectangles. The active sites and position of omega loop were detected and marked based on previous reports (Shi and Walsh 1995, Noda, Kawahara et al. 2004). ................................................................. 188

**Figure 4-19.** MIC of D-cycloserine against three E. coli strains EC121 (pGEM-T Easy), EC126 (pGEM-T Easy::intI-attI-ddl6) and EC127 (pGEM-T Easy::intI-attI-ddl7). The MIC was determined using the agar dilution method. Each experiment was repeated at least three times. ................................................................. 190

**Figure 4-20.** The country sources, lot numbers and physical forms of D-cycloserine manufactured by Sigma Aldrich used in this study. The labelled purity for both product was 96.0%. .......................................................................................... 191

**Figure 4-21.** Identification of the SNP of the integron-located ddls responsible for altering the D-cycloserine resistance phenotype of E. coli. EC121 was the vector control carrying empty pGEM-T Easy. ................................................................. 192

**Figure 4-22.** The MIC of D-cycloserine against the strains of B. subtilis 168. Expression of the inserted genes was induced by 0.5 mM IPTG. ................................................................. 194

**Figure 4-23.** Alignment of Pc promoter detected upstream of ddl6 and ddl7. The Pc promoter was found to exhibit very good similarity with the P2 promoter of class 1
integron (Collis and Hall 1995) and E. coli σ70 promoter (Ozoline, Deev et al. 1997). The identical nucleotides are highlighted with grey shading. ................................................................. 195

**Figure 4-24.** Expression of ddl6 and ddl7 under the control of Pc promoter without IPTG induction. No change in D-cycloserine susceptibility was observed in the clones carrying only the coding sequence of ddls cloned into pET-28a. Resistance to D-cycloserine increases in the clones carrying the putative Pc region. The Pc of both ddl variants were 100% identical. ........................................................................................................... 197

**Figure 4-25.** (A) PCR to clone the upstream sequence of integron carrying ddl using the genetic information of T. denticola ATCC35405 strain. UK metagenomic DNA was used as a template for PCR, (B) Agarose gel electrophoresis of EcoRI digest of pGEM-T Easy vectors carrying ~5 kb inserts. ................................................................................................................. 199

**Figure 4-26.** The genetic arrangement of 4421 bp pGEM-T Easy insert carrying ddl7, full length intI and another partial ORF encoding a hypothetical protein. The primers used to amplify this product are shown with small arrows. The genetic arrangements of the equivalent site of the closest homologues including T. denticola H-22 (GenBank accession: AGDV01000005.c23656-19631), T. denticola ATCC35405 (Genbank accession: AE017226:1873126-1877351) and T. denticola US-Trep acdtB (GenBank accession: AGEB01000014.1:3405-5167) have also been shown. The most variable region among the strains was found to be the sequence in between the ORF for hypothetical protein and intI. The size of this region is also shown. ......................... 200

**Figure 4-27.** The sequence of the 4421 bp insert of pGEM-T Easy carrying upstream sequence of ddl7. The partial ORF located at the 5' end of the insert that encodes a hypothetical protein is underlined with black colour. The putative ORF encoding the intI is bolded and underlined with purple colour and the coding sequence for ddl7 at the 3' end of the insert was underlined with red colour. The putative RBS located upstream of the genes was marked with green font. The -10 and -35 sequences of the putative promoter for integrase, Pint and the Pc promoter is in red font and underlined. The putative integrase binding sites are in blue font. The recombination point on attI is shown with a red vertical arrow.................................................................................................................. 201

**Figure 4-28.** The phylogenetic tree of homologues of intI found upstream of ddl7 in the 4421-bp inserts in pGEM-T Easy. The evolutionary relationship was inferred using Neighbour-Joining Method (Saitou and Nei 1987). Evolutionary analyses were
conducted in MEGA6 (Tamura, Stecher et al. 2013). intI1 of E. coli was used as outgroup.

**Figure 4-29.** The growth curve of B. subtilis 168 overexpressing ddl6 and ddl7 under IPTG induction. The growth rate was compared with the vector control and host controls.

**Figure 4-30.** The agarose gel electrophoresis of the PCR products obtained using the TDIF and ddlR primer to detect the presence of ddl in individual saliva metagenomic DNA samples from Bangladesh (A) and the UK (B).

**Figure 4-31.** The presence of GC encoding cof-like hydrolase in individual oral metagenomic DNA samples isolated from the healthy volunteers of UK and Bangladesh.

**Figure 5-1.** A plot of the reaction velocity ($V_0$) as a function of the substrate concentration [S] for an enzyme that obeys Michaelis-Menten kinetics. The Michaelis-Menten constant (Km) is the substrate concentration yielding a velocity of Vmax/2. The figure was adapted from Berg, Tymoczko et al. (2002).

**Figure 5-2.** SDS-PAGE analysis of the cell lysates collected at different time points for optimization of time for expression of ddl6 (A) and ddl7 in E. coli BL21 (DE3) after induction with 0.2 mM IPTG. The red arrows indicate the region where the 38 kDa recombinant Ddl6 and Ddl7 appear on the gel.

**Figure 5-3.** Western blot analysis of EC306 (pET-28a::ddl6) and EC307 (pET-28a::ddl7) expressing Ddl6 and Ddl7, respectively. The 6x-his-tagged recombinant proteins were detected using anti-His-Tag antibody. Expression was induced using 0.2 mM IPTG.

**Figure 5-4.** SDS-PAGE of different fractions of proteins obtained during optimization of purification. A) Fractions obtained during purification under native conditions from the cell-lysate of IPTG-induced EC307 (pET-28a::ddl7), B) Fractions obtained during purification under native conditions from the cell-lysate of EC307 without IPTG induction. Where, CL= raw cell lysates, S= supernatants after centrifugation, W1 and W2= washed fractions with buffer A (native/denatured), E1 and E2= two elutes eluted with buffer B, A= the agarose boiled with 1x sample buffer. His-Select Nickel Affinity Gel was used to bind the 6x-tagged recombinant Ddl7. The red horizontal arrow indicates the expected location of the expressed Ddl.
Figure 5-5. Chromatogram of purification of Ddl6 using FPLC system fitted with Ni-NTA superflow cartridge (1mL). The different parts of the chromatogram were labelled. .............................................................................................................................................231

Figure 5-6. SDS-PAGE of purified Ddl6, Ddl7, DdlAEc and DdlTd. 80 µL of the purified protein was mixed with 20 µL of 5x Protein Loading Buffer and boiled for 5 minutes at 95°C. 10 µL of the denatured proteins was loaded. .................................................................................................................232

Figure 5-7. Standard curve for inorganic phosphate.................................................................................233

Figure 5-8. (A) Release of inorganic phosphate in the reactions catalysed by Ddl6 and Ddl7 in the presence of 20 mM D-alanine. The differences of the release of Pi in between the negative controls (no Ddl) and Ddl catalysed reactions were analysed by one-way ANOVA (**P<0.001; ***P<0.0001). (B) In-vitro Inhibition of the ligase activity of recombinant Ddl7 in the presence of D-cycloserine (10 to 80 mM). Significant differences of the release of Pi in between the negative control (no D-cycloserine) and D-cycloserine-inhibited reactions were analysed by one-way ANOVA (****P<0.0001, *P<0.05). Data represent mean values of three independent experiments. Only the significant differences are shown by the asterisks.................................................................233

Figure 5-9. Ascending paper chromatography to detect the formation of D-ala-D-ala and D-ala-D-ser dipeptide as well as D-ala-D-lac depsipeptide. The chromatogram was developed using the solvent system n-hexane : acetic acid : water (12:3:5) supplemented with 0.5% ninhydrin. In the first two reactions (lane 1 and lane 2) 20 mM D-alanine was used, whereas in other reactions it was reduced to 10 mM and supplemented with 10 mM D-serine or D-lactate........................................................................................................234

Figure 5-10. Initial velocity versus substrate concentration [ATP] plots for Ddl6, Ddl7, DdlAEc and DdlTd. A saturating concentration of D-alanine (100 mM) was used with a variable amount of ATP to determine the $K_{m_{\text{ATP}}}$ of the enzymes. The plots generated and the $K_{m_{\text{ATP}}}$ values calculated by the curve fitting options of GraphPad Prism software.*The scale at the vertical axis of DdlAEc is different. .............................................................236

Figure 5-11. Initial velocity versus substrate concentration [D-alanine] plots for Ddl6, Ddl7, DdlTd and DdlAEc. Variable concentrations of D-alanine were used with a fixed saturating concentration of ATP. The plots were generated and $K_{m_{\text{D-ala2}}}$ values were calculated using GraphPad Prism.*Different scale of the X-axis. .........................................................238
**Figure 5-12.** Substrate specificity of VanG as D-Ala:D-X ligase (A) and D-Ser:D-X ligase (B) tested by TLC. Lane 1, no enzyme; lane 2, D-alanine, lane 3, D-serine; lane 4, D-lactate; lane 5, D,L-aminobutyrate; lane 6, D,L-hydroxyvalerate; lane 7, D-norvaline; lane 8, D-valine; lane 9, D-norleucine; lane 10, D-threonine; lane 11, D-leucine; lane 12, D-phenylalanine; lane 13, D-isoleucine; lane 14, D-methionine; lane 15, D-tryptophan. The lane 3 was boxed to highlight the reaction containing D-serine. Due to high specificity towards D-serine, no D-aladala was formed. The figure was adapted from Meziane-Cherif, Saul et al. (2012) (p.37585).

**Figure 6-1.** Steps involved in comparative modelling. The first step of the process is the identification of a related protein with known 3D structure (fold recognition or fold assignment). In the second step the target and template sequences are aligned and then a model is built based on the alignment and structure of the template. At the last step the built 3D structure is further refined and validated. When no templates are found for the target sequence, models can be generated using ab initio modelling...

**Figure 6-2.** Approximate correspondence of the algorithms, accuracy, and the biological usefulness of protein structure predictions. The figure was adapted from Zhang (2009) (p.146).

**Figure 6-3.** DdlB of E. coli complexed with ADP 1(S)-aminoethyl [2-carboxy-2(R)-methyl-1-ethyl] phosphinic acid as phosphinophosphonate that arise from phosphoryl transfer from ATP (Fan, Moews et al. 1994). The PDB file of DdlB (PDB code: 2DLN) was downloaded from protein data bank and viewed and analysed in UCSF Chimera.

**Figure 6-4.** Proposed mechanism of formation of D-ala-D-ala dipeptide by DdlB of E. coli. The figure was adapted from the article published by Fan, Moews et al. (1994) and modified. The predicted steps for formation of D-ala-D-ala dipeptide are shown.

**Figure 6-5.** Flowchart of the I-TASSER Suite pipelines that consist of three steps: template identification, full-length structure assembly and structure-based function annotation. The figure was taken from Yang and Zhang (2015).

**Figure 6-6.** Predicted 3D-structures of Ddl6 (A) and Ddl7 (B) monomer as determined by I-TASSER. The model was viewed by UCSF Chimera. The residues at 164 and 259 are shown. The N-terminal domain and C-terminal domains were shown in blue and red colour, respectively.
**Figure 6-7.** Superimposition of predicted 3D models of Ddl6 (blue) and Ddl7 (red). The alignment of the 3D structures of Ddl6 and Ddl7 were done using TM-align program (http://zhanglab.ccmb.med.umich.edu/TM-align/). ................................................................. 262

**Figure 6-8.** Alignment of putative 3D structures of Ddl6 and Ddl7 (blue colour) with crystal structure of DdlB of E. coli (PDB code 2DLN) (red colour). The location of the Ser150 and Tyr216 of DdlB of E. coli and the corresponding residues, Ser184 and Tyr250, of Ddl6 and Ddl7 are shown. The structure of the omega-loop of Ddl6 and Ddl7 were found to be very similar to that of DdlB of E.coli. ................................................................. 264

**Figure 6-9.** Alignment of predicted 3D structures of Ddl6 and Ddl7 with DdlA of S. typhimurium (PDB code 3I12) and VanG of E. faecalis (PDB code 4FU0). The Ddl6 and Ddl7 are shown in blue colour and the homologues in red colour. ...................... 267

**Figure 6-10.** Spatial positions of the putative active/binding sites of Ddl6 and Ddl7. The active sites were mapped according to experimentally verified conserved active sites of DdlB of E. coli (Fan, Moews et al. 1994, Shi and Walsh 1995)................................................. 268

**Figure 6-11.** 2D structures of D-alanine and D-cycloserine. The structures were obtained from ChemDraw Professional 16.0 (PerkinElmer Informatics). ..................... 269

**Figure 6-12.** Mode of interaction of D-alanine and D-cycloserine with WT Ddl6 and its W259C mutant. Interaction of D-alanine with WT Ddl6 (A), D-ala with W259C mutant (B), D-cycloserine with WT Ddl6 (C) and D-cycloserine with W259C mutant. .......... 270

**Figure 6-13.** 2D structures of quercetin, apigenin and salvicine. The structures were obtained using ChemDraw Professional 16.0 (PerkinElmer Informatics)......................... 271

**Figure 6-14.** Mode of interaction of apigenin and quercetin with the D-alanine binding sites of WT Ddl6 and W259C mutant. A) Apigenin with WT Ddl6, B) Apigenin with W259C mutant, D) Quercetin with WT Ddl6 and D) Quercetin with W259C mutant. . 272

**Figure 6-15.** Model of interaction of apigenin and quercetin with the ATP binding sites of WT Ddl6 and W259C mutant. A) Apigenin with WT Ddl6, B) Apigenin with W259C mutant, D) Quercetin with WT Ddl6 and D) Quercetin with W259C mutant. .......... 273

**Figure 6-16.** 2D structures of cysteine and tryptophan. The structures were obtained using ChemDraw Professional 16.0 (PerkinElmer Informatics). ......................... 276
Figure 6-17. Multiple sequence alignment of the four proteins in PDB which have highest structural similarity to Ddl6 and Ddl7 (PDB codes: 3I12, 4FU0, 3TQT and 1EHI). The E. coli DdlB (PDB code: 2DLN) and E. faecium VanA (PDB code: 1E4E) were included in the alignment to compare the active/binding sites. PDB code 3I12: Ddl of S. enterica subsp. enterica Serovar Typhimurium Str. LT2; 4FU0: D-ala-D-ser ligase (VanG) of E. faecalis; 3TQT: Ddl of Coxiella burnetii; 1EHI: D-ala-D-lac ligase of L. mesenteroides (LmDdl2). The vancomycin resistant proteins are marked with red arrows.

Figure 7-1. The representation of the hypothesis that how a single mutation in Ddl6/Ddl7 could led to vancomycin resistance. A) Y250F mutation in Ddl6/Ddl7 is hypothesized to gain VanG-type vancomycin resistance phenotype by gaining the ability to bind D-serine at the second subsite; B) Y250H mutation in Ddl6/Ddl7 could also led to gaining a VanA-type vancomycin resistance.

Figure 7-2. Diagrammatic representation of the hypothetical model that plant flavonoids could be responsible for selecting the ddls and maintaining within the first GC of the integron of T. denticola. The integron encoded Ddl6/Ddl7 could play a dual role: sequestration of flavonoids or other molecules that are able to bind Ddl and catalysing the formation of D-ala-D-ala dipeptide in presence of inhibited Ddl encoded by the house-keeping genes.

List of Tables

Table 1-1. Mobile genetic elements involve in dissemination of antibiotic resistance.

Table 1-2. Different variants of Pc and P2 promoters detected in class 1 integrons.

Table 1-3. List of gram-negative bacterial species carrying class 1 integrons.

Table 1-4. Some representative bacteria carrying class 3 integrons.

Table 1-5. Bacterial species that were identified to carry chromosomal integrons.

Table 1-6. The programs and tools available to detect integron and associated GCs in the gene sequence.
Table 1-7. ORFs within the CI of T. denticola ATCC35404 that encode proteins with known functions. ................................................................................................................. 84

Table 1-8. The COG functional category distributions of the integron gene cassettes identified in the different locations of oral cavity ................................................................. 86

Table 2-1. Bacterial strains and plasmids used in this study ................................................................. 91

Table 2-2. List of primers used in this study .......................................................................................... 98

Table 2-3. Software and tools used for nucleotide and amino acid sequence analysis .............................................................................................................................................. 112

Table 3-1. Percentage identity of partial intIs with their closest homologues in GenBank .................................................................................................................................................. 131

Table 3-2. Orientation of the first GCs in the cassette library (not mapped to scale). TDIF and MARS 2 are the primer binding sites. The size of the inserts has been shown with numbers at ends .................................................................................................................. 135

Table 3-3. Characterization of all gene cassettes detected in the saliva metagenomic DNA from UK and Bangladesh using TDIF and MARS2 primer combination to detect the first gene cassette ........................................................................................................... 138

Table 3-4. Orientation of the putative ORFs in the gene cassettes (For orientation type A and B, please see the Table 3-2) ......................................................................................... 141

Table 3-5. Core sites R’ (1R) and R” (1L) abutting the forward and reverse attC primer sequence on the gene cassettes ................................................................................................. 143

Table 3-6. Characterization of all gene cassettes detected in the saliva metagenomic DNA from UK and Bangladesh using attC-based primers .................................................................................. 146

Table 5-1. Steady-state kinetic parameters of Ddl6, Ddl7, DdlTd and DdlEc when determined with a fixed saturating concentration of D-alanine (100 mM) and variable concentration of ATP ........................................................................................................................................... 236

Table 5-2. Steady-state kinetic parameters of Ddl6, Ddl7, DdlEc and DdlTd determined with a fixed saturating concentration of ATP (500 µM) and variable concentration of D-alanine .......................................................................................................................................... 238
Table 5-3. Comparison of protein parameters of Ddl6 and Ddl7 with Ddls encoded by housekeeping genes.................................................................241

Table 5-4. Comparison of kinetic parameters of Ddl6 and Ddl7 with DdlA and DdlB of E. coli as well as Ddl of S. lavendulae and Ddl of M. tuberculosis..............................245

Table 6-1. Values of C-scores, estimated TM-scores and RMSD of the model 1 of Ddl6 and Ddl7 as estimated by I-TASSER.................................................................261

Table 6-2. Top 10 structurally homologous protein of Ddl6 and Ddl7 in PDB. The I-TASSER models of Ddl6 and Ddl7 were used to compare the structural similarity with the PDB structures.........................................................................................................................265

Table 6-3. Free energy of binding (ΔG) and GOLD scores of D-alanine and D-cycloserine with native Ddl6 and its two single mutants .........................................................269

Table 6-4. Free binding energy (ΔG) and GOLD fitness score of apigenin, quercetin and salvicine for the D-alanine binding sites of native Ddl6 and its two single mutants. ........................................................................................................................................272

Table 6-5. Free binding energy (ΔG) and GOLD fitness score of apigenin, quercetin and salvicine for the ATP binding sites of native Ddl6 and its two single mutant, L164F and W259C........................................................................................................................................273
**List of important abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amp</td>
<td>Ampicillin</td>
</tr>
<tr>
<td>ARG</td>
<td>Antibiotic resistance gene</td>
</tr>
<tr>
<td>Bp</td>
<td>Base pair</td>
</tr>
<tr>
<td>CI</td>
<td>Chromosomal Integron</td>
</tr>
<tr>
<td>Cm</td>
<td>Chloramphenicol</td>
</tr>
<tr>
<td>CS</td>
<td>Conserved segment</td>
</tr>
<tr>
<td>D-cycloserine</td>
<td>D-cycloserine</td>
</tr>
<tr>
<td>Ddl</td>
<td>D-alanine-D-alanine ligase</td>
</tr>
<tr>
<td>ESBL</td>
<td>Extended-spectrum beta-lactamases</td>
</tr>
<tr>
<td>GC</td>
<td>Gene cassette</td>
</tr>
<tr>
<td>HGT</td>
<td>Horizontal gene transfer</td>
</tr>
<tr>
<td>intI</td>
<td>Integrase</td>
</tr>
<tr>
<td>IS</td>
<td>Insertion sequence</td>
</tr>
<tr>
<td>I-TASSER</td>
<td>Iterative Threading ASSEmbly Refinement</td>
</tr>
<tr>
<td>MDR</td>
<td>Multidrug-resistant</td>
</tr>
<tr>
<td>MGE</td>
<td>Mobile genetic element</td>
</tr>
<tr>
<td>MI</td>
<td>Mobile integron</td>
</tr>
<tr>
<td>MIC</td>
<td>Minimum Inhibitory Concentration</td>
</tr>
<tr>
<td>mL</td>
<td>Millilitre</td>
</tr>
<tr>
<td>ORF</td>
<td>Open reading frame</td>
</tr>
<tr>
<td>OTU</td>
<td>Operational taxonomic unit</td>
</tr>
<tr>
<td>Pc</td>
<td>Gene cassette promoter</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>Pint</td>
<td>Integrase promoter</td>
</tr>
<tr>
<td>SDS-PAGE</td>
<td>Sodium dodecyl sulphate polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>Tn</td>
<td>Transposon</td>
</tr>
<tr>
<td>XDR</td>
<td>Extensively drug-resistant</td>
</tr>
<tr>
<td>µL</td>
<td>Microliter</td>
</tr>
</tbody>
</table>
Chapter One: Introduction
1.1 Antibiotic resistance: A global crisis

Global health is currently under severe threat by the emergence of antibiotic resistance in clinically important pathogens. The WHO has warned that if proper actions are not taken, there is a potential of us entering a post-antibiotic era in the 21st century where once again a once treatable minor infection may kill (WHO 2014). The mobile antibiotic resistance genes (ARGs) carried on various mobile genetic elements (MGEs) including plasmids, transposons and integron gene cassettes (GCs) have been implicated in worldwide rapid emergence and spread of multidrug-resistant (MDR) pathogens. Identification of the principal reservoirs of ARGs has been suggested as one of the top research priorities to tackle the problem of the emergence of antibiotic resistance by a group of scientists who gathered in a meeting in New York, USA in 2011 (Bush, Courvalin et al. 2011). Some other recommendations from that meeting were the collaborative approach for the discovery of new antibiotics, re-engineering of old antibiotics, controlling the non-therapeutic use of antibiotics and improving people awareness and hygiene (Bush, Courvalin et al. 2011).

1.2 Genetic basis of antibiotic resistance

In general, antibiotic resistance means the continued growth of microorganisms in the presence of cytotoxic concentrations of antibiotics (Wright 2007). Bacteria can be intrinsically resistant to certain antibiotics but can also acquire resistance to antibiotics via mutations in chromosomal genes or via by horizontal gene transfer (HGT) (Brown and Reynolds 1980, Griggs, Gensberg et al. 1996).
1.2.1 Intrinsic resistance

Intrinsic resistance is the innate ability of a bacterial species to resist activity of a particular antimicrobial agent through its inherent structural or functional characteristics. It may be due to the absence of a drug target, inaccessibility of the drug into the bacterial cell, extrusion of the drug by chromosomally encoded active exporters or innate production of enzymes that inactivate the drug (Alekshun and Levy 2007, Mullany 2014, Blair, Webber et al. 2015). For example, gram-negative bacteria are naturally resistant to vancomycin due to the presence of an outer-membrane which the antibiotic cannot penetrate. Aerobic bacteria are intrinsically resistant to metronidazole because of the inability to anaerobically reduce the drug to its active form in the presence of oxygen (Reysset 1996).

1.2.2 Acquired resistance

Acquired resistance is heritable and involves mutations in antibiotic targets and transfer of resistant determinants usually on plasmids, transposons and other MGEs.

1.2.2.1 Mutation mediated resistance

Resistance related to point mutations may involve both pre-existing genetic determinants and acquired genes. Point mutations of pre-existing genes can affect both structural (e.g., DNA gyrase, topoisomerase IV) and regulatory genes (e.g., mexR encoding an efflux pump). Mutations in a structural gene are an effective mechanism of resistance for β-lactams, fluoroquinolones, streptomycin and rifampicin. Point mutations in regulatory genes have been found to confer resistance to β-lactams, fluoroquinolones and tetracyclines. The
evolution of extended-spectrum beta-lactamases (ESBLs) by mutation in the wild-type beta-lactamases is a common example of point mutations in acquired genes.

1.2.2.2 Horizontal gene transfer (HGT)

The acquisition of foreign DNA through HGT is one of the major important drivers of bacterial evolution and the development of antibiotic resistance. Acquired resistance by HGT occurs in various ways, including transformation, transduction and conjugation (Figure 1-1) (Furuya and Lowy 2006).

Transformation is the process by which naked DNA released on lysis of an organism is taken up by another organism. The ARGs can be integrated into the chromosome or plasmid of the recipient cell. In transduction, the ARGs are transferred from one bacterium to another by means of bacteriophages. During the conjugation process two bacteria come in close contact and form a mating bridge through which different mobile genetic elements such as plasmids and transposons can be transferred from the donor to the recipient cell (Furuya and Lowy 2006).
**Figure 1-1.** Mechanisms of HGT in bacteria. a) Transformation b) Transduction, c) Conjugation. The figure was reproduced from Furuya and Lowy (2006).

Examples of MGEs which are involved in the dissemination of antibiotic resistance by HGT are shown in Table 1-1.

**Table 1-1.** Mobile genetic elements involve in dissemination of antibiotic resistance.

<table>
<thead>
<tr>
<th>Genetic Element</th>
<th>General Characteristics</th>
<th>Resistance Determinant(s) Specified/ Examples²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plasmid</td>
<td>Variable size (1- &gt;100 kb), conjugative, and mobilizable</td>
<td>R factor: multiple resistances</td>
</tr>
<tr>
<td>Insertion sequence</td>
<td>Small (&lt;2.5 kb), contains terminal inverted repeats, and specifies a transposase</td>
<td>IS1, IS3, IS4, etc.</td>
</tr>
<tr>
<td>Composite (compound) transposon</td>
<td>Flanked by insertion sequences and/or inverted repeats</td>
<td>Tn5: Kan, Bleo, and Str</td>
</tr>
<tr>
<td>Complex transposon</td>
<td>Large (&gt;5 kb), flanked by short terminal inverted repeats, and specifies a transposase and recombinase</td>
<td>Tn1 and Tn3: β-lactamase; Tn7: Tmp, Str, Spc; Tn1546: glycopeptides</td>
</tr>
<tr>
<td>Conjugative transposon</td>
<td>Promotes self-transfer</td>
<td>Tn916: Tet and Mino; Tn1545: Tet, Mino, Ery, and Kan</td>
</tr>
<tr>
<td>Genetic Element</td>
<td>General Characteristics</td>
<td>Resistance Determinant(s) Specified/ Examples²</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Transposable bacteriophage</td>
<td>A bacterial virus that can insert into the chromosome</td>
<td>Mu</td>
</tr>
<tr>
<td>Other transposable elements</td>
<td>Other than composite, complex, and conjugative transposons</td>
<td>Tn4: Amp, Str, Sul, and Hg; Tn1691: Gen, Str, Sul, Cm, and Hg</td>
</tr>
<tr>
<td>Integron gene cassettes</td>
<td>Integrons facilitates acquisition and dissemination of gene cassettes; specifies an integrase, attachments sites, and transcriptional elements to drive expression of multiple resistance genes</td>
<td>Class 1 integron: Multiple single determinants and MDR efflux pump (Qac); Class 2 integron: Tmp, Strp, Str, and Spc (Tn7); Class 3 integron: carbapenems; Class 4 integron: <em>Vibrio</em> spp. super-integron</td>
</tr>
</tbody>
</table>

¹The information on the table was adapted from Alekshun and Levy (2007) (p.139).

²Abbreviations: Amp, ampicillin; Bleo, bleomycin; Cm, chloramphenicol; Ery, erythromycin; Fus, fusidic acid; Gen, gentamicin; Hg, mercury; Kan, kanamycin; Mino, minocycline; Spc, spectinomycin; Str, streptomycin; Strp, streptothricin; Sul, sulfonamide; Tet, tetracycline; Tmp, trimethoprim; Van, vancomycin.

### 1.3 Integrons

Integrons are genetic elements capable of capturing and expressing open reading frames (ORFs) embedded in gene cassettes (GCs) (Cambray, Guerout et al. 2010, Gillings 2014). The term “integron” or “DNA Integration elements” was first proposed by Stokes and Hall (1989) to describe a novel type of MGE carrying different ARGs. They were predicted to be mobile elements due to their wide distribution in different locations such as plasmids and transposons. Class 1 integrons from multidrug-resistant (MDR) pathogens are embedded in a MGE such as transposons or plasmids, however, integrons located on bacterial chromosomes, known as chromosomal integrons (CIs), are also common (Stokes and Gillings 2011).
1.3.1 Structure of integrons

Integrons are composed of three basic features which are essential for the gene capturing and expression functionality: a gene encoding a DNA integrase (intI); an integron associated recombination site, attI; and a GC promoter, Pc, which expresses the GCs (Figure 1-2). IntI catalyses recombination between the attC of the incoming GCs and attI of the integron platform (Collis, Grammaticopoulos et al. 1993, Collis and Hall 1995, Recchia and Hall 1995).

The attI site of the integron is the primary recombination site where a gene cassette is inserted by the action of IntI (Collis, Grammaticopoulos et al. 1993). The length of attI1 (attI of class 1 integron) site is 65 bp and contains a simple integrase binding site which consists of a pair of inversely oriented integrase binding domains called L and R (Figure 1-3) (Partridge, Recchia et al. 2000). The recombination point is located between G and TT of the 5' GTT-3' triplet of the R site (consensus 5' GTTRRRY-3') of the attI. There are two other directly...
oriented integrase binding site named DR1 (strong binding site) and DR2 (weak binding site) (Partridge, Recchia et al. 2000) (Figure 1-3).

A. attI

5’-TTTGAGTTATGACGAGCAGTATTAGCAGGAGGCTGTCGCTGACGAGGGCAGTTAGTGACGAGG-3’
3’-AAACTACAATACCTCGTGCTAGTACGAGTGCTCCCGCTCAGGTCGATTTTTTTTCAATACGCTAGG-5’

B. attCrecA

5’-ATCTCTAGATATTTACTACGAGTGCTGTCGAGGGCAGTTAGTGACGAGG-3’
3’-TACAGATTATCTTATGAGTCCGATTGTCGAGGAGGCGAGGAGGCTAATTTGGTTCAATACGCTAGG-5’

Recombination point

Figure 1-3. Structure and sequence of attI and attC. A. Sequence of attI (Cambray, Guerout et al. 2010). L and R, the simple integrase binding sites; DR1 and DR2, direct repeat sequences. B. Sequence of attC (Cambray, Guerout et al. 2010). R’ and R” are the inverted repeats located at the boundaries and L’ and L” are the inverted repeats located 6 bp upstream of R’ and 5 bp downstream of R”, respectively.

1.3.2 Gene cassettes (GCs)

GCs are small mobile genetic elements usually consisting of a single ORF and a recombination site called 59-base element or attC which is specifically recognised by IntI (Figure 1-2). Successive integration of cassette genes at the integron recombination site, attI results in the formation of cassette array (Collis, Grammaticopoulos et al. 1993).

GCs can exist in two forms: a free circular form which is unable to remain stable during cell division and a linear form integrated in the integron (Figure 1-2). The defining features of a typical GC in a cassette array are as follows: i) an open reading frame (ORF), ii) a short non-coding region upstream of the start codon (ATG, GTG, TTG) which usually contains a ribosome binding site (RBS), iii) a
stop codon of the gene usually located in the *attC* and iv) the *attC* itself located downstream of the gene (Recchia and Hall 1995, Partridge, Tsafnat et al. 2009, Gillings 2014). However, GCs carrying two or three different genes or ORFs or no ORFs have also been identified (Holmes, Gillings et al. 2003, Elsaied, Stokes et al. 2011). Similarly, not all GCs encode proteins or have RBS preceding the ORFs. For instance, Stokes, Holmes et al. (2001) identified 21 empty GCs (non-protein coding) out of 123 cassettes detected in the environmental samples. In the library of GCs they also found a total of 107 ORFs of which 50 had no RBS. The function of these empty or non-protein coding GCs in the integron is not known, however, it is assumed that they may encode promoters or regulatory RNAs (Holmes, Gillings et al. 2003). These non-protein-coding GCs are usually found in the CIs. For instance, the empty cassettes comprise between 4 and 49% of *Vibrio* CI cassette arrays (Boucher, Nesbo et al. 2006). It has also been found that GCs are present in some bacterial genomes without their association with integron. These GCs are called “solo cassettes” (Hall 2012).

1.3.3 *attC*: the cassette associated recombination sites

The *attC* sites were originally named as 59-base elements (59-be) and identified as an “insertional hot-spot” in plasmids and Tn21 (Cameron, Groot Obbink et al. 1986). Hansson, Sköld et al. (1997) used the term *attC* for 59-be for the first time and this was subsequently adapted in the literature. The presence of an *attC* site is an essential feature of a GC. The *attC* of a GC is recognised by IntI (Hall 2012).
The length of the \textit{attC} sites can vary from 50 to 150 bp (Hansson, Sköld et al. 1997). All \textit{attC} sites detected so far were characterized by the presence of two simple sites for integrase binding, each composed of a pair of conserved “core sites” and named 1L (R’’) and 2L (L’’) (located at 5’-end), 2R (L’) and 1R (R’) (located at 3’-end) (Recchia and Sherratt 2002) (Figure 1-3). Generally, R’’ and R’ are nearly perfect inverted repeats with a consensus sequence of RYYYYAAC and GTTRRRY (where R is a purine, Y is a pyrimidine), respectively. L’’ and L’ are also complementary except for an extra base pair present in L’’ (Partridge, Tsafnat et al. 2009). The central region is highly variable with a size between 20-104 bp (Cambray, Guerout et al. 2010).

\textbf{1.4 Classification of integrons}

Integrons are broadly classified into two groups: mobile integrons (MIs) and chromosomal integrons (CIs), also called superintegrons (SIs) (Fluit and Schmitz 2004). MIs are associated with mobile DNA elements such as transposons or plasmids and usually possess a few cassettes in the cassette array. The MIs typically carry the ARGs which confer resistance to most classes of antibiotics, whereas the CIs contain GCs containing genes encoding proteins predicted to carry out a range of diverse functions (Gillings 2014). A total of 130 different antibiotic resistant GCs have been identified in the MIs (based on 98\% nucleotide identity threshold) (Partridge, Tsafnat et al. 2009, Cambray, Guerout et al. 2010). CIs typically have long arrays of GCs with related \textit{attC} sites that show certain species specificity (Rowe-Magnus, Guerout et al. 2001). However, the CIs carrying a few GCs have also been found. For example, the \textit{Vibrio cholerae} N16961 genome contains a CI with a cassette array of 179 GCs, while
the CIs of *Geobacter metallireducens* contain an array of only 3 GCs (Cambray, Guerout et al. 2010). Unlike class 1 integrons, the association of CIs with MGEs has not been found. However, the existence of CIs carrying ARGs has been reported (Rowe-Magnus, Guerout et al. 2001, Elbourne and Hall 2006). It is been suggested that the resistant genes in the mobile integrons were recruited from the superintegrons (Rowe-Magnus, Guerout et al. 2002).

MIs are further classified into five classes: 1 to 5 based on the divergence of their integrase genes (Rowe-Magnus, Guerout et al. 2003, Gillings 2014). The IntI proteins encoded by different integron types are 34 to 94% identical or 57 to 96% similar in pairwise comparisons (Nield, Holmes et al. 2001). The first three MIs (class 1, 2 and 3) are involved in the development of multidrug resistance (MDR) phenotype of gram-negative pathogens. Among them class 1 and class 2 integrons are the most common in resistant bacteria compared to class 3 integrons and their mobility led to their wide dissemination into many different bacterial species (Collis, Kim et al. 2002). The structural organisation of class 1, 2 and 3 MIs are shown in Figure 1-4.
Figure 1-4. Structure of different class of integrons. A) Class 1 integron in Tn21 (Krin, Cambray et al. 2014) carrying aadA1 cassette encoding streptomycin/spectinomycin resistance where intI and the GCs in the array are oriented in the opposite direction. B) A class 2 integron (Hussein, Ahmed et al. 2009) carrying three resistance GCs. C) A class 3 integron (Collis, Kim et al. 2002).

The classification system of integrons based on similarity of integrase genes is widely used, however, this approach has some serious limitations (Boucher, Labbate et al. 2007). So far it has not been clearly defined what degree of nucleotide sequence identity will be required to be the part of an integron class,
however, similar to differentiate ARGs, a 98% cut-off value is generally used (Boucher, Labbate et al. 2007). The recent discovery and detection of hundreds of different integrons in different environments, has proved that the traditional classification system is insufficient. To avoid discrepancies, Boucher, Labbate et al. (2007) proposed an alternative classification system based on the phylogenetic tree of known integron integrases (Figure 1-5). In this method, three main evolutionary groups of integrons were defined: 1) the soil/freshwater proteobacteria group that includes class 1 and class 3 integrons, 2) the marine γ-proteobacteria group which includes class 2 integrons and the integrons found on SXT (sulphamethoxazole and trimethoprim resistant) integrative conjugative element and pRSV1 plasmid of Vibrio and 3) inverted integrase group where the intI gene is in the same orientation of cassette encoded genes (Boucher, Labbate et al. 2007). The only integron found in the oral cavity (Coleman, Tetu et al. 2004) belongs to the last group. The integrons of inverted integrase group are also called reverse integrons or unusual integrons.
**Figure 1-5.** Phylogenetic tree of known integron integrases (IntI). A single integron IntI is included for each bacterial species, provided that all IntI from that species cluster together in a preliminary analysis. Only species for which most of the GC array has been sequenced are included. Black boxes indicate integrons that are associated with antibiotic resistance GCs, with the particular cassette identified in the box. Class 1, 2 and 3 integrons can contain multiple ARGs. The accession number of each integron integrase is in parentheses next to the taxon name of its host and the number of GCs associated with it is in brackets. The tree and bootstrap support values were inferred by maximum likelihood using PHYML. The different colours represent different families of proteobacteria. The position of IntI of Treponema denticola in the tree has been shown in a red-lined box. The tree and description were adapted from Boucher, Labbate et al. (2007) (p. 304).
1.5 Variants and locations of Pc promoters

Most of the class 1 integrons carry a single promoter called Pc (formerly P\text{ant}), however, in some cases a second active promoter P\text{2} has also been detected (Collis and Hall 1995). In general, the Pc promoter of class 1 and class 3 integrons is located within the coding sequence of the intI and the P\text{2} promoter, if present, is embedded within the sequence downstream of Pc promoter (Lévesque, Brassard et al. 1994, Collis and Hall 1995) (Figure 1-4). Unlike class 1 integron, the Pc promoter for class 2 integron is located within the attI2 (da Fonseca, dos Santos Freitas et al. 2011) (Figure 1-4).

The Pc of the CI of V. cholerae, a well-studied integron, was found to be co-localized with P\text{int} (intI\text{A} promoter) in the opposite orientation in a highly conserved region in between intI\text{A} and attI region (Krin, Cambray et al. 2014). However, the Pc of the CI of Pseudomonas stutzeri is localized at the 5'-end of the intI, resembling Pc of class 1 integrons (Coleman and Holmes 2005). The putative Pc of the CI of T. denticola was detected in silico within the attI (Coleman, Tetu et al. 2004).

The analysis of -35 and -10 hexamer sequences and their relative strength has revealed the existence of four variants (Table 1-2) of Pc promoters in the class 1 integrons: strong promoter (PcS), weak promoter (PcW), hybrid promoter (PcH1) and weak promoter + second promoter or P2 promoter (Lévesque, Brassard et al. 1994). The relative strength of these promoters was analysed by determining the percentage of acetylated chloramphenicol in the reaction as a result of the expression of a chloramphenicol acetyltransferase (cat) gene located downstream of the promoters (Lévesque, Brassard et al. 1994). In another comprehensive study, Jové, Da Re et al. (2010) analysed the complete
sequences of 321 distinct class 1 integrons and detected a total of 13 Pc variants including the four variants identified by Lévesque, Brassard et al. (1994) and a super strong promoter (PcSS) identified by Brízio, Conceição et al. (2006). The prevalence of newly detected promoter and PcSS were very low (a total of 1.6%) compared to the prevalence (98.4%) of the previously described promoters (PcS, PcW, PcH1 and PcW + P2) (Jové, Da Re et al. 2010). The pattern of -35 and -10 sequence of these promoters are shown in Table 1-2.

Table 1-2. Different variants of Pc and P2 promoters detected in class 1 integrons.

<table>
<thead>
<tr>
<th>Promoter</th>
<th>-35 sequence</th>
<th>Spacing</th>
<th>-10 region</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>PcS (Strong)</td>
<td>TTGACA</td>
<td>17</td>
<td>TAAACT</td>
<td>Lévesque, Brassard et al. (1994)</td>
</tr>
<tr>
<td>PcW (Weak)</td>
<td>TTGACA</td>
<td>17</td>
<td>TAAGCT</td>
<td>Lévesque, Brassard et al. (1994)</td>
</tr>
<tr>
<td>PcH1 (Hybrid)</td>
<td>TGGACA</td>
<td>17</td>
<td>TAAACT</td>
<td>Lévesque, Brassard et al. (1994)</td>
</tr>
<tr>
<td>PcSS (Super-strong)</td>
<td>TTGATA</td>
<td>17</td>
<td>TAAACT</td>
<td>Brízio, Conceição et al. (2006)</td>
</tr>
<tr>
<td>PcIn42*</td>
<td>TTGGCA</td>
<td>17</td>
<td>TAAACT</td>
<td>Jové, Da Re et al. (2010)</td>
</tr>
<tr>
<td>PcIn116*</td>
<td>TTGACA</td>
<td>17</td>
<td>TAAACT</td>
<td>Jové, Da Re et al. (2010)</td>
</tr>
<tr>
<td>PcPUO*</td>
<td>TCGACA</td>
<td>17</td>
<td>TAAACT</td>
<td>Jové, Da Re et al. 2010)</td>
</tr>
<tr>
<td>P2 (second promoter)</td>
<td>TTGGTA</td>
<td>17</td>
<td>TACAGT</td>
<td>Lévesque, Brassard et al. (1994)</td>
</tr>
<tr>
<td>P2m1 (mutated)</td>
<td>TTGGTA</td>
<td>17</td>
<td>GACAGT</td>
<td>Jové, Da Re et al. (2010)</td>
</tr>
<tr>
<td>P2m2 (mutated)</td>
<td>TTGGTA</td>
<td>17</td>
<td>TACACA</td>
<td>Jové, Da Re et al. (2010)</td>
</tr>
</tbody>
</table>

*PcIn42, PcIn116 and PcPUO, as they are carried by integrons In42 and In116 and by plasmid pUO901, respectively (Jové, Da Re et al. 2010).

It has also been shown that the strength of the Pc promoters correlates with integrase excision activity. A weak promoter was found to be associated with a strong integrase (Jové, Da Re et al. 2010).
1.6 Expression of *intI* gene: Role of the SOS response and cAMP-CRP complex

The expression of integron *intI* by the integrase promoter, Pint is controlled by a transcriptional repressor protein known as LexA (Guerin, Cambray et al. 2009) as well as the cyclic AMP (cAMP)-cAMP receptor protein (CRP) (cAMP-CRP) complex (Baharoglu, Krin et al. 2012). LexA governs the SOS response which is a global regulatory network responsible for addressing DNA damage by repairing or bypassing DNA lesions (Erill, Campoy et al. 2007). The cAMP-CRP complex is also a global regulator which is involved in activation of transcription at promoters of a vast number of genes. The CRP (also known as catabolite activator protein, CAP) functions by binding, in the presence of the allosteric effector cAMP, to specific DNA sites in or near target promoters and enhancing the ability of RNA polymerase holoenzyme (RNAP) to bind and initiate transcription (Lawson, Swigon et al. 2004). LexA represses SOS genes by binding to a highly specific 16 bp long binding site (known as LexA box) located upstream of the promoter regions of the SOS genes (Walker 1984). In *E. coli* and most β- and γ-proteobacteria, the LexA binding site consist of a palindromic structure: CTGTatatatatACAG (Walker 1984) (Figure 1-6). When a cell’s DNA is damaged by environmental stresses, single-stranded DNA (ssDNA) fragments are produced which bind non-specifically to the universal recombination protein RecA. The resulting RecA-ssDNA nucleofilaments activates the proteolytic activity of RecA, thus the cleavage of the LexA repressor protein is triggered (Sassanfar and Roberts 1990). Once cleaved,
LexA is unable to bind at the LexA box, which allows the expression of the SOS regulon.

**Figure 1-6.** Alignment of the promoter regions of intI genes from the CI of V. cholerae (Vch), V. metschnikovii (Vme), V. parahaemolyticus (Vpa), and V. natriegens (Vna) and from class 1 (cl1-MI), class 2 (cl2-MI), class 3 (cl3-MI), and class 5 (cl5-MI) MIs (1). Putative LexA-binding sequences are boxed, whereas putative σ70 promoter elements (–35 and –10) are underlined and the translation start site of intI is boxed in red. The location of the LexA motif within a class 1 integron (on the top) is shown. The alignment section of the figure was adapted from Guerin, Cambray et al. (2009).

The role of the SOS response in controlling intI expression was first discovered by Guerin, Cambray et al. (2009). By aligning the upstream region of intI of different CI and MI, they identified a conserved LexA binding motif overlapping the putative Pint promoter regions (Figure 1-6). Later on, Cambray, Sanchez-Alberola et al. (2011) found that putative LexA regulation is a widespread phenomenon in integrons, as they detected E. coli-like LexA binding motifs in almost all Vibrionaceae super integrons and all but one of the mobile integrons.
By using electro-mobility shift assays (EMSA), it was confirmed that the LexA binding motif of CI of *V. cholerae* can bind to LexA (Guerin, Cambray et al. 2009). They also found that when the SOS response is induced by treating the cells with antibiotics including ciprofloxacin, mitomycin, trimethoprim and ampicillin, the expression of a β-galactosidase reporter of integrase is increased by 4.5-fold in *E. coli* and 37-fold in *V. cholerae*. Additionally, the expression of *intI* upon SOS induction increases and the rate of excision of GCs of class 1 MI and CI of *V. cholera* increased by 141-fold and 340-fold, respectively (Guerin, Cambray et al. 2009). In another study, the induction of SOS response by ssDNA was confirmed in *V. cholerae* A1552 and it was also shown that along with SOS response, the expression of *intI* is regulated by the CRP-dependent catabolite repression pathway (Baharoglu, Krin et al. 2012). These findings demonstrate that the regulation of *intI* is strictly dependent on the SOS response and that SOS induction controls the rates of cassette recombination. This also suggests that under stressful conditions, SOS regulation enhances cassette swapping and capture, while stabilizing the integron in stable environments (Cambray, Sanchez-Alberola et al. 2011). The regulation of *intI* expression by cAMP-CRP complex indicates that the extracellular environment influence on chromosomal gene content.

### 1.7 Expression of cassette-located genes and its control by cAMP-CRP complex

GCs are usually promoterless, so, the cassette genes must rely on the Pc promoter for their expression (Collis and Hall 1995), however, a few cassettes were identified in MIs carrying their own promoters including *sul1* (Guerineau,
Brooks et al. 1990), cmlA (Bissonnette, Champetier et al. 1991) and ereA (Biskri and Mazel 2003). The toxin-antitoxin cassettes present in the CI also carry their own promoters (Guerout, Iqbal et al. 2013). Collis and Hall (1995) investigated the expression of ARGs encoded in class 1 integrons. Their study showed that the expression of cassette-encoded ARGs is mediated by the common Pc promoter and the level of resistance is affected by mutations in the Pc promoter sequence. They also found that all of the transcripts begin at Pc, however, transcripts originating at P2 (second promoter of class 1 integron) are also found when P2 is associated with a weak Pc. They observed that the level of antibiotic resistance is largely affected by the position of the ARGs in the cassette array. When the ARGs were in the first cassette, the level of resistance was found to be the highest. For instance, when the aadA2 cassette which confers resistance to streptomycin is relocated from the first position of the cassette array (aadA2-aacC1-orfE-cmlA) to third position (aacC1-orfE-aadA2-cmlA), the IC_{50} value of streptomycin changed sharply from 1,120 µg/mL to 60 µg/mL. Another important finding from the Northern blot analysis was the identification of different size transcripts starting from Pc and large transcripts carry the distal genes only. They suggested that early transcription termination occurs within the GCs and perhaps the attC sites function as transcription terminators along with its role in GC recombination (Collis and Hall 1995).

The expression of the genes encoded by the cassettes in CIs carrying hundreds of GCs were first studied by Michael and Labbate (2010). They studied the expression of genes in the CI of *Vibrio spp.* DAT722 carrying 116 GCs in the array by RT-PCR using primers designed from attC sites. The location of the internal or intra-array promoters were predicted by comparing the PCR results
on cDNA and genomic DNA prepared from the same culture used for RNA preparation. They found that the majority of the detectable cassette size classes (39 of 62 size classes) were expressed and those expressed cassettes were distributed throughout the array. The cassettes were expressed in different ‘blobs’ with a similar level of expression within a blob but different level of expression with other blobs. These observations suggested that along with the Pc, there are intra-array promoters that catalyse the expression of genes located on distal GCs and the ability of these intra-array promoters to express the GCs were different (Michael and Labbate 2010).

Although it is well-established that the activity of Pint promoter which mediates the expression of intI is tightly controlled by the SOS regulon and cAMP-CRP complex (Guerin, Cambray et al. 2009, Cambray, Sanchez-Alberola et al. 2011), the control of the Pc promoter was not known until 2014. Krin, Cambray et al. (2014) showed that the expression of GCs in V. cholerae integron is controlled by the catabolite repression cAMP-CRP complex, but not by the SOS response. The regulatory region of intI and cassette array is shown in Figure 1-7. Both the Pc promoter and Pint were found to be co-regulated and induced in rich medium, at high temperature, high salinity and at the end of the exponential growth phase (Krin, Cambray et al. 2014).
Figure 1-7. The regulatory regions of intI and cassette array are shown in the alignment of V. cholerae and V. mimicus strains. The intIÁ gene sequence is shown in bold italicized characters. Regulator binding site are framed. -10 and -35 promoter boxes are shown in grey. The attI recombination point is indicated by a vertical bar. The figure was taken from Krin, Cambray et al. (2014) (Supplementary figure S1).

1.8 Recombination reactions in integron-gene cassette systems

The recombination mechanism in integron-gene cassette systems have been studied extensively (Collis and Hall 1992, Collis, Grammaticopoulos et al. 1993, Hall and Collis 1995). There are three types of site-specific recombination reactions possible: attIxattC, attCxattC and attIxattI (arranged here in order of decreasing efficiency) (Collis, Recchia et al. 2001). The efficiency of these reactions were determined by analysing the formation of cointegrates and the change in sensitivity of the hosts to trimethoprim (Collis, Recchia et al. 2001). The attIxattC recombination was the most efficient. In this recombination, a circular form of a GC is inserted in the attI site catalysed by Intl. Recombination occurs at the R site of attI between G and T residues of the canonical
G↓TTRRRY sequence (Collis, Grammaticopoulos et al. 1993, Collis, Recchia et al. 2001).

The recombination between two attC (attCxattC) is also possible when there are one or more GCs already present in the cassette array, however, this recombination reaction was found to be far less efficient than attIxattC recombination (Collis, Recchia et al. 2001). The recombination of two attC sites is probably responsible for excision of a gene cassette as a circular form (Collis and Hall 1992) (Figure 1-8).

The recombination between two attI sites (attIxattI) is the least efficient among the three recombination systems in the integron. The recombination efficiency of attIxattI recombination was found to be at least 10-fold lower than the attIxattC recombination reaction (Collis, Recchia et al. 2001).
Figure 1-8. Model of insertion and excision of circular GCs into integron platforms by site-specific recombination. The 5'-CS contains the core integron consisting of intI, attI and Pc and the 3'-CS contains the qacEΔ1 and sul1. The first event of insertion of a circular GC is an example of attIxattC recombination and the second insertion is an example of attCxattC recombination.

1.9 Bacteria carrying integrons

More than 10% of bacterial genomes including the partially or completely sequenced genomes carry integrons (Boucher, Labbate et al. 2007). Although there is no report of the presence of integrons in archaea, Elsaied, Stokes et al.
(2014) identified some cassette proteins having very good homologies with the archaeal genomes.

The INTEGRALL database (Moura, Soares et al. 2009) (which can be accessed at http://integrall.bio.ua.pt/?) is a comprehensive database which collects and organises information on integrons. To date (17 June 2017), INTEGRALL comprises more than 8,000 gene cassette sequences from 382 bacterial species which belong to 171 genera. In 2009, the number of integron related sequences was 4800 which indicates the rapid increase in entry into the database in the past 7 years which may be due to wide use of high throughput sequencing technologies. As of 2009, 70% and 27% of integron-related sequences in the INTEGRALL were related to uncultured bacteria and the phylum γ-Proteobacteria, respectively and the remaining 3% included the representative from the phylum α-, β-, δ- and ε-Proteobacteria, Firmicutes, Cyanobacteria, Actinobacteria, Chlamydiae/Verrucomicrobia group, Bacteroidetes/Chlorobi group, Spirochaetes and Planctomycete (Moura, Soares et al. 2009). This indicates that integrons are found in a diverse range of hosts, however, the reason for their higher prevalence in γ-Proteobacteria may be linked with the number of pathogens (such as Escherichia, Shigella, Salmonella, Pseudomonas, etc.) in this phylum.

1.9.1 Class 1 integron carrying bacteria

Class 1 integrons are most common in clinically important pathogens and they have been playing a central role in the worldwide dissemination of antibiotic resistance. van Essen-Zandbergen, Smith et al. (2007) carried out a study to detect class 1, 2 and 3 integrons in a subset of MDR pathogens including Salmonella spp. isolated from humans and animals and Campylobacter spp.
and *E. coli* isolated from broilers. They found that 76% of the *E. coli* and 43% of the *Salmonella* isolates were positive for class 1 integrons. Whereas the class 2 integrons were found in 11% of the *E. coli* and 1% of the *Salmonella* isolates and no class 3 integrons were detected in any of the bacterial isolates tested (van Essen-Zandbergen, Smith et al. 2007). In another screening study to detect class 1 integrons in the isolates collected from different European hospitals, it was found that 71.4% of *Klebsiella oxytoca*, 62.0% *E. coli*, 44.0% *Enterobacter aerogenes* and 30.8% of *K. penumoniae* isolates were positive for class 1 integrons (Martinez-Freijo, Fluit et al. 1998). Li, Hu et al. (2013) investigated the presence of class 1 integrons and associated GCs in a sample of 176 *K. pneumoniae* isolated from a tertiary-care hospital in Beijing. They found that 51.1% (90/176) of the isolates were positive for class 1 integrons.

Table 1-3 represents some gram-negative bacteria frequently found to be associated with class 1 integrons. It can be seen from the Table 1-3 that *dfr* and *aadA* cassettes that confer resistance to trimethoprim and streptomycin-spectinomycin, respectively are predominantly present in class 1 integrons.

<table>
<thead>
<tr>
<th>Host</th>
<th>Resistance gene cassettes</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Aeromonas salmonicida</em></td>
<td><em>aadA2, aadA1, dfrIic</em></td>
<td>(L'Abée-Lund and Sørum 2001, Schmidt, Bruun et al. 2001)</td>
</tr>
<tr>
<td><em>Achromobacter xylosoxydans</em></td>
<td><em>blaVIM-1, aacA4, aphA15, aadA1</em></td>
<td>(Riccio, Pallecchi et al. 2001)</td>
</tr>
<tr>
<td><em>Burkholderia cepacia</em></td>
<td><em>oxa, aac(6')-1a</em></td>
<td>(Crowley, Daly et al. 2002)</td>
</tr>
<tr>
<td>Organism</td>
<td>Gene(s)</td>
<td>Reference</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>Campylobacter jejuni</td>
<td>dfr1</td>
<td>(Gibreel and Sköld 2000)</td>
</tr>
<tr>
<td>Citrobacter freundii</td>
<td>dfrXII, ant(3&quot;)-l (aadA)</td>
<td>(Nørskov-Lauritsen, Sandvang et al. 2001)</td>
</tr>
<tr>
<td>Enterobacter cloacae</td>
<td>blaICB-1</td>
<td>(Kartali, Tzelepi et al. 2002)</td>
</tr>
<tr>
<td>Enterobacter aerogenes</td>
<td>aacA4-arr-3-dfrA27- aadA16</td>
<td>(Cheng, Sun et al. 2016)</td>
</tr>
<tr>
<td>E. coli</td>
<td>dfrIa, dfrV, dfr12, dfr17, aadA1a, aadA2, aadA4, aadB, cmlA ereA2, aadA7, dfrA12, orfF and aadA2, dfrA1, dfrA5, dfrA7, dfrA12, dfrA17, aadA1, aadA, aadA5, aadA2</td>
<td>(Chang, Chang et al. 2000) (Mazel, Dychinco et al. 2000) (Sunde 2005) (Vinue, Saenz et al. 2008)</td>
</tr>
<tr>
<td>K. oxytoca</td>
<td>aacA4-arr-3-dfrA27- aadA16</td>
<td>(Cheng, Sun et al. 2016)</td>
</tr>
<tr>
<td>K. pneumoniae</td>
<td>blaKPC-2, blaVIM-2, aadA2</td>
<td>(Falco, Ramos et al. 2016)</td>
</tr>
<tr>
<td></td>
<td>blaOXA-1, blaOXA-1, blaSHV-1, aac(6(^\prime))-lb-c</td>
<td>(Pérez-Moreno, Estepa et al. 2012)</td>
</tr>
<tr>
<td>Proteus mirabilis</td>
<td>blaPSE-1</td>
<td>(Cheng, Sun et al. 2016)</td>
</tr>
<tr>
<td>P. aeruginosa</td>
<td>aadA1, aadA6</td>
<td>(Severino and Magalhães 2002)</td>
</tr>
<tr>
<td></td>
<td>blaVIM-2, cmlA6, catB11, blaGES-7, aacA7, aacA4</td>
<td>(Papa Ezdra, Bado et al. 2016)</td>
</tr>
<tr>
<td>S. enterica Serovar Typhi</td>
<td>dfr7, aacA4, blaOXA-1</td>
<td>(Lee, Yong et al. 2004)</td>
</tr>
<tr>
<td></td>
<td>dfrA12, aadA27</td>
<td>(Ahmed, El-Hofy et al. 2016)</td>
</tr>
<tr>
<td>Serratia marcescens</td>
<td>blaMP-10, aacA31, aadA1</td>
<td>(Silva, Cayo et al. 2015)</td>
</tr>
<tr>
<td>Shigella flexneri</td>
<td>dfrA5, aadA1, dfrA1, dfrA17, aadA5, aacA4 and cmlA</td>
<td>(Xu, Zhuang et al. 2016)</td>
</tr>
<tr>
<td>Stenotrophomonas maltophilia</td>
<td>qacl, aadB, cmlA, aad2</td>
<td>(Huang, Hu et al. 2015)</td>
</tr>
<tr>
<td>V. cholerae O1</td>
<td>aadA1</td>
<td>(Dalsgaard, Forslund et al. 1999)</td>
</tr>
</tbody>
</table>

Abbreviations: VIM, Verona Integron-mediated metallo-beta-lactamase; IMP, imipenemase; aad, aminoglycoside adenyl transferase; dfr, dihydrofolate reductase; aph, aminoglycoside phosphotransferase; OXA, oxacillinase; KPC, K. pneumoniae carbapenemase, cmlA, chloramphenicol acetyltransferase, ereA2, erythromycin esterase; CTX-M, cefotaximase; TEM, temoneria; SHV, sulphhydryl variable; CARB, carbenicillinase.
Between 1989 (the year of the first report of integrons) and 1998, all of the integrons that were reported were found in gram-negative bacteria. In 1998, for the first time, Nesvera, Hochmannova et al. (1998) reported the detection of a class 1 integron (Figure 1-9) in gram-positive bacterium, Corynebacterium glutamicum which was identical to the integron, InC of P. aeruginosa present on the plasmid pSA1700. Since then several other gram-positive bacterial genera including Enterococcus (Clark, Olsvik et al. 1999, Shi, Zheng et al. 2006), Arthrobacter (Agersø and Sandvang 2005) and Streptococcus (Shi, Zheng et al. 2006) were found to carry class 1 integrons.

**Figure 1-9.** The structure of the class 1 integron detected on the plasmid pCG4 of gram-positive bacterium Corynebacterium glutamicum. The white box represents the integrated GC and black boxes the 5'- and 3'-conserved segments (CS). Numbers below the vertical arrows represent positions of the base pairs differing from those of integron InC. Pant (now named as Pc) indicates the integron promoter involved in expression of the streptomycin/spectinomycin resistance gene associated with the cassette. The following genes are indicated: int, site-specific integrase; aadA2a, streptomycin/spectinomycin resistance; qacEΔ1, antiseptics resistance; sul1, sulfonamide resistance; ORF5, unknown function. The figure was adapted from Nesvera, Hochmannova et al. (1998).

### 1.9.2 Class 2 integron carrying bacteria

The prevalence of class 2 is much less than class 1 integrons in clinical and environmental bacteria. Class 2 integrons have been reported in Burkholderia cenocepacia (Ramírez, Vargas et al. 2005), S. marcescens (Crowley, Cryan et al. 2008), S. enterica serovar Typhimurium (Macedo-Vinas, Cordeiro et al.
2009), E. coli (Solberg, Ajiboye et al. 2006, Vinue, Saenz et al. 2008), Enterobacteriaceae (E. coli, K. pneumoniae, C. freundii, E. cloacae) (Machado, Coque et al. 2008, Ramírez, Piñeiro et al. 2010) as well as A. baumannii (Ramírez, Piñeiro et al. 2010). In a survey to detect class 2 integrons, a set of 726 bacterial isolates from environmental and clinical samples were analysed, it was found that 17.9% (130/726) of the isolates were positive for class 2 integrons (Ramírez, Piñeiro et al. 2010). A total of 32 arrays of GCs which were grouped into eight types of class 2 integrons were detected (Figure 1-10) and among them 6 types (In2-0, In2-2, In2-3, In2-5, Tn7::IS1, and In2-6::IS26) were described for the first time.

**Figure 1-10.** Schematic representation of arrays of class 2 integrons found in the bacterial isolates of clinical and environmental samples (n = 126). The thin vertical closed bar represents the attI2 site, and the ovals represent the attC sites of the GCs. All intI2 genes from these arrays were sequenced, and they revealed the usual internal stop codon. The figure and description were adapted from Ramírez, Piñeiro et al. (2010) (p.703).
Class 2 integrons are commonly found to be associated with Tn7. Another important feature of these integrons is that the IntI2 is usually inactive due to an internal stop codon (at amino acid position 179), however, an unusual type of class 2 integron with a functional IntI2 in *Providencia stuartii* has also been reported (Barlow and Gobius 2006). In some strains such as *A. baumannii* BM4431 (Ploy, Denis et al. 2000), both class 1 and 2 integrons were found to co-exist.

### 1.9.3 Class 3 integron carrying bacteria

Class 3 integrons are less frequently detected in bacterial isolates compared to class 1 and class 2 integrons. Class 3 integrons carrying *bla*IMP and *aacA* GCs were reported for the first time in 1995 in a strain of carbapenem-resistant *S. marcescens* (Arakawa, Murakami et al. 1995). Their presence has also been detected in *E. cloacae* (Barraud, Casellas et al. 2013), *E. coli* (Kargar, Mohammadalipour et al. 2014), *K. pneumoniae* (Correia, Boavida et al. 2003, Jones-Dias, Manageiro et al. 2016), *Acinetobacter johnsonii*, *Aeromonas allosaccharophila* and *C. freundii* (Jones-Dias, Manageiro et al. 2016). The structure of a class 3 integron detected in *K. pneumoniae* is shown in Figure 1-11. It can be seen from the Table 1-2 that all of the class 3 integrons carry a metallo-beta-lactamase gene in the cassette array.
Table 1-4. Some representative bacteria carrying class 3 integrons

<table>
<thead>
<tr>
<th>Host</th>
<th>Resistance gene cassettes</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. marcescens</td>
<td>bla_{IMP}, aac</td>
<td>(Arakawa, Murakami et al. 1995)</td>
</tr>
<tr>
<td>E. cloacae</td>
<td>bla_{OXA-256}, aac(6')-lb</td>
<td>(Barraud, Casellas et al. 2013)</td>
</tr>
<tr>
<td>K. pneumoniae</td>
<td>bla_{GES-1}, bla_{OXA-10'-aac(6')-lb}</td>
<td>(Correia, Boavida et al. 2003)</td>
</tr>
<tr>
<td>K. pneumoniae</td>
<td>dfrB3, bla_{GES-11}</td>
<td>(Jones-Dias, Manageiro et al. 2016)</td>
</tr>
</tbody>
</table>

Figure 1-11. Schematic representation of a class 3 integron detected in a K. pneumoniae isolate. The figure was adapted from Jones-Dias, Manageiro et al. (2016) (p.10).

It is found that the same GCs are present in different classes of integrons. All of the GCs of class 2 and class 3 integrons have also been found in class 1 integrons. This is due to the fact that integron IntIs can recognize different types of attC sites (non-cognate) which allows cassettes to move from one integron class to another (Hall 2012).

1.9.4 Chromosomal integron (CI) carrying bacteria

The 126-kb long SI of V. cholerae carrying at least 179 cassettes with 179 attC sites or V. cholera repeats (VCRs) (Rowe-Magnus, Guérout et al. 1999) were first reported in 1998 (Mazel, Dychinco et al. 1998), although the presence of gene-VCRs in V. cholerae O1 genome were reported earlier (Barker, Clark et
al. 1994). Retrospective analysis of the presence of integrons in other Vibrio isolates indicates that they are present in a different Vibrio species such as V. mimicus, V. anguillarum, V. fischerii (Rowe-Magnus, Guérout et al. 1999) as well as in a strain of V. metschnikovii isolated in 1888 (Mazel, Dychinco et al. 1998). CIs were also detected in Pseudomonas spp. (Vaisvila, Morgan et al. 2001), S. oneidensis (Drouin, Melancon et al. 2002), Xanthomonas spp. (Rowe-Magnus, Guerout et al. 2001) and Treponema denticola (Coleman, Tetu et al. 2004). Table 1-5 lists the bacterial species known to carry chromosomal integrons.

**Table 1-5. Bacterial species that were identified to carry chromosomal integrons**.

<table>
<thead>
<tr>
<th>Bacterial group</th>
<th>Bacterial strain</th>
<th>Number of gene cassettes</th>
<th>Typical length of attC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>γ-proteobacteria</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vibrionaceae</td>
<td>Vibrio vulnificus CMCP6</td>
<td>217</td>
<td>126–129</td>
</tr>
<tr>
<td></td>
<td>Vibrio cholerae N16961</td>
<td>179</td>
<td>126–129</td>
</tr>
<tr>
<td></td>
<td>Vibrio sp. DAT722</td>
<td>116</td>
<td>126–128</td>
</tr>
<tr>
<td></td>
<td>Vibrio parahaemolyticus RIMD2210633</td>
<td>69</td>
<td>125–128</td>
</tr>
<tr>
<td></td>
<td>Vibrio alginiticus 12G1</td>
<td>51</td>
<td>126–129</td>
</tr>
<tr>
<td></td>
<td>Vibrio fischemi ES114</td>
<td>38</td>
<td>125–127</td>
</tr>
<tr>
<td></td>
<td>Vibrio splendidus LGP32</td>
<td>25</td>
<td>126–129</td>
</tr>
<tr>
<td>Pseudoalteromonas</td>
<td>Pseudoalteromonas tunicata D2</td>
<td>7</td>
<td>78–129</td>
</tr>
<tr>
<td></td>
<td>Pseudoalteromonas haloplanktis TAC125</td>
<td>5</td>
<td>98–102</td>
</tr>
<tr>
<td>Xanthomonadaceae</td>
<td>Xanthomonas campestris pv. campestris ATCC 33913</td>
<td>22</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Xanthomonas campestris pv. vesicatoria</td>
<td>3</td>
<td>79–94</td>
</tr>
<tr>
<td>Pseudomonadaceae</td>
<td>Pseudomonas stutzeri Q</td>
<td>&gt;7</td>
<td>76–77</td>
</tr>
<tr>
<td></td>
<td>Pseudomonas alcaligenes</td>
<td>32</td>
<td>76–90</td>
</tr>
<tr>
<td>Shewanellaceae</td>
<td>Shewanella sp. MR-7</td>
<td>2</td>
<td>90–92</td>
</tr>
<tr>
<td></td>
<td>Shewanella denitrificans OS217</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>Bacterial group</td>
<td>Bacterial strain</td>
<td>Number of gene cassettes</td>
<td>Typical length of attC</td>
</tr>
<tr>
<td>----------------------</td>
<td>-------------------------------</td>
<td>--------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Alteromonadaceae</td>
<td>Saccharophagus degradans 2–40</td>
<td>73</td>
<td>111–140</td>
</tr>
<tr>
<td>β-proteobacteria</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nitrosomonas europaea ATCC19718</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Nitrosomonas eutropha C71</td>
<td>3</td>
<td>60–120</td>
</tr>
<tr>
<td></td>
<td>Azoarcus sp. EbN1</td>
<td>3</td>
<td>59–73</td>
</tr>
<tr>
<td></td>
<td>Rubrivivax gelatinosus PM1</td>
<td>2</td>
<td>70–126</td>
</tr>
<tr>
<td>δ-proteobacteria</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Geobacter metallireducens GS-15</td>
<td>3</td>
<td>59–63</td>
</tr>
<tr>
<td>Planctomycetes</td>
<td>Rhodopirellula baltica SH1</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>Spirochaetales</td>
<td>Treponema denticola ATCC35405</td>
<td>45</td>
<td>63–68</td>
</tr>
</tbody>
</table>

*The information on the table was taken from Cambray, Guerout et al. (2010) (p.145).

Unlike MIs, the attC sites of CIs in the cassette array are very similar to each other and they are species specific (Mazel 2006). It has been hypothesized that MIs evolved from CIs by entrapping intI genes and their cognate attI sites into highly mobile structures like transposons (Rowe-Magnus, Guerout et al. 2001). This hypothesis was based on two observations: i) the cassettes of SI of V. cholerae were found to be the substrates for IntI1 and ii) the attC sites of 12 different resistance cassettes including the blaCARB and dfr6 cassettes were identical to the attC sites located in the SIs of Xanthomonas and Vibrio species (Mazel, Dychinco et al. 1998, Mazel 2006). However, the origins of cassette genes and attC recombination sites are still to be determined.

1.10 Detection of integrons

The methods used for detection of integrons can be divided into the following approaches: i) Sequence-based, ii) PCR-based, and iii) capturing of functional
GCs by Intl-mediated recombination. Among these methods, the sequence and PCR-based methods are widely used to detect and screen for the presence of integrons in clinical and environmental samples.

### 1.10.1 Sequence-based

In this approach, the genome sequence of bacteria is used to identify integrons and associated GCs. The genetic environment of integrons can also be easily recovered. This method is applicable for both the pure DNA of a single isolate or the metagenomic DNA of a mixed population of bacteria.

There are some programs and tools available to identify integron cassettes in the genome sequence which are listed in Table 1-6.

**Table 1-6. The programs and tools available to detect integron and associated GCs in the gene sequence**

<table>
<thead>
<tr>
<th>Program/Tool</th>
<th>Application</th>
<th>Access</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>XXR program</td>
<td>attC sites of <em>Vibrio</em> integrons</td>
<td>-</td>
<td>Rowe-Magnus, Guerout et al. (2003)</td>
</tr>
</tbody>
</table>

### 1.10.2 PCR-based

This is a quick and cost-effective way to detect integrons from clinical isolates as well as environmental DNA. There are four different ways to do it: i) Gene
cassette array PCR using primers targeting conserved segments (CS) of 5'- and 3'-ends of the integrons, ii) attC-PCR/GC PCR using primers targeting conserved regions of attC, iii) IntI-attC PCR using a primer from intI gene of 5'-CS and another primer from attC and iv) intI-PCR using primers targeting a specific class of intI. The approximate locations of the widely used primers to detect integrons are shown in Figure 1-13.

1.10.2.1 Gene cassette array PCR

The whole array of GCs of class 1, class 2 and 3 integrons can be PCR amplified using one primer from the 5'-CS and another primer from 3'-CS. The amplified product is sequenced to detect the genes and associated attC sites in the cassette array.

Use of PCR to detect class 1 integrons and the associated cassettes was first reported in 1995 (Levesque, Piche et al. 1995). Widespread dissemination of integrons in the then recently isolated clinical strains of Enterobacteriaceae and Pseudomonas spp. using a set of primers namely 5'-CS and 3'-CS (Levesque, Piche et al. 1995) was found. An image of the agarose gel electrophoresis from this report is shown in Figure 1-12. Since then these primer combinations have been widely used for detection of class 1 integrons in clinical isolates of human and animal sources as well as environmental DNA which is evidenced by more than 1,000 citations of this article. Another primer combination namely hep58 and hep59 are also commonly used for amplifying the whole cassette array of class 1 integrons (White, McIver et al. 2000).
Figure 1-12. Agarose gel electrophoresis of the variable size PCR products obtained using 5'-CS and 3'-CS primers. Lane 1, 1-kb DNA ladder; lane 2, K. pneumoniae 154; lane 3, P. aeruginosa 702; lane 4, E. aerogenes 177; lane 5, E. cloacae 588; lane 6, S. marcescens 616; lane 7, S. marcescens 946; lane 8, S. marcescens 947; lane 9, S. marcescens 909; lane 10, S. typhimurium 101; lane 11, E. coli 801; lane 12, Proteus mirabilis 820. The figure was adapted from (Levesque, Piche et al. 1995).

The primer combinations hep74, which binds to attI2, and hep51 which binds to orfX (Figure 1-13) have been used by many researchers for detecting the cassette carry of class 2 integron (White, McIver et al. 2001).
1.10.2.2 \textit{AttC-PCR/gene-cassette PCR}

In this approach, primers are designed to target the conserved \textit{attC} sites. The variable size PCR products are sequenced to detect the genes and associated \textit{attC} sites. This PCR strategy was first reported by Stokes, Holmes et al. (2001). Using a novel set of degenerate \textit{attC} primers namely HS286 and HS287 they detected a myriad of full-length novel genes in the environmental DNA samples of which majority had no known functions (Stokes, Holmes et al. 2001). This primer combination is commonly used to investigate environmental GCs (EGC) (Holmes, Gillings et al. 2003, Nemergut, Martin et al. 2004). Very recently a novel high-throughput PCR amplicons sequencing method has been developed to analyse the integron GCs in environmental sample using the HS286 and
HS287 primers (Gatica, Tripathi et al. 2016). The approximate location of HS286 and HS287 primers are shown in Figure 1-13. The attC-PCR can also be used for strain typing (Tokunaga, Yamaguchi et al. 2010). It has been successfully used to differentiate the very closely related strains of *V. cholera* (Labbate, Boucher et al. 2007, Tokunaga, Yamaguchi et al. 2010).

### 1.10.2.3 *IntI*-attC PCR

This approach is used to recover the complete attI site of integrons along with the first GCs when the 3’-CS of integrons are not known. The most common approach is designing the forward primer based on the conserved regions of *intI* gene and reverse primer from the left hand simple site of attC. This strategy was first used by Nield, Holmes et al. (2001) to recover integrons from environmental DNA using HS298 designed based on the conserved C-terminal sequence in *IntI* and HS286 designed based conserved sequences in attC (Nield, Holmes et al. 2001). Using this method, they identified several new classes of integrons in soil bacteria. A similar strategy was used in many other studies (Elsaied, Stokes et al. 2007, Elsaied, Stokes et al. 2011, Elsaied, Stokes et al. 2014).

### 1.10.2.4 *intI*-PCR

The presence of integrons in samples can be detected only by targeting the presence of *intI*. Either the degenerate primers targeting *intI1*, *intI2* and *intI3* in a single reaction (White, Mclver et al. 2000) or different primers for distinct class of *intIs* can be used (Ploy, Lambert et al. 2000, Goldstein, Lee et al. 2001, Nemergut, Martin et al. 2004). Using degenerate integrase primers novel *intIs* can be recovered from environmental bacteria (Nemergut, Martin et al.)
Using multiplex PCR the presence of intI1-IntI3 can be detected in a single reaction (Ren, Zhao et al. 2013).

**Figure 1-14.** Multiplex PCR amplification of the class 1, 2 and 3 integrons (Lane 1 and 2). PCR bands show that the amplification products of class 1, 2 and 3 integrons were 280, 788 and 979 bp, respectively. The figure was adapted from Ren, Zhao et al. (2013).

The intI-PCR provides quick information of the presence of integrons in bacterial isolates. Once it is confirmed that the isolates are positive for integrons, subsequent PCR can be done to detect the presence of GCs in the cassette array by using GC-array PCR or attC-PCR or intI-attC PCR depending on the objective of the study and suitability of the PCR primers.

### 1.10.3 Capturing of functional GCs by IntI-mediated recombination

This method of detection of functional GCs of integrons has been described by Rowe-Magnus (2009). This system used the ability of IntI to integrate the functional GCs carried on a substrate plasmid with an integron platform carried on a conjugative R plasmid. The insertions of GCs in the R plasmid can be sequenced by using specific primers (Rowe-Magnus 2009). Due to the
difficulties in carrying out of this process and low throughput, this method did
not receive a lot of attention by the researchers.

1.11 Investigation of ARGs using metagenomics

Metagenomics is the study of nucleic acids obtained directly from the microbial
inhabitants of a particular environment bypassing the need for isolation or
cultivation (Chen and Pachter 2005). The total DNA obtained from a particular
environment is called metagenomic DNA or the metagenome. The term
"metagenome" was first used by Handelsman, Rondon et al. (1998).

Although culture-based studies have characterized most of the resistance
mechanisms, to understand the role of the vast majority of not-yet-cultivable
bacteria in antibiotic resistance, the use of metagenomics is very important
(Riesenfeld, Goodman et al. 2004). Our understanding about the distribution
and diversity of ARGs has advanced a lot by the application of metagenomics
(Sukumar, Roberts et al. 2016).

There are three different metagenomic approaches to study the reservoir of
ARGs: PCR-based, sequence-based and function-based (Figure 1-15). In the
PCR-based approach, the PCR primers are designed to target specific
resistance genes in the metagenome. Koike, Krapac et al. (2007) used a
combination of normal PCR and real-time PCR to detect and quantify the
presence of tetracycline resistance genes in groundwater and lagoon samples
(Figure 1-15). The major disadvantage of the PCR -based approach is that it is
a low throughput technique and the application is restricted to specific DNA
targets (Mullany 2014, Sukumar, Roberts et al. 2016).

The sequence-based approach is a high throughput process and can be used
to detect and quantify ARGs in a sample. In this technique, the total DNA is
extracted directly from an environmental or clinical sample and sequenced using high-throughput DNA technologies. The sequences are then compared with public databases to identify resistance genes (Schmieder and Edwards 2012) (Figure 1-15). The major limitation of this method is that the expression of the putative ARGs identified using this approach needs to be studied.

In a functional metagenomics, the metagenomic DNA from a sample of microbial community is sheared and cloned into a suitable vector and then transformed into a suitable host (such as *E. coli*). The library is then functionally screened by plating on agar media supplemented with antibiotics at a concentration that will kill the host if it does not contain a resistance gene for the antibiotic used (Figure 1-15). The major advantage of this method is that completely novel antibiotic resistance genes can be detected.
According to a recent estimate, an adult human contains on average about 30 trillion human cells and 39 trillion bacteria, thus, in a 'reference man' the ratio of bacterial and human cells is close to 1:1 (Sender, Fuchs et al. 2016). The majority of these human commensal live in our gut and oral cavity. These huge numbers of oral and gut microbes are very diverse and vary widely among healthy individuals (Nasidze, Li et al. 2009, Human Microbiome Project 2012).

**Figure 1-15.** Flowchart showing the steps for different metagenomic approaches to investigate the presence of ARGs in an environmental or clinical sample. The common steps for all different approaches are shown in orange colour on the top.

### 1.12 Microbial communities in the oral cavity

According to a recent estimate, an adult human contains on average about 30 trillion human cells and 39 trillion bacteria, thus, in a 'reference man' the ratio of bacterial and human cells is close to 1:1 (Sender, Fuchs et al. 2016). The majority of these human commensal live in our gut and oral cavity. These huge numbers of oral and gut microbes are very diverse and vary widely among healthy individuals (Nasidze, Li et al. 2009, Human Microbiome Project 2012).
Figure 1-16. The anatomy of the human mouth. The photo was adapted from Wikipedia on 20 June 2017.

The oral cavity contains the second largest community of commensal bacteria in a human body. It is composed of multiple structures ranging from soft tissues such as the tongue, buccal mucosa, tonsils and cheek to hard, non-shedding surfaces of teeth as well as hard and soft palates (Figure 1-16). Each of these structures are colonised by distinct types of microorganisms and continuously bathed in saliva. So, the saliva contains a mix of bacteria shed from the different ecosystems in the oral cavity (Lamont 2006).

To date the Human Oral Microbiome Database (HOMD)(www.homd.org) has recorded a total of 688 taxa which belong to 13 phyla including Actinobacteria, Bacteroidetes, Chlamydiae, Chloroflexi, Euryarchaeota, Firmicutes, Fusobacteria, Proteobacteria, Spirochaetes, SR1, Synergistetes, Tenericutes, and TM7 of which the six major phyla Firmicutes, Bacteroidetes, Proteobacteria, Actinobacteria, Spirochaetes, and Fusobacteria constitute 96% of the taxa (Dewhirst, Chen et al. 2010). Figure 1-17 represents the diversity of oral bacteria according to their phylum. According to HOMD, among the 688 taxa, 444 taxa (65%) are cultivable and the remaining 244 taxa (35%) are uncultivable (Dewhirst 2015).
Some bacterial species in the oral cavity are not only responsible for oral diseases including dental caries and periodontal diseases, but are also attributed to other systemic infections such as bacterial endocarditis, pneumonia as well as other conditions including low birth-weight and coronary heart diseases (Lockhart and Durack 1999, Curtis, Zenobia et al. 2011, Ling, Liu et al. 2013).

**Figure 1-17.** Microbial diversity in the oral cavity according to the phylum. The number before the name of the phyla indicates the number of taxa in each phylum. The photo was adapted from Dewhirst (2015).

The microbiota in the oral cavity can be present as planktonic cells or in the form of biofilms. Many of the bacteria present in the mouth grow in biofilms. The compact and stable microbial biofilms formed on the different surfaces in the oral cavity provide an ideal environment for HGT of ARGs as well as other adaptive genes (Roberts and Kreth 2014).
Figure 1-18 shows the architecture of a typical subgingival biofilm.

**Figure 1-18.** A subgingival biofilm with Actinomyces sp. (green), bacteria (red) and eukaryotic cells (large green cells on top). The four different layers of the biofilm are depicted with the root surface orientated to the bottom of the image. Scale bar = 10 μm. The image was adapted from Zijing, van Leeuwen et al. (2010) (p.3).

### 1.13 ARGs in the cultivable oral bacteria

The majority of the early reports on antibiotic resistance in oral bacteria involved oral streptococci. *Streptococcus mutans*, resistant to penicillin, vancomycin, chloramphenicol and tetracyclines were reported in 1974 (Baker and Thornsberry 1974). In the early 1980s, *tet(M)* was identified as the cause of resistance to tetracyclines in viridans group streptococci where some strains
were found to contain a Tn916-like conjugative element (Burdett, Inamine et al. 1982, Hartley, Jones et al. 1984). In one study carried out in our laboratory to identify the genetic basis of tetracycline resistance in the oral bacteria of healthy volunteers, 105 tetracycline-resistant oral isolates were screened for the presence of tetracycline resistance genes by multiplex PCR (Villedieu, Diaz-Torres et al. 2003). It was found that 79% of the isolates carried tet(M) which is followed by tet(W) (21%), tet(O) (10.5%), tet(Q) (9.5%) and tet(S) (2.8%). A low percentage of isolates (4.8%) carried tetracycline resistance genes encoding an efflux protein that includes tet(L) (2.8%), tet(A) (1%) and tet(K) (1%) (Villedieu, Diaz-Torres et al. 2003). In another study, Warburton, Roberts et al. (2009) detected two variants tet(32) in the oral cavity.

The oral bacteria were also found to carry erythromycin resistance genes, erm(B) and mef. Villedieu, Diaz-Torres et al. (2004) found that on average 7% of the cultivable oral bacteria were resistant to erythromycin. Screening of the resistance genes showed that 67% of the erythromycin-resistant isolates carried mef encoding efflux pumps and most of the mef positive isolates were identified as Streptococcus spp. The methyltransferase gene, erm(B) was identified in 31% of the erythromycin-resistant streptococci and erm(F) was identified in one Veillonella isolate (Villedieu, Diaz-Torres et al. 2004).

Tetracycline resistant bacteria carrying tet(M), tet(B), tet(K), tet(O), tet(Q), tet(S), tet(W) as well as tet(32) were also identified in dental plaque samples collected from a group of children aged 4-6 years (Lancaster, Bedi et al. 2005). tet(M) was the most common and southern blotting analysis showed that the gene was located on a Tn916-like element in different bacterial genera such as Streptococcus, Granulicatella, Veillonella and Neisseria (Lancaster, Bedi et al. 2005). Some other resistance genes were also found to be carried on mobile genetic elements. For instance, mef(A) was found on a conjugative transposon Tn1207.3 in Streptococcus pyogenes (Santagati, Iannelli et al. 2003). In a strain
of *Streptococcus oralis*, Ciric, Mullany et al. (2011) detected a novel gene named, *qrg* [quaternary ammonium compound (QAC) resistance gene] located on Tn6087, a transposon of Tn916-family.

1.14 ARGs in the oral cavity detected by different metagenomic approaches

It has been estimated that approximately 35% of the bacterial taxa that inhabit the oral cavity cannot be cultured in laboratory conditions (Dewhirst 2015). So, the role of the uncultivable fraction of microbiota in the development and spread of antibiotic resistance remains elusive. In the last two decades, many studies were carried out to investigate the presence of ARGs in the metagenomic DNA of the oral cavity which results in the detection of previously unknown resistance genes. For instance, using a functional metagenomic approach, a novel tetracycline resistant determinant, *tet(37)* has been identified in the saliva and dental plaque metagenome. This gene was found to confer resistance by inactivating the antibiotic (Diaz-Torres, McNab et al. 2003). In another function-based study using saliva and dental plaque metagenome of healthy human volunteers, the inserts in clones (*E. coli*) resistant to tetracycline, amoxicillin and gentamicin were sequenced and tetracycline resistance genes including *tet(M), tet(O)*, *tet(Q), tet(W), tet(37)* and *tet(A)* were identified (Diaz-Torres, Villedieu et al. 2006).

Sommer, Dantas et al. (2009) constructed metagenomic libraries from human saliva and faecal samples by cloning into pZE21 MCS 1. The clones in these libraries were screened against 13 different classes of antibiotics and a total of 93 functional ARGs were identified, the majority of which were novel. In the
saliva library, Sommer, Dantas et al. (2009) identified two novel beta-lactamase
genes (including a HOA-1), a novel aacA gene and eight novel ddl (D-alanine-
D-alanine ligase) genes responsible for resistance to penicillin G,
aminoglycosides and D-cycloserine, respectively (Sommer, Dantas et al. 2009).
Recently, Reynolds, Roberts et al. (2016) constructed a metagenomic library
from a pooled saliva sample of 11 healthy human volunteers. A novel
tetracycline and tigecycline transporter, tetAB(60) was identified which
conferred resistance to tetracycline and tigecycline.

1.15 Integrons in the oral cavity

In 2004, the presence of an unusual integron (intI genes and ORFs in the GCs
are in the same orientation) harbouring an array of 45 gene cassettes (GCs)
encoding 70 complete ORFs (TDE1843-TDE1773) was identified in silico within
the chromosome of T. denticola ATCC35405, an oral spirochete known to
cause periodontitis (Coleman, Tetu et al. 2004) (Figure 1-19). Later on, Wu,
Rho et al. (2012) identified two more attC sites immediately downstream of the
integron detected by Coleman, Tetu et al. (2004), thus the total number of GCs
in that integron is now 47. The GCs of the integron of T. denticola ATCC35405
mostly encode proteins of unknown functions and no known ARGs were found
in the GC array. Among the 70 ORFs, only 5 ORFs were found to encode
proteins with known function (Table 1-7).

Table 1-7. ORFs within the CI of T. denticola ATCC35404 that encode proteins with
known functions.

<table>
<thead>
<tr>
<th>ORF</th>
<th>Predicted function</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDE1784</td>
<td>AraC family transcription regulator</td>
</tr>
<tr>
<td>TDE1793/TDE1827*</td>
<td>Carbon-nitrogen family hydrolase</td>
</tr>
<tr>
<td>TDE1795</td>
<td>FunZ protein</td>
</tr>
<tr>
<td>TDE1805</td>
<td>Radical SAM protein</td>
</tr>
</tbody>
</table>
*These two ORFs are >99% identical.

**Figure 1-19.** Structure of the unusual chromosomal integron of *T. denticola* ATCC35405. The intI gene and cassette ORFs (block arrows, sizes not to scale) are numbered according to GenBank accession no. NC_002967. White ORFs have no known function, light grey ORFs are related to conserved hypothetical proteins and dark grey ORFs are related to proteins of known function (blast E values<0.00001). Circles indicate putative 59-be recombination sites. Underlined cassettes (A–G) are duplicates, defined as containing ORFs with >95% amino acid identity. The figure was taken from Coleman, Tetu et al. (2004) (p.3524).

In an *in silico* analysis of the metagenomic datasets of NIH Human Microbiome Project (HMP), Wu, Rho et al. (2012) identified 826 integron GCs related to *Treponema* species in the oral cavity. The genes were assigned to Clusters of Orthologous Groups (COGs) families according their predicted functions (https://www.ncbi.nlm.nih.gov/COG/). Among the 826 genes, approximately 60% (501/826) could not be assigned to any COG families. Of the 325 genes that were assigned, approximately 60% (195/325) were categorised R (general function prediction only) and S (function unknown) (Table 1-8). Thus, combining these two categories (R and S) with the genes that were remained unassigned,
85% of the 826 genes were found to encode proteins of unknown function. Their analysis also showed that gene sharing among the samples is minimal and unique to individuals (Wu, Rho et al. 2012). Their analysis also showed that approximately 80% of the normal human population carry integron-containing *Treponema* sp.

**Table 1-8.** The COG functional category distributions of the integron gene cassettes identified in the different locations of oral cavity

<table>
<thead>
<tr>
<th>COG Functional Categories</th>
<th>Supragingival plaque*</th>
<th>Tongue dorsum*</th>
<th>Subgingival plaque*</th>
</tr>
</thead>
<tbody>
<tr>
<td>[C] Energy production and conversion</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[D] Cell cycle control, cell division, chromosome partitioning</td>
<td>8 (11)</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>[E] Amino acid transport and metabolism</td>
<td>2 (4)</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>[F] Nucleotide transport and metabolism</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[G] Carbohydrate transport and metabolism</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>[H] Coenzyme transport and metabolism</td>
<td>1</td>
<td>1(8)</td>
<td>0</td>
</tr>
<tr>
<td>[I] Lipid transport and metabolism</td>
<td>1 (3)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[J] Translation, ribosomal structure and biogenesis</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>[K] Transcription</td>
<td>9</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>[L] Replication, recombination and repair</td>
<td>10 (12)</td>
<td>10 (14)</td>
<td>5 (6)</td>
</tr>
<tr>
<td>[M] Cell wall/membrane/envelope biogenesis</td>
<td>3 (4)</td>
<td>2 (3)</td>
<td>0</td>
</tr>
<tr>
<td>[N] Cell motility</td>
<td>8 (10)</td>
<td>2 (5)</td>
<td>1</td>
</tr>
<tr>
<td>[O] Posttranslational modification, protein turnover, chaperones</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>[P] Inorganic ion transport and metabolism</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>[R] General function prediction only</td>
<td>45 (67)</td>
<td>13 (14)</td>
<td>7 (8)</td>
</tr>
<tr>
<td>[S] Function unknown</td>
<td>59 (72)</td>
<td>18 (21)</td>
<td>14 (15)</td>
</tr>
<tr>
<td>[T] Signal transduction mechanisms</td>
<td>6 (10)</td>
<td>3</td>
<td>2 (3)</td>
</tr>
<tr>
<td>[U] Intracellular trafficking, secretion, and vesicular transport</td>
<td>2 (3)</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>[V] Defence mechanisms</td>
<td>2</td>
<td>3 (4)</td>
<td>1</td>
</tr>
</tbody>
</table>

*Number of genes is obtained by clustering the genes at a 97% identity threshold for each functional category within each location. Numbers within parentheses indicate the number of genes before clustering. The data was taken from.
In another *in silico* study based on HMP dataset, 14 *T. denticola* strains carrying integrons were identified in the oral cavity (Wu, Doak et al. 2013). The proportion of integron genes shared between any two strains was calculated and among the 14 *T. denticola* strains, a few strains were found to share the integron genes. The strains which were phylogenetically close shared more genes. For instance, the strains ATCC 33521 and ATCC 35404 were very close in the phylogenetic tree (Figure 1-20) which share 89.7% of their integron genes.

*Figure 1-20.* The phylogenetic tree of the 14 integron-carrying *T. denticola* strains found in the oral cavity. The tree was built using the 31 marker genes. Red numerals indicate that the groupings are consistent with the phylogenetic relationship and the blue numeral highlights an unusual sharing of GCs between not-so-closely-related strains OTK and SP37. The figure was adapted from Wu, Doak et al. (2013) (p.7).

Although the closely related strains of *T. denticola* were found to share a high percentage of integron-located genes, the overall integron gene sharing among the strains was only 24.79%. Based on this observation, they argued that the *T. denticola* integrons are fully active and they are undergoing active insertion and deletion of cassettes.
1.16 Aims of this study

The main aim of this study was to investigate the presence of integrons, gene cassettes (GCs) and integron-associated antibiotic resistance genes (ARGs) in the oral metagenomic DNA of healthy humans.

The specific aims are:

i) To screen of the presence of integrons and GCs in the oral metagenome obtained from the healthy volunteers of the UK and Bangladesh using a PCR-based metagenomic approach

ii) To determine the resistance phenotype of the putative ARGs located within the GCs
Chapter Two: Materials and Methods
2.1 Introduction

The materials and methods that are common to the subsequent chapters are discussed here. Methods used for a specific chapter are outlined at the beginning of each chapter.

2.2 Source of chemicals, reagents and antibiotics

Chemicals, solvents, culture media and antibiotics were purchased from Sigma-Aldrich Ltd (Dorset, UK), BDH (UK) and Life Technologies (USA) unless stated otherwise. Plasmid preparation kits were obtained from Qiagen (UK). The restriction enzymes were obtained either from New England Biolabs (UK) Ltd or from Promega (UK). *E. coli* α-Select competent cells used for cloning were obtained from Bioline Reagents Ltd (UK). The primers were obtained from Sigma-Aldrich (UK).

2.3 Bacterial strains and plasmids used in this study

The bacterial strains and plasmids used in this study are listed in Table 2-1. pGEM-T Easy vector (Promega) was used for constructing the metagenomic library of PCR amplicons of integrons and gene cassettes. *E. coli* α-select silver efficiency (Bioline) was used to manipulate the pGEM-T Easy library of gene cassettes. *E. coli* BL21 (DE3) (New England Biolabs, NEB) and pET-28a(+) plasmid (Novagen) were used for cloning and expression of *ddls* and purification of His-tagged proteins. *Bacillus subtilis* 168 and pHCMC05 (a *B. subtilis*-*E. coli* shuttle vector) were used for cloning and expression of *ddl6* and *ddl7* to determine the susceptibility of D-cycloserine and vancomycin. *E. coli*
strains were grown in LB medium aerobically at 37°C. B. subtilis 168 was grown in Brain-heart infusion (BHI) medium unless stated otherwise. The bacterial strains were preserved at -80°C in 20-30% (v/v) sterile glycerol.

Table 2-1. Bacterial strains and plasmids used in this study

<table>
<thead>
<tr>
<th>Strain or plasmid</th>
<th>Genotype/description</th>
<th>Resistance marker</th>
<th>Reference /Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bacterial Strains</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| E. coli α-Select* | F- deoR endA1 recA1 relA1 gyrA96 hsdR17(rk-,
|                   | mk+) supE44 thi-1 phoA Δ(lacZYA-argF)U169
|                   | Φ80lacZAM15 λ-       |                   | Bioline, UK       |
| EC121             | E. coli α-Select containing empty pGEM-T Easy
|                   | vector               | AmpR              | This work         |
| EC126             | E. coli α-Select transformed with pGEM-T Easy::intI
|                   | attI-ddl6            | AmpR              | This work         |
| EC127             | E. coli α-Select containing pGEM-T Easy::intI-attI
|                   | ddl7                 | AmpR              | This work         |
| EC206             | E. coli α-Select containing pGEM-T Easy::intI-attI
|                   | ddl6 c.490 C>T       | AmpR              | This work         |
| EC207             | E. coli α-Select containing pGEM-T Easy::intI-attI
|                   | ddl6 c.777 G>T       | AmpR              | This work         |
| EC507             | E. coli α-Select containing pGEM-T Easy::ORF-intI
|                   | attI-ddl7            | AmpR              | This work         |
| EC606             | E. coli α-Select containing pET-28a(+)::attI-ddl6
|                   |                     | KanR              | This work         |
| EC607             | E. coli α-Select containing pET-28a(+)::attI-ddl7
|                   |                     | KanR              | This work         |
| E. coli BL21 (DE3)* | F-- ompT hsdSB(rB-,
|                   | mB-) gal dcm (DE3)   |                   | Invitrogen, UK    |
| EC300             | E. coli BL21 (DE3) containing empty pET-28a(+)
|                   | vector               | KanR              | This work         |
| EC306             | E. coli BL21(DE3) containing pET-28a(+): ddl6
|                   | plasmid              | KanR              | This work         |
| EC307             | E. coli BL21(DE3) containing pET-28a(+): ddl7
|                   | plasmid              | KanR              | This work         |
| EC308             | E. coli BL21(DE3) containing pET-28a(+): ddlTd
<p>|                   | plasmid              | KanR              | This work         |</p>
<table>
<thead>
<tr>
<th>Strain or plasmid</th>
<th>Genotype/description</th>
<th>Resistance marker</th>
<th>Reference /Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC309</td>
<td><em>E. coli</em> BL21(DE3) containing pET-28a(+)::ddlAEc plasmid</td>
<td>Kan&lt;sup&gt;R&lt;/sup&gt;</td>
<td>This work</td>
</tr>
<tr>
<td>EC310</td>
<td><em>E. coli</em> BL21(DE3) containing pET-28a(+)::attl-ddl6 plasmid</td>
<td>Kan&lt;sup&gt;R&lt;/sup&gt;</td>
<td>This work</td>
</tr>
<tr>
<td>EC311</td>
<td><em>E. coli</em> BL21(DE3) containing pET-28a(+)::attl-ddl7 plasmid</td>
<td>Kan&lt;sup&gt;R&lt;/sup&gt;</td>
<td>This work</td>
</tr>
<tr>
<td>B. subtilis 168</td>
<td>Laboratory strain (wild-type)</td>
<td></td>
<td>Dr. Haitham Hussain</td>
</tr>
<tr>
<td>BS700</td>
<td><em>B. subtilis</em> 168 containing empty pHCMC05 vector</td>
<td>Cm&lt;sup&gt;R&lt;/sup&gt;, Amp&lt;sup&gt;R&lt;/sup&gt;</td>
<td>This work</td>
</tr>
<tr>
<td>BS706</td>
<td><em>B. subtilis</em> 168 containing pHCMC05::ddl6 plasmid</td>
<td>Cm&lt;sup&gt;R&lt;/sup&gt;, Amp&lt;sup&gt;R&lt;/sup&gt;</td>
<td>This work</td>
</tr>
<tr>
<td>BS707</td>
<td><em>B. subtilis</em> 168 containing pHCMC05::ddl7 plasmid</td>
<td>Cm&lt;sup&gt;R&lt;/sup&gt;, Amp&lt;sup&gt;R&lt;/sup&gt;</td>
<td>This work</td>
</tr>
</tbody>
</table>

**Plasmids**

<table>
<thead>
<tr>
<th>Plasmid</th>
<th>Description</th>
<th>Resistance marker</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>pGEM-T Easy</td>
<td>Cloning vector</td>
<td>Amp&lt;sup&gt;R&lt;/sup&gt;</td>
<td>Promega, UK</td>
</tr>
<tr>
<td>pHCMC05</td>
<td><em>E. coli</em>-<em>B. subtilis</em> shuttle vector</td>
<td>Cm&lt;sup&gt;R&lt;/sup&gt;, Amp&lt;sup&gt;R&lt;/sup&gt;</td>
<td>Bacillus Genetic Stock Centre (BGSC)</td>
</tr>
<tr>
<td>pET-28a(+)</td>
<td>Expression vector, N-terminal His-tag</td>
<td>Kan&lt;sup&gt;R&lt;/sup&gt;</td>
<td>Novagen</td>
</tr>
<tr>
<td>pGEM-T Easy::intl-attl-ddl6</td>
<td>2024 bp PCR product containing partial sequence of <em>intl</em>, full length <em>attl</em> and <em>ddl6</em> (1032 bp) cloned into pGEM-T Easy</td>
<td>Amp&lt;sup&gt;R&lt;/sup&gt;</td>
<td>This work</td>
</tr>
<tr>
<td>pGEM-T Easy::intl-attl-ddl7</td>
<td>2024 bp PCR product containing partial sequence of <em>intl</em>, full length <em>attl</em> and <em>ddl7</em> (1032 bp) cloned into pGEM-T Easy</td>
<td>Amp&lt;sup&gt;R&lt;/sup&gt;</td>
<td>This work</td>
</tr>
<tr>
<td>pGEM-T Easy::intl-attl-ddl6 c.490 C&gt;T</td>
<td>2024 bp PCR product containing partial sequence of <em>intl</em>, full length <em>attl</em> and <em>ddl6</em> with C&gt;T substitution mutation at 490 position of nucleotide sequence cloned into pGEM-T Easy</td>
<td>Amp&lt;sup&gt;R&lt;/sup&gt;</td>
<td>This work</td>
</tr>
<tr>
<td>pGEM-T Easy::intl-attl-ddl6 c.777 G&gt;T</td>
<td>2024 bp PCR product containing partial sequence of <em>intl</em>, full length <em>attl</em> and <em>ddl6</em> with G&gt;T substitution mutation at 777 position of nucleotide sequence cloned into pGEM-T Easy</td>
<td>Amp&lt;sup&gt;R&lt;/sup&gt;</td>
<td>This work</td>
</tr>
<tr>
<td>pET-28a(+)::attl-ddl6</td>
<td><em>ddl6</em> and upstream 198 bp sequence containing <em>attl</em> and <em>Pc</em> cloned into pET-28a(+)</td>
<td>Kan&lt;sup&gt;R&lt;/sup&gt;</td>
<td>This work</td>
</tr>
<tr>
<td>Strain or plasmid</td>
<td>Genotype/description</td>
<td>Resistance marker</td>
<td>Reference /Source</td>
</tr>
<tr>
<td>------------------------------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>pET-28a(+)::attl-ddl7</td>
<td>ddl7 and upstream 198 bp sequence containing attl and Pc cloned into pET-28a(+)</td>
<td>KanR</td>
<td>This work</td>
</tr>
<tr>
<td>pET-28a(+)::ddl6</td>
<td>ddl6 (1032 bp) cloned into pET-28a(+)</td>
<td>KanR</td>
<td>This work</td>
</tr>
<tr>
<td>pET-28a(+)::ddl7</td>
<td>ddl7 (1032 bp) cloned into pET-28a(+)</td>
<td>KanR</td>
<td>This work</td>
</tr>
<tr>
<td>pHCMC05::ddl6</td>
<td>ddl6 (1032 bp) cloned into pHCMC05</td>
<td>CmR, AmpR</td>
<td>This work</td>
</tr>
<tr>
<td>pHCMC05::ddl7</td>
<td>ddl7 (1032 bp) cloned into pHCMC05</td>
<td>CmR, AmpR</td>
<td>This work</td>
</tr>
<tr>
<td>pET-28a(+)::ddlAEc</td>
<td>ddlA of E. coli cloned into pET-28a(+)</td>
<td>KanR</td>
<td>This work</td>
</tr>
<tr>
<td>pET-28a(+)::ddlTd</td>
<td>ddl of T. denticola cloned into pET-28a(+)</td>
<td>KanR</td>
<td>This work</td>
</tr>
</tbody>
</table>

*KKey to the genotypes: F-, host does not contain the fertility plasmid; deoR, constitutive expression of genes for deoxyribose synthesis; endA1, mutation in the non-specific endonuclease I which eliminates non-specific endonuclease activity, resulting in improved plasmid preps; recA1, mutation in a DNA-dependent ATPase that is essential for recombination and general DNA repair, thus reduces plasmid recombination and increases plasmid stability; relA1, RNA is synthesized in absence of protein synthesis (relaxed phenotype); gyrA96, DNA gyrase mutant produces resistance to nalidixic acid; hsdR17(rk-, mk+), unmethylated DNA not degraded, cell still can methylate DNA; supE44, suppression of the amber (UAG) stop codon by inserting glutamine; thi-1, requires thiamine for growth on minimal media, lacZΔM15, element required for β-galactosidase complementation when plated on X-gal; Φ80, cell carries the lambdoid prophage φ80, ompT, mutation in outer membrane protein protease VII, reducing proteolysis of expressed proteins; gal, mutation in galactose metabolism pathway, thus, the cell cannot grow on galactose only; dcm, mutation in DNA Cytosine methylase, thus, prepare unmethylated DNA; DE3, contains a lysogen that encodes T7 RNA polymerase.

### 2.4 Ethical Approval

Ethical approvals for this project were obtained from University College London (UCL) Ethics Committee (project number 5017/001) and the Institutional Animal, Medical Ethics, Biosafety and Biosecurity Committee (IAMEBBC) for Experimentations on Animal, Human, Microbes and Living Natural Sources, University of Rajshahi, Bangladesh (project number 54/320/IAMEBBC/IBSC). Both ethics committees approved the consent procedure for the sample
collection and processing. The signed consent forms from all participants are securely stored.

**2.5 Collection and storage of saliva samples**

Saliva samples were collected from 11 healthy volunteers from the UK and 10 from Bangladesh, respectively. The volunteers included both male and female with age between 21 and 65. They were given an information sheet prior to sample collection to read and decide if they want to participate in the study. The individuals who agreed to participate in the study were provided with the consent form to sign. None of the volunteers had received antibiotic treatment for 3 months before the sample collection day and were generally healthy.

The UK samples (n=11) were collected from the staff and international postgraduate students from the UCL Eastman Dental Institute and represent various ethnic and cultural backgrounds including Asian, Australian, European, African and Middle-Eastern, some of which had moved to the UK in the past few months. Therefore, the UK samples represent an international metagenome. 2 ml of saliva were collected from the volunteers (based in the UK) in a sterile plastic tube. They were instructed to spit into a sterile plastic tube until the volume of 2 ml is reached. The samples were stored on ice immediately after collection and processed immediately in a biosafety cabinet.

The Bangladeshi samples (n=10) were collected from the staff, undergraduate and post-graduate students of Department of Pharmacy of Rajshahi University, Bangladesh and all of the volunteers were Bangladeshi. The samples from these volunteers were collected using Saliva DNA Collection, Preservation and Isolation Kit (Cat. RU35700) (Norgen, Canada) (Figure 2-1) following the
manufacturer’s guidelines. Briefly, saliva samples are collected by spitting inside the Collection Funnel which was assembled with the Collection Tube. After collecting 2 ml of saliva, the Collection Funnel was removed and the contents of the Preservative Ampoule were then added and mixed with the collected saliva. The Saliva Collection Tubes were subsequently sent to the laboratory of Microbial Diseases at UCL for DNA isolation and analysis. All of the samples were anonymized, but were labelled with a serial number.

![Collection Funnel, Collection Tube, Preservative](image)

**Figure 2-1.** Saliva DNA collection, Preservation and Isolation Kit (Norgen, Canada) used for collecting saliva from Bangladeshi volunteers. The photo was modified from Norgen’s website.

### 2.6 Extraction of metagenomic DNA from saliva

The freshly collected UK saliva samples were pooled together into a sterile plastic tube in a class I microbiological safety cabinet. The pooled saliva sample was then divided into 1.5ml aliquots and centrifuged at 16,168 x g for 1 min. The UK oral metagenomic DNA was then extracted by using the Gentra Puregene Yeast/Bact. Kit (Qiagen, Germany), following the protocol for Gram-
positive bacteria with the modification in the final step, where the DNA pellets were dissolved in 400µL DNA hydration solution instead of 100 µL.

The Bangladeshi oral metagenomic DNA was extracted from the Norgen’s Saliva DNA storage buffer using the ethanol precipitation technique according to manufacturer’s protocol. The preservative buffer of Norgen devices is designed for rapid cellular lysis and subsequent preservation of DNA from fresh saliva samples. Prior to DNA isolation, the storage devices were incubated for 1h at 50°C and mixed by inversion and gentle shaking for 10 seconds. DNA was then extracted from 500 µL of the pooled saliva in preservative buffer by taking 50 µL aliquots from 10 saliva samples. 15 µL proteinase K (Norgen, Canada) was added and mixed by vortexing for 10 seconds followed by incubating 1hr at 55°C. An equal volume of room temperature isopropanol (515 µL) was added to the samples and mixed gently by inversion 10 times. The tubes were centrifuged at room temperature for 5 min at 16,168 x g and supernatants were discarded. The residual isopropanol was drained by keeping the tubes upside down on an absorbent paper. The DNA was then washed by 500 µL of 70% ethanol and gently swirled and allowed to stand for 1 min. The tubes were then centrifuged for 1 min at 16,168 x g and the supernatant was discarded. The tubes were air dried for 15 mins by keeping the tubes upside down. The dried DNA pellets were re-dissolved in 50 µL DNA hydration buffer (from Puregene kit) and vortexed for 30 seconds. The tubes were incubated at 55°C for 10 min and centrifuged for 1 min at 16,168 x g to remove any insoluble material. The clear supernatant was transferred to new tubes and stored at -20°C. An aliquot of the extracted DNAs was run on a 1% agarose gel to see the
quality and the size of the DNA. Concentration was measured using NanoDrop (Thermo Fisher Scientific, USA).

### 2.7 Purification of genomic DNA from bacterial culture

Genomic DNA extraction from bacterial culture was carried out using the Gentra Puregene Yeast/Bact. Kit (Qiagen, Germany) with slight modifications. All centrifugation steps were carried out at 16,168 × g (13,200 rpm) in a bench-top microcentrifuge. 1.5 mL of overnight culture was centrifuged for 5 seconds and the supernatant was removed. The cell pellet was re-suspended in 300 µl of Cell Suspension Solution by gentle pipetting. 1.5 µL of Lytic Enzyme Solution was added and the tubes were inverted 25 times and incubated at 37°C for 30 min to weaken the cell wall. The cells were centrifuged for 1 min and the supernatant was discarded. The cell pellet was re-suspended in 300 µl of Cell Lysis Solution and pipetted to lyse the cells. In most of the cases the cells were heated at 80°C and allowed to cool for 5 minutes to room temperature. To remove RNA from the solution, 1.5 µl of RNase A Solution was added to the cell lysate which was mixed by inverting 25 times followed by incubation at 37°C for 60 mins. The tubes were kept on ice for 1 min to quickly cool down the samples and then 100 µl of Protein Precipitation Solution was added. The samples were vortexed vigorously at high speed for 20 seconds and centrifuged for 3 min. The supernatant was transferred to another Eppendorf tube containing 300 µl 100% isopropanol and mixed by inverting the tubes 50 times and centrifuged for 1 min. The supernatant was discarded and the tubes were drained on a clean absorbent paper. 300 µl of 70% ethanol was added and the tube was inverted several times to wash the DNA pellet. The tube was centrifuged for 1 min and
supernatant discarded. The DNA pellet was air dried for 10-15 min and 100 µl of DNA Hydration Solution was added to re-dissolve the DNA. The tubes were vortexed at medium speed for 5 seconds and incubated at 65°C for 1 hour. Finally, the tubes are incubated overnight with a gentle shaking. The dissolved DNA was then stored at 20°C.

2.8 Synthesis of oligonucleotides used in this study

The primers or oligonucleotides were designed and ordered for synthesis at Sigma Aldrich (UK). Unmodified primers were synthesized at the scale of 0.025 µM and purified by desalting and supplied in dry form. The phosphorylated primers (5’-modification with a phosphate group) were synthesized at 0.05 µM scale and purified by HPLC and supplied in dry form. 100 µM stock solution was prepared using the indicated amount of molecular grade water (Sigma, UK) and diluted 10 times to prepare 10 µM working solution. The stock and working oligonucleotide solutions were stored in -20°C and thawed on ice before use. The names, sequences, targets and the references of the primers used in this study are given in Table 2-2.

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Primer Sequence (5’-3’)</th>
<th>Target</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>MARS1</td>
<td>CGYAATRTCAKGGTGAAAG*</td>
<td>attC site in reverse direction</td>
<td>This work</td>
</tr>
<tr>
<td>MARS2</td>
<td>GCAATGTCAGGTTGAAAGC</td>
<td>attC site in reverse direction</td>
<td>This work</td>
</tr>
<tr>
<td>MARS3</td>
<td>CRCRAMYRYWGGTYAAAGCG*</td>
<td>attC site in forward direction</td>
<td>This work</td>
</tr>
<tr>
<td>MARS4</td>
<td>CRCAAATGCAGGTYAAAGCG*</td>
<td>attC site in forward direction</td>
<td>This work</td>
</tr>
<tr>
<td>MARS5</td>
<td>CGCAAATTGCGGTTAAGCG</td>
<td>attC site in forward direction</td>
<td>This work</td>
</tr>
<tr>
<td>TDIF</td>
<td>TCAAGGCCAAATCAGGCTCT</td>
<td>intI in forward direction</td>
<td>This work</td>
</tr>
<tr>
<td>TDIR</td>
<td>CCTGAATCCTTAACCGCAAG</td>
<td>intI in reverse direction</td>
<td>This work</td>
</tr>
<tr>
<td>Primer name</td>
<td>Primer Sequence (5’-3’)</td>
<td>Target</td>
<td>Source</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------------------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>HS458</td>
<td>GTTGTGATTAGGACGAGCAACG</td>
<td>5’-conserved segment (5’-CS) (intI1 and attI1)</td>
<td>(Betteridge, Partridge et al. 2011)</td>
</tr>
<tr>
<td>HS459</td>
<td>GCAAAAAGGCAGCAATTAGGACC</td>
<td>3’-conserved segment (3’-CS) (qacEΔ1 and sul1)</td>
<td></td>
</tr>
<tr>
<td>hep74</td>
<td>CGGGATCCCGGACGCAACTGCA</td>
<td>att2</td>
<td>(White, McIver et al. 2001)</td>
</tr>
<tr>
<td>hep51</td>
<td>GATGCCATCGCAAGTACGAG</td>
<td>orfX located downstream of the cassette array</td>
<td></td>
</tr>
<tr>
<td>intI-864R</td>
<td>YACGAGATGNGTGGCRAAVSWRT GSCG*</td>
<td>intI gene conserved region</td>
<td>(Elsaied, Stokes et al. 2011)</td>
</tr>
<tr>
<td>GCP1</td>
<td>GCSGCTKANCTCVRRCGTTRRY*</td>
<td>attC site in forward direction</td>
<td></td>
</tr>
<tr>
<td>GCP2</td>
<td>TCSGCTKGARCGAMTTGTTTRRY*</td>
<td>attC site in reverse direction</td>
<td></td>
</tr>
<tr>
<td>ICC48</td>
<td>GCAACTGGTGACAGACCTTGC</td>
<td>intI</td>
<td>(Huang, Cagnon et al. 2009)</td>
</tr>
<tr>
<td>ICC21</td>
<td>GTCCGGTCTGGYGAATTGTTAGRC*</td>
<td>attC of class 1 and class 2 integrons</td>
<td></td>
</tr>
<tr>
<td>AR1F</td>
<td>GCAAGTAGCGTGATCGTCACGC</td>
<td>intI</td>
<td>This study</td>
</tr>
<tr>
<td>ARC1</td>
<td>CGATTGTATTGCGACGATC</td>
<td>attC of class 1 integron</td>
<td>This work</td>
</tr>
<tr>
<td>ARC2</td>
<td>GTTGTGACATGAGGCGGGCGG</td>
<td>attC of class 1 integron</td>
<td>This work</td>
</tr>
<tr>
<td>intI1 F</td>
<td>CCTCCCCAGCATGATCCGCA</td>
<td>intI1 gene</td>
<td>(Goldstein, Lee et al. 2001)</td>
</tr>
<tr>
<td>intI1 R</td>
<td>TCCACGCATGCAGTGCGC</td>
<td>intI1 gene</td>
<td></td>
</tr>
<tr>
<td>intI2 F</td>
<td>TTATTGCTGGGTATTAGCC</td>
<td>intI2 gene</td>
<td></td>
</tr>
<tr>
<td>intI2 R</td>
<td>ACGGCTACCTCCTGTATCT</td>
<td>intI2 gene</td>
<td></td>
</tr>
<tr>
<td>intI3 F</td>
<td>AGTGGGTGGCGGGAGGTG</td>
<td>intI3 gene</td>
<td></td>
</tr>
<tr>
<td>intI3 R</td>
<td>TGTTCCTGTGATCGGAGG</td>
<td>intI3 gene</td>
<td></td>
</tr>
<tr>
<td>Sul1F</td>
<td>GTATTGCGGCGCTCTTAGAC</td>
<td>sul1 gene</td>
<td>This work</td>
</tr>
<tr>
<td>Sul1R</td>
<td>CCGACTTCAGGCTCTTGAAGG</td>
<td>sul1 gene</td>
<td></td>
</tr>
<tr>
<td>Sul2Fw</td>
<td>TTTCCGGATCGCTCAACATAA</td>
<td>sul2 gene</td>
<td>This work</td>
</tr>
<tr>
<td>Sul2Rv</td>
<td>GCCGGTTGATAACTGTCGAG</td>
<td>sul2 gene</td>
<td></td>
</tr>
<tr>
<td>Qac1F</td>
<td>GGCAGAGTAATCGAACACT</td>
<td>qacEΔ1 gene</td>
<td>This work</td>
</tr>
<tr>
<td>Qac1R</td>
<td>AACCAGGCAATGGCTGTAAT</td>
<td>qacEΔ1 gene</td>
<td></td>
</tr>
<tr>
<td>TDIF</td>
<td>TCAAGCCAAAATCAGGCTCT</td>
<td>intO of T. denticola (forward primer)</td>
<td>This work</td>
</tr>
<tr>
<td>MARS2</td>
<td>GCAATGTACGGTGAGGCA</td>
<td>attC in reverse direction</td>
<td>This work</td>
</tr>
<tr>
<td>ddlF</td>
<td>GTAGTACTTGTGAGGAGG</td>
<td>ddl6/ddl7 (Forward primer)</td>
<td>This work</td>
</tr>
<tr>
<td>ddlR</td>
<td>GCCCTTTAATTTATTGCAATTG</td>
<td>ddl6/ddl7 (Reverse primer)</td>
<td>This work</td>
</tr>
<tr>
<td>Primer name</td>
<td>Primer Sequence (5’-3’)</td>
<td>Target</td>
<td>Source</td>
</tr>
<tr>
<td>-------------</td>
<td>--------------------------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>TddlF</td>
<td>GCTAGGATCCATGAAGGTAGTACTTGC</td>
<td>ddI6/ddI7 (Forward primer, BamHI site underlined)</td>
<td>This work</td>
</tr>
<tr>
<td>TddlR</td>
<td>ACGCTCTAGATTAGCCTCTAATTTTTATTTGC</td>
<td>ddI6/ddI7 (Reverse primer, XbaI site underlined)</td>
<td>This work</td>
</tr>
<tr>
<td>Tddl28aR</td>
<td>GCCGCTCGAGTTAGCCTCTAATT TTATTGC</td>
<td>ddI6/ddI7 (Reverse primer, XhoI site underlined)</td>
<td>This work</td>
</tr>
<tr>
<td>Ddi6-490F</td>
<td>TATCAGAAGTTAATTTTGATACAATTAAAGAGAGA</td>
<td>ddI6 cloned into pGEM-T Easy (forward primer to change C of ddI6 at c.490 to T)</td>
<td>This work</td>
</tr>
<tr>
<td>Ddi6-490R</td>
<td>ATGAATAAGTTTTCCATTTCAATT TATTGCG</td>
<td>ddI6 cloned into pGEM-T Easy (reverse primer to couple with Ddi6-490F primer)</td>
<td>This work</td>
</tr>
<tr>
<td>Ddi6-777F</td>
<td>CAACGAAATATGTCCTGCTGAAATT AAAGAGAGA</td>
<td>ddI6 cloned into pGEM-T Easy (forward primer to change G of ddI6 at c.777 to T)</td>
<td>This work</td>
</tr>
<tr>
<td>Ddi6-777R</td>
<td>GATAATCCTTTTTTGATACCTTTTTATAATCTAAATTTT</td>
<td>ddI6 cloned into pGEM-T Easy (reverse primer to couple with Ddi6-777F primer)</td>
<td>This work</td>
</tr>
<tr>
<td>Upint-5100F</td>
<td>GGCGGAGATGAAGATACCTT</td>
<td>2.7 kb upstream of integrase of T. denticola ATCC35404</td>
<td>This work</td>
</tr>
<tr>
<td>Upint-3685F</td>
<td>TCACAGCCTGCAAGCAATGT</td>
<td>4.2 kb upstream of integrase of T. denticola ATCC35404</td>
<td>This work</td>
</tr>
<tr>
<td>Upint-721F</td>
<td>GGTGGGGAGCTCATCATAAA</td>
<td>7.2 kb upstream of integrase of T. denticola ATCC35404</td>
<td>This work</td>
</tr>
<tr>
<td>Upint-122F</td>
<td>GCTTCAAATCTATACCGGAAC</td>
<td>7.8 kb upstream of integrase of T. denticola ATCC35404</td>
<td>This work</td>
</tr>
<tr>
<td>TddlF</td>
<td>GCTAGGATCCATGAAGGTAGTACTTGC</td>
<td>ddI6/ddI7 (Forward primer, BamHI site underlined)</td>
<td>This work</td>
</tr>
<tr>
<td>TddlR</td>
<td>ACGCTCTAGATTAGCCTCTAATTTTTATTTGC</td>
<td>ddI6/ddI7 (Reverse primer, XbaI site underlined)</td>
<td>This work</td>
</tr>
<tr>
<td>Tddl28aR</td>
<td>GCCGCTCGAGTTAGCCTCTAATT TTATTGC</td>
<td>ddI6/ddI7 (Reverse primer, XhoI site underlined)</td>
<td>This work</td>
</tr>
<tr>
<td>EC-ddIA-F</td>
<td>GCTAGGATCCATGGGAAAATGCGG TAGGAAT</td>
<td>ddIA of E. coli (forward primer, BamHI site underlined)</td>
<td>This work</td>
</tr>
<tr>
<td>EC-ddIA-R</td>
<td>ACGCAAGCTTCAATTGGATTTTTCAATGCGTTATC</td>
<td>ddIA of E. coli (reverse primer, HindIII site underlined)</td>
<td>This work</td>
</tr>
<tr>
<td>DdiTD-F</td>
<td>GCCGCAATTCTATGAATATAGCAATCATTACCAGCGG</td>
<td>ddI6 of T. denticola (forward primer, EcoRI site underlined)</td>
<td>This work</td>
</tr>
<tr>
<td>Primer name</td>
<td>Primer Sequence (5ʹ-3ʹ)</td>
<td>Target</td>
<td>Source</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------------------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>DdlTD-R</td>
<td>ACGCAAGCTTAGATGACGGCAGGTGGTGGTTC</td>
<td><em>ddl</em> of <em>T. denticola</em> (reverse primer, HindIII site Underlined)</td>
<td>This work</td>
</tr>
<tr>
<td>M13F</td>
<td>GTAAAACGACGGCCAG</td>
<td>M13 forward sequencing</td>
<td>Universal</td>
</tr>
<tr>
<td>M13R</td>
<td>CAGGAAACAGCTATGAC</td>
<td>M13 reverse sequencing</td>
<td>Universal</td>
</tr>
<tr>
<td>T7 promoter</td>
<td>TAATACGACTCATATAGGG</td>
<td>T7 forward sequencing</td>
<td>Universal</td>
</tr>
<tr>
<td>T7 Terminator</td>
<td>GCTAGTTATTGCTACGCGG</td>
<td>T7 reverse sequencing</td>
<td>Universal</td>
</tr>
<tr>
<td>pHCMC05-F</td>
<td>CAAGGTGTGGCATAATGTGTG</td>
<td>pHCMC05 forward sequencing</td>
<td>This work</td>
</tr>
<tr>
<td>pHCMC05-R</td>
<td>TCATCTCCAAATTTCTTCCGAG</td>
<td>pHCMC05 reverse sequencing</td>
<td>This work</td>
</tr>
</tbody>
</table>

*Degenerate nucleotides: D = A, G, or T; H = A, C, or T; I or N= A, C, G, or T; K = G or T; M = A or C; R = A or G; S = G or C; V = A, C, or G; Y = C or T.*

### 2.9 General PCR protocol

For PCR, two different ready-to-use mix of polymerases and dNTPS were used to amplify the targets:

i) BioMix Red (Bioline, UK), a complete ready-to-use 2x reaction mix containing an ultra-stable Taq DNA polymerase and a red-coloured inert loading dye, 2.5 mM Mg++ and dNTPs was used to amplify products for TA cloning. This was also used for routine PCR such as colony PCR, confirmation of the presence of inserts in a plasmid or presence of a gene in a DNA sample etc.

ii) The Q5® High-Fidelity 2X Master Mix (NEB, UK), which contains a high-fidelity, thermostable DNA polymerase with 3´→ 5´ exonuclease activity and 2.0 mM Mg++ and dNTPs, was used for cloning the PCR products into pET vectors and to amplify large DNA fragments.
A typical 50 µL PCR reaction contained 25 µL 2x BioMix Red or Q5® High-Fidelity 2X Master Mix, 2 µL of each 10 µM primer, 50-100 ng of DNA template, and molecular water up to 50 µL. All PCR reactions were performed in Biometra Thermocycler T3000 machine. The PCR conditions were set according to the type of mix used. The following PCR conditions were used for BioMix Red: an initial denaturing temperature of 95°C for 2 minutes followed by 30-35 cycles at 95°C for 30 seconds (denaturing), 50-65°C for 30 seconds (annealing), and 72°C for 30-120 seconds depending on the expected size of the amplicon (extension). The PCR run ended with a final 72°C elongation step for 5-10 minutes before holding the reaction at 4°C.

The reaction condition for Q5® High-Fidelity 2X Master Mix was as follows: an initial denaturing temperature of 98°C for 30 seconds followed by 30-35 cycles at 98°C for 10 seconds (denaturing), 50-75°C for 30 seconds (annealing), and 72°C for 30 seconds to 200 seconds depending on the expected size of the amplicon (extension). The PCR run ended with a final 72°C elongation step for 2-10 minutes before holding the reaction at 4°C. The annealing temperature was used based on the calculated Tm of the oligos by the supplier.

**2.10 Agarose gel electrophoresis**

Agarose gels were cast using 1x TAE buffer and stained with 1:10,000 dilution of GelRed nucleic acid stain (Biotium, UK). 1% gels were used in most of the cases. PCR products obtained with BioMix Red were loaded directly as they already contained a loading dye, but, the products obtained using Q5® High-Fidelity 2X Master Mix were loaded using 1x Loading Dye (Qiagen, Germany). 1x TAE buffer was used for electrophoresis with a volt range of 4-10 V/cm (the
distance between anode and cathode). The amount of PCR product loaded depends on the purpose of electrophoresis. For gel extraction, the whole amount of the products is loaded after the initial checking with 2-3 µL sample. HyperLadder I (Bioline) was used as a size marker (200 bp-10 kb). The gels were imaged in an Alphalmager gel documentation system (Alpha Innotech).

2.11 PCR purification

The PCR products were purified by using the QIAquick PCR Purification Kit (Qiagen, UK) before sequencing or cloning. This kit can be used to purify single or double stranded DNA fragments from PCR and other enzymatic reactions. All centrifugation steps were carried out at 16,168 x g (13,200 rpm) in a bench-top microcentrifuge. 5 volumes of Buffer PB was added to 1 volume of the PCR sample and mixed by inverting several times. Sample was applied to the spin column and centrifuged for 1 min to bind the DNA. The flow-through was discarded and 0.75 ml of Buffer PE was added to wash the column and centrifuged for 1 min. The flow-through was discarded and the column was centrifuged for an additional 1 min to remove the residual wash buffer. The spin column was placed into a clean 1.5 ml Eppendorf tube and the DNA was eluted by adding 30 µl molecular grade water to the centre of the QIAquick membrane, left to stand for 1 min at room temperature and then centrifuged for 1 min. The concentration of DNA was measured using NanoDrop (Thermoscientific, USA) and stored at -20°C until further use.
2.12 Extraction of DNA from agarose gel

The QIAquick Gel Extraction Kit (Qiagen, UK) was used to purify a particular PCR band after cutting from a gel. This protocol was designed for DNA extraction of up to 10 µg. All centrifugation steps were carried out at 16,168 x g (13,200 rpm) in a table-top microcentrifuge. The gels were run with a low voltage (4 V/cm) to separate the fragments properly. When the run was completed, the gels were viewed briefly under the UV light to see the separation. The desired DNA fragment was excised from the agarose gel with a clean and sharp scalpel and placed in an Eppendorf tube and weighed. 3 volumes of Buffer QG was added to 1 volume gel (100 mg ~ 100 µL) and incubated at 50°C with occasional shaking in vortex until the gel slice completely dissolved. 1 gel volume of isopropanol (300 uL isopropanol for 300 mg gel) was then added and mixed by inverting several times. For the large gel slice, this mixing step is carried out in a 15 mL conical tube. The sample was applied to the spin column and centrifuged for 1 min to bind the DNA. The flow-through was discarded and the spin column was placed back into the same tube. 0.75 ml of Buffer PE was added to the spin column and centrifuged for 1 min to wash the DNA. The flow-through was discarded and once more, the spin column was placed back into the 2 ml collection tube and centrifuged for an additional 1 min to remove the residual wash buffer. To elute the DNA the spin column was placed into a clean 1.5 ml Eppendorf tube. The DNA was eluted by adding 30 µl molecular grade water and centrifuged for 1 min. The concentration of eluted DNA was measured using a NanoDrop. The purity and integrity of the extracted DNA was checked again in 1% agarose gel.
2.13 Cloning of PCR products

2.13.1 Cloning into pGEM-T Easy

Purified PCR products were ligated into pGEM-T Easy vector (Promega, UK). pGEM-T Easy vectors are linearized vectors with a single 3’-terminal thymidine at both ends (Figure 2-2). The T-overhangs at the insertion site ligate with the 3'-A overhangs of PCR products. The ligation reaction (10 µL) was performed according to the manufacturer’s protocol. In a cold Eppendorf tube kept on ice, 5 µL of 2x Rapid Ligation Buffer, 1 µL of pGEM-T Easy vector (50ng), the calculated amount of PCR product, and 1 µL of T4 DNA ligase (3 Weiss units/µl) were taken and the volume was adjusted to 10 µL with molecular water. The reaction was mixed by pipetting and incubated for 1hr at room temperature or overnight at 4°C.

Figure 2-2. Map of pGEM-T Easy vector (Promega, UK) (30 May 2016).
2.13.2 Cloning into pET-28a(+) vector

The pET-28a(+), one of the widely-used vectors used for expression and purification of proteins in *E. coli*, was used in this study. This expression vector has a T7lac promoter for efficient expression, has a multiple cloning site with a wide range of restriction sites and it adds an N-terminal 6xHis-tag to the expressed protein, thus facilitating purification by affinity chromatography. pET vectors also contain a gene encoding a *lac* repressor. In *E. coli* BL21 (DE3) or similar strains with a lysogen, this lac repressor acts both at the *lacUV5* promoter in the host chromosome to repress transcription of the T7 RNA polymerase gene by the host polymerase and at the T7lac promoter in the vector to block transcription of the target gene by any T7 RNA polymerase (pET Manual, Novagen).

The cloning of PCR products into pET-28a(+) (Figure 2-3) was done according to the pET system manual (Novagen). The vectors were prepared by digesting with specific restriction enzymes following the protocol of the enzyme manufacturers and then gel-purified to remove uncut plasmid. A typical 30 µL vector digestion reaction contains: 3 µg of plasmid, 3 µL of 10x reaction buffer, 10-20 U of each restriction enzymes and molecular grade water up to 30 µL. The reactions were incubated at 37°C for 1 hour and heat inactivated at 65°C for 15 minutes followed by PCR purification. The inserts were prepared by digesting the PCR products (amplified using primers with desired restriction enzyme sites and high-fidelity polymerase) with the restriction enzymes and PCR purified using the same method for vector preparation.

The next step was ligation of digested insert with sticky ends into the digested pET-28a(+) vector. A typical 20 µL ligation reaction contained the following
components: 2 uL 10x ligation buffer, 100 ng prepared pET vector, 80-100 ng prepared target gene insert, 1 uL T4 DNA ligase and the required amount of water to make the total volume 20 uL. A negative control reaction was prepared without adding T4 DNA ligase. The reaction was incubated at 16°C for 30 minutes in a PCR machine and heat inactivated at 65°C for 15 minutes. The ligation was confirmed by running an aliquot of negative reaction and original reaction on a 1% agarose gel.

2.13.2.1 Cloning of ddl6, ddl7, ddlAEc and ddlTd into pET-28a(+) vector

The coding sequences of ddl6 and ddl7 were amplified using TddlF and Tddl28aR primers and cloned into BamHI and XhoI sites of pET-28a(+) vector. The plasmids pGEM-T Easy::intI-attI-ddl6 and pGEM-T Easy::intI-attI-ddl7 were used as templates to amplify ddl6 and ddl7, respectively. The ddlA of E. coli (GenBank accession number: AM946981.2) was amplified using the forward (EC-ddlA-F) and reverse (EC-ddlA-R) primers (Chapter 2). The genomic DNA of E. coli BL21 (DE3) was used as a template for PCR amplification of ddlAEc. The primers for PCR amplification of ddl of T. dentiocla ATCC35405 (1103 bp) (GenBank accession number: AE017226:2414439-2415542) using DdlTd-F and DdlTd-R. The metagenomic DNA of UK and Bangladesh were used as template to amplify the coding sequence for ddlTd. All of the PCR was performed using Q5 High Fidelity 2X Master Mix using 100-200 ng of the genomic or plasmid DNA as template and 0.4 µM of each primer.

The plasmids were purified and the inserts were sequenced to check the accuracy and orientation. The purified plasmids with correct inserts were then transformed into E. coli BL21 (DE3) using the same protocol. E. coli BL21 (DE3)
strains are deficient of Lon protease (cytoplasm) and outer membrane protease, OmpT and is compatible with T7lac promoter system (Structural Genomics, Architecture et Fonction des Macromolécules et al. 2008). The strain also contains a lysogen of bacteriophage DE3 that carries the gene carries the lacI gene, the lacUV5 promoter, and the gene for T7 RNA polymerase. lacUV5 promoter is IPTG inducible and thus addition of IPTG is required to maximally induce expression of the T7 RNA polymerase in order to express recombinant genes cloned downstream of a T7 promoter.

Figure 2-3. Vector map of pET-28a(+) used for cloning and expression of ddl genes in E.coli BL21 (DE3). The map was created using SnapGene software (from GSL Biotech; available at snapgene.com).
2.14 Preparation of *E. coli* competent cells

Competent *E. coli* cells were prepared by one-step protocol described by Chung, Niemela et al. (1989). Briefly, a 5 ml overnight culture of *E.coli* α-select or *E. coli* BL21 (DE3) was prepared in LB broth. 500 µL of this culture was added to 50 ml fresh LB broth and grown at 37°C in a shaking incubator (200 rpm) until the OD<sub>600</sub> becomes 0.5. The cells were centrifuged at 1910 x g for 10 minutes at 4°C and the supernatant was discarded. The precipitated cells were re-suspended in 5 ml TSS (10% polyethylene PEG 8000, 10 mM MgSO<sub>4</sub>.7H<sub>2</sub>O, 5% DMSO and LB broth). The cells were kept on ice for 10 minutes and dispensed in 100 µL volume to Eppendorf tubes and stored in -80°C.

2.15 Preparation of *Bacillus subtilis* competent cells

Competent *B. subtilis* cells were prepared using the method described by Hardy (1985). A single fresh colony of *B. subtilis* 168 was inoculated in 10 mL SPI broth [a 100 mL SPI media contains 25 mL 4x SP broth, 2.5 mL 20% glucose, 2.86 mL Thymine (3.5 mg/mL), 5 mL amino acid solution, 64.6 mL autoclaved H<sub>2</sub>O] and incubated overnight at 30°C at 200 rpm. The 4x SP broth (pH 7.2) was prepared by adding 8 gm/L (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>, 56 gm/L K<sub>2</sub>HPO<sub>4</sub>, 24 gm/mL KH<sub>2</sub>PO<sub>4</sub>, 4 gm/L C<sub>6</sub>H<sub>5</sub>O<sub>7</sub>Na<sub>3</sub>H<sub>2</sub>O, 0.8 gm/L MgSO<sub>4</sub>.7H<sub>2</sub>O, 0.8 gm/L casamino acids (BD Biosciences), 4 gm/L yeast extract and H<sub>2</sub>O up to 1L. A 20 mL amino acid solution in water contained 20 mg histidine, 20 mg threonine and 20 mg methionine (the strain is auxotrophic for these amino acids) which was filter sterilized using a 0.22 µM filter. 10 mL of the overnight culture were inoculated in 100 mL prewarm SPI broth and were grown at 37°C until the end of the log phase. 10 mL of this culture was added to 90 mL prewarm SPII broth (identical
to SPI except the amino acid solution is excluded) and incubated for 90 minutes in a 37°C shaking incubator. The cells were harvested by centrifugation at 1910 x g for 10 minutes at 20°C and re-suspended in 10 mL of the supernatant containing 10% glycerol. 500 µL of the suspension was dispensed in Eppendorf tubes kept on ice and stored at -80°C.

2.16 Transformation of ligation reactions into *E. coli* α-Select

The ligation mixtures were transformed into *Escherichia coli* α-select competent cells (Bioline, UK) by heat shock at 42°C in a bench-top water bath for 40 seconds following the manufacturer’s protocol. For pGEM-T Easy cloning, the transformants were selected on Luria-Bertani (LB) agar containing 100 µg/mL ampicillin, 40 µg/ml X-Gal and 0.4 mM IPTG for blue-white colony screening. For pHCMC05 and pET-28a(+) cloning, the same method for transformation was used, however, the transformants were selected on LB plates containing only the antibiotics for selection i.e. 100 µg/mL ampicillin and 30 µg/mL kanamycin, respectively.

2.17 Transformation of *E. coli* BL21 (DE3)

The pET-28a(+) or pGEM-T Easy vectors containing the correct inserts were transformed into competent *E.coli* BL21 (DE3) cells. About 200-300 ng of pET-28a(+) was transformed into 200 µL of competent *E.coli* BL21 (DE3) cells using similar protocol used for the transformation of ligation reaction. The transformants were selected on LB agar plates supplemented with 30 µg/mL kanamycin or 100 µg/mL ampicillin depending on the types of plasmids used.
2.18 Transformation of *B. subtilis*

Chemically competent *B. subtilis* 168 cells were transformed with pHCMC05 plasmids. For transformation, an aliquot (500 µL) of competent cells were quickly thawed at 37°C and were added to a 50 ml conical tube containing 5 µg of plasmids. The tube was incubated with gentle agitation (50 rpm) for 1 hour at 37°C. After the incubation 5 ml room temperature LB broth was added and incubated for further 1.5 h at 37°C with vigorous shaking. The whole transformation was spread on BHI plates supplemented with 10 µg/mL of chloramphenicol (200 uL of transformants per plate) and incubated for 16-18 hours at 37°C incubator.

2.19 Colony PCR

The colony PCR method was used for determining the presence or absence of insert DNA in plasmid constructs. To do this a 30 µL PCR reaction using BioMix Red was prepared containing primers targeting the insert gene or the primers targeting the vector DNA flanking the insert. A single colony was picked from the plate using a sterile micropipette tip and mixed in the PCR reaction and then the tip was touched on a fresh LB plate containing appropriate antibiotic for selection. The colonies were grown on the plates at 37°C for 12-16 hours. The initial heat step of the PCR lyses the cells and the DNA is released which the primers use as template. After the PCR was finished, the presence or absence of insert in the plasmids were checked on a 1% agarose gel.
2.20 Plasmid DNA isolation

White colonies (for blue-white screening) were picked from the agar plates using sterile micropipette tips and inoculated in 5 mL of LB broth supplemented with 100 μg/mL ampicillin (for pGEM-T Easy vector) or with 30 μg/mL kanamycin (for pET vector) and were incubated overnight. Plasmid DNA was isolated by using QIAprep Spin Miniprep Kit (Qiagen, UK) following the manufacturer’s instructions.

The presence of the insert in a plasmid was verified by a 10μl DNA digestion reaction, containing appropriate restriction enzyme, 50-200 ng of DNA and molecular grade water (Sigma, UK) up to 10 μL. The reactions were incubated at 37°C for 1 hour and electrophoresed on 1% agarose gel.

2.21 DNA sequencing and analysis

The inserts in the plasmids or the purified PCR products were sequenced by the commercial service at the Beckman Coulter Genomics (Beckman Coulter Genomics, UK). The primers used for sequencing the inserts are listed in Table 2-2. For sequencing large inserts additional primers were designed. Different bioinformatics tools used for sequence analysis are listed in Table 2-3.

Table 2-3. Software and tools used for nucleotide and amino acid sequence analysis

<table>
<thead>
<tr>
<th>Program/Software/Tools</th>
<th>Use/Purpose</th>
<th>Reference/Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bio Edit Sequence Alignment Editor</td>
<td>Sequence alignment and restriction mapping</td>
<td>(Hall 1999)</td>
</tr>
<tr>
<td>Nucleotide BLAST (megablast, discontiguous blast and blatsn)</td>
<td>To identify the query sequence or to find sequences similar to query sequence</td>
<td>(Altschul, Gish et al. 1990)</td>
</tr>
<tr>
<td>Translated BLAST (blastx)</td>
<td>To find similar proteins to translated query in a protein database</td>
<td></td>
</tr>
<tr>
<td>Program/Software/Tools</td>
<td>Use/Purpose</td>
<td>Reference/Source</td>
</tr>
<tr>
<td>------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------------------------------------------------</td>
</tr>
<tr>
<td>Standard Protein BLAST (blastp)</td>
<td>To identify the query sequence or find protein sequences similar to the query</td>
<td><a href="http://blast.ncbi.nlm.nih.gov/Blast.cgi">http://blast.ncbi.nlm.nih.gov/Blast.cgi</a></td>
</tr>
<tr>
<td>PSI-BLAST</td>
<td>To find members of a protein family or build a custom position-specific score matrix</td>
<td></td>
</tr>
<tr>
<td>CD-search (RPS-BLAST)</td>
<td>To find conserved domains in the query</td>
<td></td>
</tr>
<tr>
<td>VecScreen</td>
<td>Screen for vector contamination</td>
<td></td>
</tr>
<tr>
<td>ORF finder</td>
<td>To find Open Reading Frames in the sequence</td>
<td><a href="http://www.ncbi.nlm.nih.gov/gorf/gorf.html">http://www.ncbi.nlm.nih.gov/gorf/gorf.html</a></td>
</tr>
<tr>
<td>EMBOSS Transeq</td>
<td>To translate nucleic acid sequences to their corresponding amino acid sequences</td>
<td><a href="http://www.ebi.ac.uk/Tools/st/emboss_transeq/">http://www.ebi.ac.uk/Tools/st/emboss_transeq/</a></td>
</tr>
<tr>
<td>Reverse Complement</td>
<td>To convert a DNA sequence into its reverse, complement, or reverse-complement counterpart.</td>
<td><a href="http://www.bioinformatics.org/sms/rev_comp.html">http://www.bioinformatics.org/sms/rev_comp.html</a></td>
</tr>
<tr>
<td>BPROM</td>
<td>To predict the presence of bacterial promoters</td>
<td>(V. Solovyev 2011)</td>
</tr>
<tr>
<td>Clustal Omega</td>
<td>To generate alignments between three or more sequences</td>
<td><a href="http://www.ebi.ac.uk/Tools/msa/clustalo/">http://www.ebi.ac.uk/Tools/msa/clustalo/</a></td>
</tr>
<tr>
<td>MEGA6</td>
<td>Sequence alignment, inferring phylogenetic trees, and testing evolutionary hypotheses.</td>
<td>(Tamura, Stecher et al. 2013)</td>
</tr>
<tr>
<td>T-Coffee</td>
<td>To align sequences or to combine the output of different alignment methods (Clustal, Mafft, Probcons, Muscle, etc.) into one unique alignment (M-coffee).</td>
<td>(Notredame, Higgins et al. 2000)</td>
</tr>
</tbody>
</table>

The phylogenetic analyses, based on amino acid and nucleotide sequences were performed by applying the Neighbour-joining method (Saitou and Nei 1987). The phylogenetic trees were drawn using the MEGA6 software (http://www.megasoftware.net/) (Tamura, Stecher et al. 2013).
2.22 Agar dilution method for determination of minimum inhibitory concentrations (MIC)

The MIC of D-cycloserine, vancomycin and beta-lactams (ampicillin, amoxicillin and penicillin G) was determined by agar dilution method following the guidelines of Clinical and Laboratory Standards Institute (CLSI). Three identical fresh and pure colonies of *E. coli* or *B. subtilis* 168 strains were inoculated in Mueller-Hinton (MH) broth supplemented with appropriate antibiotics and grown at 37°C for 4-5 hrs. The OD_{600} was adjusted to 0.08 to 0.1 using sterile saline. MH agar plates supplemented with various concentration of antibiotics starting from 0.125 to 256 µg/ml were prepared and inoculated with the adjusted bacterial suspensions with a multipoint inoculator. The spots were allowed to dry for 30 minutes and the plates were then incubated aerobically at 37°C for 16-20 hrs. The MIC value was defined as the lowest concentration of the compounds giving rise to no visible growth. To confirm the results all MIC determinations were performed at least three times in triplicate. As a control, *E. coli* or *B. subtilis* 168 strains transformed with empty vectors were used.

2.23 Preparation of sample for dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE)

The pellets in the Eppendorf tubes were thawed on ice. The supplied 5x Protein Loading Buffer (National Diagnostics, UK) was diluted to 1x with water and 100 µL of the 1x buffer is added to the pellet, mixed by pipetting and boiled at 95°C for 5 minutes. 10 µL of the solution was then loaded into the cast sodium SDS-PAGE gel.
To prepare purified protein samples for SDS-PAGE, 80 µL of protein were mixed with 20 µL of 5x Protein Loading Buffer and boiled for 5 minutes at 95°C.

2.24 Casting a discontinuous SDS-PAGE Gel

All of the SDS-PAGE gels used in the study were prepared using the discontinuous buffer systems according to the protocol developed by Laemmli (Laemmli 1970). In this system, the gel is divided into an upper "stacking" gel of low percentage (i.e. large pore size) and low pH (6.8) and a resolving gel with a pH of 8.8 with much smaller pores. Both gels contain only Cl⁻ as the mobile anion. The tank buffer contains glycine as its anion, at a pH of 8.8.

As the predicted size of proteins to be analysed in this study falls in the range of 20-80 kDa, a 10% gel was prepared in all cases. A 10 ml casting solution for resolving gel contained: 3.3 mL Protogel, 2.5 mL ProtoGel Buffer and 4.06 mL Deionized H₂O. The components are mixed in a 50 mL conical tube and 100 µL of fresh 10% ammonium persulfate (APS) solution was added and swirled gently to mix. Then 10 µL of TEMED was added and swirled gently to mix. Maximum care was taken to avoid bubble formation. The solution was poured into the gel cassette keeping a space on top to pour stacking gel. Immediately after pouring the resolving gel into the cassette, the gel was overlayed with water to exclude O₂ and to ensure a flat interface between the resolving and stacking gels. The gel was allowed to polymerize for 30 minutes.

When the resolving gel was polymerized, a 5 mL of casting solution for stacking gel was prepared by mixing 650 µL Protogel, 1.25 ml ProtoGel Stacking Buffer and 3.05 mL Deionized H₂O. The water on the top of resolving gel in the cassette is drained by inverting the gel. Then 50 µL of 10% APS and 5 µl
TEMED was added to the stacking gel solution and the top of the cassette was filled with this mixture. The comb was inserted and the gel was allowed to polymerize for 30-60 minutes. The tank was filled with 1X Tris-Glycine SDS and 10 µL of the prepared proteins samples was loaded into the wells using a Hamilton syringe (Sigma, UK). The gels were run at 200 volts for 1 hour or until the run is complete.

2.25 16S metagenomic profiling

To verify that the extracted metagenomic DNA of pooled samples of UK and Bangladesh represents the diverse group of oral bacteria, PCR amplicon libraries of the V5-V7 hypervariable region of the 16S rRNA were generated. This study was done in collaboration with Dr. Morgana Vianna of Cardiff University, UK. PCR was performed using the forward primer 785F (GGATTAGATACCCBRGTAGTC) and the reverse primer 1175R (ACGTCRTCCCCDCCTTCCTC) (Kraneveld, Buijs et al. 2012). The primers included adaptors fused to the 5′ end of the 16S rDNA bacterial primer sequence and a unique 10 nt sample identification key.

The PCR reactions were performed using MolTaq 16S basic Master Mix (Molzyme). Each 25 µL PCR reaction contained 2.5 µL mix, 2 µL of each primer, 5 ng of the metagenomic DNA and molecular grade water up to 25 µL. The PCR condition was as follows: 95°C for 5 minutes, followed by 29 cycles at 95°C for 30 seconds, 55°C for 40 seconds and 72°C for 60 seconds. The final elongation was done at 72°C for 10 minutes. PCR products were cleaned up by using AMPpure XP beads (Beckman Coulter). The quality and size of the amplicons were analysed by Agilent Bioanalyzer using Agilent DNA 1000 kit.
The quantity of DNA was measured using Qubit® dsDNA BR Assay Kit (Thermofisher Scientific). Sequencing was performed using MiSeq Illumina systems. Analysis of the sequences was done using Quantitative Insights into Microbial Ecology (QIIME)-pipeline (Caporaso, Kuczynski et al. 2010). The figures were generated using Microsoft Excel 2016.

### 2.26 Submission of DNA sequences to GenBank

The DNA sequences were submitted to GenBank to get accession numbers for each sequence. The sequences of inserts including the sequences which contained the first GCs were deposited in the DNA database under accession numbers from KT921469 to KT921473. The accession numbers from KT921496 to KT921509 represent sequences from Bangladeshi samples and KT921510 to KT921531 represent sequences from UK the samples. The sequences of the 2024 bp PCR amplicons containing *ddl*6 and *ddl*7 were submitted separately to GenBank with the accession number of KU886208 and KU886209, respectively. The 4421 bp sequence of PCR product containing the upstream sequence of *intI* along with the downstream *ddl*7 was submitted to GenBank with the accession number KY039278. This whole-genome shotgun sequence of *Eggerthia catenaformis* has been deposited GenBank under the accession no. NCVR00000000.
Chapter Three: Detection of integrons and associated gene cassettes in the human oral metagenome
3.1 Introduction

This chapter considers the detection of integrons in oral metagenomic DNA using a PCR-based metagenomic approach. The diversity of orientations and putative functions carried by the ORFs embedded within the GCs are also presented. The features of the conserved attC sites have also been discussed.

3.2 Materials and Methods

3.2.1 Design of primers to target the integron gene cassettes

The primers targeting the class 1, class 2 and class 3 integrons were designed based on information in published literature. The list of primers and their sequences are shown in Table 2-2 and the target sites for the primers are shown in Figure 3-1.

The sequences of intI and attC of the *T. denticola* ATCC 35405 reverse integron (Coleman, Tetu et al. 2004) were used to design novel primers to target the conserved sequences of intI and attC in different *Treponema* species. For instance, the TDIF primer was designed from the sequence that encodes the conserved amino acid sequence SSQNQAL of the IntI (Appendix-IV) which is conserved in almost all species of *Treponema*. The attC-based primers were designed using the conserved sequence of the boundaries of attC. The location of the primers on the sequences are shown in Figure 3-1. As the attC sites associated with the GC of the *T. denticola* integron are nearly identical (Coleman, Tetu et al. 2004), it is expected that forward attC primers will bind to the 3'-end of attC of one GC and the reverse primers will bind to the 5'-end of
attC of other GCs, giving rise to products with single or multiple GCs flanked by the primer binding sites.

**Figure 3-1.** The target sites of primers used to detect different types of integrons in the oral metagenomic DNA. The primers MARS2 and MARS3 designed based on the attC of the unusual integron of *T. denticola* are shown aligned with their target binding sites. The core sites R" and L", and R' and L' have been indicated by bold fonts.

### 3.2.2 PCR amplification of GCs and cloning into pGEM-T Easy

The protocol to amplify the integrons and associated GCs from the metagenomic DNA of saliva has been described in Chapter 2.
3.2.3 Analysis of the sequences for the features of GCs

A sequence obtained using the attC-based primers was considered a putative GCs if (i) it contains both of the primer sequences designed from conserved nucleotides of attC (ii) at the downstream of the forward primers there is a presence of R' core site and at the upstream of reverse primers there is a presence of R'' core site (iii) the primer binding sites flank one or more putative ORF beginning with ATG, TTG or GTG (Elsaied, Stokes et al. 2007). The sequences which did not contain an ORF, but contained the attC sites, were considered as empty GCs. The putative translated sequences were subjected to BlastX searches and matches were considered significant if the e-value was <0.001.

A flow diagram of the detection and analysis of integrons in the oral cavity is shown in the Figure 3-2.
3.2.4 Nomenclature of the gene cassettes

The GCs were named according to the name of the primers and the source of saliva sample. The first two letters indicate the forward and reverse primers used for amplification, respectively. The third letter indicates the source of oral metagenomic DNA from which GCs were detected (U for UK; B for Bangladesh). This is followed by a numerical code for the number of clone. For example, TMB1 means this GC is obtained from clone 1 of the PCR products of the Bangladeshi samples amplified by using the primers TDIF and MARS2.

Figure 3-2. Flow chart of the methods used to detect integrons and associated gene cassettes in the oral metagenomic DNA.
3.2.5 Submission of the sequences to GenBank

The sequences were deposited to the GenBank database for accession numbers. The accession numbers for all submissions are given in Chapter 2.

3.3 Results

3.3.1 High molecular weight metagenomic DNAs were isolated from the saliva samples collected from UK and Bangladesh

The high molecular weight DNA with a size larger than 10 kb was isolated from the pooled saliva of UK and Bangladeshi sample (Figure 3-3). In 1% agarose gel the band for the extracted DNA appeared above the 10 kb band of the ladder.

![Agarose gel electrophoresis of the metagenomic DNA isolated from pooled saliva samples of UK (left) and Bangladeshi (right) volunteers. HyperLadder I (Bioline) was used as size marker (200 bp-10 kbp). The gel was imaged in Alphalmager gel documentation system (Alpha Innotech). Extraction was done in triplicate.](image)
3.3.2 16S metagenome profiling of pooled saliva samples of the UK and Bangladesh shows that they are truly metagenomic

To confirm that the metagenomic DNA isolated from pooled saliva samples of the UK and Bangladesh are truly metagenomic and contain bacteria of different, representative phyla, the V5-V7 region of 16S rRNA were amplified and sequenced. The results show that among the 13 phyla reported to be present in the human oral cavity (Dewhirst, Chen et al. 2010), 10 can be detected in the saliva of both cohorts (Figure 3-4). Among them the predominant phyla were Bacteroidetes, Firmicutes, Proteobacteria, Actinobacteria, Fusobacteria and TM7 (Figure 3-4). The abundance of Bacteroidetes and Firmicutes was found to be higher in the UK sample compared to the Bangladeshi, however, three other phyla showed the opposite pattern. The abundance of Proteobacteria was three-fold and Fusobacteria and TM7 was approximately two-fold higher in the Bangladeshi sample than the UK (Figure 3-4).

![Abundance of different bacterial phyla in pooled saliva of UK and Bangladeshi individuals.](image)

**Figure 3-4.** Abundance of different bacterial phyla in pooled saliva of UK and Bangladeshi individuals.
When the abundance OTUs were analysed according to the phylogenetic order, bacterial species of 18 and 17 orders were detected in the UK and Bangladeshi sample, respectively. No reads were found for *Bifidobacteriales* in the Bangladeshi sample. The *Bacteroidales* and *Clostridiales* were found to represent the majority of the OTUs (Figure 3-5). The abundance of *Neisseriales*, which belong to the phylum *Proteobacteria* was found to be 4-fold higher in the Bangladeshi sample, compared to the UK sample (Figure 3-5).

**Figure 3-5.** Abundance of bacteria in the oral cavity in the pooled saliva sample of UK and Bangladesh grouped according to their order.

### 3.3.3 Class 1 and class 2 integrons could not be recovered in the salivary metagenomic DNA of UK and Bangladeshi samples

The published primers that were successfully used in previous studies (Figure 3-1) to amplify the GC array of integrons in both clinical and environmental samples were used to screen the presence of class 1 and class 2 integrons in
saliva. The PCR products generated by these primers (Figure 3-6) were cloned and sequenced. It was found that none of the PCR amplicons carries the features of GCs and they were produced due to nonspecific binding of the primers. Many of the inserts analysed matched to human DNA. For instance, 11 clones were analysed from the library of PCR amplicons (Figure 3-6) produced by the degenerate primers intI-864R (binds conserved intI sequence) and GCP2 (binds conserved attC), and found that 9 of them aligned to bacterial DNA and 2 aligned to human DNA. However, the bacterial sequences did not carry the features of a GC and they were apparently the product of non-specific binding of the primers. Similarly, the GC-PCR approach was applied by using the PCR primers GCP1 and GCP2 targeting the flanking regions of attC site, however, the PCR failed to generate any product.

**Figure 3-6.** PCR products generated by previously published primers using the UK metagenomic DNA as template. The name of the primers used are given under the figures. The PCR were run in triplicate.

The library of clones constructed using the products of HS458 and HS459 (Figure 3-6B); ICC48 and ICC21 primers (Figure 3-6C) (Gillings, Labbate et al.)
2009, Huang, Cagnon et al. 2009, Betteridge, Partridge et al. 2011) also
produced amplicons not related to integrons and GCs. The primers hep74 and
hep51 (White, McIver et al. 2001) were used to detect class 2 integrons,
however, analyses of the sequences of the clones also showed that they are
the products of non-specific binding. In addition to the published primers, some
novel primers (such as ARC1 and ARC2) were also designed targeting the
class 1 integron cassette array, but they also did not work.

![Image]

**Figure 3-7.** A) Products of integrase PCR using the UK metagenomic DNA as
template. B) EcoRI digestion of plasmids prepared from randomly picked clones to find
the clone containing 280 bp insert which is the expected size of the intI1 PCR.

Although the published primers failed to amplify the GC array, using the primers
targeting the intI1 [intI1F and intI1R (Goldstein, Lee et al. 2001)], the presence of
intI1 in the oral metagenome could be detected. The 280 bp PCR product of
intI1 (Figure 3-7) was cloned into pGEM-T Easy and sequenced. The insert was
found to be 100% identical to the intI1 of *E. coli*. This result prompted the use of
the \textit{intI1F} primer with the novel \textit{E. coli attC}-based reverse primer ARC1 and ARC2 as well as published primer ICC21 to amplify the cassette array of class1 integron. This attempt also failed to produce integron related amplicons. This finding confirms that the oral cavity harbours \textit{intI1}, but, the gene cassette array associated with them could not be recovered. The PCR products generated by the \textit{intl2} and \textit{intl3} specific primers (Figure 3-7 A) were also cloned and the right size inserts were sequenced, but all of the sequences were the result of non-specific binding of the primers.

3.3.4 PCR products having all the structural features of an integron were recovered from both UK and Bangladeshi samples

When the published primers failed to detect the presence of GC array of mobile integrons, new primers were designed based on the reverse integron of the oral spirochete \textit{T. denticola} (Coleman, Tetu et al. 2004). The PCR was performed by using the \textit{intI}-based primer TDIF (designed based on the sequence that encodes the conserved amino acid sequence SSQNLQR of IntI of the \textit{T. denticola} integron) coupled with the \textit{attC}-based primer MARS2 (Figure 3-1).

The resulting amplicons (Figure 3-8A; Figure 3-9D) were cloned into pGEM-T Easy vector and plasmids were prepared from a total of 35 randomly selected clones and digested with EcoRI (Figure 3-8B; Figure 3-8C and Figure 3-9B, Figure 3-9C) to see the size of the inserts. The inserts of 22 clones were sequenced. Inserts had a size range of 1244 bp to 2024 bp. All of these inserts contained the basic features of an integron. Within the amplicons, a major part of \textit{intI} (768 bp of 1266 bp), the full length putative \textit{attI} site and a putative integron promoter, \textit{Pc} were detected. A total of 7 unique amplicons containing 6
different GCs including one empty GC with no identifiable ORF were found. Each of the amplicons had only one GC downstream of the *attI* implying that MARS2 primed to the first *attC* site in the array. The predicted *attI* sites were located at c. 96-98 bp downstream of the stop codon of the integrase.

The putative ORFs detected within the GCs had a size range of 258 to 1035 bp. All but one ORF detected within the first GCs had putative RBS at less than 8-bp upstream of the ORFs. The ORFs were oriented at the same direction of the *intI* similar to a reverse integron (Table 3-2).

![Figure 3-8](image)

**Figure 3-8.** PCR products obtained using TDIF and MARS 2 primers using the metagenomic DNA of UK saliva samples (A). Agarose gel electrophoresis of the EcoRI digest (B and C).
digest of the plasmids containing different size inserts (B-C). The numbers on the top indicate the number of the clones. The inserts were sequenced and analysed.

**Figure 3-9.** PCR products obtained using TDIF and MARS2 primers using the metagenomic DNA of Bangladeshi saliva samples (A). Agarose gel electrophoresis of the EcoRI digest of the plasmids containing different size inserts (B-C). The numbers on the top indicates the number of the clones.

3.3.5 The partial *intI* sequences associated with *attI* and first GCs were related to different *Treponema sp.*

The recovered partial *intI* sequences located at the 5'-ends of the inserts were analysed to see their percentage identity with other *intI* in the NCBI database. The percentage of nucleotide identity of the *intIs* with the NCBI nr/nt and NCBI
whole genome sequence (WGS) databases falls in the range of 85-100% (Table 3-1). The *int* of clone TMU3 (Table 3-1) exhibited 100% identity to the *int* of *T. denticola* MYR-T (GenBank accession no.: AGDX01000006.1: 604170-604934) and *T. denticola* H1-T (GenBank accession no.: AGDW01000016.1: 20283-21047).

**Table 3-1.** Percentage identity of partial *int*Is with their closest homologues in GenBank

<table>
<thead>
<tr>
<th><em>int</em></th>
<th>Maximum Identity (%) (BlastN)</th>
<th>Homology source</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>int</em> <em>TMB1</em></td>
<td>95</td>
<td><em>Treponema denticola</em> H-22</td>
</tr>
<tr>
<td><em>int</em> <em>TMB3</em></td>
<td>99</td>
<td><em>Treponema denticola</em> H-22</td>
</tr>
<tr>
<td><em>int</em> <em>TMB4</em></td>
<td>85</td>
<td><em>Treponema vincentii</em> ATCC 35580</td>
</tr>
<tr>
<td><em>int</em> <em>TMU3</em></td>
<td>100</td>
<td><em>Treponema denticola</em> MYR-T, <em>Treponema denticola</em> H1-T</td>
</tr>
<tr>
<td><em>int</em> <em>TMU18</em></td>
<td>96</td>
<td><em>Treponema socranskii</em> subsp. paredis ATCC 35535</td>
</tr>
<tr>
<td></td>
<td>95</td>
<td><em>Treponema denticola</em> ATCC35405</td>
</tr>
<tr>
<td><em>int</em> <em>TMU6</em></td>
<td>99</td>
<td><em>Treponema denticola</em> H-22</td>
</tr>
<tr>
<td><em>int</em> <em>TMU7</em></td>
<td>99</td>
<td><em>Treponema denticola</em> H-22</td>
</tr>
<tr>
<td><em>int</em> <em>TMB7</em></td>
<td>98</td>
<td><em>Treponema denticola</em> H-22</td>
</tr>
</tbody>
</table>

The *int* of the clone TMB3 was aligned with the genome of *T. denticola* H-22 along with the putative *att* and the associated GC encoding cof-like hydrolase (Accession: AGDV01000005.1: 20559-18805=1755 bp). This is the only amplicon in the library of first GCs that can be mapped across their whole length to an entry in the GenBank database. For all other cases, the gene in the first GC detected in this study could not be found in the first GC of the *Treponema* integrons in the NCBI nr/nt and NCBI WGS databases, however,
they were found to be located within the genome of different *Treponema* sp., except for the first GC of the clone TMB1. For instance, the GC associated with the *intI* of the clone TMU18 that encodes a hypothetical protein (Table 3-2) could not be found in the GC array of the integron of *T. socranskii*. The GC was aligned with a GC of *T. denticola* ATCC 33520 with 99% nucleotide identity (Table 3-3).

The phylogenetic tree constructed using the partial *intI*s showed that the *intI* of clone TMB4 which was associated with the empty GCs was diverged from other *intI*s in the tree (Figure 3-10).

![Phylogenetic tree of partial intI sequences detected in the oral metagenomic DNA.](image)

**Figure 3-10.** Phylogenetic tree of partial intI sequences detected in the oral metagenomic DNA. The evolutionary history was inferred using the Neighbour-Joining method and the evolutionary distances were computed using the Maximum Composite Likelihood method. Evolutionary analyses were conducted in MEGA6. The intI1 of *P. aeruginosa* was used as outgroup.

### 3.3.6 The cassette promoters, Pc were found to be located upstream of first GCs and within the attI sites

The putative *attI* sites were detected upstream of the first GCs in all inserts of the first GC library. The typical features of an *attI* site (Partridge, Recchia et al.
were detected by inspection of the sequences. In agreement with the
definition of attI site (Partridge, Recchia et al. 2000), all of the attI sites in our
library of first gene GCs were found to contain two putative integrase binding
sites S1 and S2 (also termed R and L, respectively) (Partridge, Recchia et al.
2000, Elsaeid, Stokes et al. 2011) and two direct repeat (DR) sequences (DR1
and DR2) which are known to be involved in integrase binding (Collis, Kim et al.
1998) (see chapter 1). The conserved 5'-GTT-3' triplet where the recombination
of GCs takes place between G and TT (Partridge, Recchia et al. 2000) were
found to be located at S1 (Figure 3-11).

\[ \text{Figure 3-11. Multiple sequence alignment of putative attI sites detected in the} \\
\text{clones containing all the features of integrons located upstream of the first gene} \\
cassettes with attI1, attI2 and attI3 (Partridge, Recchia et al. 2000, Collis and
Hall 2004). The inversely oriented integrase binding sites were labelled as S1
and S2. The recombination point as well as -35 sequence and -10 sequence of
the putative Pc have also been indicated. The sequences after the
recombination point are shown in lower case.} \]

The DR sequences located at the upstream of the core recombination site had
a nearly conserved typical sequence of GTTTTTTA and GTTTAGR (R=A or G)
for DR1 and DR2, respectively (Figure 3-11). Unlike class 1 integrons, the putative \( P_c \) promoters were found to be located within the \( attl \) sites (Figure 3-11). One putative strong promoter with a consensus sequence of -35 TTGYAA|17bp|-10 TATWGT (Y = C or T; W = A or T) was identified which was found to be similar to one of the putative \( P_c \) promoters found on \( T. \) denticola integron (Coleman, Tetu et al. 2004).

### 3.3.7 The core sites of \( attC \) were identified at the 3'-end of first GCs

The 7 bp core site \( R'' \) (1L) of \( attC \) was detected upstream of the reverse MARS2 primer having the consensus sequence RYY(/R)YAAC (Figure 3-12) (R = A or G; Y = C or T). In most cases, the stop codons of the ORFs were located at these \( R'' \) integrase binding sites of \( attC \) (Stokes, O'Gorman et al. 1997, Mazel 2006).

![Diagram of integron structure](attachment:image.png)

**Figure 3-12.** Detection of the core site \( R'' \) with a nearly consensus RYYYAAC sequence on the first GCs in the library of PCR amplicons. The location of the reverse primer (MARS2) which contains the core site \( L'' \) has been indicated with an arrow.
3.3.8 The first GCs encoding proteins of different types of functions were recovered

Among the 22 clones sequenced from both cohorts, 8 clones (TMB3/5/6/10/11/13/14/16) harboured an ORF (768 bp) encoding a protein homologous to cof-like hydrolase of *T. putidum*. The 1757 bp insert containing cof-like hydrolase GC along with its upstream sequence containing *attI* and *intI* could be mapped to WGS contig of *T. denticola* H-22 (GenBank accession no: AGDV01000005.1: 18805-20559). The whole insert exhibited 94% nucleotide sequence identity to this contig of *T. denticola* H-22. The coding sequence of the cof-like hydrolase gene exhibits 98% nucleotide sequence identity.

**Table 3-2.** Orientation of the first GCs in the cassette library (not mapped to scale). TDIF and MARS 2 are the primer binding sites. The size of the inserts has been shown with numbers at ends.

<table>
<thead>
<tr>
<th>Clones</th>
<th>Type of orientation</th>
<th>Size of the inserts and orientation of the putative ORFs within the GCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMB1/8/1 2/15</td>
<td>A</td>
<td><img src="https://via.placeholder.com/150" alt="Diagram" /></td>
</tr>
<tr>
<td>TMB3/5/6/10/11/13/1 4/16</td>
<td>A</td>
<td><img src="https://via.placeholder.com/150" alt="Diagram" /></td>
</tr>
<tr>
<td>TMB4</td>
<td>B</td>
<td><img src="https://via.placeholder.com/150" alt="Diagram" /></td>
</tr>
<tr>
<td>Clones</td>
<td>Type of orientation</td>
<td>Size of the inserts and orientation of the putative ORFs within the GCs</td>
</tr>
<tr>
<td>------------</td>
<td>---------------------</td>
<td>-----------------------------------------------------------------------</td>
</tr>
<tr>
<td>TMU6/7/19</td>
<td>A</td>
<td>[\text{TDIF} \quad 1 \quad \text{intI} \quad \text{attI} \quad \text{Pc} \quad \text{ddl (1035-bp)} \quad \text{MARS2} \quad \text{attC} \quad 2024]</td>
</tr>
<tr>
<td>TMU3/4/11</td>
<td>A</td>
<td>[\text{TDIF} \quad 1 \quad \text{intI} \quad \text{attI} \quad \text{Pc} \quad \text{ORF (780-bp)} \quad \text{MARS2} \quad \text{attC} \quad 1775]</td>
</tr>
<tr>
<td>TMB7/9</td>
<td>A</td>
<td>[\text{TDIF} \quad 1 \quad \text{intI} \quad \text{attI} \quad \text{Pc} \quad \text{ddl (1035-bp)} \quad \text{MARS2} \quad \text{attC} \quad 2024]</td>
</tr>
<tr>
<td>TMU18</td>
<td>A</td>
<td>[\text{TDIF} \quad 1 \quad \text{intI} \quad \text{attI} \quad \text{Pc} \quad \text{ORF (390-bp)} \quad \text{MARS2} \quad \text{attC} \quad 1692]</td>
</tr>
</tbody>
</table>

The 258 bp ORF present within the first GC of the clones TMB1/8/12/15 had no homologous sequence in the NCBI database. No conserved domains could be found in this ORF. Another GC detected within clones TMU3/4/11 with an ORF of 780 bp was found to encode a hypothetical protein of *T. denticola* (Table 3-3). The ORF along with its 50 bp upstream sequence until the G↓TT recombination site and partial *attC* site at the downstream were found to be located within the WGS contigs of different strains of *T. denticola* including US-Trep acdtB (GenBank: AGEB01000011: 46082-46934), ASLM (GenBank: AGDR01000041:32839-33691), AL-2 (GenBank: AGDQ01000016:2133-2986)
and SP32 (GenBank: AHAC01000019:105-964). However, none of the contigs containing the homologues of this gene were found to harbour an *intI* gene. The GC carrying the 390 bp ORF in the clone TMU18 was also found to be present in four strains of *T. denticola* including MYR-T (GenBank: AGDX01000008:11445-12210), H1-T (GenBank: AGDW01000016: 2328-3093), ATCC33520 (GenBank: GDS01000012: 38777-39542) and F0402 (GenBank: ADEC01000014: 597526-598311). The homologue located within contig 12 of the strain ATCC33520 was found to be located within a GC array, however, the distance between the *intI* and the homologue gene was approximately 18 kb. The upstream sequence of the ORF up to the G↓TT recombination site and the downstream partial *attC* were found to 100% identical, which suggests a potential rearrangement of the array of GC as shown above for the GC in the clone TMU3/4/11.

One clone was found to carry an empty first GC (TMB4) at the first position of the integron GC array (Table 3-3).

A 1035 bp ORF encoding D-alanine-D-alanine ligases (Ddl5) was found within the GC of 5 clones: TMB7/9 and TMU6/7/19. By analysing sequences of *ddl* (1035 bp) within the first GC of 5 clones, two variants of the gene were identified. The variants were named as *ddl6* and *ddl7* and they were different from each other by two nucleotide changes at c.490 and c.777 position of the coding sequence.
<table>
<thead>
<tr>
<th>Gene cassettes/clone code</th>
<th>Primer pair</th>
<th>Size of insert including the integrase gene and first cassette (bp)</th>
<th>GenBank accession number of the submitted sequence</th>
<th>Distance between integrase gene and ORF of first gene cassette (bp)</th>
<th>Size of the ORFs</th>
<th>BlastN (ORF of first gene cassette)</th>
<th>BlastX (ORF of first gene cassette)</th>
<th>GenBank accession number of the homologous proteins (BlastX)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMB1/8/12/15</td>
<td>TDIF-MARS2</td>
<td>1499</td>
<td>KT921469</td>
<td>271</td>
<td>258</td>
<td>No significant similarity</td>
<td>No significant similarity</td>
<td>-</td>
</tr>
<tr>
<td>TMB3/5/6/10/11/13/14/16</td>
<td>TDIF-MARS2</td>
<td>1757</td>
<td>KT921470</td>
<td>198</td>
<td>768</td>
<td>Treponema denticola H-22</td>
<td>Cof-like hydrolase [Treponema denticola]</td>
<td>96</td>
</tr>
<tr>
<td>TMB4</td>
<td>TDIF-MARS2</td>
<td>1244</td>
<td>KT921471</td>
<td>-</td>
<td>-</td>
<td>No ORF</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TMU6/7/19</td>
<td>TDIF-MARS2</td>
<td>2024</td>
<td>KU886208</td>
<td>198</td>
<td>1035</td>
<td>Treponema pedis B. 683</td>
<td>D-alanine-D-alanine ligase [Treponema pedis]</td>
<td>97</td>
</tr>
<tr>
<td>TMB7/9</td>
<td>TDIF-MARS2</td>
<td>2024</td>
<td>KU886208</td>
<td>198</td>
<td>1035</td>
<td>Treponema pedis B. 683</td>
<td>D-alanine-D-alanine ligase [Treponema pedis]</td>
<td>97</td>
</tr>
<tr>
<td>TMU18</td>
<td>TDIF-MARS2</td>
<td>1692</td>
<td>KT921473</td>
<td>161</td>
<td>387</td>
<td>Treponema denticola ATCC 33520</td>
<td>Hypothetical protein [Treponema denticola]</td>
<td>99</td>
</tr>
</tbody>
</table>
3.3.9 GCs with different size and orientations were amplified using *attC*-based primers

GCs (other than the first GCs) were amplified using different combinations of *attC*-based forward (MARS3/MARS4/MARS5) and reverse (MARS1/MARS2) primers (Figure 3-1). Almost all of the amplicons produced by primers with degenerate nucleotides (MARS1, MARS2 and MARS3) were found to be the product of non-specific binding. However, the primer combination, MARS5 and MARS2, having no degenerate nucleotides results in amplicons carrying the features of a typical GC (Figure 3-13A).

![Figure 3-13. A) PCR products generated using *attC*-based forward (MARS5) and reverse (MARS2) primers using Bangladeshi metagenomic DNA. B, C and D) EcoRI digestion of the plasmids to see the size of the inserts.](image-url)
A total of 47 clones were sequenced from the library of PCR amplicons obtained using a MARS2 and MARS5 primers (Figure 3-13 and Figure 3-14). Analysis of the sequences results in the detection of a total of 36 unique GCs having the features of an integron GC. The cassettes were found to be flanked by the primer binding sites.

All of the sequenced inserts from the library of GC-PCR were shown to contain a single GC. This strong bias towards single GC can be explained by the competitive nature of PCR to favour shorter PCR products. The size of the

Figure 3-14. A) PCR products generated using attC-based forward (MARS5) and reverse (MARS2) primers using UK metagenomic DNA. B and C) EcoRI digestion of the plasmids to see the size of the inserts.
cassettes ranged from 423 to 1144 bp. Of the 36 GCs, 10 had no identifiable ORFs (empty cassettes) and the remaining 26 GCs contained one or more putative ORFs giving a total of 39 different ORFs with a size range between 117 to 804 bp.

By analysing the arrangement of genetic features within the GCs recovered using attC-based primers, it was found that they were arranged in five different ways (Table 3-4) as defined by the direction, position and number of ORFs within the GCs. The type C orientation having a single ORF accounted for the majority found in 15 cassettes. A total of 9 GCs harboured two ORFs and only 2 GCs had 3 ORFs in tandem (Table 3-4). The sequences of the clones containing two or more ORFs were examined for the presence of other putative attC sites in between the ORFs and none of which were found to have attC in between them.

Table 3-4. Orientation of the putative ORFs in the gene cassettes (For orientation type A and B, please see the Table 3-2.)

<table>
<thead>
<tr>
<th>Orientation type</th>
<th>Orientation of ORFs</th>
<th>Number of unique gene cassettes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>UK</td>
<td>Bangladesh</td>
</tr>
<tr>
<td>C</td>
<td><img src="attachment" alt="attC" /></td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>D</td>
<td><img src="attachment" alt="attC" /></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>E</td>
<td><img src="attachment" alt="attC" /></td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>F</td>
<td><img src="attachment" alt="attC" /></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td><img src="attachment" alt="attC" /></td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Total GCs</td>
<td></td>
<td>14</td>
<td>22</td>
</tr>
</tbody>
</table>
Out of 39 putative ORFs detected in 36 GCs, 34 had a RBS located less than 8 bp upstream of the predicted start codons (Table 3-6). An obvious promoter could be identified for the ORFs encoding toxin-antitoxin modules only.

### 3.3.10 The core sites of attC (R' and R'"") were identified on the GCs

The sequences obtained from the library of GC-PCR were also analysed to detect the presence of core sites of attC. As discussed in the methods section of this chapter, the attC primers were designed such that the amplified products contain the R' (1R) core site in all GCs after the forward primer sequence (MARS5) and the complementary R" (1L) core sites upstream of the reverse primer sequence (MARS2) (Figure 3-15).

![Figure 3-15](image)

**Figure 3-15.** Core sites R' (1R) and R" (1L) abutting the forward (MARS5) and reverse attC primers (MARS2) in two representative GCs (MMU26 and MMB3/9) recovered in this study. The location of primers binding sites are marked with arrow.

It was found that all the putative GCs had the R' and R" core sites flanked by the ORFs in GCs and majority of the R' and R" core sites (34 out of 36) were 100% complementary with each other (Table 3-5). The consensus 5'-GTTRRY-3' and 5'-RYYYAAC-3' motifs for R' and R" were most common and represented by 66.66% of GCs (24 out of 36). Approximately
16.0% GCs (6 out of 36) had the R' with a consensus 5'-GTTRYRY-3' and the R" with consensus 5'-RYRYAAC-3' i.e., for these sets, the variation with the most common pattern occurs at the 5th and 3rd base pair of the R' and R" core sites, respectively (Table 3-5). The identification of these core sites flanked by the ORFs confirms that the putative GCs are not PCR artefacts and is consistent with the attC structure of a GC (Stokes, O'Gorman et al. 1997).

**Table 3-5. Core sites R' (1R) and R" (1L) abutting the forward and reverse attC primer sequence on the gene cassettes.**

<table>
<thead>
<tr>
<th>Sequence of R' after the forward primer sequence of the attC on GCs</th>
<th>Pattern of R' sequence of the GC</th>
<th>Sequence of R&quot; before the reverse primer of the attC on GCs</th>
<th>Pattern of R&quot; sequence of the attC on GCs</th>
<th>Clones</th>
<th>Complementarity between the pattern of R' and R&quot; core sites of the attC on GCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTTAGAC GTTRRRY GTCTAAC RYYYYAC MMU2 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY GTCTAAC RYYYYAC MMU19 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY GTCTAAC RYYYYAC MMU23 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY GTCTAAC RYYYYAC MMU24 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY GTCTAAC RYYYYAC MMB2 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY GTCTAAC RYYYYAC MMB4 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY GTCTAAC RYYYYAC MMB5 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY GTCTAAC RYYYYAC MMB19 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY GTCTAAC RYYYYAC MMB20 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY GTCTAAC RYYYYAC MMB32 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY GTCTAAC RYYYYAC MMB34 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY GTCTAAC RYYYYAC MMB36 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY GTCTAAC RYYYYAC MMB37 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY GTCTAAC RYYYYAC MMB38 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY GTCTAAC RYYYYAC MMB39 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY ATCTAAC RYYYYAC MMU9 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY ATCTAAC RYYYYAC MMB28 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY ATCTAAC RYYYYAC MMB15 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY ATCTAAC RYYYYAC MMB17 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY ATCTAAC RYYYYAC MMB39 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGGC GTTRRRY GCTTAAC RYYYYAC MMB18 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGGT GTTRRRY ACCTAAC RYYYYAC MU18 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGGT GTTRRRY ACCTAAC RYYYYAC MU20 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGGT GTTRRRY ACCTAAC RYYYYAC MMB22 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTAGAC GTTRRRY ACCTAAC RYYYYAC MU7 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTATAC GTTRYRY ATCTAAC RYYYYAC MMB14 6/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTATAC GTTRYRY GCTAAC RYYYYAC MMB28 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTATGT GTTRYRY ACATAAC RRYYAC MU11 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTATGT GTTRYRY ACCTAAC RYYYYAC MU25 6/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTTATGT GTTRYRY ACATAAC RRYYAC MMB11 7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sequence of R’ after the forward primer sequence of the attC on GCs

<table>
<thead>
<tr>
<th>Sequence of R’ after the forward primer sequence of the attC on GCs</th>
<th>Pattern of R’ before the reverse primer of the attC on GCs</th>
<th>Sequence of R” sequence of the attC on GCs</th>
<th>Pattern of R” core sites of the attC on GCs</th>
<th>Clones</th>
<th>Complementarity between the pattern of R’ and R” core sites of the attC on GCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTTATGT</td>
<td>GTTRYRY</td>
<td>ACATAAC</td>
<td>RYRYAAC</td>
<td>MMB23</td>
<td>7/7</td>
</tr>
<tr>
<td>GTTAGGA</td>
<td>GTTRRRR</td>
<td>TCCTAAC</td>
<td>YYYYAAC</td>
<td>MU27</td>
<td>7/7</td>
</tr>
<tr>
<td>GTTAGCA</td>
<td>GTTRRRY</td>
<td>TGCTAAC</td>
<td>YRYYAC</td>
<td>MU3</td>
<td>7/7</td>
</tr>
<tr>
<td>GTTAGCA</td>
<td>GTTRRRY</td>
<td>TGCTAAC</td>
<td>YRYYAC</td>
<td>MMB33</td>
<td>7/7</td>
</tr>
<tr>
<td>GTTAGCT</td>
<td>GTTRRYY</td>
<td>AGCTAAC</td>
<td>RRYYAC</td>
<td>MMB30</td>
<td>7/7</td>
</tr>
<tr>
<td>GTTAGT</td>
<td>GTTRRRY</td>
<td>ACCTAAC</td>
<td>RRRYAC</td>
<td>MU26</td>
<td>7/7</td>
</tr>
<tr>
<td>ATTGAAC</td>
<td>ATTRRRY</td>
<td>ATCTAAC</td>
<td>RYYYYAC</td>
<td>MMB3</td>
<td>7/7</td>
</tr>
</tbody>
</table>

3.3.11 Diversity of the functions of putative proteins encoded by ORFs within the GCs detected by novel attC primers

Out of 39 putative ORFs detected on 36 unique GCs amplified by using attC primers, 36 (92.3%) had at least one homologue in the NCBI database and 3 ORFs did not have any homologue. Only 17 of the 36 ORFs (47.2%) were found to encode proteins that matched with the proteins of known function and the remaining 22 matched to hypothetical proteins. In terms of sequence similarity of the ORFs with GenBank, it was found that the predicted amino acid sequence of 27 of the 39 ORFs (69.2%) exhibited >90% identity with their homologous proteins. When the nucleotide sequence identity was considered, only 9 ORFs were found to exhibit >98% sequence identity with their homologues in GenBank. This indicated that the majority of the GCs recovered in this study using attC-based primers are novel and have not been detected previously.

The putative ORFs detected on the GCs were predicted to encode proteins of diverse functions including antibiotic resistance, host adaptation to stress and competence (Table 3-6). Two putative antibiotic resistance genes were found
among the cassette ORFs. BlastX searches showed that the clone MMB22 contained an ORF that encoded a protein with 99% identity to streptogramin A O-acetyltransferase of *T. denticola*. The single ORF (390-bp) present in the clone MMU7 was predicted to encode a glyoxalase/bleomycin antibiotic binding protein. Proteins related to adaptation to stress include different toxin-antitoxin systems and a twitching motility protein, PilT. The clones containing the ORFs encoding toxin-antitoxin system includes MMB23 and MMB38 which encoded peptidase (antitoxin)-PemK (toxin) and higA (antitoxin)-higB (toxin), respectively. The PilT protein is encoded by MMB3/9.

Most of the proteins encoded by the ORFs on GCs showed a similarity with proteins in the database of *Treponema* spp., mostly from *T. denticola* (14 out of 36) followed by *T. putidum, T. medium, T. pedis, T. socranskii, T. vincentii, T. phagedenis, T. maltophilum* and Treponema sp. OMZ-838. This observation supports the previous reports that *T. denticola, T. vincentii and T. phagedenis* carry chromosomal integrons (Wu, Rho et al. 2012, Wu, Doak et al. 2013). However, 15 ORFs has been identified related to other *Treponema* spp. that were not reported previously to carry GC including *T. putidum, T. medium, T. pedis, T. maltophilum* and *T. socranskii*. Only three ORFs out of 36 were predicted to encode proteins related to non-treponemes including those from *Klebsiella oxytoca, Clostridium drakei* and *Mariprofundus ferrooxydans*, however, the homologies of the proteins with these species were low (<70%) at the amino acid level.
<table>
<thead>
<tr>
<th>Gene cassettes/clone code</th>
<th>Primer pair</th>
<th>Cassette Size (bp)</th>
<th>Orientation Types</th>
<th>BlastN (Whole GC)</th>
<th>BlastX (ORF only)</th>
<th>Distance between attC and ORF (bp)</th>
<th>Accession number of the homologous proteins (BlastX)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMU2</td>
<td>MARS5-MARS2</td>
<td>519</td>
<td>G</td>
<td>Treponema denticola 81 72 6e-94</td>
<td>-</td>
<td>-</td>
<td>- KT921496</td>
</tr>
<tr>
<td>MMU3</td>
<td>MARS5-MARS2</td>
<td>826</td>
<td>C</td>
<td>No significant similarity found - - -</td>
<td>hypothetical protein SK92_04217 [Klebsiella oxytoca] 753 32 6e-28</td>
<td>Yes</td>
<td>KLY26699 63 KT921497</td>
</tr>
<tr>
<td>MMU7</td>
<td>MARS5-MARS2</td>
<td>765</td>
<td>C</td>
<td>Treponema denticola MYR-T 100 100 0.0</td>
<td>Bleomycin resistance family protein [Treponema] 390 100 4e-87</td>
<td>Yes</td>
<td>WP_0026888988 148 KT921498</td>
</tr>
<tr>
<td>MMU9/13</td>
<td>MARS5-MARS2</td>
<td>725</td>
<td>C</td>
<td>Treponema phagedenis 62 73</td>
<td>Hypothetical protein [Treponema phagedenis] 474 53 1e-45</td>
<td>Yes</td>
<td>WP_044634793 127 KT921499</td>
</tr>
<tr>
<td>MMU11</td>
<td>MARS5-MARS2</td>
<td>734</td>
<td>G</td>
<td>Treponema denticola 81 97 0.0</td>
<td>-</td>
<td>-</td>
<td>- KT921500</td>
</tr>
<tr>
<td>MMU18</td>
<td>MARS5-MARS2</td>
<td>690</td>
<td>C</td>
<td>Treponema denticola F0402 97 100 0.0</td>
<td>Hypothetical protein [Treponema denticola] 459 97 1e-77</td>
<td>Yes</td>
<td>EGC77532 174 KT921501</td>
</tr>
<tr>
<td>MMU19</td>
<td>MARS5-MARS2</td>
<td>431</td>
<td>G</td>
<td>Treponema denticola F0402 83 84 2e-80</td>
<td>-</td>
<td>-</td>
<td>- KT921502</td>
</tr>
<tr>
<td>MMU20</td>
<td>MARS5-MARS2</td>
<td>943</td>
<td>C</td>
<td>Treponema denticola MYR-T 88 31 2e-77</td>
<td>No significant similarity found 780 - -</td>
<td>No</td>
<td>- 72 KT921503</td>
</tr>
<tr>
<td>Gene cassette/Clone code</td>
<td>Primer pair</td>
<td>Cassette Size (bp)</td>
<td>Orientation Types</td>
<td>BlastN (Whole GC)</td>
<td>BlastX (ORF only)</td>
<td>Distance between attC and ORF (bp)</td>
<td>Accession number of the GCs</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------------</td>
<td>--------------------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>-----------------------------------</td>
<td>----------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Closest homologue</td>
<td>Percentage identity (%)</td>
<td>Coverage (%)</td>
<td>E-value</td>
</tr>
<tr>
<td>MMU24</td>
<td>MARS5-MARS2</td>
<td>104</td>
<td>D</td>
<td>Treponema socranski subsp. paredis ATCC 35535</td>
<td>96</td>
<td>81</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>114</td>
<td>F</td>
<td>Treponema maltophilum ATCC 21939</td>
<td>93</td>
<td>44</td>
<td>2e-166</td>
</tr>
<tr>
<td>MMU25</td>
<td>MARS5-MARS2</td>
<td>972</td>
<td>C</td>
<td>Treponema denticola SP44</td>
<td>98</td>
<td>100</td>
<td>0.0</td>
</tr>
<tr>
<td>MMU27</td>
<td>MARS5-MARS2</td>
<td>896</td>
<td>E</td>
<td>Treponema denticola SP32</td>
<td>99</td>
<td>99</td>
<td>0.0</td>
</tr>
<tr>
<td>Gene cassette/Clone code</td>
<td>Primer pair</td>
<td>Cassette Size (bp)</td>
<td>Orientation Types</td>
<td>BlastN (Whole GC)</td>
<td>BlastX (ORF only)</td>
<td>Distance between attC and ORF (bp)</td>
<td>Accession number of the homologous proteins (BlastX)</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>----------------------------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>MMU28</td>
<td>MARS5-MARS2</td>
<td>987 E</td>
<td>Treponema putidum 758</td>
<td>Prevent-host-death protein [Treponema putidum]</td>
<td>282 91 1e-58 Yes WP_044979047</td>
<td>320</td>
<td>KT921509</td>
</tr>
<tr>
<td>MMU23</td>
<td>MARS5-MARS2</td>
<td>634 G</td>
<td>Treponema denticola F0402</td>
<td>Hypothetical protein (Plasmid stabilization protein) [Treponema putidum]</td>
<td>213 85 1e-53 Yes AIN94307</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MMB1/2</td>
<td>MARS5-MARS2</td>
<td>633 G</td>
<td>Treponema denticola F0402</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MMB3/9</td>
<td>MARS5-MARS2</td>
<td>846 C</td>
<td>Treponema pedis B.683</td>
<td>Twitching motility protein PilT [Treponema putidum]</td>
<td>414 70 5e-59 No AIN93467</td>
<td>369</td>
<td>KT921511</td>
</tr>
<tr>
<td>MMB5/3 5</td>
<td>MARS5-MARS2</td>
<td>423 G</td>
<td>Treponema denticola</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MMB4/1 2/31</td>
<td>MARS5-MARS2</td>
<td>877 D</td>
<td>Treponema denticola</td>
<td>Toxin [Treponema denticola]</td>
<td>288 95 2e-56 No WP_01069226</td>
<td>271</td>
<td>KT921512</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hypothetical protein [Treponema pedis]</td>
<td>258 99 9e-32 Yes WP_024468276</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene cassette/Clone code</td>
<td>Primer pair</td>
<td>Cassette Size (bp)</td>
<td>Orientation Types</td>
<td>BlastN (Whole GC)</td>
<td>BlastX (ORF only)</td>
<td>Ribosomal binding site</td>
<td>Accession number of the homologous proteins (BlastX)</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>-----------------------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>MMB11/13/21/25/29</td>
<td>MARS5-MARS2</td>
<td>629 C</td>
<td>Treponema denticola</td>
<td>98 81 0.0 Membrane protein [Treponema denticola]</td>
<td>312 94 2e-44 Yes WP_002669840</td>
<td>98 KT921514</td>
<td></td>
</tr>
<tr>
<td>MMB15</td>
<td>MARS5-MARS2</td>
<td>889 C</td>
<td>No significant similarity found</td>
<td>- - -</td>
<td>Hypothetical protein [Mariprofundus ferrooxydans]</td>
<td>804 46 6e-70 Yes WP_009850619</td>
<td>31 KT921516</td>
</tr>
<tr>
<td>MMB22</td>
<td>MARS5-MARS2</td>
<td>854 C</td>
<td>Treponema denticola US-Trep acdB</td>
<td>99 100 0.0 Vat family streptogramin A O-acetyltransferase [Treponema denticola]</td>
<td>645 99 2e-152 Yes WP_00267814</td>
<td>160 KT921521</td>
<td></td>
</tr>
<tr>
<td>MMB14</td>
<td>MARS5-MARS2</td>
<td>862 E</td>
<td>Treponema vincentii F0403</td>
<td>89 88 0.0 RelB/DinJ antitoxin [Treponema putidum]</td>
<td>417 97 3e-57 Yes WP_044978223</td>
<td>86 KT921515</td>
<td></td>
</tr>
<tr>
<td>MMB17</td>
<td>MARS5-MARS2</td>
<td>838 E</td>
<td>Treponema denticola</td>
<td>99 100 0.0 Hypothetical protein [Treponema denticola]</td>
<td>396 98 2e-89 Yes WP_010957066</td>
<td>49 KT921517</td>
<td></td>
</tr>
<tr>
<td>MMB18</td>
<td>MARS5-MARS2</td>
<td>792 C</td>
<td>Treponema putidum</td>
<td>98 100 0.0 Hypothetical protein [Treponema denticola]</td>
<td>117 100 8e-16 Yes WP_010957065</td>
<td>26 KT921518</td>
<td></td>
</tr>
<tr>
<td>MMB19</td>
<td>MARS5-MARS2</td>
<td>817 C</td>
<td>Treponema denticola</td>
<td>94 100 0.0 Hypothetical protein [Treponema denticola OTK]</td>
<td>708 90 5e-147 Yes EMB24958</td>
<td>45 KT921519</td>
<td></td>
</tr>
<tr>
<td>Gene cassette/Clone code</td>
<td>Primer pair</td>
<td>Cassette Size (bp)</td>
<td>Orientation Types</td>
<td>BlastN (Whole GC)</td>
<td>BlastX (ORF only)</td>
<td>Distance between attC and ORF (bp)</td>
<td>Accession number of the homologous proteins (BlastX)</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------------</td>
<td>--------------------</td>
<td>-------------------</td>
<td>------------------</td>
<td>-------------------</td>
<td>-------------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Closest homologue</td>
<td>Percentage identity (%)</td>
<td>Coverage (%)</td>
<td>E-value</td>
<td>Closest homologue</td>
</tr>
<tr>
<td>MMB20</td>
<td>MARS5-MARS2</td>
<td>531</td>
<td>G</td>
<td>Treponema denticola ATCC 35405</td>
<td>82</td>
<td>100</td>
<td>2e-96</td>
</tr>
<tr>
<td>MMB23/27</td>
<td>MARS5-MARS2</td>
<td>901</td>
<td>D</td>
<td>Treponema denticola ATCC 35405</td>
<td>65</td>
<td>69</td>
<td>9e-53</td>
</tr>
<tr>
<td>MMB28</td>
<td>MARS5-MARS2</td>
<td>543</td>
<td>C</td>
<td>Clostridium scatologenes strain ATCC 25775</td>
<td>75</td>
<td>42</td>
<td>2e-33</td>
</tr>
<tr>
<td>MMB30</td>
<td>MARS5-MARS2</td>
<td>808</td>
<td>C</td>
<td>Treponema pedis isoM1224</td>
<td>99</td>
<td>-</td>
<td>0.0</td>
</tr>
<tr>
<td>MMB32</td>
<td>MARS5-MARS2</td>
<td>632</td>
<td>G</td>
<td>Treponema denticola FO402</td>
<td>84</td>
<td>98</td>
<td>5e-108</td>
</tr>
<tr>
<td>MMB33</td>
<td>MARS5-MARS2</td>
<td>810</td>
<td>E</td>
<td>Treponema denticola</td>
<td>99</td>
<td>99</td>
<td>0.0</td>
</tr>
<tr>
<td>MMB34</td>
<td>MARS5-MARS2</td>
<td>714</td>
<td>C</td>
<td>Treponema medium ATCC 700293</td>
<td>95</td>
<td>100</td>
<td>0.0</td>
</tr>
<tr>
<td>Gene cassette/Clone code</td>
<td>Primer pair</td>
<td>Cassette Size (bp)</td>
<td>Orientation Types</td>
<td>BlastN (Whole GC)</td>
<td>BlastX (ORF only)</td>
<td>Distance between attC and ORF (bp)</td>
<td>Accession number of the GCs</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>----------------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Closest homologue</td>
<td>Percentage identity (%)</td>
<td>Coverage (%)</td>
<td>E-value</td>
<td>Closest homologue</td>
</tr>
<tr>
<td>MMB36</td>
<td>MARSS-MARS2</td>
<td>732</td>
<td>G</td>
<td>Treponema denticola ATCC 35404</td>
<td>80</td>
<td>100</td>
<td>9e-84</td>
</tr>
<tr>
<td>MMB37</td>
<td>MARSS-MARS2</td>
<td>635</td>
<td>G</td>
<td>Treponema denticola</td>
<td>82</td>
<td>98</td>
<td>1e-87</td>
</tr>
<tr>
<td>MMB38</td>
<td>MARSS-MARS2</td>
<td>947</td>
<td>F</td>
<td>Treponema socranski subsp. paredis ATCC 35535</td>
<td>96</td>
<td>100</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Treponema socranski</td>
<td>282</td>
<td>98</td>
<td>4e-67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>318</td>
<td>97</td>
<td>5e-56</td>
</tr>
<tr>
<td>MMB39</td>
<td>MARSS-MARS2</td>
<td>857</td>
<td>E</td>
<td>Treponema medium ATCC 700293</td>
<td>96</td>
<td>94</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>357</td>
<td>96</td>
<td>5e-76</td>
</tr>
</tbody>
</table>

151
3.4 Discussion

Among the different metagenomic approaches to detect integrons, the PCR-based approach is widely used. The use of PCR primers targeting the conserved regions of integrons and selectively amplifying the GC array have been used successfully in previous studies (Nield, Holmes et al. 2001, Stokes, Holmes et al. 2001, Holmes, Gillings et al. 2003, Elsaied, Stokes et al. 2007, Elsaied, Stokes et al. 2011). However, all of these metagenomic studies were carried out using non-human environmental samples. Most of the metagenomic studies involving human microbiota, were either sequence-based (Qin, Li et al. 2010) focusing on the recovery of all genetic features or focusing on a function of interest such as antibiotic resistance (Sommer, Dantas et al. 2009). No studies have been reported previously on metagenomes obtained from human saliva to detect integrons using a PCR approach.

In this chapter, an analysis of the diversity of integron GCs amplifiable in the metagenomic DNA samples collected from healthy volunteers of the UK and Bangladesh has been provided. The previously published primers were found to be ineffective in amplifying products carrying integron GCs. However, as there was no positive control to test the primers, the presence of class 1 and class 2 integrons in the saliva DNA samples should not be ruled out. Although the presence of intI1 was detected, the associated GC array could not be recovered. The reason for the failure to detect the associated GC array by using known and newly designed primers may be that attC-sites associated with the GCs are novel and have not been found before. When the known primers for class 1 and class 2 integrons failed to amplify GC array, some novel primers were designed based on the structural features of the reverse integron of T.
*denticola* ATCC 35405 (Coleman, Tetu et al. 2004) and these primers successfully amplified a diverse array of GCs including those in the first position, most of which were novel.

A total of six different GCs have been detected on the first position of integrons including an empty GCs. The *intI*s located upstream of these GCs were closely related and all of them are oriented in the same direction of the first GCs, thereby they belong to the family of reverse integrons.

Unlike class1 and class 3 integrons and the CI of *Pseudomonas stutzeri* (Coleman and Holmes 2005) and *V. cholera* (Krin, Cambray et al. 2014), all of the integrons detected in this study were found to harbour the putative *Pc* within the *attI* site which is a property of class 2 integrons (Collis and Hall 1995). The -35 and -10 sequences of the putative *Pc* were found to be highly conserved and exhibited close similarity with canonical sequence of *E. coli* promoter. The integrase binding sites and the recombination sites were detected at the 3'-end of the *attI* sequences.

Although the chromosomal integron of *T. denticola* ATCC 35405 is the only integron described from oral bacteria (it has 45 gene cassettes in the array), *in silico* analysis of metagenomic data sets from the Human Microbiome Project (HMP) showed that two other *Treponema* species, including *T. vincentii* ATCC 35580 and *T. phagedenis* F0421, also carry integron GCs (Coleman, Tetu et al. 2004, Wu, Rho et al. 2012, Wu, Doak et al. 2013). However, the PCR strategies used in this study, recovered novel GCs that were predicted to encode proteins related to those from genera other than *Treponema* spp.
Analysis of the proteins encoded by the GCs amplified from the oral cavity showed several interesting ORFs. GCs encoding two novel variants of Ddl were found in the library of first GCs of both cohorts. Ddls are one of the key enzymes in peptidoglycan synthesis in bacteria (Neuhaus 1962, Neuhaus 1962). It catalyses the formation of D-ala-D-ala dipeptide in early stage of peptidoglycan synthesis. D-cycloserine (D-cycloserine) is the only clinically available antibiotic that targets Ddl (Neuhaus and Lynch 1964). Generally, *ddl* is a single copy house-keeping gene in bacteria. Although some bacteria such as *E. coli* (Zawadzke, Bugg et al. 1991) and *E. gallinarum* (Ambur, Reynolds et al. 2002) were reported to carry two copies of *ddl*, there is no previous report of identification of *ddl* within an integron. In the next three chapters (Chapter 4, 5 and 6), the in-depth characterisation of these two variants of *ddl* and the proteins they encode are presented.

The GC MMU7 was predicted to encode a protein with 100% amino acid identity to the glyoxalase from multiple species of *Treponema* (WP_002688988.1). Using CDD (Conserved Domain Database) of NCBI (Marchler-Bauer, Bo et al. 2017), the sequence was found to contain a BLMA (bleomycin binding protein)-like domain. Bleomycin is a glycopeptide antibiotic, which inhibits the peptidoglycan synthesis in bacteria, and also used as an antitumour drug which binds to DNA and generates free radicals that result in both double-strand and single-strand DNA breaks (Chen and Stubbe 2005, Kahne, Leimkuhler et al. 2005).

Another ORF (645 bp) found on GC MMB22 detected in the Bangladeshi sample was predicted to encode a Vat family streptogramin A O-acetyltransferase (215 aa) and exhibits 99.0% amino acid identity with the Vat
family streptogramin A O-acetyltransferase from *T. denticola*. Streptogramins are natural antibiotics produced by various *Streptomyces* spp. There are two subgroups of this family of antibiotics, type A (macrolactones) and type B (cyclic hexadepsipeptides) and they are produced simultaneously in the same bacterial species in a ratio of approximately 7:3 (Pechere 1996). The semisynthetic combination of streptogramins, quinupristin/dalfopristin (Synercid), was approved by USFDA in 1999 to treat infections due to vancomycin-resistant *E. faecium* (Winston, Emmanouilides et al. 2000, Hayes, Wagner et al. 2005), however, the natural streptogramin, virginiamycin which is a mixture of virginiamycin M (a group A streptogramin) and virginiamycin S (a group B streptogramin) has been used in agriculture as a growth promoter for more than 45 years (Acar, Casewell et al. 2000). Streptogramin A O-acetyltransferases mediate resistance to type A streptogramins by modifying A-type compounds by acetylation (Allignet, Loncle et al. 1993). Perhaps the streptogramin A O-acetyl transferase in the oral cavity plays an important role in detoxifying the acetyl group containing xenobiotics including residual streptogramins present in our diet.

A cof-like hydrolase gene (a member of haloacid dehalogenase superfamily) was found within a GC amplified using first gene cassette primers (GC TMB3). Cof-like hydrolases are a group of enzymes that inactivate halogenated aliphatic hydrocarbons by hydrolysing the carbon-halogen bonds. They are essential for detoxification of many chlorinated compounds (Hardman 1991, Koonin and Tatusov 1994). Therefore, a cof-like hydrolase in the oral cavity could play a role in detoxifying or inactivating antimicrobials or other compounds with carbon-halogen bonds that are used as antibiotics, pesticides and food
preservatives such as chloramphenicol, atrazine and brominated vegetable oil, respectively.

Another ORF with the GC MMU26 encodes a nitrilase or carbon-nitrogen hydrolase (295 aa), an enzyme belonging to C-N hydrolases superfamly of enzymes. The nitrilases are widely expressed in both prokaryotes and eukaryotes. They catalyse the hydrolysis of various nitriles to the corresponding carboxylic acid and ammonia (Thuku, Brady et al. 2009). In some bacteria such as of *B. subtilis* Strain ZJB-063 (a nitrile-amide-degrading strain), the gene is constitutively expressed and found to specifically hydrolyse arylacetonitrile compounds (Zheng, Chen et al. 2007). The nitrilases present in the oral bacteria may play an important role to detoxify toxic nitriles present in foods (Kupke, Herz et al. 2016).

Another function of predicted GC ORFs was related to the adaptation of bacteria to environmental stress. For example, the twitching motility PilT protein was predicted to be encoded by the ORF in GC of clone MMB3 and MMB9. It has been shown to be involved with type IV fimbria-mediated twitching motility and protease secretion (Han, Kennan et al. 2008). Twitching motility was also shown to play a key role in the development of biofilm from *P. aeruginosa* (Chiang and Burrows 2003). As many oral bacteria can form biofilms on the surfaces in the human oral cavity, having a PilT-encoded GC could help them to develop biofilms and survive environmental stress.

As in previous metagenomic studies to detect integron GCs (Elsaied, Stokes et al. 2007, Koenig, Boucher et al. 2008, Elsaied, Stokes et al. 2011), ORFs predicted to encode proteins with regulatory functions such as toxin-antitoxin (TA) systems have been detected. Two different TA operons including the
HigBA and RelBE were detected on GCs in this study. TA cassettes are usually abundant in chromosomal integrons and are thought to have a role in the stability of the integron GC arrays (Szekeres, Dauti et al. 2007). All of the detected TA cassettes are the members of type II toxin-antitoxin systems (Leplae, Geeraerts et al. 2011). The toxins (HigA and RelE) work by cleaving mRNA, inhibiting translation and exhibit bacteriostatic activity, and the antitoxins (HigB and RelE) can inhibit the action of toxin by protein-protein complex formation (Pandey and Gerdes 2005, Christensen-Dalsgaard and Gerdes 2006, Makarova, Grishin et al. 2006, Jorgensen, Pandey et al. 2009). Two HigBA TA systems have been detected in these GCs (MMU24 and MMB38).

Several recovered GCs did not contain any ORFs. This kind of ORF-less GCs was found both as the first position GC and other GC positions in the integron such as clone TMB4, MMU2, MMB11, MMB23 etc. The presence of empty cassettes have been previously found in cassette arrays comprising, for example, between 4 and 49% of Vibrio spp. cassette arrays (Boucher, Nesbo et al. 2006). Elsaied, Stokes et al. (2011) identified an empty first GC (SuezGC14.1) in the metagenomic DNA of soil samples of Suez Bay. It was previously shown that a Xanthomonas campestris integron GC encoded trans-acting small RNA, which was capable of regulating the virulence in Xanthomonas (Chen, Tang et al. 2011).

Another interesting observation of this study was that none of the GCs except that encodes Ddl were found in both the UK and Bangladeshi library. It could be due to small number of GCs that were picked from a vast pool of GCs. However, it was also assumed that the differences in diet, lifestyle and antibiotic use among the individuals of the two cohorts may have an effect in the GC
diversity in the oral cavity. This interesting finding should be investigated using more samples from two countries and analysing the GC diversity in individual volunteers.

### 3.5 Conclusion

The results discussed in this chapter show that bacterial communities in the human oral cavity harbour integrons and integron GCs encoding proteins predicted to carry out diverse functions which reflects the highly variable physicochemical and stressful environment of the oral cavity. Most of the GCs detected in this study were novel. This suggests that the pool of GCs in the oral cavity is quite large and if more clones would be picked for screening from the library, more novel GCs could be detected. However, the discovery of *ddl* on an integron for the first time was interesting which formed the remaining investigations for my PhD.
Chapter 4: Genetic features and resistance phenotypes of integron-located *ddls*
4.1 Introduction

4.1.1 The role of Ddls in peptidoglycan synthesis in bacteria

Ddls are ATP-dependent enzymes [Enzyme Commission (EC) number: 6.3.2.4)] which belong to the superfamily of D-Ala-D-X ligases. They catalyse the synthesis of D-ala-D-ala dipeptide at the early stage of peptidoglycan synthesis by the following reaction (Neuhaus 1960):

$$\text{ATP} + 2 \text{D-alanine}^{\text{Mg}^{2+}} \rightarrow \text{ADP} + \text{Pi} + \text{D-ala-D-ala}$$

The substrate of Ddl, D-alanine is produced by the alanine racemase (Alr) enzyme which converts L-alanine to D-alanine at the previous step (Figure 4-1). The resulting D-ala-D-ala dipeptide formed by Ddl is polymerised with the UDP-MurNAc-tripeptide to form UDP-MurNAc-pentapeptide precursor units by MurF ligase (Figure 4-1).

When the MurNAc-pentapeptide precursor units reach the inner side of the membrane, they are assembled with undecaprenyl phosphate by translocase to form lipid I intermediates (lipid-anchored disaccharide–pentapeptide monomer) followed by the addition of GlcNAc by the transferase (MurG) to form lipid II intermediates. The lipid II intermediates are then flipped across the membrane and subsequent polymerisation occurs by a transglycosylation reaction. Crosslinking occurs by the actions of transpeptidases and carboxypeptidases which results in the formation of peptide bonds between D-alanine residues of one glycan strand and diaminopimelic acid residues of the adjacent strand (Figure 4-1). Thus, D-alanine plays a critical role in the cross-linking steps during peptidoglycan synthesis.
4.1.2 Mechanism of action of D-cycloserine

Ddl is one of the targets of D-cycloserine, a broad-spectrum antibiotic discovered in 1955 from *Streptomyces garyphalus* (Harris, Ruger et al. 1955). D-cycloserine also inhibits Air which converts L-alanine to D-alanine (Strominger, Ito et al. 1960, Lambert and Neuhaus 1972) (Figure 4-1; Figure 4-2). D-cycloserine is a structural analogue of D-alanine. Strominger, Ito et al. (1960) showed that the inhibitory activity of D-cycloserine is competitively reversed by D-alanine in *S. aureus*. 

*Figure 4-1.* The common biosynthetic pathway of peptidoglycan in bacteria. The role of Ddl and other enzymes in synthesis of peptidoglycan has been shown. The figure was adapted from Bugg, Braddick et al. (2011).
D-cycloserine is a broad-spectrum antibiotic and was found to exhibit inhibitory activity against a wide range of pathogens including *E. coli*, *K. pneumoniae*, *S. typhimurium*, *S. dysenteriae*, *P. aeruginosa*, *Micrococcus pyogenes* and *Diplococcus pneumoniae* as well as rickettsias (Harris, Ruger et al. 1955). Later its effectiveness to treat pulmonary tuberculosis in humans was also found (Epstein, Nair et al. 1956). As D-cycloserine causes severe neurological side effects, its use is currently restricted to treat multi-drug resistant (MDR) and extensively drug resistant (XDR) species of *M. tuberculosis*. Although it is one of the antibiotics discovered during the golden age of antibiotics (1950-1960)(Davies 2006), there is no record of the use of D-cycloserine in agriculture.

### 4.1.3 Mechanisms of D-cycloserine resistance

The known mechanisms of resistance to D-cycloserine in *M. tuberculosis* and other bacteria are shown in Figure 4-3. Overexpression of the D-cycloserine
targets (*alr* and *ddl*) in *M. smegmatis* is a well-studied mechanism of D-cycloserine resistance (Cáceres, Harris et al. 1997, Feng and Barletta 2003). The MIC to D-cycloserine increased two-fold when *ddl* of *M. smegmatis* was cloned into a pMV262 (a multi-copy *E.coli-Mycobacterium* shuttle vector) and overexpressed in *M. smegmatis* (Feng and Barletta 2003). Functional analyses of oral and gut metagenomes of healthy human volunteers have also shown that the expression of heterologous *ddls* cloned into multi-copy expression vector (pZE21) can confer resistance to D-cycloserine at high concentrations (100 µg/mL) in *E. coli* (Sommer, Dantas et al. 2009). Some *ddls* recovered from a paediatric faecal metagenomic library were found to be associated with the genes encoding proteins involved in homologous recombination such as *recR* and *recG* (Moore, Patel et al. 2013). Forsberg, Reyes et al. (2012) identified a novel ORF of unknown function in the soil metagenomic library whose overexpression confers a high level of D-cycloserine resistance. Additionally a nonsynonymous mutation in *Alr* (S22L) in *M. tuberculosis* has been found to be associated with D-cycloserine resistance (Merker, Kohl et al. 2013). Recently some other nonsynonymous mutations in *Alr* including K133E, L89R, M319T, Y364D and R373G were also identified to confer moderate to high level resistance to D-cycloserine in clinical strains of *M. tuberculosis* (Desjardins, Cohen et al. 2016). However, the detailed mechanisms of how these mutations in *Alr* confer resistance to D-cycloserine have not been investigated.
There are several other mechanisms of D-cycloserine resistance in *Mycobacterium* and other bacteria. Mutations in *cycA* gene (G122S) of *E. coli* (Curtiss, Charamella et al. 1965, Russell 1972) and *Mycobacterium bovis* BCG (Chen, Uplekar et al. 2012) have also been found to confer D-cycloserine resistance. CycA is a bacterial D-serine/L- and D-alanine/glycine/D-cycloserine: a proton symporter of the amino acid transporter (AAT) family. Mutation in the transporter gene may lead to the reduced uptake of D-cycloserine into the cell, thus increasing D-cycloserine resistance. However, in *E. coli* the resistance to D-cycloserine due to *cycA* mutation can only be detected in minimal medium. No change in D-cycloserine sensitivity was observed when the susceptibility was determined in non-minimal media such as LB (Baisa, Stabo et al. 2013).

*Figure 4-3. The model of mechanisms of resistance of D-cycloserine in M. tuberculosis and other bacteria including E. coli. Overproduction of targets, target modification by mutation, mutation in the transporter genes, increased efflux and mutation metabolic genes are the known mechanisms of D-cycloserine. Ddl and Alr were represented by blue and orange shapes. D-cycloserine is in blue squares.*
Recently, Baisa, Stabo et al. (2013) observed that mutation in \textit{dadA} and \textit{pnp} which encodes a D-amino acid dehydrogenase and a RNA polynucleotide phosphorylase (PNPase), respectively confers high level resistance to D-cycloserine in \textit{E. coli}. DadA is an enzyme responsible for oxidative deamination of D-amino acids including D-alanine and the by-products of D-alanine catabolism works as a source for carbon, nitrogen and energy (Franklin and Venables 1976). It has been shown that due to mutation in \textit{dadA}, the intracellular concentration of D-ala increases and thus more D-cycloserine is required to inhibit its two targets: Ddl and Alr (Baisa, Stabo et al. 2013) and as a result MIC of D-cycloserine increases.

PNPase is involved in mRNA maturation and degradation (Walter, Kilian et al. 2002). The mechanism of D-cycloserine resistance in \textit{pnp} mutants has not been studied yet, however, it has been hypothesized that in the mutants the abundance and stability of mRNA is altered which may result in increased expression of Ddl and Alr (Baisa, Stabo et al. 2013). They also found that mutations in the genes including \textit{ubiE}, \textit{ubiF}, \textit{ubiG}, \textit{ubiH} and \textit{ubiX} that encode enzymes involved in ubiquinone and menaquinone metabolism in bacteria increased D-cycloserine resistance up to 8-fold compared to their wild-type counterpart of \textit{E. coli} (Baisa, Stabo et al. 2013).

Very recently a novel mechanism of resistance to D-cycloserine has been detected in clinical strains of \textit{M. tuberculosis}. Desjardins, Cohen et al. (2016) have identified that loss of function mutation of \textit{ald} encoding L-alanine dehydrogenase, due to insertion or deletions, confer a moderate level of resistance to D-cycloserine in \textit{M. tuberculosis}. It has been hypothesized that in a non-functional Ald, the conversion L-alanine to pyruvate is inhibited which
increases the pool of available L-alanine to be converted to D-alanine by Alr. Increased D-alanine competitively inhibits D-cycloserine, thus, cell-wall biosynthesis is continued. Although several mutations in Alr were found to increase D-cycloserine resistance, any mutations in Ddl, another target of D-cycloserine, was found to be associated with alteration of D-cycloserine resistance.

As discussed in Chapter 3, during screening the library of GCs, two natural variants of ddl (named as ddl6 and ddl7) were detected in the library of first GCs of both the UK and Bangladeshi sample. In this chapter, it has been shown that expression of these gene variants in E. coli and B. subtilis confer resistance to D-cycloserine. The upstream sequence of the integron carrying ddl7 has also been determined. Based on the sequence homology of the upstream sequence of integron-located ddl, the source of the integron was predicted to be a strain of T. denticola. The activity of the putative Pc located within the attI-site of the integron has also been analysed experimentally and it was found that the Pc is active and can efficiently induce the expression of downstream GCs.

4.2 Materials and methods

4.2.1 Determination of minimum inhibitory concentration (MIC)

The MIC of D-cycloserine, vancomycin and β-lactams was determined using the agar dilution method discussed in Chapter 2.

4.2.2 Construction of ddl6 mutants by site-directed mutagenesis

A Phusion Site-Directed Mutagenesis Kit (Thermo Scientific, USA) was used for mutagenesis reactions. Two different mutants of ddl6 were constructed using
the protocol for point mutations provided with the kit and the mutants were
named according to the position of nucleotide changed by site-directed
mutagenesis. The purified plasmid pGEM-T Easy::intl-attl-ddl6 was used as
template. The pGEM-T Easy::intl-attl-ddl6 c.490 C>T plasmid was constructed
by using the forward mutagenic primer Ddl6-490F (containing C>T substitution)
with the reverse primer Ddl6-490R (Figure 4-4A). To construct the pGEM-T
Easy::intl-attl-ddl6 c.777 G>T mutant the forward mutagenic primer Ddl6-777F
(with a G>T substitution) was coupled with Ddl6-777R (Figure 4-4B). All of the
primers were phosphorylated at the 5’-end by the producer (Sigma-Aldrich, UK).
The PCR programme was set according to the protocol provided with the kit: 30
seconds at 98°C followed by 25 cycles for 10 seconds at 98°C, 30 seconds at
58°C, and 2 minutes at 72°C. The final elongation was done for 10 minutes at
72°C. The PCR products were purified and ligated to form circular plasmids
using T4 DNA ligase. The plasmids were transformed into E. coli α-Select silver
efficiency cells (Bioline, UK) for MIC determination.
4.2.3 Cloning of *ddl6* and *ddl7* into pHCMC05 and pET-28a

The plasmids pGEM-T Easy::intl-attl-ddl6 and pGEM-T Easy::intl-attl-ddl7 were used to amplify only the coding sequences of *ddl6* and *ddl7*, respectively using the primers TddlF and TddlR. To reduce the chance of PCR error Q5 High-Fidelity 2X master mix was used. The amplified PCR products were cloned into *BamHI* and *XbaI* sites of pHCMC05 (Figure 4-5) and transformed into *B. subtilis* 168. For cloning into pET-28a, the coding sequences of *ddl6* and *ddl7* were amplified using TddlF and Tddl28aR primers and cloned into *BamHI* and *XhoI* sites of pET-28a(+) vector. The methods for ligation and transformation are discussed in detail in Chapter 2.
Figure 4-5. Cloning of ddl6 and ddl7 into pHCMC05 vector. The vector is digested with BamHI and XbaI and purified with PCR purification kit. The coding sequence of ddl6 and ddl7 are PCR amplified using TddlF and TddlR primers having BamHI and XbaI restriction sites. The PCR products are purified and digested with BamHI and XbaI and purified again. The digested PCR products were then ligated to the digested vector and transformed into E. coli α-select for manipulation of the plasmids. After confirming the correctness of the insert by sequencing, the plasmids are transformed to B. subtilis 168.

4.2.4 Determination of the upstream region of integron carrying ddl6 and ddl7

With an aim to clone to full-length of intI and its upstream region, four different forward primers (Figure 4-6) were designed randomly at different locations of the chromosome of T. denticola ATCC35405 (NC_002967.9). PCR was performed by combining these primers with the reverse primer, ddIIR designed from the 3'-end of ddls. It was assumed that if the integrons carrying ddI GCs
are located on the chromosome of *T. denticola*, the PCR products will yield a
different length of upstream region of *intI*.

**Figure 4-6.** A) Approximate location of the primers (small arrow) on the genome
of *T. denticola* ATCC35405 (NC_002967.9) (ORFs were annotated according
GenBank database). Four forward primers were designed upstream of the *intI*
of *T. denticola*. B) These forward primers were coupled with a reverse primer,
ddlR designed from the end of ddls located on the 2,024 bp PCR amplicon
recovered from the library.

For instance, Upint_5100F primer, designed approximately 2.7 kb upstream of
the start codon of *intI* of *T. denticola* ATCC 35405 and was combined with ddlR
to produce amplicon containing 2.7 kb sequence upstream of *intI* followed by
the full-length *intI* and *ddl*. Q5® High-Fidelity 2X Master Mix (NEB, UK) was
used for PCR using the pooled metagenomic DNA of UK saliva samples as
template. The expected size PCR products were purified by gel extraction using
QiAquick Gel Extraction Kit (Qiagen, Germany) and A-tailed following the
protocol for cloning blunt-ended PCR products described in pGEM®-T Easy
Vector Systems Technical Manual (Promega, UK). A 30 µL reaction mixture for
A-tailing contained about 30 ng purified PCR product, 3 µL of 10x standard *Taq*
reaction buffer, 1.5 µL 100 mM dATP, 2 µL *Taq* DNA polymerase (NEB, UK)
and molecular grade water up to 30 µL. The reaction mixture was then
incubated for 30 minutes at 72°C in a block of PCR machine. An aliquot of A-
tailed PCR products was cloned into pGEM-T Easy and then transformed into
E. coli alpha select cells. The transformants were selected on LB agar plates supplemented with ampicillin, IPTG and X-gal. Plasmids were prepared from 4 randomly selected white colonies and the inserts were sequenced completely by using M13F, M13R and newly designed bridge primers.

4.2.5 Determination of functionality of putative gene cassette promoter, Pc located upstream of ddl cassettes

To confirm that the putative Pc is located within the putative attl region and to test if the Pc can mediate the expression of downstream ddls, the coding sequences of ddl6 and ddl7 and the upstream sequences that contain the attl and Pc were PCR amplified using TdattIF and Tddl28aR primers (Figure 4-7). The amplified PCR products were cloned into the BamHI and XhoI sites of pET-28a(+) vector (Figure 4-8).

Figure 4-7. Primers used to clone ddl with their upstream attl sequence containing putative Pc promoter.
The inserts in the vectors were sequenced and checked for any errors and then transformed into *E. coli* BL21 (DE3) cells (New England Biolabs). The transformants were selected on LB plates supplemented with 30 µg/mL kanamycin. MIC of D-cycloserine against the *E. coli* BL21 (DE3) cells was determined using agar dilution method without adding IPTG into the media. An increase in the MIC of D-cycloserine indicates that the expression of *ddl* is driven by the *Pc*. The strain EC300 [*E. coli* BL21 (DE3) transformed with pET-28a(+)]) was used as vector control and EC306 and 307 which harbours plasmids pET-28a(+)::*ddl6* and pET-28a(+)::*ddl7*, respectively were used as negative control.

4.2.6 Growth curve analysis of *B. subtilis* overexpressing *ddl7*

To investigate if the overexpression of *ddl* s affect the growth rate of the host, the growth curve of BS707 (*B. subtilis* 168 transformed with pHCMC05/ddl7)
overexpressing \textit{ddl7} was determined and compared with wild-type \textit{B. subtilis} 168 and BS700 carrying empty pHCMC05. A single pure colony of BS700 and BS707 was inoculated into 10 ml of BHI broth supplemented with 10 µg/mL of chloramphenicol and incubated overnight at 37° C with shaking (220 rpm). \textit{B. subtilis} 168 was grown in an antibiotic free medium. On the following day, 500 µL of the overnight culture was inoculated into 50 mL fresh BHI broth supplemented with 10 µg/mL of chloramphenicol and 0.5 mM IPTG to induce expression of \textit{ddl7}. 1ml culture was removed every half an hour and OD$_{600}$ was measured. The average OD$_{600}$ of three individual biological replicates were plotted against time and change in the growth rate was compared.

4.2.7 PCR to detect \textit{ddl} in individual samples of oral metagenomic DNA

To detect if the ORFs encoding \textit{ddl6} and \textit{ddl7} are present in individual samples PCR was performed by coupling the primer TDIF that target the \textit{intI} with the reverse primer ddlR that bind the 3'-end end of the \textit{ddl6/ddl7}. So, the PCR amplicons that will be produced will contain a part if \textit{intI}, \textit{attI} and GC encoding \textit{ddl} on the first position or other location of the cassette array.

4.2.8 Isolation of the putative host of the integron carrying \textit{ddl6} and \textit{ddl7}

To isolate the original host of the integron carrying \textit{ddl}, 11 UK volunteers who provided saliva samples for metagenomic screening of integrons, were again contacted to provide 2 ml of saliva. This time the saliva samples (n=5) were pooled immediately after collection. To reduce oxygen exposure the pooled saliva was processed immediately. 500 µL of the pooled saliva were inoculated into the Oral Treponeme Enrichment Broth (OTEB) (Anaerobe System, USA).
Before inoculation, the OTEB was supplemented with 5 µg/mL rifampicin with or without 64 µg/mL D-cycloserine to select both D-cycloserine resistant and D-cycloserine sensitive treponemes. Oral *Treponema* species are naturally resistant to rifampicin and can be used as a selective agent for isolation of oral spirochaetes (Canale-Parola 1980). The inoculation of OTEB was carried out inside an anaerobic chamber. OTEB is an enriched medium for the isolation, cultivation and maintenance of oral treponemes. It is a complex peptone-yeast extract medium containing volatile fatty acids and serum for a source of long chain fatty acids. Volatile fatty acids and serum are necessary nutrients for recovery of treponemes from the oral flora. The tubes were grown for 4 days in anaerobic chamber at 37°C. After 4 days, 20 µL of the growth was diluted 50 times into pre-reduced BHI. 100 µL of the diluted suspension was spread onto pre-reduced fastidious anaerobic agar (FAA) plates supplemented with 10% Foetal Calf Serum (FCS, Gibco) and 5% sheep blood (E&O laboratories, UK) and the same antibiotics used in OTEB. The plates were grown anaerobically for two weeks to support the growth of D-cycloserine resistant oral treponemes. The colonies that appeared within 2-3 days (non-treponemes) were subcultured and grown in fastidious anaerobic broth supplemented with the antibiotics as mentioned above for genomic DNA extraction. Genomic DNA was extracted from the pellet obtained from 1 mL broth culture. The presence of integron carrying *ddl* was screened by PCR using *intI* and *ddl* specific primers.

**4.2.9 Tools and software used for bioinformatics analysis**

Multiple sequence alignment of nucleotide sequences was performed using the T-Coffee server (Notredame, Higgins et al. 2000). Alignment of protein sequences was carried out with ClustalW program on www.uniprot.org/align.
The promoters were detected by using two promoter prediction programmes: BPROM (http://www.softberry.com/berry.phtml) and Neural Network for Promoter Prediction (NNPP) (http://www.fruitfly.org/seq_tools/promoter.html). The phylogenetic tree was constructed using MEGA6 software (Tamura, Stecher et al. 2013). The homologous sequences *ddl6/ddl7* and *intI* were downloaded from NCBI database and aligned with ClustalW. The default settings for the Neighbour-Joining method were applied to construct the phylogenetic tree.

### 4.3 Results

#### 4.3.1 Genetic features of integron-located *ddl* genes were studied

In the libraries of PCR amplicons of integrons and associated GCs generated using TDIF and MARS2 primers, a total of five clones (three from UK and two from Bangladesh) were recovered containing an ORF predicted to encode a Ddl (Figure 4-9). The genetic organisation of the insert was as follows: a partial *intI* (768 bp) followed by a putative *attI* (70-bp, 867-936 bp), a putative *Pc* and a GC carrying *ddl* (1035 bp, 967-2001) (Figure 4-9). A putative RBS was identified at 6 bp upstream of the start codon of the gene: aggtggtgtaaaatg (RBS underlined and start codon of *ddl* in bold and in italics).
Figure 4-9. (A) Genetic organisation of 2,024 bp inserts in pGEM-T Easy vector containing a partial intI, Pc, and ddl within the first cassette. (B) Comparison of putative attI sequence preceding ddl in the pGEM-T Easy inserts (TMU6/7) with the putative attI of integron of T. denticola ATCC 35405 (Coleman, Tetu et al. 2004). Location of putative Pc with their consensus -35 and -10 sequences on attI has been shown. The integrase binding sites S1 and S2 as well as DR1 and DR2 are also shown. The recombination point G↓TT and the putative transcription start site (TSS) located at the 3ʹ-end of attI are also shown. (C) Comparison of partial sequence of attC detected at the 3ʹ-end of the 2,024 bp insert with a typical complete attC associated with Tde1837 of T. denticola integron (Coleman, Tetu et al. 2004).

Similar to the reverse integron of T. denticola, the intI and the ORF encoding ddl were oriented in the same direction (Coleman, Tetu et al. 2004). The attI site harbours two putative integrase binding sites [L (S2) and R (S1)], two direct repeat sequences (DR1 and DR2) and a putative attI/attC recombination site (G^TT at 926 position of the 2,024-bp fragment) (Figure 4-9B). The stop codon of the ORF encoding ddl was located on the 7 bp attC core site, R" (GGCTAAC or RRYYAAC; stop codon of ddl underlined). The R" site is followed by the L" (CTTCAACC) which covers the priming site of the reverse primer MARS2.

By analysing the sequence of ddls detected in this study, two variants of ddl were identified and they were named as ddl6 and ddl7. The variants differ from
each other by two nucleotides. At c.490 and c.777 position of the coding sequence, the *ddl6* has cytosine (C) and guanine (G), respectively, whereas the corresponding nucleotides of *ddl7* contain thymine (T) (Figure 4-10). These nonsynonymous substitutions alter the codon **C**TT (Leu164) and **T**GG (Trp259) of *ddl6* to **T**TT (Phe164) and **T**GT (Cys259) of *ddl7*, respectively (the substitutions are bold and underlined). To confirm that these variants are real and not the result of error during amplification by PCR, the PCR to amplify the amplicons containing *ddl* was repeated using Q5 High-Fidelity polymerase. The PCR products were sequenced and analysed. Both variants were recovered again having 100% sequence identity with the initial sequences recovered from the metagenomic libraries. The sequences were submitted to GenBank with the accession number KU886208 and KU886209 for the inserts carrying *ddl6* and *ddl7*, respectively.

Pairwise sequence alignment of the whole 2,024 bp containing *ddl6* with that of *ddl7* shows that the partial *intI*S of the two sequences are different from each other by 6 nucleotides, however, the putative *attI* harbouring the *Pc* are 100% identical to each other.
Figure 4-10. The sequence of ddl6 (in bold font) with its upstream and downstream sequence within the 2,024 bp insert. The positions of SNPs (marked with red) by which ddl6 differed from ddl7 have been shown underneath. The non-coding sequence in between partial intI (underlined red at the 5'-end) and ddl6 is shown in lower-case letter. This region contains the putative attI site (underlined green), RBS, IntI-binding sites (S1 and S2) and the -10 and -35 sequences of the putative Pc. The recombination point is shown by a vertical arrow. The R'' and L'' sequences within the partial attC at the downstream of the gene have also been shown (grey shaded). The location of priming sites of TDIF and MARS2 primers is shown by arrows.
4.3.2 The second-copy of *ddl* in *T. pedis* B683 is the closest homologue of *ddl*6 and *ddl*7

In the NCBI nr/nt database, no homologous sequence of *ddl*s was found, however, when the NCBI WGS database was searched by limiting with *Treponema* (taxid: 157), a single homologue of the gene was found to be located on a 5699 bp contig of *T. pedis* B 683 (NZ_AOTN01000179) (Figure 4-11 and Figure 4-12). This homologue *ddl* of *T. pedis* B683 was 98% identical with the integron-located *ddl*6 and *ddl*7. As shown in Figure 4-11 and Figure 4-12, within that WGS contig of *T. pedis* B683, there are six ORFs including the ORFs that encode a multidrug transporter, hypothetical proteins, a HDIG-domain containing protein and the *ddl* homologous to *ddl*6 and *ddl*7. This *ddl* of *T. pedis* is named as *ddl*(GC) in the next sections of this thesis. The alignment of *ddl*6, *ddl*7 and *ddl*(GC) is shown in Figure 4-14.

![Diagram](image-url)

*Figure 4-11. The genetic organisation of *ddl* of *T. pedis* found on a wgs contig (GenBank accession number: NZ_AOTN01000179) which shows 98% nucleotide sequence identity to the integron-located *ddl*s. This homologous *ddl* was found to be flanked by two *attC* sites.*
Figure 4-12. The location of the ddl homologous to ddl6 and ddl7 (bolded and red underlined) within the 5699 bp contig of T. pedis B663 (GenBank accession number: NZ_AOTN01000179). The noncoding sequences are marked with small case letters. The R'', L'', L' and R' core-sites of the putative attC sites are bolded and marked with blue, green, purple and red coloured font, respectively. The different ORFs located upstream and downstream of ddl are underlined with different colour (orange-multidrug...
A total of three different putative attC sites were identified within the contig carrying _ddl(GC)_ (Figure 4-11 and Figure 4-12). The three attCs were variable in terms of size and sequence (Figure 4-13) which is a feature GC array of MI (Rowe-Magnus, Guerout et al. 2001). However, as there was no _intI_ within that contig, the association of this contig with an integron, although likely, could not be confirmed.

**Figure 4-13.** Sequence and size of the putative attC sites detected within the contig of _T. pedis_ B683 that contains a _ddl_. The attC sequences of the contig were compared with attC of _T. denticola_ (Tde1837) and partial attC located downstream of _ddl6_ or _ddl7_.

The B683 strain of _T. pedis_ was found to contain another copy of _ddl_ (GenBank accession: AOTN01000124: 8689-9912bp), presumably the house-keeping _ddl_, as it was found to be present in other strains of _T. pedis_ including str. T A4 and T M1, whereas the _ddl(GC)_ was found to be present only in B683. Pairwise sequence alignment of _ddl(GC)_ and the house-keeping _ddl_ of B683 using Emboss Water (Rice, Longden et al. 2000) showed that they are only 43.7% identical.

At the amino acid level, Ddl6 and Ddl7 (344 aa) are 97% identical to the Ddl(GC) (WP_029410393) of _T. pedis_. The next closest homologue was Ddl of
Syntrophobotulus glycolicus (ADY55260) with 55% amino acid identity, followed by Clostridium sp. (54%; WP_033164556), Lachnoclostridium phytofermentans (53%, WP_029502590) and Paenibacillus pini (52%; WP_036650467). Ddl6 or Ddl7 shared a low homology with the Ddl of E. coli [(28%; BAE76162; DdlA), (36%; EGT67290; DdlB)], and M. tuberculosis (25%; WP_003912011). Ddl of T. denticola (AAS12903) and B. subtilis (CAB12263) were 29% identical to Ddl6 and Ddl7. Ddl proteins with altered specificity that confer resistance to vancomycin resistance such as VanA (AAA65956), VanB (YP_009076352) and VanC (P29753) also shared low sequence identity (27 to 30%) to Ddl6 and Ddl7.
Figure 4-14. Multiple alignment of the coding sequences of ddl6, ddl7 and ddl(GC) of T. pedis. The alignment was done by using T-Coffee server (Notredame, Higgins et al. 2000). The position of two mutations at c.490 and c.777 by whichddl6 is different from ddl7 is shown by the blue arrows.
4.3.3 The *ddl*(GC) of *T. pedis* B683, the only homologue of *ddl*6 and *ddl*7, is likely to be acquired by HGT

The GC content of the *T. pedis* genome was calculated as 36.9% which was very similar to the GC content of house-keeping *ddl* of *T. pedis* (38%). However, the GC content of *ddl*(GC) was only 30.6%, approximately 6% lower than the overall GC content of the *T. pedis* B683 genome. This variability of GC content of the overall genome and the *ddl*(GC) suggests that the *ddl*(GC) may have been acquired by HGT by *T. pedis* B683.

Next, to predict if there has been a likely genetic exchange occurred between *T. pedis* B683, the host of *ddl*(GC) and the host of integron-located *ddl*6/*ddl*7, the upstream sequence of *ddl*6/*ddl*7 up to the recombination point G↓TT located on putative S1 of *attI* and the downstream partial *attC* were aligned with the upstream and downstream sequences of *ddl*(GC) of *T. pedis* B683. It was found that the 40 bp upstream of *ddl*6/*ddl*7 is 100% identical to the upstream sequence of *ddl*(GC) (Figure 4-15). The partial 29 bp *attC* of *ddl*6 and *ddl*7 (*attC*ddl6/ddl7) located downstream of the genes was 90% identical to the *attC*ddl(GC), however, the core sites R” and L” were 100% identical to each other. The identity of the flanking sequence including the recombination points indicates a possible HGT of the genes among their hosts.
Figure 4-15. The identity of the flanking sequence of ddl6/ddl7 (40 bp upstream and 29bp downstream) with their closest homologue, ddl(GC) located on a 5699 bp contig of T. pedis B683 genome. The putative core sites of attC (R'' and L''), the simple integrase binding site (S1) are shown. The recombination points located on S1 and core site, R' of the attC located upstream of ddl(GC) (attCHDIG) are marked with a down arrow. The primer binding sites are shown by small horizontal arrows. The start codons (ATG) of the homologous ddls are bolded. The attC sites downstream of the genes are the attC for this gene.

4.3.4 Integron located ddls are phylogenetically diverged from house-keeping ddls of Treponema

A phylogenetic tree was constructed to observe the relationship of the ddls of different Treponema spp. with the ddl6, ddl7 and ddl(GC). This shows that the integron-located ddl6 and ddl7 and ddl(GC) of T. pedis clustered together and was distantly related to house-keeping ddls of different species Treponema (Figure 4-16). The results of phylogenetic analysis, the nucleotide sequence identity of among ddl6/ddl7 and ddl(GC) as well as the flanking sequence of the genes, supports the hypothesis that the GCs carrying ddls have been transferred horizontally among their host strains.
Figure 4-16. Phylogenetic tree of ddl6, ddl7 and ddl(GC) along with the ddls of different strains of T. denticola and some other Treponema species. The ddIA of E. coli was used as an out group. The evolutionary history was inferred using the Neighbor-Joining method. Evolutionary analyses were conducted in MEGA6 (Tamura, Stecher et al. 2013). The ddls found within GCs are shown in rectangular box.

Another phylogenetic tree was constructed based on the deduced amino acid sequences of Ddl6 and Ddl7 (Appendix-II and III) and their evolutionary relationships were compared with Ddl of many different species including the Van proteins that confer resistance to vancomycin. Similar to the gene tree (Figure 4-16), this shows that the Ddl6 and Ddl7 form a separate cluster along with the DdlGC of T. pedis (WP_029410393) (Figure 4-17). The tree showed
that the house-keeping Ddls of *T. pedis* and *T. denticola* are distantly related to Ddl6 and Ddl7.

Figure 4-17. The phylogenetic tree of deduced amino acid sequences of Ddl6 and Ddl7 (highlighted in red rectangle) along with the Ddl homologues of other species. The evolutionary history was inferred using the Neighbour-Joining method. Evolutionary analyses were conducted in MEGA6 (Tamura, Stecher et al. 2013).

4.3.5 The putative active sites of Ddl6 and Ddl7 are conserved

Alignment of the amino acid sequences of Ddl6 and Ddl7 with that of Ddl of *T. pedis*, DdlA and DdlB of *E. coli*, VanA of *E. faecium* and VanB of *E. faecalis* shows that the known active sites of the proteins that bind with ATP and D-alanine are conserved in all Ddls (Figure 4-18).
The active sites and position of the omega loop were detected and marked based on previous reports (Shi and Walsh 1995, Noda, Kawahara et al. 2004). The putative 3D-structures of Ddl6 and Ddl7 were determined by homology modelling are discussed in chapter 6.

Figure 4-18. Multiple sequence alignment of Ddl and Ddl7 with the Ddl(GC) of T. pedis (Ddl_Tped) (WP_029410393), DdlA (BAE7676162) and DdlB (EGT67290) of E. coli, Van A (AAA65956) of E. faecium as well as VanB (YP_009076352) of E. faecalis. The amino acid residues known to bind D-Ala and ATP have been marked on their top. The conserved residues are marked with an asterisk and grey shaded. The position of the omega loop has been indicated by a box. The changes between Ddl6 and Ddl7 are
numbered on top and marked with red rectangles. The active sites and position of omega loop were detected and marked based on previous reports (Shi and Walsh 1995, Noda, Kawahara et al. 2004).

4.3.6 Expression of *ddl6* and *ddl7* located within the 2,024 bp insert in pGEM-T Easy confers a different level of resistance to D-cycloserine

To determine if the *ddls* located within the 2,024 bp inserts in pGEM-T Easy are expressed and their expression can confer resistance to D-cycloserine, the MIC of D-cycloserine was determined against the *E. coli* strains EC126 and EC127 carrying pGEM-T Easy::*intI-attI-ddl6* and pGEM-T Easy::*intI-attI-ddl7* plasmids, respectively. It was observed that the constitutive expression of *ddl6* in *E. coli* increased the MIC of D-cycloserine. The MIC of D-cycloserine against *E. coli* expressing *ddl6* and *ddl7* was 16 and 64 µg/mL, respectively, whereas the MIC against the vector control was 8 µg/mL (Figure 4-19). As the upstream region of *ddl6* and *ddl7* that contains putative *attI* and *Pc* are 100% identical, it was hypothesized that the difference in MIC in between the two variants was due to the SNPs at c.490 and c.777 (Figure 4-10).
The above MIC data were obtained using D-cycloserine produced by Sigma-Aldrich (Lot no. 115M411V, Produced in Israel). However, when the D-cycloserine manufactured by the same company but produced in Taiwan (Lot no. BCBK1338V) was used, the MIC was found to be changed by four-fold. Using the lot from Taiwan, the MIC against EC121, EC126 and EC127 increased proportionately and was determined as 32, 64 and 256 µg/mL, respectively. This variation of MIC of D-cycloserine due to the lot change prompted the checking of the physical form of the powders. The D-cycloserine sourced from Taiwan was a fine powder, but the powder sourced from Israel was crystalline (Figure 4-20). Although both lots are labelled with a purity of 96.0%, presumably they were not and the Israeli product was more pure, which causes this variation of MIC.
Figure 4.20. The country sources, lot numbers and physical forms of D-cyclosorine manufactured by Sigma Aldrich used in this study. The labelled purity for both product was 96.0% (The photos were taken on 15 June 2016).

4.3.7 Site-directed mutagenesis confirmed that c.777 G>T substitution at ddl6 is responsible for the alteration of MIC of D-cycloserine

To determine if the SNPs among the two integron-located ddl are responsible for the alteration of D-cycloserine susceptibility in E. coli, site-directed mutagenesis was used. Mutagenic primers were designed to change the two nucleotides of ddl6 that made it different from ddl7. In one construct, using the pGEM-T Easy::intl-attI-ddl6 as template, the cytosine (C) at 490 was substituted with thymine (T) (pGEM-T Easy::ddl6-c.490C>T) and in another construct using the same template, the guanine (G) at 777 was substituted with thymine (T) (pGEM-T Easy::intl-attI-ddl6-c.777G>T) (Figure 4-4). The newly constructed plasmids were then transformed into E. coli. The MICs of D-cycloserine was measured against EC206 and EC207 carrying pGEM-T Easy::ddl6-c.490C>T
and pGEM-T Easy::intI-attl-ddl6-c.777G>T, plasmids, respectively. The EC126 and EC127 harbouring plasmids with wild-type ddl6 and ddl7, respectively were used as positive control.

It was found that the MIC of D-cycloserine for the strain EC206 carrying C>T substitution at c.490 of ddl6 and the strain EC126 carrying wild-type ddl6 was 16 µg/mL (Figure 4-21). Thus, the C>T substitution which alters Leu164 to Phe164 of Ddl6 does not have any effect on the alteration of the D-cycloserine resistance phenotype. However, the MIC of D-cycloserine against the strain EC207 carrying the G>T substitution at c.777 of ddl6 was found to increase by four-fold (64 µg/mL) compared to EC126 and EC206 (Figure 4-21). This suggests that the G>T substitution which alters the Trp259 to Cys259 of Ddl6 is responsible for changing susceptibility of D-cycloserine in E. coli.

**Figure 4-21.** Identification of the SNP of the integron-located ddls responsible for altering the D-cycloserine resistance phenotype of E. coli. EC121 was the vector control carrying empty pGEM-T Easy.
4.3.8 Expression of *ddls* does not confer cross-resistance to other antibiotics targeting cell-wall biosynthesis

To determine if the overexpression of *ddl* can also confer resistance to D-cycloserine, vancomycin and β-lactam antibiotics in gram-positive bacteria, the 1032 bp coding sequence of *ddl6* and *ddl7* was cloned into a medium-copy (10-20 copy) expression vector, pHCMC05 (an *E. coli*-B. subtilis shuttle vector) and transformed into *B. subtilis* 168 (Figure 4-5). The pHCMC05 vector carries an IPTG-inducible Pspac promoter and the genes were cloned downstream of this. The MICs of the antibiotics were determined in the presence of 0.5 mM IPTG added into the media to induce the expression of the genes. The MIC of D-cycloserine against the strains BS706 and BS707 which carry pHCMC05::*ddl6* and pHCMC05::*ddl7* vectors, respectively was found to increase by eight-fold (64 µg/mL) compared to the BS700 strain which carries the empty plasmid (8 µg/mL) (Figure 4-22). No difference in the MIC of D-cycloserine could be spotted in between BS706 and BS707 carrying *ddl6* and *ddl7*, respectively, when the expression of the genes was induced by IPTG.

The MIC of D-cycloserine against the strains EC306 and EC307 carrying pET-28a::*ddl6* and pET-28a::*ddl7*, respectively also exhibited a similar phenotype. Like BS706 and BS707, the MIC of D-cycloserine against EC306 and EC307 was determined to be 64 µg/mL and no difference between MIC could be observed in between the strains overexpressing *ddl6* and *ddl7* under IPTG induction (Figure 4-22).

The MIC against the EC300 carrying empty pET-28a(+) (negative control) was 8 µg/mL.
Figure 4-22. The MIC of D-cycloserine against the strains of B. subtilis 168. Expression of the inserted genes was induced by 0.5 mM IPTG.

As *E. coli* is intrinsically resistant to vancomycin, therefore the MIC of vancomycin was determined against the *B. subtilis* strains only. A vancomycin resistant *E. fecalis* strain was used as a positive control. It was found that the overexpression of *ddl6* and *ddl7* does not confer vancomycin resistance. The MIC of vancomycin against all strains of *B. subtilis* was found to be <0.125 µg/mL, whereas the MIC of vancomycin against positive control *E. faecalis* was >128 µg/mL.

The MIC of penicillins including penicillin G, amoxicillin, oxacillin and ampicillin against both *E. coli* and *B. subtilis* strains expressing the *ddl* was also found to be unaffected. This suggests that overproduction of Ddl does not confer cross-resistance to the antibiotics targeting cell-wall biosynthesis used in this study.
4.3.9 The putative Pc promoter is located within the attI site and can express downstream GCs

As discussed previously, the putative Pc of the integron carrying ddl is embedded within the attI site. The putative Pc promoter exhibited high similarity with the E. coli σ70 promoter as well as the P2 promoter of class 1 integrons (Figure 4-23).

\[
\begin{align*}
Pc \text{ promoter (this study)} & \quad \text{TTGCA} \quad (17 \text{ bp}) \quad \text{TATAGT} \\
P2 \text{ promoter (class 1 integrons)} & \quad \text{TTGT} \quad (17 \text{ bp}) \quad \text{TACAGT} \\
E. \text{ coli } \sigma^{70} \text{ promoter} & \quad \text{TTGACA} \quad (16-18 \text{ bp}) \quad \text{TATAAT}
\end{align*}
\]

**Figure 4-23.** Alignment of Pc promoter detected upstream of ddl6 and ddl7. The Pc promoter was found to exhibit very good similarity with the P2 promoter of class 1 integron (Collis and Hall 1995) and E. coli σ70 promoter (Ozoline, Deev et al. 1997). The identical nucleotides are highlighted with grey shading.

To confirm that the Pc of the integron carrying ddl is located within the attI site and can express the downstream genes, two constructs were generated in the pET-28a(+) vector by excluding the partial intI from the 5'-end of the 2,024 bp insert (please see methods). The functionality of Pc preceded by ddl on the first gene cassette was determined by its ability to drive the expression of ddl6 and ddl7 without IPTG induction and the expression of ddl5 driven by the Pc was assessed by determining the change in the susceptibility of the hosts to D-cycloserine. To observe any leaky expression by the T7 promoter, the strains of E. coli harbouring only the coding sequences of ddl6 and ddl7 cloned into pET-28a(+) were used as negative controls. E. coli BL21 (DE3) containing empty pET-28a was used as a vector control.
It was observed that without IPTG induction, $Pc$ could express the downstream $ddl6$ and $ddl7$ and as a result, the MIC of D-cycloserine against $E. coli$ strains increased. The MIC for EC310 and EC311 carrying the plasmid pET-28a(+)::attI-$ddl6$ and pET-28a(+)::attI-$ddl7$, respectively was found to be 32 and 64 µg/mL, whereas the MIC for EC306 and EC307 carrying the pET-28a(+)::$ddl6$ and pET-28a(+)::$ddl7$ plasmids, respectively was 16 µg/mL (Figure 4-24). As this expression occurred without IPTG induction, it is assumed that $E. coli$ RNA polymerase binds to the $Pc$ in the clones containing the attI sites and initiates transcription of $ddl6$ and $ddl7$ even in the absence of IPTG.

These observations supply three important pieces of information: 1) the putative $Pc$ is functional and can drive the expression of the downstream gene cassettes; 2) unlike the class 1 and class 3 integrons, the $Pc$ of unusual or reverse integron carrying $ddl6$ and $ddl7$ is located within the attI region, and 3) the differences in MIC of D-cycloserine for $ddl6$ and $ddl7$ are due to the SNP at c.777 position in between the two variants, not related with the difference in the associated intI1s upstream of the genes.
Figure 4-24. Expression of ddl6 and ddl7 under the control of Pc promoter without IPTG induction. No change in D-cycloserine susceptibility was observed in the clones carrying only the coding sequence of ddls cloned into pET-28a. Resistance to D-cycloserine increases in the clones carrying the putative Pc region. The Pc of both ddl variants were 100% identical.

4.3.10 The upstream region of integron carrying ddl7 was found to be similar to T. denticola

The 2,024 bp pGEM-T Easy inserts contained a part of intI, full-length attI, and a ddl within the first GC. As the partial intI and attI associated with ddl GC exhibited >99% nucleotide sequence identity with the integron of T. denticola ATCC35405, it was assumed that T. denticola is the likely host of this reverse integron. To confirm this hypothesis, the upstream sequence of the integron was amplified by PCR using forward primers designed based on the genome sequence of T. denticola ATCC 35405. The four forward primers (Upint_5100F, Upint_3685F, Upint_721F and Upint_122F) were designed around 2.7, 4.2, 7.2
and 7.8 kb upstream of the intI (Section 4.2.4). The forward primers were coupled with a reverse primer (ddlR) that targeted the 3'-end of ddl6/ddl7 so that the primer pair would selectively amplify a region containing ddi6. Agarose gel-electrophoresis of the PCR products of four different sets of reactions showed that the primer combination Upint_122F and ddlR yielded a single ~10 kb band (Figure 4-25 A). The other primer combination Upint_5100F and ddlR which is expected to produce a ~5 kb amplicon, also produced a band approximately 5 kb in size, however, some other smaller size products were also formed. The PCR products of two other primer combinations were found to appear as a smear of bands on the gel (Figure 4-25 A). The ~10 kb and ~5 kb bands from the first and 4th set of reactions were extracted from agarose gel and the purified products were A-tailed and cloned into pGEM-T Easy with T-overhangs. Perhaps due to the large size product, the cloning of ~10kb products failed and no positive colonies were found in two separate experiments. However, the cloning of ~5kb products was successful. The digestion of purified pGEM-T Easy with EcoRI showed that all of the four clones were positive for the ~5kb insert (Figure 4-25 B).

The inserts of four clones were completely sequenced and the actual size of the sequenced insert was found to be 4,421 bp and they were found to be 100% identical to each other. So, this PCR amplification has added another 2,423 bp sequence upstream of the 2,024 bp amplicon containing ddl obtained during initial screening. Now, the new 4,421 bp sequence contains a partial ORF encoding a hypothetical protein (1-1,362 bp), a complete ORF encoding IntI (1,925 - 3,191 bp, 421 aa) (Appendix-IV), a putative attI, the Pc promoter followed by a cassette carrying ddl7 (3,389 - 4,421 bp) (Figure 4-26).
The partial sequence of the ORF (1,362 bp) encoding a hypothetical protein was found to be a maximum of 97% identical to the *T. denticola* US-Trep acdtB (GenBank accession number: AGEB01000014: 742-6053). The *intI* (1,266 bp) showed a maximum of 96% identity with the *intI* of *T. denticola* H-22 (GenBank accession no: AGDV01000005:19792-21072). The *intI* of *T. denticola* ATCC 35405 exhibits 95% nucleotide sequence identity with the *intI* upstream of *ddl7*. At the amino acid level, the identity ranges between 81-97% with other *Treponema* IntI. Maximum identity was found with the IntI of *T. denticola* AL-2 (97%, EMB42956).

The Figure 4-26 shows that the 4,421 bp amplicon is different from the closely related hits in the NCBI databases in terms of the genetic organisation and size and number of the putative ORFs. The *intI* located upstream of *ddl7* was 1,266 bp (422 aa), which is 25 bp less compared to the *intI* of H-22 and ATCC35405 strains of *T. denticola* and 72 bp larger than the *intI* of US-Trep acdtB. The
intergenic 556 bp sequence flanked by the *intI* and the partial ORF encoding a hypothetical protein is non-protein-coding. However, the similar region in closely related strains has 2 to 3 ORFs (Figure 4-26).

**Figure 4-26.** The genetic arrangement of 4421 bp pGEM-T Easy insert carrying ddl7, full length *intI* and another partial ORF encoding a hypothetical protein. The primers used to amplify this product are shown with small arrows. The genetic arrangements of the equivalent site of the closest homologues including *T. denticola* H-22 (GenBank accession: AGDV01000005:c23656-19631), *T. denticola* ATCC35405 (GenBank accession: AE017226:1873126-1877351) and *T. denticola* US-Trep acdtB (GenBank accession: AGEB01000014.1:3405-5167) have also been shown. The most variable region among the strains was found to be the sequence in between the ORF for hypothetical protein and *intI*. The size of this region is also shown.
Figure 4-27. The sequence of the 4421 bp insert of pGEM-T Easy carrying upstream sequence of ddl7. The partial ORF encoding the intI is bolded and underlined with purple colour and the coding sequence for ddl7 at the 3'-end of the insert was underlined with red colour. The putative RBS located upstream of the genes was marked with green font. The -10 and -35 sequences of the putative promoter for integrase, Pint and the Pc promoter is in red font and underlined. The putative integrase binding sites are in blue font. The recombination point on attI is shown with a red vertical arrow.

The non-protein-coding 556 bp sequence located in between the partial ORF and the intI harbours the putative promoter for the intI (Pint). The sequence of the putative Pint was found be conserved with a sequence of 5'-ATGAAT [19 bp]
TAAACT-3' (Figure 4-27). The upstream sequence of the \textit{intI} was analysed \textit{in silico} for potential LexA binding sites similar to that found in the upstream of \textit{intI} of mobile and CIs (Guerin, Cambray et al. 2009). However, no LexA binding motif could be detected. Additionally, this region contains several inverted repeat and direct repeat sequences as well as a putative transcription terminator sequence.

4.3.11 The \textit{intI} of the integron carrying \textit{ddl7} is closely associated with \textit{intI} of different strains of \textit{T. denticola}

A phylogenetic tree was constructed to compare the evolutionary relationship of \textit{intIs} of the integron carrying \textit{ddl} with the \textit{intIs} of different species of \textit{Treponema}. The tree in the Figure 4-28 showed that the \textit{intI} associated with \textit{ddl7} is closely related with other \textit{intIs} of \textit{T. denticola}. The results obtained from the \textit{intI} tree as well as the identity of the upstream sequence \textit{intI} with the other strain of \textit{T. denticola} suggest that the likely host of the integron associated with \textit{ddl7} is a species of the genus \textit{Treponema} and most likely a strain of \textit{T. denticola}. 
Figure 4-28. The phylogenetic tree of homologues of intl found upstream of ddl7 in the 4421-bp inserts in pGEM-T Easy. The evolutionary relationship was inferred using Neighbour-Joining Method (Saitou and Nei 1987). Evolutionary analyses were conducted in MEGA6 (Tamura, Stecher et al. 2013). intl1 of E. coli was used as outgroup.

4.3.12 Growth curve of *B. subtilis* 168 overexpressing *ddl6* or *ddl7*

To determine if the overexpression of *ddl* affects the growth of the surrogate host, the growth of *B. subtilis* 168 overexpressing *ddl7* was compared with the vector controls and host control. It was observed that overexpression of *ddl7* under IPTG induction does not affect the growth rate of the host *B. subtilis* 168. This suggests that overexpression of *ddl* does not affect the overall fitness of the host (Figure 4-29).
Figure 4.29. The growth curve of B. subtilis 168 overexpressing ddl6 and ddl7 under IPTG induction. The growth rate was compared with the vector control and host controls.

4.3.13 The GCs encoding *ddl* are present in almost all individual samples of the UK and BD

As the presence of *ddl* cassettes was detected in the separate pooled samples of both countries, the individual samples which represent *ddl* were unknown. To investigate how many individual samples are positive for *ddl*, the metagenomic DNA from saliva samples were isolated separately from 10 Bangladeshi and 8 UK samples. PCR was performed using TDIF (targets the *intI*) and ddlR (targets both *ddl* variants).
Figure 4-30. The agarose gel electrophoresis of the PCR products obtained using the TDIF and ddlR primer to detect the presence of ddl in individual saliva metagenomic DNA samples from Bangladesh (A) and the UK (B).

It was found that 80.0% of Bangladeshi (8 out of 10) and 100% of the UK samples were positive for ddl GCs (Figure 4-30). The two ddl negative samples from Bangladesh were further examined to see if the ddl is present in another position of the integron GC array, but they could not be detected.

To compare if the other genes located on the first GC such as cof-like hydrolase which was found in the clones TMB3/5/6/10/11/13/14/16 are also widespread like ddl, the primer TDIF was coupled with a reverse primer clhR that targets the 3'-end of cof-like hydrolase gene. It was found that unlike ddl GC, the GC encoding cof-like hydrolase was not common and was represented by one Bangladeshi (sample 4) and three UK individuals (sample 1, 2 and 8) (Figure 4-31).
4.3.14 The original host of the integron carrying *ddl* could not be recovered from the oral cavity

Experiments were designed to isolate the original host of *ddl* by culture-based method based on three assumptions: i) the host is a culturable species of the genus *Treponema*; most likely a strain of *T. denticola* ii) the ORF of GC encoding Ddl is maintained in the first position of the integron and expressed constitutively and iii) expression of *ddl* confers resistance to D-cycloserine to the host, thus can be selected by supplementing the isolation medium with D-cycloserine.

More than 50 D-cycloserine resistant isolates from saliva were screened for the presence of *ddls* and the associated integron, however, none of them were positive. The 16S rDNA of the isolates were found to >99.0% identical to the *Eggerthia catenaformis* DSM20559 (previously *Lactobacillus catenaformis*), a commensal of human oral cavity. One isolate (MAR1) was selected for whole genome shot-gun sequencing. The draft genome of the strain *E. catenaformis* Strain MAR1 was annotated and the the sequence was submitted to GenBank (accession no. NCVR00000000). A Genome Announcements paper is published based on this draft genome which is attached (Appendix-VII).
4.4 Discussion

This chapter covers the in-depth analysis of the *ddl*s detected within the first GC of a reverse integron. Expression of these genes increase D-cycloserine resistance of the surrogate host. Using site-directed mutagenesis the SNP (c.777G→T) between the two variants of the gene that alters the D-cycloserine resistance phenotype in *E. coli* were identified. The genes associated with integrons were present in the saliva of nearly all healthy volunteers of the UK and Bangladesh.

The closest homologue of *ddl6* and *ddl7* with 98% nucleotide identity was found to be located on a WGS contig of a *T. pedis* B683. This second copy *ddl* in the genome of *T. pedis* B683 is strain specific and was not present in the genome of other *T. pedis* strains, thus this gene is a part of the accessory genome of *T. pedis*. The strain B683 was isolated from a shoulder ulcer of pigs (NCBI Reference Sequence: NZ_AOTN00000000.1) (Svartstrom, Mushtaq et al. 2013).

*T. pedis* is an animal pathogen and is associated with digital dermatitis (DD) in cattle and necrotic ulcers in pigs and has been isolated from gingiva of pigs (Svartstrom, Mushtaq et al. 2013). This suggests that this treponeme is capable of colonizing both the oral cavity and skin lesions. Additionally, *T. denticola*, the likely host of the integron carrying *ddl6/ddl7*, is only found in the human oral cavity and is associated with human periodontitis (Simonson, Goodman et al. 1988). *T. denticola* is one of the best-characterised species among the 49 species of *Treponema* found in the human oral cavity (Dashper, Seers et al. 2011). It is also a member of the ‘red complex’ associated with clinical
progression of periodontitis (Holt and Ebersole 2005). Although *T. pedis* and *T. denticola* colonise a different host, they share extensive genomic similarities. The genomes of *T. pedis* T A4 and *T. denticola* ATCC 35405 were found to be very similar in terms of genome size, number of genes and G+C content as well as homology of the genes (Svartstrom, Mushtaq et al. 2013). It was observed that each of the strains of two species carried a unique set of virulence related genes (Svartstrom, Mushtaq et al. 2013).

The flanking sequence of *ddl*6/*ddl*7 (40 bp upstream and 29 bp downstream) was found to be nearly identical with flanking sequence the homologous *ddl*(GC) of *T. pedis* B683 suggesting genetic exchange between *T. pedis* and the host of *ddl*6/*ddl*7, perhaps a species of *Treponema*. The exchange could also involve an intermediate host. However, how and when this predicted genetic transfer occurred between microorganisms of two different ecological niche demands more investigation. The identification of the *ddl*6/*ddl*7 on an integron among most of the individuals of two cohorts is also interesting. This suggests that the host carrying the genes evolved to colonise the human oral cavity in ancient times and a common selection pressure may exist which maintains the GC encoding *ddl* at the first position of the integron GC array.

Now, the next questions are: i) why does the host of *ddl*6/*ddl*7 need another copy of *ddl* and ii) what is the selective pressure to maintain these genes on an integron? The presence of two copies of *ddl* has been reported only in a few bacterial genomes such as *E. coli*, *Salmonella enterica* serover Typhimurium (Zawadzke, Bugg et al. 1991) and *Enterococcus gallinarum* BM4174 (Ambur, Reynolds et al. 2002). However, in these strains, none of the two genes were found on a mobile GC. Zawadzke, Bugg et al. (1991) proposed two plausible
explanations for the duplication of *ddl* in *E. coli*: a) perhaps one of the two genes (*ddlA*) is expressed constitutively whereas the expression of the other gene (*ddlB*) is controlled tightly and coupled with the expression of other cell-wall biosynthesis genes; b) the D-ala-D-ala dipeptide may be used for another pathway other than for peptidoglycan synthesis. Assuming that the *T. denticola* is the host of the integron carrying *ddl6/ddl7*, the genetic environment of the housekeeping *ddl* was analysed. It was found that the gene was located within a cluster of biosynthetic genes, so, the expression of this gene might be tightly controlled and perhaps the host needed another copy of the gene which was expressed constitutively. The possibility that the acquired *ddls* within integron might have another function other than peptidoglycan metabolism will be considered in Chapter 5 and 6.

The most important finding in this chapter is the observation that the expression of *ddl6* and *ddl7* cloned into pGEM-T Easy confers a different level of resistance to D-cycloserine. As Ddl is the target of D-cycloserine, this mechanism of resistance is very likely due to overproduction of the target (Figure 4-3). As the Pc promoters of both *ddl6* and *ddl7* were 100% identical, it was assumed that the two nonsynonymous substitutions between *ddl6* and *ddl7* cause the MIC of D-cycloserine to be altered by four-fold. Site-directed mutagenesis confirmed that the G↔T SNP at c.777 position which changes the Trp259 of Ddl6 to Cys259 is solely responsible for this change. The mechanism of how the W259C mutation in Ddl6 alters MIC of D-cycloserine is not clear. In the next chapter, the mechanism of the alteration of D-cycloserine resistance due to this mutation will be further discussed. It is the first report of the alteration of MIC of D-cycloserine due to natural mutation in *ddl*.  

209
Another important result of this chapter is the confirmation that putative $P_c$ is located within the intergenic region that harbours $attl$. The $P_c$ promoter of class 1 and class 3 integrons are located within the coding sequence of the $intl$. Some class 1 integrons that carry a weak $P_c$ promoter, a second promoter $P_2$ is also found located within the $attl$ region (Collis and Hall 1995, Jové, Da Re et al. 2010). However, unlike these class 1 integrons, a single $P_c$ promoter within the $attl$ region upstream of $ddl6/ddl7$ could be detected. The 200 bp sequence located upstream of the putative $P_c$ (-35 TTGCAA |17 bp| -10 TATAGT) was examined for the presence of another promoter, however, no other predicted promoters could be found. This $P_c$ is likely to bind RNA polymerase of $E. coli$ as it conserves 9 of the 12 nucleotides of hexamer of $E. coli \sigma^{70}$ promoter (-35 TTGACA |16-18 bp|-10 TATAAT) (Ozoline, Deev et al. 1997).

It was also determined that the putative $P_c$ can mediate the expression $ddl6$ and $ddl7$ without IPTG induction. In the absence of IPTG, the T7 promoter of the vector pET-28a remains inactive due to the binding of the lac repressor encoded by lacI of the vector to the lac operator. So, the leaky expression of the inserted gene is tightly controlled. If the genes cloned into the pET vector do not have their own promoter and associated RBS, expression is unlikely to happen in the absence of IPTG. It was observed that without IPTG, the genes in the insert associated with upstream $attl$ where $P_c$ is embedded are expressed and their expression led to increase the MIC of D-cycloserine. It is therefore likely that the $E. coli$ RNA polymerase binds to the $P_c$ of the integron in the absence of IPTG induction.

Another interesting observation of this study is that when the $ddl6$ and $ddl7$ are expressed without IPTG-induction, the alteration of MIC of D-cycloserine due to
the nonsynonymous SNP at c.777 could be detected. However, when the genes are overexpressed by IPTG induction, they confer a similar level of resistance to both *E. coli* and *B. subtilis*. Perhaps, due to the very large amount of protein in the expression host produced by IPTG induction, the difference in MIC among the strain carrying *ddl* variants could not be observed. However, when no IPTG is added into the medium, the expression of *ddl6/ddl7* in the vector is probably mediated by RNA polymerase encoded by the house-keeping gene of the host, not by the T7 RNA polymerase unless there is a leaky expression. So, in this condition, the expression is likely to be mediated by *Pc* located upstream of the *ddls* and this simulates the natural expression of the genes and thus, the amount of protein in the cell remains low. Due to low the concentration of Ddl in the host, perhaps the change in MIC of D-cycloserine among the hosts expressing *ddl6* and *ddl7* can be detected. In the next chapter, kinetic analysis of Ddl6 and Ddl7 will be considered to investigate if there is any difference in catalytic activity among the Ddls which could explain the change in D-cycloserine resistance phenotype of their surrogate host.

### 4.5 Conclusion

This study reports for the first time that *ddls* that were detected within the first GC of a *T. denticola*-like reverse integron can confer variable resistance to D-cycloserine, an important antibiotic for TB management. The integron-mediated resistance to D-cycloserine is a concern as there is a possibility of dissemination of the cassette-located *ddl* genes by HGT to other bacteria in the oral cavity and perhaps to the intestinal microbes when they pass through the oral cavity. Moreover, as Ddl is a popular target in antibiotic drug discovery research, the presence of these mobile *ddls* should be considered as a possible
threat to newly developed Ddl inhibitors. It was also found for the first time that a nonsynonymous mutation of *ddl* at the positions other than the known active sites can alter the D-cycloserine resistance phenotype. The putative *Pc* located upstream of the first GC was determined to be functional and can mediate the expression of the downstream GCs.
Chapter Five: Functional characterization of DdIs encoded by integron gene cassettes
**5.1 Introduction**

Neuhaus (1960) showed for the first time that Ddl is the enzyme that catalyses the synthesis of D-ala-D-ala dipeptide in *Streptococcus faecalis* (currently *Enterococcus faecalis*). It was shown that the formation D-ala-D-ala dipeptide requires ATP, D-alanine, a divalent metal cation (Mn$^{+2}$ or Mg$^{+2}$) and an enzyme, named D-alanyl-D-alanine synthetase. Later, the purification, properties and kinetic studies of *S. faecalis* Ddl were reported in a series of two articles (Neuhaus 1962a, Neuhaus 1962b). It was shown that in addition to a divalent metal cation, the enzyme needs a K$^+$ for efficient catalytic activity. Subsequently, the Ddls encoded by the housekeeping genes from many other different species of bacteria including *S. typhimurium* (Daub, Zawadzke et al. 1988), *E. coli* (Zawadzke, Bugg et al. 1991), *Leuconostoc mesenteroides* (Park and Walsh 1997), *Streptomyces lavendulae* (a D-cycloserine producing strain) (Noda, Kawahara et al. 2004), *Helicobacter pylori* (Wu, Zhang et al. 2008), and *M. tuberculosis* (Prosser and de Carvalho 2013) as well as the Ddl homologues responsible for vancomycin resistance such as VanA (Bugg, Dutka-Malen et al. 1991), VanB (Meziane-Cherif, Badet-Denisot et al. 1994), VanC (Healy, Park et al. 1998) and VanG (Meziane-Cherif, Saul et al. 2012) were also studied extensively.

The kinetic parameters of Ddl proteins including $V_{\text{max}}$, $K_m$ and $K_{\text{cat}}$ are usually estimated by measuring continuous release of ADP in the reaction by using PK/LDH-coupled spectrophotometric system (Neuhaus 1962b, Zawadzke, Bugg et al. 1991, Feng and Barletta 2003). $K_m$ is defined as the substrate concentration at which the reaction rate is half its maximal value ($V_{\text{max}}$) and $K_{\text{cat}}$ is the number of substrate molecules converted into product by an enzyme molecule in a unit time when the
enzyme is fully saturated with substrate. The $K_{cat}$ is also known as turnover number (Berg, Tymoczko et al. 2002).

$$V_{max} \quad \text{Reaction velocity (V_o)}$$

$$V_{max}/2 \quad [S] \quad \text{Substrate concentration}$$

**Figure 5-1.** A plot of the reaction velocity ($V_o$) as a function of the substrate concentration $[S]$ for an enzyme that obeys Michaelis-Menten kinetics. The Michaelis-Menten constant ($K_m$) is the substrate concentration yielding a velocity of $V_{max}/2$. The figure was adapted from Berg, Tymoczko et al. (2002).

For the Ddl enzymes, two $K_m$ values can be determined for two D-alanine as they bind to separate binding sites within the active sites of the enzyme in N- and C-terminal (Zawadzke, Bugg et al. 1991, Wu, Zhang et al. 2008, Prosser and de Carvalho 2013). The first D-alanine (N-terminal) binds 20-400 fold higher affinity compared to the second D-alanine (C-terminal). For instance, the $K_m, D$-ala1 and $K_m, D$-ala2 for DdlAEc was found to be 5.7 µM and 0.55 mM, respectively (Zawadzke, Bugg et al. 1991). Similarly, $K_m, D$-ala1 (0.075 mM) of Ddl of *M. tuberculosis* was 48 times lower than the $K_m, D$-ala2 (3.6 mM) (Prosser and de Carvalho 2013). As values of $K_m, D$-ala1 are very low, its measurement is often erroneous (Prosser and de Carvalho 2013) and the value for $K_m, D$-ala1 is independent of the C-terminal substrate (D-ala2). Most often only $K_m, D$-ala2 is determined at a saturating concentration of ATP and a relatively high concentration of D-alanine (Noda, Kawahara et al. 2004).
In the previous chapter, the detection of two variants of ddl (ddl6 and ddl7) in a library of PCR amplicons of first GCs and their D-cycloserine resistance phenotype were discussed. In this chapter, the overexpression of these two gene variants and purification of the 6x-his-tagged proteins by fast-protein liquid chromatography (FPLC) will be described. The purified proteins have been used to confirm that they are Ddl and can catalyse D-ala-D-ala dipeptide formation. The substrate specificity has also been checked to confirm their inability to confer vancomycin resistance. The enzyme kinetic parameters of both enzymes have been determined and the kinetic properties have been compared with an aim to explain the alteration of MIC of D-cycloserine between these two variants. The DdIA of *E. coli* (DdIAEc) has also been purified to use as a positive control for kinetic assays. The house-keeping *ddl* of *T. denticola* (ddlTd) has been cloned, expressed and purified and the kinetic parameters of DdITd have been compared with integron-encoded Ddl6 or Ddl7.

### 5.2 Materials and Methods

#### 5.2.1 Chemicals and reagents

The EnzChek Phosphate assay kit and Pierce BCA Protein assay kits were purchased from Life Technologies (USA). The phosphokinase-lactate dehydrogenase (PK-LDH) enzyme used for kinetic analysis was purchased from Sigma, UK. Mono-potassium phosphoenol-pyruvate (PEP, monopotassium salt) (Roche, Germany), NADH disodium salt Grade II (Roche, Germany) and ATP disodium salt (Roche, Germany) were supplied by Sigma, UK. MgCl₂, KCl and D-alanine were purchased from Sigma, UK.
5.2.2 Preparation of the reagents and buffers

5.2.2.1 Reagents for EnzCheck phosphate assay kit

The kit contains 4 components: MESG (2-amino-6-mercapto-7-methylpurine riboside) substrate (Component A), Purine nucleoside phosphorylase (PNP) (Component B), 20x reaction buffer (Component C) and phosphate standard (Component D). 1 mM stock solution of MESG was prepared by adding 20 mL of dH₂O directly to the bottle containing the MESG substrate (Component A). The bottle was shaken vigorously to mix the substrate. The solution was then aliquoted in 200 µL in Eppendorf tubes and the tubes were stored immediately at -20°C. An aliquot of MESG substrate was thawed immediately before use by placing the tubes in a 37°C water bath until just melted (no more than 5 minutes). The tubes were vortexed vigorously and placed on ice until added to the reaction. A 100 U/mL stock solution of purine nucleoside phosphorylase (Component B) was prepared by adding 500 µL of dH₂O to each vial and the vial was stored at 4°C. A 500 µM working solution of the phosphate standard was prepared by diluting a portion to 100-fold with dH₂O.

5.2.2.2 Reagents for kinetic analysis

0.1M working solutions of KCl, MgCl₂, ATP, PEP and D-alanine were prepared in water and used fresh or stored at -20°C. 15 mM solution of NADH disodium salt was prepared in sodium bicarbonate buffer and stored at 4-8°C. The PK-LDH enzymes were supplied in solution and were used directly.

5.2.3 Optimization of time for maximum expression of Ddl

A single fresh colony of EC306 (pET-28a(+)::ddl6), EC307 (pET-28a(+)::ddl7), EC308 (pET-28a(+)::ddlTd) and EC309 (pET-28a(+)::ddlAEc) was inoculated separately in 5 ml LB broth supplemented with 30 µg/mL of kanamycin and
incubated overnight at 37°C with shaking. On the following day, 500 µL of the overnight culture were inoculated into 50 ml fresh LB broth and grown until the OD$_{600}$ became 0.5 to 0.7. At this stage 0.2 mM IPTG was added into the media and incubated again. 1 ml of sample was withdrawn in an Eppendorf tube at 1, 2, 3 and 20 hours. The cell pellets were boiled in 1x protein buffer for 5 minutes and SDS-PAGE was carried out in 1x running buffer to check the time for optimum expression. Optimum level of expression was found in 3 hours after IPTG induction.

5.2.4 Western blotting

5.2.4.1 Preparation of samples for western blotting

The EC306 (pET-28a(+)::ddl6) and EC307 (pET-28a(+)::ddl7) strains were grown to their log-phase and 0.2 mM IPTG was added to induce the expression of the genes. In the negative control of expression, no IPTG was added. 1 mL of the sample was withdrawn after 3 hrs and the pellet was boiled for 5 minutes in 1x protein sample buffer. 10 µL of the sample was loaded into 10% polyacrylamide gel.

5.2.4.2 Electrophoresis of protein using SDS-PAGE

SDS-PAGE was carried out using 1 x running buffer. Proteins were separated using a Mini-PROTEAN® Electrophoresis cell (Biorad) using a 10% separating gel, with a 4% stacking gel (Chapter 2) and ECL™ Rainbow™ Full Range Marker, RPN800E (GE Healthcare). Gels were run at 200 V for approximately 60-70 minutes.

5.2.4.3 Electrophoretic transfer

Proteins were transferred to a PVDF membrane (GE Healthcare) using the Mini Trans-Blot® Electrophoretic Transfer cell (Biorad) in 1 x transfer buffer. The electrophoretic transfer was performed at 15 V for 20 minutes. Transfer was checked by Ponceau S stain (Sigma) and the dye was washed off with water after detection.
5.2.4.4 Detection of proteins

After the proteins were transferred onto the PVDF membrane, the membrane was blocked with membrane blocking agent (GE Healthcare) dissolved in TBS supplemented with 0.05% Tween-20. The blot was then incubated overnight with an anti-His-tag antibody (Biolegend, USA) (stock 0.5 mg/ml) with a working dilution of 0.2 µg/mL in blocking solution.

After overnight incubation, the blot was washed 3 x 10 minutes in TBS-T to remove unbound primary antibody. The membrane was incubated for 1 hour with the secondary antibody, a goat-anti-mouse-IgG HRP-conjugate (Southern Biotech) which was used at a dilution of 1:2000 in blocking buffer. The membrane was washed for 3 x 10 minutes in TBS-T to remove secondary antibody. The membrane was then exposed to Lumigen ECL Plus (PS-3) detection reagent (Lumigen, USA) wrapped in Saran Wrap and was exposed to autoradiography film (GE Healthcare) between 30 secs and 5 minutes.

5.2.5 Optimization of purification

The batch purification method was used to determine the solubility of Ddl proteins and to determine if the recombinant proteins bind with the nickel affinity column. The cells were grown (50 ml) to an exponential phase in a 37°C shaking incubator and expression was induced by addition of 0.2 mM IPTG. Cells were harvested after 3 hours of incubation and cell pellets were stored in -20°C. The cells were lysed with B-PER bacterial protein extraction reagent (4 ml per gram of cell pellet) (Thermo Scientific, USA) supplemented with lysozyme (to 1.0 mg/ml), DNase (to 1 U/ µL) and cOmplete EDTA-free protease inhibitor cocktail (Roche, 1 tablet/10 ml). After incubation at room temperature for 15 minutes, the suspension was centrifuged at
15,000 x g for 30 minutes at 4° C. The clear supernatant containing the soluble fraction of recombinant proteins was transferred to a clean tube. The cell debris which contain the insoluble fraction of the protein were treated with 1 ml denaturing buffer (50 mM NaH$_2$PO$_4$, 300 mM NaCl, 6M guanidine hydrochloride and 10 mM imidazole, pH 7.4) and mixed by pipetting and centrifuged at 15,000 x g for 10 minutes at 4° C. The supernatant containing the denatured recombinant proteins was collected and kept on ice.

The HIS-Select Nickel Affinity Gel (Sigma, UK) was used for batch purification following the manufacturer instructions. All centrifugations were done at 5,000 x g. Briefly, 50 µL of gel suspension was taken in an Eppendorf tube and centrifuged for 5 minutes the supernatant was discarded. The gel was washed with 500 µL (10 gel volume) of buffer A (50 mM NaH$_2$PO$_4$, 300 mM NaCl, 10 mM imidazole, pH 7.4) and centrifuged for 5 minutes and the supernatant was discarded. 500 µL of supernatant from native and denaturing conditions were added separately and mixed in an orbital shaker at 175 rpm for 15 mins. The mixture was centrifuged for 5 minutes and the supernatant was retained in a clean Eppendorf tube for SDS-PAGE analysis. This step was repeated twice with same volume of buffer and the supernatants were stored. At the last step 100 µL of elution buffer (native/denaturing) was added to the gel and mixed in the orbital shaker for 10 minutes. The supernatant was collected after centrifugation and this step was repeated again. All of the wash fractions and elutes from both native and denaturing conditions were analysed on SDS-PAGE. As a negative control, cell lysate of E.coli BL21::pET/ddl7 without induction was treated in the same manner and the washes and elutes were analysed in SDS-PAGE.
5.2.6 Preparation and cell lysates and purification of Ddl6, Ddl7, DdlEc and DddlTd using FPLC system

5.2.6.1 Preparation of cleared E. coli lysates under native conditions

A single pure colony of EC306, EC307, EC308 and EC309 was inoculated separately into 20 ml LB broth supplemented 30 µg/mL kanamycin and grown overnight at 37°C. 1L LB was inoculated with 10 ml overnight culture, and expression was induced in exponential growth phase with 0.2mM IPTG for 3h at 37°C. Cells were harvested by centrifugation at 15,000 x g at 4°C and stored at -20°C until protein purification. Preserved cells were thawed on ice for 30 minutes and resuspended in buffer A (50 mM NaH₂PO₄, 300 mM NaCl, 10 mM imidazole, pH 7.4) containing complete EDTA-free protein inhibitor cocktail (Roche, 1 tablet/10 ml), lysozyme (to 1.0 mg/ml) and DNAse (to 1 U/ µL). The suspension was incubated on ice for 30 min and sonicated (6 pulse, each 10 second). Cellular debris was removed by centrifugation at 15,000 x g for 30 minutes at 4°C and the clear supernatant was transferred to a new tube and kept on ice before loading into Fast Protein Liquid Chromatography (FPLC) system (BioRad). An aliquot of the cell lysate was saved for SDS-PAGE analysis.

5.2.7 FPLC to purify 6x His-tagged proteins

FPLC is a widely used technique for protein purification which was first developed in Sweden by Pharmacia Lab in 1982 (Madadlou, O’Sullivan et al. 2011). One of the major advantages of FPLC is that the whole process can be monitored on the computer screen and the chromatogram can be recorded. In this study, the Ni-NTA Superflow cartridge (1ml, Qiagen) compatible with FPLC system was used to bind and release nickel-bound hexahistidine-tagged Ddl proteins upon loading elution buffer with 250 mM imidazole into the FPLC system.
For purification of expressed DdlS, the FPLC system was filled with buffer A and the Ni-NTA Superflow cartridge (1ml, Qiagen) was attached to the system and the cartridge was equilibrated for 10 minutes with a flow rate of 1 ml/min. The supernatant was loaded into the pump and when the loading was done, the system was filled again with buffer A and washed until the $A_{280\, \text{nm}}$ reached to the baseline. The wash fraction was collected for SDS-PAGE analysis. Finally, the proteins were eluted with buffer B (50 mM NaH$_2$PO$_4$, 300 mM NaCl, 250 mM imidazole, pH 7.4). The elutes and washes were immediately kept on ice.

Both of the buffer A and buffer B were filter sterilized through 0.22 µm filter before running into the system.

5.2.8 Buffer exchange to remove imidazole from the protein solutions

The Thermo Scientific Slide-A-Lyzer Dialysis Cassette (size: 3-12 ml; membrane molecular cut-off:10 kDa) was used to exchange the buffer and remove any contaminants. The cassettes are constructed from two sheets of low-binding, regenerated-cellulose dialysis membrane that are hermetically sealed on either side of a silicone-like gasket inside an inert plastic frame. The dialysis buffer used was same as buffer A (see above) excluding the imidazole. The pH of the dialysis buffer was adjusted according to the pI of the proteins. The pI of Ddl6 and Ddl7 is 5.55, so the pH of the buffer was adjusted to 7.0 (the recommended pH is 1.5 higher than the pI). The membrane was hydrated by immersing the cassette into the buffer for two minutes. The protein samples kept on ice in Eppendorf tubes were taken into the syringe provided with the kit and the sample was injected slowly through one of the syringe ports of the cassette. The cassette was then slipped into the groove of a buoy and floated in the dialysis solution and swirled at 200 rpm overnight at 4°C. On the next day, the buffer was changed again and dialysis was continued for another
10 hours. After the dialysis is finished, the sample was transferred from the cassette using a fresh syringe.

5.2.9 Determination of protein concentration

The concentration of dialysed protein samples was determined using Pierce BCA Protein Assay Kit (ThermoFisher Scientific) following the manufacturer’s protocol for microplates. As reference standard, the Pierce Bovine Serum Album (BSA) standard (2 mg/ml) was used. Briefly, certain amount of working reagent was prepared by mixing 50 parts of BCA reagent A with 1 part of BCA reagent B. BSA standards were prepared by double dilution method so that the concentration range from 2 mg/ml to 1.95 µg/ml. 25 µL of each standard and test sample were pipetted into 96 well plate and 200 µL of the BCA working reagent was added to each well and mixed. The plate was covered and incubated for 30 minutes at 37°C. The plate was then cooled to bring the temperature to RT and the absorbance was measured at 590 nm using a plate reader.

5.2.10 Enzymatic activity assay

The D-ala-D-ala ligase activity of purified Ddl enzymes was assayed by the methods described previously with some modifications (Daub, Zawadzke et al. 1988, Bruning, Murillo et al. 2011). The enzyme activity was checked by determining the formation of D-ala-D-ala dipeptide into the reaction as well as by measuring the release of inorganic phosphate. The methods are described below:

5.2.10.1 Detection of formation of D-ala-D-ala dipeptide by paper chromatography

To detect the formation of D-ala-D-ala dipeptide in the reactions catalysed by Ddl7, the standard reaction mixture containing the following components were incubated
for 4 hours at 37° C: 50 mM Tris-HCl, 10 mM MgCl₂, 10 mM KCl, 20 mM D-alanine, 2 mM ATP, 2.5 mM Glutathione and 20 µg/mL of purified Ddl. To determine if the enzyme can catalyse the formation of D-ala-D-lac depsipeptide or D-ala-D-ala dipeptide, 10 mM of each was added to the above reaction and the amount of D-alanine was reduced to half (10 mM). To stop the reaction, the mixtures were boiled for 5 mins at 95°C. Using a micropipette, 4 µl of the reactions was loaded onto Whatman cellulose chromatography paper (grade 1) and ascending chromatography was developed for 3 hours in butanol-acetic acid-water (12:3:5) supplemented with 0.5% ninhydrin. The paper was dried for 5 minutes at 95°C. D-ala-D-ala dipeptide formation in the reaction catalysed by Ddl was detected by comparison with commercially available D-ala-D-ala (Sigma). The Rf values of pure amino acids and dipeptides were determined using the following equation:

\[
R_f = \frac{\text{Distance travelled by the compound}}{\text{Distance travelled by the solvent}}
\]

5.2.10.2 Measurement of inorganic phosphate, Pi in the Ddl-catalysed reaction

The enzymatic activity of Ddl was also assayed by measuring the release of Pi in the reaction (same reaction D-ala-D-ala dipeptide formation assay) with EnzChek phosphate assay kit (Life Technologies). 10 µl of the reaction product was added to 790 µl of standard reaction mixture containing 230 µl dH₂O, 50 µl 20x reaction buffer, 200 µl 2-amino-6-mercpto-7-methylpurine riboside (MESG) substrate solution and 10 µl purine nucleoside phosphorylase. The reaction was performed by incubation at 22° C for 30 minutes. After incubation absorbance of the reactions was measured at 360 nm with a spectrophotometer. The amount of inorganic phosphate released into the reactions was calculated from the standard curve of inorganic phosphate.
5.2.11 Inhibition of Ddl activities by D-cycloserine

To observe the inhibitory activity of D-cycloserine on Ddl, different concentrations of the drug (5 mM to 80 mM) were added into the reaction catalysed by the enzymes and incubated for 4 hours at 37° C. The release of Pi was measured according to the method described above.

5.2.12 Standard curve for Pi

A standard curve of Pi was generated using the phosphate standard supplied with the EnzCheck Phosphate Assay kit. Seven concentrations of inorganic phosphate starting from 0 to 150 µM were used. The different concentrations of phosphate standard were added to the standard reaction mixture (please see the composition under section 5.2.10.1) by adjusting the amount of water. The reaction mixtures were incubated for 30 mins at 22° C and absorbance was measured at 360 nm. The absorbance value of no-phosphate control was subtracted from the absorbance of each sample and a standard curve was plotted in MS-Excel 2013.

5.2.13 Measurement of $K_m$, $V_{\text{max}}$ and $K_{\text{cat}}$ of Ddl6 and Ddl7, DdlAEc and DdlTd

The $V_{\text{max}}$ and $K_m$, D-ala2 of Ddl6, Ddl7, DdlTd and DdlAEc were determined by using pyruvate kinase (PK)/lactate dehydrogenase (LDH)-coupled spectrophotometric system (Neuhaus 1962, Daub, Zawadzke et al. 1988). The ADP-release coupled assay estimates ADP formation with the coupled assay using phosphoenolpyruvate-ADP transphosphorylase and lactate dehydrogenase according to the following reaction (Neuhaus 1962):
ADP + phosphoenolpyruvate → ATP + pyruvate  \quad (\text{Reaction 1})

Pyruvate + NADH → Lactate + NAD$^+$  \quad (\text{Reaction 2})

The rate of oxidation of NADH is equivalent to formation of D-ala-D-ala dipeptide in µM. The oxidation rate of NADH is measured by rate of decrease in absorbance at 340 nm. This method is still used widely to study the kinetic parameters of Ddl (Meziane-Cherif, Saul et al. 2012, Prosser and de Carvalho 2013, Prosser and de Carvalho 2013).

All assays were conducted at 37° C in 96-well plate system using CLARIOstar® (BMG Labtech, Germany) in 200µL reaction buffer. The assay mixture (200µL) contained 100 mM HEPES (pH 7.5), 10 mM MgCl$_2$, 10 mM KCl, 0.2 mM NADH, 6-10 U/mL PK, 9-14 U/mL LDH, 2 mM phosphoenolpyruvate (PEP) and variable concentrations of ATP and D-alanine (saturating concentrations were 500 µM and 100 mM for ATP and D-alanine, respectively). The reactions were initiated by adding Ddl enzymes at a final concentration of 0.0025 µg/µL. The initial velocity ($v_0$) of the reactions were determined by measuring the changes in the absorbance at 340 nm for 500 secs (8.3 mins). The slope of the reaction progression curves (time (s) vs change in absorbance at 340 nm) is equivalent to the $v_0$ ($v_0 = \text{slope} = \Delta Y/\Delta X$)(Brooks, Geeganage et al. 2004). The slope was calculated by the MARS data analysis software (BMG Labtech, Germany).

After the $v_0$ values for each substrate concentration (ATP and D-alanine) were obtained, GraphPad Prism version 7 (GraphPad Software, Inc., USA) was used to determine the $V_{\text{max}}$ and $K_m$ values by fitting the curves for Michaelis-Menten model. \textbf{$V_{\text{max}}$} represents the maximum reaction velocity achieved by the system, at maximum saturation of the substrate concentration and \textbf{$K_m$} is the substrate concentration at
which the reaction rate is half of $V_{\text{max}}$. The following steps were followed to determine $K_m$ and $V_{\text{max}}$ values in GraphPad prism:

a. An XY data table was created.

b. The concentrations of substrate, $[X]$ (mM) were entered into X column and the enzyme velocities, $v_0$ (mM/s) determined by MARS software were entered into Y column.

c. Then the Analyze icon was clicked which was followed by selection of non-linear regression and the Michaelis-Menten enzyme kinetics from the panel of enzyme kinetic equation.

The Ddl enzyme binds two D-alanine molecules in two separate binding sites and thus two $K_m$ values ($K_m, \text{D-ala}_1$ and $K_m, \text{D-ala}_2$) can be determined (Zawadzke, Bugg et al. 1991). The general scheme for steady-state kinetics of Ddl-catalysed reaction is as follows, where $K_{m_1}$ and $K_{m_2}$ indicates the $K_m, \text{D-ala}_1$ and $K_m, \text{D-ala}_2$, respectively:

$$
E+S \rightleftharpoons ES+S \rightleftharpoons ESS \rightleftharpoons E+P \quad (1)
$$

By application steady-state kinetics model, a rate equation for Ddl-catalysed reaction can be obtained by using the following equation:

$$
V = \frac{V_{\text{max}} [S]^2}{K_{m_1}K_{m_2} + K_{m_2}[S] + [S]^2} \quad (2)
$$

Kinetic studies of Ddls have shown that $K_{m_1}$ value of the above equation is very small compared to $K_{m_2}$ (Zawadzke, Bugg et al. 1991, Noda, Kawahara et al. 2004). At high concentration of D-alanine, the N-terminal site becomes saturated and under this condition the rate equation (2) can be simplified to equation (3)(Park and Walsh 1997):
The rate equation thus reduces to the general equation for the Michaelis-Menten model and the kinetic parameters such as \( V_{\text{max}} \), \( K_{\text{m,D-ala2}} \), \( K_{\text{cat}} \) can be calculated using the standard steady-state velocity reactions (Prosser and de Carvalho 2013).

The values for \( K_{\text{m,ATP}} \) of different Ddls were determined at a fixed saturating concentration of D-alanine (100 mM) and varying concentrations of ATP starting from 5 \( \mu \)M to 320 \( \mu \)M. The \( K_{\text{m}} \) values for second D-alanine binding (\( K_{\text{m,D-ala2}} \)) was determined by using a fixed saturating concentration of ATP (500 \( \mu \)M; approximately 10 times of \( K_{\text{m,ATP}} \)).

The values of \( K_{\text{cat}} \) (\( s^{-1} \)) for all enzymes were determined using the equation (4). The values of \( V_{\text{max}} \) (mM/s) obtained in GraphPad were changed to \( \mu \)M/s before calculation of \( K_{\text{cat}} \).

\[
K_{\text{cat}} = \frac{V_{\text{max}}}{[E]} \tag{4}
\]

where the values of \( V_{\text{max}} \) and concentration of enzymes, \([E]\) were expressed in \( \mu \)M/s and \( \mu \)M, respectively.

To determine the catalytic efficiency, \( K_{\text{cat}}/K_{\text{m,D-ala2}} \) (M\(^{-1}\)s\(^{-1}\)), the unit of \( K_{\text{m,D-ala2}} \) was changed to Mole (M) and the \( K_{\text{cat}} \) was divided by \( K_{\text{m,D-ala2}} \).

5.2.14 Statistical Analysis

Statistical analysis was performed by using the GraphPad Prism (GraphPad Software, Inc., USA; version 7.0). Statistical significance was calculated using ordinary one-way ANOVA. \( P \) values of \( \leq 0.05 \) were considered significant.
5.3 Results

5.3.1 Optimization of time for expression and the purification process

Although the biological activity of Ddl6 was previously studied using raw cell-lysate (Feng and Barletta 2003), this study aimed to purify Ddl6 and Ddl7 to confirm their ligase activity without interference of the endogenous Ddl and other protein contaminants from the host. The coding sequences of ddl6 and ddl7 were cloned into a pET-28a expression vector containing an IPTG-inducible T7 promoter. The time for optimal expression under after IPTG induction was determined as 3 hours (Figure 5-2).

![Figure 5-2. SDS-PAGE analysis of the cell lysates collected at different time points for optimization of time for expression of ddl6 (A) and ddl7 in E. coli BL21 (DE3) after induction with 0.2 mM IPTG. The red arrows indicate the region where the 38 kDa recombinant Ddl6 and Ddl7 appear on the gel.](image)

Western blotting was performed to see if the variations in the two amino acids between Ddl6 and Ddl7 alter the expression of proteins. A similar level expression of Ddl6 and Ddl7 indicated that the amino acid substitutions between these two variants have no effect on the expression of the proteins (Figure 5-3).
Figure 5-3. Western blot analysis of EC306 (pET-28a::ddl6) and EC307 (pET-28a::ddl7) expressing Ddl6 and Ddl7, respectively. The 6x-his-tagged recombinant proteins were detected using anti-His-Tag antibody. Expression was induced using 0.2 mM IPTG.

The results of optimization of purification under both native and denaturing conditions showed that the proteins were soluble and could be purified in their native forms (Figure 5-4).

**Figure 5-4.** SDS-PAGE of different fractions of proteins obtained during optimization of purification. A) Fractions obtained during purification under native conditions from the cell-lysate of IPTG-induced EC307 (pET-28a::ddl7), B) Fractions obtained during purification under native conditions from the cell-lysate of EC307 without IPTG induction. Where, CL = raw cell lysates; S = supernatants after centrifugation; W1 and W2 = washed fractions with buffer A (native/denatured); E1 and E2 = two elutes eluted with buffer B; A = the agarose boiled with 1x sample buffer. His-Select Nickel Affinity Gel was used to bind the 6x-tagged recombinant Ddl7. The red horizontal arrow indicates the expected location of the expressed Ddl.

5.3.2 Purification of Ddl6, Ddl7, DdlAEc and DdlTd using FPLC

Based on the results of optimization of purification, the 6x-his-tagged Ddl6 and Ddl7 were purified using FPLC using the elution buffer containing 250 mM imidazole at pH 7.4. The chromatogram of FPLC for purification of Ddl6 is shown in Figure 5-5. SDS-PAGE analysis shows that the purity of the proteins were very good (>95%).
molecular mass of Ddl6 and Ddl7 was found to be approximately 38 kDa which is similar to the calculated size (Figure 5-6 A).

Purification of the DdlAEc and DdlTd was also performed following the similar protocol for Ddl6 and Ddl7. SDS-PAGE analysis showed that both DdlAEc and DdlTd could also be purified near to purity (>95%) (Figure 5-6).

![Chromatogram of purification of Ddl6 using FPLC system fitted with Ni-NTA superflow cartridge (1mL). The different parts of the chromatogram were labelled.](image)

**Figure 5-5.** Chromatogram of purification of Ddl6 using FPLC system fitted with Ni-NTA superflow cartridge (1mL). The different parts of the chromatogram were labelled.
Figure 5-6. SDS-PAGE of purified Ddl6, Ddl7, DdlAEc and DdlTd. 80 µL of the purified protein was mixed with 20 µL of 5x Protein Loading Buffer and boiled for 5 minutes at 95°C. 10 µL of the denatured proteins was loaded.

5.3.3 The integron-encoded Ddl7 catalyse the formation of D-ala-D-ala dipeptide, not D-ala-D-ser or D-ala-D-lac

The pure proteins were used to check their ability to catalyse the formation of D-ala-D-ala dipeptide using a two-level approach: determination of the release of Pi and detection of dipeptide formation using paper chromatography.

A standard curve of inorganic phosphate was generated (Figure 5-7) to calculate the amount of phosphate released into the reaction. The Pi assay (Figure 5-8) showed that both of the isoforms release a significant amount of Pi into the reaction (P>0.05) which indicates that ATP is consumed in the reaction and Pi is formed as a byproduct. The ability of D-cycloserine to inhibit Ddl’s activity has also been confirmed by adding different concentration D-cycloserine into the reactions catalysed by Ddl7. A concentration dependent reduction of the release of inorganic phosphate was observed in the presence of D-cycloserine. 10 mM D-cycloserine
inhibited Ddl activity to almost half (Figure 5-8B). An inhibition plateau after 50 mM D-cycloserine was observed.

![Image](image_url)  
**Figure 5-7.** Standard curve for inorganic phosphate.

![Image](image_url)  
**Figure 5-8.** (A) Release of inorganic phosphate in the reactions catalysed by Ddl6 and Ddl7 in the presence of 20 mM D-alanine. The differences of the release of Pi in between the negative controls (no Ddl) and Ddl catalysed reactions were analysed by one-way ANOVA (***P<0.001; ****P<0.0001). (B) In-vitro Inhibition of the ligase activity of recombinant Ddl7 in the presence of D-cycloserine (10 to 80 mM). Significant differences of the release of Pi in between the negative control (no D-cycloserine) and D-cycloserine-inhibited reactions were
analysed by one-way ANOVA (****P<0.0001, *P<0.05). Data represent mean values of three independent experiments. Only the significant differences are shown by the asterisks.

The paper chromatography showed that Ddl7 can catalyze the ligation of D-ala resulting in the synthesis of D-ala-D-ala dipeptide. The substrate specificity assay indicated that Ddl7 could catalyse the synthesis of D-ala-D-ala dipeptide, not the D-ala-D-ser or D-ala-D-lac. However, the ability of Ddl7 to catalyse formation of peptide bonds between D-alanine and other amino acids such as phenylalanine, aminobutyrate, D-norvaline etc. has not been tested. The Rf values for D-alanine and D-serine were found to be 0.46 and 0.33, respectively, whereas the Rf value of D-ala-D-ala dipeptide was calculated as 0.60. The D-ala-D-ala produced in the reaction had identical Rf value of the commercial D-ala-D-ala.

![Figure 5-9](image)

**Figure 5-9.** Ascending paper chromatography to detect the formation of D-ala-D-ala and D-ala-D-ser dipeptide as well as D-ala-D-lac depsipeptide. The chromatogram was developed using the solvent system n-hexane: acetic acid: water (12:3:5) supplemented with 0.5% ninhydrin. In the first two reactions (lane 1 and lane 2) 20 mM D-alanine was used, whereas in other reactions it was reduced to 10 mM and supplemented with 10 mM D-serine or D-lactate.
5.3.4 Kinetic properties of Ddl6, Ddl7, DdlAEc and DdlTd

The kinetic parameters of Ddl6, Ddl7, DdlTd and DdlAEc were determined to answer the following questions:

i) Is there any difference between Ddl6 and Ddl7 in terms of $K_m$ values for ATP and D-alanine?

ii) Is Ddl7 more catalytically active than Ddl6 which could explain the alteration of D-cycloserine resistance phenotype due to Trp259$\rightarrow$Cys substitution?

iii) Are the integron-encoded Ddls more efficient compared to the Ddl of *T. denticola* (the likely host of the integron carrying *ddl*)?

5.3.4.1 The reactions catalysed by Ddls follow the Michaelis-Menten kinetics and no major difference in $K_m$,ATP and $K_m$,D-ala2 was observed between Ddl6 and Ddl7

To determine the $K_m$,ATP of the Ddls, the initial velocities of the reactions were measured using a saturating concentration of D-alanine (100 mM) and variable concentrations of ATP (5, 10, 20, 40, 80, 160 and 320 µM). The initial velocities (the slopes of the reaction progression curves) of the reactions vary with substrate concentration and the plot of velocity versus substrate concentration followed Michaelis-Menten kinetics (Figure 5-10). The values of $K_m$,ATP for Ddl6, Ddl7, DdlTd and DdlEc are shown in Table 5-1. Based on the values of $K_m$,ATP, the saturating concentration ATP for determination of $K_m$,D-ala2 was determined.
**Table 5-1.** Steady-state kinetic parameters of Ddl6, Ddl7, DdlTd and DdlEc when determined with a fixed saturating concentration of D-alanine (100 mM) and variable concentration of ATP

<table>
<thead>
<tr>
<th>Proteins</th>
<th>Vmax (mM s⁻¹)</th>
<th>Kₘₐₚ (µM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ddl6</td>
<td>0.0003875</td>
<td>54.96</td>
</tr>
<tr>
<td>Ddl7</td>
<td>0.0004074</td>
<td>43.28</td>
</tr>
<tr>
<td>DdlTd</td>
<td>0.0003139</td>
<td>19.08</td>
</tr>
<tr>
<td>DdlAEc</td>
<td>0.0005868</td>
<td>48.99</td>
</tr>
<tr>
<td>DdlAEc*</td>
<td>NR</td>
<td>38*</td>
</tr>
</tbody>
</table>

* The values of DdlAEc reported by Zawadzke, Bugg et al. (1991). NR, not reported.

**Figure 5-10.** Initial velocity versus substrate concentration [ATP] plots for Ddl6, Ddl7, DdlAEc and DdlTd. A saturating concentration of D-alanine (100 mM) was used with a variable amount of ATP to determine the Kₘₐₚ of the enzymes. The plots generated and the Kₘₐₚ values calculated by the curve fitting options of GraphPad Prism software. *The scale at the vertical axis of DdlAEc is different.
The $K_{m,D\text{-ala}_2}$ of the DdlS were determined using a fixed saturating concentration of ATP (500 µM, approximately 10-fold of the $K_{m,ATP}$) with a variable concentration of the D-alanine concentration (0.1, 0.2, 0.4, 0.8, 1.6 and 3.2 mM). The plots of initial velocity versus substrate concentration were also found to follow Michaelis-Menten kinetics (Figure 5-11). The $K_{m,D\text{-ala}_2}$ of Ddl6, Ddl7, DdlTd and DdlAEc were similar with the values of 0.39, 0.54, 0.45 and 0.52 mM, respectively (Table 5-2). So, the $K_{m,D\text{-ala}_2}$ was not useful to answer the first question.

5.3.4.2 The $K_{cat}$ of Ddl7 was approximately 3-fold higher than Ddl6

Although the values of $K_{m,D\text{-ala}_2}$ were not helpful to explain the alteration of D-cycloserine phenotype, the turnover numbers ($K_{cat}$) or molecular activity were found to be indicative. The $K_{cat}$ indicates the number of substrate molecules converted to product per molecule of enzyme per unit time and was calculated using the equation (4). The $K_{cat}$ of Ddl7 was calculated as 3.85 s⁻¹ which is about three fold higher than Ddl6 (1.20 s⁻¹). The value of $K_{cat}/K_{m,D\text{-ala}_2}$ (catalytic efficiency) for Ddl6 was 3008.85 M⁻¹ s⁻¹, whereas this value for Ddl7 was approximately double (7010.321 M⁻¹ s⁻¹).

These differences in the values of $K_{cat}$ and $K_{cat}/K_{m,D\text{-ala}_2}$ between Ddl6 and Ddl7 correlate well with the D-cycloserine resistance phenotype of the surrogate host strains where it was found that the MIC of D-cycloserine for *E. coli* expressing Ddl7 was four-fold higher than Ddl6 (Chapter 4). However, how the increased rate for the formation of D-ala-D-ala dipeptide affect D-cycloserine resistance in *E. coli* is not clear and needs more investigation.
Table 5-2. Steady-state kinetic parameters of Ddl6, Ddl7, DdlEc and DdlTd determined with a fixed saturating concentration of ATP (500 µM) and variable concentration of D-alanine.

<table>
<thead>
<tr>
<th></th>
<th>Vmax (mM/s)</th>
<th>Kcat (s⁻¹)</th>
<th>Kₘ, D-ala2 (mM)</th>
<th>Kcat/Kₘ, D-ala2 (M⁻¹ s⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ddl6</td>
<td>0.0000778</td>
<td>1.19</td>
<td>0.39</td>
<td>3008.85</td>
</tr>
<tr>
<td>Ddl7</td>
<td>0.0002503</td>
<td>3.85</td>
<td>0.55</td>
<td>7010.32</td>
</tr>
<tr>
<td>DdlTd</td>
<td>0.0002246</td>
<td>3.74</td>
<td>0.45</td>
<td>8294.55</td>
</tr>
<tr>
<td>DdlEc</td>
<td>0.0003728</td>
<td>6.43</td>
<td>0.52</td>
<td>12517.21</td>
</tr>
<tr>
<td>DdlEc*</td>
<td>NR</td>
<td>NR</td>
<td>0.55*</td>
<td>NR</td>
</tr>
</tbody>
</table>

* The values of DdlEc reported by Zawadzke, Bugg et al. (1991). NR, not reported.

The Kcat and Kcat/Kₘ, D-ala2 for DdlTd and Ddl7 were similar. The value of Kₘ, D-ala2 of DdlEc was determined as 0.51 mM which is similar to the Kₘ, D-ala2 determined previously using the same protocol (Zawadzke, Bugg et al. 1991). This validates the method used for determination of kinetic parameters in this study.

Figure 5-11. Initial velocity versus substrate concentration [D-alanine] plots for Ddl6, Ddl7, DdlTd and DdlEc. Variable concentrations of D-alanine were used with a fixed saturating
concentration of ATP. The plots were generated and \( K_{mD-\alpha\lambda2} \) values were calculated using GraphPad Prism. *Different scale of the X-axis.

5.3.5 The instability index, aliphatic index (AI) and grand average of hydropathicity (GRAVY) of Ddl6, Ddl7 were compared with Ddls encoded by house-keeping genes

The protein parameters of Ddl6, Ddl7, DdlTd and DdlAEc were determined using ProtParam (Gasteiger E. 2005). The instability index (an estimate of stability of protein in vitro) predicted that Ddl6 and Ddl7 were unstable in vitro with instability index values of 45.18 and 45.66, respectively (a value above 40 indicates that the protein may be unstable) (Guruprasad, Reddy et al. 1990). However, the DdlTd encoded by housekeeping gene of T. denticola were predicted to be stable with index value of 31.47 (well-below the threshold value 40) (Table 5-3). The difference of instability index among the two groups of Ddls was approximately 13.

To see if the predicted stability between the two copies of Ddl of E. coli (DdlA and DdlB) and S. enterica serover Typhimurium follows the same pattern, the protein parameters for them were also calculated. Interestingly, the tool predicted DdlBEc as stable and DdlAEc as unstable. Similarly, among the two Ddls of S. enterica, one (GenBank accession no.: KKE09049) was predicted to be stable and other one as unstable (GenBank accession no.: KKE09290). These results suggest that when bacteria carry two copies of ddl, the product of one gene is stable and the other is unstable.

The Ddl6 and Ddl7 was also found to be different in terms of aliphatic index (AI) and Grand Average of Hydropathicity (GRAVY) (Table 5-3). When the AI of two isoforms was compared, Ddl7 was found be more thermostable than Ddl6. The W259C
mutation in Ddl6 is mainly responsible for this change. The GRAVY of Ddl7 was also found to be higher than Ddl6. The pI of Ddl6 and Ddl7 was identical (5.55), however, the pI of DdlTd was 7.05.
### Table 5-3. Comparison of protein parameters of Ddl6 and Ddl7 with Ddl5s encoded by housekeeping genes

<table>
<thead>
<tr>
<th></th>
<th>Ddl6</th>
<th>Ddl6 W259C</th>
<th>Ddl7</th>
<th>DdlTd</th>
<th>DdlAEc</th>
<th>DdlBEc</th>
<th>DdlSen (KKE09049)</th>
<th>DdlSen (KKE09290)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of amino acids</strong></td>
<td>344</td>
<td>344</td>
<td>344</td>
<td>368</td>
<td>364</td>
<td>306</td>
<td>306</td>
<td>364</td>
</tr>
<tr>
<td><strong>MW (KDa)</strong></td>
<td>38.30</td>
<td>38.22</td>
<td>38.22</td>
<td>41.19</td>
<td>39.31</td>
<td>32.89</td>
<td>32.63</td>
<td>39.35</td>
</tr>
<tr>
<td><strong>Theoretical pl</strong></td>
<td>5.55</td>
<td>5.55</td>
<td>5.55</td>
<td>7.05</td>
<td>5.02</td>
<td>4.71</td>
<td>4.73</td>
<td>5.39</td>
</tr>
<tr>
<td><strong>Total number of negative charge residue</strong></td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>54</td>
<td>46</td>
<td>39</td>
<td>37</td>
<td>40</td>
</tr>
<tr>
<td><strong>Total number of positive charge residue</strong></td>
<td>39</td>
<td>39</td>
<td>39</td>
<td>54</td>
<td>31</td>
<td>25</td>
<td>23</td>
<td>30</td>
</tr>
<tr>
<td><strong>Instability index</strong></td>
<td>45.18</td>
<td>45.74</td>
<td>45.66</td>
<td>31.47</td>
<td>40.91</td>
<td>39.27</td>
<td>42.17</td>
<td>38.34</td>
</tr>
<tr>
<td><strong>Stability</strong></td>
<td>Unstable</td>
<td>Unstable</td>
<td>Unstable</td>
<td>Stable</td>
<td>Unstable</td>
<td>Stable</td>
<td>Unstable</td>
<td>Stable</td>
</tr>
<tr>
<td><strong>Aliphatic index</strong></td>
<td>98.53</td>
<td>103.72</td>
<td>102.59</td>
<td>89.10</td>
<td>102.36</td>
<td>98.53</td>
<td>100.6</td>
<td>100.49</td>
</tr>
<tr>
<td><strong>GRAVY</strong></td>
<td>0.067</td>
<td>-0.092</td>
<td>-0.95</td>
<td>-0.342</td>
<td>0.016</td>
<td>0.067</td>
<td>0.120</td>
<td>-0.009</td>
</tr>
</tbody>
</table>
5.4 Discussion

This study has shown for the first time that paper chromatography and ninhydrin can be used to detect the formation of D-ala-D-ala dipeptide. The dipeptide formation assay is generally performed using thin-layer chromatography (TLC) using radio-labelled D-alanine, D-serine or D-lactate ($^{14}$C-labelled) (Marshall and Wright 1998, Feng and Barletta 2003, Meziane-Cherif, Saul et al. 2012). Paper chromatography is widely used for its simplicity and reliability to separate amino acids and peptides in a mixture. This technique was introduced by Consden, Gordon et al. (1944) for qualitative analysis of protein using a cellulose filter paper. A solvent system based on n-butanol: acetic acid: water is most commonly used to separate the amino acids (Heimer 1972). For detection of the amino acid, ninhydrin is mixed with the solvent system or sprayed on the paper after the run is complete. Ninhydrin reacts with amino acids and forms an aldehyde and carbon dioxide in stoichiometric amounts and different amounts of ammonia, hydrindantin and Ruhemann’s Purple (diketohydrindylidenediketohydrindamine). The detection and quantification of α-amino depends on this purple pigment (Bottom, Hanna et al. 1978). Ninhydrin was also used to detect separation of amino acids by TLC (Qiu, Li et al. 2010). It was envisaged that as the amino acids can be detected with ninhydrin by paper chromatography, it can also be used to detect the dipeptide/depsipeptide formation assay. Using the same solvent system used for TLC to detect D-ala-D-ala dipeptide [(n-butanol-acetic acid-water (12:3:5)] (Feng and Barletta 2003) and the same concentration of ninhydrin used for amino acid detection by TLC (Qiu, Li et al. 2010), this study attempted to detect the formation of D-ala-D-ala
dipeptide and D-ala-D-ser or D-ala-D-lac depsipeptide in the reactions catalysed by Ddl. The results showed that this method can be used for the analysis of Ddl activity without using radioactive isotopes. Moreover, this method is cheap and does not need a quarantine area.

It was shown previously that when the Ddls or their homologues show weak specificity for alternative substrates at subsite 2 (C-terminal), D-ala-D-ala dipeptide is produced in the reactions which can be seen on the TLC. However, when the enzymes at subsite 2 show very strong specificity for alternative substrate such as D-serine or D-lactate, the formation of D-ala-D-ala cannot not be detected. For instance, VanG (a Ddl homologue that confers vancomycin resistance) exhibits very high specificity for D-ser for the second subsite, thus, it forms D-ala-D-ser dipeptide as the only reaction product (Meziane-Cherif, Saul et al. 2012) (Lane 3, Figure 5-12). However, for the other substrates such as aminobutyrate and D-norvaline to which VanG exhibits a weak specificity, the formation of D-ala-D-ala was also observed (Figure 5-12).

The formation of D-ala-D-ala was more prominent when VanG could not bind the second substrate as was seen for D-lactate (lane 4) or D,L-hydroxyvalerate (lane 6) (Figure 5-12) (Meziane-Cherif, Saul et al. 2012). This happens due to the free D-alanine into the reaction that occupy the subsite 2. This explains why the reactions catalysed by Ddl7, where D-lactate or D-serine were added along with D-alanine, the D-ala-D-ala dipeptide was produced.
It was shown in chapter 4 that the MIC of D-cycloserine against *E. coli* expressing Ddl7 was four-times higher than *E. coli* expressing Ddl6. This chapter also aimed to determine the kinetic parameters of Ddl6 and Ddl7 to explain this alteration of D-cycloserine resistance phenotype of *E. coli*. One of the assumptions was that increased resistance to D-cycloserine could be related to higher catalytic activity of the Ddl7 than Ddl6. The assumption was based on the relation of $K_{cat}$ and $K_i$ for D-cycloserine at subsite 2 (C-terminal) ($K_i,DCS2$) of two variants of Ddl of *E. coli* (Zawadzke, Bugg et al. 1991). It was shown that $K_i,DCS2$ for DdlA and DdlB of *E. coli* were proportional to the values of $K_{cat}$. The $K_{cat}$ for DdlB was 2.5-fold higher than DdlA (444 and 1018 min$^{-1}$, respectively) and proportionately, the $K_i,DCS2$ for DdlB was also approximately 3-fold higher than DdlA (8.9 and 27 $\mu$M, respectively) (Zawadzke, Bugg et al. 1991).
The similar relation of $K_{cat}$ and $K_{i, DCS2}$ was also observed for DdlA of *S. typhimurium* ($K_{cat} = 444 \text{ min}^{-1}$ and $K_{i, DCS2} = 14 \text{ µM}$) (Zawadzke, Bugg et al. 1991) and Ddl of *M. tuberculosis* ($K_{cat} = 9.7 \text{ s}^{-1}$ or $582 \text{ min}^{-1}$ and $K_{i, DCS2} = 14 \text{ µM}$) (Prosser and de Carvalho 2013)(Table 5-4). However, this positive correlation of higher $K_{cat}$ with higher $K_{i, DCS2}$ is not followed by the Ddl of *Streptomyces lavendulae* (DdlS), one of the natural producers of D-cycloserine (Noda, Kawahara et al. 2004) where these two parameters were inversely proportional. The $K_{cat}$ for DdlS was only 100 min$^{-1}$ (approximately 10 times less than DdlBEc) whereas the $K_{i, DCS2}$ was 920 µM (approximately 34 times higher than DdlBEc) (Noda, Kawahara et al. 2004). The high $K_{i}$ for D-cycloserine suggested that the DdlS evolved as a self-resistant determinant to protect the host from its product, D-cycloserine. So, if the Ddl6 and Ddl7 would have evolved within integron as a mechanism to confer D-cycloserine resistance, they should also exhibit high $K_{i}$ values for D-cycloserine. So, to understand the mechanism of evolution of Ddl6 and Ddl7 within integron as well to understand the reason for alteration of D-cycloserine resistance due to the mutation at p.259, it is important to determine the $K_{i, DCS2}$ for these proteins.

Table 5-4. Comparison of kinetic parameters of Ddl6 and Ddl7 with DdlA and DdlB of *E. coli* as well as Ddl of *S. lavendulae* and Ddl of *M. tuberculosis*.

<table>
<thead>
<tr>
<th>Kinetic parameters</th>
<th>Ddl6</th>
<th>Ddl7</th>
<th>DdlA (<em>S. typhimurium)</em></th>
<th>DdlA (<em>E. coli)</em></th>
<th>DdlB (<em>E. coli)</em></th>
<th>Ddl (<em>M. tuberculosis)</em></th>
<th>Ddl (<em>S. lavendulae)</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_{cat}$ (min$^{-1}$)</td>
<td>71*</td>
<td>231*</td>
<td>644</td>
<td>444</td>
<td>1018</td>
<td>582*</td>
<td>100</td>
</tr>
<tr>
<td>$K_m$, D-alal2 (mM)</td>
<td>0.39</td>
<td>0.55</td>
<td>0.54</td>
<td>0.55</td>
<td>1.2</td>
<td>3.6</td>
<td>2.0</td>
</tr>
<tr>
<td>$K_{i, DCS2}$ (µM)</td>
<td>ND</td>
<td>ND</td>
<td>14</td>
<td>8.9</td>
<td>27</td>
<td>14</td>
<td>920</td>
</tr>
</tbody>
</table>

*Zawadzke, Bugg et al. (1991); Prosser and de Carvalho (2013); Noda, Kawahara et al. (2004). The values were originally determined in s$^{-1}$, so, the units in s$^{-1}$ were converted to min$^{-1}$ by multiplying with 60. ND, not determined. DCS2, D-cycloserine.
In chapter 4, it was shown that the likely host of the integron carrying $ddls$ is a strain of *T. denticola*. This assumption was based on the identity of the upstream sequence of the integron with the genome of other *T. denticola* strains. In this chapter, the kinetic parameters of Ddl of *T. denticola* (DdlTd) were determined to test the hypothesis that *T. denticola*, the putative host of the integron, may have acquired another $ddl$ within integron by HGT that has higher catalytic activity. Gene acquisition by HGT is very unlikely to be related only to increased production of the enzyme with similar level of catalytic activity or same function. Because, it was shown previously that gene transfer is mainly related to acquisition of new functions, whereas the gene duplication results in higher gene dosage (Treangen and Rocha 2011). As the $K_{cat}$ (s$^{-1}$) of DdlTd was very close to Ddl7 and was 3 times higher than Ddl6, it was concluded that the acquisition of a second copy $ddl$ within the integrons was not related to gain a Ddl with higher activity. So, it is likely that the acquired gene product have other functions which is not related to increased production of the enzyme. The next chapter tested another interesting hypothesis of evolution of Ddl within integrons by using molecular docking approach.

This chapter also calculated some protein parameters to see if they provide some clues regarding the evolution of $ddls$ within integron GCs. The instability index predicted Ddl6 and Ddl7 as unstable *in vitro* but Ddl of *T. denticola* was as stable and the difference in the instability index was approximately 13 units. This pattern of *in-vitro* instability can also be seen in *E. coli* and *S. enterica* carrying two copies of the gene encoding Ddl where the product of one copy of $ddl$ was stable and the other one was unstable. This suggests that perhaps to maintain the cell wall integrity of the host, the gene product of one copy of the
genes needs to be degraded quickly. As the DdlS encoded by the integron-located genes were predicted to be more unstable than the Ddl encoded by the house-keeping *ddl* of the putative host, it was hypothesized that perhaps the Ddl6/Ddl7 are overexpressed to protect the host from a short-term stress and when the stress is tackled, the proteins are degraded quickly. However, it should be stressed here that the instability index is calculated to predict the instability of the proteins *in vitro*, so, the stability of the proteins in the cytoplasm of the host could be different.

Additionally, as the thermostability of globular proteins is proportional to the AI (Ikai 1980), the higher values of AI of Ddl6/Ddl7 than DdlTd suggest that the integron-encoded Ddls may also have a role in adaptation of the host to high temperatures. Furthermore, the pI values of Ddl6/Ddl7 and DdlTd suggest that the integron-encoded proteins are more stable at low pH (pI 5.55) compared to DdlTd (pI 7.05). The acidic nature of Ddl6 and Ddl7 could be related with the adaptation of the host to low pH caused by the breakdown of carbohydrates by oral bacteria.

**5.5 Conclusion**

The results covered in this chapter showed that the two isoforms of Ddl encoded by GC of a reverse integron are functional which was proved by the measuring the release of Pi and paper chromatographic detection of dipeptide formation for the first time. Kinetic studies showed that Ddl7 is more catalytically active than Ddl6. Although the catalytic activity of Ddl6 and Ddl7 correlates the D-cycloserine resistance phenotype, the mechanism how it affects MIC is not clear and needs further investigation. The protein parameters suggest that the
integron GC-encoded proteins may have a role in adaptation of the host to high temperature as well as low pH. However, as these observations are based on \textit{in silico} analysis, they need to be validated experimentally.
Chapter Six: Determination of predicted 3D structures of DdIs and molecular docking
6.1 Introduction

The biological function of a protein depends on its three-dimensional (3D) structure; hence, the determination of 3D protein structure is the key to understanding how the protein functions in the cell. The 3D structures of proteins are experimentally obtained by using one of the following techniques: crystallography (X-ray diffraction), nuclear magnetic resonance (NMR) or electron microscopy (Dorn, e Silva et al. 2014). However, these methods are expensive and time consuming. Due to these impediments, the gap between the number of amino acid sequences of proteins generated by genome projects and the number of proteins with solved 3D structures is increasing day by day. It has been determined that only a very small fraction of the proteins with known sequences have their experimentally solved 3D structures in the Protein Data Bank (PDB) library (0.6% in 2009 and 1.2% in 2007) (Roy, Kucukural et al. 2010, Dorn, e Silva et al. 2014). However, the recent developments in the field of computer algorithms to predict the 3D structure of a protein have much alleviated this problem (Zhang 2008). Using computer algorithms, biologists now can obtain a near accurate 3D structure of a protein depending on the identity of the target sequence with the template structures in PDB whose structures were experimentally determined (Zhang 2009).

The computational methods currently used for predicting 3D structures can be divided into three categories (Zhang 2009): comparative modelling (or homology modelling), threading, and free modelling or \textit{ab initio} modelling. In comparative modelling, a 3D model of a protein (the target) is generated by matching the amino acid sequence of the target to an evolutionary related
A protein whose structure is known (the template) (Bajorath, Stenkamp et al. 1993, Bordoli, Kiefer et al. 2009). The steps involved in comparative modelling are shown in Figure 6-1.

**Figure 6-1.** Steps involved in comparative modelling. The first step of the process is the identification of a related protein with known 3D structure (fold recognition or fold assignment). In the second step the target and template sequences are aligned and then a model is built based on the alignment and structure of the template. At the last step the built 3D structure is further refined and validated. When no templates are found for the target sequence, models can be generated using ab initio modelling.

The threading or fold recognition method is based on the fact that there are many proteins which show high structural similarity even though they have very
little or no amino acid sequence identity (Jones, Taylort et al. 1992). In this method, the 3D structure of a protein is predicted by using the structural information such as the environment of each residue in the structure and solvent accessibility, so, this method can detect more distant sequence-structure relationships (Jones, Taylort et al. 1992). As both comparative and threading methods are based on the availability of templates in PDB, they are also known as the template-based method. The ab initio or free modelling is used when no suitable templates are found in PDB (Figure 6-1). This method is based on the fact that the native structure of a protein corresponds to the most thermodynamically stable state (Liwo, Lee et al. 1999). As this approach does not use any template, it also called free-modelling.

The availability of templates in PDB and percentage identity among the target and templates largely determines the accuracy of the models and their biological usefulness (Zhang 2009) (Figure 6-2). High-resolution models with root mean square deviation (RMSD) in 1-2 Å are typically generated by comparative modelling and they can be used for computational ligand-binding experiments and virtual screening. The medium-resolution models with RMSD in the range of 2.5-5 Å are generally generated by comparative modelling or by the fold recognition method and they can be used for locating functionally important residues in the protein structures and designing site directed mutagenesis (Figure 6-2) (Arakaki, Zhang et al. 2004, Zhang 2009).
Molecular docking is a tool used to predict the predominant binding mode(s) of a ligand with a protein of known 3D structure (target) (Morris and Lim-Wilby 2008). The process of molecular docking was pioneered during the late 1970s (Wodak and Janin 1978) and the early 1980s (Kuntz, Blaney et al. 1982). The structure of the protein target to be used for docking should ideally be determined experimentally by X-ray crystallography or NMR, however, 3D structures determined by computational methods are also widely used (Hillisch, Pineda et al. 2004, Morris and Lim-Wilby 2008). As the results of docking strongly depend on the accuracy of the protein structure used as a target, it is very important to use good quality models for docking studies (Bordogna, Pandini et al. 2011).
Crystal structures of Ddl of many different bacterial species have been solved in the past 23 years since the first structure of DdlB of *E. coli* was determined co-crystallized with ATP and 1(S)-aminoethyl [2-carboxy-2(R)-methyl-1-ethyl] phosphinic acid (Fan, Moews et al. 1994) (Figure 6-3). As of today, a total of 41 atomic resolution structures of Ddl and its homologues have been published in RCSB Protein Data Bank (PDB) (http://www.rcsb.org/pdb/home/home.do) (searched using D-alanine-D-alanine ligase as query). Most of the structures were determined in complex with phosphonate and phosphinate inhibitors, ADP, magnesium and sulphate as ligands. ADP was found to bind the central and C-terminal domains in DdlB of *E. coli*. ADP appeared to be sandwiched between two β-sheets, one from central and another from C-terminal domains (Fan, Moews et al. 1994).

The analysis of these crystal structures of Ddls and their homologues reveals that they are comprised of three domains: N-terminal, central and C-terminal. Each of these domains is formed by β-sheets surrounded by α-helices and loops (Tytgat, Colacino et al. 2009).
Based on the crystal structure of DdlB of *E. coli* (PDB code: 2DLN) the active site residues of Ddl that bind two D-alanine were identified and the mechanism of formation of D-ala-D-ala dipeptide formation was proposed (Fan, Moews et al. 1994). The molecular basis of vancomycin resistance by VanA was also predicted based on this structure (Fan, Moews et al. 1994) which was later confirmed by atomic resolution structures of VanA (Roper, Huyton et al. 2000).

The crystal structure of *E. coli* DdlB revealed that N-terminal Glu15 is responsible for positioning the first D-alanine (D-ala1) by hydrogen bonding. This causes a polarisation of the carboxylate oxygen atom which subsequently attacks the γ-phosphate of ATP and the transfer of phosphate from ATP to D-ala1 to form D-alanyl-acylphosphate (Figure 6-4). This transfer is coordinated by Mg\(^{2+}\) bound to Glu270. The second D-alanine (D-ala2) is positioned by hydrogen bonding to three residues Tyr216, Ser281 and Leu282 on the helix 11.

**Figure 6-3.** DdlB of *E. coli* complexed with ADP 1(S)-aminoethyl [2-carboxy-2(R)-methyl-1-ethyl] phosphinic acid as phosphinophosphonate that arise from phosphoryl transfer from ATP (Fan, Moews et al. 1994). The PDB file of DdlB (PDB code: 2DLN) was downloaded from protein data bank and viewed and analysed in UCSF Chimera.
(H11) of DdlB. The phenolic –OH of Tyr216 is positioned by hydrogen bonds with Glu15 and Ser150. In the next step, the nucleophilic α-NH₂ attacks the electrophilic carbonyl centre of the D-alanyl-acylphosphate which produces a tetrahedral intermediate. This intermediate is then decomposed to release inorganic phosphate and D-ala-D-ala dipeptide as the reaction product (Figure 6-4) (Fan, Moews et al. 1994).

**Figure 6-4.** Proposed mechanism of formation of D-ala-D-ala dipeptide by DdlB of E. coli. The figure was adapted from the article published by Fan, Moews et al. (1994) and modified. The predicted steps for formation of D-ala-D-ala dipeptide are shown.

The characterization of the crystal structures of different proteins showed that omega-loops are usually found on the protein surface and are believed to play a central role in function and molecular recognition of the proteins (Fetrow 1995). The length of an omega-loop can vary between 6 and 16 amino acid residues, however, the definition has been relaxed to include segments larger than 16.
residues (Fetrow 1995). In the Ddl6s and related proteins, the omega-loop was found to have a critical role in catalytic activity and substrate specificity (Fan, Moews et al. 1994, Roper, Huyton et al. 2000, Meziane-Cherif, Saul et al. 2012). The mobile omega-loop of DdlB of *E.coli* is composed of 16 residues extending from 205-220 which harbours the two important residues Lys215 and Tyr216 that are responsible for binding to ATP and D-ala2, respectively (Fan, Moews et al. 1994). The trio of residues including Tyr216, Ser150, and Glu15 that are hydrogen bonded connect the mobile omega-loop to the amino group of D-ala1 when it is activated by phosphorylation by ATP (Fan, Moews et al. 1994, Park, Lin et al. 1996).

The Tyr216 and Ser150 residues of the trio were found to be very important for catalytic activity and substrate specificity of DdlB of *E. coli* (Fan, Moews et al. 1994). The Y216F and S150A mutants of *E.coli* DdlB were found to gain a weak D-ala-D-lac depsipeptide activity, but also keep their original D-ala-D-ala ligase activity (Park, Lin et al. 1996). At higher pH, the formation of D-Ala-D-Ala is favoured, however, at the pHs less than 7, the formation of D-Ala-D-lactate is favoured in both VanA and the Y216F DdlB mutant (Park, Lin et al. 1996).

This chapter considers the generation of 3D structures of Ddl6 and Ddl7 using computational methods. The predicted 3D structures have been used to observe any conformational change in the putative tertiary structures between Ddl6 and Ddl7. The spatial position of putative active/binding sites of Ddl6 and Ddl7 have been located by aligning their 3D structures with their structural homologues in PDB. This chapter also covers molecular docking studies to identify if there is any change in the binding of ligands due to Trp259→Cys substitution in Ddl6. The molecular docking approach has also been used to
predict if the plants flavonoids such as apigenin and quercetin can bind to both D-alanine and ATP binding sites of Ddl6 and its mutants.

6.2 Materials and methods

6.2.1 Prediction of 3D structures of Ddl6, Ddl7 and DdlTd

The 3D models of Ddl6, Ddl7 and DdlTd were constructed using I-TASSER (Iterative Threading ASSEmble Refinement) (http://zhanglab.ccmb.med.umich.edu/I-TASSER) (Yang and Zhang 2015). I-TASSER uses a composite approach (combination of threading and \textit{ab initio} modelling) to generate 3D models. The steps of I-TASSER modelling are shown in Figure 6-5.

\textit{Figure 6-5. Flowchart of the I-TASSER Suite pipelines that consist of three steps: template identification, full-length structure assembly and structure-based function annotation. The figure was taken from Yang and Zhang (2015).}

To generate models using I-TASSER, the target sequences were submitted to the server and among the top five models built by the server, the best model was selected based on the highest C-score and TM-score. The C-score is a confidence score for estimating the quality of predicted models by I-TASSER.
The C-score is typically in the range of -5 to 2, where a C-score of higher value signifies a model with a high confidence and vice-versa. On the other hand, the TM-score is a scale for measuring the structural similarity between two structures. A TM-score >0.5 indicates a model of correct topology and a TM-score<0.17 means a random similarity (Zhang and Skolnick 2004). The model quality was also assessed by the root-mean square division (RMSD). RMSD is an average distance of all residue pairs in two structures. The lower the values of RMSD, the higher the resolution of the models (Zhang and Skolnick 2004). Molecular graphics and analyses were performed by using the UCSF Chimera package v. 1.10.2 (University of California, San Francisco) (Pettersen, Goddard et al. 2004) which is free to download at http://www.rbvi.ucsf.edu/chimera/.

6.2.2 Molecular docking

The predicted 3D-structures of Ddl6 and its two single mutants, L164F and W259C, as well as Ddl encoded by house-keeping gene of T. denticola (DdlTd) were used as the targets for molecular docking. The 3D structure of Ddl6 was manipulated in UCSF Chimera to generate the L164F and W259C. These protein targets were used for molecular docking of six ligands including D-alanine, D-cycloserine, ATP, apigenin, quercetin and salvicine. AutoDock SMINA (Koes, Baumgartner et al. 2013) was used to find the best binding pocket by exploring all probable binding cavities in the proteins. Then GOLD (Genetic Optimization for Ligand Docking) (Jones, Willett et al. 1995, Jones, Willett et al. 1997) was used for docking the ligands into the SMINA-located best binding site to perform flexible molecular docking and determine more precise and evaluated energies and scores. Based on the fitness function scores and ligand binding position, the best-docked poses for each ligand are
selected. The less fitness function energy of poses, generated using the GOLD program that has the highest GOLD fitness score reveals the best-docked pose for each system.

Genetic algorithm (GA) is used in GOLD ligand docking to thoroughly examine the ligand conformational flexibility along with partial flexibility of the protein (Nissink, Murray et al. 2002). The maximum number of runs was set to 20 for each compound and the default parameters were selected (100 population size, 5 for the number of islands, 100,000 number of operations and 2 for the niche size). Default cut-off values of 2.5Å (dH-X) for hydrogen bonds and 4.0Å for van-der-Waals distance were employed. When the top solutions attained the RMSD values within 1.5 Å, the GA docking was terminated. The docking analysis was performed by our collaborators at Kings College London (KCL).

6.3 Results

6.3.1 Predicted 3D structures of Ddl6 and Ddl7 obtained using I-TASSER could be superimposed with each other

The best ranked model (model 1) among the five generated by I-TASSER were used for further analysis. The values for C-scores, estimated TM scores and RMSD for the models of Ddl6 and Ddl7 are shown in Table 6-1. The values of C-score suggest that the models were built with high confidence. The TM-scores for both models were much higher than the threshold level (>0.5) for correct topology. However, when the quality parameters of the 3D model of Ddl6 were compared with Ddl7, it was found that the overall quality of 3D model of Ddl7 was better than Ddl6 (Table 6-1). These variations in the quality
parameters were reproducible when the models were regenerated two more times and the values were compared.

**Table 6-1. Values of C-scores, estimated TM-scores and RMSD of the model 1 of Ddl6 and Ddl7 as estimated by I-TASSER**

<table>
<thead>
<tr>
<th>Proteins</th>
<th>C-score</th>
<th>TM-score</th>
<th>RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ddl6</td>
<td>0.92</td>
<td>0.84±0.08</td>
<td>4.6±3.0</td>
</tr>
<tr>
<td>Ddl7</td>
<td>1.15</td>
<td>0.87±0.07</td>
<td>4.1±2.8</td>
</tr>
</tbody>
</table>

Next, to observe if the substitutions at p.164 and p. 259 between Ddl6 and Ddl7 cause a conformational change, the predicted 3D structure of Ddl6 was superimposed with Ddl7 using TM-align (Zhang and Skolnick 2005). They were found to superimpose very well with a TM-score of 0.97 and no major difference in conformation could be found except in the omega-loop region (Figure 6-7). This suggests that the substitutions at 164 and 259 positions of Ddl6 and Ddl7 may have an impact on the structure of the omega-loop.
6.3.2 The putative omega-loop of Ddl6 and Ddl7 are located within the C-terminal domain of the proteins and contain two important residues for catalytic activity and substrate specificity

Three putative domains of Ddl6 and Ddl7 were identified (Figure 6-6) by comparing the homology-modelled 3D structures of Ddl6 and Ddl7 with the crystal structure of D-ala-D-ser ligase (VanG, PDB code: 4FU0) (Meziane-Cherif, Saul et al. 2012), DdlB of *E. coli* (Fan, Moews et al. 1994), VanA of *E. faecium* (Roper, Huyton et al. 2000) and Ddl of *S. aureus* (Liu, Chang et al. 2006). The predicted N-terminal domain of Ddl6 and Ddl7 runs from the N-terminus to Gly125 which contains 3 predicted α-helices and four β-sheets. The predicted central domain runs from Ser126 to Gly219 and contains four α-helices and four β-sheets and the predicted C-terminal domain contains four α-
helices and 6 β-sheets and it spans the residues from Arg220-Ile340 (Figure 6-6). The Trp259 of Ddl6 whose substitution with Cys259 causes an increase in 4-fold D-cycloserine resistance was found to be located at the end of the 4th β-sheet (residues Asn256-Ile258) of C-terminal domain (Figure 6-6).

The putative omega-loop, which plays a very important role in catalytic activity and substrate specificity of Ddl proteins, is located on C-terminal domain of Ddl6 and Ddl7 and runs from Lys240-Ser255 (Figure 6-6). In DdlB of *E. coli* this loop spans the residues Pro205-Glu220 (Fan, Moews et al. 1994) and in VanG of *E. faecalis* residues Ser248-Ser262 (Meziane-Cherif, Saul et al. 2012). The Ser150 and Tyr216 residues of *E. coli* DdlB are predicted to bind the second D-alanine, which corresponds to the Ser184 and Tyr250 of Ddl6 and Ddl7 (Figure 6-8). As these residues are conserved in Ddl6 and Ddl7, this supports the *in-vitro* assays (Chapter 5) that Ddl6 and Ddl7 failed to produce D-ala-D-lac or D-ala-D-ser.
6.3.3 Structural homologues of Ddl6 and Ddl7 as identified by I-TASEER include vancomycin resistance proteins

After the structure assembly simulation, the models of Ddl6 and Ddl7 were matched to all structures in the PDB library to identify the top 10 proteins that
have the closest structural similarity to the predicted I-TASSER model. The results of the analysis are shown in Table 6-2. Due to structural similarity these structural homologous proteins are expected to have a similar function to the target (Ddl6/Ddl7). It can be seen from Table 6-2 that among the top 10 homologous proteins, six were Ddl of different species of bacteria and the remaining four are related to vancomycin resistance (PDB codes 4FU0, 1EHI, 1E4E and 3SE7). The RMSD values obtained from superimposition of Ddl6 and Ddl7 with their closest homologues were in the range of 1.20-2.68 Å (Table 6-2) which suggests the reliability of the predicted 3D structures of Ddl6 and Ddl7.

**Table 6-2.** Top 10 structurally homologous protein of Ddl6 and Ddl7 in PDB. The I-TASSER models of Ddl6 and Ddl7 were used to compare the structural similarity with the PDB structures.

<table>
<thead>
<tr>
<th>Rank</th>
<th>PDB hits</th>
<th>Resolution (Å)</th>
<th>Protein name, size and the host</th>
<th>TM-score</th>
<th>RMSD</th>
<th>% identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3i12</td>
<td>2.2</td>
<td>DdlA of <em>Salmonella enterica</em> subsp. <em>enterica</em> serovar Typhimurium str. LT2 (364 aa)</td>
<td>0.943</td>
<td>0.966</td>
<td>1.82</td>
</tr>
<tr>
<td>2</td>
<td>4L1K</td>
<td>2.3</td>
<td>Ddl of <em>Xanthomonas oryzae</em> pv. <em>Oryzae</em> (384 aa)</td>
<td>0.910</td>
<td>0.922</td>
<td>1.59</td>
</tr>
<tr>
<td>3</td>
<td>4FU0</td>
<td>2.35</td>
<td>VanG (D-ala-D-ser ligase) of <em>Enterococcus faecalis</em> (351 aa)</td>
<td>0.901</td>
<td>0.906</td>
<td>1.99</td>
</tr>
<tr>
<td>4</td>
<td>1EHI</td>
<td>2.38</td>
<td>D-ala-D-lac ligase (LmDdl2) of vancomycin-resistant <em>Leuconostoc mesenteroides</em> (377 aa)</td>
<td>0.907</td>
<td>0.890</td>
<td>2.45</td>
</tr>
<tr>
<td>5</td>
<td>3TQT</td>
<td>1.88</td>
<td>Ddl of <em>Coxiella burnetii</em></td>
<td>0.886</td>
<td>0.596</td>
<td>1.89</td>
</tr>
<tr>
<td>Rank</td>
<td>PDB hits</td>
<td>Resolution (Å)</td>
<td>Protein name, size and the host</td>
<td>TM-score</td>
<td>RMSD&lt;sup&gt;b&lt;/sup&gt;</td>
<td>% identity&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>------</td>
<td>----------</td>
<td>----------------</td>
<td>---------------------------------</td>
<td>----------</td>
<td>----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>6</td>
<td>1E4E</td>
<td>2.5</td>
<td>VanA (D-ala-D-lac ligase) of E. faecium BM4147 (343 aa)</td>
<td>0.885</td>
<td>0.882</td>
<td>2.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>3SE7</td>
<td>3.07</td>
<td>VanA, metagenomics (346 aa)</td>
<td>0.872</td>
<td>0.876</td>
<td>2.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2ZDH</td>
<td>1.9</td>
<td>Ddl of <em>Thermus thermophilus</em> HB8 (319 aa)</td>
<td>0.867</td>
<td>0.859</td>
<td>1.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2I8C</td>
<td>2.46</td>
<td>Ddl of <em>S. aureus</em> (358 aa)</td>
<td>0.852</td>
<td>0.858</td>
<td>2.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1IOW</td>
<td>1.9</td>
<td>Y216F mutant of DdlB of <em>E. coli</em> (306 aa)</td>
<td>0.812</td>
<td>0.812</td>
<td>1.90</td>
</tr>
</tbody>
</table>

<sup>a</sup> Ranking of proteins is based on TM-score of the structural alignment between the query structure (Ddl6/Ddl7) and known structures in the PDB library.
<sup>b</sup> RMSD is the RMSD between residues that are structurally aligned by TM-align;
<sup>c</sup> % identity is the percentage sequence identity in the structurally aligned region

The alignment of the predicted 3D structures of Ddl6 and Ddl7 with two homologous proteins (PDB code 3I12 and 4L1K) (Table 6-2) are shown in Figure 6-9. It can be seen that although the overall structure of Ddl6 and Ddl7 with their homologues were very similar, the omega-loops of the structures were quite different (Figure 6-9).
Figure 6-9. Alignment of predicted 3D structures of Ddl6 and Ddl7 with DdlA of S. typhimurium (PDB code 3I12) and VanG of E. faecalis (PDB code 4FU0). The Ddl6 and Ddl7 are shown in blue colour and the homologues in red colour.

6.3.4 The residues at 259 position of Ddl6 (Trp) and Ddl7 (Cys) are not part of putative ligand binding sites of DdlS, but are located very close to the omega loop

The spatial positions of the putative ligand/substrate binding sites were located within the 3D structures of Ddl6 and Ddl7 based on the experimentally verified active sites of DdlB of E. coli (Fan, Moews et al. 1994, Shi and Walsh 1995, Fan, Park et al. 1997). It was found that the residue at 259 whose substitution from Trp259→Cys (Ddl6) alters D-cycloserine susceptibility is located close to the omega-loop, however, spatially far from the putative active sites of Ddl6 and Ddl7 (Figure 6-10).
6.3.5 Molecular docking experiments showed that W259C mutation in Ddl6 does not alter the binding affinity of D-alanine and D-cycloserine at the D-alanine binding site

The best binding sites for D-alanine and D-cycloserine in the homology-modelled 3D structures of Ddl6, its two mutant structures and DdlTd were identified by AutoDock SMINA (methods). Then the GOLD was used to dock the D-alanine and D-cycloserine to bind the SMINA-located best binding sites of the proteins. The binding affinity was measured based on the values of predicted free energy of binding (ΔG) and the fitness scores as determined by GOLD. The lowest is the binding energy, the most favourable pose of the ligands in the binding sites.
The values of binding energies and GOLD fitness scores of D-alanine and D-cycloserine for Ddl6, L164F Ddl6, W259C Ddl6 and DdlTd showed that D-alanine binds more favourably to the Ddls compared to D-cycloserine (Table 6-3). For instance, the fitness score for D-alanine for native Ddl6 was 32.81, whereas the score for D-cycloserine for this protein was 26.63. No significant change in the binding energies and scores of D-alanine and D-cycloserine for the native Ddl6 and its mutants could be observed. However, the fitness scores of D-alanine and D-cycloserine to the Ddl encoded by the house-keeping gene of T. denticola (DdlTd) was found to be at least 5 points lower than Ddl6 and its two mutants.

Table 6-3. Free energy of binding ($\Delta G$) and GOLD scores of D-alanine and D-cycloserine with native Ddl6 and its two single mutants

<table>
<thead>
<tr>
<th></th>
<th>Ddl6</th>
<th>L164F</th>
<th>W259C</th>
<th>DdlTd</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\Delta G$ (kcal/mol)</td>
<td>Score</td>
<td>$\Delta G$ (kcal/mol)</td>
<td>Score</td>
</tr>
<tr>
<td>D-alanine</td>
<td>-40.55</td>
<td>32.81</td>
<td>-42.07</td>
<td>34.08</td>
</tr>
</tbody>
</table>
The mode of interactions of D-alanine and D-cycloserine with Ddl6 and its W259C mutant have been shown in Figure 6-12.

Figure 6-12. Mode of interaction of D-alanine and D-cycloserine with WT Ddl6 and its W259C mutant. Interaction of D-alanine with WT Ddl6 (A), D-ala with W259C mutant (B), D-cycloserine with WT Ddl6 (C) and D-cycloserine with W259C mutant.

6.3.6 Quercetin and apigenin bind to D-ala binding sites with higher affinity than the ATP-binding sites

Next, we also docked apigenin and quercetin to D-alanine and ATP binding sites of Ddl6 and Ddl7 to answer the following research question: Could these flavonoids bind to Ddl6 encoded by integron GCs? Salvicine (a terpenoid isolated from a Chinese medicinal plant, *Salvia prionitis*) was used as a positive control of the docking because this terpenoid was
previously identified as a potent inhibitor of Ddl of *T. pallidum* by using only a molecular docking tool (Dwivedi, Tiwari et al. 2015).

![Quercetin](image1.png) ![Apigenin](image2.png) ![Salvicine](image3.png)

**Figure 6-13.** 2D structures of quercetin, apigenin and salvicine. The structures were obtained using ChemDraw Professional 16.0 (PerkinElmer Informatics).

The results of GOLD molecular docking showed that quercetin and apigenin can dock to both D-alanine (Figure 6-14) and ATP binding sites (Figure 6-15) of Ddl6 and its mutants. It was previously shown that these two flavonoids are competitive inhibitors of ATP binding sites of Ddl of *H. pylori* and are non-competitive for D-alanine (Wu, Kong et al. 2008). Accordingly, it was expected that these flavonoids would exhibit a higher fitness score towards the ATP binding sites of Ddl6/Ddl7 than the D-alanine binding sites. In contrast to expectations, it was found that both apigenin and quercetin bind to D-alanine sites with higher fitness scores (Table 6-4) than the ATP binding sites of native Ddl6 (Table 6-5). For instance, the scores of quercetin and apigenin to D-alanine binding sites of WT Ddl6 were 33.42 and 35.48, respectively. However, when quercetin and apigenin were docked to ATP binding sites of WT Ddl6, the scores were calculated as 25.30 and 25.09, respectively which are approximately 8-10 units lower than the D-alanine binding sites.
Table 6-4. Free binding energy (ΔG) and GOLD fitness score of apigenin, quercetin and salvicine for the D-alanine binding sites of native Ddl6 and its two single mutants.

<table>
<thead>
<tr>
<th>D-alanine binding site</th>
<th>Quercetin</th>
<th>Apigenin</th>
<th>Salvicine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ΔG (kcal/mol)</td>
<td>Score</td>
<td>ΔG (kcal/mol)</td>
</tr>
<tr>
<td>Ddl6</td>
<td>-40.20</td>
<td>33.42</td>
<td>-43.41</td>
</tr>
<tr>
<td>L164F</td>
<td>-39.90</td>
<td>32.33</td>
<td>-44.61</td>
</tr>
<tr>
<td>W259C</td>
<td>-39.85</td>
<td>32.95</td>
<td>-41.58</td>
</tr>
<tr>
<td>DdlTd</td>
<td>-40.20</td>
<td>33.43</td>
<td>-39.90</td>
</tr>
</tbody>
</table>

Figure 6-14. Mode of interaction of apigenin and quercetin with the D-alanine binding sites of WT Ddl6 and W259C mutant. A) Apigenin with WT Ddl6, B) Apigenin with W259C mutant, D) Quercetin with WT Ddl6 and D) Quercetin with W259C mutant.
Table 6-5. Free binding energy ($\Delta G$) and GOLD fitness score of apigenin, quercetin and salvicine for the ATP binding sites of native Ddl6 and its two single mutant, L164F and W259C.

<table>
<thead>
<tr>
<th>ATP binding site</th>
<th>Quercetin</th>
<th></th>
<th></th>
<th>Apigenin</th>
<th></th>
<th></th>
<th>Salvicine</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\Delta G$ (kcal/mol)</td>
<td>Score</td>
<td>$\Delta G$ (kcal/mol)</td>
<td>Score</td>
<td>$\Delta G$ (kcal/mol)</td>
<td>Score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ddl6</td>
<td>-34.12</td>
<td>25.30</td>
<td>-27.29</td>
<td>25.09</td>
<td>-39.70</td>
<td>35.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L164F</td>
<td>-35.96</td>
<td>33.98</td>
<td>-29.76</td>
<td>29.69</td>
<td>-46.95</td>
<td>38.99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W259C</td>
<td>-37.84</td>
<td>34.52</td>
<td>-30.55</td>
<td>29.78</td>
<td>-39.54</td>
<td>37.28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DdlTd</td>
<td>-31.57</td>
<td>24.89</td>
<td>-31.98</td>
<td>27.04</td>
<td>-38.27</td>
<td>36.66</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 6-15. Model of interaction of apigenin and quercetin with the ATP binding sites of WT Ddl6 and W259C mutant. A) Apigenin with WT Ddl6, B) Apigenin with W259C mutant, D) Quercetin with WT Ddl6 and D) Quercetin with W259C mutant.

Interestingly, the fitness score for apigenin and quercetin were found to remain unaffected towards the Ddl6 mutants when they were docked to D-ala binding.
sites, however, the values increased with the ATP binding (Table 6-4 and Table 6-5). It is notable that the fitness scores for quercetin for the ATP binding sites of WT Ddl6 and DdlTd were similar (25.30 and 24.89, respectively) however, the score increased by at least 8 units for L164F and W259C mutants of Ddl6 (Table 6-5). The similar trend was also observed for apigenin for the ATP binding sites. Salvicine was found to bind both D-ala and ATP binding sites of all four proteins (Ddl6, L164f Ddl6, W259C Ddl6 and DdlTd) and the fitness scores for this terpenoid was higher than quercetin and apigenin (Table 6-4 and Table 6-5).

6.4 Discussion

In this study, the I-TASSER, a stand-alone software package for protein structure and function modelling (Yang, Yan et al. 2015) was used to generate 3D structures of the two natural variants of Ddl discovered during this study; Ddl6 and Ddl7, both encoded by integron GCs. Although there are many other popular software and tools to predict 3D structure of proteins including SWISS-MODEL, Phyre2, MODELLER etc., the reasons for the selection of I-TASSER were twofold: i) it was ranked as the best method by the community wide critical assessment of protein structure prediction (CASP) experiments including CASP7 and CASP8 (Roy, Kucukural et al. 2010), ii) the percentage sequence identity of the targets and available templates in PDB were less than 50%, so, it was important to use a tool that combines the template-based and free modelling methods. I-TASSER is a composite tool that uses threading and ab initio folding (Yang, Yan et al. 2015).
The predicted 3D structures generated by I-TASSER were found to be overall very good quality as they had a high C- and TM-score. Emphasis was given on the estimated TM-score rather than RMSD as TM-score is generally more reliable than RMSD particularly when the RMSD value is high for the I-TASSER models (Roy, Kucukural et al. 2010). The TM-scores of the models of Ddl6 and Ddl7 were >0.80, so, the models are likely to be suitable to use in molecular docking simulations (Figure 6-2). The parameters of the 3D models of Ddl6 and Ddl7 suggest that Ddl7 was better than Ddl6 in terms of C-score, TM-score and estimated RMSD. This variation was found to be the same when the models were generated two more times. This suggests that perhaps the changes of the amino acid residues at these positions affects the threading-alignment step of I-TASSER and also the overall topology of the models. These differences are also seen when the 3D models of Ddl6 and Ddl7 were aligned with each other by the TM-align tool. The aligned structures showed that the amino acid variations between these isoforms affect the conformation of the omega-loop of the proteins (Figure 6-7). If this conformational change occurs in real-time, it can also affect the catalytic efficiency of the proteins. However, more experiments are necessary to confirm these observations.

The 3D structures of Ddl6 and Ddl7 also support the results of site-directed mutagenesis discussed in chapter 3 where it was shown that the C>T substitution in *ddl6* which alters the Leu164 to Phe164 did not have any effect on the alteration D-cycloserine resistance phenotype. The 3D structures of the proteins show that the residues at p.164 are located within the central domain of the protein (Figure 6-6) and located spatially very far from the active site pocket of Ddl6 and Ddl7 (Figure 6-10A), thus this residue is very unlikely to affect the
ligand binding and catalytic activity of the proteins. On the contrary, the G>T substitution in *ddl6* which alters Trp259 to Cys259 was found to increase D-cycloserine resistance by four fold (Chapter 4) and it can be seen in the 3D structures that this residue is located within the C-terminal domain (Figure 6-6) and very close to the pocket of active sites (Figure 6-10). However, these residues at p.259 and the corresponding residues in other Ddl proteins are not known to be involved in ligand binding and catalytic activity. So, how the W259C mutation in Ddl6 increases D-cycloserine resistance is not clear. Perhaps, this nonsynonymous mutation alters the binding of D-alanine and ATP with the putative active sites of the proteins by altering the movement and/or conformation of the omega-loop after substrate binding, thus, alters D-cycloserine resistance phenotype. This assumption is based on very different physicochemical properties of tryptophan and cysteine. Cysteine is a polar amino acid and it has a very reactive thiol (-SH) group (Figure 6-16) which is very important for activity in many proteins including thiol proteinases (Whitford 2005).

![Figure 6-16. 2D structures of cysteine and tryptophan. The structures were obtained using ChemDraw Professional 16.0 (PerkinElmer Informatics).](image)
The cysteine residues form disulphide bridges in the proteins and play an important role in the final 3D conformation of the proteins (Sevier and Kaiser 2002). Tryptophan is a non-polar aromatic acid and is hydrophobic in nature. Although it plays an important role in protein folding, it is not reactive like cysteine is. Thus, it is understandable that Trp→Cys substitution may have a major impact on protein structure and function due to the very opposite nature of the amino acids which may also affect ligand binding and catalytic activity.

Next, a molecular docking approach was used to test if the mutations at p.164 and p.259 alter the binding characteristics of D-alanine, ATP to the predicted 3D structures of Ddl6 and its two mutants, L164F and W259C.

It is interesting that although the RSCB PDB database has more than 41 crystal structures of Ddl from 16 different species of bacteria including some clinically important pathogens such as A. baumannii (PDB code 5DMX), Yersinia pestis (PDB code 5BPF), Coxiella burnetii (PDB code 3TQT), S. aureus (PDB code 3N8D) and M. tuberculosis (PDB code 3LWB), among the top 10 structural homologues of Ddl6/Ddl7 in PDB, four homologues were found to be related to vancomycin resistance that includes VanA of E. faecium, VanG of E. faecalis and D-alanine-D-lactate ligase of L. mesenteroides (Table 6-2). The very good TM-scores (>0.9 for VanG and >0.85 for VanA) also indicate their close similarity in the 3D structures. It is known that the predicted function of a protein is more related to the structure rather than the sequence (Sanchez, Pieper et al. 2000). So, this inclination of the predicted 3D structures of Ddl6 and Ddl7 towards vancomycin resistant determinants suggests that if the corresponding residues of VanG or VanA which are responsible for binding and positioning D-ser (VanG-type) or D-lac (VanA-type) instead of D-ala2 are
substituted to the corresponding residue of Ddl6 and Ddl7 (Tyr250), they could gain D-ala-D-ser dipeptide or D-ala-D-lac depsipeptide ligase activity. However, only the substitution of Tyr250 of Ddl6/Ddl7 with corresponding residue of VanA (His244) or VanG (Phe252) may not be sufficient to gain vancomycin resistance. Because, the vancomycin resistant determinants in VRE (vancomycin-resistant enterococci) including \textit{vanA} and \textit{vanG} are located in a gene cluster (Arthur, Molinas et al. 1993), where the associated genes are involved in the regulation and expression of vancomycin resistance. For instance, the \textit{vanH}, located upstream of \textit{vanA} encodes a VanH which is responsible for the synthesis of D-lactate, the substrate for VanA (Arthur, Molinas et al. 1991). The VanX encoded by \textit{vanX} gene located downstream of \textit{vanA} is a D,D-dipeptidase which hydrolyses D-ala-D-ala, thus reduces the amount of D-ala-D-ala in the cell (Arthur, Molinas et al. 1991). However, as the first few GCs in the integrons are efficiently expressed, it can be assumed that if the DdlS encoded by the integrons acquire D-ala-D-lac or D-ala-D-ser ligase activity by mutation, the alternative dipeptides may be outnumbered by the D-ala-D-ala. However, not all vancomycin resistant bacteria carry a \textit{van} operon. For instance, the vancomycin resistance in \textit{L. mesenteroides} is caused by the chromosomally encoded Ddl (LmDdl2) which has both D-ala-D-lac depsipeptide and D-ala-D-ala dipeptide ligase activity (Park and Walsh 1997). It was shown that \textit{E. coli} DdlB gained a weak D-ala-D-lac depsipeptide activity when the Tyr216 and Ser150 residues were substituted with the corresponding residue, phenylalanine and alanine of LmDdl2, respectively (Park and Walsh 1997). This supports the hypothesis that the Ddl6/Ddl7 could be able to confer resistance to
vancomycin alone if the corresponding residue of VanA, VanG or LmDdl2 are substituted with Tyr250.

Similar to DdlB of *E. coli* (Fan, Moews et al. 1994), the putative omega loop of Ddl6 and Ddl7 was found to contain 16 amino acid residues (Lys240-Ser255). The omega-loop of VanA of *E. faecium* extends from Arg242-Asn255 (14 residues) and the residue His244 (equivalent residue of Tyr216 of DdlBEc) which is located within the omega-loop was found to be principally involved in binding and D-lactate selectivity by forming a hydrogen bonding network (Roper, Huyton et al. 2000). Similarly, the omega-loop of VanG of *E. faecalis* is 15 residues-long and extends from Ser248-Ser262. The residue Phe252 which was found to play a major role in D-ser selectivity of VanG was also located within the omega-loop (Meziane-Cherif, Saul et al. 2012). These examples clearly demonstrate that the substrate selectivity of Ddls and their Van homologues is largely determined by the omega-loop. A comparison of the structures of the omega-loops of DdlB of *E. coli* with Ddl6 and Ddl7 revealed that the residue Tyr216 of DdlB of *E. coli* which is responsible for binding and positioning D-ala2, is conserved in Ddl6 and Ddl7 (Tyr250) and the overall structure of the omega-loops of Ddl6/Ddl7 and DdlB of *E. coli* were also very similar (Figure 6-17). This observation supports the *in vitro* results discussed in chapter 5 as to why Ddl6/Ddl7 only catalyze the formation of D-ala-D-ala dipeptide.
Figure 6-17. Multiple sequence alignment of the four proteins in PDB which have highest structural similarity to Ddl6 and Ddl7 (PDB codes: 3I12, 4FU0, 3TQT and 1EHI). The E. coli DdlB (PDB code: 2DLN) and E. faecium VanA (PDB code: 1E4E) were included in the alignment to compare the active/binding sites. PDB code 3I12: Ddl of S. enterica subsp. enterica Serover Typhimurium Str. LT2; 4FU0: D-ala-D-ser ligase (VanG) of E. faecalis; 3TQT: Ddl of Coxiella burnetii; 1EHI: D-ala-D-lac ligase of L. mesenteroides (LmDdl2). The vancomycin resistant proteins are marked with red arrows.
Another important part of this chapter that covers the molecular docking studies. The goal of the ligand-protein docking was twofold: i) to identify if the Trp→Cys substitution alters the binding mode and/or affinity of the Ddl substrates such as D-alanine and ATP as well as Ddl inhibitor; D-cycloserine; ii) to see if the plant flavonoids, quercetin and apigenin, having Ddl-inhibitory activity can bind at the D-alanine and ATP binding sites of Ddl6 as well as L164F and W259C mutants of Ddl6. The Ddl6 and its two single mutants were chosen to simulate the site-directed mutagenesis experiments (Chapter 4). The docking result showed that there is no significant difference in binding energy for D-alanine and D-cycloserine among the three isoforms. However, D-alanine was found to bind with higher affinity compared to D-cycloserine. Quercetin and apigenin belong to a family of natural compounds known as flavonoids. They naturally occur in fruits, vegetables and beverages such as tea. Quercetin is particularly rich in tea, onions, apples, berries and vegetables including broccoli, tomato, lettuce and pepper. It is one of the most studied flavonoids and has been shown to have many beneficial effects to human health (Nijveldt, van Nood et al. 2001, Boots, Haenen et al. 2008). Most of health benefits of quercetin are related to their strong antioxidant properties (Boots, Haenen et al. 2008). It was shown to have anticancer (Scambia, Ranelletti et al. 1990), antibacterial (Ohemeng, Schwender et al. 1993) and antiviral activity (Kim, Woo et al. 1998) and act by binding to multiple targets (Cushnie and Lamb 2005). Apigenin also exert antibacterial (Aljančić, Vajs et al. 1999), anticancer (Caltagirone, Rossi et al. 2000, Sung, Chung et al. 2016) and antiviral activities (Critchfield, Butera et al. 1996). The antibacterial activity of quercetin is partially attributed to their inhibitory activity of DNA gyrase.
(Ohemeng, Schwender et al. 1993, Cushnie and Lamb 2005), whereas the antiviral activity against HIV was attributed to inhibition of HIV-1 integrase (Kim, Woo et al. 1998). Another target of quercetin as well as apigenin is Ddl of bacteria. They are competitive inhibitors of ATP in the Ddl binding sites and are non-competitive for D-alanine (Wu, Kong et al. 2008). The IC$_{50}$ values of quercetin and apigenin were 19.9±1.8 and 163.0±26.2 µM, respectively. The potential of quercetin to exhibit Ddl inhibitory activity has also been demonstrated in a recent molecular docking study (Alves, Froufe et al. 2014). It was shown that quercetin inhibits Ddl of Thermus thermophilus HB8 (PDB code: 2ZDQ) with a predicted Ki value of less than 1 µM (0.35 µM). As these flavonoids were shown to have Ddl inhibitory activity, this study has tried to link these observations to explain the evolutionary advantage of acquiring $ddl$ as a part of integron by the host (Chapter 4). It was hypothesized that if the Ddl encoded by the house-keeping gene of the host is prone to be inhibited by these compounds, the host may need another copy of $ddl$ to overcome this inhibition by sequestration. In this study, this hypothesis has been tested using molecular docking. The results of docking of quercetin and apigenin to Ddl6 showed that quercetin and apigenin can be positioned within both D-alanine and ATP binding pockets of all of the 3D models tested including the Ddl from T. denticola; DdlTd. This suggests that these integron-encoded Ddls have the potential to titrate out these flavonoids by being expressed from an integron, thus allowing the host bacterium to overcome the potential constitutive inhibitory activity, due to the ubiquitous nature of flavonoids. However, the docking results do not accord with the in-vitro study by Wu, Kong et al. (2008) where they have shown that apigenin and quercetin inhibits Ddl of H. pylori by binding to the ATP
binding sites only. It was found that these flavonoids bind with the D-alanine pocket with higher fitness than ATP binding sites.

6.5 Conclusion

The I-TASSER package could generate 3D models of Ddl6 and Ddl7 with high confidence and reasonable quality despite the low levels of target—template sequence identity. The identification of the spatial locations of the active site residues within the 3D models helped to predict why among the two substitutions in between Ddl6 and Ddl7, only the Trp→Cys substitution alters the D-cycloserine resistance phenotype. The models also confirmed the biological function of Ddl6 and Ddl7 as the closest structural homologues of the proteins in protein data bank (PDB) were Ddl of different bacteria. Some structural homologues include Van enzymes, however, the residues within the omega-loop that confer vancomycin resistance were not conserved in both Ddl6 and Ddl7. The docking experiments showed that the plant flavonoids, specifically apigenin and quercetin can dock with reasonable fitness scores with the predicted D-alanine and ATP binding sites, thus there is a potential that these flavonoids would be able to exert a selective pressure to maintain the ddl GC within the first position of an integron GC array as the exposure to flavonoids would be frequent. However, it should be mentioned that these observations are based on 3D models based on low target-template homology, so, the analysis will be repeated with experimentally verified 3D structures of Ddl6 and Ddl7. The crystallization process of Ddl6 and Ddl7 is underway.
Chapter Seven: Final conclusion and future directions
Final conclusion and future directions

The main aim of this study was to investigate the presence of integrons, gene cassettes (GCs) and integron-associated antibiotic resistance genes (ARGs) in the oral metagenomic DNA of healthy humans. Saliva-derived metagenomic DNA of the healthy volunteers from two different countries, the UK and Bangladesh, were used. Using a PCR-based metagenomic approach, it has been shown that the oral bacteria harbour integrons and GCs carrying ARGs as well as genes that encode proteins related to competence and host-adaptation to stress. The putative ARGs detected in the libraries of attC-PCR amplicons included a Vat-family streptogramin A O-acetyltransferase and a bleomycin binding protein. The majority of the integrons and GCs that were detected in this study were novel and some genes were found to have no homologues in the GenBank database. Interestingly, among the GCs recovered from the library of PCR amplicons none of the gene cassettes (except GCs carrying ddl6/ddl7), were found to be common in the UK and Bangladesh, which suggests that difference in life-style, antibiotic use and diets may have an implication in the types of GC carried in the oral cavity.

A major part of the work presented in this thesis focused on characterising the function and resistance phenotype of the two variants of integron-located D-alanine-D-alanine ligase (ddl) identified in the libraries of intl-attC PCR amplicons. The variants were named as ddl6 and ddl7. There were seven reasons for focusing on ddls:

i) They were found to be always located within the first GCs of a reverse integron similar to the integron found in T. denticola
ii) The genes were novel and phylogenetically distant from their closest
ddl homologues in GenBank

iii) Before this study, no ddls were found to be located within integrons.
So, it was interesting to investigate why a house-keeping gene
evolved as an integron GC

iv) The ddl6/ddl7 GCs were detected in both UK and Bangladeshi library
and were found to be present in almost all of the samples analysed
which suggests that a common selective pressure exists in both
countries; selecting these genes in the first position of the integrons

v) The evidence discussed in chapter 4 suggests that the host of the
integron GC carrying ddl6 and ddl7 may have been acquired by HGT
from a strain of T. pedis or vice-versa either by direct transfer or via
an intermediate

vi) The related enzymes of Ddl are known to confer vancomycin
resistance by producing D-ala-D-ser dipeptide (VanC-, VanE-, VanG-, VanL, and VanN-type) or D-ala-D-lac depsipeptide (VanA-, VanB-, VanD-, and VanM-type). So, it was interesting to see if the integron
GC-encoded Ddl6 and Ddl7 also have the capacity to confer
vancomycin resistance by producing alternative products

vii) It was shown previously that overexpression of Ddl conferred high
level of resistance to D-cycloserine in M. smegmatis (Feng and
Barletta 2003). So, it was assumed that overexpression of Ddl6 and
Ddl7 would also confer D-cycloserine resistance by sequestration of
the drug.
The results of the experiments discussed in Chapter 4 showed that expression of both Ddl variants (Ddl6/Ddl7) increased the MIC of D-cycloserine by two to four-fold. Using site-directed mutagenesis, it has been shown that the SNP at c.777 which changes the Trp of Ddl6 to Cys in Ddl7 is responsible for this phenotypic alteration. The possible mechanisms for this were tested using different approaches described in chapter 5 and 6. These compared i) the kinetic parameters of Ddl6 and Ddl7 including the $K_m$, $K_{Dala2}$ and $K_{cat}$; ii) the predicted 3D structures of Ddl6 and Ddl7, iii) binding affinity of D-alanine and D-cycloserine in the wild-type and mutated Ddl6 using molecular docking and iv) the chemical and physical properties of tryptophan and cysteine and predicting how these residues could alter the binding of the ligands. Although this analysis provided some clues that helped us to explain the phenotypic alteration, nothing was conclusive and more experiments, described below, are needed to confirm the mechanism. Determination of inhibitory constant, Ki of D-cycloserine is necessary to see if the two variants exhibit difference in Ki. The crystal structures of Ddl6 and Ddl7 are also needed to be determined to see if the p.259W$\rightarrow$C mutation could alter the structural conformation of the proteins, particularly the structure of the omega-loop which in turn could change the binding of D-alanine, ATP as well as D-cycloserine to the active sites and alter MIC of D-cycloserine. The crystallization of Ddl6 and Ddl7 is under way and we are expecting to get the structures soon.

Although the expression of Ddl6 and Ddl7 conferred D-cycloserine resistance, they could not confer resistance to vancomycin when expressed in *B. subtilis*. The reason for this are discussed in Chapter 5 and 6. It was found that the amino acid residues that are responsible for binding the alternative substrate,
D-ser or D-lac, are not present in Ddl6 and Ddl7. However, the predicted 3D structures of Ddl6 and Ddl7 were found to be very similar to VanG and VanA. This suggests that if the corresponding residue of VanG (Phe252) or VanA (His244) is substituted in Ddl6 and Ddl7 (Figure 7-1), they might gain alternative substrate specificity conferring vancomycin resistance. This hypothesis will be tested very soon in our laboratory where the target residues of Ddl6 and Ddl7 will be changed by site-directed mutagenesis. Only a point mutation in the coding sequence of ddl6/ddl7 is enough to obtain either the Y250F or Y250H mutants (Figure 7-1). The vancomycin resistance phenotype will be tested by expressing the genes in B. subtilis. The formation of alternative dipeptides will also be tested by paper chromatography.

A.

**Ddl6/Ddl7 Y250F**

D-alanine + D-serine + ATP → D-ala-D-ala + ADP + Pi → D-ala-D-ser (??) + ADP + Pi (VanG phenotype?)

B.

**Ddl6/Ddl7 Y250H**

D-alanine + D-lactate+ ATP → D-ala-D-ala + ADP + Pi → D-ala-D-lac (??) + ADP + Pi (VanA phenotype?)

**Figure 7-1.** A) The Y250F mutant of Ddl6/Ddl7 is predicted to gain VanG-type vancomycin resistance phenotype; B) The Y250H mutant of Ddl6/Ddl7 could also led to gaining a VanA-type vancomycin resistance. Only a single base pair substitution is sufficient to obtain these Ddl6/Ddl7 mutants with potential vancomycin resistance phenotype. To obtain the Y250F mutant of Ddl6/Ddl7 (A), the adenine at c.749 of the coding sequence of ddl6/ddl7 will be substituted with thymine which will change the codon TAT (Tyr or Y) to TTT (Phe or F). Similarly, to obtain the Y250H mutant (B), the thymine at c.748 of the coding sequence of ddl6/ddl7 will be substituted with cytosine which will change the codon CAT (Tyr or Y) to CAT (His or H).

As mentioned earlier, one of the intriguing findings of this study was the identification of *ddl* within the first GC of an integron and their presence in
almost all individuals from the UK and Bangladesh. As D-cycloserine is currently used only in the treatment of MDR and XDR-tuberculosis and there is no history of environmental use of D-cycloserine in agriculture for example, so it is very unlikely that D-cycloserine could act as the driving force for evolution of \(ddl\) within integron. In this thesis, a hypothesis was developed to understand this mechanism. It was assumed that the dietary flavonoids or inhibitory small molecules present in our diet that are able to bind Ddl and to which the bacterial host is repeatedly exposed could drive the evolution of an \(ddl\) containing gene cassette. The flavonoids were targeted as some flavonoids (e.g. apigenin, quercetin and catechin) are known to have antibacterial activity (Cowan 1999) and they are rich in fruits and vegetables, beverages (e.g., tea). Moreover, it has been shown that apigenin and quercetin can inhibit the activity of Ddl of \(H. pylori\) (Wu, Kong et al. 2008) by competitively inhibiting the binding of ATP to the ATP-binding sites of \(H. pylori\) Ddl. The results of molecular docking discussed in chapter 6 also showed that these flavonoids can bind to both D-ala and ATP binding sites of Ddl6 and Ddl7. These observations supports our hypothesis that the hosts of integrons carrying \(ddl6/ddl7\) may have acquired these genes and maintained at the first position to ensure their efficient expression so as to titrate out these toxins from their host and to help catalysing the formation of D-ala-D-ala in the presence of inhibited Ddl encoded by the house-keeping gene. A model of this hypothesis is shown in Figure 7-2. This needs to be further tested using the experimentally verified structures of Ddl6 and Ddl7 which is now under way.
Figure 7-2. Diagrammatic representation of the hypothetical model that plant flavonoids could be responsible for selecting the ddls and maintaining within the first GC of the integron of T. denticola. The integron encoded Ddl6/Ddl7 could play a dual role (marked with asterisk): sequestration of flavonoids or other molecules that are able to bind Ddl and catalysing the formation of D-ala-D-ala dipeptide in presence of inhibited Ddl encoded by the house-keeping genes.

This study has a number of potential limitations which need to be mentioned. First, the library of PCR products obtained from the intI-attC PCR as well as the attC-attC PCR could not be analysed until the point of saturation as the focus of the study moved to ddl detected during the screening of GCs. We believe that more interesting and novel GCs could be recovered in the oral cavity if more clones are screened. Second, the study was conducted using two separate pooled samples from the UK and Bangladesh, thus, we could not understand the variation of the type GCs in the oral cavity of the volunteers. Third, the phenotype conferred by the genes were studied in the surrogate host, E. coli or B. subtilis. So, the isolation of the original host is necessary to investigate if the
same phenotype could also be seen in host. Fourth, the predicted 3D structures of the Ddl were determined based on the templates having low sequence identity, so, the hypotheses made based on these structures need to be tested using experimentally determined structures. Future studies should address these limitations.

I believe that the results presented in this thesis will improve our current knowledge about the role of human microbiota as a potential source of mobile ARGs that could potentially disseminate, by HGT, to other bacteria in the oral cavity including the transient pathogens. Moreover, the findings may also help to improve our present understanding about the structure and types of integrons and GCs in the oral cavity and the bacteria which carries them. Furthermore, we have developed a hypothesis that flavonoids present in our diet may act as a major driver for evolution of ddls within integrons and select within the first position of the integron to protect the host by sequestering these toxins.


resistant Citrobacter freundii isolates in a Danish county." Journal of Medical Microbiology 50(7): 636-641.


Pérez-Moreno, M. O., V. Estepa, Y. Sáenz, M. Cortell-Ortolá, I. Fort-Gallifa, J. Ruiz and C. Torres (2012). "Intrahospitalary dissemination of Klebsiella pneumoniae carrying blaDHA-1 and qnrB4 genes within a novel complex class 1 integron." Diagnostic Microbiology and Infectious Disease 73(2): 210-211.


Appendix- I: Multiple sequence alignment of empty gene cassettes

detected in this study
Appendix - II: Translated sequence of *ddl6*

> The residues of Ddl6 that are different from Ddl7 (164L and 259W) are marked. The 250Y is the putative binding site for second D-alanine which if mutated to the corresponding residue of VanA and VanG could lead to gaining a vancomycin resistance phenotype.
Appendix- III: Translated sequence of *ddl7*

The residues of Ddl7 that are different from Ddl6 (164F and 259C) are marked. The 250Y is the putative binding site for second D-alanine which if mutated to the corresponding residue of VanA and VanG could lead to gaining a vancomycin resistance phenotype.

*The residues of Ddl7 that are different from Ddl6 (164F and 259C) are marked. The 250Y is the putative binding site for second D-alanine which if mutated to the corresponding residue of VanA and VanG could lead to gaining a vancomycin resistance phenotype.*
Appendix- IV: Translated sequence of \textit{intI} gene located upstream of \textit{ddl7}*

* The binding site of the TDIF primer is shown in the box.
<table>
<thead>
<tr>
<th>Type of presentation</th>
<th>Title</th>
<th>Event</th>
<th>Year</th>
<th>City and Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oral</td>
<td>Antibiotic resistance genes in the oral cavity of healthy individuals in the UK and Bangladesh.</td>
<td>Commonwealth Scholars Regional Network Conference: Scholars’ Presentation Day Symposium (20-07-2015)</td>
<td>2014</td>
<td>Nottingham, UK</td>
</tr>
<tr>
<td>Oral</td>
<td>Discovery of antibiotic resistance encoding integrons from the human oral cavity.</td>
<td>UK Plasmid and Mobile Genetic Element Workshop</td>
<td>2015</td>
<td>Birmingham, UK</td>
</tr>
<tr>
<td>Oral</td>
<td>A Novel D-Alanine-D-Alanine Ligase Gene From Human Oral Metagenomic DNA Confers High-Level Resistance.</td>
<td>FEMS Congress</td>
<td>2015</td>
<td>Maastricht, Netherlands</td>
</tr>
<tr>
<td>Poster</td>
<td>Integron Gene Cassettes Encoding D-alanine-D-alanine ligases Confer High-level Resistance to D-cycloserine.</td>
<td>Annual Conference of Society of General Microbiology</td>
<td>2015</td>
<td>Birmingham, UK</td>
</tr>
<tr>
<td>Poster</td>
<td>Investigation of integrons in human saliva metagenomic DNA.</td>
<td>Annual Conference of Society of General Microbiology</td>
<td>2015</td>
<td>Birmingham, UK</td>
</tr>
<tr>
<td>Poster</td>
<td>Characterization of novel genes conferring resistance to D-cycloserine detected in oral human oral metagenomic DNA.</td>
<td>BSAC Antimicrobial Resistance Workshop</td>
<td>2015</td>
<td>Birmingham, UK</td>
</tr>
<tr>
<td>Oral</td>
<td>Characterisation of a novel integron associated gene cassette detected in the human oral metagenomic DNA.</td>
<td>Lunch-time Research Presentation at Eastman Centre for Postgraduate Research</td>
<td>2016</td>
<td>London, UK</td>
</tr>
<tr>
<td>Poster</td>
<td>Detection and Characterization of a Novel Integron Gene Cassette in the Human Oral Metagenomic DNA that Confers Resistance to D-Cycloserine.</td>
<td>ASM Microbe 2016</td>
<td>2016</td>
<td>Boston, USA</td>
</tr>
<tr>
<td>Oral</td>
<td>A single base pair mutation alters the susceptibility of the host to D-cycloserine.</td>
<td>EDI-Research Away Day-2016</td>
<td>2016</td>
<td>London, UK</td>
</tr>
</tbody>
</table>
Appendix- VI: List of awards, prizes and travel grants

1. FEMS-ASM Makela-Cassel Award for Early Career Scientists (June 2016)
2. ASM Student and Postdoctoral Travel Award (June 2016)
3. Postgraduate Research Competition Prize (4M Thesis) at the EDI-Research Away Day (November 2016)
4. SGM Travel Grant to Join the Annual Conference of Society of General Microbiology, 30 April-2 May, 2015, ICC Birmingham, UK
5. Funded place to join BSAC Antimicrobial Resistance Workshop-2013, Birmingham, UK
6. Funded place to join BSAC Antimicrobial Resistance Workshop-2014, Birmingham, UK
7. Funded place to join BSAC Antimicrobial Resistance Workshop-2015, Birmingham, UK
Appendix- VII: Research articles published as a result of the work presented in the thesis
