



University of
Nottingham

UK | CHINA | MALAYSIA

An Investigation of *Helicobacter pylori* Integrative and Conjugative Elements

Ibikunle O Idowu

Thesis submitted to the University of Nottingham for
the degree of Doctor of Philosophy

September, 2022

Declaration

Unless otherwise acknowledged the work presented in this thesis is my own; no part has been published for another degree at the University of Nottingham or any other institute of learning.

Ibikunle Idowu.

Dedication

This work is dedicated to Almighty God, to whom by His mercy and grace had brought me thus far and to my lovely mum, Mrs Alice Idowu for her support and prayers.

Abstract

Several key *Helicobacter pylori* virulence factors are encoded within three major genomic islands namely, the *cag* pathogenicity island (*cagPAI*), *tfs3* and *tfs4*. Although the *cagPAI* is fairly well characterised, understanding the functional role played by *tfs3* and *tfs4* in disease outcomes remains unknown. Also, the possibility of functional interchangeability between the components of both *tfs3* and *tfs4* subtypes by interacting with one another to form a functional type IV secretion system (T4SS) remains to be investigated.

The main aim of this study was to investigate interactions between proteins encoded by *tfs3*, *tfs4* and *tfs4* variant types that are important for T4SS assembly and function. A second aim of this study was to evaluate approaches for virulence typing, comparing genomics with PCR typing using an available rich dataset including *cagA* and *cagE* PCR typing data, and phenotypic data. Additionally, the study aimed to explore the association of *tfs* T4SS genes in *H. pylori* strains isolated from patients with their disease status.

Initially, the yeast two-hybrid assay (Y2H) was used to establish a network of interactions between *tfs3*, *tfs4* and *tfs4* variant encoded proteins comprising the T4SS structure. Other interactions involving proteins that are not part of the T4SS assembly, *tfs4* VirD2 and orphan VirB11 together with proteins of *tfs3*, *tfs4* and *tfs4* variant were also investigated using Y2H assay to assess whether the *tfs* T4SSs might be functional for export/transfer of any *tfs* integrative and conjugative elements (ICEs). The ABRicate tool was configured with custom scripts of both *tfs3* and *tfs4* genes from reference *H. pylori* strains. Then, the association between disease status and *tfs3* and *tfs4* clusters of genes in genomic data from clinical isolates was examined using the ABRicate bioinformatics tool to identify virulence genes. Prior to using ABRicate to search for *tfs3* and *tfs4* genes in isolates, its efficacy was assessed by comparing its output on the presence of *cagA* and *cagE* genes with PCR genotype and phenotypic data available from clinical isolates from patients.

The Y2H analyses identified several novel interactions between paired proteins of *tfs3*, *tfs4* and *tfs4* variant type (*tfs4*VirB11/*tfs4*VirB11v, *tfs4*VirB11/*tfs4*VirB4v, *tfs4*VirB11v/*tfs4*VirB4, *tfs4*VirB4/*tfs4*VirB4v, *tfs4*VirB4/*tfs4*VirB11v, *tfs3*VirB11/*tfs4*VirB8v, *tfs3*VirB11/*tfs4*VirB4v) suggesting the possibility of functional complementation between *tfs* ICE proteins in the assembly of a T4SS. Other interactions involving *tfs4* VirD2, identified novel interactions with components of *tfs3* T4SS (*tfs4*VirD2/*tfs3*VirB4), which could suggest some functional flexibility of

the *tfs* T4SSs for inter ICE transfer especially when both ICEs are present within a single strain.

The ABRicate tool detected the *cagA* gene in the genome of 17 out of 18 clinical isolates that were previously found to be *cagA*-positive by PCR. Also, the tool found the *cagE* gene in the genome of all 15 isolates that were previously reported to be *cagE*-positive by PCR. Additionally, PCR typing, and genomic data were in agreement for 18 *dupA*-negative isolates and for 6 *dupA*-positive isolates analysed.

The presence of *dupA*, a *virB4* homolog in *tfs4* and some *tfs4* *vir* genes (*virB2*, *virB4*, *virB8* and *virB10*) that form the “*dupA* cluster”, in clinical isolates was compared with disease status in the patients infected with them. Although the ABRicate tool detected a *dupA* cluster in four strains isolated from two patients, who presented with duodenal ulcer and gastric ulcer disease, genome sequencing of isolates from many more patients would be required to establish associations between disease status and cluster of *tfs4* genes in isolates.

Conclusions

The Y2H results provided an insight of system specificity in protein-protein interactions of *tfs3* and *tfs4* and, variable cross-complementation between *tfs3* and *tfs4* VirB proteins to form a functional T4SS. Also, the comparison between genomic and PCR typing data in this study, provides some support in the reliability of both PCR typing and genomics analyses in the detection of virulence factors in *H. pylori*. Lastly, genomic analysis could not be used to establish association between disease status and *tfs* T4SS genes in isolates due to the small sample size but it indicated the feasibility for using genomics in future studies. Therefore, genomics analysis approaches can readily be used to investigate the genes present in *tfs* ICEs.

Acknowledgements

My supervisors, Dr Rob Delahay, Dr Karen Robinson and Dr Jane Grove for their proper guidance and support.

All the members of the *Helicobacter* research group for their support.

Presentations

Idowu, I., Delahay, R.

Functional characterisation of the *Helicobacter pylori* *tfs* Integrative and Conjugative Elements

School of Medicine Sue Watson Oral Presentation Event, November 2018.

Poster Presentation

Idowu, I., Delahay, R.

Functional characterisation of the *Helicobacter pylori* *tfs* Integrative and Conjugative Elements

E-poster presentation: Sue Watson Oral Presentation Event, School of Medicine. November 2018.

List of Abbreviations

ATP	Adenosine Triphosphate
CagA	Cytotoxin-associated Gene A
<i>cagPAI</i>	Cytotoxin-associated Gene Pathogenicity Island
CI	Circular Intermediate
dATP	Deoxyadenine triphosphate
DBD	DNA-Binding Domain
dCTP	Deoxycytidine triphosphate
dGTP	Deoxyguanosine triphosphate
DNA	Deoxyribonucleic Acid
dTTP	Deoxythymidine triphosphate
DupA	Duodenal ulcer promoting gene A
EGF	Epidermal Growth Factor
FCS	Foetal Calf Serum
GEC	Gastric Epithelial Cell
GSP	General Secretory Pathway
HF	High Fidelity
ICEs	Integrative and Conjugative Elements
IL	Interleukin
IM	Inner membrane
IMP	Inner membrane protein
Kb	Kilobase
KDa	Kilo Daltons
LB	Luria Bertani
LFA-1	Lymphocyte function associated antigen – 1

LJ	Left Junction
MALT	Mucosal-Associated Lymphoid Tissue
MGEs	Mobile Genetic Elements
MU	Methionine Uracil
MUA	Methionine Uracil Adenine
MUHA	Methionine Uracil Histidine Adenine
NADPH	Nicotinamide Adenine Dinucleotide Phosphate Hydrogen
NF- κ B	Nuclear factor kappa light chain enhancer of activated B cells
NOD1	Nucleotide-binding oligomerisation domain 1
NSAID	Non-steroidal anti-inflammatory drug
OD	Optical density
Oip	Outer inflammatory protein
OM	Outer membrane
OMP	Outer membrane protein
<i>oriT</i>	Origin of Transfer
PBS	Phosphate buffer saline
PCR	Polymerase Chain Reaction
PEG	Polyethylene glycol
PI	Protease Inhibitor
PMSS1	Parent Mouse Sydney Strain 1
PPI	Proton Pump Inhibitor
PUD	Peptic Ulcer Disease
PZ	Plasticity Zone
RJ	Right Junction

ROS	Reactive Oxygen Species
SS1	Sydney strain 1
ssDNA	single-stranded Deoxyribonucleic Acid
T-DNA	Transfer-Deoxyribonucleic Acid
T4CP	Type 4 Coupling Protein
T4SS	Type 4 Secretion System
Tfs3	Type 4 secretion system 3
Tfs4	Type 4 secretion system 4
TMP	Tyrosine phosphorylation motifs
TNF- α	Tumour Necrosis Factor alpha
tRNA	transfer Ribonucleic Acid
UV	Ultraviolet
VacA	Vacuolating Cytotoxin A
Vir	Virulence
YMM	Yeast Minimal Medium
YPD	Yeast Peptone Dextrose

Table of Contents

1.0	Introduction	2
1.1	<i>Helicobacter pylori</i>	2
1.1.1	The Discovery.....	2
1.1.2	Characteristics of <i>H. pylori</i>	3
1.1.2	Epidemiology and transmission of <i>H. pylori</i>	5
1.1.3	Disease Spectrum of <i>H. pylori</i>	10
1.1.3.1	Acute and chronic gastritis	11
1.1.3.2	Gastric Cancer.....	13
1.1.3.2.1	Gastric mucosa-associated lymphoid tissue (MALT) lymphoma.....	14
1.1.3.3	Peptic Ulcer Disease (PUD)	15
1.1.4	<i>H. pylori</i> diagnosis and treatment	16
1.1.4.1	Diagnosis	16
1.1.4.2	Methods for diagnosis of <i>H. pylori</i> infection	19
1.1.4.3	Treatment.....	21
1.2	Factors that contribute to disease spectrum of <i>H. pylori</i>	21
1.2.1	<i>H. pylori</i> virulence factors	21
1.2.1.1	Vacuolating cytotoxin A (VacA).....	22
1.2.1.2	Cag pathogenicity island (cagPAI)	25
1.2.1.3	Host genetic factors.....	29
1.2.1.4	Environmental factors	30
1.2.1.4.1	<i>H. pylori</i> and nutrition	30
1.3	Type IV Secretion Systems (T4SSs).....	36
1.3.1	Structural components of T4SSs	37
1.3.2	The components of Type IV Secretion Systems	38
1.3.3	Biochemical and functional interactions of VirB proteins in T4SSs. 43	
1.3.4	Mechanism of substrate translocation in T4SS	44
1.4	<i>H. pylori</i> T4SSs	45
1.4.1	Conjugative T4SSs.....	47
1.4.2	The Effector T4SS	48
1.4.3	ComB T4SS.....	49
1.5	Integrative and Conjugative Elements (ICEs).....	50
1.5.1	The life cycle of the ICEs	52
1.5.2	ICE integration and excision	53
1.5.3	Identification of Tfs ICEs and variant subtype of Tfs4 ICE ...	54

1.5.3.1	<i>tfs3</i> and <i>tfs4</i> ICEs	54
1.5.4	Geographical distribution of Tfs ICEs.....	58
1.5.5	Modular classification of Tfs ICEs.....	59
1.5.6	Modules and hybrid functions of Tfs ICEs	61
1.5.7	Disease association of Tfs ICEs	63
1.6	<i>Rationale, Hypotheses and Aims</i>	64
CHAPTER TWO		68
2.0	Introduction	72
2.1	Rationale	75
2.2	Materials and Methods.....	77
2.2.1	Yeast strains, bacterial strains, growth media and conditions 77	
2.2.2	Oligonucleotides and bioinformatics resources.....	79
2.2.3	Polymerase Chain Reaction – amplification of genomic DNA 83	
2.2.4	Agarose Gel electrophoresis	84
2.2.5	Purification of PCR products	84
2.2.6	Extraction of Plasmid DNA.....	85
2.2.7	Restriction Digest of Purified plasmids and PCR products..	85
2.2.8	Ligation of DNA fragments.....	85
2.2.9	Preparation of <i>E. coli</i> competent cells and transformation ..	86
2.2.10	Screening and Identification of transformants.....	87
2.2.10.1	Colony PCR.....	87
2.2.10.2	Construct Screening	88
2.2.11	The Yeast-two hybrid Assay	88
2.2.11.1	Lithium acetate transformation of <i>S. cerevisiae</i>	88
2.2.11.2	Beta-galactosidase assay using ONPG substrate	89
2.2.12	Protein analysis and purification.....	91
2.2.12.1	Construction of pBAD:: <i>tfs3virB11</i> , pBAD:: <i>tfs4virB11</i> and pBAD:: <i>gB11</i>	91
2.2.12.2	Expression, solubility and purification of VirB proteins	91
2.2.12.3	Sodium Dodecyl Sulphate- Polyacrylamide Gel Electrophoresis (SDS-PAGE)	92
2.2.12.4	Western Blot analysis	93
2.2.12.5	Coomassie blue staining	94
2.2.12.6	Small-scale purification of VirB proteins	94
2.2.12.6.1	Protein purification optimisation	95
2.2.12.7	Large-scale purification of VirB proteins.....	96
2.3	Results	98

2.3.1	Bioinformatics: percentage identity and similarity in VirB11 proteins of <i>tfs</i> ICEs.....	98
2.3.2	Polymerase Chain reaction and Gel Electrophoresis.....	100
2.3.3.	Colony PCR of Amplicons	101
2.3.4	Yeast Two Hybrid Cross-Complementation Interactions	104
2.3.5	Validation of VirB8 interactions	108
2.3.7	Interactions of <i>tfs4</i> ATPases proteins with <i>tfs4</i> variant proteins important for T4SS assembly and function.	111
2.3.8	Interactions of <i>tfs3</i> ATPases proteins with <i>tfs4</i> variant proteins important for T4SS assembly and function	114
2.3.9	Heterologous interactions of <i>tfs4</i> VirB11v with other T4SS VirB proteins	115
2.3.10	Interactions between the <i>tfs4</i> variant proteins to form hybrid type.	117
2.3.11	Investigation of VirD2 interactions as a secretory substrate	118
2.3.12	Screening of orphan genomic gene (<i>gB11</i>) interaction with <i>tfs</i> ICE VirB proteins.....	121
2.3.11	Protein expression and purification study.	125
2.3.11.1	Small-scale expression.....	125
2.3.11.2	Small-scale purification assay	126
2.3.11.3	Large-scale purification assay.....	128
2.3.12	<i>In silico</i> bioinformatics analysis of Novel protein-protein interactions	130
2.4	Discussion	132
2.4.1	Reproducibility of previous VirB8 interactions	133
2.4.2	ATPases interactions for system assembly and secretion of <i>tfs</i> T4SSs	134
2.4.3	Interactions of orphan VirB11 (<i>gB11</i>).....	138
2.5	Study strength and weaknesses.....	141
2.6	Conclusions.....	143
CHAPTER THREE		144
3.0	Introduction	145
3.1	Rationale and Aims	147
3.2	Materials and Methods.....	150
3.2.1	Collection of Clinical Isolates	150
3.2.2	Clinical Isolates selection criteria.....	151
3.2.3	PCR typing for <i>cagA</i> and <i>cagE</i> genes.....	157
3.2.4	CagA IgG ELISA	158
3.2.5	Interleukin 8 induction as a marker for presence of <i>cagE</i> ..	158

3.2.6	Sample preparation.....	159
3.2.7	Genomic Extraction and Whole genome sequencing	159
3.2.7.1	Quality control of sequencing reads	160
3.2.8	Sequence retrieval, ABRicate set-up, and genomic analysis of clinical isolates.....	161
3.2.8	Analyses of Open Reading Frames and motifs upstream in selected clinical isolates.....	165
3.2.9	Determination of percentage accuracy, sensitivity, and specificity	166
3.3	Results	167
3.3.1	Detection of <i>cagA</i> genes in Clinical Isolates	167
3.3.1.1	Correctness of open reading frames.....	170
3.3.2	Detection of multiple copies of <i>cagA</i> in the genome of the PMSS1 strain.....	172
3.3.3	Detection of <i>cagE</i> genes in Clinical Isolates	173
3.3.4	Analysis of <i>cagE</i> gene sequences in selected clinical isolates, detected by ABRicate software.	175
3.3.6	Determination of percentage accuracy, sensitivity, and specificity	176
3.3.6.1	Percentage Accuracy, Sensitivity and Specificity of <i>cagA</i> findings, using ABRicate software, PCR typing and serological data.	176
3.4	Discussion	176
3.4.1	Comparison of <i>cagA</i> PCR typing and genomic data (% accuracy, sensitivity, and specificity)	178
3.4.2	Comparison of <i>cagA</i> PCR typing and <i>cagA</i> detection in genomes	179
3.4.3	Comparison of CagA serology and genomic data	182
3.5	General applicability of ABRicate - Pros and Cons	183
3.6	Conclusion.....	185
CHAPTER FOUR		186
4.0	Introduction	187
4.0.1	Analysis of <i>tfs vir</i> genes using bioinformatics to predict virulence effect.	187
4.1	Rationale	190
4.2	Materials and Methods.....	191
4.2.1	Analysis of <i>dupA</i> in clinical isolates	191
4.2.3	PCR detection of <i>dupA</i>	192
4.2.4	PCR typing for some selected <i>tfs4</i> genes.....	193
4.2.5	Analysis of <i>tfs4</i> homologous genes in isolates using ICEfinder.....	194

4.3	Results	195
4.3.1	Detection of <i>dupA</i> in clinical isolates.....	195
4.3.2	Identification of <i>tfs4</i> genes in clinical isolates using Shi470 as reference strain.....	197
4.3.3	Pairwise Alignment of <i>vir</i> genes detected in isolates with homologues genes in Shi470	201
4.3.4	Identification of <i>tfs4</i> genes in clinical isolates using P12 as the reference strain	203
4.3.5	Identification and comparison of <i>tfs4</i> genes using P12 as reference strain with ICEfinder tool	205
4.3.6	Comparison of PCR and genome data in selected isolates - <i>tfs4</i> genes.....	207
4.4	Discussion	209
4.4.1	<i>dupA</i> status in genomes of isolates agreed with PCR typing data	209
4.4.2	Detection of <i>tfs</i> ICE genes and association with disease status in clinical isolates.	210
4.4.2.1	Detection of <i>tfs4 vir</i> genes and association with disease status in clinical isolates	210
4.4.2.2	Comparison of <i>tfs4 vir</i> genes using PCR typing and genome methods	212
4.4.2.3	Detection of <i>tfs3 vir</i> genes and association with disease status in clinical isolates	212
4.4.3	XercD: tyrosine recombinases in clinical isolates.....	213
4.4.4.	Challenges involved in sequencing methods of <i>Helicobacter pylori</i>	214
4.4.5	Comparison of ABRicate with a second algorithm (ICEfinder) used to detect <i>tfs vir</i> genes in isolates.....	216
4.4.6	Study conclusion, limitations, and future work.	217
	CHAPTER FIVE	219
	GENERAL DISCUSSION.....	219
5.1	Important and Novel Findings.....	222
5.2	Critique of Experimental Design.....	224
	REFERENCES	227
	APPENDIX	266

List of Figures

Fig 1.1: Electron micrograph of <i>H. pylori</i> .	4
Fig 1.2: Microscopic images of coccoid forms of <i>H. pylori</i> .	4
Fig 1.3: <i>H. pylori</i> disease spectrum showing stomach infection regions in chronic infection states.	11
Fig 1.4: Schematic representation of the allelic diversity that exist in three regions of <i>vacA</i> gene and their associations with disease.	23
Fig 1.5: Vacuolation effect of VacA on HeLa cells <i>in vitro</i> .	24
Fig. 1.6: Cluster of genes that form the <i>cag</i> T4SS in <i>H. pylori</i> strain P12.	26
Fig 1.7: Structural model for <i>cagPAI</i> T4SS in <i>H. pylori</i> .	26
Fig 1.8: Schematic representation of type IV secretion system assembly.	38
Fig 1.9: Schematic representation of interaction network of VirB/D homologues protein from <i>Agrobacterium tumefaciens</i> .	44
Fig 1.10: Schematic arrangement of genes encoded within the four chromosomally T4SSs gene clusters in <i>H. pylori</i> .	46
Fig 1.11: Global pairwise alignment of distinct <i>tfs4</i> ICE and modular representation.	61
Fig. 2.1: Schematic representation of interaction network of VirB/D homologues protein from <i>A. tumefaciens</i> .	75
Fig 2.2: Multiple protein sequence alignment showing conservation of motifs in <i>tfs</i> VirB11 proteins.	99
Fig 2.3. DNA amplicons electrophoresed by agarose gel 0.8% at 80v for 30 minutes.	100
Fig 2.4: Colony PCR of ligated transformants electrophoresed using 0.8% agarose at 80v for 30 minutes.	102
Fig. 2.5: Schematic summary of protein-protein interactions detected using the Y2H assay	124
Fig 2.6: Showing 10% SDS-page gel electrophoresed at 180v for 45 mins.	126
Fig 2.7: Showing 10% SDS-page gel electrophoresed at 180v for 45 mins.	127

Fig 2.8: Showing 10% SDS-page gel electrophoresed at 180v for 45 mins.....	128
Fig 2.9: 10% SDS-page gel electrophoresed at 180v for 45 mins.	129
Fig 2.10: Positions of secondary structure predictions in novel interaction partners.....	131
Figure 3.1: Multiple sequence alignment of predicted CagA amino acid sequences from selected genomes of <i>H. pylori</i> clinical isolates and reference strain 26695.	171
Figure 3.2: The aligned <i>cagA</i> gene of selected isolates and, -10 and -35 positions upstream from transcription start site (ATG).	172
Figure 3.3: Multiple alignment of four copies of CagA amino acid sequences detected using ABRicate software in <i>Helicobacter pylori</i> strain PMSS1. The reference strain 26695 is also shown.....	173
Figure 4.1: Schematic representation of T4SS VirB/D proteins assembly in the membrane.	189

List of Tables

Table 1.1: Recommendations for individual and population-based testing for <i>H. pylori</i> infections.	18
Table 1.2: Tests for <i>H. pylori</i> infection.....	20
Table 1.3: Specific features typical to ICEs and conjugative plasmids	52
Table 2.1: Bacterial and Yeast strains used in this study.....	78
Table 2.2: Plasmids used in this study.	79
Table 2.3: Oligonucleotide primer sequences used in this study.	81
Table 2.4: Bioinformatics resources used in this study.	82
Table 2.5: Tfs3 and Tfs4 constructs used in this study for protein-protein interaction in the yeast two hybrid assay.....	103
Table 2.6: Preliminary screening for self-activation screening of <i>tfs4 virB</i> constructs.....	106
Table 2.7: Preliminary screening for self-activation of <i>tfs3 virB</i> constructs.	108
Table 2.8: Screening for validation of previous VirB interactions.	110
Table 2.9: Summary of homodimerisation interactions between VirB11 proteins.....	111
Table 2.10: Screening for Tfs4 ATPases proteins with Tfs4 variant proteins.....	113
Table 2.11: Screening for cross-complementation between Tfs3 and Tfs4 variant proteins.....	115
Table 2.12: Screening for cross-complementation between <i>tfs4</i> VirB11v and other tfs component proteins.....	116
Table 2.13: Screening for cross-complementation between <i>tfs4</i> variant proteins.....	117
Table 2.14: Screening for <i>tfs4</i> VirD2 interaction with other <i>tfs4</i> VirB and <i>tfs4</i> variant VirB proteins.	119
Table 2.15: Screening for <i>tfs4</i> VirD2 interaction with <i>tfs3</i> VirB proteins	120
Table 2.16: Screening for gB11 interactions with tfs3, tfs4 and tfs4 variant proteins.....	122

Table 3.1: Demographic characteristics of patients who donated gastric biopsies.	153
Table 3.2: Patient disease status	155
Table 3.3: Comparative analyses of <i>cagA</i> in sequenced genomes of clinical isolates, PCR typing and serology data.....	169
Table 3.4: IL-8 secretion by AGS cells in response to co-culture with <i>H. pylori</i> clinical isolates, as a marker for a functional <i>cag</i> type IV secretion system (T4SS).....	174
Table 4.1: cycling conditions for PCR with <i>dupA</i> _RI primers	193
Table 4.2: Comparison of <i>dupA</i> status in genome sequences and PCR genotyping data of <i>H. pylori</i> clinical isolates.....	196
Table 4.3: Identification of <i>tfs4</i> homologues (using Shi470 reference sequence) in clinical isolates.....	199
Table 4.4: Percent identity from pairwise sequence comparison of T4SS nucleotide sequences – Shi470 <i>tfs4</i> vir genes against the sequences of vir genes from clinical isolates	202
Table 4.5: Identification of <i>tfs4</i> homologues in clinical isolates, using strain P12 as the reference sequence.....	204
Table 4.6: Identification and comparison of <i>tfs4</i> homologues using ABRicate and ICEfinder methods.....	206
Table 4.7A and 4.7B: Comparison of PCR typing data of <i>tfs4</i> genes with genomic data on <i>tfs4</i> genes from reference strains Shi470 (A) and P12 (B), including patient gastro-duodenal disease status.	208

CHAPTER ONE

General Introduction

1.0 Introduction

1.1 *Helicobacter pylori*

1.1.1 The Discovery

Helicobacter pylori is a spiral, Gram-negative bacterium believed to have inhabited the stomachs of humans for millennia and co-evolved with human populations. *H. pylori* was first discovered in 1982 by Marshall and Warren following earlier reports of spiral shaped bacteria in the stomachs of patients with peptic ulcer disease (Kusters et al., 2006). They demonstrated positive culture and eradicated the bacteria with amoxicillin and bismuth. Self-ingestion of the bacteria by Marshall made it clear that the spiral-shaped organisms cause chronic gastritis (Yuan et al., 2005) which, due to underlying persistent inflammation, was presented as a likely cause for more serious outcomes such as peptic ulcer disease and distal gastric adenocarcinomas (Kusters et al., 2006, Ernst and Gold, 2000, Leonardo et al., 2014). The discovery of *H. pylori* by Marshall and Warren led to them being awarded the Nobel Prize in Medicine in 2005.

In the years following the discovery of *H. pylori*, several studies characterised the organism (Anyanwu and Jukes, 1991). The bacterium was originally named *Campylobacter pylori* and was later renamed under the genus *H. pylori* based on its distinct features such as growth requirements, cellular fatty acid, respiratory quinones analyses, enzyme profiles and genetic composition. The new name originated from Latin and means 'spiral rod of the lower part of the stomach' (Velázquez and Feirtag, 1999).

1.1.2 Characteristics of *H. pylori*

The Gram-negative bacterium is microaerophilic and spiral-shaped in nature (Ando et al., 2004, Robinson and Atherton, 2021) (Fig 1.1), but when in culture, *H. pylori* can be seen in two different forms: bacillary and coccoid forms (Baj et al., 2021, Robinson and Atherton, 2021). The cell structure of the bacillary form is helical in shape, and it transforms to a distinctly spherical coccoid form after many days in culture or with antibiotic treatment (Sharndama and Mba, 2022) (Fig 1.2). The bacillary forms have several unipolar flagella and are highly motile. The bacillary type is reported to be the predominant viable, culturable form. The unique ability of *H. pylori* to change from spiral to coccoid shape is one of the major mechanisms that enables the bacterium to survive in the stomach of the host (Sharndama and Mba, 2022). However, the biological importance of the coccoid form is yet to be fully understood. To date, the coccoid form of the bacterium continues to be difficult to eradicate with treatment (Sharndama and Mba, 2022).



Fig 1.1: Electron micrograph of *H. pylori*.

Several flagella at one pole allow motility of the bacterium in the gastric mucosa. Reproduced from Robinson and Atherton, 2021.

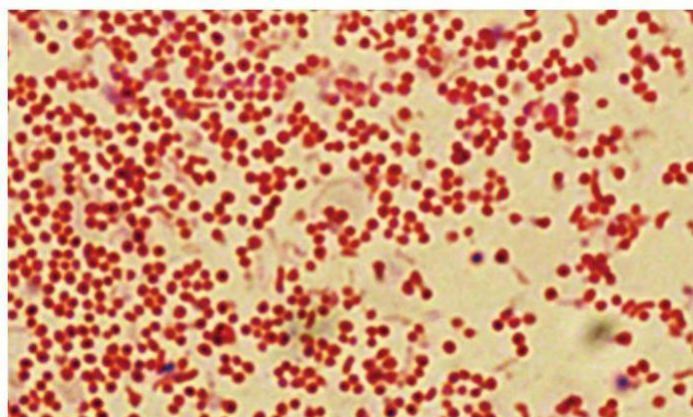


Fig 1.2: Microscopic images of coccoid forms of *H. pylori*.

Reproduced from Reshetnyak and Reshetnyak, 2017.

Successful colonization of *H. pylori* in the mucosal layer of the stomach is achieved by producing enzymes such as phospholipase, catalase, urease and thioredoxin reductase (Ernst and Gold, 2000, Robinson and

Atherton, 2021). Urease is abundantly produced by *H. pylori* and contributes to the bacterium's survival in the gastric environment of the stomach by hydrolysing urea in the gastric juice to ammonia and carbon dioxide (Hunt, 1996, Robinson and Atherton, 2021). This process of hydrolysis has been reported to release hydrogen ions from the cytoplasm which *H. pylori* uses to regulate gastric pH and create an acid-neutral (alkaline) environment for itself, so that it survives the acidic medium of the stomach. *H. pylori* also produces other enzymes including phospholipase A2 and C, which break down phospholipids and gastric mucus secreted by the gastric epithelial cells, to further aid *H. pylori* colonisation (Hunt, 1996). After colonisation, *H. pylori* produces adhesins such as BabA and SabA to mediate its attachment to gastric epithelial cells (Cover and Blaser, 2009, Robinson and Atherton, 2021).

1.1.2 Epidemiology and transmission of *H. pylori*

H. pylori primarily resides in humans (Blaser, 1999, Leonardo et al., 2014) although other known carriers of the bacterium include houseflies, domestic cats and water (Ernst and Gold, 2000, Everhart, 2000). The bacterium infects more than half of the world's population, >4.4 billion individuals, and its prevalence may be >80% among middle-aged adults in developing countries (Miller and Williams, 2021). Epidemiological studies have reported variable prevalence of *H. pylori* infection between developing and developed countries (Perez-Perez et al., 2004, Robinson and Atherton, 2021) as a consequence of socioeconomic differences.

High *H. pylori* infection rates have been reported in Africa (79.1%), Latin America and the Caribbean (63.4%), and Asia (54.7%) compared to low rates in Oceania (24.4%) and Northern America (37.1%) (Miller and Williams, 2021). However, the *H. pylori* infection rate across different regions is not strongly associated with gastric cancer development (Miller and Williams, 2021). The reason for the manifestation of *H. pylori* infection in some individuals and not others still remains unclear. This prompted a proposal that the presence of *H. pylori* in the stomach is not enough for disease manifestation (Alam et al., 2020). The lack of correlation between disease development (gastric cancer) and infection rate has led to the creation of different “Enigmas”.

In Africa, the majority of those infected with *H. pylori* remain asymptomatic with only approximately 15–20% of infected individuals going on to develop peptic ulcer in the long-term (Alam et al., 2020), compared to in Asia, Europe, and the Americas (Smith et al., 2021). This is termed the “African Enigma”. The “African Enigma” phenomenon was first described by Holcombe as a situation of high prevalence of *H. pylori* infection in the continent of Africa but without correlation with severe pathology outcomes such as gastric cancer (Smith et al., 2019). This led to the proposal that *H. pylori* infection in Africa has a different pattern to that of the Western populations, in terms of genetic make-up, environmental factors, age, sex, and dietary factors (Smith et al., 2019).

Despite the availability of population studies of people presenting with disease related to *H. pylori* infection, only a few studies have reported that gastritis is the most common health problem linked with *H. pylori* infection in Africa, followed by peptic ulcer (Smith et al., 2019). For example, a study by Boukhris and colleagues in Morocco (Boukhris et al., 2012) reported an association between gastritis and *H. pylori* infection and with peptic ulcer, but no correlation between infection and gastric cancer. Another study in Ethiopia conducted by Alebie and Kaba (Alebie and Kaba, 2016) reported an increase of *H. pylori* infection (71.0%) among student populations who presented with gastritis. In central African countries, data mining suggests *H. pylori* infection is scarce especially among patients with gastroduodenal diseases (Smith et al., 2019). However, studies from Cameroon have established a strong association between *H. pylori* infection and gastritis. A study by Ankouane and colleagues (Ankouane et al., 2015) found that 71.2% of patients who presented with atrophic gastritis were positive for *H. pylori* infection. The findings were supported by a study by Ebule and colleagues from the same country (Ebule et al., 2017), where they reported *H. pylori* infection in 72.8% of patients who presented with superficial gastritis. In West African countries such as Ghana and Nigeria, the most common reported health problem associated with *H. pylori* infection is gastritis. A study in Nigeria by Harrison and colleagues (Harrison et al., 2017) reported high prevalence of *H. pylori* infection in patients with chronic gastritis, but no such correlation was reported with other disease outcomes. Conversely, a study by Darko (Darko et al.,

2015) found gastritis and duodenal ulcer to be associated with *H. pylori* infection among patients in Ghanaian population.

In comparison with the “African Enigma” theory, the “Asian Enigma” is more controversial because the actual distribution and association between *H. pylori* infection and disease development is unclear in both developing Asian countries, such as Thailand, Malaysia, Pakistan, Bangladesh and India, and in developed Asian countries, such as Japan, China and South Korea (Zaidi, 2016). *H. pylori* infection is more common in developing Asian countries while prevalence has been reported to be low in developed Asian countries. A high incidence of gastric cancer associated with *H. pylori* infection has been reported in developed Asian countries with low infection rates, whereas gastric cancer incidence is low in developing Asian countries with high *H. pylori* infection levels (Nguyen et al., 2021). For example, Japan has the highest prevalence of gastric cancer while the incidence of gastric cancer has been reported to be low in developing Asian countries such as India. These contradictory findings have been attributed to *H. pylori* genetic types in these regions. Different strains of *H. pylori* have been described and although not all strains are pathogenic, its genetic diversity has been associated with clinical outcomes in individuals (Zaidi, 2016).

The prevalence of *H. pylori* infection also differs within countries (Smith et al., 2021) due to differences in age, ethnicity, diet and hygiene

standards (Alam et al., 2020). In developing countries, such as those in South and East Europe, Southern America and Asia, the infection rate is between 80-90% of the whole population (Leonardo et al., 2014). While in developed countries, a lower rate of 40-50% among adults above 50 years and 5% among children below 10 years is more common (Hooi et al., 2017).

H. pylori infection is usually acquired during childhood, with hosts remaining largely asymptomatic until adulthood and the infection invariably persists throughout the lifetime of the host (Malaty, 2007, Bruce and Maaroos, 2008, Robinson and Atherton, 2021). *H. pylori* infection varies between children according to geographical location and socioeconomic factors. For example, histology screening of 129 children aged between 2-19 years in Brazil revealed 41.1% had *H. pylori* infection. Also histology screening of 1634 children aged between 1-18 years in China reported 32.1% had *H. pylori* infection (Leonardo et al., 2014).

Although the mode of transmission of *H. pylori* remains unclear, interpersonal transmission seems to be the major pathway (Leonardo et al., 2014). The transmission of *H. pylori* during childhood is assumed to occur via the orofaecal route involving parents to children or from between children themselves (Sonnenberg, 2022). Also, transmission has been attributed to drinking contaminated water soiled by faecal

waste in countries with poor public hygiene management (Sonnenberg, 2022). In Nigeria, the prevalence rate was reported as 87.8% among the adult population in the northern region of the country, associated with smoking, contaminated water sources and low socioeconomic status. The prevalence rate in other countries in Africa such as Burundi, Rwanda, Togo, Congo, Morocco, Ghana and Egypt have been reported as 70.8%, 75%, 70.41%, 93.1%, 63.8%, 88% and 66.12%, respectively. This is higher than in other locations such as Europe, North America, Australia, and Asia, where *H. pylori* infection rate reported as 20–40% (Europe), 23.1% (North America), 24.6% (Australia) and 48.8% (Asia).

1.1.3 Disease Spectrum of *H. pylori*

A well-described disease spectrum following *H. pylori* infection has been reported across different populations even though not many of the people infected develop a disease state such as gastritis (Yamaoka, 2010, Robinson and Atherton, 2021, Atherton, 2006). The bacterium actively causes mucosal damage that may subsequently result in gastric cancer, peptic ulcer disease and other gastric-related diseases (Fig 1.3). *H. pylori* infection accounts for approximately 75% of gastric cancers (Mégraud et al., 2015, Amieva and Peek Jr, 2016), 70-75% gastric ulcers, and 90-95% of duodenal ulcers globally (Ernst and Gold, 2000). The risk of developing peptic ulcer diseases has been recently reported as approximately 7% in *H. pylori* infected individuals throughout their life time (Lee et al., 2022).

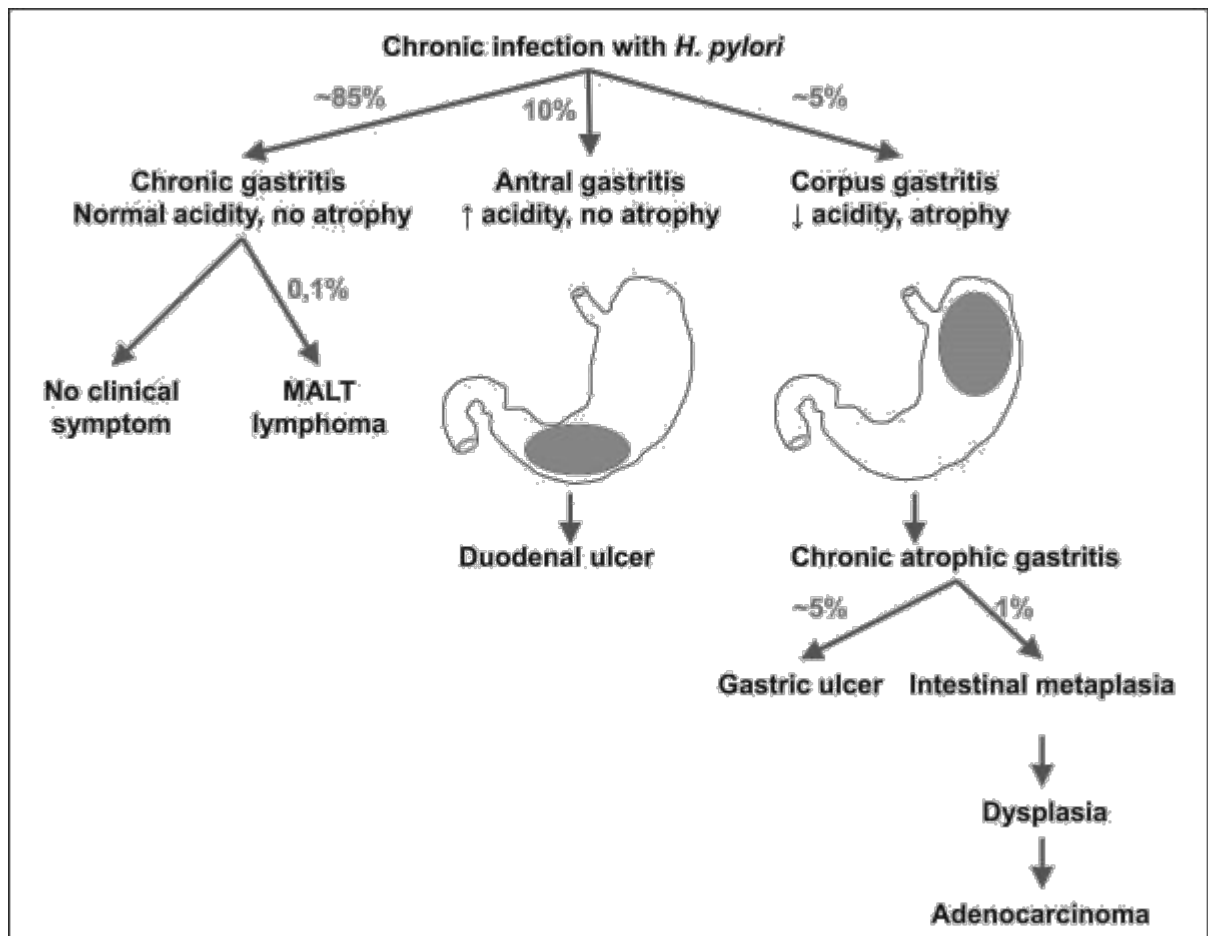


Fig 1.3: *H. pylori* disease spectrum showing stomach infection regions in chronic infection states.

Reproduced from Mégraud et al., 2015.

1.1.3.1 Acute and chronic gastritis

The word “gastritis” has been defined as the inflammation of the gastric mucosa (Jensen et al., 2019). *H. pylori* is one of the major pathogenic organisms responsible for gastritis globally. Infections with *H. pylori* are usually early in life (childhood) and capable of remaining throughout the life span if no treatment or eradication therapy is administered (Robinson et al., 2007). *H. pylori*-related gastritis is one of the most common inflammatory disorders and it affects more than two-third of the global population (Jensen et al., 2019). Individuals who develop gastritis display symptoms relating to either or both acute and chronic gastritis. Patients

presenting with the acute form of gastritis can develop symptoms such as vomiting, nausea, dyspepsia, feeling of fullness and reduced gastric acid secretion (Robinson and Atherton, 2021). Healthy individuals who deliberately or accidentally ingested *H. pylori* provide data correlating infection with symptoms manifested (Robinson and Atherton, 2021). In one study (Graham et al., 2004), more than 90% of the healthy individuals administered with *H. pylori* presented with mild to moderate dyspeptic symptoms which later resolved within days. It remains unclear whether children presenting with primary *H. pylori* infection experience similar symptoms to adults (Robinson and Atherton, 2021) despite acute gastritis being one of the most common illnesses in children (Nakamura et al., 2021).

Continuous persistence of *H. pylori* infection (Robinson and Atherton, 2021) leads to chronic gastritis. The development of chronic gastritis, is usually dependent on the level of gastric acid secretion and the location of the fully inflamed lymphatic cells in the stomach. This is an important determinant of the outcome of *H. pylori* infection in patients. For example, *H. pylori* colonisation is usually in the antrum of the stomach in people with normal level of acid production (Kuipers et al., 1995). Gastritis related to this is usually antrum predominant. In contrast, people with limited acid production generally have *H. pylori* colonisation equally distributed between both the antrum and the corpus. The colonisation of *H. pylori* is generally close to the gastric mucosa, thus making it corpus predominant pangastritis (Kuipers et al., 1995). Also, individuals with low

acid secretion caused by reduced production of parietal cells can develop gastric atrophy, which can later result in the development of intestinal metaplasia and adenocarcinoma (Robinson and Atherton, 2021).

1.1.3.2 Gastric Cancer

Gastric cancer has been defined as a stomach tumour which may arise from the connective tissues of the gastric wall, gastric mucosa, neuroendocrine tissue, or lymphoid tissue. Gastric cancer is the fourth most prevalent cause of cancer-related death globally (Robinson and Atherton, 2021, Sukri et al., 2020, Cordova-Marks et al., 2022) with prevalence higher in East Asian countries such as China, Japan and South Korea (Hatakeyama, 2014, Robinson and Atherton, 2021) than in Europe, the United States and Africa. Gastric cancer development is more common in men than in women. It is the 4th most prevalent cancer-type in men and 5th most common cancer in women among age range 60-84 years as reported by the American Cancer Society. In comparison with other forms of cancer such as prostate and breast cancer, gastric cancer fatality is higher, at 70%, compared to 30% and 33% for prostate and breast cancer, respectively. Similar to other types of cancer, gastric cancer develops via multiple progressions (Hatakeyama, 2014). These progressions may be regulated by either environmental or host genetic factors (Robinson and Atherton, 2021) amongst which *H. pylori* is the major risk factor for development of both the intestinal or diffuse form of gastric adenocarcinoma (Hatakeyama, 2014, Robinson and Atherton, 2021). Lifestyle behaviours such as smoking, high alcohol intake,

coupled with bacterial infection increases the risk of gastric cancer development.

The association between colonisation of *H. pylori* and the emergence of antral gastritis in the mucosa of both adults and children has been well established (Robinson and Atherton, 2021). Examination of infected adult mucosa by biopsy invariably reveals chronic inflammation and increased influx of neutrophils (Robinson and Atherton, 2021). Persistent infection may result in development of other gastric inflammation and some positive gastritis patients may develop a rise in gastrin and gastric acid secretion (Öztekin et al., 2021). This possibly increases the risk of gastric cancer development by allowing the passage of pepsin and luminal acid to stomach tissue.

1.1.3.2.1 Gastric mucosa-associated lymphoid tissue (MALT) lymphoma

H. pylori infection is also the leading cause for gastric mucosa-associated lymphoid tissue (MALT) lymphoma, although this is much less common than gastric cancer (Robinson and Atherton, 2021). MALT lymphoma originates from the proliferation of the B-cells which can spread uncontrollably in the gastric mucosa (Robinson and Atherton, 2021). Antibiotics can cure most early stage MALT lymphoma cases as demonstrated by several studies (Zullo et al., 2009, Nakamura et al., 2012), as long as it has not spread outside of the stomach. *H. pylori* plays an important role during proliferation in MALT lymphoma (Robinson and

Atherton, 2021): bacterial virulence factors and inflammatory microenvironment have been implicated in triggering proliferation of the B cells.

1.1.3.3 Peptic Ulcer Disease (PUD)

Peptic ulcer disease (PUD) refers to both duodenal and gastric ulcers (Malfertheiner et al., 2009) which result from injury to the digestive tract, culminating in mucosal damage characterised with bleeding or perforation which if untreated may kill the host (Ramakrishnan and Salinas, 2007). Bleeding, the most severe complication of PUD, has been reported in 50-170 per 100000 cases, with adults aged ≥ 60 years having the highest risk. Both duodenal and intestinal forms of PUD share the same characteristics, making it difficult to differentiate them - both manifest as abdominal uneasiness after food intake and epigastric pain (Ramakrishnan and Salinas, 2007). Nevertheless, approximately 90-100% of duodenal ulcer patients are positive for *H. pylori* infection, and 60-100% in gastric ulcer patients (Kusters et al., 2006). *H. pylori* infection and use of non-steroidal anti-inflammatory drugs (NSAIDs) result in increased peptic ulcer progression. More than 80% of the peptic ulcer disease cases are caused by *H. pylori* infection while the use of NSAIDs contributes to the remaining percentage (Atherton, 2006). Among other factors that contribute to peptic ulcer pathogenesis are host and environmental factors (excessive alcohol consumption, smoking), although none of these have been single-handedly reported to cause

ulceration (Malfertheiner et al., 2009). Though the infection is generally acquired during childhood, *H. pylori* induced-peptic ulcer disease is rare in children and its associated peptic dysfunctions – gastropathy, gastritis, oesophagus and duodenitis are uncommon in children (Urs et al., 2014). Notwithstanding the extent of *H. pylori* infection, the incidence of peptic ulcerations is relatively low, approximately 10-15% in infected patients [30]. Ulcer development in *H. pylori* infected patients is dependent on factors such as the strain of *H. pylori*, immunopathogenesis, *H. pylori* interactions with the stomach barrier, changes in acid secretion (Malfertheiner et al., 2009) smoking and genetic make-up (Kusters et al., 2006).

1.1.4 *H. pylori* diagnosis and treatment

1.1.4.1 Diagnosis

Recently, the policy guiding who to test, when to test, and how to test for *H. pylori* infection has been reviewed by the gastroenterology community. The community recommended that eradication therapy should follow immediately after the diagnosis of *H. pylori* infection. This is due to the recognition of the benefits provided by eradication therapies of *H. pylori* infection over the years. The recommendation has been confirmed since agreement in 2015. Nevertheless, the most recent consensus on testing for *H. pylori* infection agreed on swift proactive testing and eradication for *H. pylori* infection. Table 1.1 below adapted from Lee and colleagues (Lee et al., 2022) provided recommendations for whom to test (individual testing and population-based testing) from

Houston and Taipei consensus conferences respectively. In their recommendations, *H. pylori* diagnosis was recommended for patients with a history of either peptic or gastric ulcer and un-investigated dyspepsia.

Table 1.1: Recommendations for individual and population-based testing for *H. pylori* infections.

Houston Consensus Conference recommendations: which individuals to test?	Agreement (%)	Evidence level
Individuals with suspected <i>H. pylori</i> infections (example active DU)	100	High
Individuals with current or past gastric or duodenal ulcers	100	High
Individuals with un-investigated dyspepsia	100	High
Individuals with gastric mucosal-associated lymphoma	100	Moderate
Family members residing in same household of patients with proven active <i>H. pylori</i> infections	91	Moderate
Individuals with family history of peptic ulcer disease	91	Moderate
Individuals with family history of gastric cancer	100	Moderate
First-generation immigrants from high-prevalence cancer regions	82	High
High-risk groups(example, in the US, Latino and African American, and other racial and ethnic groups)	91	Low
Taipei Global Consensus recommendations: which specific populations to screen?	Agreement (%)	Evidence level
Population with high prevalence of gastric cancer	84	low
Young adults in high-prevalent populations to reduce the transmission to their children	92	low
Young adults in high-prevalent populations before the development of atrophic gastritis	84	low
Populations with high prevalence being integrated or included into the national healthcare priorities	92	low

Reproduced from Lee et al., 2022. Level of agreement was defined as 80% or more agreed or strongly agreed. Abbreviations: US - United States of America, DU- duodenal ulcer.

1.1.4.2 Methods for diagnosis of *H. pylori* infection

Several tests have been developed to detect *H. pylori* infection (Dore and Pes, 2021). These diagnostic tests are broadly divided into invasive and non-invasive tests (Granstrom et al., 2008, Dore and Pes, 2021). Apart from the patient's need, the testing method used should also reflect the cost of the test and local availability (Lee et al., 2022). The invasive test entails culture and histology of gastric samples, while the non-invasive testing uses faeces, blood, saliva or urine for detection of bacterial antigens, antibodies or urease activity, the enzyme that converts urea in the stomach to ammonia and carbon-dioxide (Kusters et al., 2006). For large population studies, serology testing is mostly employed while routine diagnostic testing uses the urea breath test, culture and histology tests. In children, detection of *H. pylori* is by faecal antigen testing, bypassing the need for a blood sample or endoscopy. Table 1.2 provides a summary of the current available tests for *H. pylori* infection, their strengths and weaknesses as reported by Lee et al. (Lee et al., 2022).

Table 1.2: Tests for *H. pylori* infection.

Tests	Strengths	Weakness
Non-invasive		
Serology	Widely available, least expensive and does not require medication modifications prior to testing	Does not reliably delineate between active and previous infection and cannot be used to confirm eradication
Stool antigen test	High sensitivity and specificity, can be used to test for active infection and evaluate for eradication	Stool samples needed, patients' aversion, requires prior cessation of antibiotics, bismuth products, or proton-pump inhibitors to reduce risk of false negative results
Urea breath test	High sensitivity and specificity; can be used to test for active infection and evaluate for eradication	Resources and trained personnel needed to reliably reproduce test; Require prior cessation of antibiotics, bismuth products, or proton-pump inhibitors to reduce risk of false negative results
Invasive (Endoscopic)		
Culture	Allows testing antibiotics susceptibilities	Poor availability in some countries
Molecular-based testing	Detects infection and can assess susceptibility/resistance for all six commonly used antibiotics; Stool can be used; Rapid results in days.	May not be covered by insurance; Availability only as "send out"
Histology	Can be used to test for infection and evaluate for eradication; Provides additional information such as degree of inflammation and associated pathology (e.g. intestinal metaplasia, atrophic gastritis)	Accurate results require interested pathologist and use of special stain, preferably immunohistochemical
	Rapid; inexpensive; Good sensitivity and specificity	Requires prior cessation of antibiotics, bismuth products, or proton-pump inhibitors to reduce risk of false negative results

Reproduced from Lee et al., 2022.

1.1.4.3 Treatment

H. pylori infection is usually treated by triple therapy, comprising a combination of two antibiotics such as amoxicillin and clarithromycin together with a proton pump inhibitor (PPI) or bismuth compound (Bytzer and O'Morain, 2005, Wolle and Malfertheiner, 2007). A normal successful treatment option should eradicate 90-95% of *H. pylori* infection (Graham and Fischbach, 2010). Many strains of *H. pylori* are becoming resistant to antibiotics, therefore quadruple therapy (consisting of a bismuth compound, two antibiotics - clarithromycin and metronidazole, and a PPI) may be considered when eradication is unsuccessful with triple therapy (Kusters et al., 2006). *H. pylori* resistance to antibiotics is a growing challenge in the clinical setting (Bytzer and O'Morain, 2005) where the eradication rate is below 80% and in countries such as Turkey, Italy, France and Spain where clarithromycin resistance is high (Graham and Fischbach, 2010).

1.2 Factors that contribute to disease spectrum of *H. pylori*

The disease outcome of *H. pylori* infection is influenced by several factors including bacterial virulence factors, host genetic make-up and environmental factors.

1.2.1 *H. pylori* virulence factors

H. pylori strains are highly genetically diverse, and this diversity may account for some of the different disease outcomes across different geographical populations. Several virulence factors have been reported

to contribute to pathogenicity (Salih et al., 2014) including – the *cytotoxin-associated gene (cag)* pathogenicity island (*cagPAI*), adhesins, outer membrane proteins (OMP) and the vacuolating cytotoxin (VacA). Adherence of the bacteria to the mucosal layer activates both bacterial chromosomal outer inflammatory protein (Oip) and *cagPAI* that promote pathogenicity and possibly ulcerogenic effects (Ramakrishnan and Salinas, 2007). *H. pylori* virulence factors that promote bacterial colonisation include urease, flagella, and outer membrane proteins, while the *cag* pathogenicity island (encoding CagA and VacA) is important for disease induction and immune escape. Historically, VacA was discovered first (Covacci et al., 1993) followed by CagA, which was named in relation to VacA as cytotoxin-associated gene A (Knorr et al., 2019).

1.2.1.1 Vacuolating cytotoxin A (VacA)

The Vacuolating cytotoxin, VacA (88kDa) is an important virulence factor of *H. pylori* and the most extensively studied in terms of *H. pylori* toxins. VacA, given the name because of its capability to induce vacuole formation in the cytoplasm of host cells (Palframan et al., 2012), possesses two major subunits - p33 and p55 (Foegeding et al., 2016). It is encoded by the *vacA* gene, which is present in all strains of *H. pylori* with notable sequence diversity between different geographical strains. Studies of *vacA* polymorphisms have revealed three distinct variable regions of the gene – the signal sequence region (*s-region*), the mid-region (*m-region*) and recently discovered third intermediate region (*i-*

region) (Chung et al., 2010) (Fig 1.4). Within both the *s* and *m* regions, variations have been classified: *s1* or *s2* and *m1* or *m2*, respectively (Chung et al., 2010). Both variations differ in each subtype (*s* and *m* regions) in that *s1* toxin possesses no additional segment of N-terminal hydrophilic amino acid unlike the *s2* which has the additional segment. This segment prevents *s2* toxin from triggering vacuole formation (Fig. 1.5).

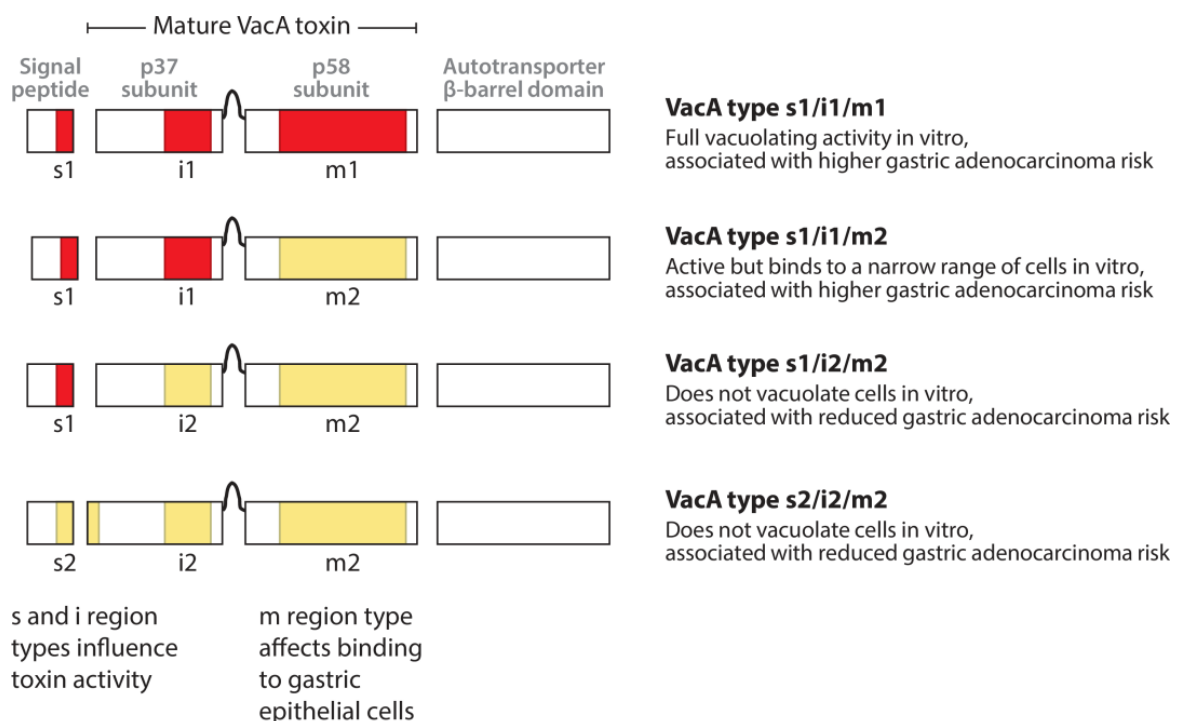


Fig 1.4: Schematic representation of the allelic diversity that exist in three regions of *vacA* gene and their associations with disease.

The *vacA* gene varies most in the signal (*s*) region (which determines toxin activity), the intermediate (*i*) region located within the p37 subunit (which determines toxin activity), and the mid (*m*) region within the p58 subunit (which determines binding to host cells). Only commonly occurring combinations of regions are shown. Type 1 regions are shown in red, and type 2 in yellow. Reproduced from Robinson and Atherton, 2021.

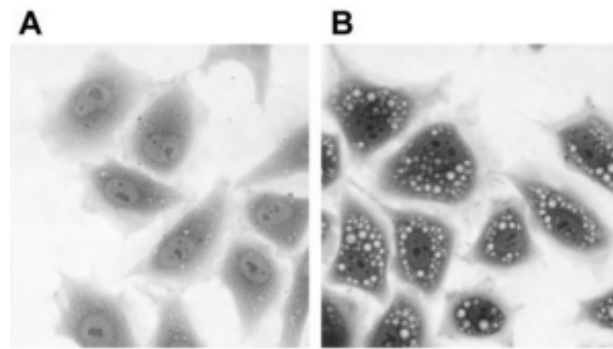


Fig 1.5: Vacuolation effect of VacA on HeLa cells *in vitro*.

(A) Untreated HeLa cells with VacA. (B) HeLa cells treated with VacA. White patches are vacuoles pores formed after treatment with VacA. Reproduced from Schraw et al., 2002.

Studies have reported that *H. pylori* strains that harbour the *s1/m1* VacA toxin contribute to increased disease outcomes in terms of chronic inflammation (Rhead et al., 2007). The recently discovered intermediate region (i-region) has been reported to associate strongly with VacA production from various geographical strains of *H. pylori*. This association may suggest *i-region* importance in disease outcomes of *H. pylori* infection. Furthermore, epidemiological studies have reported two novel polymorphic regions called the deletion (d1 and d2) and c-region (c1 and c2). Both regions are located at the 3'- end of VacA and the variants of these regions have been associated with increased risk of gastric cancer in patients.

Vacuolation is the most recognised effect of VacA entry into host epithelial cells, which occurs because VacA produces anion-selective channels that facilitate chloride ions transfer, thus promoting the increase in intra-luminal chloride concentrations. Its vacuolating activity has been

linked to triggering cell apoptosis via the release of cytochrome c from the mitochondria (Palframan et al., 2012, Chung et al., 2010). It has also been shown to induce gastroduodenal damage in mouse models, and causes an increase in gastric ulceration risk in infected Mongolian gerbils (Rhead et al., 2007). VacA not only binds to host epithelial cells but also to multiple other cell types including T-cells. The main receptor that binds VacA to T-lymphocytes is beta-2 integrin lymphocyte function associated antigen-1 (LFA-1). In contrast, epidermal growth factor (EGF), tyrosine phosphatase alpha, glycosphingolipids, heparin sulphate and sphingomyelin are among surface receptors on gastric epithelial cells that bind VacA (Palframan et al., 2012).

1.2.1.2 *Cag* pathogenicity island (*cagPAI*)

The *cagPAI* is approximately 40kb in size and encodes approximately 30 genes, depending on the strain (Robinson and Atherton, 2021, Fischer et al., 2020). Sixteen of the genes are found in the upstream region of the PAI, termed *cagII* (type-1 *cagPAI*) and 14 genes are located in the downstream region termed *cagI* [35], including *cagA* gene (Fig 1.6). Many of the genes within the *cagPAI* encode the *cag* type IV secretion system (T4SS) (Robinson and Atherton, 2021) (Fig 1.7), which translocates the effector protein called cytotoxin associated antigen A (CagA) (Olbermann et al., 2010, Fischer et al., 2020) into host epithelial cells. The *cagPAI* T4SS is the most well characterised T4SS in the membrane of *H. pylori*, assembled from the products of an orthologous

complement of *virB/D* genes (B1–B11 and D4) found in the model organism *Agrobacterium tumefaciens*.

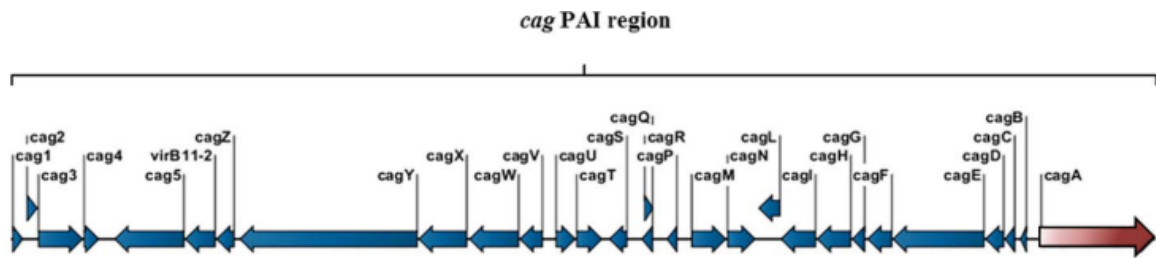


Fig. 1.6: Cluster of genes that form the *cag* T4SS in *H. pylori* strain P12.

The structure of the *cagPAI* region comprising of 28 genes that encode the *cagPAI* T4SS, including the CagA protein. The genes encoded are similar to the cluster of genes in integrative and conjugative elements – *tfs* ICEs and *tfs* ICEs, (discussed later in this chapter). Reproduced from Kabamba et al., 2018.

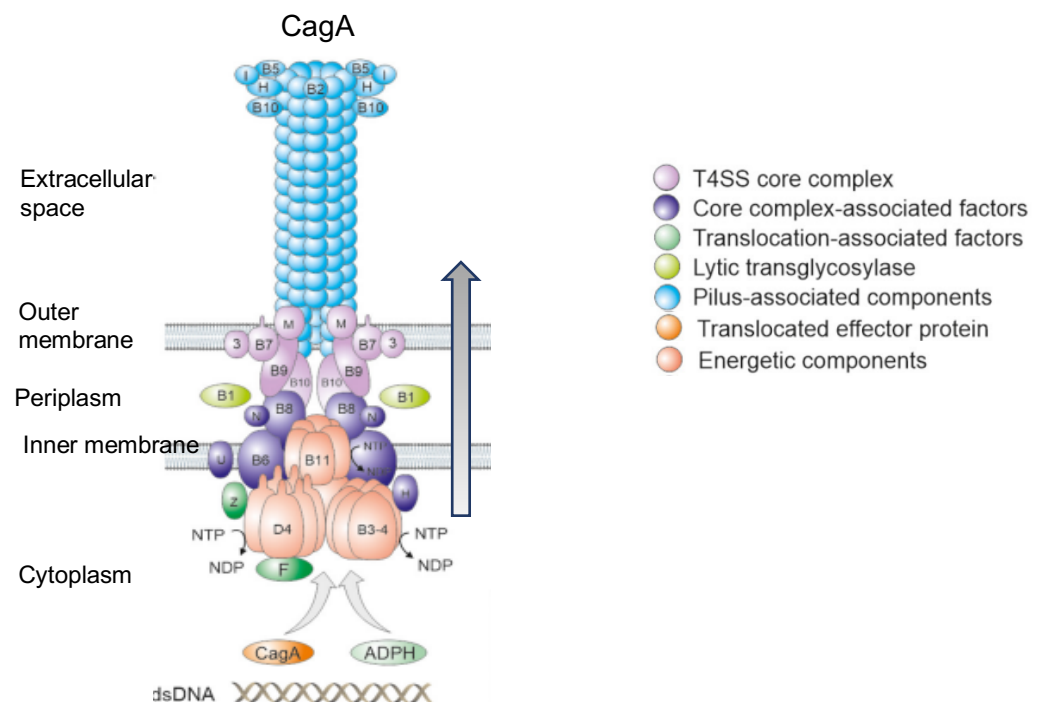


Fig 1.7: Structural model for *cagPAI* T4SS in *H. pylori*.

VirB proteins with homology to the prototype *Agrobacterium tumefaciens* are denoted (B1–B11 and D4). Accessory Cag proteins that form the structure of *cagPAI* T4SS are represented in colours with their simplified positions. Upward arrows represent effector (CagA) delivery direction in *cagPAI* T4SS. Reproduced from Fischer et al., 2020.

The *cagA* gene has been reported to be present in more than 50-70% of *H. pylori* strains and is known to increase strain virulence (Kusters et al., 2006, Maeda et al., 1999, Robinson and Atherton, 2021). *H. pylori* strains that harbour the *cagPAI* are called *cag* positive strains while strains that do not carry a *cagPAI* are referred to as *cag* negative. Patients infected with a *cag*⁺ strain are at greater risk of disease outcomes such as gastric cancer or peptic ulcer disease. These strains cause increased inflammatory responses by inducing expression of interleukin 8 (IL-8) and other inflammatory cytokines and chemokines by gastric epithelial cells (GECs) in the host (Robinson and Atherton, 2021). Upon translocation into GECs via the T4SS, CagA is phosphorylated by Src and Abl dependent tyrosine kinases (Krisch et al., 2016) leading to activation of SHP-2 (Src homology region 2 domain-containing phosphatase-2), resulting in elongation of the cell and cell motility (Krisch et al., 2016). Therefore, the gastric epithelial cell is dysregulated via homeostatic signal transduction which results in cell apoptosis, distortion of cell proliferation and polarity.

The *cagA* gene encodes a family of proteins with molecular weights ranging from 120 to 140 kD which are highly immunogenic and induce a specific serological response in infected individuals (Robinson and Atherton, 2021, Robinson et al., 2007). The presence of CagA serum antibodies in infected individuals with CagA⁺ strains of *H. pylori* indicates an increased risk of developing duodenal ulcers and gastric cancer (Cover et al., 1995). CagA has been reported to exist in variant forms.

The presence of the variant forms is highly linked to disease progressions such as gastric adenocarcinoma due to alterations in the host cell signalling pathway. At the C-terminal region of CagA are conserved motifs termed EPIYA being composed of Glu-Pro-Ile-Tyr-Ala (Altanbayar et al., 2022, Robinson and Atherton, 2021). The EPIYA motif has been classified into four namely– EPIYA A, B, C and D (Altanbayar et al., 2022). These four subtypes of EPIYA motifs have been reported in different strains of *H. pylori* from different geographical locations. For example, strains from the United States, Australia and Europe harbours Western type, that is, the EYIPA-A, B and C while EPIYA - A, B and D are commonly found in East Asian strains (Altanbayar et al., 2022, Robinson and Atherton, 2021). *H. pylori* strains that harbour the latter induce increased level of IL-8 in epithelial cells of the host compared to strains that harbour EPIYA - A, B, and C subtypes. Correlation between the EPIYA motif and disease progression of *H. pylori* infection has been reported in infected populations. A study by (Tserentogtokh et al., 2019) reported the presence of EPIYA motif (Western type) in CagA⁺ strains of patients who presented with gastritis.

The cytotoxin associated gene E (*cagE*), a homolog of *virB4* and *ptlC* in *A. tumefaciens* and *Bordetella pertussis*, respectively, is another important component of the *cagPAI* T4SS located close to *cagA* (Yamazaki et al., 2005, El Khadir et al., 2021) in the *cagPAI*. *cagE* is the second well-characterised marker gene of *cagPAI* (El Khadir et al., 2021) after *cagA* and, has been reported to be well conserved among different

strains of *H. pylori* (Yamazaki et al., 2005). The presence of CagE is indispensable to the translocation of CagA and its phosphorylation (El Khadir et al., 2021). It is also important in the biogenesis of the *cagPAI* T4SS system. Therefore, the removal of *cagE* has been reported to affect the synthesis of functional *cagPAI* T4SS (El Khadir et al., 2021). The presence of *cagA*, *cagE* and EPIYA motifs have been implicated in disease outcomes such as duodenal ulcer and gastric cancer development (El Khadir et al., 2021). A study by Bakhi and co-workers (Bakhti et al., 2020) concluded that both *cagA* and *cagE* presence increased the risk of gastric cancer development in an Iranian population. Another recent study carried out in a Moroccan population reported that people infected with isolates with both *cagA* and *cagE* presented with increased risk of duodenal ulcer development (El Khadir et al., 2021).

1.2.1.3 Host genetic factors

Factors other than *H. pylori* virulence factors are considered to contribute to disease pathogenesis because most infected people remain asymptomatic. This has led to the investigation of host genetic factors that might be linked to disease outcomes in some infected patients. For example, interleukin 1 beta (IL-1 β) is a pleiotropic proinflammatory molecule that interacts directly with some of the *H. pylori* virulence factors (Wroblewski et al., 2010). The IL-1 β gene cluster comprises of *IL-1 β* and *IL-1RN*, that encodes IL-1 β receptor antagonist (Wroblewski et al., 2010). IL-1 β polymorphisms resulting in high-expression have been linked to increased risk of hypochlorhydria, distal gastric

adenocarcinoma and gastric atrophy. It is a well established risk factor in infected patients but not in uninfected individuals (Wroblewski et al., 2010). In individuals with high-expression IL-1 β alleles, the risk of gastric cancer has been reported to increase by 87-fold in those infected with *vacA* s1-type *H. pylori* (Wroblewski et al., 2010) as compared with uninfected individuals.

Also a polymorphism in IL-8, which affects the innate immune response, has been reported to contribute to an increased risk of gastric cancer outcomes. In Caucasian and Asian populations, a variant in the promoter section of IL-8 increases expression which has been linked to increased inflammation and gastric cancer rate, though this is debatable in some studies (Wroblewski et al., 2010). In white populations, a functional polymorphism in TLR4 has been linked to increased gastric and atrophic cancer incidence and may possibly be attributed to decreased IL-10 production.

1.2.1.4 Environmental factors

1.2.1.4.1 *H. pylori* and nutrition

In addition to the ability of *H. pylori* virulence factors to trigger proinflammatory responses upon infection in the host epithelial cells, infection by *H. pylori* also affects the brain-gut connection indirectly, similar to other microorganisms (Öztekin et al., 2021). The two important hormones, ghrelin and leptin, which control both growth and appetite,

can be affected alongside *H. pylori* infection causing immunological responses in infected individuals (Öztekin et al., 2021). This can result in changes to food consumption and appetite. *H. pylori* infection has been associated with malnutrition and growth retardation in children in developing countries (Franceschi et al., 2014, Robinson and Atherton, 2021).

Environmental factors are also important in disease outcomes of *H. pylori* infection. For example, several studies have reported the correlation between dietary salts and gastric cancer development. Studies conducted both in Japan and South Korea reported increased risk of gastric cancer in *H. pylori* infected individuals with increased dietary salt intake compared with infected patients with reduced salt intake. In an animal model using Mongolian gerbils, a high correlation between *H. pylori* infection with increased salt intake, which resulted in gastric cancer and premalignant lesions development, was observed (Wroblewski et al., 2010). Another study by Gaddy and colleagues (Gaddy et al., 2013) reported that *H. pylori* infected gerbils have increased levels of the proinflammatory cytokines IL-1, IL-6 and TNF-alpha when fed with more dietary salt compared to the gerbils given lower dietary salt. In another mouse study, increased salt intake has been shown to boost the colonisation of *H. pylori* in C57BL/6 mice, which elevated proliferation in epithelial cells (Wroblewski et al., 2010). In humans, the process by which increased salt intake correlates with gastric cancer in *H. pylori*

infected people is poorly studied, though studies suggest that gastric mucosa is damaged by salt, thus permitting the passage of carcinogens (Wroblewski et al., 2010). Another recent suggestion is that increased dietary salt intake may possibly modulate *H. pylori* gene expression, which has been validated by expression of CagA (Gaddy et al., 2013).

A growing body of evidence has identified that smoking cigarettes plays a significant role in the development of duodenal ulcer disease in individuals infected with *H. pylori* (Chung and Shelat, 2017). Several studies have identified that tobacco inhibits pancreatic bicarbonate secretion, which consequently results in increased acidity in the duodenum, and can contribute to peptic ulcer development.

Iron is an essential important nutrient involved in the maintenance of various metabolic activities such as electro transport, nucleotide synthesis, oxidation-reduction processes and amino acid synthesis, where it functions as a cofactor (Öztekin et al., 2021). Also, iron is needed by all bacteria for their growth. Therefore, a race between bacteria and the host for iron absorption is inevitable (Frawley and Fang, 2014). *H. pylori* infection in some individuals can result in iron deficiency in the long term as the host body competes and holds the available iron to prevent the bacterium growth and oxidative damage possible through interaction of iron with oxygen molecules, which is capable of generating oxygen radicals (Öztekin et al., 2021). It has been reported that iron deficiency

may also be effective in *H. pylori* virulence (Öztekin et al., 2021). A study by Noto and colleagues reported that the deficiency of iron enhances expression of *H. pylori* virulence factors, and it is a measurable biomarker to screen for population at high risk of developing gastric cancer (Noto et al., 2012).

Another important micronutrient which is essential is vitamin C. Vitamin C is a chemical agent which protects against oxidative stress in cancerous cells with its antioxidant effect and also increases the risk of cancer via its pro-oxidant activity (Öztekin et al., 2021). Therefore, infection with *H. pylori* can lead to reduction in the level of vitamin C in the host body. In a study reported by Woodward and colleagues, the level of vitamin C in plasma and gastric juice was low in individuals infected with *H. pylori* and with history of reduced intake of vitamin C (Woodward et al., 2001). In another study reported by Waring and colleagues (Waring et al., 1996), it was found that high dose of vitamin C (500 mg twice per day for 2 weeks), in 32 individuals presenting with gastritis caused by *H. pylori* infection, could reduce the risk of gastric cancer when compared with control group of 48 gastritis patients not given vitamin C supplement. However, high doses of vitamin C (500 mg) may result in gastrointestinal disturbances, and increased uric acid and oxalate excretion.

Upon *H. pylori* arrival in the gastric epithelium, it activates nicotinamide adenine dinucleotide phosphate hydrogen (NADPH) oxidase and produces reactive oxygen species (ROS) such as superoxide (O_2^-), hydrogen peroxide (H_2O_2), hypochlorous acid (HOCl) and hydroxyl radicals (OH) (Öztekin et al., 2021). The productions of these compounds trigger inflammatory responses. The ROS produced by *H. pylori* in the gastric mucosa then activates the oxidant-sensitive transcription factor NF- κ B. The NF- κ B factor then induces the expression of oncogenes which cause damage to the epithelial cells and induces cell apoptosis and DNA damage. Increase in ROS in *H. pylori* infected individuals has been associated with decreased production of antioxidants such as vitamin C, and development of gastric cancer caused by DNA damage (Poplawski et al., 2013). Since *H. pylori* can affect the oxidative stress process, it is thoughtful to assume that the presence of antioxidants in foods may be effective towards *H. pylori* infection (Öztekin et al., 2021). Therefore, several studies have suggested the addition of antioxidants in diet may have protective effects against *H. pylori* infection (Öztekin et al., 2021).

For example, garlic (*Allium sativum* L.), a medicinal plant cultivated in Asia contain important sulphur compound called allicin. It was first demonstrated in 1998 by Chung and colleagues that garlic can suppress the growth of *H. pylori* (Chung et al., 1998). Garlic suppressed *H. pylori* growth by inhibiting the activation of NF- κ B caused the Toll-like receptor

4 (TLR4) pathway (Öztekin et al., 2021). Another antioxidant turmeric (*Curcuma longa*), a polyphenolic compound contains curcumin has been reported to have various beneficial properties such anti-inflammatory, antioxidant, anticancer, antimicrobial and antifungal effects (Öztekin et al., 2021). A study conducted by Judaki and colleagues tested the effect of curcumin, (700 mg) one of the two groups of infected individuals with *H. pylori* they treated with the same protocol (Judaki et al., 2017). Curcumin was administered to one group three times per day for three months. After three months, the curcumin group reported reduced level of oxidative DNA damage compared to non-curcumin group (Judaki et al., 2017). Additionally, lycopene, a natural antioxidant that gives red colouration in fruits and vegetables (Öztekin et al., 2021, Grabowska et al., 2019) was reported to inhibit DNA damage and *H. pylori* induced gastric carcinomas (Jang et al., 2012).

Furthermore, it has been identified that N-nitroso compounds play a role in gastric carcinogenesis as they are recognised as mutagens and carcinogens (Wiseman, 2008). Nitrates which are endogenously produced in gastric acid contribute to the endogenous production and exposure to N-nitroso compounds. Wiseman *et.al* Identified in their study that N-nitroso compounds significantly increased gastrointestinal cancer risk. Red meat has been identified to promote the formation of N-nitroso compounds via reaction between myoglobin, haemoglobin, and nitric oxide. In addition, a component of red meat is iron which can lead to the production of free radicals. Taken together, these factors contribute to

the development of gastrointestinal cancer in individuals infected with *H. pylori* (Jakszyn et al., 2006, Loh et al., 2011).

1.3 Type IV Secretion Systems (T4SSs)

Type IV secretion systems (T4SSs) are complex macromolecular systems present in both Gram-negative and Gram-positive bacteria and some *Archaea* species (Fronzes et al., 2009b, Grohmann et al., 2018). T4SSs requires direct host cell contact to translocate macromolecules including protein and DNA substrates across the cell membrane(s) of a recipient cell (Alvarez-Martinez and Christie, 2009). They are important for survival and virulence, and although diverse in substrates composition and functions they share similar structural components and have similar functional characteristics (Green and Mecsas, 2016). The two major subfamilies of T4SSs reported to date, based on their functions, are the conjugative systems and effector translocators (Costa et al., 2021). Unlike the conjugative systems that are responsible for conjugative transfer of mobile genetic elements (MGEs), the effector translocators mainly delivers effector proteins to target cells in eukaryotic organisms (Costa et al., 2021, Fronzes et al., 2009a). The conjugative systems form the larger subfamily and usually found in Gram-negative bacteria. In addition, conjugative systems can spread antibiotic resistance genes in different pathogens. A third subtype of T4SSs are those responsible for DNA uptake from the environment and delivers the DNA into the extracellular milieu (Costa et al., 2021).

1.3.1 Structural components of T4SSs

Type IV secretion systems (T4SSs) in Gram-negative bacteria have been variably reported, but they all share similarity with the *A. tumefaciens* VirB/D T4SS which primarily consists of twelve proteins (Green and Mecsas, 2016) (Fig 1.8). Generally, in Gram-negative bacteria T4SSs are divided into two phylogenetic families namely the “typeIVA” and “typeIVB” (Grohmann et al., 2018). The model organisms used to represent the typeIVA are *Agrobacterium tumefaciens* VirB/D T4SS and the *Escherichia coli* conjugative machineries encoded by plasmids R388 and pKM101 respectively. Conversely, the paradigm for typeIVB is the *L. pneumophila* Dot/Icm system where more than 25 proteins are required for its assembly (Grohmann et al., 2018). The typeIVA T4SSs are mainly made up of 12 subunits called VirB1 – VirB11 and VirD4 coupling proteins (Fronzes et al., 2009a) based on the prototype from *A. tumefaciens* VirB/D T4SS (Grohmann et al., 2018). The proteins named VirB1-11 and VirD4, are organised on one or a few operons. Generally, T4SSs comprise of an extracellular pilus which has both minor (VirB5) and major (VirB2) subunits, several outer membrane proteins including VirB9 and the VirB7 lipoprotein, and the conserved inner membrane proteins VirB6, VirB8 and VirB10 (Fronzes et al., 2009a). Three energetic components, VirB4, VirB11 and the T4CP VirD4 generally associate with the cytoplasmic side (Grohmann et al., 2018) of the inner membrane and are required for T4SS structural assembly and substrate export (Fronzes et al., 2009a). Both VirB1 and VirB2 have been

reported to not be a part of the T4SS components, but they may assist in the assembly of the system (Fronzes et al., 2009a).

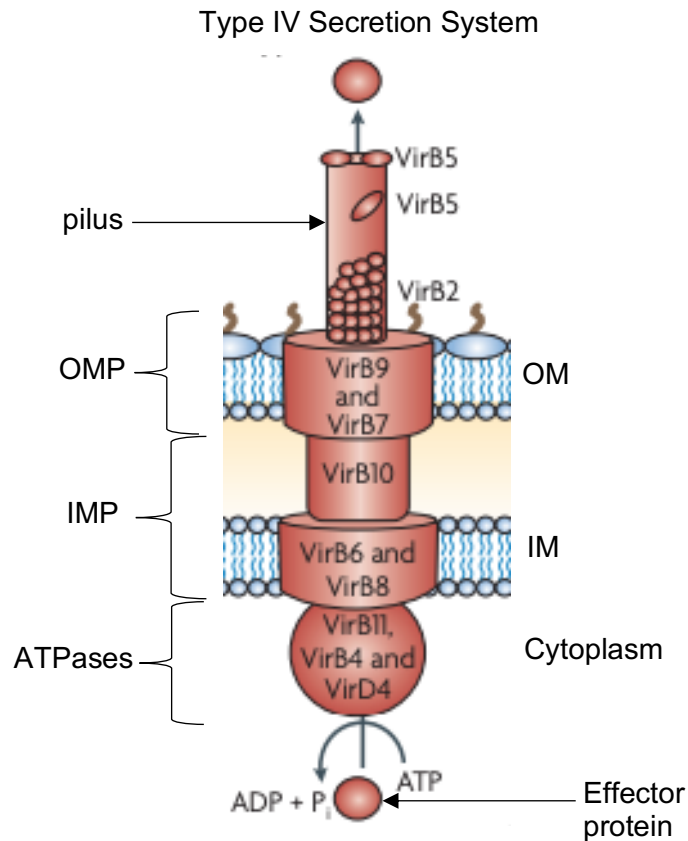


Fig 1.8: Schematic representation of type IV secretion system assembly. T4SSs are made up of 12 subunits VirB1–VirB11 and VirD4. Of all these, three are energetic ATPases (VirB4, VirB11 and VirD4), five membrane proteins located between OM and IM respectively. VirB2 and VirB5 form the pilus subunit. Abbreviations: OM – outer membrane, IM – inner membrane, OMP – outer membrane proteins, IMP – inner membrane proteins. Reproduced from Fronzes et al., 2009a.

1.3.2 The components of Type IV Secretion Systems

The twelve major subunits that form the T4SSs in bacteria have been extensively described. T4SSs are composed of important three ATPases (VirB4, VirB11 and VirD4) called the membrane energetics (Christie,

2004, Fronzes et al., 2009a, Costa et al., 2021). As their name implies, they form the power units that drive the secretion pathway. The trio of ATPases are crucial for substrate secretion (Christie, 2004) while VirB4 and VirB11 are known to participate in pilus biogenesis (Wallden et al., 2010). VirD4 has been given the name coupling protein (CP) due to its crucial engagement to recruit substrate secretion system via translocation route (Fronzes et al., 2009a). The recruitment of substrates by VirD4 has been presumed to occur through direct interaction with the substrates via binding secretion signal sequences, followed by its ability to link the substrates to other components of the secretion system (Fronzes et al., 2009a). VirD4-like proteins have also been reported in other bacteria such as the *Brucella spp.* The T4SS in this species uses the effector translocation system and the recruitment and secretion of substrates are totally independent of the coupling proteins. Alternatively, substrate recruitment and secretion occur directly via the secretory pathway or via a receptor with no sequence homology to VirD4 (Fronzes et al., 2009a). The coupling proteins have two motifs (Walker A and B) crucial for hydrolysis and nucleotide binding (Costa et al., 2021, Gomis-Ruth et al., 2004, Gomis-RuÈth et al., 2001). Studies have confirmed that any alterations or mutations of these motifs will prevent the translocation of substrate in the secretion system. Furthermore, the VirD4 homologue (TrwB) encoded by the *E. coli* plasmid R388 has been shown to display ATPase activity (Tato et al., 2007).

VirB11 another member of the ATPases family belongs to the “traffic ATPases” subgroup which are usually linked to secretion system types II, III , IV and VI (Wallden et al., 2010, Fronzes et al., 2009a) in Gram-negative bacteria. The VirB11 proteins are capable of self-display of ATPase activities. Also, these activities can be stimulated via interactions with lipids, which suggest that VirB11 association with components of the membrane is biologically important (Fronzes et al., 2009a). Extensive work carried out using electron microscopy to visualise the structure of VirB11-like proteins present in different bacteria such as *H. pylori* and *B. suis.*, revealed that the VirB11 homologue (HP0525) of Cag T4SS in *H. pylori* show hexameric rings of approximately 100 – 120Å in diameter (Hare et al., 2006). Further crystal structure analysis revealed that individual monomers consist of two domains, the amino and carboxyl termina denoted as N-terminal domain (NTD) and C- terminal domain (CTD) respectively (Fronzes et al., 2009a). Located in-between the domains is an interface known as the nucleotide-binding site. In the absence of the site, the CTD domain usually remains unchanged, thus maintaining the hexameric structure while the NTD domain is flexible and present rigid-body conformations. When nucleotide-binding site is present, the ATP molecules, thus promoting the movement of NTD to form a rigid conformation simultaneously with the hydrolysis of the ATP molecules. Upon complete hydrolysis, the hexameric structure of VirB11 is relaxed to its original nucleotide-free state (Fronzes et al., 2009a, Savvides et al., 2003).

VirB4 is another important member of T4SSs ATPases present in both Gram-negative and Gram-positive bacteria. VirB4 proteins are usually associated with conjugative T4SSs (Grohmann et al., 2003) and in comparison, with CPs, they have conserved Walker A and B motifs which are indispensable in secretion, hydrolysis, and nucleotide binding. VirB4 homologs such as TrwK from *E. coli* plasmid R388 have been shown to have up to four transmembrane segments (Fronzes et al., 2009a). The ATPase activity of VirB4 has been reported in hexameric form which suggests that VirB4 exhibit hexameric function similar to VirB11 and VirD4 (Wallden et al., 2010). In *A. tumefaciens*, VirB4 is attached to the inner membrane through the membrane spanning segments and its interaction with other membrane components enable its peripheral binding to the membrane (Wallden et al., 2010).

The T4SSs core complex is composed of the VirB7, VirB9 and VirB10 proteins. Structural analysis of the complex with Cryoelectron microscopy revealed the arrangement of the complex in a cylindrical manner approximately 185Å. Also, the complex was reported to harbour two distinct layers (O and I layers) (Fronzes et al., 2009a). Comparison between the two layers revealed that the I layer comprises of the NTD of VirB9 and VirB10 present in the inner membrane while the O layer is made up of the main complex body. The VirB10 is a single-spanning inner membrane protein made up of one transmembrane segment, short N-terminal cytoplasmic region, and a C-terminal periplasmic domain which is proline-rich (Fronzes et al., 2009a). The domain consists of

flexible helix-turn-helix repeat and the proline-rich region is crucial for core complex assembly and secretion of substrate. VirB7 has been reported to interact with VirB9 (Wallden et al., 2010, Fronzes et al., 2009a). This interaction helps to stabilise VirB9 protein. Both VirB7 and VirB9 proteins consist of signal peptides and they both form disulphide bridge upon interaction in the prototype organism (*A. tumefaciens*) (Fronzes et al., 2009a). Also, structural analysis shows that VirB10 forms part of the outer membrane channel and is capable of inserting into the inner membrane channel. Thus, VirB10 is capable of interaction with VirB components of the inner membrane and with the energetics ATPases. Therefore, VirB10 is unique in its function in that, it can monitor the opening and closing of the outer membrane and relate any conformational changes that may occur with the energetics ATPases (Fronzes et al., 2009a).

VirB8 is a bitopic inner membrane protein which consists of a N-terminal transmembrane segment and a large periplasmic C-terminal domain (Fronzes et al., 2009a). VirB8 functions in both pilus biogenesis and the secretion of T4SS substrate. The analysis of its crystal structure revealed that the periplasmic domains its homologous in *B. suis* and prototype organism (*A. tumefaciens*) consists of well-extended beta sheet usually flanked by five alpha helices (Fronzes et al., 2009a). VirB8 undergoes flexible interactions with other components of T4SS including outer membrane channel proteins. Furthermore, the ability for VirB8 protein to

dimerize as revealed by crystal structure and oligomerisation via mutational studies has been reported (Buhrdorf et al., 2003).

Both VirB1 and VirB3 are classified as “orphan proteins” due to evidence that they might not be part of the membrane secretion components, and they remain indispensable in the formation of the T4SS system (Fronzes et al., 2009a). In *A. tumefaciens*, the deletion of the *virB1* gene has been reported to disrupt T-pilus biogenesis although the secretion pathway is not affected (Fronzes et al., 2009b).

1.3.3 Biochemical and functional interactions of VirB proteins in T4SSs.

The VirB/D T4SS in *A. tumefaciens* has been used to establish VirB/D protein-protein interactions in the T4SS assembly pathway (Fronzes et al., 2009a) (Fig 1.9). In VirB/D T4SS, interactions have been reported between the membrane energetic components (VirB4, VirB11 and VirD4), VirB6 interacts with outer membrane components VirB7 and VirB9, VirB8 interact with VirB1, VirB4, VirB5, VirB11, VirB7 and VirB9 (VirB7, a small lipoprotein outside the membrane which helps stabilise VirB9). These arrays of interaction help in the translocation pathway of T-DNA into the plant (Fronzes et al., 2009a).

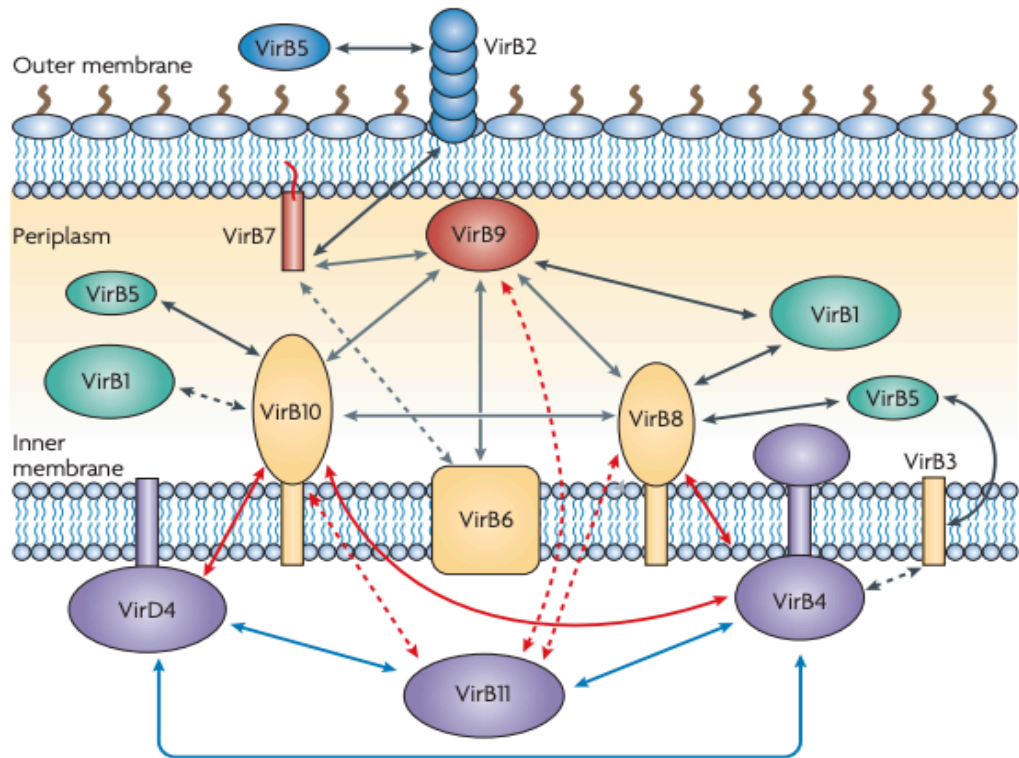


Fig 1.9: Schematic representation of interaction network of VirB/D homologues protein from *Agrobacterium tumefaciens*.

Cytoplasmic ATPases are shown in purple, IMPs shown in blue, PPs in green, OMPs in red and Eps in blue. Blue arrows indicate interactions between ATPases, reds arrows indicate interactions between ATPases and non-ATPases, grey arrows indicate interactions between core T4SS components, black arrows indicate other interactions and dotted arrows indicate possible interactions. Reproduced from Fronzes et al., 2009a.

1.3.4 Mechanism of substrate translocation in T4SS

In the prototype organism *A. tumefaciens* VirB/D T4SS, the transfer DNA (T-DNA) makes contact with VirD4, VirB11, VirB6, VirB8, VirB9 and with VirB2 (Fronzes et al., 2009a). The first T4SS component that makes contact with the substrate is the VirD4 which acts as substrate receptor at the cytoplasmic entrance. VirD4 then transfers the substrate to VirB11 (Fronzes et al., 2009b). Substrate transfer to VirB11 could not be

detected when the small lipoprotein VirB7 was deleted. The deletion shows the importance of VirB7 in the formation of core complex in T4SS assembly. The translocation of substrate has been reported not to require energy (ATP) because mutation in the important motifs (Walker A and B) of both VirD4 and VirB11 have no inhibitory effect on substrate transfer. The T-DNA substrate from VirB11 is then delivered to inner membrane T4SS components VirB6 and VirB8. The deletion of either of these components (VirB6 or VirB8) have been reported to block the translocation process suggesting that both components work together (Fronzes et al., 2009a). In the final translocation process, the substrate is transferred through the periplasm to the outer membrane components VirB2 and VirB9. VirB3, VirB5 and VirB10 are all important for the final translocation process. VirB10 is crucial for the final step as it links both inner and outer membranes where it detects ATP hydrolysis by VirB11 and VirD4 at the inner membrane, and subsequently undergoes conformational changes that allow the passage of substrate through the periplasm (Fronzes et al., 2009a).

1.4 *H. pylori* T4SSs

Genes encoded with the genomes of *H. pylori* are exceptionally diverse in different strains of *H. pylori* described to date. Genetic diversity has been linked to frequent mutation and recombination events in the host, which facilitates bacterial adaption in the host. Apart from gene mutation and recombination events, horizontal gene transfer (HGT) also results in genetic diversity in *H. pylori*, thus, making HGT an essential feature of

the bacterium (Fischer et al., 2020). Therefore, *H. pylori* has been reported to be highly recombination-proficient and naturally competent for transformation (Fischer et al., 2020). Three genome islands which are variably present, all of which encode either full or partial T4SSs and, a fourth T4SS within the core genome of *H. pylori* have been reported (Fig 1.10). According to functional characteristics, T4SSs in *H. pylori* have been divided into three main types – conjugative, effector translocators and DNA uptake or release T4SSs (Christie, 2016, Christie et al., 2005).

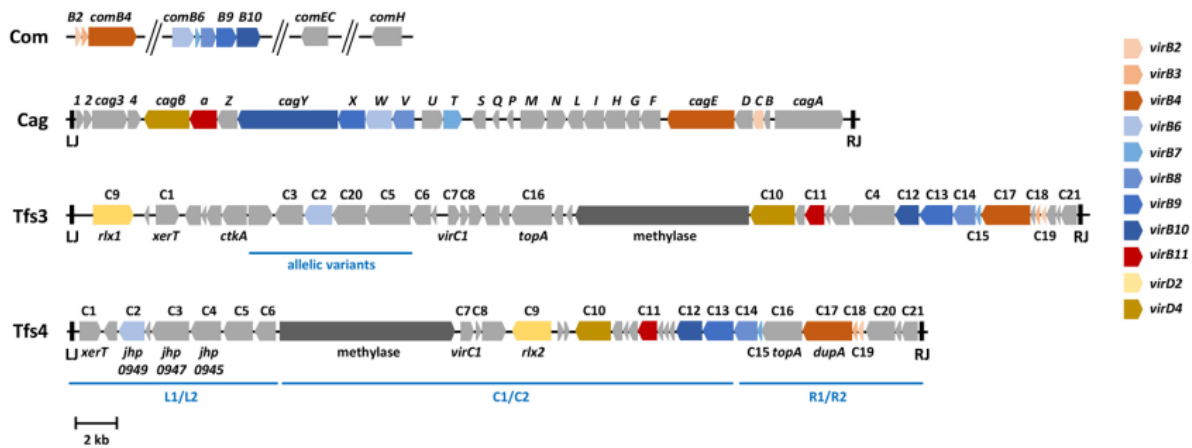


Fig 1.10: Schematic arrangement of genes encoded within the four chromosomally T4SSs gene clusters in *H. pylori*.

The arrows (forward and reverse) show typical direction of genes encoded by *comB*, *cag*, *tfs3* and *tfs4*, with colouring based on sequence similarities with prototype genes from *virB/virD* genes of *A. tumefaciens*. The Cag, Tfs3 and Tfs4 systems are on junctions (Left and Right Junctions – LJ and RJ) on the genome islands while the competence (*comB*) genes are not organised as a contiguous unit, instead made up of two operons – *comB2* to *comB4* and *comB6* to *comB10* with gene homologues to corresponding *vir* genes. The remaining genes of *comB* (*comEC* and *comH*) are located elsewhere in the genome. The *cagPAI* genes are represented with commonly used letters, while genes conserved (C1 to C21) within both *tfs3* and *tfs4* are shown and coloured to demonstrate similarities between both *tfs* ICEs as reported by Delahay and colleagues (Delahay et al., 2018b). Individual genes such as

methylase and cell translocation kinase protein (*ctkA*) may be absent in some *tfs* ICEs, and most *tfs4* genes and some *tfs3* genes (C2, C3 and C5) may occur as distinct allelic variants (shown with blue bar). On *tfs4* ICE, variants are usually grouped in modules (left – L1/L2, Central – C1/C2 and Right – R1/R2) respectively, which can combine to form distinct ICE variants such as L1C1R1, L2C2R2, L1C1R2, etc as reported by Delahay *et al.* (Delahay *et al.*, 2018b). Reproduced from Fischer *et al.*, 2020.

1.4.1 Conjugative T4SSs

The conjugative T4SSs form the largest group of type IV secretion systems and are broadly distributed in bacteria and some species of Archaea (Grohmann *et al.*, 2018, Alvarez-Martinez and Christie, 2009). Plasmids, transposons and integrative and conjugative elements (ICEs) are transferred as single-stranded DNA intermediates into a recipient host cell by conjugation. The process of conjugative DNA translocation increases bacterial adaptability to environmental changes and strengthens its genome plasticity. Thus, conjugative DNA transfer increases the emergence of antimicrobial resistance genes among bacteria implicated in disease outcomes and, thereby increasing bacteria resistance strains (Fronzes *et al.*, 2009a). The conjugation process has been divided into three biochemical reactions – (i) processing of DNA substrates, (ii) substrate recruitment and (iii) substrate translocation (Alvarez-Martinez and Christie, 2009). The DNA substrate processing step has been reported to be conserved in almost all conjugation systems. During DNA processing, initiation is triggered by DNA transfer and replication (*Dtr*) proteins, including a relaxase, which binds to an origin of transfer (*oriT*) sequence to form a DNA-protein complex known

as the relaxosome (Alvarez-Martinez and Christie, 2009). The relaxase nicks transfer DNA at the *oriT* sequence and remains bound to it, subsequently facilitating its recognition by a cognate T4SS. The relaxase is also responsible for piloting the T-DNA through the transfer channel across the membrane. The translocation mechanism of both ICEs and conjugative DNA plasmids is triggered by interactions between a relaxosome and well-conserved substrate receptor known as type IV coupling protein (T4CP) (Alvarez-Martinez and Christie, 2009). T4CP, typically encoded by *VirD4*, is an ATPase that interacts physically with the transfer pathway which is composed of mating pair formation (Mpf) proteins (Gomis-Ruth et al., 2004). Two types of Mpf are present in all types of T4SSs – an ATPase and polytopic membrane protein (Alvarez-Martinez and Christie, 2009). Collectively, the Mpf proteins generate the translocation pathway and the surface pilus in gram-negative bacteria. The surface pilus/filament increases attachment to host target cells (Lawley et al., 2003) whereas surface adhesins are responsible for attachment in gram-positive bacteria (Grohmann et al., 2018, Grohmann et al., 2003).

1.4.2 The Effector T4SS

The effector T4SSs are widely distributed in gram-negative bacteria and have attracted much attention because of their important roles in disease outcomes as reported in many pathogenic bacteria (Nagai and Roy, 2003). Effector T4SSs are responsible for delivering macromolecules and effector proteins to the cytosol of eukaryotic cells, which increases

successful bacterial habitation and survival in host target tissues or cells (Alvarez-Martinez and Christie, 2009). For example, the gram-negative bacteria *H. pylori*, *Legionella pneumophila* and *Brucella suis* employ effector T4SSs to deliver virulence proteins into host cells (Fronzes et al., 2009a). Unlike the conjugative systems, these systems lack the Dtr proteins required to process DNA substrates. Despite this, many of the bacteria that employ these systems are still able to use the T4CP to bind protein substrates. Interestingly, *Bordetella pertussis* and *Brucella sp.* are deficient of T4CPs but instead use the general secretory pathway (GSP) to translocate substrates across the inner membrane (Alvarez-Martinez and Christie, 2009).

1.4.3 ComB T4SS

A good example of bacteria that use this system are *H. pylori* and *Neisseria gonorrhoeae*. In *H. pylori*, the ComB system mediates DNA uptake from the extracellular milieu, while DNA is released into the milieu in *N. gonorrhoea* (Fronzes et al., 2009a). *H. pylori* is known to be naturally competent for genetic modification and transformation, similar to other Gram-negative bacteria such as *Neisseria gonorrhoea* (Stingl et al., 2010). In *H. pylori*, the ComB system is a minimal T4SS that mediates the uptake of DNA from the environment (Karnholz et al., 2006). The ComB T4SS comprises of homologues of seven *A. tumefaciens* T4SS proteins namely, VirB2, VirB3, VirB4, VirB7, VirB8, VirB9 and VirB10 which extend across both inner and outer membranes in *H. pylori* (Stingl et al., 2010). The *comB4* encodes the putative energetic ATPase, while

ComB8 – ComB10 are important for the bacterial transformation competence, and the ComB7, though has no function in bacterial competence, assists in stabilizing the complex of the ComB proteins (Stingl et al., 2010). The transformation process is complex and includes uptake and binding DNA.

1.5 Integrative and Conjugative Elements (ICEs)

ICEs are present in both Gram-negative and Gram-positive bacteria ranging in size from between 18 to over 600Kbp with majority carrying cargo of adaptive genes (Johnson and Grossman, 2015, Burrus, 2017), that initiated their discovery (Johnson and Grossman, 2015). Like conjugative plasmids and bacteriophages, ICEs can integrate within and excise from genomes (Johnson and Grossman, 2015, Burrus and Waldor, 2004) to process conjugation of DNA via the XerT recombinase and transfer mechanism through VirD2 relaxase (Delahay et al., 2018b). Early reports of ICE's function indicated them to be vectors for transferring genes responsible for antimicrobial resistance, but recent evidence has shown that ICEs often carry a cargo of virulence determinants into host cells (Delahay et al., 2018b, Burrus and Waldor, 2004). Thus, ICEs contribute to shaping bacterial genomes by promoting variability in pathogenic bacterial strains and species including *H. pylori* (Delahay et al., 2018b).

ICEs are typically mosaic with modular structure (Johnson and Grossman, 2015). Modules present in ICEs are functional, with genes of

similar function generally arranged together. The regulatory process controlling the gene expression of the ICEs have been reported to be similar to those present in phages and conjugative plasmids (Johnson and Grossman, 2015). In addition, the mechanisms controlling ICEs integration and excision are similar to those found in transposons and viruses, and the conjugation system used by the ICEs (T4SS) is homologous to those present in conjugative plasmids. Studies of conjugative plasmids such as F plasmids from *E. coli* and Ti plasmid from *A. tumefaciens* have provided extensive knowledge about their mechanisms (Johnson and Grossman, 2015). Currently, many important aspects of ICE biology remain unknown despite outstanding breakthroughs in distinct properties of the ICE such as combining the features of transposons and phages due to ICE ability to integrate and excise in the chromosome using either tyrosine or serine recombinases (Botelho and Schulenburg, 2021). Also, ICEs can compete for the same integration site in the chromosome which may result in strong competition for limited attachment sites especially among coinfecting ICEs. Table 1.3 below highlights specific features associated with ICEs and conjugative plasmids.

Table 1.3: Specific features typical to ICEs and conjugative plasmids

Features	ICEs	Conjugative plasmids
Differences		
Location	Integrated in the chromosome	Extrachromosomal
Signature modules	Integration / Excision	Replication
GC content (in comparison with host genome)	Closer	More distinct
Size	Less variable	More variable
Density of DNA repeats	Lower	Higher
Segregational loss	Immune	susceptible
Similarities		
Type of mobility	Intercellular	
Mobility mechanism	Conjugation	
Shared modules	Maintenance, conjugative transfer	
Transfer system	Type IV secretion system	

Reproduced from Botelho and Schulenburg, 2021.

1.5.1 The life cycle of the ICEs

ICEs generally are integrated into the host chromosome and the genes capable of conjugation are not expressed under normal conditions. But upon induction by cellular factors, the ICEs excise from the host chromosome and form circularised double stranded DNA (dsDNA) (Johnson and Grossman, 2015). After circularisation, several encoded genes of the ICE assemble into a mating pore which is capable of translocating the ICE DNA. Other ICE-encoded proteins are capable of

recognising the origin of transfer (*oriT*) and thus, processing the ICE DNA to form a single stranded DNA-protein complex which is linear (Johnson and Grossman, 2015). This linear ssDNA-protein complex is called the transfer DNA (T-DNA) which later translocates into the recipient host cell via the mating pore machinery. Upon arrival in the recipient host, the ICE is capable of re-circularising to become dsDNA which later recombines into the recipient host chromosome through recombinases. Integration of many ICEs occur at specific sites in the chromosome but, some ICEs are capable of integrating at different positions in the chromosome (Johnson and Grossman, 2015). Nevertheless, for the maintenance of ICEs in their original state, they must be integrated back into the chromosomes of their host.

1.5.2 ICE integration and excision

ICEs can integrate into the chromosome using the ICE-encoded recombinases (also called integrases). Both processes, integration and excision require recombinases which are often homologous to the phage integrases. Many ICEs are capable of integrating at specific sites on the bacterial chromosome known as the attachment site (*attB*) which is majorly located in a tRNA gene of the ICEs (Johnson and Grossman, 2015). The attachment site on the ICE is usually called *attICE* or *attP* (*attP* was given based on the phage classification) (Johnson and Grossman, 2015). The recombinases present in most ICEs are either tyrosine recombinases or serine recombinases. The reactions catalysed by both tyrosine and serine recombinases are slightly different, but they

both confer the same function on the ICE. The main function of both tyrosine and serine recombinases is to mediate site-specific integration of the ICE between dsDNA at short DNA sequences. If similar short sequences are present in the circularised ICE and the host chromosome, integrated ICE will be flanked by two attachment sites, one to the left and the other to the right (*attL* and *attR*). In addition to the ICE recombinases, a directional factor called recombination directionality factor (RDF) is also needed for ICE excision (Johnson and Grossman, 2015).

The specificity of ICEs to integration sites are generally variable. For example, many ICEs target a single *att* site on the host chromosome similar to *att* site in the ICE. Alternatively, if the *att* site is not present, the ICE can target alternative sites at reduced efficiency. Some ICEs are capable of inserting at multiple *att* sites in the chromosome such as the CTnDOT ICE of *Bacteroides* spp, while integration by others such as Tn916 ICE from *Enterococcus faecalis* is not site-specific (Johnson and Grossman, 2015).

1.5.3 Identification of Tfs ICEs and variant subtype of Tfs4 ICE

1.5.3.1 *tfs3* and *tfs4* ICEs

The genome of *H. pylori* is highly variable in terms of gene sequence and content (Delahay et al., 2018b). The gene clusters are broadly present inside three genomic islands – *cag* pathogenicity island, *tfs3* and *tfs4* (*tfs3* is called “type IV secretion system 3”, as it was the 3rd T4SS cluster and, *tfs4* is “type IV secretion system 4”, as it was the 4th T4SS cluster

reported in *H. pylori*) (Delahay et al., 2018b). The latter two are ICEs recently identified in the chromosome of some *H. pylori* strains. Before the discovery of *tfs* ICEs in *H. pylori*, early studies using genome sequence analysis of *H. pylori* (strain 26695) reported two variable regions called the plasticity zones (PZ). Located within these zones are the genes specific for the strain as reported by Tomb and colleagues (Tomb et al., 1997). Subsequent genome analyses revealed variability in PZs between different strains of *H. pylori*, differentiating them with respect to location in the genome, presence, and gene content (Alm et al., 1999). Additionally, the PZs were reported to have low G + C content in comparison to the rest of the genome (Tomb et al., 1997). This provided insight that PZs have been acquired through horizontal gene transfer (Tomb et al., 1997).

Recent studies have identified substantial phylogeographical diversity and mosaicism of *tfs* ICEs from global *H. pylori* strains which may explain different disease outcomes in some infected patient populations (Delahay et al., 2018b, Phuc et al., 2021a). Both *tfs* ICEs are present in all seven geographically and genetically diverse *H. pylori* populations (hpAfrica1, hpAfrica2, hpNEAfrica, hpEurope, hpAsia2, hpEAsia and hpSahul (Delahay et al., 2018b) except for a notable absence of *tfs3* in the ancestral hpAfrica2 population, which also lacks the *cagPAI*. Both *tfs* ICEs contain genes that encode complete T4SSs homologous to the VirB/D T4SS in *A. tumefaciens* and may coexist in a single strain of *H. pylori*. Within the gene clusters of both *tfs3* and *tfs4* have been reported

21 putative conserved genes (Delahay et al., 2018b) (section 1.4 above, Fig 1.10). In comparison, the *tfs4* ICEs have multiple variants broadly contained and distributed across the three modules (left, central and right modules) while clusters of variable genes are contained within the left segment of *tfs3* ICE (Delahay et al., 2018b). Often, one or both *tfs* ICE are fragmented in different strains of *H. pylori*, which may result in loss of function of the ICE. Fragmentation is apparently higher in strains from Europe (hpEurope) but reduced in strains from African origin (hpAfrica) (Delahay et al., 2018b). Thus, the *tfs* ICE with these *H. pylori* descents may be unstable.

Tfs3, named as the third putative T4SS to be discovered in *H. pylori*, was first identified as a 16 kb element in strain PeCan18B isolated from a Peruvian gastric cancer patient (Kersulyte et al., 2003). More recently, full-length *tfs3* has been revealed to be a 37.5 - 46 kb element (Alandiyjany et al., 2017). Although not present in all strains of *H. pylori*, *tfs3* has been implicated in disease outcomes possibly through stimulation of pro-inflammatory responses in some populations. A recent study by (Alandiyjany et al., 2017) has revealed that the *tfs3* ICE encodes the variable host interacting cell translocation kinase protein (CtkA). Several studies have reported *ctkA* as a marker for gastric cancer development in some populations and CtkA as a pro-inflammatory protein that triggers cytokine secretion from both host immune and epithelial cells, which may contribute to gastric inflammation. (Alandiyjany et al., 2017) also reported involvement of the *tfs3* T4SS in

mediating the pro-inflammatory activity of CtkA. These findings provide interesting insight into the potential function(s) of the *tfs3* T4SS. The C-terminal of CtkA seems to be important for secretion and is conserved in several other variable *tfs3* proteins – PZ39 and Fic. These could therefore be secretion substrates. PZ39 is most common in strains and interestingly is encoded adjacent to and overlapping a small protein PZ38 that could be hypothesised to be a chaperone.

The majority of genes encoded within the cluster of *tfs4* are like those of the *tfs3* cluster. However, in *tfs4* the T4SS *virB4* gene has been reported as a disease marker for duodenal ulcer disease in some populations and referred to as duodenal ulcer promoting gene A (*dupA*) (Lu et al., 2005). *dupA* gene encoded by both *jhp0917* and *jhp0918* in *H. pylori* infected patients from East Asia and North America contributes to the development of duodenal ulcer among 42% of the infected patients (Lu et al., 2005). Their findings also reported that *dupA* stimulates increased neutrophil infiltration and IL-8 concentration in gastric epithelial cells (GECs), which then stimulates pro-inflammatory cytokines secretion in GECs. Also encoded by both *tfs* ICE clusters is a XerD family tyrosine recombinase which may function in excision to generate extrachromosomal ICE circular intermediates (CIs) which can be transferred to a recipient bacterial cell by conjugative transfer mediated by the Tfs type IV secretion systems (Fischer et al., 2010).

As both ICEs are often present in a partial state in many strains and *tfs4* present as one of several variant or hybrid forms, it may be the case that many ICE T4SSs are non-functional, limiting the ability for ICE transfer and other activities such as likely CtkA effector secretion. However, as all types of *tfs* ICEs encode many similar proteins, there is a possibility that proteins from different *tfs* systems which have a conserved function, such as T4SS assembly, may be sufficiently flexible in their interactions to generate functionally active hybrid systems.

1.5.4 Geographical distribution of Tfs ICEs

Early studies of *tfs* ICEs reported ICE clusters in the few available *H. pylori* strains with complete genome sequences. Other *H. pylori* strains with draft genome sequences deposited online have been overlooked over the years due to difficulties in establishing a full representation of larger genomic segments (Delahay et al., 2018b). Additionally, the fact that *tfs* ICEs are also fragmented and one *tfs* ICE may coexist with another in a single strain confers additional challenges of data mining in the draft genome sequences. A recent study used a robust bioinformatics workflow to extract and precisely define *tfs* ICE populations in a global collection of *H. pylori* strains (221 draft and complete *H. pylori* genome sequences). The study (Delahay et al., 2018b) reported a comprehensive global prevalence of *tfs* ICEs in the seven phylogeographical populations of *H. pylori*. Tfs ICE clusters in a global subset of *H. pylori* strains identified 21 putative genes conserved between both *tfs* ICEs (*tfs3* and *tfs4*) (Delahay et al., 2018b). Additionally,

tfs4 ICE cluster of genes was highly present either as complete in 96 or fragmented in 74 global *H. pylori* strains. In contrast, *tfs3* was absent in 60 strains and both *tfs* ICEs were present in 18 global strains (Delahay et al., 2018b). An apparent increase in fragmentation was reported in strains of European origin, while strains of African descent (HpAfrica1) displayed lesser level of fragmentation (Delahay et al., 2018b). Coexistence of *tfs3* and *tfs4* ICEs was reported to be less common in strains from East Asia descent (HpEAsia) (Delahay et al., 2018b).

1.5.5 Modular classification of Tfs ICEs

All ICEs generally have unique mobile units called modules used for maintenance, conjugative transfer, excision, and integration into the host genome (Botelho and Schulenburg, 2021). Both *tfs* ICEs (*tfs3* and *tfs4*) also have distinct modular classification. *Tfs4* ICE and its variant type have the L-C-R (Left-Central-Right) classification (Fig 1.11) while *tfs3* ICE has the L-R representation (Delahay et al., 2018b). The L-C-R modules have been reported to be capable of interchanging to generate a particular *tfs4* type. In correlation to the three types of *tfs4* ICE (ICE*tfs4a-c*) previously reported using comparative analysis, clustering of *tfs4* genes by (Delahay et al., 2018b) in the global *H. pylori* strains show *tfs4* ICE gene content similarly identifies the three types in addition to several additional novel types (Delahay et al., 2018b). The identified types show orthologous gene complements represented as L2-C1-R1 for ICE*tfs4a*, L1-C1-R1 for ICE*tfs4b* and L1/L2-C2-R2 for ICE*tfs4c*. Further analysis of the L1/L2, C1/C2 and R1/R2 revealed that the principal modular representation of *tfs4* ICE in the global *H. pylori* population is

L1-C1-R1 (Delahay et al., 2018b) (Fig 1.11). Another modular configuration (L2-C2-R2) yet to be described although previously reported, was present in nine *H. pylori* strains. Additionally, other modular representations, where genes encoded in the R1 segment appear truncated and/or fragmented, were also discovered. Therefore, both modular representations L1-C1-R1 and L2-C2-R2 (predominantly present in *H. pylori* population), were used to represent the two predominant *tfs4* types (*tfs4/tfs4* variant) in the *H. pylori* population (Delahay et al., 2018b). The remnants of *tfs4* modular representations were classified as hybrid forms that resulted from either genomic recombination and rearrangement between two ICEs in the same strain or partial integration of heterologous genes of the ICE (Delahay et al., 2018b). Therefore, the ability of the two *tfs4* types (*tfs4* and *tfs4* variant) to form a functional hybrid type remain unknown and to date, never investigated.

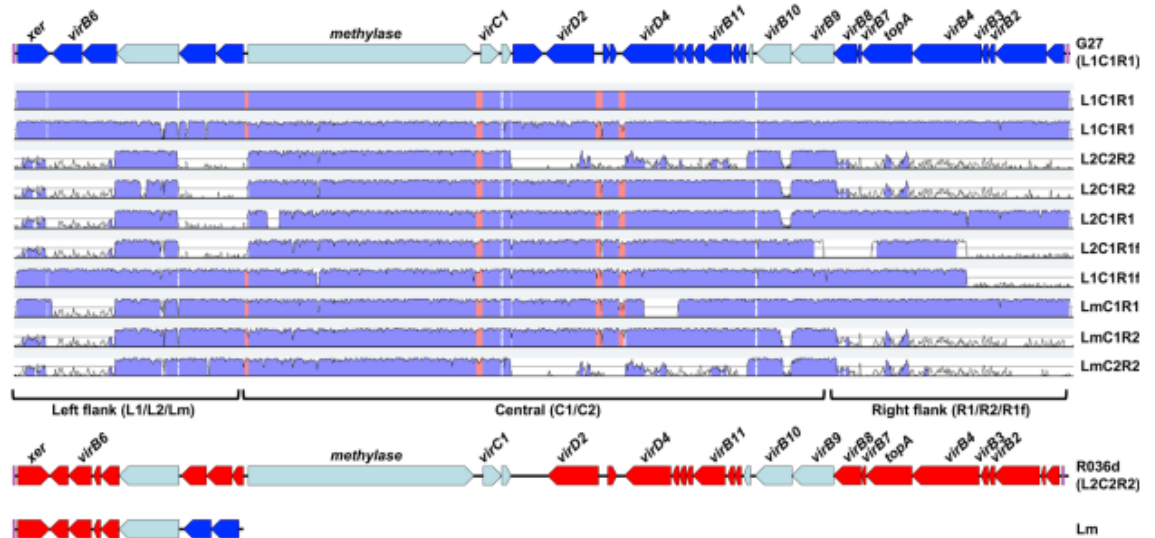


Fig 1.11: Global pairwise alignment of distinct *tfs4* ICE and modular representation.

H. pylori strain G27 is used as reference strain with *vir* genes *tfs4* cluster (top) with modular subtype L1C1R1. *vir* genes coloured blue are L1, C1 and R1 module gene complements whereas orthologous genes characteristic of L2C2R2 modular representation from strain R036d are coloured red. Reproduced from Delahay et al., 2018b.

1.5.6 Modules and hybrid functions of Tfs ICEs

ICEs are common in different bacterial strains with varying advantages to their host. A common feature to ICEs is the presence of modular structure where genes with similar function are grouped together (Bellanger et al., 2014). The function(s) of a particular module is relevant to different activities carried out by the ICEs. For example, ICE with conjugative ability encoded in its modular subtype mating pair T4SS genes, transfer ability encodes VirD2 and VirD4 proteins and recombination encodes relaxase and integrase functions (Delahay et al., 2018b).

As earlier stated, both *tfs* ICEs have modular subtypes (Fig 1.9) which are prominent in *tfs4* and represented as L-C-R. The L- subtype encodes for recombination function mediated by the *xer* gene, C and R encodes for transfer and mating pair function (Delahay et al., 2018b, Grove et al., 2013). Contrary to *tfs4* modular representation, *tfs3* ICE is divided into two by its right and left modules respectively. The L module is mostly intact in both *tfs* ICEs, but they are presented with fragmentation of the module encoded with mating pair function. It has been speculated that the presence of the L module in majority of global *H. pylori* strains may confer a specific benefit via the encoded proteins, which may be independent of the *tfs3* ICE T4SS. Therefore, it was suggested that *tfs* L-C-R modules can encode for ICE-dependent and ICE-independent activities (Delahay et al., 2018b). The *H. pylori* strains from African descent (HpAfrica1) have high prevalence of the L2C1R1f *tfs4* ICE type, a modular subtype which lacks several key T4SS genes in the R1 region.

Inter-strain transfer of *tfs* ICEs has been reported in global strains of *H. pylori*. For example, *tfs* transfer was reported in strains of African and European origins. This transfer may provide possibility for functional complementation of *tfs3* and *tfs4* ICEs (Delahay et al., 2018b). This possibility remains unknown. Also, inter-population of *tfs* ICEs heterologous genes may also affect the virulence outcomes on the host organism. Understanding how the latter might be possible remains to be elucidated.

1.5.7 Disease association of Tfs ICEs

Both *tfs* ICE-encoded genes have been implicated in disease outcome in some populations infected with *H. pylori*. In particular, the genes encoded by *tfs4* R1 module and L1 module (*dupA*, *jhp0947* and *jhp0949*) or gene encoded by both L1/L2 modules (*jhp0945*) and *tfs3* L module encoded gene (*ctkA*) have all been implicated in disease outcomes. In the context of *tfs4*, it is unclear whether the association of these genes with disease progression is dependent on a particular module, the activity of individual proteins or the function of ICE T4SSs (Delahay et al., 2018b). The study reported that both L and R modules are prevalent in global *H. pylori* populations in reference to *tfs4* and, often isolated from other modules from the ICE. Also, they reported that truncated modular representation (L2C1R1f) is highly conserved in HpAfrica1. Their report suggested that both L and R modules of *tfs4* ICE may be functionally independent of *tfs4* T4SS. The notion supported a study where they reported a low incidence of disease outcomes in *H. pylori* infected population with strains having L2C1R1f modular subtype and highly virulent *cagPAI* genes (*cagA* and *vacA*) (Delahay et al., 2018b). Furthermore, R1 module that encodes *dupA* gene is considerably absent from *H. pylori* strains having the L2C1R1f modular representation, which may suggest that *dupA* alone cannot be used as a disease marker for strains that harbour full cluster of *tfs4* T4SS proteins.

1.6 Rationale, Hypotheses and Aims

Rationale

Based on homology of genes encoded on *tfs* ICEs (*tfs3* and *tfs4*), both encode full complement of Vir- homologous T4SS proteins. To date, there is very little experimental work to fully support the role of these proteins in T4SS assembly due to their dynamic nature as membrane-bound proteins. Previous work done in the group by Alandiyjany (Alandiyjany, 2015) established a network of interactions between *tfs4* VirB proteins (*tfs4B8/tfs4B8*, *tfs4B11/tfs4B11*, *tfs4B4/tfs4D4* and *tfs4B4/tfs4B8*). The studies concluded that these interactions are similar to known networks of interactions (section 1.3.3, Fig 1.8) reported by Fronzes and colleagues (Fronzes et al., 2009a) and, they are important in the assembly of the T4SSs. Furthermore, the study by Alandiyjany (Alandiyjany, 2015) also selected the *tfs4* VirB8 variant protein (*tfs4B8v*) for interaction with *tfs4B8* and *tfs4B4* respectively. The study reported strong interactions between *tfs4B8v/tfs4B4*, *tfs4B8v/tfs4B4* and *tfs4Bv/tfs4B8*. The conclusions from this work suggested that *tfs4B8v* could conceivably functionally complement for *tfs4B8* in the assembly of T4SS.

Virulence genes encoded in *tfs3* and *tfs4* ICEs have been implicated in disease outcomes. Although, genes encoded on different modules, for example R1 module of *tfs4* ICE (*dupA+*) and L module of *tfs3* ICE (*ctkA*) (section 1.7.4.5) have been implicated as markers of disease

progression, the associations between disease outcomes and *tfs* ICEs remain poorly understood.

Therefore, the study by Alandiyjany (Alandiyjany, 2015) set the pace to investigate the potential for cross complementation integrations between *tfs* ICEs encoded proteins especially with proteins encoded on *tfs4* ICE and its variant form. The work in this thesis aimed to explore the potential for heterologous protein-protein interactions of encoded VirB proteins on *tfs3* and *tfs4* ICEs to establish potential functional complementation between the proteins. In addition, bioinformatics approaches were be used to search for full complements of *tfs* ICEs genes in clinical isolates to establish correlation with disease status.

Hypotheses

Hypothesis 1: Similar components of *tfs* ICEs (*tfs3* and *tfs4*) are capable of interacting with one another to enable a functional T4SS Therefore, it was hypothesised that interactions may be possible between *tfs3* and *tfs4* VirB-homologous proteins and with the *tfs4* variant form.

Hypothesis 2: It was hypothesised that a bioinformatics approach could be used to search for and define complete sets of *tfs* T4SS genes and predict their impact on disease developments in patients.

Therefore, the aims of this thesis were set out.

Aim 1: To investigate interactions between proteins encoded by *tfs3*, *tfs4* and *tfs4* variant types important for assembly and function:

Use the yeast two-hybrid assay to:

- a. Validate the *tfs4* VirB8 interactions already reported in the group.
- b. Investigate heterologous protein-protein interactions of *tfs* ICE proteins that mediate ICE T4SS assembly. Here, I aimed to investigate the potential for cross complementation between *tfs3* and *tfs4* encoded proteins .
- c. Investigate the functional activity of *tfs* ICEs energetic proteins (VirB4, VirB11 and VirD4) by identifying *tfs4* variant proteins they interact with. This aimed to establish if any of the energetics from any of the *tfs*'s ICEs can be used in any system (for example if *tfs4B4* can functionally complement for *tfs4B4v*), or whether they are strictly system specific.
- d. Investigate interactions of *tfs4* VirD2 protein with VirB proteins of *tfs3*, *tfs4* variant aimed to establish if any ICE (*tfs3* or *tfs4*) can use any *tfs* T4SS for transfer of substrate.
- e. Investigate interactions of orphan genomic protein (gB11) aimed to establish if it can complement for loss of function of any *tfs* T4SS ATPases.

Then, I aimed to validate novel interactions identified using a secondary approach:

- a. recombinant protein purification and characterisation. I aimed to clone novel VirB interactions proteins into an expression vector

and perform a co-expression pull-down assay after expression, solubility and purification have been confirmed. Characterisation of purified protein will be done using sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE), protein staining with Coomassie blue and Western Blot analysis of co-expressed proteins.

- b. *in-silico* bioinformatics analysis to validate reported interactions from yeast two-hybrid study.

Aim 2: To explore the association of *tfs* T4SS genes in *H. pylori* isolated from patients with disease status. This section of the thesis will use genomic analysis using the ABRicate software to search for full complement of *tfs* T4SS genes in isolates and correlation between cluster of genes of *tfs* ICEs and disease status will be established.

CHAPTER TWO

PROTEIN-PROTEIN INTERACTIONS
BETWEEN T4SS PROTEINS OF Tfs3,
Tfs4 AND Tfs4 VARIANT TYPE

2.0 Introduction

Several studies of the bacterium *H. pylori* and the genome analysis of the bacterium have revealed a subset of variable genes present in different strains of *H. pylori* (Romo-González et al., 2009, Muñoz-Ramírez et al., 2017, Farnbacher et al., 2010, Baltrus et al., 2009). Subsequently, some of these variable genes have been implicated in disease status in some *H. pylori* infected populations (Alandiyjany et al., 2017, Delahay et al., 2018b). For example, one of the most extensively studied, *dupA*, has been reported to be associated with duodenal ulcer development. The *dupA* gene was reported to be homologous to the T4SS virulence *virB4* gene of *A. tumefaciens* (de Lima Silva et al., 2021). Until now, the function of the *tfs* ICEs (*tfs3* and *tfs4*) remains poorly understood in *H. pylori*. Nevertheless, both *tfs* ICEs encode the T4SS proteins *virB1-virB11* and coupling protein *virD4*, similar to the T4SS VirB/D in the prototype organism *A. tumefaciens*. Recent studies have established that *tfs3* and *tfs4* T4SS can harbour a complete set of genes, such as *virD4*, *virB4* and *virB11* required for T4SS structural assembly, as well as genes involved in DNA processing and transfer (*virD2*) (Fischer et al., 2020, Alandiyjany et al., 2017).

Protein-protein interactions (PPI), a network of protein interconnections upon which many biological activities are regulated, have been established in many organisms by studying interaction partners (Chakraborty et al., 2021). In Gram-negative bacteria including *H. pylori*, several studies have reported protein-protein interactions (Terradot and Waksman, 2011, Busler et al., 2006, Kutter et al., 2008) using biological

techniques such as pull-down assay, co-immunoprecipitation, and the yeast two-hybrid assay (Y2H).

The Y2H method is a popular and power technique used for the analysis of protein-protein interactions (PPIs) *in vivo* (Paiano et al., 2019). The method developed by Fields and Song in 1989 (Paiano et al., 2019) uses two candidate proteins, where one is the “bait” and the other “prey”. The bait protein is fused to the DNA binding domain (DBD) and the prey is fused to the activation domain (AD) of a transcription activator (Erffelinck et al., 2018). Upon interaction of the paired proteins, the transcription factor drives the expression of the reporter genes in the yeast strain which can later be scored by the ability of the yeast cells to grow on selection plates (Erffelinck et al., 2018). In this study, the DNA binding domain (DB) of the yeast (*Saccharomyces cerevisiae* strain PJ69-4A) *GAL4* transcription activator was encoded by plasmid pBGT9 and, the activation domain of the yeast strain was encoded by plasmid pGAD424. The yeast strain used in this study encodes for three reporter genes namely *HIS3*, *ADE2* and *lacZ*.

In comparison to other biochemical procedures such as pull-down assays and co-immunoprecipitation, the Y2H system is simple to use, inexpensive and able to detect weak interactions among paired proteins. On the contrary, other biochemical techniques may require highly purified and good quality antibodies before interactions can be detected in paired proteins.

Several studies (Busler et al., 2006, Ward et al., 2002, Yuan et al., 2005, Iqbal et al., 2021, Zupan et al., 2007) have demonstrated protein-protein interactions between the VirB/D homologues using the yeast two-hybrid method. Several interaction networks that include the T4SS VirB/D homologues from *A. tumefaciens* have been reported according to Fronzes *et al.*, (Fronzes et al., 2009a). Fig 2.1 below from Fronzes and colleagues shows the interaction network between the VirB/D homologues. The VirB8, VirB9, VirB10 and the ATPases (VirB4, VirB11 and VirD4), classified as structural proteins of the T4SS, demonstrate multiple interactions. For example, a positive interaction was demonstrated between VirB8-VirB1, VirB8-VirB4, VirB8-VirB5, VirB8-VirB8, VirB8-VirB9, VirB8-VirB10, and VirB8-VirB11 (Fronzes et al., 2009a). In addition, VirB11 demonstrated positive interactions with VirB4, VirB8, VirB9, VirB10, VirB11 and VirD4 (Fronzes et al., 2009a). This model of established interaction network will be used as the basis for subsequent modelling of interactions between T4SS proteins in *H. pylori* identified in this study. Furthermore, VirD2 is a known secretion substrate present in both *tfs3* and *tfs4* ICEs. As a secretion substrate, VirD2 has been reported to be tightly linked with T-DNA and aid in its direction towards the plant in *Agrobacterium tumefaciens*. Both amino and carboxyl terminals of VirD2 are important targets for proteins within the secretory pathway, a major reason for its classification as a secretory substrate.

The nomenclature adopted for proteins of *tfs3*, *tfs4*, *tfs4v* and the orphan genomic VirB11 included in this study are given as: *tfs3VirB4*, *tfs3VirB8*,

tfs3VirB11, *tfs4VirB4*, *tfs4VirB8*, *tfs4VirB9*, *tfs4VirB10*, *tfs4VirB11*, *tfs4VirD2*, *tfs4VirD4*, *tfs4VirB4v*, *tfs4VirB8v*, *tfs4VirB11v* and the orphan gB11 respectively. The orphan genes of different bacteria have been reported to enable their adaptation to specific environments.

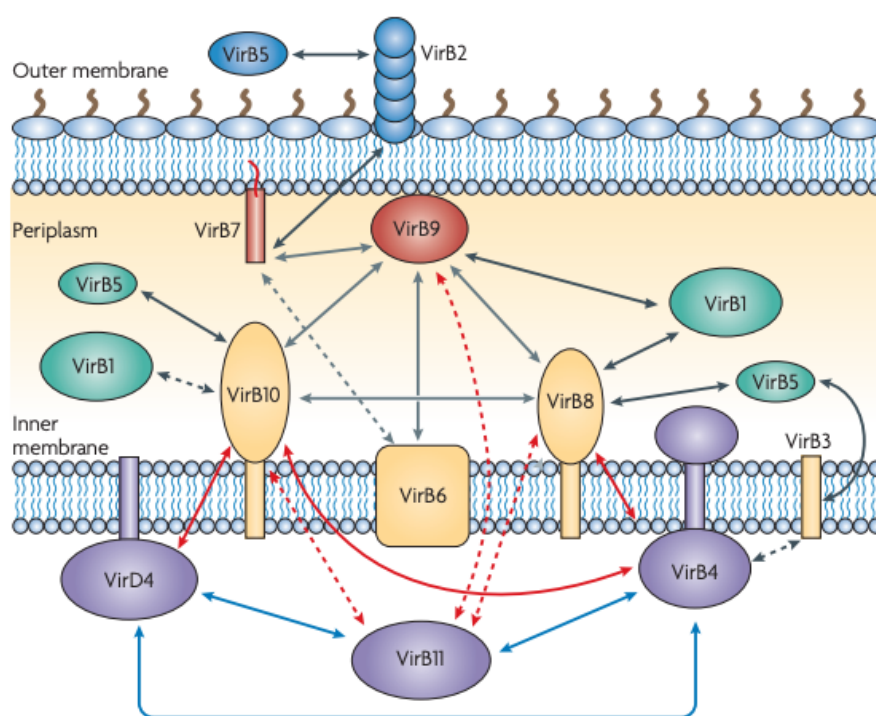


Fig. 2.1: Schematic representation of interaction network of VirB/D homologues protein from *A. tumefaciens*.

Cytoplasmic ATPases are shown in purple, IMPs shown in blue, PPs in green, OMPs in red and Eps in blue. Blue arrows indicate interactions between ATPases, reds arrows indicate interactions between ATPases and non-ATPases, grey arrows indicate interactions between core T4SS components, black arrows indicate other interactions and dotted arrows indicate possible interactions. Reproduced from Fronzes et al., 2009a.

2.1 Rationale

A previous study in our group by Alandijjany (Alandijjany, 2015) established a potential for cross complementation interactions between

proteins encoded on *tfs3*, *tfs4* and *tfs4* variant type. The study reported a network of interactions between *tfs4* VirB proteins - *tfs4B8/tfs4B8*, *tfs4B11/tfs4B11*, *tfs4B4/tfs4D4* and *tfs4B4/tfs4B8* and selected only the *tfs4* VirB8 variant protein which interacted with *tfs4* VirB4 and *tfs4* VirB8 respectively. The study suggested that interactions between *tfs4* VirB8 variant and *tfs4* VirB8 could potentially provide functional complementation between the proteins in the assembly of a hybrid/variant Tfs4 T4SS. The study only reported interactions of *tfs4* VirB8v and not possible interactions of other *tfs4* variant VirB proteins such as the ATPases (VirB4v, VirB11v) with other T4SS components. Whether interaction(s) is or are possible with other VirB proteins in the case of *tfs4* where variant subtype is apparent, remains unknown. Also, as all types of *tfs* ICEs encode many similar proteins, there is a possibility that proteins from different *tfs* systems which have a conserved function, such as T4SS assembly, may be sufficiently flexible in their interactions to generate functionally active hybrid systems.

Therefore, this chapter of the thesis aimed to use the high throughput yeast two-hybrid assay to investigate interactions between proteins encoded by *tfs3*, *tfs4* and *tfs4* variant type to achieve the objectives – to first validate the VirB8 interactions reported previously in the group, then, investigate functional activity of the *tfs* ICEs that mediate ICE T4SS assembly by investigating potential for cross-complementation between proteins encoded by *tfs3*, *tfs4* and *tfs4* variant type. Also, this study aimed to investigate the functional activity of the orphan VirB11 with encoded proteins of *tfs3*, *tfs4* and *tfs4* variant to explore a possible

functional link with *tfs* ICEs in the genome of *H. pylori*. Furthermore, this study aimed to use secondary approaches such as recombinant protein purification and *in silico* bioinformatics analysis to validate novel interactions.

2.2 Materials and Methods

2.2.1 Yeast strains, bacterial strains, growth media and conditions

The yeast strain used in this study, *Saccharomyces cerevisiae* (PJ69-4A) (James et al., 1996) was cultured in either Yeast Peptone Dextrose or Yeast Minimal Medium (YPD or YMM respectively) supplemented with a final concentration of 2% glucose (Fischer-Scientific). YPD was prepared in 200 ml of distilled water with 2 g of yeast extract (Oxoid) and 4 g bacteriological peptone (Oxoid), adjusted to pH 5.8.

For YMM, 1.34 g yeast nitrogen base (Difco 0919-15) was added with 4 g of Agar technical No 3. (Oxoid) to 200 ml distilled water. Supplementary solutions were added to YMM minimal media as appropriate to make a particular selective media. These included L-methionine (Met), Uracil (Ura), L-Histidine (His) and Adenine (Ade) - all at 20 µg/mL (Sigma-Aldrich).

E. coli XL1-Blue (Stratagene) used for general cloning was grown in Lysogeny Broth (LB) or on LB-agar supplemented with 100 µg/mL ampicillin (Sigma-Aldrich) and incubated overnight (16-18 hours) at 37°C with shaking at 250 g.

Genomic DNA of *H. pylori* strains P12, 247 and J166 was provided by Dr Rob Delahay together with available VirB constructs from previous work.

Table 2.1 and 2.2 below show bacterial, yeast and plasmid strains used in this study.

Table 2.1: Bacterial and Yeast strains used in this study

Strain/Vector	Description	Cloned genes or genotype	Source
P12	Clinical isolate <i>H. pylori</i>	<i>tfs4 virB11, tfs3 virB11 and gB11,</i>	Provided by Dr Rob Delahay
247		<i>tfs4 virB11v</i>	
XL-1 Blue	<i>E. coli</i> used to make competent cells for transformation	<i>recA1 endA1 gyrA96 thi-1 hsdR17 supE4 relA1 lac [F' proAB lacI^qZΔM15 Tn10 (Tet^r)]</i>	Stratagene, provided by Dr Rob Delahay
<i>E. coli</i> Top10	High Efficiency competent cells (> 10 ⁹ cfu/μg)	recombinant products of <i>tfs4 virB11, tfs3 virB11 and gB11</i>	Novagene, provided by Dr Rob Delahay
<i>E. coli</i> soluble BL21	Optimised <i>E. coli</i> strain for protein expression in soluble form.	recombinant products of <i>tfs4 virB11, tfs3 virB11 and gB11</i>	provided by Dr Rob Delahay
<i>Saccharomyces cerevisiae</i> strain PJ69-4A	(<i>MATa trp1-901 leu2-3 112 ura3-52 his3-200 gal4Δ gal80Δ LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ.</i>	Supplied by Dr Rob Delahay; (James et al., 1996)	<i>Saccharomyces cerevisiae</i> strain PJ69-4A

Table 2.2: Plasmids used in this study.

(Illustrated in Appendix Fig 1 and 3)

Plasmid	Description	Source and Reference
pGAD424 (prey)	<i>oriColE1 ori2μ LEU1 PADH : GAL4 activator domin: (amp¹⁰⁰)</i>	Supplied by Dr R. Delahay; (James et al., 1996)
pGBT9 (Bait)	<i>oriColE1 ori2μ TRP1 PADH : GAL4 binding domin: (amp¹⁰⁰)</i>	
pBAD cMyc/His expression vector	<i>araBAD promoter region, Multiple cloning sites, initiation (ATG), myc epitope, polyhistidine tag, Ampicillin, AraC, pBR322</i>	<i>tfs4 virB11, tfs3 virB11 and gB11</i>

2.2.2 Oligonucleotides and bioinformatics resources

Oligonucleotide primers used in the polymerase chain reactions (PCR) were purchased from Sigma-Genosys and the nucleotide sequences of *tfs3* and *tfs4* genes were obtained from the National Centre for Biotechnology Information (NCBI). The reference *H. pylori* strain P12 was used as the source of genes of *tfs4 virB11*, *tfs3 virB11* and *gB11*, while strain 247 was used for genes of *tfs4 virB11v*. Oligonucleotides used for PCR are listed in Table 2.3. All primers were diluted to 10 pmol μl^{-1} for use in PCR reactions. Sequencing of all constructs generated in this study was done by Geneservice and sequences were assessed to

ensure that no base changes had occurred during PCR amplification. Sequence files and downloaded nucleotide sequences from NCBI were formatted using BioEdit program and relevant alignments were performed using ClustalW. BLAST search of orphan genomic gene (*gB11*) of strain P12 was performed using the NCBI BLAST via the Geneious Prime software (dotmatics) and percentage identity of gB11 with other *H. pylori* strains was determined with multiple alignment available in the software. A list of websites hosting bioinformatics tools used in this study is provided in Table 2.4.

Table 2.3: Oligonucleotide primer sequences used in this study.

Region	Primer Name	Primer Sequence '5 – 3'	PCR Size (bp)
Tfs4 <i>virB11</i>	Tfs4B11EcoF	GAACTGGAATTCATTCTAATCTTTTAAAAAGTTACCTTG	942
	Tfs4B11SalR	GTACTAGTCGACTCATAACATCAACTCCTTTTTAAAT	
Tfs4 <i>virB11v</i>	Tfs4B11vEcoF	GAACTGGAATTC AATGGTCTTTTTAATTCTAAACAAG	983
	Tfs4B11vSalR	GTACTTTCGAC TCATTTCGTATCCCAAAAGTTG	
<i>gB11</i>	gB11EcoF	TTACCGGAATTCGAAACTTTACAAACCCATAGAGTTTTAC	915
	gB11SalR	GTACTAGTCGACTTAAAGCTCTTTAGTCCATAAGACTTC	
Tfs3 <i>virB11</i>	3B11BADNcoF	GAACTGCCATGGGTGACTTAGACAACTCAAAGATTATAGG	945
	3B11BADEcoR	GTACTAGGAATCTAGGCTATCTCTTAAATCTTTTGCT	
Tfs4 <i>virB11</i>	4B11BADNcoF	GAACTGCCATGGGTATTTCTAATCTTTTAAAAAGTTACCTTG	942
	4B11BADEcoR	GTACTAGGAATCTAACATCAACTCCTTTTTAAATTCG	
<i>gB11</i>	gB11BADNcoF	GAACTGCCATGGGTGAAACTTTACAAACCCATAGAGTTTTAC	915
	gB11BADEcoR	GTACTAGGAATCAAGCTCTTTAGTCCATAAGACTTCAG	
pBAD	pBADF	CTATGCCATAGCATTTTTATCC	
	pBADR	CGTTCTGATTTAATCTGTATCAGG).	

Red text shows an EcoRI restriction site, blue text shows a SalI restriction site, magenta shows an NcoI restriction site and green bases correct the frame for expression

Table 2.4: Bioinformatics resources used in this study.

Resources	Use	Web address
National Center for Biotechnology (NCBI)	Source for DNA and protein sequence BLAST Search	https://www.ncbi.nlm.nih.gov/ (National Library of Medicine, 16 February 2022.)
EXPASY Translate	Translate nucleotide (DNA/RNA) sequence to a protein sequence	http://web.expasy.org/translate/ (Expasy, 1 July 2020.)
ClustalW	Sequence alignment	https://www.ebi.ac.uk/Tools/msa/clustalo/ (Clustal Omega, 2022)
Netprimer	Primer design tool	http://www.premierbiosoft.com/netprimer/index.html (PREMIER Biosoft, 1994 - 2022)
New England Biolabs (NEB) Tm tool	Tm calculator tool	http://tmcalculator.neb.com/#/ (New England Biolabs, 2022)
BioEdit	Sequence analysis software	http://www.mbio.ncsu.edu/bioedit/bioedit.html

Geneious Prime	Sequence analysis tools	https://www.geneious.com/prime/ (Dotmatics, 2022)
Max-Planck Institute Bioinformatics Toolkits (MPI)	Sequence analysis tools with databases	https://toolkit.tuebingen.mpg.de/ (Max-Planck Institute, 2022)

2.2.3 Polymerase Chain Reaction – amplification of genomic DNA

Appropriate primers were synthesised from Sigma-Genosys to amplify required genes (*tfs3 virB11*, *tfs4 virB11*, *tfs4 virB11v* and *gB11*). The primers were designed to have restriction enzyme sites for cloning as required (Table 2.3). The PCR mixtures comprised 30.5 µl of molecular biology grade water (Sigma), 10 µl 5x buffer (New England Biolabs), 4 µl dntps (dATP, dTTP, dGTP, dCTP) (Promega) at 2.5 mmol concentration of stock, 2.5 µl of each primer (P1 and P2) at 10 pmol each, 1 µl of template genomic DNA and 0.5 µl of high fidelity Phusion polymerase enzyme (New England Biolabs) was used as required. Typical PCR cycling parameters comprised of 28 cycles of template denaturation at 98°C for 30 seconds, followed by 30 seconds annealing of primers at an appropriate temperature (T_m°C) for the primer pairs, calculated using the NETprimer and NEB T_m tools (Table 2.4), then extension at 72°C for 20 seconds per kb. The PCR was completed with a final extension step at 72°C for 5 minutes to ensure complete synthesis of all strands, before being cooled at 4°C until analysed.

2.2.4 Agarose Gel electrophoresis

Agarose gel electrophoresis (Bio-Rad) was performed to resolve PCR amplicons and plasmids – pGAD424 (*oriColE1 ori2 μ LEU1* PADH: *GAL4* activating domain amp100) and pGBT9 (*oirColE1 ori2 μ TRP1* PADH: *GAL4* binding domain amp100). The gel concentration was prepared to 0.8% w/v agarose in 1x Tris Acetate-EDTA buffer (TAE) (40 mM Tris-acetate, 2 mM EDTA at pH 8.0). Ethidium bromide (500 μ g/ml, Sigma-Andrich) was added to the mix to allow visualisation of DNA under ultraviolet light. Electrophoresis was carried out at 80 volts for 30 minutes. The samples and loading markers (Promega) were loaded appropriately into the gel using 6x loading dye (Promega). DNA products were visualised with the Uvidoc documentation system.

2.2.5 Purification of PCR products

Purification of the PCR products using NucleoSpin Gel and PCR clean-up kit (Macherey-Nagel), following the manufacturer's manual was performed. The same kit was used for purification of DNA extracted from agarose gels. DNA exposure to UV light was short as UV degrades DNA upon long exposure. Therefore, DNA fragments were carefully cut out with a scalpel and quickly transferred into 1.5 ml Eppendorf tube, then melted with hot block at 50°C.

2.2.6 Extraction of Plasmid DNA

The NucleoSpin Plasmid miniprep kit (Macherey-Nagel) was used to extract plasmid DNA pGAD424 and pGBT9 from *E. coli* XL1-Blue according to the manufacturer's procedure.

2.2.7 Restriction Digest of Purified plasmids and PCR products

Different restrictions endonuclease enzymes (New England Biolabs) were used as appropriate to digest PCR products and plasmids. In general, reaction mixes comprised 1.5 µl of each enzyme, 12 µl of plasmid DNA or 15 µl of PCR products, 5 µl of 10x Cut Smart buffer (New England Biolabs) and molecular biology grade water (Sigma) to 50 µl. Reaction mixes were incubated at 37°C in a hot block for 2 hours. The reaction was stopped by addition of the DNA loading buffer (Promega). The digested fragments were separated in agarose gels (Bio-Rad) and products were purified using appropriate NucleoSpin Gel and PCR clean-up kit (Macherey-Nagel) and stored at -20°C.

2.2.8 Ligation of DNA fragments

PCR products were cloned into appropriate yeast two-hybrid vectors (pGAD424 and pGBT9) using T4 DNA ligase (New England Biolabs) as described by Weiss and colleagues (Weiss et al., 1968) to catalyse the ligation of DNA ends by forming a phosphodiester bond between the 3'-OH group and the 5'-OH phosphate group of another end. In general, the ligation mix used an excess of insert DNA mixed with plasmid vector in

a 10 µl standard reaction volume. Vector and insert were premixed and heated to 65°C for 5 mins prior to addition of 1 µl of 10x T4 DNA ligase buffer (New England Biolabs) and 1 µl T4 DNA ligase (New England Biolabs). The ligation mixtures were incubated at 4°C for 2-3 days or 16°C overnight (16-18 hours).

2.2.9 Preparation of *E. coli* competent cells and transformation

The *E. coli* strain XL1-Blue (Stratagene) was used to make competent cells for transformation. An overnight culture of the bacteria was grown in 10 ml of LB broth (Oxoid). 5 ml of the overnight culture was added to 500 ml of LB broth in a conical flask, followed by incubation at 37°C for 2-3 hours until the OD₆₀₀ read between 0.4-0.6. At the desired OD₆₀₀ reading, cells were pelleted at 4,500 x g (Beckman Allegra, X-22R) for 5 minutes at 4°C. All cells were kept on ice. The pelleted cells were resuspended in 1/10th of the original culture with ice-cold TFB1 (Transformation Buffer 1) buffer (100 mM KCL, 50 mM MnCl₂, 30 mM C₂H₃O₂K, 10 mM RbCl, 15% glycerol and pH 5.6-6.2) in two separate falcon tubes, then incubated on ice for 30 minutes. After that, cells were pelleted at 4°C, 4,500 x g (Beckman Allegra, X-22R) for 5 minutes. Finally, the cells were resuspended in 3 ml of ice-cold TFB2 (10 mM MOPS, 10 mM KCl, 75 mM RbCl, 15% glycerol and pH 7.00). Resuspended cells were incubated on ice for 30 minutes, then 100 µl of cells was aliquoted into 1.5 ml Eppendorf tubes and stored at -80°C.

Transformation of plasmids into competent cells was performed as described by Sambrook and Russell 2002. Concisely, the competent

cells were thawed on ice and 100 µl of cells aliquoted into pre-chilled tubes as required. After this, 5 µl of the ligation mix was added to the competent cells and incubated on ice for 30 minutes. The mixture was heat-shocked at 42°C for 45 seconds and immediately cooled on ice for 2 minutes. After that, 900 µl of LB broth (Oxoid) was added and incubated at 37°C for at least 1 hour with shaking (250 rpm). Cells were pelleted after incubation and re-suspended in 100 µl of LB broth, then plated on LB agar plates supplemented with the appropriate antibiotics selection and incubated overnight at 37°C.

2.2.10 Screening and Identification of transformants

2.2.10.1 Colony PCR

Single bacterial colonies from overnight transformation plates were picked with sterile tips into 25 µl PCR reaction mixtures. Single colonies were picked because each colony developed from a single bacterium that has undergone transformation with an antibiotic resistance-encoding plasmid. PCR analysis showing a positive result will indicate that the bacteria harbour a plasmid which contains the desired cloned gene. The colony PCR of amplicons was performed with the following reaction mixtures - 18.5 µl of molecular biology grade water (Sigma), 5 µl 10x buffer, 8 µl dNTPs (Promega) (dATP, dTTP, dGTP, dCTP) at 10 mmol concentration of stock, 5 µl of each primer (P1 and P2) at 10 pmol each, 1 µl 50% glycerol and 0.5 µl of Taq polymerase enzyme (New England Biolabs). Typical colony PCR cycling parameters comprised; initiation

step involved 1 cycle of 94°C for 5 minutes. The subsequent amplification involved 30 cycles of 58°C for 30 seconds, followed by annealing of primers at 55°C for 40 seconds or as appropriate for primer pairs, then extension at 68°C for 1 minute per kb. The PCR was finished with a final step at 68°C for 5 minutes to make sure that all amplicons were synthesised completely, before being cooled.

2.2.10.2 Construct Screening

Individual colonies were incubated onto LB (Oxoid) plates and grown in LB broth overnight containing appropriate antibiotic for selection. Plasmids were recovered using the NucleoSpin Plasmid kit (Machery-Nagel). After, plasmids were digested with appropriate restriction endonuclease enzymes and the resulting fragments were visualised on a UV Uvidoc documentation system following separation through a 0.8% w/v agarose gel (detailed in section 2.2.4).

2.2.11 The Yeast-two hybrid Assay

2.2.11.1 Lithium acetate transformation of *S. cerevisiae*

The transformation procedure as described by Gietz and colleagues (Gietz et al., 1995, Grove et al., 2013) was used to introduce both pGAD424 (*oriColE1 ori2 μ LEU1 PADH: GAL4* activator domain: amp100) and pGBT9 constructs into yeast *S. cerevisiae* PJ69-4A. Initially, yeast was grown overnight (16-18 hours) at 30°C with shaking at 250 rpm to reach a stationary phase (OD₆₀₀ >1.5) in 10 ml yeast complete media

(YPD supplemented with 2% w/v glucose). 1.5 ml of overnight culture was transferred into 25 ml YPD supplemented with 2% glucose. The mixture was incubated at 30°C for 3-4 hours with shaking (250 rpm) until the OD₆₀₀ was recorded at 0.46, then cells were centrifuged at 4000 x g (Allegra 22/X) for 5 minutes, then washed in distilled water and resuspended in 1ml of freshly prepared 100mM Lithium acetate. The transformation mix included 36 µl of 1M Lithium acetate (LiAC), 25 µl single stranded DNA (ss-DNA), 240 µl of PEG (polyethylene glycol 50% w/v) and 50 µl water and plasmid DNA (0.1-10 µg). The transformation mix was vortexed vigorously to ensure no cell clumps remained in the mix then pelleted and incubated at 30°C for 30 minutes. The mixture was then heat-shocked for 45 seconds at 42°C in a water bath, then cells pelleted and re-suspended in 400 µl sterile water then 110 µl plated onto MUHA plates (YMM agar supplemented with methionine, uracil, histidine and adenine; to select for co-transformants) and MU plates (YMM agar plates supplemented with methionine, uracil; to select for reporter genes *His3* and *ADE2*) respectively. After 2-3 days, co-transformants were replica-plated from MUHA onto MUA plates (which select for weak or low stringent interactions) and MU plates (which select for strong stringent interactions and reporters *His3* and *ADE2*).

2.2.11.2 Beta-galactosidase assay using ONPG substrate

The co-transformants on both MUA and MU plates were further tested for the Beta-galactosidase enzyme activity using the O-nitrophenyl-β-D-galactopyranoside as substrate (Zubay et al., 1972). Co-transformants

were grown in 5 ml of appropriate selection medium e.g., colonies from MUA were picked to YMM MUA and MU colonies into YMM MU broth supplemented with 2% glucose w/v at 30°C overnight with shaking (230 rpm). Then, 2 ml of overnight broth was added to 8 ml of YPD (containing 2% glucose) and incubated at 30°C with shaking (250 rpm) to reach OD₆₀₀ at 0.62. At required OD₆₀₀, 1.5 ml of cells was harvested and resuspended in 1.5 ml Z buffer. Then, pelleted cells were resuspended in 300 µl of Z-buffer at pH 7.0 (16.1 g/L Na₂HPO₄·7H₂O, 5.5 g/L NaH₂PO₄·H₂O, 0.75 g/L KCL and 0.24 g/L MgSO₄·7H₂O). Aliquots, 100 µl each were added with 0.7 ml mixture of Z-buffer and dichlorodiphenyltrichloroethane (DDT), then addition of 160 µl of ONPG (ortho-Nitrophenyl-β-galactoside) in Z-buffer then tubes were incubated at 30°C until a yellow colour developed in the solution. At this point, the reactions were stopped by adding 0.4 ml of 1M Na₂CO₃ (Sigma-Aldrich) and the elapse time was recorded. The OD of the samples was measured (at OD₄₂₀) and beta-galactosidase activity was calculated using the Miller equation as appropriate.

Miller equation is given as: $\text{Miller Units} = 1000 \times \text{OD}_{420} / (t \times V \times \text{OD}_{600})$

Where t = elapsed time (in minutes) of the incubation

V = 0.1 ml x concentration factor (for this V = 0.5)

2.2.12 Protein analysis and purification

2.2.12.1 Construction of pBAD::*tfs3virB11*, pBAD::*tfs4virB11* and pBAD::*gB11*

The genes of *tfs3virB11*, *tfs4virB11* and *gB11* from *H. pylori* strain P12 were amplified using primers listed in Table 2.3. The PCR products and pBAD (ThermoFisher Scientific) expression vector were digested with NcoI and EcoRI high fidelity (HF) restriction enzymes (New England Biolabs) as in 2.2.7. The digested vector was gel extracted as in 2.2.6 and insert : vector ligations (7:1), was performed as described in method 2.2.8. The pBAD::*tfs3virB11*, pBAD::*tfs4virB11* and pBAD::*gB11* constructs obtained from these ligations using methods described in 2.2.10 were finally transformed into *E. coli* expression host strain Top10 (Novagene).

2.2.12.2 Expression, solubility and purification of VirB proteins

The expression host (Top10 *E. coli* strain (Novagene)) harbouring the pBAD::*tfs3virB11*, pBAD::*tfs4virB11* and pBAD::*gB11* constructs were cultured overnight in 10 ml LB broth containing an appropriate concentration of the antibiotic (10 µl Ampicillin¹⁰⁰) at 37°C with shaking at 250 rpm. The following day, 1 ml was transferred from overnight culture into 100 ml LB broth containing 100 µl Ampicillin¹⁰⁰ as selection and incubated at 37°C with shaking for 2 hours 30 minutes until the O.D₆₀₀ reached 0.4. At the appropriate O.D₆₀₀, the culture was induced with 100 µl arabinose¹⁰⁰ and incubated for 2 hours 30 minutes at 37°C in a shaker (250 rpm). After this, cells were pelleted at 4°C for 5 mins at

6500 x g (Beckman Allegra, X-22R) and the supernatant discarded. The cell pellets were then resuspended in 1 ml sonication buffer (20 mM Tris (2.42 g / L), 200 mM NaCl (11.6 g / L) at pH 7.0 with 300 mM and protease inhibitor (PI) (one tablet of protease inhibitor per 20 ml of sonication buffer). The resuspended cells were sonicated with SoniPrep 150 sonicator (Sanyo) at 10 microns amplitude. Sonicated samples were then centrifuged at 17,000 x g (Beckman Allegra, X-22R) for 3 mins and the supernatant transferred into pre-labelled Eppendorf tubes as soluble fractions (supernatant). The insoluble proteins (pellets) were resuspended in 100 µl of sonication buffer and 20 µl of 6x protein loading dye (Promega) was added and 10 µl of the same protein dye (Promega) was added to 50 µl of soluble samples. Samples were resolved on 10 % SDS-PAGE gel [(2.5 ml of 30% (w/v) acrylamide (Sigma-Aldrich), 1.875 ml of 1.5 M Tris pH 8.8/ 0.4% (w/v) SDS, 3.125 ml of distilled water, 25 µl of 10% (w/v) ammonium persulfate (APS) (Sigma-Aldrich) and 5 µl of N,N,N,N'-tetramethylethylenediamine (TEMED) (Sigma-Aldrich)] described below (section 2.2.12.3).

2.2.12.3 Sodium Dodecyl Sulphate- Polyacrylamide Gel Electrophoresis (SDS-PAGE)

SDS-PAGE was used to analyse the supernatant, recombinant proteins and cell lysates as described by Laemmli (Laemmli, 1970). Unless otherwise stated, 10 % w/v resolving gels comprising 2.5 ml of 30% (w/v) acrylamide (Sigma-Aldrich), 1.875 ml of 1.5 M Tris pH 8.8/ 0.4% (w/v) SDS, 3.125 ml of distilled water, 25 µl of 10% (w/v) ammonium persulfate (APS) (Sigma-Aldrich) and 5 µl of N,N,N,N'-tetramethylethylenediamine

(TEMED) (Sigma-Aldrich) were used to separate proteins. The mixture was poured into a 1mm gel cassette (Invitrogen) and overlaid with 1 ml of water saturated butan-1-ol (Fisher-Scientific) for 20 minutes or until the gel mixture solidified. After that, the butan-1-ol was removed, and the gel interface washed three times with distilled water. 30% (w/v) acrylamide stacking gel (0.65 ml acrylamide 30% (w/v), 1.25 ml of 0.5 M Tris pH 6.8/ 0.4% SDS, 3.05 ml of distilled water, 25 µl of 10% (w/v) APS and 5 µl of TEMED) was then poured on the top of the resolving gel and a 12 or 15 wells comb embedded in the mixture as required. The comb was removed from the polymerised stacking gel and the gel cassette placed into an electrophoresis tank (Bio-Rad). Protein samples were denatured by mixing with 6x SDS loading buffer (Promega) and boiled at 100°C for 5 minutes in a heat block. The electrophoresis chamber was filled with 1x SDS-PAGE running buffer then 6 µl of the samples (both soluble and insoluble samples preparation is described below in section 2.2.12.6) loaded into the wells. The proteins were separated by applying 180 Volts for approximately 1 hour.

2.2.12.4 Western Blot analysis

After SDS-PAGE analysis, the gel for the Western blot analysis (Burnette, 1981) was equilibrated in the blotting buffer (5.8 g glycerine, 11.6 g Tris base, 7.4 ml 10% SDS, 400 ml methanol and 2L distilled water) for 10 minutes before transblotting. The nitrocellulose paper cut to the size of the gel (8.5 cm x 6.5 cm) was soaked into the blotting buffer. After that, the Scotch Brite pad, the gel and nitrocellulose paper were assembled in a sandwich manner on a mini blotter cassette (Bio-Rad)

and trapped air bubbles were removed by rolling a roller or pipette gently over the sandwich and the top was aligned with the nitrocellulose. The cassette was placed in the electrode assembly with the nitrocellulose facing the anode and ice pack inserted as the cooling unit in the buffer. The cell was connected to the power pack (Bio-Rad) through the lid and ran at a constant current of 180 mA for 60 minutes. Proper transfer of the pre-stained molecular weight marker was used to adjudge the efficiency of the transfer.

2.2.12.5 Coomassie blue staining

Coomassie Brilliant Blue solution (0.1% (w/v) Coomassie brilliant blue (Bio-Rad), 45% (v/v) methanol, 9% (v/v) glacial acetic acid) was used to stain separated proteins. SDS- PAGE gels were incubated overnight in staining solution then washed and de-stained with distilled water.

2.2.12.6 Small-scale purification of VirB proteins

The soluble fractions (supernatants) of recombinant proteins after expression as described in section 2.2.12.2, were purified using a flow-gravity column chromatography containing 1 ml of TALON metal affinity resin (Clontech) as recommended by the manufacturer's instructions. In brief, the columns (three) were prepared by washing the TALON resin thrice with sonication buffer (20 mM Tris (2.42 g / L), and 200 mM NaCl (11.6 g / L) at pH 7.0 with protease inhibitor (PI). Before loading samples with TALON, 50 µl of soluble samples were treated with 10 µl protein dye (Promega) and kept on ice as pre-TALON samples. After that, 120 µl of TALON was pipetted into each tube containing the remaining soluble

samples of each recombinant protein. TALON was then washed (twice) with 100 µl of sonication buffer at pH 7.0. TALON samples were then centrifuged at reduced speed of 5 x g for 30 seconds. The lysates collected after sonication (on ice) were loaded into appropriate tubes with TALON and incubated for at least 1 hour at 4°C. Samples were then centrifuged at a reduced speed of 5 x g for 30 seconds, supernatants collected into prelabelled tubes and placed on ice as post-TALON samples. 50 µl of each post-TALON sample was treated with 10 µl 6x protein dye and kept on ice. A final wash (10 times) step was done with sonication buffer at pH 7.0. Finally, samples were eluted in 50 µl of elution buffer (300 mM imidazole - 0.102g / 5 ml of sonication buffer), and 10 µl of 6x protein dye was added to each elute and placed on ice as purified samples. The elution step was repeated four times and each time the flow-through was collected into a new separate Eppendorf tube, after the elutes were pooled together. All samples (50 µl each) prepared with 6x protein dye were loaded on to 10 % SDS-PAGE gel and run for 45 minutes. The concentration of proteins were determined using BCA assay kit (Thermo-Scientific) in a 96 well plate following the manufacturer's instructions. Purified proteins were stored in aliquots at -80°C.

2.2.12.6.1 Protein purification optimisation

To further optimise purification and recombinant protein yield, the cultures were incubated at different pH (7.0, 8.0) and three sonication buffers A, B and C were used [(Buffer A containing 20 mM Tris pH 7.0, 200 mM NaCl, 5 mM imidazole (0.087 g / 250 ml) with PI, Buffer B

containing 20 mM Tris pH 8.0, 200 mM NaCl, 300 mM imidazole (0.204 g / 10 ml) with PI and Buffer C was a lysis buffer BugBuster (Sigma-Aldrich) with PI. Each protein was eluted in four different concentrations of sonication buffer [(0 mM imidazole, 100 mM imidazole – 0.034g / 5 ml in sonication buffers A, B and C separately, 300 mM imidazole - 0.102g / 5 ml sonication buffers A, B and C; and 600 mM imidazole - 0.0204g / 5 ml sonication buffers A, B and C) (Fig. 2.2). The third lysis buffer called BugBuster (Sigma-Aldrich) was used alternatively to sonication by SoniPrep 150 sonicator (Sanyo) because it allows gentle lysis or disruption of *E. coli* cell wall in comparison to vigorous lysis by SoniPrep 150 sonicator (Sanyo).

2.2.12.7 Large-scale purification of VirB proteins

After overnight incubation of recombinant proteins in 10 ml LB broth containing 10 µl Ampicillin¹⁰⁰ at 37°C with shaking (250 g), three conical flasks were labelled accordingly as *tfs3B11*, *tfs4B11* and *gB11* to represent each of the recombinant proteins, and each containing 250 ml of LB broth and 250 µl of Ampicillin¹⁰⁰. From the overnight culture, 1 ml of each sample was added to pre-labelled conical flasks and incubated at 37°C for 3 hours until the O.D₆₀₀ reached 0.4. To optimise the purification and protein yield, the cultures were incubated at pH 7.0, 8.0 and room temperature (RT), elution was performed in four concentrations of elution buffer as listed in section 2.2.12.2.1.2. The purification process was repeated as described in section 2.2.12.2.1 with little adjustments as stated below.

After, the supernatants were discarded, and pellets were re-suspended in 10 ml of sonication buffer at pH 7.0 and 8.0 containing protease inhibitor (PI) (Roche). The third tube was re-suspended in 5 ml of lysis buffer (BugBuster buffer (BB) (Sigma-Aldrich) and left at room temperature for 5 minutes. This sample was not treated with the SoniPrep 150 sonicator machine (Sanyo) but the sample was treated as a 'sonicated sample' because after 5 minutes the lysis buffer had broken the cells as an alternative to sonication. The tubes labelled pH 7.0 and 8.0 were sonicated with SoniPrep 150 sonicator (Sanyo) at 10 microns amplitude for three 3-second bursts and the samples were placed on ice after each burst. After sonication, all samples were centrifuged at 4°C for 10 minutes (6,500 x g). The supernatant of each sample was collected and placed on ice. After that, three 1.5 ml Eppendorf tubes were labelled accordingly – pH 7.0, 8.0 and BB. 120 µl of TALON was pipetted into each tube. TALON was then washed (twice) with 100 µl of sonication buffer according to pH use. TALON samples were then centrifuged at reduced speed of 5 x g for 30 seconds. The lysates collected after sonication (on ice) were loaded into appropriate tubes (pH 7.0, 8.0 and BB) with TALON and incubated for at least 1 hour at 4°C. Samples were then centrifuged at a reduced speed of 5 x g for 30 seconds, supernatants collected into prelabelled tubes (pH 7.0, 8.0 and BB) and placed on ice as post-TALON samples. A final wash step was done with appropriate sonication buffer (pH 7.0, 8.0, and BB). The wash step was repeated 10 times. Finally, samples were eluted in 50 µl of each elution buffer (as stated in 2.2.12.2.1.2), and 10 µl of 6x protein dye was added

to each eluted sample, before placing on ice. Eluted samples prepared with 6x protein dye were loaded on to 10 % SDS-PAGE gel and resolved for 45 minutes.

2.3 Results

2.3.1 Bioinformatics: percentage identity and similarity in VirB11

proteins of *tfs* ICEs.

To further investigate the possibility of other components of *tfs4* and its variant types which may be important for assembly and function of the *tfs4* T4SS, the *tfs4* VirB11 and *tfs4* VirB11v ('v' represents 'variant type') proteins were selected. The *tfs4* proteins were selected on the basis of known interactions of homologs of these proteins in other T4SSs, and because the knowledge into their protein-protein interactions might provide insight into the function of *tfs4* T4SS. A third *virB11* known as gB11 ('g' represents 'genomic type') which is not part of the *tfs4* T4SS cluster of genes was also included to investigate the association of orphan gene with *tfs* ICEs cluster of proteins important for function and assembly of T4SS. The *tfs3* VirB11 was also included to test for potential cross-complementation interactions between *tfs3* and *tfs4* ICEs. Pairwise protein alignment of the *tfs3*VirB11, *tfs4*VirB11, *tfs4*VirB11v and gB11 from *H. pylori* strains was carried out as described in section 2.2.1. The result of the protein alignment shows 43.5% identity and 61.0% similarity between *tfs4*VirB11 and *tfs3*VirB11 of *H. pylori* strain P12, 39.0% identity; 57.7% similarity between *tfs4*VirB11 and an orphan genomic VirB11 (gB11) of strain P12; 52.5% identity; 71.7% similarity

between *tfs3*VirB11 and gB11 of strain P12; 53.7% identity and 70.4% similarity between the two variant *tfs4* VirB11s (4B11 and 4B11v) of strains P12 and 247 respectively. The BLAST search of the orphan gene (gB11) detected more than 100 homologs present in different strains of *H. pylori* with >90% sequence identity with inputted gB11 of reference strain P12 (Appendix Fig 3 A-C). The multiple sequence alignment of the VirB11 proteins with VirB11 from *A. tumefaciens* plasmid pTiA6 show conserved Walker A motif and, an essential motif for homodimerization (GXXG) (Fig 2.2).

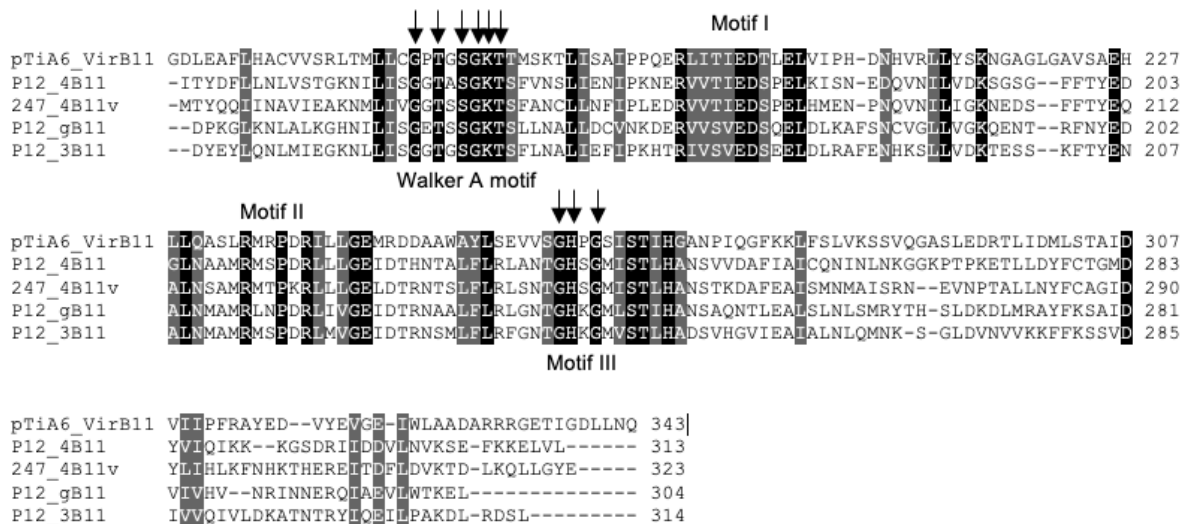


Fig 2.2: Multiple protein sequence alignment showing conservation of motifs in *tfs* VirB11 proteins.

Aligned sequences show conserved Walker A motif and motif III (arrow facing down indicating position of conservation in the motifs) in *H. pylori* strain P12 *tfs3*VirB11, *tfs4*VirB11, gB11 and *tfs4*VirB11v (strain Shi470) in comparison with VirB11 represented by *A. tumefaciens* pTiA6 (Accession number AF242881.1).

2.3.2 Polymerase Chain reaction and Gel Electrophoresis

To examine the potential for cross-complementation interaction between heterologous T4SS proteins, *tfs3*, *tfs4* and other encoded VirB11 were initially selected for investigation. Three different *virB11* (*tfs4virB11*, *tfs4virB11v* and *gB11*) genes were PCR amplified from the genomic DNAs of a selection of *H. pylori* strains (section 2.2.2; Table 2.1) and were electrophoresed on 0.8% agarose gel to ascertain their sizes (Fig 2.3). All other *virB* constructs used in this study were supplied by Dr Rob Delahay from previous work done in the group.

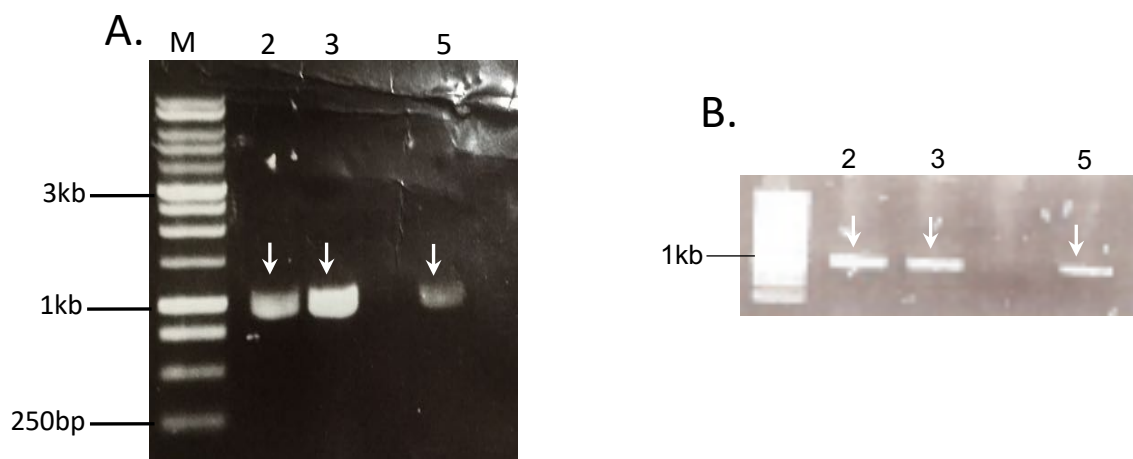


Fig 2.3. DNA amplicons electrophoresed by agarose gel 0.8% at 80v for 30 minutes.

(A) successful amplification of *tfs4gB11*, *tfs4virB11* and *tfs4virB11v* respectively; Lanes 2 and 3 - *gB11* and *tfs4virB11* genes using genomic DNA isolated from *H. pylori* strain P12, and lane 5 – *tfs4virB11v* gene using genomic DNA isolated from *H. pylori* strain 247. PCR products correspond to the predicted size (arrow) of each gene (*gB11* – 915bp, *tfs4virB11* – 942bp and *tfs4virB11v* – 983bp respectively). MW ladder is shown in Lane 1. (B) successful amplification of *tfs3virB11*, *tfs4virB11* and *gB11* selected for recombinant protein assay. Primers were specific for sequence regions flanking the cloning sites of the pBAD recombinant vector and products corresponding to the expected size (arrow) can be

seen in Lane 2 (*tfs3virB11* - 945 bp), 3 (*tfs4virB11* - 942 bp) and 5 (*gB11* -915 bp) respectively.

2.3.3. Colony PCR of Amplicons

Upon successful amplification of *tfs virB11* genes from *H. pylori* strains, the PCR products were digested with restriction enzymes (EcoRI and Sall) and ligated into the cloning vectors pGAD242 and pGBT9 and transformed into the *E. coli* XL1-Blue. Colony PCR as described in section 2.2.10.1 was used to determine the transformants with inserts (Fig 2.4).

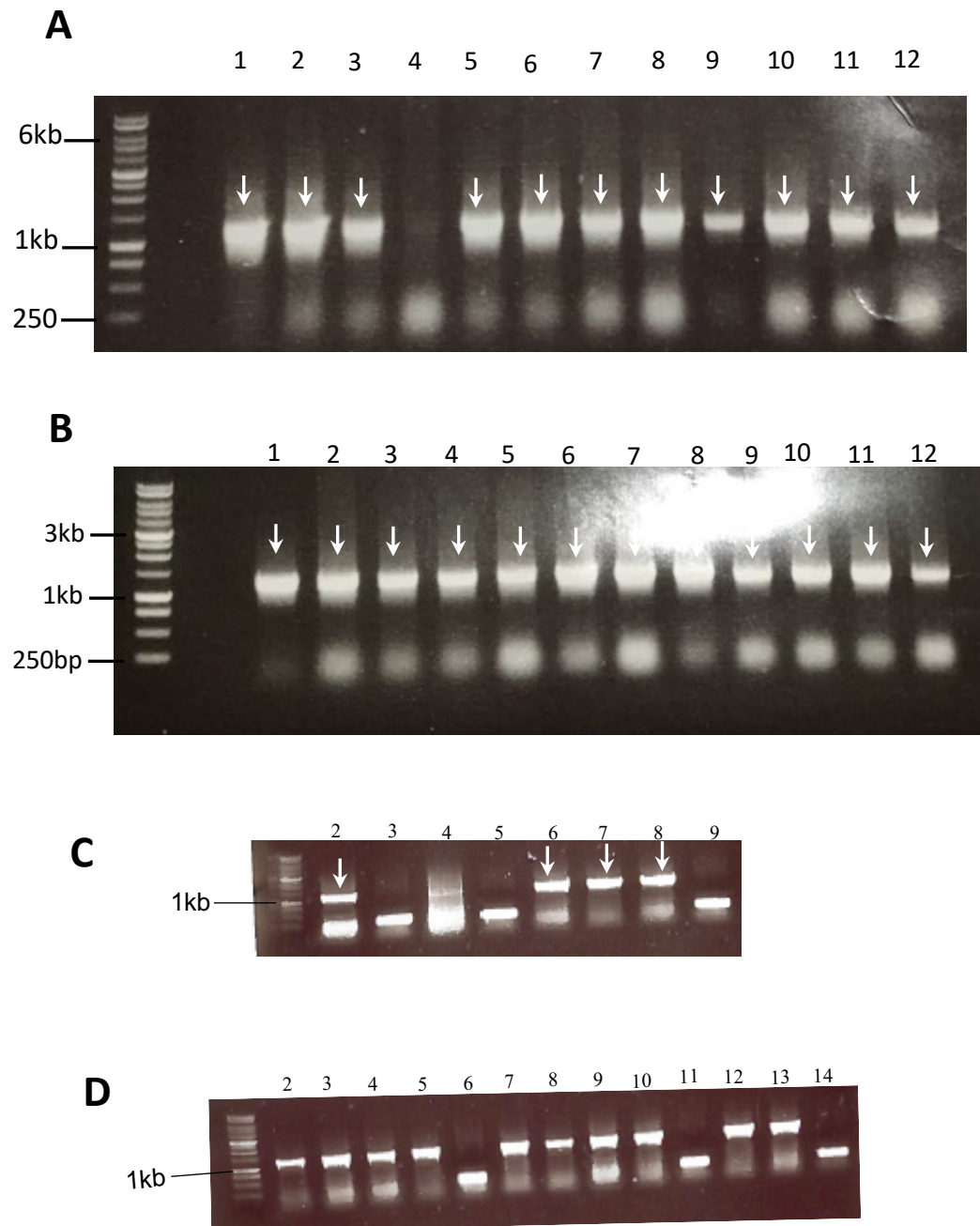


Fig 2.4: Colony PCR of ligated transformants electrophoresed using 0.8% agarose at 80v for 30 minutes.

(A) Lanes 1-4 showing transformants of pGAD424*gB11* size (arrow) 915bp, pGAD424*tfs4virB11v* (lanes 5-8) size 983bp, pGAD424*tfs4virB11* (lanes 9-12) size 942bp. (B) Cloned inserts into pGBT9 – lanes 1-4 pGBT9*gB11* size 915 bp, lane 5-8 pGBT9*tfs4virB11v* size 983 and lanes 6-12 pGBT9*tfs4ivrB11* size 942bp. (C) Lanes 2-9 showing recombinant pBAD::*gB11* construct with predicted size (arrow) 915bp. (D) Lanes 2-6 showing recombinant pBAD::*tfs3virB11* construct

(945bp) and Lanes 7-14 recombinant pBAD::*tfs4virB11* (942bp) construct. Colony PCR performed with both pBAD forward and reverse primers as listed in Table 2.3. MW is shown in the first Lane of each gel image.

Table 2.5: Tfs3 and Tfs4 constructs used in this study for protein-protein interaction in the yeast two hybrid assay.

Strain	Description	Source
pGAD424:: <i>virB4 tfs3</i>	pGAD424 vector containing <i>virB4</i> from <i>tfs3</i>	Alandiyjany, 2015
pGAD424:: <i>virB8 tfs3</i>	pGAD424 vector containing <i>virB8</i> from <i>tfs3</i>	Alandiyjany, 2015
pGAD424:: <i>virB11 tfs3</i>	pGAD424 vector containing <i>virB11</i> from <i>tfs3</i>	Alandiyjany, 2015
pGAD424:: <i>virD4 tfs3</i>	pGAD424 vector containing <i>virD4</i> from <i>tfs3</i>	Alandiyjany, 2015
pGBT9:: <i>virB4 tfs3</i>	pGBT9 vector containing <i>virB4</i> from <i>tfs3</i>	Alandiyjany, 2015
pGBT9:: <i>virB8 tfs3</i>	pGBT9 vector containing <i>virB8</i> from <i>tfs3</i>	Alandiyjany, 2015
pGBT9:: <i>virB11 tfs3</i>	pGBT9 vector containing <i>virB11</i> from <i>tfs3</i>	Alandiyjany, 2015
pGBT9:: <i>virD4 tfs3</i>	pGBT9 vector containing <i>virD4</i> from <i>tfs3</i>	Alandiyjany, 2015
pGAD424:: <i>virB4 tfs4</i>	pGAD424 vector containing <i>virB4</i> from <i>tfs4</i>	Alandiyjany, 2015
pGAD424:: <i>virB8 tfs4</i>	pGAD424 vector containing <i>virB8</i> from <i>tfs4</i>	Alandiyjany, 2015
pGAD424:: <i>virB9 tfs4</i>	pGAD424 vector containing <i>virB9</i> from <i>tfs4</i>	Alandiyjany, 2015
pGAD424:: <i>virB10 tfs4</i>	pGAD424 vector containing <i>virB10</i> from <i>tfs4</i>	Alandiyjany, 2015
pGAD424:: <i>virB11 tfs4</i>	pGAD424 vector containing <i>virB11</i> from <i>tfs4</i>	This Study
pGAD424:: <i>virB11v tfs4</i>	pGAD424 vector containing <i>virB11v</i> from <i>tfs4</i>	This Study
pGAD424:: <i>virD2 tfs4</i>	pGAD424 vector containing <i>virD2</i> from <i>tfs4</i>	Alandiyjany, 2015

pGAD424:: <i>virD4 tfs4</i>	pGAD424 vector containing <i>virD4</i> from <i>tfs4</i>	Alandiyjany, 2015
pGAD424:: <i>virB4v tfs4</i>	pGAD424 vector containing <i>virB4v</i> from <i>tfs4</i>	Alandiyjany (Alandiyjany, 2015)
pGAD424:: <i>virB8v tfs4</i>	pGAD424 vector containing <i>virB8v</i> from <i>tfs4</i>	Alandiyjany, 2015
pGAD424:: <i>gB11</i>	pGAD424 vector containing <i>gB11</i>	This Study
pGBT9:: <i>virB4 tfs4</i>	pGBT9 vector containing <i>virB4</i> from <i>tfs4</i>	Alandiyjany, 2015
pGBT9:: <i>virB4 tfs4</i>	pGBT9 vector containing <i>virB4</i> from <i>tfs4</i>	Alandiyjany, 2015
pGBT9:: <i>virB8 tfs4</i>	pGBT9 vector containing <i>virB8</i> from <i>tfs4</i>	Alandiyjany, 2015
pGBT9:: <i>virB9 tfs4</i>	pGBT9 vector containing <i>virB9</i> from <i>tfs4</i>	Alandiyjany, 2015
pGBT9:: <i>virB10 tfs4</i>	pGBT9 vector containing <i>virB10</i> from <i>tfs4</i>	Alandiyjany, 2015
pGBT9:: <i>virB11 tfs4</i>	pGBT9 vector containing <i>virB11</i> from <i>tfs4</i>	This Study
pGBT9:: <i>virB11v tfs4</i>	pGBT9 vector containing <i>virB11v</i> from <i>tfs4</i>	This Study
pGBT9:: <i>virD2 tfs4</i>	pGBT9 vector containing <i>virD2</i> from <i>tfs4</i>	Alandiyjany, 2015
pGBT9:: <i>virD4 tfs4</i>	pGBT9 vector containing <i>virD4</i> from <i>tfs4</i>	Alandiyjany, 2015
pGBT9:: <i>virB4v tfs4</i>	pGBT9 vector containing <i>virB4v</i> from <i>tfs4</i>	Alandiyjany, 2015
pGBT9:: <i>virB8v tfs4</i>	pGBT9 vector containing <i>virB8v</i> from <i>tfs4</i>	Alandiyjany, 2015
pGBT9:: <i>gB11</i>	pGBT9 vector containing <i>gB11</i>	This Study

Abbreviations: g=genomic, *vir*=virulence, v=variant, pGAD424 (prey with *GAL4* activation domain), pGBT9 (bait with *GAL4* binding domain).

2.3.4 Yeast Two Hybrid Cross-Complementation Interactions

To investigate potential for cross-complementation between *tfs* VirB proteins in this study, all selected genes of both *tfs3*, *tfs4* and *tfs4v* were cloned in both pGAD424 (AD) and pGBT9 (BD) plasmids for all possible pairwise combinations in the Y2H assay. In the preliminary screening

process, all constructs were initially tested for self-activation of the *GAL4* gene by testing for individual construct using Y2H method as described in section 2.2.11.1 with empty plasmids (empty pGAD424 and pGBT9). The results in Table 2.6 and Table 2.7 show that individual fusion constructs did not show strong interaction of the *GAL4* reporter gene. This provides confidence that subsequent interactions between any of the paired construct would not provide false positive results.

Table 2.6: Preliminary screening for self-activation screening of *tfs4 virB* constructs.

Plasmid combination tested in <i>Saccharomyces cerevisiae</i> PJ69-4A		Test/Control	Replica-plates	
pGAD424	pGBT9		MUA	MU
<i>virB4 tfs4</i>	pGBT9	Neg	+	x
<i>virB4v tfs4</i>	pGBT9	Neg	+	x
<i>virB8 tfs4</i>	pGBT9	Neg	+	x
<i>virB8v tfs4</i>	pGBT9	Neg	+	x
<i>virB10 tfs4</i>	pGBT9	Neg	+	x
<i>virB11 tfs4</i>	pGBT9	Neg	+	x
<i>virB11v tfs4</i>	pGBT9	Neg	+	x
<i>virD4 tfs4</i>	pGBT9	Neg	+	x
<i>virD2 tfs4</i>	pGBT9	Neg	+	x
pGAD424	<i>virB4 tfs4</i>	Neg	+	x
pGAD424	<i>virB4v tfs4</i>	Neg	+	x
pGAD424	<i>virB8 tfs4</i>	Neg	+	x
pGAD242	<i>virB8v tfs4</i>	Neg	+	x
pGAD424	<i>virB11 tfs4</i>	Neg	+	x
pGAD424	<i>virB11v tfs4</i>	Neg	+	x
pGAD424	gB11	Neg	+	x
gB11	pGBT9	Neg	+	x

Initial screening was performed twice and selections plates were scored accordingly to number of yeast cells (colonies) that grow on the plates: no colony growth (X), 1-3 colonies (+), 4-10 colonies (++), 11-20 colonies (+++) and >20 colonies (++++). pGAD424 – empty prey plasmid with activation binding domain (AD), pGBT9 – empty bait plasmid with binding domain (BD), MUA – low stringency replica plate supplemented with methionine, uracil and adenine, MU – high stringency replica plate supplemented with methionine and uracil. The screen result shows no auto activation of all included *tfs4* genes and

the orphan genomic gene. Beta galactosidase assay was not done because no *GAL4* activity was reported. Abbreviations: Neg – negative control.

For the pairwise interaction screens using the Y2H assay, yeast co-transformed with both pGAD424 and pGBT9 with individual *vir* gene were initially recovered on MUHA plates that select for yeast cells transformed with both vectors. After, colonies from MUHA plates were then replica plated onto both MUA and MU plates which select for interactions between the activation domain and binding domain fusions as a consequent of reporter activity from functional restoration of *GAL4*. Yeast cells growth on MU plates indicates high stringency selection as this requires activation of two independent reporters (*HIS3* and *ADE2*) whereas growth on MUA plates is less stringent because it only requires activation of *HIS3* reporter activity.

Table 2.7: Preliminary screening for self-activation of *tfs3 virB* constructs.

Plasmid combination tested in <i>Saccharomyces cerevisiae</i> PJ69-4A		Test/Control	Replica-plates	
			MUA	MU
pGAD424	pGBT9			
<i>tfs3virB4</i>	pGBT9	Neg	+	x
<i>tfs3virB8</i>	pGBT9	Neg	+	x
<i>tfs3virB10</i>	pGBT9	Neg	+	x
<i>tfs3virB11</i>	pGBT9	Neg	+	x
<i>tfs3virD4</i>	pGBT9	Neg	+	x
pGAD424	<i>tfs3virB4</i>	Neg	+	x
pGAD424	<i>tfs3virB8</i>	Neg	+	x
pGAD424	<i>tfs3virB10</i>	Neg	+	x
pGAD424	<i>tfs3virB11</i>	Neg	+	x

Initial screening was performed twice, and selections plates were scored accordingly to number of yeast cells (colonies) that grow on the plates as follows: no colony growth (X), 1-3 colonies (+), 4-10 colonies (++), 11-20 colonies (+++) and >20 colonies (++++). pGAD424 – empty prey plasmid with activation binding domain (AD), pGBT9 – empty bait plasmid with binding domain (BD), MUA – low stringency replica plate supplemented with methionine, uracil and adenine, MU – high stringency replica plate supplemented with methionine and uracil. The screen result shows no auto activation of all included *tfs3* genes. Beta galactosidase assay was not done because no *GAL4* activity was reported. Abbreviations: Neg – negative control.

2.3.5 Validation of VirB8 interactions

After screening for self-activation of *tfs* VirB proteins, VirB8 was selected as a good candidate for potential cross-complementation interactions because VirB8 proteins have been reported to interact from a previous study carried out in the group. Repeat analysis of the interactions between VirB8 proteins was carried out to confirm reproducibility of the

results reported from previous studies. Strong homodimerization interactions were reported between putative core structural proteins of *tfs4virB8*, *tfs4virB8v* and *tfs3virB8* proteins (Table 2.8). This is expected since the protein assembles as a multimer (Fig 1.7).

Also, interactions were recorded between *tfs4virB8/ tfs4virB8v* in both orientations. Investigation of potential interchangeability between proteins of *tfs3* and *tfs4*, found there were no interactions between the putative core structural proteins from both *tfs* ICEs (*tfs3virB8/tfs4virB8*), and no interactions were reported between *tfs3virB8* and *tfs4virB8v* (Table 2.8).

Table 2.8: Screening for validation of previous VirB interactions.

Pairwise screening was performed in triplicate, and selection plates were

Plasmid combination tested in <i>Saccharomyces cerevisiae</i> PJ69-4A		Test/Control	Replica-plates		DP
pGAD424	pGBT9		MUA	MU	MU
<i>tfs4virB8</i>	<i>tfs4virB8</i>	TI	++	+	+
<i>tfs4virB8v</i>	<i>tfs4virB8v</i>	TI	++	++	+++
<i>tfs4virB8</i>	<i>tfs4virB8v</i>	TI	+++	++	++
<i>tfs4virB8v</i>	<i>tfs4virB8</i>	TI	+++	+	++
<i>tfs4virB8v</i>	<i>tfs3virB8</i>	TI	+	x	x
<i>tfs3virB8</i>	<i>tfs4virB8v</i>	TI	+	x	x
<i>tfs3virB8</i>	<i>tfs3virB8</i>	TI	++	++	+++
<i>tfs3virB8</i>	<i>tfs4virB8</i>	TI	+	x	x
<i>tfs4virB8</i>	<i>tfs3virB8</i>	TI	+	x	x

scored accordingly to number of yeast cells (colonies) that grow on the plates as follows: no colony growth (X), 1-3 colonies (+), 4-10 colonies (++) , 11-20 colonies (+++) and >20 colonies (++++). The table showed that *tfs4virB8* interacts with itself in the same system and with *tfs4virBv*. Also, *tfs3virB8* was found to interact with *tfs3virB8* but no interaction was reported between *tfs3virB8/tfs4virB8* and *tfs3virB8/tfs4virBv*. Abbreviations: MUA – low stringency replica plate supplemented with methionine, uracil and adenine, MU – high stringency replica plate supplemented with methionine and uracil. DP – direct MU plates (In black = test interactions with one reporter activation and in red = test interaction with more than one reporter activated).

2.3.6 Interactions of ATPases T4SS proteins – for assembly and secretion of *tfs* T4SSs

To investigate the functional activity of *tfs* ICEs proteins that mediate ICE T4SS, the proteins encoded by selected *virB11* genes (*tfs3virB11*, *tfs4virB11* and *tfs4virB11v*) and the orphan *gB11* were tested for

homodimerisation. The three VirB11s were shown to homodimerize (Table 2.9) as expected since they are multimeric ATPases.

Table 2.9: Summary of homodimerisation interactions between VirB11 proteins.

Plasmid combination tested in <i>Saccharomyces cerevisiae</i> PJ69-4A		Test/Control	Replica-plates		DP	β -gal (Miller units)
pGAD424	pGBT9		MUA	MU	MU	
<i>tfs4virB11</i>	<i>tfs4virB11</i>	TI	++	+	+	-
<i>tfs4virB11v</i>	<i>tfs4virB11v</i>	TI	++	++	++	+
<i>tfs3virB11</i>	<i>tfs3virB11</i>	TI	+++	+++	++	++++
<i>gB11</i>	<i>gB11</i>	TI	+++	+++	+	++++

Pairwise screening was performed in triplicate and selection plates were scored accordingly to number of yeast cells (colonies) that grow on the plates as follows: no colony growth (X), 1-3 colonies (+), 4-10 colonies (++) , 11-20 colonies (+++) and >20 colonies (++++). The table showed strong homodimerisation between the three VirB11s proteins. Abbreviations: MUA – low stringency replica plate supplemented with methionine, uracil and adenine, MU – high stringency replica plate supplemented with methionine and uracil. DP - direct MU plate (In red = test interaction with more than one reporter activated).

2.3.7 Interactions of *tfs4* ATPases proteins with *tfs4* variant proteins important for T4SS assembly and function.

Previous results have identified that *tfs4* ICE can exist as a variant type in different strains of *H. pylori* and fragmentation of both genes clusters in many strains background has been reported. Since individual fragmented ICEs may not encode all components required for assembly and a functional T4SS, it is possible that VirB proteins present in related ICE could functionally complement for the fragmented one, which may produce an assembled and functional T4SS. To investigate this

hypothesis, the *tfs* ICEs ATPases were selected (VirB4, VirB11, and VirD4) as they are known to interact strongly with each other and with several other *tfs* T4SS proteins. *tfs4* ATPases were individually paired with the variant proteins using the Y2H as described in section 2.2.11.1. The results show that *tfs4virB4* interacts with *tfs4virB4v*, *tfs4virB8v* and *tfs4virB11v*. Interactions were also reported between *tfs4virB11/tfs4virB4v*, *tfs4virB11/tfs4virB11v* and *tfs4virD4/tfs4virB8v* (Table 2.10). No interaction was detected between *tfs4virB8v/tfs4virB4* in both orientations tested using the Y2H assay (Table 2.10), Also, *tfs4virD4* showed no interaction with *tfs4virB4v* or *tfs4virB11v* in both orientations.

Table 2.10: Screening for Tfs4 ATPases proteins with Tfs4 variant proteins.

Plasmid combination tested in <i>Saccharomyces cerevisiae</i> PJ69-4A		Test/Control	Replica-plates		DP
pGAD424	pGBT9		MUA	MU	MU
<i>tfs4virB4</i>	<i>tfs4virB4v</i>	TI	++	+	x
<i>tfs4virB4</i>	<i>tfs4virB8v</i>	TI	+	+	x
<i>tfs4virB4</i>	<i>tfs3virB11v</i>	TI	+++	+	x
<i>tfs4virB4v</i>	<i>tfs4virB4</i>	TI	+	+	x
<i>tfs4virB8v</i>	<i>tfs4virB4</i>	TI	+	x	x
<i>tfs4virB11v</i>	<i>tfs4virB4</i>	TI	++	+	x
<i>tfs4virB11</i>	<i>tfs4virB4v</i>	TI	++	+	x
<i>tfs4virB11</i>	<i>tfs4virB8v</i>	TI	+	x	x
<i>tfs4virB11</i>	<i>tfs4virB11v</i>	TI	+	+	x
<i>tfs4virB4v</i>	<i>tfs4virB11</i>	TI	++	+	x
<i>tfs4virB8v</i>	<i>tfs4virB11</i>	TI	+	x	x
<i>tfs4virB11v</i>	<i>tfs4virB11</i>	TI	+	+	x
<i>tfs4virD4</i>	<i>tfs4virB4v</i>	TI	+	x	x
<i>tfs4virD4</i>	<i>tfs4virB8v</i>	TI	+++	++	x
<i>tfs4virD4</i>	<i>tfs4virB11v</i>	TI	+	x	x
<i>tfs4virB4v</i>	<i>tfs4virD4</i>	TI	+	x	x
<i>tfs4virB8v</i>	<i>tfs4virD4</i>	TI	++	++	x
<i>tfs4virB11v</i>	<i>tfs4virD4</i>	TI	+	x	x

Pairwise screening was performed in triplicate, and selection plates were scored accordingly to number of yeast cells (colonies) that grow on the plates as follows: no colony growth (X), 1-3 colonies (+), 4-10 colonies (++), 11-20

colonies (+++) and >20 colonies (++++). The table showed that *tfs4virB4* interacted with *tfs4virB4v*, *tfs4virB8v* and *tfs4virB11v* in both orientations. Also, no interaction was detected between paired *tfs4virD4/tfs4virB4v*, *tfs4virD4/tfs4virB8v* and *tfs4virD4/tfs4virB11v*. β -gal assay was not done due to no reported growth of yeast cells on direct MU plates where 100 μ l of tested interaction was plated. Abbreviations: MUA – low stringency replica plate supplemented with methionine, uracil and adenine, MU – high stringency replica plate supplemented with methionine and uracil. DP – direct MU plate (In red = test interaction with more than one reporter activated and).

2.3.8 Interactions of *tfs3* ATPases proteins with *tfs4* variant proteins important for T4SS assembly and function.

To investigate the potential for interchangeability between *tfs3* and *tfs4* variant and establish if the interactions of the ATPases are system specific or not, *tfs3virB11* was selected and paired with *tfs4* variant proteins in this study. *Tfs3 virB11* was selected because *VirB11* was known to interact strongly with itself and with other components of T4SS. The result showed that *tfs3virB11* interacted with *tfs4virB4v* and *tfs4virB8v* (Table 2.11).

Table 2.11: Screening for cross-complementation between Tfs3 and Tfs4 variant proteins.

Plasmid combination tested in <i>Saccharomyces cerevisiae</i> PJ69-4A		Test/Control	Replica-plates		DP
pGAD424	pGBT9		MUA	MU	
<i>tfs3virB11</i>	<i>tfs4virB4v</i>	TI	+	+	x
<i>tfs3virB11</i>	<i>tfs4virB8v</i>	TI	+++	+	x
<i>tfs3virB11</i>	<i>tfs4virB11v</i>	TI	+	x	x
<i>tfs4virB4v</i>	<i>tfs3virB11</i>	TI	+	x	x
<i>tfs4virB8v</i>	<i>tfs3virB11</i>	TI	+	+	x
<i>tfs4virB11v</i>	<i>tfs3virB11</i>	TI	+	x	x

Pairwise screening was performed in triplicate, and selection plates were scored accordingly to number of yeast cells (colonies) that grow on the plates as follows: no colony growth (X), 1-3 colonies (+), 4-10 colonies (++), 11-20 colonies (+++) and >20 colonies (++++). The table shows that *tfs3virB11* interacted with *tfs4virB4v* and *tfs4virB8v* respectively. β -gal assay was not done due to no reported growth of yeast cells on direct MU plates where 100 μ l of tested interaction was plated. Abbreviations: MUA – low stringency replica plate supplemented with methionine, uracil and adenine, MU – high stringency replica plate supplemented with methionine and uracil. DP – direct MU plate (In black = test interactions with one reporter activation and in red = test interaction with more than one reporter activated).

2.3.9 Heterologous interactions of *tfs4* VirB11v with other T4SS VirB proteins

To further investigate the importance of the T4SS proteins for assembly and function of a secretion system, *tfs4 virB11v* was selected for interaction with other components of T4SS. The results showed that

tfs4virB11 interacted with *tfsvirB4* and *tfs4virB8* (Table 2.12). Also, no interaction was detected between paired proteins of *tfs4virB11v/tfs3virB4*, *tfs4virB4v/tfs4virB9* and *tfs4virB11v/tfs4virB10*.

Table 2.12: Screening for cross-complementation between *tfs4* VirB11v and other *tfs* component proteins.

Plasmid combination tested in <i>Saccharomyces cerevisiae</i> PJ69-4A		Test/Control	Replica-plates		DP
pGAD424	pGBT9		MUA	MU	
<i>tfs4virB11v</i>	<i>tfs3virB4</i>	TI	+	x	x
<i>tfs4virB11v</i>	<i>tfs4virB4</i>	TI	+	+	x
<i>tfs4virB11v</i>	<i>tfs4virB8</i>	TI	++	+	x
<i>tfs4virB4v</i>	<i>tfs4virB9</i>	TI	+	x	x
<i>tfs4virB11v</i>	<i>tfs4virB10</i>	TI	+	x	x

Pairwise screening was performed in triplicate, and selection plates were scored accordingly to number of yeast cells (colonies) that grow on the plates as follows: no colony growth (X), 1-3 colonies (+), 4-10 colonies (++) , 11-20 colonies (+++) and >20 colonies (++++). Interactions were detected between *tfs4virB11/tfs4virB4* and *tfs4virB11v/tfs4virB8*. β -gal assay was not done due to no reported growth of yeast cells on direct MU plates where 100 μ l of tested interaction was plated. Abbreviations: MUA – low stringency replica plate supplemented with methionine, uracil and adenine, MU – high stringency replica plate supplemented with methionine and uracil. DP – direct MU plate (In black = test interactions with one reporter activation and in red = test interaction with more than one reporter activated).

2.3.10 Interactions between the *tfs4* variant proteins to form hybrid type.

Previous results showed that *tfs4* ICE is present as several variant or hybrid types. Therefore, I hypothesised that proteins from two different variant forms may combine to form a functional T4SS hybrid type. To investigate this hypothesis, the proteins encoded on the variant types (*tfs4* variant proteins) were paired and tested using the Y2H assay. The results in Table 2.13 below show interactions between *tfs4virB4v/tfs4virB8v*, *tfs4virB4v/tfs4virB11v* and *tfs4virB11v/tfs4virB8v*.

Table 2.13: Screening for cross-complementation between *tfs4* variant proteins.

Plasmid combination tested in <i>Saccharomyces cerevisiae</i> PJ69-4A		Test/Control	Replica-plates		DP
pGAD424	pGBT9		MUA	MU	
<i>tfs4virB4v</i>	<i>tfs4virB8v</i>	TI	++	++	x
<i>tfs4virB8v</i>	<i>tfs4virB4v</i>	TI	+	+	x
<i>tfs4virB4v</i>	<i>tfs4virB11v</i>	TI	+	+	x
<i>tfs4virB11v</i>	<i>tfs4virB4v</i>	TI	++	+	x
<i>tfs4virB11v</i>	<i>tfs4virB8v</i>	TI	++	++	x
<i>tfs4virB8v</i>	<i>tfs4virB11v</i>	TI	+	+	x

Pairwise screening was performed in triplicate, and selection plates were scored accordingly to number of yeast cells (colonies) that grow on the plates as follows: no colony growth (X), 1-3 colonies (+), 4-10 colonies (++) , 11-20 colonies (+++) and >20 colonies (++++). β -gal assay was not done due to no reported growth of yeast cells on direct MU plates where 100 μ l of tested interaction was plated. Abbreviations: MUA – low stringency replica plate supplemented with methionine, uracil and adenine, MU – high stringency replica plate supplemented with methionine and uracil. DP – direct MU plate (In black = test interactions with one reporter activation and in red = test interaction with more than one reporter activated).

2.3.11 Investigation of VirD2 interactions as a secretory substrate

To test for interactions involving other non-assembly proteins such as VirD2, I hypothesised that VirD2 as a secretory substrate, should interact with other VirB proteins within the secretory pathway using Y2H assay. To test the hypothesis, *tfs4* VirD2 was selected as an interaction partner with other *tfs3*, *tfs4* and *tfs4* variant VirB proteins. The results showed a positive interaction between *tfs4virD2* and *tfs4virB11* with activation of more than two reporter genes (in red) Table 2.14, while no interaction was detected between *tfs4virD2* and other *tfs4* and *tfs4* variant proteins (shown in black where only one reporter gene was activated in low stringency replica plate) Table 2.14. Also, *tfs4virD2* interacted with *tfs3virB4*, *tfs3virD4* and *tfs3virB11* (Table 2.15). No interaction was detected between *tfs4virD2/tfs3virB8* and *tfs3virD2/tfs3virB10* (Table 2.15).

Table 2.14: Screening for *tfs4* VirD2 interaction with other *tfs4* VirB and *tfs4* variant VirB proteins.

Plasmid combination tested in <i>Saccharomyces cerevisiae</i> PJ69-4A		Test/Control	Replica-plates		DP
pGAD424	pGBT9		MUA	MU	MU
<i>tfs4virB4</i>	<i>tfs4virD2</i>	TI	+	x	x
<i>tfs4virB4v</i>	<i>tfs4virD2</i>	TI	++	x	x
<i>tfs4virD4</i>	<i>tfs4virD2</i>	TI	++	x	x
<i>tfs4virB8</i>	<i>tfs4virD2</i>	TI	+	x	x
<i>tfs4virB8v</i>	<i>tfs4virD2</i>	TI	++	x	x
<i>tfs4virB9</i>	<i>tfs4virD2</i>	TI	++	x	x
<i>tfs4virB10</i>	<i>tfs4virD2</i>	TI	+	x	x
<i>tfs4virB11</i>	<i>tfs4virD2</i>	TI	+	+	x
<i>Tfs4virB11v</i>	<i>Tfs4virD2</i>	TI	+	x	x

Pairwise screening was performed in triplicate, and selection plates were scored accordingly to number of yeast cells (colonies) that grow on the plates as follows: no colony growth (X), 1-3 colonies (+), 4-10 colonies (++), 11-20 colonies (+++) and >20 colonies (++++). The table shows positive interactions recorded only between *tfs4virD2* and *tfs4virB11*. β -gal assay was not done due to no reported growth of yeast cells on direct MU plates where 100 μ l of tested interaction was plated. Abbreviations: MUA – low stringency replica plate supplemented with methionine, uracil and adenine, MU – high stringency replica plate supplemented with methionine and uracil, DP – direct plate and TI – test interaction. (In black = test interactions with one reporter activation and in red = test interaction with more than one reporter activated).

Table 2.15: Screening for *tfs4* VirD2 interaction with *tfs3* VirB proteins

Plasmid combination tested in <i>Saccharomyces cerevisiae</i> PJ69-4A		Test/Control	Replica-plates		DP
pGAD424	pGBT9		MUA	MU	MU
<i>tfs3virB4</i>	<i>tfs4virD2</i>	TI	+	+	x
<i>tfs3virD4</i>	<i>tfs4virD2</i>	TI	++	+	x
<i>tfs3virB8</i>	<i>tfs4virD2</i>	TI	+	x	x
<i>tfs4virB10</i>	<i>tfs4virD2</i>	TI	+	x	x
<i>tfs3virB11</i>	<i>tfs4virD2</i>	TI	+	+	x

Pairwise screening was performed in triplicate, and selection plates were scored accordingly to number of yeast cells (colonies) that grow on the plates as follows: no colony growth (X), 1-3 colonies (+), 4-10 colonies (++), 11-20 colonies (+++) and >20 colonies (++++). The table shows positive interactions between *tfs4virD2/tfs3virB4*, *tfs4virD2/tfs3virD4* and *tfs4virD2/tfs3virB1*. β -gal assay was not done due to no reported growth of yeast cells on direct MU plates where 100 μ l of tested interaction was plated. Abbreviations: MUA – low stringency replica plate supplemented with methionine, uracil and adenine, MU – high stringency replica plate supplemented with methionine and uracil, DP – direct plate and TI – test interaction. (In black = test interactions with one reporter activation and in red = test interaction with more than one reporter activated).

2.3.12 Screening of orphan genomic gene (*gB11*) interaction with *tfs* ICE VirB proteins.

To investigate the potential of the orphan VirB11 from strain P12 to complement for loss of function for any of the *tfs* VirB11 and VirB11 variant proteins, *gB11* was paired with *tfs3*, *tfs4* and *tfs4v* ATPases and other T4SS proteins (VirB8, VirB9 and VirB10) important for secretion and assembly. The results in Table 2.16 show *gB11* interactions with *tfs3*VirB11 and *tfs4*VirB11 (in both orientation), with *tfs3*VirB4 and *tfs4*VirB4 (in single orientation) and with *tfs3*VirB8 and *tfs4*VirB11v (in single orientation).

Table 2.16: Screening for gB11 interactions with tfs3, tfs4 and tfs4 variant proteins.

Plasmid combination tested in <i>S. cerevisiae</i>		Test/Control	Replica-plates		DP
PJ69-4A			MUA	MU	MU
pGAD424	pGBT9				
<i>gB11</i>	<i>tfs3virB4</i>	TI	++	+	x
<i>gB11</i>	<i>tfs3virB8</i>	TI	+	+	x
<i>gB11</i>	<i>tfs3virB10</i>	TI	+	x	x
<i>gB11</i>	<i>tfs3virB11</i>	TI	+	+	x
<i>tfs3virB11</i>	<i>gB11</i>	TI	+	+	x
<i>gB11</i>	<i>tfs4virB4</i>	TI	++	+	x
<i>gB11</i>	<i>tfs4virB8</i>	TI	+	x	x
<i>gB11</i>	<i>tfs4virB9</i>	TI	+	x	x
<i>gB11</i>	<i>tfs4virB10</i>	TI	++	x	x
<i>gB11</i>	<i>tfs4virB11</i>	TI	++	+	x
<i>tfs4virB11</i>	<i>gB11</i>	TI	+	+	x
<i>gB11</i>	<i>tfs4virB4v</i>	TI	+	x	x
<i>gB11</i>	<i>tfs4virB8v</i>	TI	+	x	x
<i>tfs4virB11v</i>	<i>gB11</i>	TI	+	+	x
<i>gB11</i>	<i>tfs4virB11v</i>	TI	+	x	x
<i>gB11</i>	<i>tfs4virD4</i>	TI	+	x	x

Pairwise screening was performed in triplicate, and selection plates were scored accordingly to number of yeast cells (colonies) that grow on the plates as follows: no colony growth (X), 1-3 colonies (+), 4-10 colonies (++), 11-20 colonies (+++) and >20 colonies (++++). The table shows positive interactions of *gB11* with *tfs3virB4*, *tfs4virB4*, *tfs3virB11*, *tfs4virB11*, *tfs3virB8* and *tfs4virB11v* in single orientation. β -gal assay was not done due to no reported growth of yeast cells on direct MU plates

where 100 μ l of tested interaction was plated. Abbreviations: MUA – low stringency replica plate supplemented with methionine, uracil and adenine, MU – high stringency replica plate supplemented with methionine and uracil, DP – direct plate and TI – test interaction. (In black = test interactions with one reporter activation and in red = test interaction with more than one reporter activated).

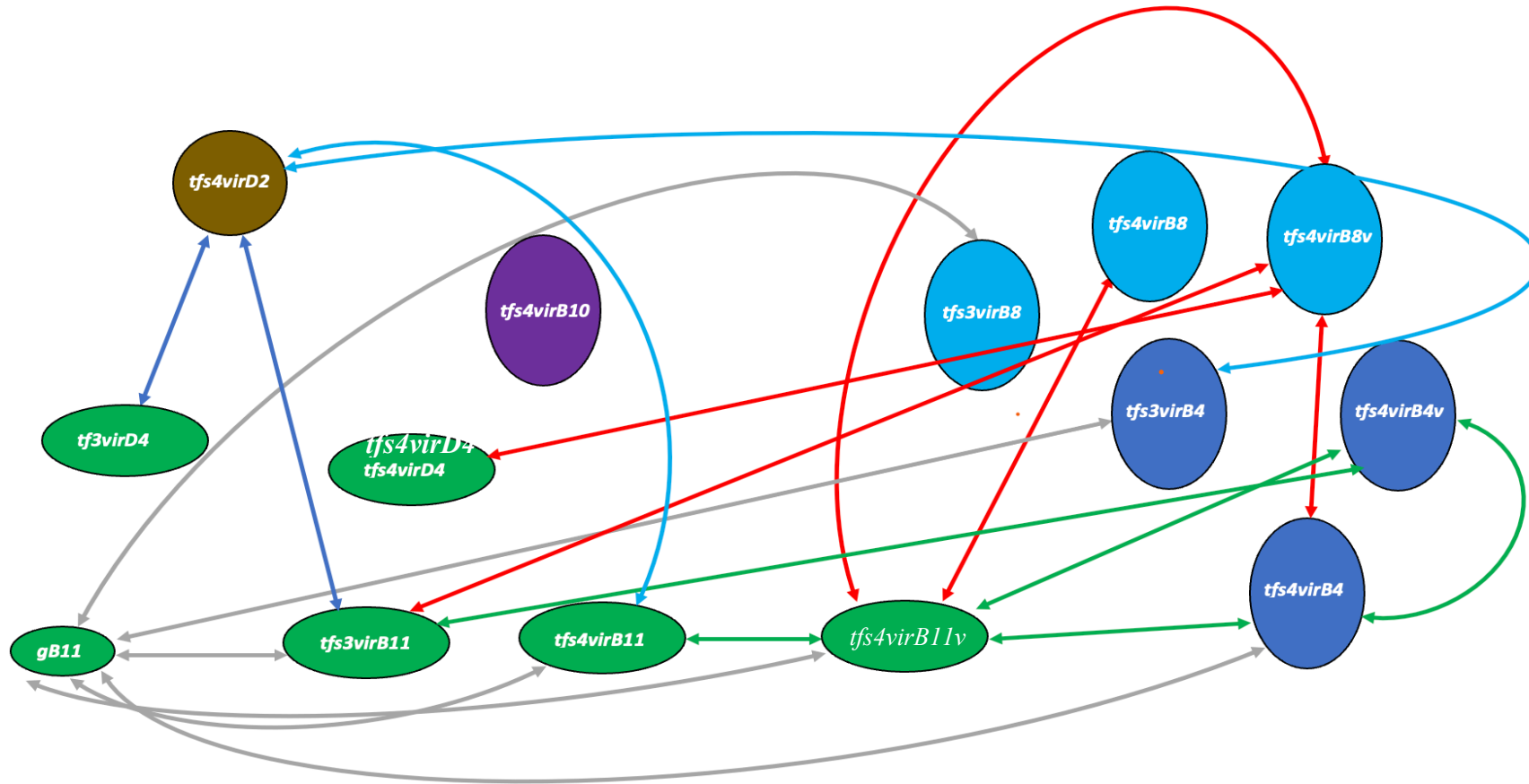


Fig. 2.5: Schematic summary of protein-protein interactions detected using the Y2H assay

Interactions between ATPases and non-ATPases are represented with red arrows, interactions between *tfs* ICEs ATPases, *tfs* ICE and *tfs4* variant type are represented with green arrows, interactions of secretory substrate *tfs4virD2* with other components of *tfs* ICEs are represented with blue arrows and interactions of orphan genomic gene (gB11) with T4SS proteins are represented with grey lines. The *tfs* ICEs ATPases include *tfs3virB11*, *tfs4virB11*, *tfs3virD4*, *tfs4virD4* and their homologues where applicable. The non-ATPases include *tfs3virB8*, *tfs4virB8*, *tfs4virB10*.

2.3.11 Protein expression and purification study.

To validate the novel protein-protein interactions of *tfs* ICEs T4SS proteins reported in Figure 2.5, the genes of *tfs3virB11*, *tfs4virB11* and *gB11* were selected because VirB11 protein is known to be soluble and stable for recombinant protein expression assay. As a first approach to using a secondary approach to validate protein-protein interactions, it was important to test if *tfs3virB11*, *tfs4virB11* and *gB11* are soluble and inducible in a protein expression vector. The genes were successfully cloned into His- tagged expression vector pBAD (sections 2.3.2 and 2.3.3) as recombinant proteins (pBAD::*tfs3virB11*, pBAD::*tfs4virB11* and pBAD::*gB11*) and transformed into expression host Top10 *E. coli* strain (Novagene).

2.3.11.1 Small-scale expression

To investigate if the recombinant VirB11 proteins are soluble and expressible, a small-scale pull-down expression assay was performed as described in section 2.2.12.2. The comparison between the pre-induced and induced samples after they were electrophoresed and stained with Coomassie blue (Bio-Rad) as described in sections 2.2.12.3 and 2.2.12.5, show that the recombinant His-tagged VirB11 proteins (pBAD::*tfs3virB11*, pBAD::*tfs4virB11* and pBAD::*gB11*) are inducible and partially soluble in Top10 *E. coli* strain (Novagene) (Fig. 2.6).

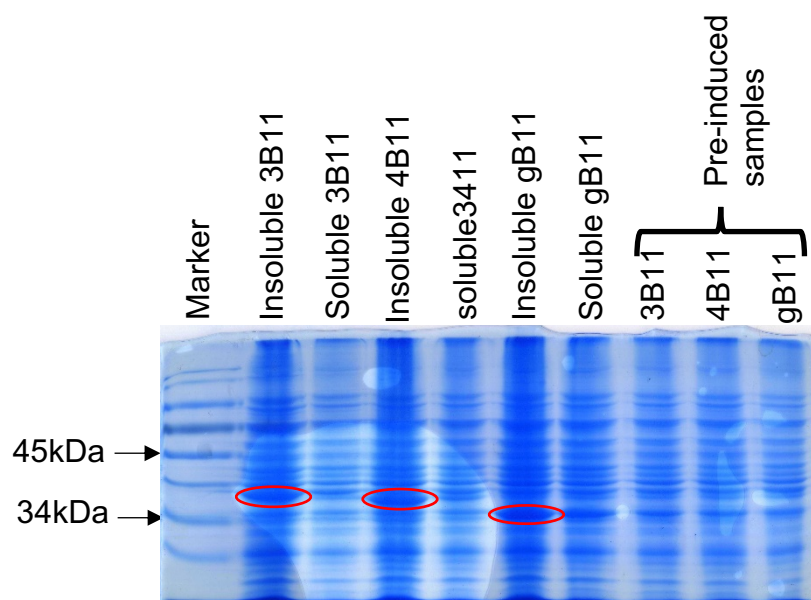


Fig 2.6: Showing 10% SDS-page gel electrophoresed at 180v for 45 mins.

Lane 2 – induced Insoluble pBAD::*tfs3virB11*, Lane 3 – induced soluble pBAD::*tfs3virB11*, Lane 4 – induced insoluble pBAD::*tfs4virB11*, Lane 5 – induced soluble pBAD::*tfs4virB11*, Lane 6 – induced insoluble pBAD::*gB11*, Lane 7 – induced soluble pBAD::*gB11*. Lanes 8-10 are pre-induced samples of pBAD::*tfs3virB11*, pBAD::*tfs4virB11* and pBAD::*gB11*. Size (red circle) correspond to the Molecular weight (Mw) of recombinant VirB11 proteins: pBAD::*tfs4virB11* (35.3kDa), pBAD::*tfs3virB11* (35.7kDa) and pBAD::*gB11* (34.5kDa) respectively. Mw marker shown in Lane 1. 3B11= pBAD::*tfs3virB11*, 4B11=pBAD::*tfs4virB11* and gB11= pBAD::*gB11* respectively.

2.3.11.2 Small-scale purification assay

Following successful induction of recombinant VirB11 proteins, the next important step was purification of the proteins. To purify the His-tagged recombinant proteins, initial small-scale purification procedure was done as described in method 2.2.12.6 without the addition of protease inhibitor. The result after the gel was electrophoresed and stained with Coomassie

blue (sections 2.2.12.3 and 2.2.12.5) showed no purified recombinant VirB11 protein (Fig 2.7).

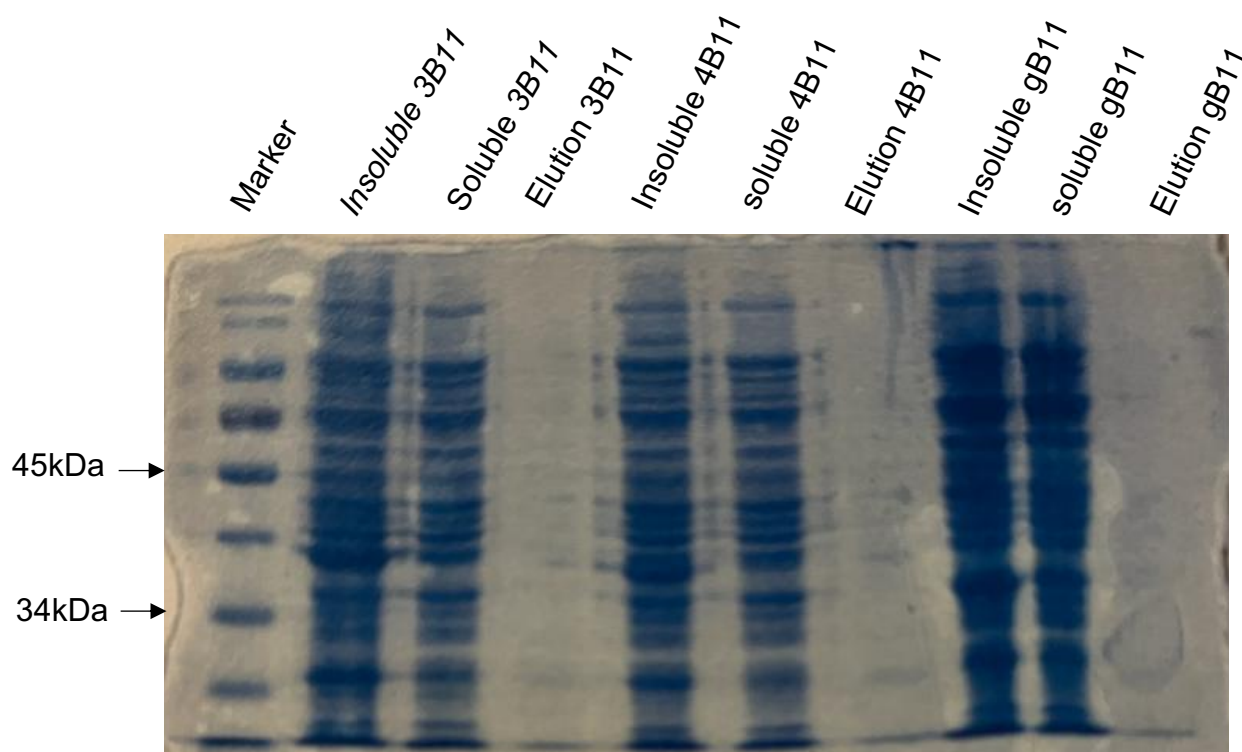


Fig 2.7: Showing 10% SDS-page gel electrophoresed at 180v for 45 mins.

Lanes 2-9 labelled accordingly (on top of image). Mw marker shown in Lane 1. 3B11= pBAD::*tfs3virB11*, 4B11=pBAD::*tfs4virB11* and gB11= pBAD::*gB11* respectively. The recombinant VirB proteins were not treated with protease inhibitor. Recombinant VirB11 proteins sizes: pBAD::*tfs4virB11* (35.3kDa), pBAD::*tfs3virB11* (35.7kDa) and pBAD::*gB11* (34.5kDa).

In order to optimise the yield of recombinant VirB11 proteins, the small-scale purification procedure was repeated as described in section 2.2.12.6 with addition of protease inhibitor (one tablet of protease inhibitor per 20 ml of sonication buffer). The result showed only recombinant pBAD::*tfs4virB11* (red circle) as an eluted fraction (Fig 2.8).

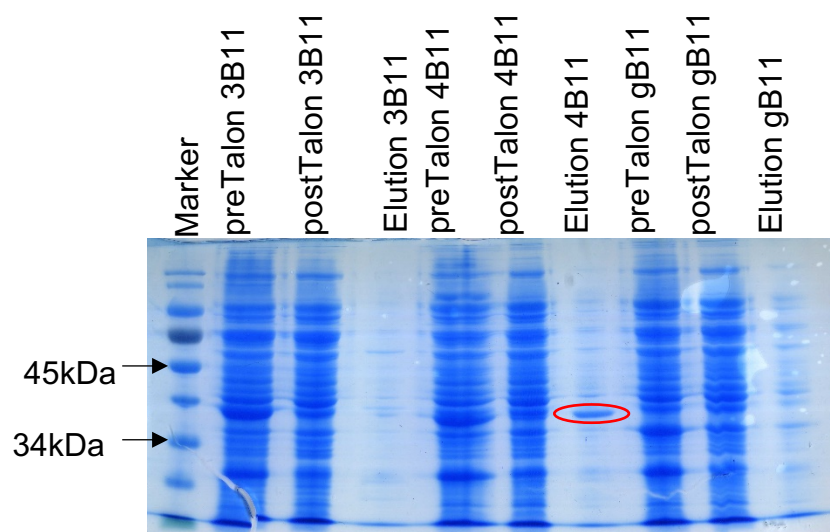


Fig 2.8: Showing 10% SDS-page gel electrophoresed at 180v for 45 mins.

Lanes 2-9 labelled accordingly (on top of image). Mw marker shown in Lane 1. 3B11= pBAD::tfs3virB11, 4B11=pBAD::tfs4virB11 and gB11= pBAD::gB11 respectively. Recombinant VirB11 proteins sizes: pBAD::tfs4virB11 (35.3kDa), pBAD::tfs3virB11 (35.7kDa) and pBAD::gB11 (34.5kDa).

2.3.11.3 Large-scale purification assay.

Following the successful purification of only pBAD::tfs4virB11 under the same purification conditions as described in section 2.2.12.6, optimisation of the purification procedure was considered necessary to increase yield of recombinant proteins pBAD::tfs3virB11 and pBAD::gB11. To do this, the optimisation conditions were adjusted and carefully followed (at every step to minimise error) as described in section 2.2.12.6.1. After that, large-scale purification was done as described in section 2.2.12.7. The result show purified pBAD::tfs3virB11 at different

concentrations of imidazole in elution buffer (Fig 2.9). In comparison with samples treated with BugBuster lysis buffer, the corresponding sizes of elutes of pBAD::*tfs3virB11* were not the same as those treated with sonication buffer and lysed with SoniPrep sonicator.

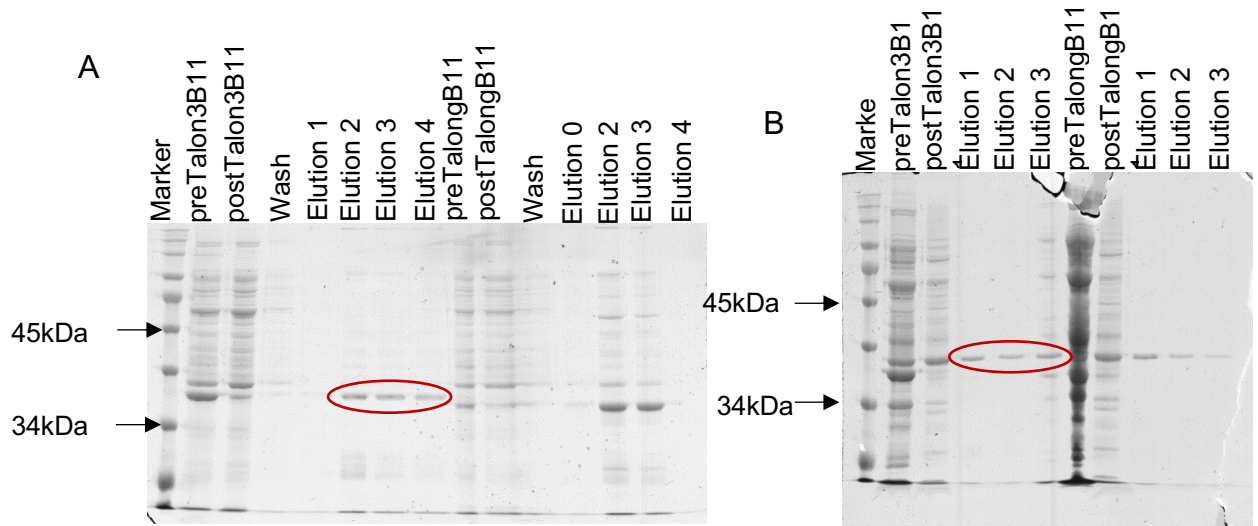


Fig 2.9: 10% SDS-page gel electrophoresed at 180v for 45 mins.

(A) Recombinant VirB11 proteins (pBAD::*tfs3virB11* and pBAD::*gB11*) treated with sonication buffer at pH 7.0 (pBAD::*tfs3virB11*) and pH 8.0 (pBAD::*gB11*). The result in (A) shows purified elutes of recombinant pBAD::*tfs3virB11* (35.3kDa) with elution buffers A, B and C. (Buffer A containing 20 mM Tris pH 7.0, 200 mM NaCl, 5 mM imidazole (0.087 g / 250 ml) with PI, Buffer B containing 20 mM Tris pH 8.0, 200 mM NaCl, 300 mM imidazole (0.204 g / 10 ml) with protease inhibitor and Buffer C was a lysis buffer BugBuster (Sigma-Aldrich) with protease inhibitor - B) Recombinant VirB11 proteins treated with BugBuster lysis buffer at room temperature for 5 minutes.

2.3.12 *In silico* bioinformatics analysis of Novel protein-protein interactions

As a second alternative approach to validate reported *tfs* protein-protein interactions detected using the Y2H assay (Fig 2.6), homology sequence analysis was performed using the bioinformatics program MPI Toolkits as listed in Table 2.1. The Quick2D tool in Toolkits was used to search for important secondary structures such as coiled-coils, transmembrane and signal peptide regions, all of which are important interaction domains in protein-protein interactions. The result of the secondary structure analysis of individual *tfs* VirB proteins that interacted with another *tfs* VirB protein, predicted coiled-coil regions well distributed in both the C- and N- termina regions of the proteins of *tfs3*VirB4, *tfs3*VirB8, *tfs3*VirD4 and, *tfs4*VirB4, *tfs4*VirB8, *tfs4*VirD2, *tfs4*VirBv and *tfs4*VirB8v (*tfs3* and *tfs4* proteins are from strain P12 and the variant proteins from strain Shi470) (Fig 2.10). Transmembrane region (TM) was detected in *tfs3*VirB8, *tfs4*VirB8v, *tfs4*VirB8 and *tfs3*VirD4 (Fig 2.10). Additionally, the region of Walker A motif was indicated in *tfs3*VirB4 (485 – 492), *tfs4*VirB4 (446 - 454) and *tfs4*VirB4v (473 – 480).

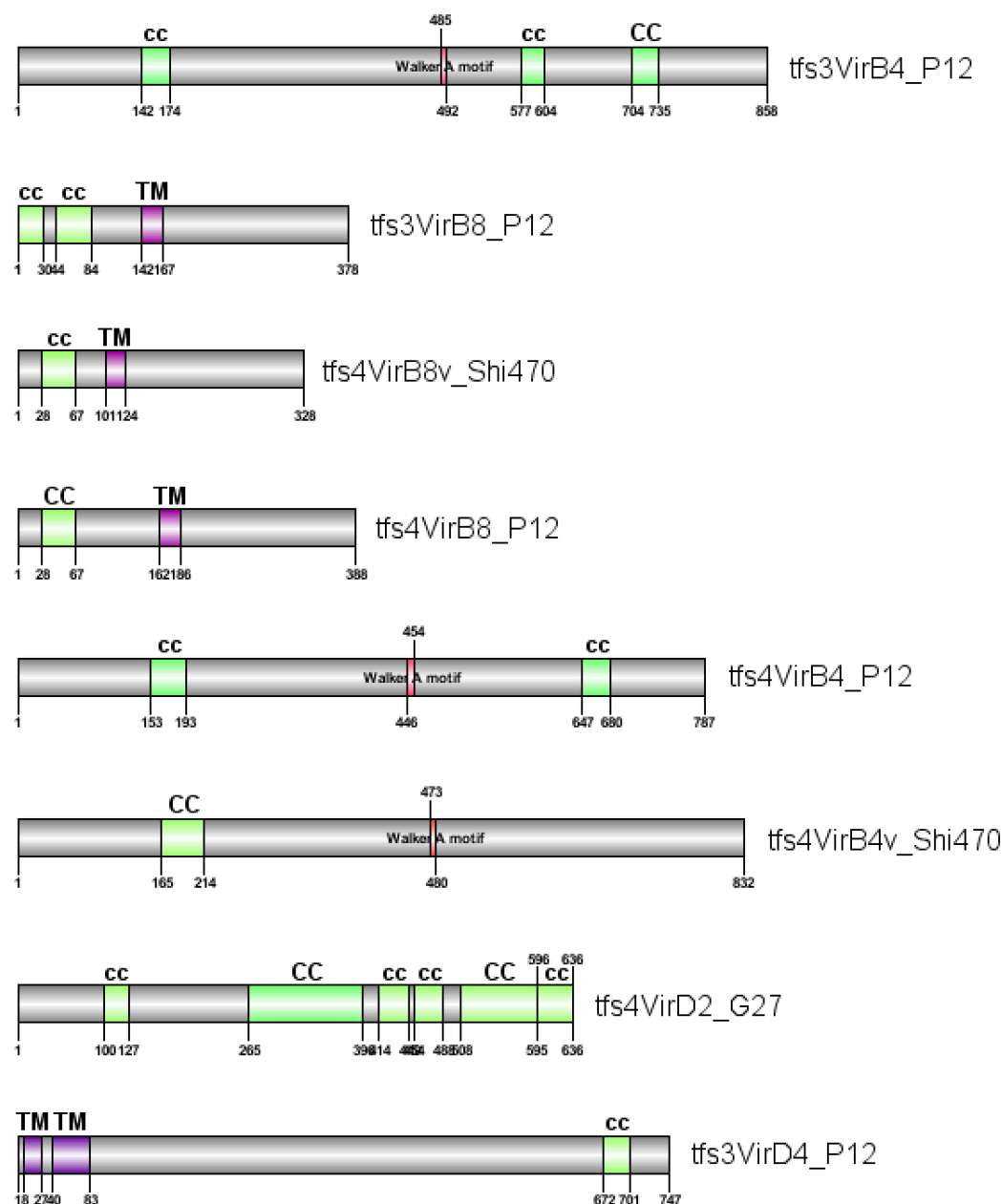


Fig 2.10: Positions of secondary structure predictions in novel interaction partners.

Coiled coil (green box) region was predicted in proteins of *H. pylori* P12 *tfs3VirB4*, *tfs3VirB8*, *tfs3VirD4* and, *tfs4VirB4*, *tfs4VirB8*, *tfs4VirD2*, *tfs4VirBv* and *tfs4VirB8v* (strain Shi470 was used for variant proteins). Transmembrane region (purple box) was predicted in proteins of *H. pylori* P12 *tfs3VirB8*, *tfs3VirD4*, *tfs4VirB8* and *tfs4VirB8v* respectively. Walker A motif was predicted in proteins of *H. pylori* P12 *tfs3VirB4*, *tfs4VirB4* and *tfs4VirBv* (strain Shi470). The grey shaded box represents sequence length of individual protein as represented. Abbreviations: confidently predicted coiled coil (COILS, PCOILS and MARCOILS) (CC), coiled coil predicted only by COILS (cc),

transmembrane predicted by TMHMM, PHOBIUS and POLYPHOBIUS represented by TM.

2.4 Discussion

Both *tfs3* and *tfs4* comprise of approximately 35 genes of which some of these are subset of Vir-homologous proteins in VirB/D T4SS in *A. tumefaciens*. Currently, the function of the *tfs* ICEs remain unknown but they both have virulence association together with genetic diversity across different populations and encode type IV secretion systems (T4SSs) that might be relevant to virulence function. Tfs encoded ICEs have been implicated in increased risk of *H. pylori* related disease outcomes possibly via the stimulation of pro-inflammatory responses in some populations (Delahay et al., 2018b). Particularly inter-ICE transfer has also been reported especially within the strains population of hpAfrica, hpEAsia and hspAmerind, which may suggest changes in virulence character and possibly disease outcomes in some populations (Delahay et al., 2018b). Therefore, this study aimed to understand the aspect of *tfs* ICE biology, including the potential for inter-ICE interactions using the Y2H assay to investigate cross-complementation interactions of *tfs* encoded ICEs among different strains.

Since individual ICEs may not encode all components required for assembly and a functional T4SS, the Y2H method was used to investigate if VirB proteins present in one ICE could functionally

complement for another one, which may then enable an assembled and functional T4SS to form.

2.4.1 Reproducibility of previous VirB8 interactions

VirB8 is an important core component of T4SS assembly which has been reported to interact with other T4SS components. Here, the results showed that interactions between VirB8 proteins validated the previous results by Alandiyjany (Alandiyjany, 2015). The study established positive interactions between VirB8 proteins and revealed some evidence that *tfs4* variant may also interact. The results here shows positive interactions between *tfs4*VirB8v/*tfs4*VirB8, *tfs4*VirB8/*tfs4*VirB8v, *tfs3*VirB8/*tfs3*VirB8, *tfs4*VirB8/*tfs4*VirB8 and *tfs4*BVir8v/*tfs4*VirB8v respectively. The results also support that of Alandiyjany's in that *tfs4*VirB8 and *tfs3*VirB8 are capable of homodimerization. Additionally, this study reports the possibility of weak interactions between *tfs4*VirB8v/*tfs3*VirB8 and *tfs3*VirB8/*tfs4*VirB8 in both orientations. This may be due to inactivation of reporters because both proteins might not interact enough for colony growth on MU plates. The results reported here support the study by Bourg and colleagues (Bourg et al., 2009) where they reported that VirB8 and its homologues TraJ from *Brucella suis* both dimerise to form homodimer and, VirB8 may not be able to interact simultaneously in a wide array of interactions based on its small size. The study also reported that the presence of TM domain resulted in stronger dimerization (Bourg et al., 2009). This study also reports the presence of TM domain in individual protein of *tfs3*VirB8, *tfs4*VirB8 and

tfs4VirB8v (Fig 2.10), which supports Bourg *et al.*, study. Additionally, potential interchangeability between VirB8 of *tfs3* and *tfs4* was not found (showed no strong interaction -no yeast colony growth on both MU replica and direct plates). But weak interactions (indicated by growth on MUA low-stringency plates with one reporter gene activation) between proteins of *tfs4VirB8v/tfs3VirB8* and *tfs3VirB8/tfs4VirB8* in both orientations was identified. This may imply that, both *tfs4VirB8* and *tfs4VirBv* may not be able to functionally replace *tfs3VirB8* in *tfs3* ICE and, this suggests that interactions of VirB8 may be strongly system-specific despite the possibility of both *tfs* ICE to co-exist in a single strain of *H. pylori*. Nevertheless, the transient interactions between *tfs3VirB8/tfs4VirB8* and *tsf3VirB8/tfs4VirB8v* also support the notion that VirB8 undergoes transient protein-protein interactions due to its promiscuity as an assembly factor. Therefore, the results here strongly validated findings of Alandiyjany (Alandiyjany, 2015) and may also suggest VirB8 from different *tfs* ICE may not be functionally flexible in the assembly of Tfs T4SS.

2.4.2 ATPases interactions for system assembly and secretion of *tfs* T4SSs

The three T4SS ATPases (VirB11, VirB4 and VirD4) are indispensable in the secretion system as they provide power to drive the secretion of the substrate. Protein-protein interactions have been demonstrated between the three members of ATPases and each member is capable of interacting with other VirB proteins in the assembly (Fronzes *et al.*,

2009a). VirB11 termed the traffic ATPase is capable of self-display of ATPase activity and known to perform dual function in pilus biogenesis and substrate translocation (Fronzes et al., 2009b).

Using the Y2H assay, this thesis found VirB11 homodimerization for *tfs4VirB11/tfs4VirB11*, *tfs4VirB11v/tfs4VirB11v*, *tfs3VirB11/tfs3VirB11* and gB11/gB11. The homodimerisation of the VirB11 has been reported to be important for its function as previous studies have reported that these proteins form hexameric ring structure with two domains, N- and C- termina domains (NTD and CTD) respectively (Ripoll-Rozada et al., 2013). In *A. tumefaciens*, mutational studies have reported interactions of individual regions with itself and not with another region, that is, NTD with NTD and CTD with CTD respectively. Additionally, VirB8 demonstrates ATPase activity and contains both Walker A and B nucleotide binding motifs (Costa et al., 2021). Mutational studies in the Walker A motif have reported a halt in the binding and hydrolysis of ATP (Tato et al., 2007). Furthermore, the presence of Walker A motif was reported to be essential for dimerization of the CTD (Rashkova et al., 2000). Therefore, the reported Walker A motif in this study (Fig 2.12) may be contribute to strong homodimerisation between *tfs* VirB11 proteins reported.

Previous results of protein-protein interactions from the group reported interactions between *tfs4VirB8v* and *tfs4VirB8* and concluded that *tfs4VirBv* could potentially functionally complement the assembly of the *tfs4* T4SS in the absence of *tfs4VirB8* (Alandiyjany, 2015). It was also

inferred that if this is the case, then the presence of incomplete fragmented *tfs* ICE may not necessarily result in T4SS inactivity provided the complementary genes are present elsewhere in the genome. Based on this assumption, I proceeded to investigate the potential for cross-complementation between *tfs4* ATPases and *tfs4* variant proteins that may be important for T4SS assembly and function. Positive interactions were reported between *tfs4VirB4/tfs4VirB4v*, *tfs4VirB4/tfs4VirB8v* and *tfs4VirB4/tfs4VirB11v*. Also, *tfs4VirB11* interacted with *tfs4VirB4v*, *tfs4VirB11v* and, interaction of *tfs4VirD4* was positive with *tfs4VirB8v*. All positive interactions reported showed activation of more than one reported gene (that is growth on high stringent MU plates). The result reported here also supports previous results from the group that there may be a possibility of potential cross-complementation between these proteins. The possibility of *tfs3* ATPases to complement for *tfs4* proteins was also investigated. The results showed interactions of *tfs3VirB11* with *tfs4VirB4v* and *tfs4VirB8v*. Interpretation of these results may imply the possibility of cross-complementation between proteins *tfs3* and *tfs4* ICE. Further experiments are needed before conclusion can be drawn with certainty that proteins of *tfs3* and *tfs4* can functionally complement. Nevertheless, interactions between *tfs4* ATPases and *tfs4* variant proteins suggest the possibility of cross-complementation between these proteins.

Also, the interaction observed for the *tfs4* variant is not strong when compared with sequence similarities which revealed 70.4% similarity between *tfs4*VirB11 and *tfs4*VirB11v, 60.7% between *tfs3*VirB11 and *tfs4*VirB11v. However, the interaction between *tfs4*VirB11v/*tfs3*VirB11 was weaker as it did not produce robust growth on high stringency selection media plate (MU) with no activation of *LacZ* for beta galactosidase activity (data not shown). The results suggest a potential for cross-complementation of *tfs* T4SS protein within the same system and also between both *tfs3* and *tfs4*, although Y2H assay may not be the best method to establish interactions in *tfs* T4SS proteins.

The Y2H assay also showed that there were interactions between *tfs4*VirB11v and *tfs4*VirB8 with minimal growth of yeast on high stringency plate (MU). VirB8 is an important core component T4SS protein demonstrated to be involved in multiple protein-protein interactions (with VirB4, VirB9, VirB8, VirB10 and VirB11) during the assembly of T4SS. The interaction of *tfs4*VirB11v with *tfs4*VirB8 from the same secretion system in this study supports a proposed possible interaction (Fronzes et al., 2009a) of VirB8 with VirB11 and the flexibility of VirB8 to interact with other VirB proteins. Furthermore, interactions between *tfs4*VirB11v and *tfs4*VirB8 may also be important for substrate secretion in the T4SS as VirB8 has been reported to participate directly in substrate secretion (Fronzes et al., 2009a). This suggests that the variant of *tfs4* may be able to act as substrate transfer in the same system. Conversely, *tfs4*VirB11v showed no interaction with *tfs3*VirB11.

This suggests that interactions of the variant *tfs4* may be system specific. Interactions of other variant *tfs4* proteins, *tfs4VirB4v* and *tfs4VirB8v* showed positive interactions with *tfs3VirB11*. These interactions suggest that there is the possibility of functional interchangeability between these *tfs* ICEs T4SS proteins in a single strain of *H. pylori*. Furthermore, Y2H assay results showed no interactions between *tfs4VirB11v/tfs4VirB9*, *tfs4VirB11v/tfs4VirB10* and *tfs4VirB11v/tfs4VirD4* respectively. This result may suggest that the proteins of the variant type of *tfs4* may be unable to form a secretion assembly with the core components of *tfs* T4SS proteins. It may also be suggested that interactions may only be possible with the ATPases and not with the core structural components.

2.4.3 Interactions of orphan VirB11 (gB11)

Although, earlier experimental studies reported that orphan genes are shorter with low GC content [6,10] and were thought to be mis-predicted protein coding genes, they have recently been demonstrated as real and functional proteins [7, 11-24]. For example, an orphan gene of *E. coli* named *neaT* was reported to play an important role in the virulence of the extraintestinal pathogenic *E. coli* in the embryo of zebrafish (Entwistle et al., 2019). Also, orphan genes have been reported to be enriched in genomic islands of bacterial genomes (Entwistle et al., 2019). Their study reported that 39% of orphan genes were discovered in cluster of genes in genomes of 119 prokaryotes. Many of the orphan genes are lineage-specific and can reside in pathogenicity islands where they may have contributed bacterial virulence (Entwistle et al., 2019). A recent study by

Entwistle and colleagues investigated orphan gene relationship with pathogenicity-related genes in pathogenic and non-pathogenic organisms and they concluded that orphan genes shared by more pathogens are more linked with pathogenicity-related genes and therefore important markers for bacterial infectious diseases. To date, experimental studies are limited on the role and functions of orphan genes present in *H. pylori* because researchers tend to focus more on conserved and virulence genes (Entwistle et al., 2019). To my understanding, this study is the first of its kind to explore the role and functional association of orphan genomic genes in *H. pylori* with cluster of T4SS genes of *tfs* ICEs using experimental methods. The Y2H result showed that gB11 interacts with *tfs3VirB11* and *tfs4VirB11* (in both orientation). Further analysis of the VirB11 sequences through multiple alignment showed conserved Walker A motif which has been reported to be important for protein-protein interaction. This implies that gB11 may be able to functionally complement for both *tfs3VirB11* and *tfs4VirB11* if one or both are fragmented in *H.pylori* infection. Also, gB11 interaction was positive with *tfs3VirB4*, *tfs4VirB4*, *tfs3VirB8* and *tfs4VirB11v* (all interactions reported in single orientation). Interactions with the VirB4s and the VirB11 variant may be due to the conserved Walker A motif, which also provides more evidence that gB11 may be able to functionally complement for these proteins in a T4SS assembly. The potential for functional complementation of orphan VirB11 (gB11) reported in this study using protein-protein interactions supports an earlier study that

most orphan genes in the genome of *E. coli* encode functional proteins (Jalan et al., 2011).

The relaxase proteins (VirD2) are important components in the mobilisation and translocation of bacterial DNA through conjugative mechanisms (Grove et al., 2013). Generally, they are responsible for transferring endogenous plasmids between strains but they have been reported to be an integral part of mobile genetic elements such *tfs* ICEs (Grove et al., 2013). The *tfs4* cluster of genes is known to be self-transmissible and fragmentation of the genes encoded on *tfs3* and *tfs4* is common in *H. pylori* strains, promoting the possibility of inactivation of the *tfs* ICE. However, the possibility of both *tfs3* and *tfs4* to co-exist in a single strain may result to inter ICE transfer of large fragments of the *tfs4* or *tfs3* gene cluster. Because VirD2 activity would be important for the process, I hypothesised that *tfs4*VirD2 can interact with *tfs3* T4SS proteins. The result showed that *tfs4*VirD2 interacted with the three ATPases of *tfs3* (VirB11, VirB4 and VirD4) in single orientation. This may suggest that there is a possibility of *tfs4*VirD2 being able to transfer *tfs3* ICE in the case when both co-exist in a single strain and complement loss of function for *tfs3*VirD2 which may be inactive due to fragmentation.

The characterisation of protein-protein interactions is important to gain further understanding into various biological processes (Reddy Chichili et al., 2013). The unfolding and unstable properties of some proteins have created difficulties to study the interactions for many protein complexes (Reddy Chichili et al., 2013). For example, it is difficult to

characterise membrane proteins due to their partially hydrophobic surfaces, flexibility, and lack of stability. Also, the exposure of the natural lipid bilayer exposes the hydrophobic bilayer which consequently results in protein denaturation or aggregation (Pan and Vachet, 2022). However, several techniques such as Y2H and pull-down have been developed to overcome these challenges. The immobilised metal-affinity chromatography (IMAC) method used to purify the *tfs* VirB11 proteins in this study used to purify his- and polyhis-tagged proteins and provide an advantage in that it uses an inexpensive matrix such as TALON to withstand multiple regeneration cycles under stringent conditions. It also allows dissociation of the bound recombinant protein by an imidazole concentration (Young et al., 2012). The use of TALON has been reported to reduce the level of non-specific binding in partner proteins and thus, increase the purity of protein elutes (Young et al., 2012). Initial expression and purification of *tfs4*VirB11 and *tfs3*VirB11 reported in this study supported earlier studies that VirB11 proteins are soluble and stable. Due to time constraint, it was not possible to perform a repeat of the expression and purification procedure to generate higher yield of purified proteins.

2.5 Study strength and weaknesses

The strength of the study was firstly the availability of a large number of cloned *tfs* T4SS proteins from *tfs3*, *tfs4* and *tfs4* variant from previous work done in the group. Also, the study used high-throughput method (Y2H) to screen for protein-protein interactions between *tfs* T4SS proteins. This method is one of the well-developed methods to determine

inter- and intramolecular interactions and networks in many organisms (Mehla et al., 2015b). Furthermore, the experimental method used in this study is inexpensive and reagents are readily available. Conversely, the Y2H method has some weaknesses, some of which include, the potential of bait (protein fused to the DNA-binding domain of yeast cell) to produce positive result due to extensive background growth of some baits. Also, there is a tendency for non-specific interactions of the prey (protein cloned to the activating domain of yeast) with different baits which may appear reproducible but may lack specificity (Mehla et al., 2015b, Mehla et al., 2015a). Furthermore, the Y2H method has been plagued between two false-positive results. The first being called the technical false-positive result. This is where the reporter gene is activated without any interaction between the bait and prey proteins, which may occur as a result of the bait acting as the transcriptional activator factor (Mehla et al., 2015b). The second type of false-positive is termed biological false-positive result. This is where there is an interaction between bait and prey but there is no biological relevance in regard to function (Mehla et al., 2015b). Due to these false-positive results, protein-protein interactions generated by Y2H method require further validations using additional experimental methods Co-immunoprecipitation (Co-IP) and pull-down assay, Native gels technique, Far Western Blot Analysis, and Label Transfer Protein Interaction Analysis. Furthermore, X-ray crystallography method can be used to show complementary structures of novel *tfs virB* proteins that interacted. Understanding of protein interaction and function

is improved by X-ray crystallography, which makes it possible to see protein structures at the atomic level.

2.6 Conclusions

The study in this chapter was able to generate the following conclusions.

- Y2H method continues to be a high-throughput method for detecting transient and strong protein-protein interactions.
- There were strong interactions (in both orientation) between ATPases of *tfs4* and *tfs4* variant proteins which suggest that interactions may be system specific.
- Interactions were reported majorly between the paired ATPases proteins of *tfs3* and *tfs4* which may suggest that interactions and potential for functional cross complementation may only be possible with the T4SS ATPases and not with structure components of the secretion system.
- Interactions of orphan genomic protein (gB11) established its potential to complement for loss of function of any of the *tfs* ATPases as it interacts strongly (in both orientation) with *tfs3VirB11* and *tfs4VirB11* respectively.

CHAPTER THREE

Methodology Development to screen
for *tfs* genes in *H. pylori* Clinical
Isolates from the Nottingham
Collection

3.0 Introduction

Many essential genes of the gastric pathogen *Helicobacter pylori* are within three major genomic islands encoding type IV secretion systems (T4SS): the *cag* pathogenicity island (*cagPAI*) and two recently discovered mobile integrative and conjugative elements (ICEs) known as *tfs3* and *tfs4* (Delahay et al., 2018a). Present in both *tfs* ICEs is a tyrosine recombinase family gene also called *xerT*, a long open reading frame (more than 2800 codon) (Delahay et al., 2018b), a cluster of T4SS genes and a subset of accessory genes with unknown functions (Phuc et al., 2021b, Delahay et al., 2018a). All three islands have been linked to virulence, but unlike the *cagPAI*, understanding how the *tfs* elements influence *H. pylori* interactions with human hosts is complicated by a lack of definition of their distribution, diversity, and structural representation in the global *H. pylori* population (Delahay et al., 2018b). The *tfs* ICEs are assumed to be similar to *cagPAI* in that, the T4SSs help with the translocation of factors into host cells (Backert et al., 2017), however the identity and nature of these factors remains unclear. Amongst the three genomic islands, the *cagPAI* T4SS has been well characterised. *cagPAI* T4SS encodes an important virulence gene called cytotoxin associated gene A (*cagA*). The gene products of the *cagPAI* T4SS assemble into a pilus that helps to deliver the effector protein (CagA) and other bacterial components into gastric epithelial cells (Robinson and Atherton, 2021). Infections with CagA positive strains of *H. pylori* have been linked to increased incidence of disease in infected patients (Vicari et al., 1998). The *cagA* gene encodes a protein of molecular weight ranging from 120

to 140 kDa, which is highly immunogenic and induces a specific serological response in infected individuals (Robinson and Atherton, 2021, Robinson et al., 2007). The presence of CagA serum antibodies in infected individuals indicates an increased risk of developing duodenal ulcers and gastric cancer from *cagA* positive strains of *H. pylori* (Cover et al., 1995). When the T4SS translocates CagA into epithelial cells it becomes phosphorylated by host kinases, and stimulates cell signalling pathways including the induction of inflammation (Buti et al., 2011). The T4SS interacts with host cell integrins, and also translocates peptidoglycan peptides into the cell. This activates further cell signalling pathways in gastric epithelial cells, which include nucleotide-binding oligomerisation domain protein 1 (NOD1), a cytoplasmic innate detection system for detection of invading bacteria (Robinson et al., 2007). The interaction of *H. pylori* peptidoglycan peptides with NOD1 activates the nuclear factor κ B (NF- κ B) signalling pathway (Ferreira et al., 2016), which mediates the expression of pro-inflammatory cytokines such as IL-8 (Robinson et al., 2007, Ferreira et al., 2016).

cagE, is another important gene that encodes a necessary component of a functional *cag* T4SS (Osman et al., 2021). *cagE* is a homologue of VirB4, and has been reported to contribute to increased risk of duodenal ulceration, but to confer protection against gastric cancer in some *H. pylori* infected patient populations (Jung et al., 2012). The disease associations with *cagPAI* positive strains are therefore complex and it may not be valid to simply assess them from the presence or absence of

a few key genes such as *cagA* using PCR. It is important to investigate whether the *cagPAI* is likely to encode a complete set of genes, leading to expression of a functional T4SS.

Therefore, this work aimed to develop a robust methodology using bioinformatics analysis of genomic data. The initial aim was to search for *cagA* and *cagE* genes of Nottingham clinical isolates and to then go on in Chapter 4 to study the correlation of *tfs3* and *tfs4* clusters of genes with disease outcomes in patients presenting with *H. pylori* infection. The search for *cagA* and *cagE* PCR genotypes in available data on clinical isolates in the Nottingham Collection was used as a proof of principle for using the bioinformatics software (ABRicate) to establish *tfs* genes and ICE representation in the next chapter.

3.1 Rationale and Aims

The polymerase chain reaction (PCR) method for typing bacterial isolates is one of the most common techniques in many microbiology research laboratories (Rahman et al., 2013) in the world. It has been useful and effective over the years to detect antimicrobial resistance genes (Liu et al., 2019a) and virulence genes (Osek et al., 1999) in different microorganisms. PCR is also used in the clinical diagnosis of some human diseases such as cancers (including gastric MALT lymphoma, of which *H. pylori* is a common cause (Robinson and Atherton, 2021), viral infections such as hepatitis, and it is also used in genetic fingerprinting (Rahman et al., 2013).

Previous PCR genotyping of *H. pylori* clinical isolates from patients by the Helicobacter Research Group in Nottingham investigated the presence of virulence factor genes such as *cagA* and *cagE*, and genes of both *tfs3* and *tfs4* ICEs (unpublished work). Similar work also done by the group, using *H. pylori* isolates of South African origin, reported that PCR typing for *cagA* and other individual *cagPAI* genes is a poor predictor of a functional *cag* T4SS (Argent et al., 2004). This is supported by findings that virulence gene PCR typing is not adequate to predict disease outcomes in patients infected with *H. pylori* (Hosseini et al., 2012, Erzin et al., 2006, Audibert et al., 2000).

The use of genomics combined with functional analyses (both proteomics and transcriptomics) has provided an alternative strategy to search for virulence genes in bacterial genomes (Bakour et al., 2016). Bacterial virulence genes may be detected by sequence homology searches of publicly available databases for known virulence genes, analyses of horizontally acquired genes, or comparative analyses of virulence genes with another known annotated strain (Bakour et al., 2016). An alternative strategy for detecting virulence genes was reported by (Phuc et al., 2021a) where they used the ABRicate bioinformatics tool created by Seemann and colleagues (Seemann, 2018) to analyse clusters of T4SS and *cagPAI* genes in Vietnamese *H. pylori* isolates. The ABRicate software was originally used for detecting antibiotic resistance genes hence, the name **ABRicate** (Seemann, 2018). A new script for

detecting virulence factors called virulence factor database (VFDB) was later incorporated into the software. VFDB is a database of virulence factors for bacterial pathogens that comprises 2505 different virulence factors (Bakour et al., 2016). The ABRicate pipeline for virulence factors uses the VFDB database, to search for sequences with homology to virulence genes. The possibility to make custom databases outside of the reference databases embedded within the software provides an opportunity to write custom scripts for genes of both *tfs3* and *tfs4*, with reference to selected *H. pylori* strains (PeCan18, P12 and Shi470) where complete clusters of *tfs* ICE genes have been defined. A major disadvantage for ABRicate's use is the unavailability of user defined parameters that may help to improve search sensitivity of virulence genes.

This chapter focuses on evaluating ABRicate in comparison to previous data from PCR and phenotypic methodologies and approaches on the well-characterised *cagPAI*. The aim was to investigate the pros and cons of the previously used approaches for virulence typing, looking at comparisons of PCR typing and genomics, based on the available rich dataset of *cagA* PCR typing, CagA serology, *cagE* PCR typing and induction of epithelial cell IL-8 as an indicator for a functional *cag* T4SS, and therefore presence of the *cagE* gene. This approach will later be used on the less well-defined *tfs* ICE system to detect potential virulence factors in the next chapter. Although disease severity cannot be linked only to virulence factors of *H. pylori*, other factors including host genetic

make-up and environmental factors also play an important role (Kandulski et al., 2008). Investigating the clusters of *tfs* genes in clinical isolates may, however, give additional insight into the contribution of *tfs* T4SSs to disease severity in patients.

To understand the functionalities of ABRicate software, ABRicate was firstly used to analyse whole genome sequences from clinical isolates for *cagPAI* genes to evaluate the efficacy of ABRicate tool. Secondly, I aimed to investigate the quality (the percentage accuracy in terms of coverage and identity, and virulence potential) of the software output, and whether the identified nucleotide sequence could be directly translated *in silico* to provide an amino acid sequence of the anticipated size and characteristics based on data from well-characterised and commonly used strains. This may provide more understanding of the pros and cons of PCR typing versus genome analysis. ABRicate was assessed by comparing its output on the presence of *cagA* and *cagE* genes with PCR and phenotypic data from a rich dataset on clinical isolates from patients attending the Queens Medical Centre, Nottingham. The utility of PCR genotyping was to be assessed, and discrepant results between PCR, phenotypic data and genomic data investigated.

3.2 Materials and Methods

3.2.1 Collection of Clinical Isolates

Gastric biopsies were collected from consenting patients attending for a routine upper gastrointestinal endoscopy, usually for investigation of

chronic dyspepsia, at the Queens Medical Centre in Nottingham by Professor John Atherton and his team at the University of Nottingham. Approval for collection and use of clinical isolates for research was granted by the NHS National Research Ethics Service, Nottingham Research Ethics Committee (Ref: 08/H0408/195) 27th January 2009. None of the patients were taking non-steroidal anti-inflammatory drugs, antibiotics or proton pump inhibitor drugs on a regular basis in the two weeks preceding the endoscopy.

Biopsies were taken from the antrum and corpus regions of the stomach for histology and for isolation of *H. pylori*. *H. pylori* infection status was determined via rapid biopsy urease test in the endoscopy suite, and confirmed by isolation and culture of the bacterium, and gastric tissue histopathology. To determine the virulence genotype, DNA was extracted from *H. pylori* isolates and PCR typing was performed.

EDTA vacutainers were used to collect 20 ml peripheral blood samples. Plasma samples were separated and kept at a temperature of -80°C. The updated Sydney scoring system (Dixon et al., 1996) was used to define the severity of inflammation, activity, atrophy and intestinal metaplasia on haematoxylin and eosin stained tissue sections, by an experienced histopathologist who was blinded to other data.

3.2.2 Clinical Isolates selection criteria

The anonymised patient data in the *H. pylori* culture collection database at the University of Nottingham was searched for one or more of the criteria listed in Table 3.1. General patient information for all *H. pylori*

isolates used within this thesis including the age, sex, ethnicity, indication for gastric endoscopy, relevant patient medical history, recent acid suppressant history, recent antibiotic use and other drug history are displayed in Table 3.2.

Table 3.1: Demographic characteristics of patients who donated gastric biopsies.

Patient ID	Age (years)	Sex	Ethnicity	Indication for Endoscopy	Relevant patient medical history
45	63	F	Caucasian	Vomiting/weight loss	N/A
77	57	F	Caucasian	Bloating/dysphagia	No
93	77	M	Caucasian	Heartburn	No
120	59	F	Caucasian	Heartburn	No
194	63	F	Asian	<i>H. pylori</i> culture	Perforated prepyloric ulcer 2 years ago
249	41	M	Pakistan	Dyspepsia and <i>H. pylori</i>	No
265	64	F	N/A	Iron deficiency anaemia	No
295	68	F	N/A	N/A	No
308	50	F	N/A	Eradication therapy failure	No
322	64	F	N/A	Epigastric pain	No
326	68	F	N/A	Eradication therapy failure	Previous eradication attempt
439	51	F	N/A	N/A	N/A
444	68	F	N/A	N/A	N/A
495	40	M	N/A	N/A	No

537	45	M	Pakistan	Eradication therapy failure	Previous eradication attempt
565	79	M	Caucasian	Dyspepsia	No
621	66	M	Caucasian	Dyspepsia, atypical chest pain	No
732	64	M	Caucasian	Epigastric pain	No

General patient information for all patients used within this thesis. The age, sex, ethnicity, indication for gastric endoscopy, and relevant patient medical history are displayed. Abbreviations used within the table include M – Male, F – Female, N/A – not available

Table 3.2: Patient disease status

Patient ID	Oesophagus	Stomach	Duodenum	Other Notes	Disease type	Rapid Biopsy Urease Test Result
45	No	No	No	Deformed pylorus, duodenal ulcer disease (scars/scraters)	Past DUD/Normal	Positive
77	Minor erosion in HH	No	Duodenal ulcer and erosions	Minor HH erosion, DU and erosions	Acute DU	Positive
93	RO(2)	No	No	N/A	Normal	Positive
120	RO(2)	No	No	N/A	Normal	Positive
194	No	No	Scar in Di	Scar in Di (PPyU, scan)	DU	Positive
249	RO(2)	No	eDi and deformed pylorus	N/A	Acute DU	Positive
265	HH and erosions	No	No	N/A	Normal	Positive
295	RO(1)	Antral erosions eGi	2 tiny erosions possible eDi	N/A	No	Positive
308	No	No	No	N/A	Normal	Positive
322	RO(2)	GU	No	N/A	Acute GU	Positive

326	RO(2)	Previous eGi	No, but previous eDi	N/A	Normal previous eGi	N/A
439	No	No	No	N/A	No	Positive
444	No	GU	No	N/A	GU	Positive
495	RO(1)	No	eDi (4 erosions)	N/A	No	Positive
537	RO(2)	No	eDi, previous DUscar	N/A	No	Positive
565	No	No	++eDi	N/A	No	Positive
621	RO(1)	Atrophic gastritis	No	N/A	No	Positive
732	RO(2)	2x GU on lesser curve	Erosive pylorus. No DU	N/A	No	Negative

Disease status of the oesophagus, stomach, duodenum is displayed for each patient as well as other notes and disease type. The CLO test refers to the rapid urease test used to indicate *H. pylori* infection. Abbreviations used within this table include; HH= hiatus hernia, RO= reflux oesophagitis (1=mild, 2=moderate, 3=severe), eGi= erosive gastritis, GU= gastric ulcer, Di= inflamed duodenum, eDi= erosive duodenitis, DUscar= scar from a previous duodenum ulcer, DU= duodenal ulcer, PPyU= pre-pylori ulcer and DUD= duodenal ulcer disease. All data was provided by Professor Atherton and colleagues at the University of Nottingham.

3.2.3 PCR typing for *cagA* and *cagE* genes.

PCR amplification of *cagA* and *cagE* from genomic DNA of *H. pylori* isolates was performed, and the results interpreted, as previously described (Kidd et al., 2001, Argent et al., 2004).

The primers used for *cagA* and *cagE* PCRs were as follows:

cag2 5' GGAACCCTAGTCGGTAATG – *cagA* forward primer

cag4 3' ATCTTTGAGCTTGTCTATCG – *cagA* reverse primer

cagE_F 5' TCTATAAAGAGAGAGGTGTT – *cagE* forward primer

cagE_R 3' GGCTAATCTTTGGTAATCAG – *cagE* reverse primer

The 15µl PCR reactions contained 7.5 µl DreamTaq (ThermoScientific DreamTaq Green PCR Master Mix (2X), 3.0 µl each of the forward and reverse primer, 0.5µl nuclease-free water, and 1 µl genomic DNA template.

After an initial denaturation step at 95°C for one minute, each reaction consisted of 35 cycles of 95°C for 30 seconds, 52°C (*cag 2/4* primers) or 53°C (*cagE_F/R* primers) for 1 minute, and 72°C for 1 minute. Following a final extension step at 72°C for 5 minutes, each PCR mixture was subjected to gel electrophoresis on 1% agarose gels and a 100 bp DNA ladder (Roche Diagnostics, Johannesburg, South Africa) was used as a size marker. Products of 888bp (*cagA*) and 508bp (*cagE*) were obtained.

3.2.4 CagA IgG ELISA

Plasma anti-CagA IgG antibody was measured using a commercial enzyme-linked immunosorbent assay (ELISA) kit (Genesis Diagnostics Ltd, Cambridgeshire, UK). Results of equal to or greater than 6.25 U/mL were defined as positive, as per the manufacturer's instructions.

3.2.5 Interleukin 8 induction as a marker for presence of *cagE*

Interleukin 8 (IL-8) secretion in response to a functional *cagPAI* T4SS was assayed using a previously reported method (Patel et al., 2013), co-culturing *H. pylori* isolates with AGS gastric epithelial cells (ATCC CRL-1739TM). Prior to co-culture, confluent T75 flasks of cells were treated with trypsin, washed, and resuspended in 10ml F12 Ham's medium with l-glutamine and 10% FCS. Cells were counted using a haemocytometer slide and seeded in 24-well plates at 1×10^5 per ml in 1ml of medium. Clinical isolates were suspended in F12 Ham's medium and added to AGS cells at a multiplicity of infection of 100 bacteria per cell, as described by (Patel et al., 2013). After incubation for 24 hours at 37°C in a 5% CO₂ humidified atmosphere, the culture supernatants were harvested (Patel et al., 2013). After co-culture of epithelial cells with *H. pylori*, supernatants were aliquoted, frozen at -80°C and thawed once only. IL-8 concentrations were determined with a Human IL-8 CytoSet™ ELISA (Fisher Scientific), according to the manufacturers' instructions and with a standard curve on each plate. Typical sensitivity limits were 5.1 pg/ml IL-8 (Patel et al., 2013).

3.2.6 Sample preparation

Initial *H. pylori* isolation and culture from gastric biopsy samples was carried out by Professor Atherton's team at the University of Nottingham. These biopsies were plated onto blood base #2 agar plates containing 5 % (v/v) horse blood and incubated for 2-3 days at 37°C under microaerophilic conditions (10% CO₂, 5% O₂, 85% N₂). Single colony isolates and/or sweeping growth from each biopsy was picked and pooled, avoiding bacterial contaminants that might also be present. Picked growth was stored in iso-sensitest medium containing 15% (v/v) glycerol at -80°C (Argent et al., 2008). The isolates were confirmed to be *H. pylori* by Gram staining, urease tests and PCR for *vacA*.

3.2.7 Genomic Extraction and Whole genome sequencing

Whole genome DNA extraction and sequencing were carried out by Dan Wilkinson at Nottingham Trent University from the clinical stocks provided by Prof Atherton. Genomic DNA was extracted using the QIAGEN (Netherlands) QIAmp DNA Mini Kit following the manufacturer's instructions with minor alterations (Wilkinson, 2019). These alterations included reduced incubation time to less than 10 minutes in lysis buffer AW, extension of incubation of proteinase K to 16-20 hours in buffer ATL, additional five minutes incubation at 70°C with 100 mg/ml RNase A (QIAGEN, Netherlands) in buffer AL, use of 4°C 100% ethanol and an additional final centrifugation step at 17,000 xg for ten minutes at room temperature. DNA was eluted in 90 µl non- DEPC treated nuclease free distilled water (Thermo Fisher Scientific, USA), after incubation of the

extraction column with the nuclease free water at room temperature for a minimum of 30 minutes (Wilkinson, 2019). All whole genome sequencing was done on the Illumina MiSeq platform (Illumina, USA).

3.2.7.1 Quality control of sequencing reads

After each sequencing run was completed, the Illumina BaseSpace (Illumina, USA) online tool generated run data results with statistical data used to determine successful completion. The statistical data analysed by the online tool included Q score distribution for all reads >Q30 above 80%, smooth intensity plots across all cycles that do not overlap, cluster density close to the desired density of 1300 k/mm², and near-equal read index distribution (Wilkinson, 2019). Because Illumina software is unable to capture all sequencing adapter sequences and cannot filter reads based on quality, the author (Wilkinson, 2019) used an additional software called Trimmomatic version 0.38 (Bolger et al., 2014) which trimmed for Illumina Nextera XT adapters and reads. The software was set-up in paired end mode to enable both palindrome and simple trimming method (Wilkinson, 2019). This additional software maintains mate pair read information and it uses the information derived from both reads for better identification and removal of sequencing adapters. When an adequately aligned match (by a user determined minimum alignment score) is found between the adapter sequence and the output sequencing reads, it is referred to as a 'simple' trimming method. Alternatively, a palindrome trim detects adapter read by ligating adapters to the beginning of mate pairs, and then aligning their biological

sequences; if the biological sequence aligns with adapters present within the read of the reverse- complement mate pair, a trimming event occurs, removing the adapter read through and leaving only the biological sequence. This is significant because these sequences are not part of the biological target sequence. To ensure that adapter read through was removed from paired-end readings, the palindrome trimming threshold was reduced to alignment score of 20.

After the Illumina sequencing adapters were removed, the resulting reads were trimmed for quality and length with Sickle version 1.33 (Joshi and Fass, 2011). Sickle employs a sliding window approach equal to 0.1 times the length of the read and trims the read when the resulting quality within the window falls below a sequence/base quality threshold (specified by user) by identifying where in the window this drop occurs. Sickle's paired-end mode accepts two paired-end inputs and outputs two trimmed paired-end sequencing read files ((Wilkinson, 2019)).

3.2.8 Sequence retrieval, ABRicate set-up, and genomic analysis of clinical isolates

Online bioinformatics tools from the National Centre for Biotechnology Information (NCBI) database found at <https://www.ncbi.nlm.nih.gov/> were used to retrieve gene sequences - *cagA* and *cagE* from reference strain *H. pylori* 26695, *tfs* cluster of genes *virB1* – *virB11* and *virD2* from reference strains P12, Shi470 - for *tfs4* representation and PeCan18 for *tfs3* representation respectively. Downloaded sequences were

annotated using BioEdit software version 7.2 (Hall, 1999) and translated into protein sequences using the EXPASY translation tool found at <http://web.expasy.org/translate/>. The *tfs* cluster of genes (*virB1* – *virB11* and *virD2*) retrieved from each reference strain (P12, Shi470 and PeCan18) (Fischer et al., 2010, Kersulyte et al.) were used to make custom searches. These *H. pylori* reference strains (P12, Shi470 and PeCan18) have been selected because; they have intact *tfs* ICE genes composition (as shown in Fig 14 Chapter 1 of this thesis) and the default databases (NCBI, CARD, ARG-ANNOT, RESFINDER, MEGARES, EcOH, PlasmidFinder, Ecoli_VF and VFDB) incorporated into ABRicate software lacks these strains.

ABRicate software was installed according to the protocol found at <https://github.com/tseemann/abricate> and virulence factor database (VFDB), one of the databases imbedded into the software was used to search for virulence genes (*cagA* and *cagE* genes specifically) by running the following codes:

```
abricate --db vfdb 45C_contigs.fasta > ../data-results/45C-result.txt
```

(code used to analyse each genome of clinical isolates using virulence factors database – vfdb) where 45C_contigs.fasta is an example of clinical isolate, data-results is a directory created to save the result and 45C-result.txt and the name and format of result file generated for the isolate.

```
cp ~/data-results/ ../../mnt/c/Users/user/Desktop/ -r (code used to save analysed genome results of isolates on computer desktop for easy access)
```

The default database (VFDB) used to search for virulence genes was embedded with *H. pylori* strain 26695 as the default strain with no possibility to change the strain. Although, the default database was used to search for virulence genes, this cannot be used to search for *tfs* genes because strain 26695 used to set-up the default database (VFDB) encodes fragmented ICEs that lack the majority of *tfs* genes. Therefore, the downloaded sequences of *tfs* genes (*virB1-virB11* and *virD2*) from well characterised genes identified from strains P12 (Fischer et al., 2010), Shi470 (Kersulyte et al.) and Pecan18 (Kersulyte et al.) were used to create custom searches for both *tfs3* and *tfs4* cluster of selected *vir* genes in ABRicate software.

These databases were created by extracting the cluster of selected *tfs4* *vir* genes from both strains (P12 and Shi470) and manually labelled by inputting the gene header name in the format (>database name~~~gene~~~accession number~~~gene description~~~any other information). Then, the file created in “fasta” format was exported into ABRicate and custom searches were created following publisher’s instructions using the following codes:

```
% cd /path/to/abricane/db # this is the --datadir default option
```

`% mkdir tinyamr` where `tinyamr` is an example of a new directory name for the custom search or database to be created.

`% cp /path/to/tinyamr.fa sequences`

`% head -n 1 sequences`

`>tinyamr~~~GENE_ID~~~GENE_ACC~~RESISTANCES` some description here

`% abricate --setupdb or`

`% makeblastdb -in sequences -title tinyamr -dbtype nucl -hash_index`
(<https://github.com/tseemann/abricate>)(Seemann, 2018)

The custom searches were used to search for *tfs* genes in clinical isolates by running the following code:

`abricate --db P12_nucl 444A_contigs.fasta > ../data-results/444A-result.txt` and `abricate --db Shi470_nucl 444A_contigs.fasta > ../data-results/444A-result.txt`. where `P12_nucl` is the name of the custom search for *tfs4 vir* genes from reference strain P12, `444A_contigs.fasta` is the sequenced genome contigs of anonymous patients tagged 444A (where A stand for antrum), `data-results` is the directory created to store result data and `444A-result.txt` is name and format used to save result data.

3.2.8 Analyses of Open Reading Frames and motifs upstream in selected clinical isolates

All detected *cagA* genes from selected isolates were translated using EXPASY translation tool found at <https://web.expasy.org/translate/> and translated proteins were aligned using Cluster Omega alignment tool found at <https://www.ebi.ac.uk/Tools/msa/clustalo/> to search for conserved motifs especially EPIYA motifs with amino acid residues Glu-Pro-Ile-Tyr-Ala. These motifs are important for CagA phosphorylation by Src kinases, once translocated inside the host cell (Robinson and Atherton, 2021). To better find the correctness of ORFs in these isolates, a manual search was performed, scanning upstream of transcriptional start sites for the Pribnow-Schaller box, and a six-nucleotide sequence motif (TATAAT) (Wasylyk, 1985) consisting of adenines and thymines and homologous to the TATA box sequence found in the promoter region of eukaryotes and archaea. The Pribnow-Schaller sequence is present in almost all gene promoters of bacteria and is bound by RNA polymerase enzyme during initiation of transcription.

Ferreira and Colleagues (Ferreira et al., 2016) reported conservation of a DNA motif (AATAAGATA) in the *cagA* gene of strains from Colombia. This motif was manually searched for in aligned *cagA* DNA sequences of selected isolates, because it could provide more understanding of the CagA serology results.

3.2.9 Determination of percentage accuracy, sensitivity, and specificity

To determine the level of agreement between data from genomic analyses, PCR typing and IL-8 screening, the percentage accuracy, sensitivity, and specificity was calculated using the following formulae:

Definitions:

True positive = where either *cagA* or *cagE* is positive in both the genome sequence and also by PCR typing.

True negative = where either *cagA* or *cagE* is negative in both the genome sequence and also by PCR typing.

Percentage Accuracy =
$$\frac{(\text{number of true positives} + \text{number of true negatives})}{\text{Total number of strains}} \times 100\%$$

Percentage Sensitivity =
$$\frac{\text{number of true positives}}{(\text{number of true positives} + \text{false negatives})} \times 100\%$$

Percentage Specificity =
$$\frac{\text{number of true negatives}}{(\text{number of true negatives} + \text{false positives})} \times 100\%$$

3.3 Results

To first gain an understanding into how, and if the ABRicate software could detect any virulence genes in clinical isolates, initial searches for *cagA* and *cagE* genes were done as described in section 3.2.8. The result was compared with AGS gastric epithelial cell IL-8 concentrations from co-culture with isolates as a marker for a *cagT4SS* capable of inducing CagA-independent cell signalling, *cagA* and *cagE* PCR typing, and patient anti-CagA IgG serology data from previous work done in the group. This approach was designed to understand how effectively ABRicate software could detect virulence genes in clinical isolates and whether there was consensus between these indicators of *cagT4SS* status. Comparison of IL-8 data from previous work done in the group with ABRicate result is important because IL-8 induction has been reported to increase in strains of *H. pylori* with a functional *cagPAI* type IV secretion system (Robinson and Atherton, 2021, Robinson et al., 2007, Crabtree et al., 1995, Yamaoka et al., 1999b).

3.3.1 Detection of *cagA* genes in Clinical Isolates

Of the genome sequences from 24 clinical isolates, the ABRicate software detected the presence of a *cagA* gene in 18. In comparison, *cagA* was detected by PCR typing in 17 of these isolates (Table 3.3). Seropositivity for CagA was reported in 6 patients and negative in 5 patients whose isolates were sequenced. Table 3.3 summarizes the data

from comparative analyses of *cagA* gene in sequenced genomes of clinical isolates, PCR typing and serology.

Table 3.3: Comparative analyses of *cagA* in sequenced genomes of clinical isolates, PCR typing and serology data.

Patient ID	<i>cagA</i> status in genomes	<i>cagA</i> PCR typing data	CagA serology	PCR agrees with genome (y/n) – for <i>cagA</i> status?	Size of <i>cagA</i> (bp) in sequenced genomes
45A	+ve	+ve	N/A	Yes	3,516
45C	+ve	+ve	N/A	Yes	3,515
77A	+ve	+ve	-ve	Yes	3,518
93C	+ve	+ve	-ve	Yes	3,518
120A	-ve	+ve	-ve	No	N/A
120C	-ve	+ve	-ve	No	N/A
194A	+ve	+ve	+ve	No	3,530
249C	+ve	+ve	+ve	Yes	3,515
265A	+ve	+ve	N/A	Yes	2,906
295C	+ve	-ve	-ve	No	3,518
308A	+ve	-ve	+ve	No	3,560
308C	+ve	+ve	+ve	Yes	3,560
322A	+ve	+ve	+ve	Yes	3,485
326C	+ve	+ve	-ve	Yes	3,518
439A	-ve	-ve	N/A	Yes	N/A
439C	-ve	-ve	N/A	Yes	N/A
444A	+ve	+ve	N/A	Yes	3,515
444C	+ve	+ve	N/A	Yes	3,515
495A	-ve	-ve	-ve	Yes	N/A
495C	-ve	+ve	-ve	No	N/A
537A	+ve	+ve	+ve	Yes	3,515
621A	+ve	+ve	-ve	Yes	3,553
732A	+ve	+ve	-ve	Yes	3,515
732C	+ve	+ve	-ve	Yes	3,515

Comparison of genome and PCR typing methods in the determination of *cagA* in clinical isolates and level of agreement. ABRicate detected *cagA* gene in 18 clinical isolates and in comparison, PCR typing detected *cagA* in 19 isolates. CagA serology was positive in 6 isolates, PCR typing data agrees with genome data in 18 isolates. *cagA* sizes were detected in 18 isolates. Abbreviations are interpreted as follows: N/A (not available), y (yes), n (no), +ve (positive), -ve (negative).

3.3.1.1 Correctness of open reading frames

Identifying Open Reading Frames (ORFs) is crucial to finding genes in the genomes of many organisms. To do this, *cagA* sequences were selected from clinical isolates where the patients' anti-CagA serology data were previously reported as being negative, however *cagA* genotyping PCR data was positive. These isolates were selected because the alignment of *cagA* genes in these isolates with reference to *cagA* gene of 26695, may explain the discrepancies. To confirm the presence of the *cagA* gene detected by ABRicate in clinical isolates, the sequence of the *cagA* gene was analysed in selected isolates (621A, 295C, 77A, 732A and 732C), as described in section 3.2.8. Open reading frame analysis of CagA amino acid sequences from isolates 77A, 295C, 621A, 732A and 732C from initiation start pointed coded with ATG (methionine – M) is shown in Fig. 3.1. Sequence deletions and insertions were observed in isolates 77A, 732A, 732C (upward arrow in Fig 1). An EPIYA motif was detected in all isolates (77A, 295C, 621A, 732A and 732C) and highlighted in red (Fig 3.1) with pink and green colour showing positions where the forward and reverse primers annealed. Furthermore, aligned *cagA* DNA of isolates with reference 26695 showed the Pribnow-Schaller TATAAT box to be completely deleted in isolate 295C (Fig 3.2), and insertions with a guanine 'G' instead of adenine 'A' were present in isolates 77A and 621A (Fig 3.2). Nucleotide insertion was also observed in isolates 732A and 732C in the first three nucleotide sequences of Pribnow-Schaller box. In these isolates, both PCR and genomics data were positive, indicating presence of the *cagA* gene, but the CagA

serology result was reported negative so the protein may not have been expressed. Therefore, analyses of their sequences may provide understanding to seronegative in these isolates and this will be address in the discussion section.

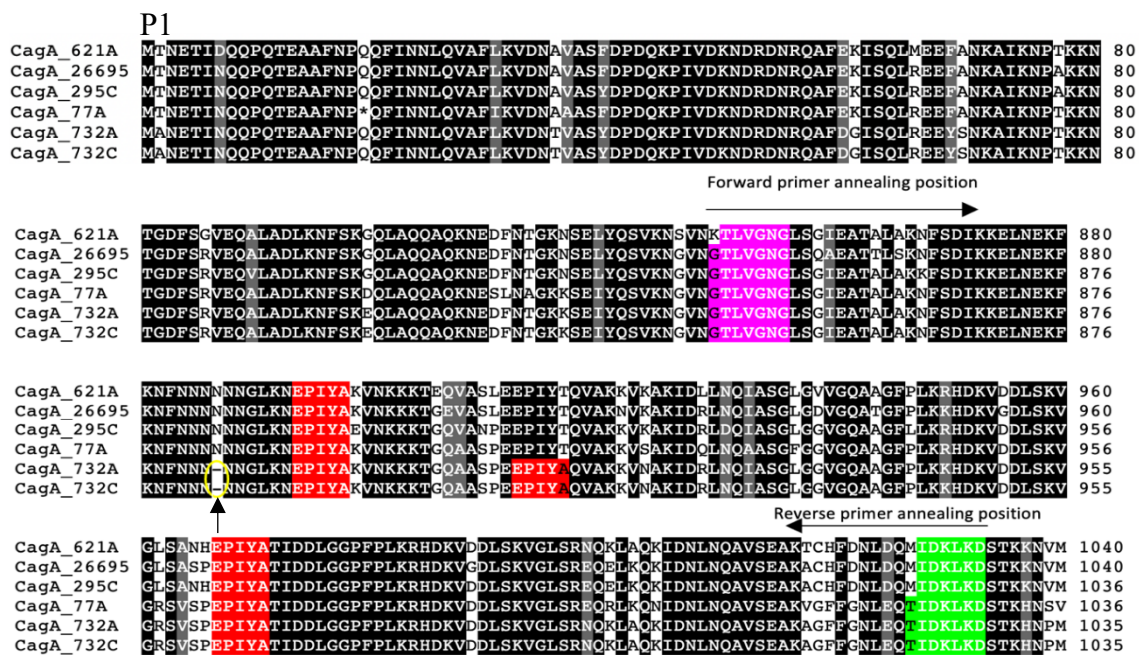


Figure 3.1: Multiple sequence alignment of predicted CagA amino acid sequences from selected genomes of *H. pylori* clinical isolates and reference strain 26695.

The aligned initiation start site (position 1 -P1) of isolates is shown. Arrow up indicates where there is deletion (circled in yellow) in isolates 732A and 732C. Coloured highlighting indicates the EPIYA motif (in red), annealing position of the forward primer (in pink) and annealing position of the reverse primer (in green).

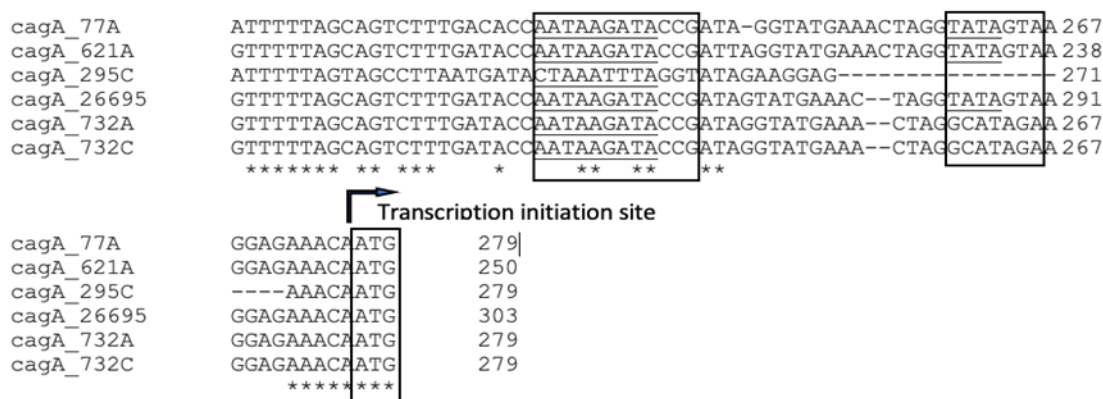


Figure 3.2: The aligned *cagA* gene of selected isolates and, -10 and -35 positions upstream from transcription start site (ATG).

-10 position show the Pribnow-Schaller box (TATAGT underlined)

3.3.2 Detection of multiple copies of *cagA* in the genome of the PMSS1 strain

To further test the efficacy of ABRicate software, the whole genome sequence of *H. pylori* strain PMSS1 was retrieved and analysed, as it is known to contain multiple copies of *cagA*. This was done according to the method in section 3.2.6. The ABRicate software detected four copies of *cagA* genes in PMSS1 strain using the default database (VFDB with 26695 as reference strain). When aligned, the four copies of the CagA predicted amino acid sequences contained an EPIYA motif (Fig 3.3 shaded red), as expected.

```

CagA_26695  KNFNNNNNGLKNEPIIYAKVNKKKTGEVASDEEPIYTQVAKNVNAKIDRLNQIASGLGDVGQATGFPLKKHDKVGDLSKV 960
CagA_1_PMSS1 GNFNNNNNGLKNEPIIYAQVNKKKAGQVVSPEEPIYAQVAKKVNNAKIDOLNQAASGLGGVGQ-AGFPLKRHDKVDDLSKV 955
CagA_2_PMSS1 GNFNNNNNGLKNEPIIYAQVNKKKAGQVVSPEEPIYAQVAKKVNNAKIDOLNQAASGLGGVGQ-AGFPLKRHDKVDDLSKV 955
CagA_3_PMSS1 GNFNNNNNGLKNEPIIYAQVNKKKAGQVVSPEEPIYAQVAKKVNNAKIDOLNQAASGLGGVGQ-AGFPLKRHDKVDDLSKV 955
CagA_4_PMSS1 GNFNNNNNGLKNEPIIYAQVNKKKAGQVVSPEEPIYAQVAKKVNNAKIDOLNQAASGLGGVGQ-AGFPLKRHDKVDDLSKV 955

CagA_26695  GLSASPEPIIYATIDDLGGPFPLKHHAKVEDLSKVGLSREQELTQKIDNLNQAVSEAKACHEDNLDOMIDKLKDSSTKKNVM 1040
CagA_1_PMSS1 GRSVSPPEPIIYATIDDLGGPFPLKHHAKVEDLSKVGLSREQELTQKIDNLNQAVSEAKAGFFGNLEQITIDNLKDSSTKKNPM 1035
CagA_2_PMSS1 GRSVSPPEPIIYATIDDLGGPFPLKHHAKVEDLSKVGLSREQELTQKIDNLNQAVSEAKAGFFGNLEQITIDNLKDSSTKKNPM 1035
CagA_3_PMSS1 GRSVSPPEPIIYATIDDLGGPFPLKHHAKVEDLSKVGLSREQELTQKIDNLNQAVSEAKAGFFGNLEQITIDNLKDSSTKKNPM 1035
CagA_4_PMSS1 GRSVSPPEPIIYATIDDLGGPFPLKHHAKVEDLSKVGLSREQELTQKIDNLNQAVSEAKAGFFGNLEQITIDNLKDSSTKKNPM 1035

CagA_26695  NLYVESAKKVPTSLSAKLDNYATNSHTRINSNVKNGAINEKATCMLTQKNPEWLKLVDNKIVAHNVGVSPLSEYDKIGFN 1120
CagA_1_PMSS1 NLWAESAKKVPAISLSAKLDNYATNSHTRINSNIONGAINEKATCMLTQKNPEWLKLVDNKIVAHNVGVSPLSEYDKIGFN 1115
CagA_2_PMSS1 NLWAESAKKVPAISLSAKLDNYATNSHTRINSNIONGAINEKATCMLTQKNPEWLKLVDNKIVAHNVGVSPLSEYDKIGFN 1115
CagA_3_PMSS1 NLWAESAKKVPAISLSAKLDNYATNSHTRINSNIONGAINEKATCMLTQKNPEWLKLVDNKIVAHNVGVSPLSEYDKIGFN 1115
CagA_4_PMSS1 NLWAESAKKVPAISLSAKLDNYATNSHTRINSNIONGAINEKATCMLTQKNPEWLKLVDNKIVAHNVGVSPLSEYDKIGFN 1115

CagA_26695  QKNMKDYSDSFKFSTKLNNAVKDIKSGFTQFLTNMFSMGSYSLMKASVEHGVKNTNTKGGFQKS 1184
CagA_1_PMSS1 QKNMKDYSDSFKFSTKLNNAVKDVRSGFTQFLANAFSTCYCYCLAEENAEHGLKKNVNTKGGFQKS 1179
CagA_2_PMSS1 QKNMKDYSDSFKFSTKLNNAVKDVRSGFTQFLANAFSTCYCYCLAEENAEHGLKKNVNTKGGFQKS 1179
CagA_3_PMSS1 QKNMKDYSDSFKFSTKLNNAVKDVRSGFTQFLANAFSTCYCYCLAEENAEHGLKKNVNTKGGFQKS 1179
CagA_4_PMSS1 QKNMKDYSDSFKFSTKLNNAVKDVRSGFTQFLANAFSTCYCYCLAEENAEHGLKKNVNTKGGFQKS 1179

```

Figure 3.3: Multiple alignment of four copies of CagA amino acid sequences detected using ABRicate software in *Helicobacter pylori* strain PMSS1. The reference strain 26695 is also shown

The conserved EPIYA motif is highlighted in red. Multiple alignment was prepared using ClusterW Omega.

3.3.3 Detection of *cagE* genes in Clinical Isolates

Presence of a *cagE* gene, which is necessary for a functional *cag* T4SS, was investigated in clinical isolate genomes. This was compared with IL-8 concentrations in the supernatants of AGS cells, previous co-cultured with these isolates (J. Cheung, unpublished), as a phenotypic indicator of a functional T4SS. The software found a *cagE* gene to be present in 18 isolates and absent in 6 clinical isolates. When these findings were compared with *cagE* PCR typing, data on 19 clinical isolates agreed with genomic detection of *cagE* genes using ABRicate software. PCR data from 5 clinical isolates did not agree (Table 3.4). IL-8 data for the presence of a functional T4SS was consistent with presence of a *cagE* gene in genome sequences of all 5 isolates tested.

Table 3.4: IL-8 secretion by AGS cells in response to co-culture with *H. pylori* clinical isolates, as a marker for a functional *cag* type IV secretion system (T4SS)

Patient ID	<i>cagE</i> status in genomes	<i>cagE</i> PCR genotype	IL-8 response indicative of an active <i>cag</i> T4SS	PCR agrees with genome for <i>cagE</i> status?	Size of <i>cagE</i> (bp) in sequenced genomes
45A	+ve	+ve	+ve	Yes	2,951
45C	+ve	+ve	N/A	Yes	2,951
77A	+ve	-ve	+ve	No	2,951
93C	+ve	+ve	N/A	Yes	2,952
120A	-ve	-ve	-ve	Yes	N/A
120C	-ve	-ve	N/A	Yes	N/A
194A	+ve	+ve	N/A	Yes	2,951
249C	+ve	+ve	N/A	Yes	2,951
265A	+ve	+ve	+ve	Yes	2,951
295C	+ve	-ve	N/A	No	2,951
308A	+ve	-ve	N/A	No	2,951
308C	+ve	+ve	N/A	Yes	2,951
322C	+ve	+ve	+ve	Yes	2,951
326C	+ve	+ve	N/A	Yes	2,951
439A	-ve	+ve	N/A	No	N/A
439C	-ve	-ve	N/A	Yes	N/A
444A	+ve	-ve	N/A	Yes	2,951
444C	+ve	+ve	N/A	Yes	2,951
495A	-ve	+ve	N/A	No	N/A

495C	+ve	+ve	N/A	Yes	2,951
537A	+ve	+ve	N/A	Yes	2,951
621A	+ve	+ve	N/A	Yes	2,951
732A	+ve	+ve	N/A	Yes	2,951
732C	+ve	+ve	N/A	Yes	2,951

Comparative analyses of *cagE* gene between PCR and sequenced genome data in selected clinical isolates. Abbreviations are interpreted as follows: Concentration of IL-8 indicative of positivity is 5.1 pg/ml. N/A (not available), +ve (positive), -ve (negative).

3.3.4 Analysis of *cagE* gene sequences in selected clinical isolates, detected by ABRicate software.

As reported in Table 3.4 above, ABRicate software detected a *cagE* gene in clinical isolates 45A, 77A, 265A and 322C. When the gene sequences from these isolates were compared to *cagE* of the 26695 reference strain (as described in section 3.2.8), only three of the four isolates aligned (45A, 77A and 322C), with percentages of identity and similarity reported as 26695 and 45A (identity 99.3%, similarity 99.8%), 26695 and 77A (identity 99.1%, similarity 99.6%), and 26695 and 322C (identity 99.4%, similarity 99.8%). It was interesting that ABRicate detected *cagE* in isolate 77A, as the PCR result was negative. Therefore, further alignment of the *cagE* sequence of strain 77A with the *cagE* primers was examined. The result found no annealing position for either of the *cagE* primers (forward and reverse) in the *cagE* sequences detected by ABRicate.

3.3.6 Determination of percentage accuracy, sensitivity, and specificity

3.3.6.1 Percentage Accuracy, Sensitivity and Specificity of *cagA*

findings, using ABRicate software, PCR typing and serological data.

Calculations of the accuracy, sensitivity, and specificity (as described in Section 3.2.7) from comparisons of genome sequence analysis using ABRicate and PCR typing of *cagA* genes in clinical isolates were 75%, 83% and 50%, respectively. *cagA* gene detection using ABRicate was also compared with available anti-CagA serology data from the patients, and the accuracy, sensitivity and specificity were 42%, 100% and 64%, respectively.

3.3.6.2 Percentage Accuracy, Sensitivity and Specificity of *cagE* findings, with ABRicate software and PCR typing, and also IL-8 data indicative of a functional *cag* T4SS.

Calculations of the accuracy, sensitivity, and specificity (as described in method 3.2.7) for comparisons of genome sequence analysis using ABRicate and PCR typing of *cagE* genes in clinical isolates were 79%, 89% and 50%, respectively. *cagE* gene detection using ABRicate was also compared with previous IL-8 data from coculturing the isolates with AGS cells *in vitro*, and the accuracy, sensitivity and specificity were 21%, 100% and 0%, respectively.

3.4 Discussion

The use of bioinformatics tools (alignment tools) has aided research analysis and it is an essential part of many workflows to compare and interpret data in research studies. These tools have helped in several

ways such as assembly of sequenced data, construction of contigs, prediction of genes and putative proteins, prediction of domains and conserved motifs (Roumpeka et al., 2017). All these predictions and functions have been possible by developing computer algorithms and bioinformatics software over the years. Several bioinformatics tools have been developed to screen for virulence genes in genomes of microorganisms. Amongst these are Virulence Finder (a tool that predicts virulence factors in bacterial genomes) (Joensen et al., 2014), BLAST (a tool that searches for regions of similarities between sequences, compares with nucleotide or proteins sequences in databases and calculates the statistical significance) (Donkor et al., 2014), MP3 (tool that predicts adhesion/colonisation genes, secretion and resistance genes) (Gupta et al., 2014), Virulence Searcher (a tool that searches raw bacterial genomes for putative virulence factors) (Underwood et al., 2005), VFDB (an integrated and comprehensive database of virulence factors) (Chen et al., 2005), VirulencePred (a tool that predicts bacterial virulence protein sequence) (Garg and Gupta, 2008). ABRicate software available at <https://github.com/tseemann/abricate> is also a bioinformatics software that screens contigs for virulence genes or antimicrobial resistance genes with eight custom databases (Phuc et al., 2021b). The software is a stand-alone tool that allows for creating custom scripts of gene(s) that can be used as a custom search database.

In this chapter, the ABRicate software was used to screen for virulence genes in 24 *H. pylori* clinical isolates and its output was compared with a dataset of *cagA* and *cagE* PCR typing, patient CagA serology, and AGS

cell co-culture supernatant IL-8 concentrations from previous work done in the group. Conclusions previously drawn from the group's work reported that PCR typing is not adequate to predict disease outcomes in patients infected with *H. pylori*. Rearrangements and deletions in *cagPAI* genes could be present, some genes may not be expressed, resulting in a non-functional T4SS. This conclusion supports another study with *H. pylori* strains from South African origin where the study reported that, the PCR typing method for *cagA* gene and other *cagPAI* genes was a poor predictor of the *cag* T4SS phenotype (Argent et al., 2004).

ABRicate software is not the only bioinformatics tool available for use to search for virulence genes, but its ability to allow user to create custom script of gene(s) made it a choice of use. ABRicate software has been reported a good bioinformatics tool used to detect *tfs* ICE, *comB* and *cagPAI* in strain *H. pylori* of Vietnamese origin (Phuc et al., 2021b).

3.4.1 Comparison of *cagA* PCR typing and genomic data (% accuracy, sensitivity, and specificity)

PCR typing a detected *cagA* gene in 18 isolates (18/24; 75%) while genome analysis with ABRicate detected *cagA* in 17 (17/24; 71%). Comparison of these methods in respect to accuracy, defined as the level of agreement between two methods (Van Stralen et al., 2009), showed the sensitivity and specificity to be 83% and 50%. Sensitivity and specificity may be used to determine accuracy of a diagnostic test in clinical setting (Van Stralen et al., 2009). The accuracy of both methods to detect a *cagA* gene in clinical isolates was calculated as 75%. As

promising as the accuracy of both methods appear, to use either of these in clinical diagnosis, several other factors need to be considered because these factors can influence the result in both clinical and laboratory settings. For example, in PCR techniques, use of insufficient amounts of template DNA of appropriate molecular weight, denaturation of DNA, inability of polymerase enzyme to bind and the presence of inhibitors in the DNA sample may affect the accuracy of PCR (Demeke and Jenkins, 2010). Presence of inhibitors may cause modifications in PCR amplification by chelating Mg^{2+} (a cofactor for DNA polymerase) or by binding directly with DNA which may reduce the efficiency or reproducibility of PCR (Demeke and Jenkins, 2010). The sensitivity of PCR typing methods has improved through technical advances such as real-time PCR, which has been reported to be more sensitive than conventional PCR (Valones et al., 2009). Despite the inexhaustible number of factors that may affect PCR typing methods, this work showed that PCR genotyping had a high accuracy in detecting the *cagA* gene in clinical isolates.

3.4.2 Comparison of *cagA* PCR typing and *cagA* detection in genomes

Five out of 7 of isolates selected for *cagA* alignment had the proper alignment with *H. pylori* reference strain 26695. This was confirmed with detection of the expected EPIYA motifs in the aligned protein sequences and the aligned ORFs. This suggests that the software was able to detect *cagA* genes in these isolates, that subsequently may express CagA protein. Conversely, in two clinical isolates, 93C and 326C, the

sequences were unable to yield a predicted CagA amino acid sequence from a bioinformatic translation tool. The predicted protein sequences from these isolates when aligned with the 26695 reference strain CagA showed percentage gaps (93C/26695: gap 97.0%, 326C/26695: 95.1%) detected from nucleotide deletions and insertions. This indicates that the software can detect a gene, but the gene may not result in expression of the protein. This would potentially explain why the patient did not make an anti-CagA antibody response. Several other factors including host genetic make-up, colonization density, and the gastric microflora (Baj et al., 2021) have also been reported to affect the immune response to *H. pylori* in infected patients.

To further evaluate the ability of ABRicate software, a well annotated *H. pylori* strain PMSS1, known to have multiple copies of *cagA* gene in its genome, was analysed with ABRicate using the default strain (26695). ABRicate detected four copies of *cagA* gene in the strain and these were predicted to result in full length CagA proteins. Also, the translated proteins aligned with the CagA protein of reference strain 26695. This detection by ABRicate further gives reliability on the use of the software to detect other virulence genes such the *tfs* genes in clinical isolates. Further analysis of these four copies revealed the conserved EPIYA motif which contributes to variable size of CagA in different strains of *H. pylori* because of tyrosine phosphorylation at this motif. Several studies have reported the association of multiple tyrosine phosphorylation motifs (EPIYA motifs) with disease outcomes (Argent et al., 2004, Yamaoka et

al., 1998, Yamaoka et al., 1999a, Stein et al., 2002, Higashi et al., 2002, Azuma et al., 2002). For example, Yamaoka and colleagues reported that strains with more than three repeat regions had significantly higher gastric mucosal atrophy and intestinal metaplasia scores than those with fewer repeat regions (Yamaoka et al., 1999a). The detection of conserved EPIYA motifs in PMSS1 strain further established more reliability on the use of ABRicate software.

ABRicate was able to detect a *cagE* gene in more than half of the clinical isolates, which agreed with previous PCR typing data. This detection gives more confidence and reliability in the use of ABRicate for further analyses to search for *tfs* genes. Furthermore, IL-8 concentrations from co-cultures of AGS cells and clinical isolates indicative of a functional T4SS agreed with detection of a *cagE* gene using ABRicate software in all five clinical isolates. IL-8, a potent activator of T cells and neutrophils, plays a crucial role in the pathogenesis of inflammatory diseases (Sharma et al., 1998). A stronger gastric IL-8 response is elicited by *cagPAI* positive strains of *H. pylori* compared to *cagPAI* negative strains, and mutation of the *cagE* gene rendered the T4SS non-functional, causing a marked reduction in IL-8 in expression by epithelial cells *in vitro* (Sharma et al., 1998). Although the sample size for the IL-8 data was small (5 clinical isolates), the genomic data was consistent. This further increases the effectiveness and reliability to use the software in further analysis to include analysis *tfs* genes.

3.4.3 Comparison of CagA serology and genomic data

ABRicate software detected a *cagA* gene in 5 clinical isolates from patients where the CagA serology data was reported as negative. This conflicting result raised important questions about whether the CagA protein might not be expressed when the gene was detected by the software. Gene expression has been reported to pave way for isolation of important genes and reveal the functional characteristics of such genes (Frommer and Ninnemann, 1995). The major controller of gene expression is the transcription initiation which is mostly dependent on core promoter, binding site for RNA polymerase II, enhancers and silencers (Frommer and Ninnemann, 1995). It is possible that non-expression of CagA protein in isolates may be due to DNA sequence differences in these isolates. Sequences (approximately 300 bp) upstream of the *cagA* gene were aligned with approximately 300 bp upstream of *cagA* DNA sequence from the 26695 reference strain. The result showed aligned DNA sequences (Fig 3.1) with conservation of the Pribnow-Schaller box DNA sequence “TATAGT box” (Wasylyk, 1985) at -10 and “AATAAGATA” at -35 positions upstream of transcription start site. In Prokaryotes, initiation of transcription is usually controlled by conserved sequence motifs at different sites on the promoter region and these sequences are different from those found in eukaryotes (Menon et al., 2021). The -10 Pribow-Schaller box sequence TATAAT followed by -35 conserved DNA sequence TTGACA upstream of transcription start site is crucial in gene expression in prokaryotes (Menon et al., 2021). Any deletion or alteration in these sequences has been reported to affect or

alter the transcription process at initiation start site as both Pribnow-Schaller box sequence and conserved -35 position sequence are spatially related (Wasylyk, 1985). Fig 3 from Section 3.3.1.1 shows nucleotide replacement TATAGT at -10 position in 3 isolates and -35 position with entirely different conserved DNA sequence AATAAGATA instead of TTGACA. Both replacements at -10 and -35 position may strongly suggest why serology data for CagA was reported for the 5 clinical isolates. Therefore, ABRicate software provided some clearer understanding of DNA sequences from these isolates, which could not be explained by PCR genotyping or serological data. Another explanation for why serology may have been reported negative in patients infected with these isolates may be due to inactivation in one of the *cagPAI* genes. Inactivation in a single gene of *cagPAI* was reported to alter CagA translocation and phosphorylation (Seydel et al., 2002).

3.5 General applicability of ABRicate - Pros and Cons

PCR typing is an enzyme-driven technique for amplifying small DNA segments. The method is based on knowing at least some target DNA sequences ahead of time and using them to construct oligonucleotide primers that hybridise selectively to those sequences. In PCR, a thermostable DNA polymerase enzyme copies the target DNA in the presence of nucleotides and primers. The target DNA is amplified exponentially using numerous heating and cooling cycles in a thermocycler to produce rounds of target DNA denaturation, primer hybridization, and primer extension (Yang and Rothman, 2004). Over the

years, PCR typing has been extensively upgraded to improve its performance and reliability. With the comparison of PCR typing and genomic methods, PCR typing data as compared with genomic data with the use of ABRicate software in this work has further given more confidence that, PCR typing continues to be an alternative in diagnostic setting due to its extreme sensitivity, incredibly low-cost, and adaptability to high-volume demands. The procedure is quick, easy, and does not require any manual labour. The approach is usually highly reproducible in any fully equipped laboratory, as long as careful attention is paid to avoid contamination (Bricker, 2002).

As reported by Phuc and Colleagues (2021), the ABRicate software was effective to search for virulence genes although setting it up was a little challenging due to no prior knowledge of bioinformatics skills and coding. These challenges were overcome through support and help from bioinformaticians from the group, and researchers at the University of Liverpool who helped with installation of the software. The software reliability was shown in its percentage accuracy, sensitivity, and specificity of both *cagA* and *cagE* genes in clinical isolates as compared to available PCR typing data of *cagA* and *cagE*. For *cagA* gene, its accuracy was 75%, sensitivity was 83% and specificity 50% while for *cagE*, accuracy, sensitivity, and specificity were calculated to be 79%, 89% and 50% respectively. One of its major setbacks is that the software cannot detect any *tfs* genes at this stage because of its default reference strain (26695). The reference strain 26695 used to set-up the virulence

factor database (VFDB) of the software encodes fragmented ICEs that lack the majority of *tfs* genes. Therefore, this is a limitation to the software, but custom searches set up with reference strains of *H. pylori* with *tfs* genes could be used as alternative ways to detect *tfs* genes in clinical isolates. The latter required more skills in the use of bioinformatics, but was explored for reporting presence of *tfs* genes in the next chapter.

3.6 Conclusion

The results in this chapter provide confidence in the reliability of using PCR typing and ABRicate genomics analyses in the determination of virulence factors present in *H. pylori*. Analysis of the genomic data explained the cases where there was no match between PCR typing and phenotypic data on clinical isolates.

CHAPTER

FOUR

Bioinformatics Analyses of Tfs ICEs
type IV secretion (T4SS) genes in
clinical isolates

4.0 Introduction

4.0.1 Analysis of *tfs vir* genes using bioinformatics to predict virulence effect.

Despite poor understanding about the functional role played by *tfs* ICEs in disease outcomes, some virulence genes encoded by *tfs3* and *tfs4* remain relevant in defining disease outcomes (Alandiyjany et al., 2017). To investigate whether any components of *tfs* ICEs are associated with severe disease in infected patients, it is necessary to establish which genes are present and functional in patient's isolates. To do this, the bioinformatics tool ABRigate (described in chapter three) was used to search for virulence genes in *tfs* ICEs and investigate their association with disease in patients. A second online bioinformatics algorithm (ICEfinder) was used to compare and validate ABRicate's results.

ICEfinder is a web-based tool developed by Liu and colleagues in 2012 with the latest update in the year 2018 (Liu et al., 2019b). It is a standalone tool designed to facilitate the detection of integrative and conjugative elements (ICEs) and integrative and mobilizable elements (IMEs) in sequenced bacterial genomes. The tool was selected because it is freely available online, easy to use and requires no prior bioinformatics skills before use. ICEfinder is incorporated in the ICEberg database that contains the known details of 1,032 ICEs (Liu et al., 2019b). The tool analysis data for signature sequences of recombination and conjugation genes using the profile of Hidden Markov Model (HMM). It then searches for the *oriT* region using the approach proposed by

oriTfinder as reported by (Li et al., 2018). Then it co-localizes, filters and groups the corresponding genes. Finally, those elements carrying an integrase gene, a relaxase gene, and T4SS gene clusters are considered as putative integrative and conjugative elements (ICEs), while those without T4SS but with integrase and relaxase are referred to as putative integrative mobilisable elements (IMEs) – mobilizable elements that do not carry their own conjugative machinery but instead depend on conjugative machinery of the ICEs. ICEfinder employs a computer program called ARAGORN based on sequence homologue search as described by (Li et al., 2018) with the default parameters to identify the 3' termini of the *tRNA* / *tmRNA* genes as the putative ICE insertion sites and Vmatch - a versatile software tool for efficiently solving large scale sequence matching tasks (Kurtz, 2003) with the default options to detect the directed repeats (DR) as the *tRNA*-distal boundaries.

The genome of *H. pylori* has been reported to encode four types of T4SSs (Phuc et al., 2021a) – the *cagPAI* , *comB* , *tfs3* and *tfs4* T4SSs. The latter two, which were more recently discovered in strains of *H. pylori* that harbour either both or one, have been implicated in disease outcomes. For example, the homologue *virB4* gene (*dupA*) (Fig. 4.1) encoded on *tfs4* has been implicated in increased risk of duodenal ulcer disease in some populations. This suggests that the *tfs4* T4SS assembly may play an important role in *H. pylori* disease severity. Despite this link with disease outcomes, the functional activity of the *tfs4* T4SS assembly remains unclear. Strains reported to have the *dupA virB4* homologue in

tfs4 (*dupA* positive strains) have been demonstrated to upregulate IL-8 expression by gastric epithelial cells (Alandiyjany et al., 2017, Weiss et al., 2019).

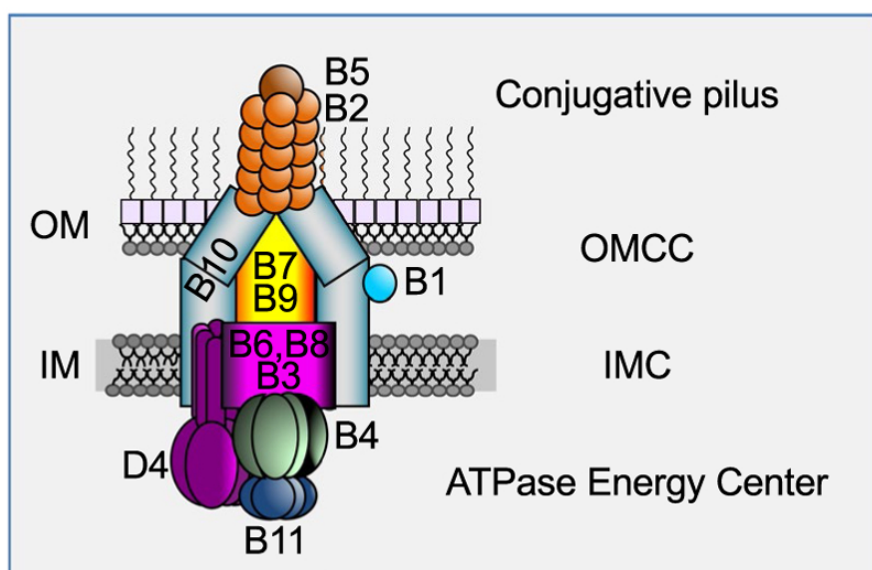


Figure 4.1: Schematic representation of T4SS VirB/D proteins assembly in the membrane.

Abbreviations used: OM – outer membrane, IM – inner membrane, IMC – inner membrane complex and OMCC – outer membrane core complex. Reproduced from (Costa et al., 2021)

The *dupA* gene was first reported by Lu and colleagues (Lu et al., 2005) as a continuous gene covering genes *jhp0917* - *jhp0918* from *H. pylori* strain J99 (Abadi and Perez-Perez, 2016). Following the primary study by Lu and colleagues, multiple studies (Nguyen et al., 2010, Shiota et al., 2010, Imagawa et al., 2010, Youssefi et al., 2020) have linked *dupA* positive status to *H. pylori* related disease (duodenal ulcer and gastric cancer) in some populations. Advances in genomic analyses have revealed the presence of an extra 600 bp in the *dupA* open reading frame of *H. pylori* G27 strain as compared the J99 strain (Abadi and Perez-

Perez, 2016). This observation revealed that the sequence length of *dupA* differs in *H. pylori* strains. Consequently, the genotypes of *dupA* are classified into two main types, namely – the long-type (with the additional 600 bp) and the short-type (Abadi and Perez-Perez, 2016). Several mutations resulting in different *dupA* gene length have also been reported, which can result in the insertion of premature stop codons, having considerable effects on the type of protein produced (Abadi and Perez-Perez, 2016). *DupA* of the long-type without any frameshift mutations has been reported to be associated with development of gastric cancer in infected patients compared with the short-type (Abadi and Perez-Perez, 2016).

4.1 Rationale

As described previously (section 3.2.8), genomic analysis using the ABRicate tool was used in comparison with PCR typing data of *cagA*, *cagE* and other data on the *cagPAI* available from previous work done in the group to characterise a collection of *H. pylori* clinical isolates. Similarly, further data from PCR typing of *dupA* was available for comparisons with bioinformatic analysis of genome sequences. Therefore, the ABRicate software was used to search for *dupA* in the genome sequences of clinical isolates and its outputs compared with available PCR typing data. The software would then be used to search for additional *tfs4* and *tfs3* genes in clinical isolates and the data will be compared with results from the second algorithm (ICEfinder). The work in this chapter aimed to use genomic analysis of Nottingham clinical

isolates to study the correlation of *tfs3* and *tfs4* clusters of genes with disease outcomes in patients presenting with *H. pylori* infection. This was the ultimate goal in this chapter, but the small sample size of clinical isolates made it impossible to establish correlations between disease outcome and virulence factors of *tfs* ICEs. Therefore, this chapter aimed to predict virulence factors in both *tfs3* and *tfs4*.

4.2 Materials and Methods

Some materials and methods have already been fully described in the previous chapter (section 3.2). In summary, ABRicate software was configured according to manufacturer's procedure (Seemann, 2018), as described in section 3.2.8 and custom searches were embedded into the software with *vir* genes of both *tfs3* and *tfs4* from reference strains PeCan18 for *tfs3* genes (database named as PeCan18_nucl) and, P12 and Shi470 for *tfs4* genes (databases named as P12_nucl and Shi470_nucl) respectively.

4.2.1 Analysis of *dupA* in clinical isolates

The custom searches created were used to search for *dupA* gene in clinical isolates by running the following code: `% abricate --db Shi470_nucl *.fasta > ../data-results/results.tab` - code representing example of a run with Shi470_nucl database. Multiple sequence alignment was done with Clustal Omega found at <https://www.ebi.ac.uk/Tools/msa/clustalo/> after translated using translate tool found at <https://web.expasy.org/translate/>.

4.2.2 Analysis of *tfs4* homologous genes in isolates.

Both custom searches of Shi470 and P12 called “Shi470_nucl” and “P12_nucl” created by the candidate with *tfs4* genes were used to search for homologues in all clinical isolates using the below code (with clinical isolates 45A as example).

```
abricate --db Shi470_nucl 45C_contigs.fasta >
../results-Shi470/45C-result.txt

abricate --db P12_nucl 45A_contigs.fasta > ../results-
P12/45A-result.txt.
```

4.2.3 PCR detection of *dupA*

For the analysis of *dupA* gene in isolates, the following primers were used for PCR amplification.

*dupA*_RI_F 5'-TACTAACAGACTTTGAGCCTA-3'

*dupA*_RI_R 5'-GATTGGATATTCCCATCGC-3'

The 20 µl PCR reactions contained 10 µl DreamTaq (ThermoScientific DreamTaq Green PCR Master Mix (2X), 4.0 µl each of the forward and reverse primer (0.12 µM in reaction), 1 µl nuclease-free water, and 1 µl genomic DNA template.

As stated in table 4.1 below, after an initial denaturation step at 95°C for three minutes, each reaction consisted of 30 cycles of 95°C for 30 seconds, 55°C for 30 seconds, and 72°C for 45 seconds. Following a final extension step at 72°C for 10 minutes, each PCR mixture was subjected to gel electrophoresis on 1% agarose gels and a 100 bp DNA

ladder (Roche Diagnostics, Johannesburg, South Africa) was used as a size marker.

Table 4.1: cycling conditions for PCR with dupA_RI primers

Step	Number cycles	Time	Temperature
Initialisation	1×	3 min	95 °C
Denaturation	30×	30 sec	95 °C
Annealing		30 sec	55 °C
Extension		45 sec	72 °C
Final Extension	1×	10 min	72 °C

The PCR amplification from genomic DNA of *H. pylori* clinical isolates was done by Joanne Rhead, as previously described by (Hussein et al., 2010).

4.2.4 PCR typing for some selected *tfs4* genes

Dr Rob Delahay and Nicola Croxall also undertook PCR typing for selected *tfs4 vir* genes from clinical isolates but unfortunately, the primers and conditions used were not available. Despite this, the PCR typing data available from Dr Delahay will be compared with genomic data to further investigate the level of agreement and reliability between both methods.

4.2.5 Analysis of *tfs4* homologous genes in isolates using ICEfinder.

A second algorithm tool, ICEfinder, was used to search for *tfs4* homologue genes in isolates. Assembled contigs of clinical isolates (as described in method 3.2.7) were uploaded on to the online tool and default subject datasets (T4SS-type ICEs / AICEs / IMEs – integrase, type IV coupling protein, replication initiator protein, *oriT*, relaxase, translocation protein and VirB4/TraU or full T4SS, and accessory modules - virulence factors and acquired antibiotic resistance determinants) were searched for with a single run. Virulence factors comparison was performed with a BLASTp search using the integrated ICEberg database with the following parameters – blast hits per gene to be mapped <100, sequence coverage of BLAST hits >60, percentage identity BLAST hits >30, distance between genes in locus <10 and weight of synteny conservation in hit sorting = 0.5. Gene cluster alignment with reference strain (P12) was performed with Multigene blast embedded in the online tool. Alignment was only performed with strain P12 for *tfs4* because the integrated ICEberg database is limited in that not every *H. pylori* sequence strain till-date is present in the database (alignment with Shi470 was not possible). ICEfinder is available at <https://bioinformml.sjtu.edu.cn/ICEfinder/ICEfinder.html>

4.3 Results

4.3.1 Detection of *dupA* in clinical isolates

Genomic analysis detected the presence of the *dupA* gene in seven isolates and it was not found in a further 18 isolates (Table 4.2). The long type *dupA*, corresponding to the size (2,500 bp) of *dupA* in reference strain (Shi470) was detected in the genome sequences of seven isolates. PCR typing and genomic data was in agreement for 18 *dupA* negative isolates and for six of the *dupA* positive isolates. However, for isolate 265A, a long-type *dupA* gene was present in the genome but a negative result was obtained with PCR typing.

Table 4.2: Comparison of *dupA* status in genome sequences and PCR genotyping data of *H. pylori* clinical isolates.

Patient ID	<i>dupA</i> status in genomes	<i>dupA</i> PCR status	PCR agrees with genome – for <i>dupA</i> status?	<i>dupA</i> type in genomes	Size of <i>dupA</i> (bp) in PCR
45A	Negative	Negative	Yes		
45C	Negative	Negative	Yes		
77A	Negative	Negative	Yes		
93C	Negative	Negative	Yes		
120A	Negative	Negative	Yes		
120C	Negative	Negative	Yes		
194A	Positive	Positive	Yes	long	2,498
194C	Positive	Positive	Yes	long	2,498
249C	Negative	Negative	Yes		
265A	Positive	Negative	No	long	2,480
295C	Negative	Negative	Yes		
308A	Negative	Negative	Yes		
308C	Negative	Negative	Yes		
322C	Negative	Negative	Yes		
326C	Negative	Negative	Yes		
439A	Positive	Positive	Yes	long	2,517
439C	Positive	Positive	Yes	long	2,499
444A	Positive	Positive	Yes	long	2,516
444C	Positive	Positive	Yes	long	2,516
495A	Negative	Negative	Yes		
495C	Negative	Negative	Yes		
537A	Negative	Negative	Yes		
621A	Negative	Negative	Yes		
732A	Negative	Negative	Yes		
732C	Negative	Negative	Yes		

dupA genomic data agreed with PCR typing in 6 isolates (shaded orange) - 194A, 194C, 439A, 439C, 444A, 444C. Isolate 265A was positive in the genome but negative by PCR typing. 17 isolates were negative for a *dupA* gene using both genomic and PCR typing methods. A *dupA* long-type was detected in 6 isolates while *dupA* sort-type was not detected in any of the clinical isolates. Abbreviations used: bp- base pair, C- corpus region, A-antrum region of the stomach.

4.3.2 Identification of *tfs4* genes in clinical isolates using Shi470 as reference strain

To search for the presence of *tfs4* genes in clinical isolates, the *tfs4 vir* genes from reference strains Shi470 and P12 as listed in the top row of Table 4.3 were retrieved and used in custom searches as described in method 4.2.2. Shi470 strain was selected because of the presence of full cluster of *tfs vir* genes in its genome, and P12 as a second reference strain for custom searches because it has intact *tfs4* genes except for *virB4* due to frameshift. T4SSs are commonly composed of a set of conserved 12 genes which are needed in Gram-negative bacteria to build a fully functional secretion system (Costa et al., 2021). These conserved genes are designated VirB1 – VirB11 and VirD4 according to the prototype in *Agrobacterium tumefaciens* VirB/VirD4 T4SS (Costa et al., 2021). Of all the 12 T4SS genes, the bolded genes in Table 4.3 *virB2* - *virB11* and *virD4* are essential for substrate translocation in both assembly and function of the T4SS system. From the 26 clinical isolates analysed, there was no detection of *tfs4 vir* genes in 6 of the isolates – 45A, 45C, 77A, 295A, 295C and 326C. Isolates 194A, 194C (from same patient) and 439A and 439C had a full cluster of *tfs4 vir* genes with exception of the *methylase* (Table 4.3). Isolates 444A and 444C (also from the same patient) only lacked the *xer*, *virB6* and *methylase* genes (Table 4.3). *virB4*, *virB11* and *virD4* the three important family of ATPases in the T4SS assembly were present in seven isolates (194A, 194C, 265A, 439A, 439C, 444A and 444C). *virB9* and *virB10* which form part of the outer membrane assembly were present in twelve isolates

(Table 4.3). *virD2*, another core relaxase in the assembly, was detected in seventeen isolates. The *topA* gene was detected in seven isolates (194A, 194C, 439A, 439C, 444A and 444C).

Table 4.3: Identification of *tfs4* homologues (using Shi470 reference sequence) in clinical isolates.

T4SS genes /Isolates	<i>xer</i>	<i>virB2</i>	<i>virB3</i>	<i>virB4</i>	<i>virB6</i>	<i>virB7</i>	<i>virB8</i>	<i>virB9</i>	<i>virB10</i>	<i>VirB11</i>	<i>virD2</i>	<i>virD4</i>	<i>pz5</i>	<i>topA</i>	<i>methylase</i>
45A															
45C															
77A															
93A								Y	Y	Y	Y				
93C								Y	Y	Y	Y				
120A	Y										Y				
120C	Y										Y				
194A	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	
194C	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	
249C	Y				Y						Y	Y			
265A		Y	Y	Y		Y		Y	Y	Y	Y	Y	Y	Y	
295A															
295C															
308A								Y	Y	Y	Y	Y			
308C								Y	Y	Y	Y	Y			
322A								Y	Y						
326C															
439A	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	
439C	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	
444A		Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	
444C		Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	
495A								Y			Y				
495C								Y			Y				

537A	Y				Y						Y				
621A	Y				Y										
732A	Y										Y				

Full cluster of *tfs4* T4SS genes, *virB2* – *virB11* and *virD4* were detected in 4 isolates – 194A, 194C, 439A and 439C. Partial cluster of *tfs vir* genes, *virB2*, *virB3*, *virB4*, *virB7-virB11* and *virD4* were detected in 2 isolates – 444A and 444C. Isolate 265A has all other *vir* genes but not *virB6* and *virB8*. >60% sequence coverage of *vir* genes of reference strain Shi470 was used as cut-off to report *vir* genes in isolates. Abbreviations used: C-corpus , A-antrum, Y=yes, T4SS-type IV secretion system (strains highlighted in red show where more than 10 *tfs* genes were discovered).

4.3.3 Pairwise Alignment of *vir* genes detected in isolates with homologues genes in Shi470

Nucleotide sequences of detected *tfs4 vir* genes (*virB2* – *virB11*, *virD2* and *virD4*) were retrieved and paired with nucleotide sequences of *vir* genes of reference strain Shi470 as described in section 4.2.1. The results showed that sequences of *virB2* – *virB11* and *virD4* in isolate 194A and 194C had over 90% identity with homologous gene sequences from reference strain (Table 4.4). Over 90% sequence identity was also observed in the ATPase genes (*virB4*, *virB11* and *virD4*) of isolate 194C and 439C when compared with those of the reference strain (Table 4.4). These isolates had the full cluster of *tfs vir* genes (*virB2* – *virB11* and *virD4*). Ten isolates had variable sequence similarities between 43.5% to 98.4%, with detection of partial clusters of *vir* genes.

Table 4.4: Percent identity from pairwise sequence comparison of T4SS nucleotide sequences – Shi470 tfs4 vir genes against the sequences of vir genes from clinical isolates

T4SS genes/ isolates	<i>xerD</i>	<i>virB2</i>	<i>virB3</i>	<i>virB4</i>	<i>virB6</i>	<i>virB7</i>	<i>virB8</i>	<i>virB9</i>	<i>virB10</i>	<i>virB11</i>	<i>virD2</i>	<i>virD4</i>	<i>pz5</i>	<i>topA</i>	<i>methylase</i>
45A															
45C															
77A															
77C															
93A								96	67.7	96	46.3				
93C								96	67.7	96.1	46.3				
120A	41.5										94.6				
120C	64.8										94.6				
194A	97.3	98.3	98.5	98.3	46.9	99.2	98.8	96.8	97.8	98.7	59.5	97.9			
194C	49.3	98.3	96.2	98.3	98.3	99.2	98.8	96.8	97.8	98.8	59.5	97.9			
249C	98.1				47.3						45.7	45.9			
265A						48.1		45.9	44.7	61.4	47	47.2			
295C															
308A								94.7	69.7	97.2	94.7	98.4			
308C								94.7	69.4	97.2	94.7	98.4			
322A								95.7	68.9						
326C															
439A	98.6	44.6	99.2	48.4	47.3	45.6	46.3	45.2	44.5	49.6	46.2	45.6			
439C	49.7	44.6	99.2	98.4	47.3	48.1	45.6	45.2	41.5	96.8	95.8	97.9			
444A		43.5	45.8	48.6		46.2									
444C															
495A															
495C															
537A															
621A															
732A															

Abbreviations used: C-corpus, A-antrum, T4SS-type IV secretion system

4.3.4 Identification of *tfs4* genes in clinical isolates using P12 as the reference strain

Table 4.5 shows the *tfs4 vir* genes detected in clinical isolates using ABRicate as described in section 3.2.2. A full cluster of *tfs4 vir* genes (*virB2-virB11*) was detected in 2 isolates (308A and 308C), however this patient had no gastro-duodenal disease. *virB6 – virB10*, known to possibly form the translocation channel in the T4SS complex (Wallden et al., 2010) were present in 6 isolates - 93A, 93C, 308A, 308C, 322A and 322C (Table 4.5). Of these patients, only 322 had gastro-duodenal disease (gastric ulceration). *virB11* and *virD4* ATPases and *virD2* relaxase were found in 9 isolates – 194A, 194C, 265A, 308A, 308C, 439A, 439C, 444A and 444C. Of these patients, 194 had a duodenal ulcer, and 444 had a gastric ulcer.

Table 4.5: Identification of *tfs4* homologues in clinical isolates, using strain P12 as the reference sequence.

T4SS genes /Isolates	Patient disease	<i>xerD</i>	<i>virB2</i>	<i>virB3</i>	<i>virB6</i>	<i>virB7</i>	<i>virB8</i>	<i>virB9</i>	<i>virB10</i>	<i>VirB11</i>	<i>virD2</i>	<i>virD4</i>	<i>pz5</i>
45A	None												
45C													
77A	DU												
93A	None	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y
93C		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y
120A	None										Y		Y
120C											Y		Y
194A	DU	Y	Y	Y				Y	Y	Y	Y	Y	Y
194C		Y	Y	Y				Y	Y	Y	Y	Y	Y
249C	DU										Y	Y	Y
265A	None		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y
295A	None												
295C													
308A	None	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
308C		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
322A	GU	Y	Y	Y	Y	Y	Y	Y	Y				Y
322C		Y	Y	Y	Y	Y	Y	Y	Y				Y
326C	None												
439A	None							Y	Y	Y	Y	Y	Y
439C								Y	Y	Y	Y	Y	Y
444A	GU	Y			Y			Y	Y	Y	Y	Y	Y
444C		Y			Y			Y	Y	Y	Y	Y	Y
495A	None	Y			Y						Y		Y
495C		Y			Y						Y		Y
537A	None										Y	Y	Y
565C	None												
621A	None												
732A	None										Y		Y
732C											Y		Y

A full cluster of *tfs4* T4SS genes (*virB2* – *virB11* and *virD4*) was detected in 2 isolates – 308A and 308C. A full cluster without *virD4* was detected in 2 isolates 93A and 93C. Partial clusters of *tfs4 vir* genes were found in other isolates. >60% sequence coverage of *vir* genes of reference strain P12 was used as cut-off to report *vir* genes in isolates. Gastro-duodenal disease status in the patients is indicated. Abbreviations used: C-corpus , A-antrum, Y=yes, T4SS – type 4 secretion system, DU – duodenal ulcer, GU – gastric ulcer

4.3.5 Identification and comparison of *tfs4* genes using P12 as reference strain with ICEfinder tool

As a second method, ICEfinder was used to search for *tfs4* cluster of genes in selected isolates with P12 as reference strain. All isolates where full or partial *tfs4* cluster of genes were detected using ABRicate software were included (isolates 93A, 93C, 194A, 194C, 265A, 308A, 308C, 322A, 322C, 444A, 444C, 495A and 495C). These isolates were selected to compare both their ABRicate and ICEfinder results. ICEfinder detected no *tfs4* genes in seven isolates (93A, 93C, 194C, 265A, 322A, 322C and 495C) but found *tfs4 vir* genes in seven isolates with majority of genes in agreement with ABRicate result (Table 4.6 coloured-coded with green and “B” to indicate where there is agreement between ABRicate and ICEfinder). Colour-coded with yellow and “A” shows where only ABRicate detected *tfs4 vir* genes in isolates.

Table 4.6: Identification and comparison of *tfs4* homologues using ABRicate and ICEfinder methods

T4SS genes / Isolates	Patient disease	<i>xer</i>	<i>virB2</i>	<i>virB3</i>	<i>virB4 (fs)</i>	<i>virB6</i>	<i>virB7</i>	<i>virB8</i>	<i>virB9</i>	<i>virB10</i>	<i>VirB11</i>	<i>virD2</i>	<i>virD4</i>	<i>pz5</i>
45A	None													
45C														
77A	DU													
77C														
93A	None	Y (A)	Y (A)	Y (A)		Y (A)	Y (A)	Y (A)	Y (A)	Y (A)	Y (A)	Y (A)		Y (A)
93C		Y (A)	Y (A)	Y (A)		Y (A)	Y (A)	Y (A)	Y (A)	Y (A)	Y (A)	Y (A)		Y (A)
120A	None											Y (A)		Y (A)
120C												Y (A)		Y (A)
194A	DU	Y (B)	Y (B)	Y (B)					Y (B)	Y (B)	Y (B)	Y (B)	Y (B)	Y (B)
194C		Y (A)	Y (A)	Y (A)					Y (A)	Y (A)	Y (A)	Y (A)	Y (A)	Y (A)
249C	DU											Y (A)	Y (A)	Y (A)
265A	None		Y (A)	Y (A)			Y (A)	Y (A)	Y (A)	Y (A)	Y (A)	Y (A)	Y (A)	Y (A)
295A	None													
295C														
308A	None	Y (B)	Y (B)	Y (B)		Y (B)	Y (A)	Y (B)	Y (B)	Y (B)	Y (B)	Y (B)	Y (B)	Y (B)
308C		Y (B)	Y (B)	Y (B)		Y (B)	Y (A)	Y (B)	Y (B)	Y (B)	Y (B)	Y (B)	Y (B)	Y (B)β
322A	GU	Y (A)	Y (A)	Y (A)		Y (A)	Y (A)	Y (A)	Y (A)	Y (A)				Y (A)
322C		Y (A)	Y (A)	Y (A)		Y (A)	Y (A)	Y (A)	Y (A)	Y (A)				Y (A)
326C	None													
439A	None								Y (A)	Y (A)	Y (A)	Y (A)	Y (A)	Y (A)
439C									Y (A)	Y (A)	Y (A)	Y (A)	Y (A)	Y (A)
444A	GU	Y (B)				Y (B)			Y (B)	Y (B)	Y (B)	Y (B)	Y (B)	Y (B)
444C		Y (B)				Y (B)			Y (B)	Y (B)	Y (B)	Y (B)	Y (B)	Y (B)
495A	None	Y (B)				Y (B)						Y (B)		Y (B)
495C		Y (A)				Y (A)						Y (A)		Y (A)
537A	None											Y (A)	Y (A)	Y (A)
565C	None													
621A	None													
732A	None											Y (A)		Y (A)
732C												Y (A)		Y (A)

Colour code: green boxes with 'B' = Both ABRicate and ICEfinder agrees and yellow boxes with 'A' represent where only ABRicate detect *vir* genes in isolates. Abbreviations: Y- yes, (A)-ABRicate, (B) – both, A – antrum, C- corpus, DU – duodenal ulcer, GU – gastric ulcer

4.3.6 Comparison of PCR and genome data in selected isolates - *tfs4* genes.

Dr Rob Delahay also carried out PCR typing of selected *tfs4 vir* genes in clinical isolates (Table 4.7) using Shi470 and P12 as reference strains. Unfortunately, his PCR typing conditions and primers are not available, however the data generated was available. Therefore, a comparison of his PCR typing data for *tfs4 vir* genes (*xer*, *virB8*, *virB4*, *virB9*, *virD2*, *pz35* and *pz5*) with genomic data was carried out. The comparison was done in 13 isolates (Tables 4.7A and B). Isolates 194A, 444A and 444C had the *tfs4 vir* genes *virD2*, *virB9*, *virB8*, *virB4* and *pz5* (Table 4.7A). Interestingly, both of these patients had peptic ulcer disease. For isolate 308A and 308C (which originated from the same patient), both methods detected the *xer* gene, *pz5* and *virB8* using P12 as reference strain (Table 4.7B).

Table 4.7A and 4.7B: Comparison of PCR typing data of *tfs4* genes with genomic data on *tfs4* genes from reference strains Shi470 (A) and P12 (B), including patient gastro-duodenal disease status.

A

Patient ID	Patient disease	4480 (<i>xer</i>)	4490 (<i>pz35</i>)	4530 (<i>virD2</i>)	4595 (<i>virB9</i>)	4600 (<i>virB8</i>)	4615 (<i>virB4</i>)	4640 (<i>pz5</i>)
93C	None	Y/N	N/N	Y/Y	Y/Y	N/N	N/N	N/N
120C	None	N/Y	N/N	N/Y	N/N	N/N	N/N	N/N
194A	DU	N/Y	Y/N	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y
249C	DU	Y/Y	Y/N	Y/Y	?/N	?/N	N/N	N/N
265A	None	?/N	N/N	Y/Y	Y/Y	?/?	Y/Y	Y/Y
295C	None	N/N	N/N	Y/Y	?/N	N/N	N/N	N/N
308A	None	N/N	N/N	Y/Y	Y/Y	N/N	N/N	N/N
308C		N/N	N/N	Y/N	Y/Y	N/N	N/N	N/N
326C	None	N/N	N/N	N/N	N/N	N/N	N/N	N/N
444A	GU	N/N	N/N	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y
444C		N/N	N/N	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y
495A	None	N/N	N/N	Y/Y	N/Y	N/N	N/N	N/N
495C		N/N	N/N	Y/Y	N/Y	N/N	N/N	N/N

B

Patient ID	Patient disease	437 (<i>xer</i>)	441 (<i>pz35</i>)	464 (<i>virB8</i>)	473 (<i>pz5</i>)
93C	None	Y/Y	Y/N	Y/Y	Y/Y
120C	None	N/N	Y/N	N/N	N/N
194A	DU	Y/Y	N/N	N/N	Y/Y
249C	DU	?/N	?/N	?/N	N/Y
265A	None	N/N	N/N	Y/Y	Y/Y
295C	None	Y/N	N/N	Y/N	N/N
308A	None	Y/Y	Y/N	Y/Y	Y/Y
308C		Y/Y	Y/N	Y/Y	Y/Y
326C	None	N/N	N/N	N/N	Y/Y
444A	GU	Y/Y	Y/N	N/N	N/Y
444C		Y/Y	Y/N	N/N	N/Y
495A	None	N/Y	N/N	N/N	Y/Y
495C		N/Y	N/N	N/N	Y/Y

tfs4 genes representation - Ref Strain Shi470; 4480 (*xer*), 4490 (*pz35*), 4530(*virD2*), 4595 (*virB9*), 4600 (*virB8*), 4615 (*virB4*), 4640 (*pz5*) and Ref strain P12; 437 (*xer*), 441 (*pz35*), 464 (*virB8*), 473 (*pz5*). Percentage nucleotide coverage used >60%. PCR typing was done by Dr Rob Delahay. Agreement in both PCR typing and genome data was represented as Y/Y and N/N to denote Y and N for PCR and genome data respectively. “?” indicates an inconclusive result. Where ABRicate results agreed with PCR typing data, results were coloured-coded in orange. This comparison was done on selected isolates with PCR data available. DU – duodenal ulcer, GU – gastric ulcer

4.4 Discussion

Following the evaluation of ABRicate software in chapter 3, this chapter describes its application in characterising the *tfs3* and *tfs4* gene complement of *H. pylori* strains. This was done in three parts. First a specific analysis of *dupA*, a well-characterised virulence factor was undertaken, comparing PCR genotype with ABRicate data. Then the presence of other *tfs4* genes was assessed. The original aim was to correlate *H. pylori* virulence genes of *tfs3* and *tfs4* in clinical isolates with patient disease outcomes, however the number of samples was too low.

4.4.1 *dupA* status in genomes of isolates agreed with PCR typing data

ABRicate was used to search for *dupA*, a homologue of the *virB4* gene that has been implicated as a virulence biomarker in those infected with a *dupA*-positive strain (de Lima Silva et al., 2021). ABRicate detected *dupA* in seven isolates, six of which agreed with previous PCR typing data for *dupA*-positive isolates. Another important finding was that *dupA* was not detected in 18 isolates, which also agreed with *dupA*-negative isolates in the PCR typing data. The level of agreement in data generated by both methods provided confidence in their reliability. This supported the report by Wang and colleagues (Wang et al., 2015) showing that the PCR method provides high level of sensitivity and specificity, with great accuracy in clinical diagnosis over conventional tests. The result with isolate 265A where *dupA* was found in the genome but PCR typing was negative requires further investigation. This may be due to acquired mutations and rearrangements in the genome of isolate 265A. The *dupA*

forward primer position was not found when searched for in the sequence detected by ABRicate. Primer mismatch is one of the major challenges in PCR typing method.

4.4.2 Detection of *tfs* ICE genes and association with disease status in clinical isolates.

The gene clusters of *tfs3* and *tfs4* have been implicated in disease outcomes in individuals infected with *H. pylori* (Phuc et al., 2021a). Disease outcomes are highly influenced by multiple factors, including environmental, host genetics and bacterial virulence factors (Robinson et al., 2007). One of the bacterial factors that contributes to disease progression is the *virB4* homologue (*dupA*), known as an important marker for duodenal ulcer development. A study by Jung and colleagues (Jung et al., 2012) concluded that the presence of a complete *dupA* cluster (that is, presence of *dupA* with neighbouring *vir* genes - B2, B4, B8 and B10) is associated with development of duodenal ulcer as compared to *dupA* alone. A *dupA* cluster was detected in the *tfs4* ICE of clinical isolates in this study.

4.4.2.1 Detection of *tfs4 vir* genes and association with disease status in clinical isolates

The primary aim was to use genomic analysis of Nottingham clinical isolates to study the correlation of *tfs3* and *tfs4* clusters of genes with disease outcomes in patients presenting with *H. pylori* infection. Notwithstanding the absence of functional data *tfs4* T4SS genes, many

genes have been implicated in disease outcomes (Fischer et al., 2020). Furthermore, the presence of further *tfs4* genes with *dupA* which form a *dupA* cluster, is a better prediction for disease progression in patients. The result using both genomics and PCR typing showed both datasets agreed in six *dupA*-positive isolates. It was then important to investigate if neighbouring *tfs4 vir* genes were reported using genomic analysis on these isolates. The neighbouring *tfs4 vir* genes (*virB2*, *virB4*, *virB8* and *virB10*) were detected in 4 isolates (194A, 194C, 444A and 444C) from two patients, where A and C denote antrum and corpus sections of the stomach that biopsies were collected from. The most interesting finding was that a *dupA* cluster (*dupA* with *tfs4 vir* genes) was reported in these two patients who presented with duodenal ulcer (patient 194) and gastric ulcer (patient 444). Patient 265, who had no gastro-duodenal disease, was also infected with a *dupA* positive strain, but the neighbouring *virB8* gene was absent (Tables 4.5 and 4.6).

The majority of the sequences of the *tfs4 vir* genes detected by ABRicate were not translatable to amino acids *in silico*, due to apparent deletions, frameshifts, and insertions. This could explain why *dupA* clusters do not readily correlate with disease outcomes. With a much larger number of clinical isolates in the future, and availability of detailed metadata such as environmental factors (e.g. diet, smoking), correlations with disease may be possible using the methodology described.

4.4.2.2 Comparison of *tfs4 vir* genes using PCR typing and genome methods

One of the main aims was to compare results from PCR typing and genomic analysis. As earlier described in Section 4.2.4, PCR typing for selected *tfs4 vir* genes (*virD2*, *virB4*, *virB8*, *virB9*, *pz35* and *xer*), was mainly in agreement with the genomic data (Tables 4.7 and 4.8). Although the primer sequences and PCR conditions were not available, the comparison of genomics data further showed good PCR typing sensitivity.

4.4.2.3 Detection of *tfs3 vir* genes and association with disease status in clinical isolates

The only isolates where full cluster of *tfs3 vir* genes was detected originated from patient 194 (isolates from both antrum and corpus of the stomach). This patient presented with a duodenal ulcer, and the genomic data on their isolates detected a complete cluster of *dupA* gene and *tfs3 vir* genes. The presence of *tfs3* in just one out of 24 isolates is not surprising, as overall distribution of *tfs3* ICE has been reported to be generally low in global populations of *H. pylori* strains (Delahay et al., 2018b). This is usually due to fragmentation which results in an incomplete *tfs3* cluster of genes. It would have been interesting to provide more data on fragmented genes in *tfs* ICEs if the ABRicate tool was able to detect fragmented genes in both *tfs3* and *tfs4*, however the tool was not configured for this. The ABRicate tool was able to detect full clusters of *tfs3* and *tfs4* T4SS genes in isolate 194. The existence of both

tfs ICE in isolate 194 is an interesting finding, which supports earlier studies that both *tfs3* and *tfs4* can co-exist in a single strain of *H. pylori* (Delahay et al., 2018b). The co-existence may correlate with disease status, but drawing this conclusion in this study was not possible due to the small sample size, and other factors such as environmental, lifestyle and dietary that may also contribute to disease development.

4.4.3 XercD: tyrosine recombinases in clinical isolates

In many bacteria, the Xer recombination system is conserved and mainly functions to allow good splitting of the chromosome during replication. Many bacterial DNA genomes are circular, and each circular chromosome must be fully replicated to ensure that a full genome complement is inherited by each daughter cell during cell division. An ICE usually encodes recombinase functions that aid in the excision of the elements from their host chromosome. The *xer* gene was detected in four of the six strains with more than 10 *tfs4* genes. *tfs4* ICE is reportedly capable to form circular intermediates depending on the presence of putative site-specific recombinase, integration site-specific motif (5'-AAGAATG-3') (Weiss et al., 2019). This excision is capable of forming a circular intermediate which subsequently, may be transferred via conjugation into host cells (Weiss et al., 2019). Tyrosine recombinase family, *xer* gene was discovered in 10 isolates with Shi470 as reference and in 12 isolates with P12 as reference strain. Upon analysis, the integration motif conserved in these isolates is 5'-AAGATG with a missing base 'A' which may result from deletion as observed in the *xer*

sequences of isolates. Although the presence of this integration motif is crucial for site-specific recombinase, a recent study by Weiss and colleagues found that excision of the *tfs4* ICE by the integration motif with the presence of T4SS genes are not necessarily required for transfer of the ICE, suggesting the possibility of alternative routes. This suggests that transfer of *tfs4* ICE via unknown routes is possible. Also, it may be that homologous recombination is more efficient in strains of *H. pylori* than site-specific recombination as *H. pylori* undergoes high recombination process.

4.4.4. Challenges involved in sequencing methods of *Helicobacter pylori*

The increasing use of whole-genome sequences is rapidly allowing scientists to comparatively study genomes of many organisms including bacteria (Thorell et al., 2017). The genomes of *H. pylori* are highly variable and diverse due to the bacterial adaptation which is important for its survival in different population (Palau et al., 2021). The most commonly used Next-Generation Sequencing methods for genomes of *H. pylori* are the Illumina MiSeq and MiniON Nanopore sequencing technologies. Both are powerful tools used to determine bacterial genome composition (Heikema et al., 2020). MiniON nanopore sequencing, launched within the last five years, is a third-generation system used for direct, long-reads sequencing of single strands of DNA. Both sequencing methods have their advantages and disadvantages with error rate being one important factor to consider in their use (Stoler

and Nekrutenko, 2021). Sequencing most bacterial genomes using Illumina method uses short reads which are less error-prone but mainly difficult to assemble due to the short length of the reads. Assembling the genomes of *H. pylori* when sequenced with Illumina MiSeq method is challenging due to biases created during library preparation such as Illumina Nextera protocol where PCR amplification process can introduce biases against some regions in the genome. These biases can cause omission in some genomic regions which may not be included in the final sequencing output. Another challenge posed from using short-read sequencing method is difficulty that may result during assembly using *de novo* assembly algorithms when it fails to assemble regions in the genome with intergenic repeats sequences (Goldstein et al., 2019). This can result in fragmentation of these regions thus, leading to non-identification of key regions in the genome such as horizontal gene transfers, mobile genetic elements and important gene cluster (Goldstein et al., 2019). In contrast, MiniON nanopore sequencing uses long-reads which provides opportunity for the reads to span throughout genomic repeats, which help to resolve the complex regions in the genomes. The long-reads also help to avoid biases caused by PCR typing (Goldstein et al., 2019). Although the use of MiniON nanopore long-reads method is helping to resolve many challenges presented by short-reads methods such as Illumina MiSeq, this method is also disadvantaged in that its assemblies generate high error rate. Fortunately, the assembly can be polished with the use of either MiniON or Illumina reads to improve its quality. Therefore, the use of MiniON nanopore method only or in

combination with Illumina MiSeq have been reported and validated for use in sequencing of bacterial genomes with low or average GC content such as *H. pylori*.

As described in the section 3.2.7, all *H. pylori* isolates in this thesis were sequenced with Illumina MiSeq method. The assembled contigs generated from each strain were presented as nodes. Present between these nodes are connections called edges which are difficult to access (Wick et al., 2015). To access these connections and join the nodes, another bioinformatics software called “Bandage” (Wick et al., 2015) is required. This final step was not carried out.

4.4.5 Comparison of ABRicate with a second algorithm (ICEfinder) used to detect *tfs vir* genes in isolates.

As a second tool to compare with ABRicate, ICEfinder was used to search for *vir* genes in isolate genomes. The tool detected virulence genes in some isolates (308A, 308C, 444A, 444C, 193A, 295A and 495A) and the majority of genes were in agreement with results from ABRicate. However, there were some issues. Firstly, ABRicate was able to detect clusters of *vir* genes in isolates 193A and 193C (both from a single patient), but the ICEfinder tool was able to detect only partial *vir* genes in 193A, and no *vir* genes in 193C. This supports the findings of (Lee and Andam, 2022), where they stated that ICEfinder was unable to find near identical genes in closely related strains. In their report, they stated that recombination events taking place at multiple times on the

same chromosomal sites are difficult to identify using the ICEfinder tool. Therefore, the use of ICEfinder may underestimate the true extent of recombination found in nature in some species of bacteria (Lee and Andam, 2022). Furthermore, they reported that although the tool was able to detect virulence genes in the genomes of the three strains of *Staphylococcus* species (*S. agalactiae* – 0.53%, *S. pyogenes* – 2.56% and *S. suis* – 0.005%), it was unable to detect intact ICEs in the strains (Lee and Andam, 2022). This may be because the tool is not able to detect an integrase (*xer* gene) that promotes site-specific integration and excision for an intact ICE. This also correlates with the result reported in this study that, ICEfinder was able to detect virulence genes of T4SS but not able to detect open reading frame (ORF) in some isolates. Finally, the inability of ICEfinder to detect *vir* genes in some isolates may be due to an out-dated database embedded within the tool because the last update was reported in 2018 (Liu et al., 2019b).

4.4.6 Study conclusion, limitations, and future work.

This study originally aimed to use genomic analysis of *H. pylori* clinical isolates to study the correlation of *tfs3* and *tfs4* gene clusters with disease status in the patients. The methodology used in this study indicated that *tfs* T4SS gene (especially *tfs4*) prediction is a feasible approach. The major limitation, however, was the small sample size (total number of isolates 24). Another limitation is the inability to connect node edges in assembled contigs of sequenced isolates to improve the accuracy of ABRicate. This was not possible due to time constraints.

Additionally, ABRicate is limited as it cannot detect fragmented or mutated genes, it has possible coverage calculation errors, sometimes heavily overlapping genes are reported for the same locus and gap reporting between query sequence and detected gene may be incomplete (Seemann, 2018). For future consideration, a larger sample size would help to improve comparisons between both PCR typing and genome analysis methods, and the use of bioinformatics software to connect the edges between nodes may improve ABRicate's ability to detect *tfs vir* genes.

CHAPTER FIVE

GENERAL DISCUSSION

General Discussion

Despite the discovery of the novel putative ICE-encoded *H. pylori* T4SSs, *tfs3* and *tfs4* over the past two decades, their function remains largely unknown. However, some of the variable genes encoded on both *tfs3* and *tfs4* have been implicated in disease outcomes in some populations (Grove et al., 2013, Alandiyjany et al., 2017). Previous studies from the group demonstrated the capacity of the ICE-encoded Vir-homologue proteins to assemble a functional T4SS known to be important for T4SS assembly and function (Alandiyjany, 2015). There has been no robust demonstration of functional complementation between the proteins in the assembly of a hybrid or variant *tfs4* T4SS, and no understanding of possibilities that proteins from different *tfs* systems with a conserved function (such as T4SS assembly) may be able to interact to generate functionally active hybrid systems.

In view of this, the initial aim was to investigate the proteins encoded by *tfs3*, *tfs4* and *tfs4* variant type with an important role in assembly and function of a T4SS. Secondary to this, it was aimed to explore the association of *tfs* T4SS genes in *H. pylori* isolates from patients with their disease status. Such data might provide some insight into correlations between clusters of genes of *tfs* ICEs with disease status.

To achieve this, several candidate *tfs* T4SS assembly proteins of *tfs3*, *tfs4* and *tfs4* variants were selected for heterologous protein-protein interactions study using the yeast-two hybrid assay to screen for cross-

complementation, which may be important for the assembly of a functional T4SS. These were selected based on information in previous work (Fig 5.1) (Delahay et al., 2018b). This was especially important as the *tfs4* T4SS is known to have several variant types.

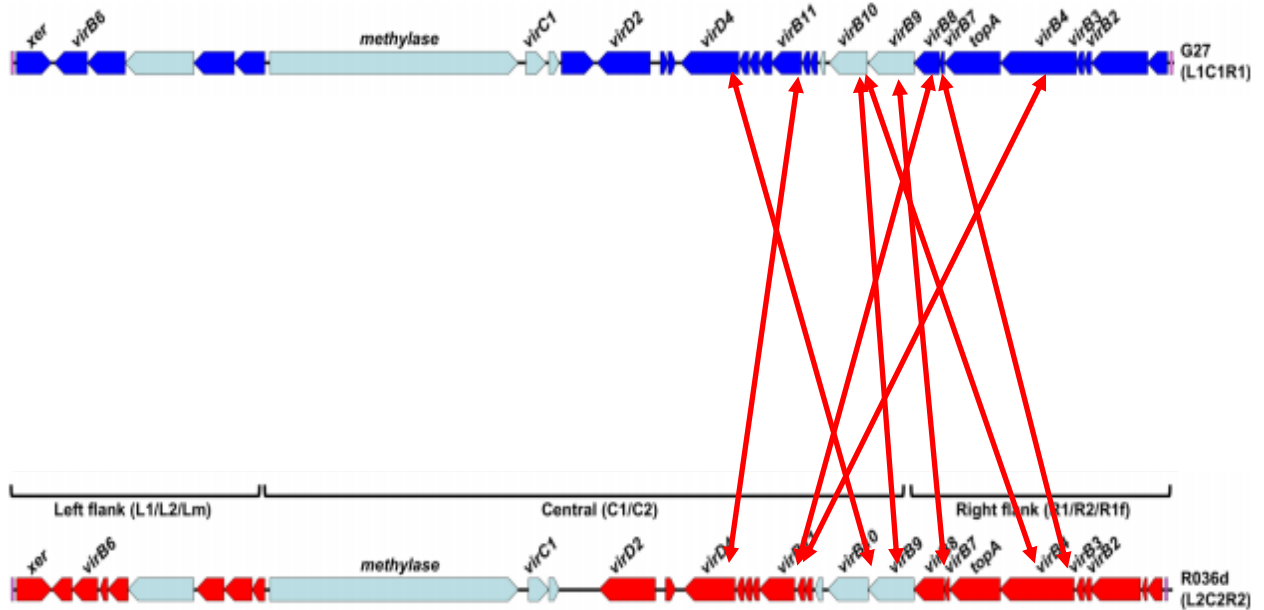


Fig 5.1: Schematic representation of the potential modular protein-protein interactions of the *tfs4* (top strain G27) and *tfs4* variant type (bottom strain R036d) encoded proteins. The full cluster of *tfs4* *vir* genes is represented in each strain (blue for strain G27; red for variant genes in strains R036d). Modular representation of *tfs* cluster of genes is indicated in each strain (L1C1R1 for G27 and L2C2R2 for R036d). Red-crossed arrows indicative of potential interactions between the proteins of *tfs4* (G27) and *tfs4* variant (R036d) which may be important for hybrid or variant types of generation and, potential for functional cross-complementation. Reproduced from (Delahay et al., 2018b).

Table 5.1: Summary of positive Y2H interactions between different structural components of *tfs3*, *tfs4* and *tfs4* variant T4SSs.

Interaction type	Reported interactions
ATPase and Non-ATPase interactions	<i>tfs4VirD4/tfs4VirB8v</i> , <i>tfs4VirB11v/tfs4VirB8</i> , <i>tfs4VirB11v/tfs4VirB8v</i> , <i>tfs3VirB11/tfs4VirB8v</i>
ATPase and ATPase interactions	<i>tfs4VirB11/tfs4VirB11v</i> , <i>tfs4VirB11v/tfs4VirB4</i> , <i>tfs4VirB11v/tfs4VirB4v</i> , <i>tfs4VirB11v/tfs4VirB8v</i> <i>tfs4VirB8v/tfs4VirB4v</i> <i>tfs3VirB11/tfs4VirB4v</i>
Relaxase and ATPase interactions	<i>tfs4VirD2/tfs4VirB11</i> , <i>tfs4VirD2/tfs3VirB4</i> , <i>tfs4VirD2/tfs3VirB11</i>
Interactions of orphan genomic VirB11 with ATPases	<i>gB11/tfs3VirB11</i> , <i>gB11/tfs4VirB11</i> , <i>gB11/tfs4VirB11v</i> , <i>gB11/tfs3VirB4</i> , <i>gB11/tfs4VirB4</i>
Interactions of orphan genomic VirB11 with Non-ATPases	<i>gB11/tfs3VirB8</i>

5.1 Important and Novel Findings

With respect to interactions between proteins of *tfs3*, *tfs4* and *tfs4* variants reported in chapter 2 (summarized in Table 5.1), cross-complementation interactions were stronger between proteins of the same system (*tfs4/tfs4*) when compared to proteins from other systems (*tfs3/tfs4*) respectively. This finding supports existing reports of strong interactions between T4SS proteins from the same system (Fronzes et al., 2009b, Alandiyany, 2015). Additionally, the interactions between ATPase and non-ATPase may suggest that the *tfs4VirB8v* is able to undergo flexible interactions with other components of the T4SS, which

is consistent with flexible interactions of VirB8 with other T4SS proteins (Fronzes et al., 2009b). Moreover, the interactions of *tfs4*VirB8v with the ATPases may be important to assist the ATPases in secretion channel formation and pilus biogenesis. The reported interaction between *tfs4*VirB8v and *tfs4*VirB4v may be important for the stabilisation of the pilus formation during substrate translocation as interactions between the homologous pair VirB8/VirB4 has been reported as key for pilus biogenesis (Ward et al., 2002). Although the interactions reported in this study and limited sample size are not able to provide enough evidence for disease status in patients, this study reported flexible interactions between *tfs* T4SS proteins important for assembly and function, and established potential for cross-complementation between the T4SS proteins. Furthermore, the data in chapter 3 showed the reliability of using both PCR typing and genomics analysis in the determination of virulence genes in *H. pylori*. The comparison of PCR typing data with genomics provided further confidence that PCR typing continues to be an alternative in diagnostic settings due to its sensitivity, low-cost, and adaptability to high volume demands. Therefore, laboratories may not need to perform genome sequencing to evaluate the virulence of isolates. In chapter 4, the use of genomics established the feasibility of using genomics to find intact virulence genes in *H. pylori* clinical isolates. Although the association with disease status in patients was impossible due to small sample size, and limitations of the ABRicate tool to search for fragmented genes, this study established that genomics analysis can

be used to develop an approach to investigate which genes are present in the *tfs* ICEs.

5.2 Critique of Experimental Design

Several reasons have been reported why the purification of membrane proteins is more difficult than water-soluble proteins. The expression of membrane proteins is usually relatively low but large quantities of proteins are needed for many structural studies (Smith, 2017). Although higher protein yield can be obtained by overexpression of the protein of interest, the difference in post-translational modifications such as phosphorylation, glycosylation and acylation can lead to reduced activity of the recombinant protein compared to the native form (Smith, 2017). Since the *tfs* T4SS proteins in this study are membrane proteins, it would have been important to carry out functional assays (for example measuring the enzyme activity or immune- or ligand binding activity) to detect the integrity of the proteins at different stages during the purification process (Smith, 2017). Also, one of the crucial ways to purify recombinant proteins is the addition of extra amino acid sequences to the N- or C-terminus of the protein of interest. This added amino acid sequences is usually called a 'tag'. The recombinant protein used in this study was a His-tagged protein for expression and purification. Earlier studies by Booth and colleagues (Booth et al., 2018) reported differences in the stability and functions of His-tagged recombinant proteins. The use of the His-tagged recombinant protein may, consequently, affect the expression and stability of the *tfs* T4SS proteins. Furthermore, the choice

of data used in this study (the Nottingham Collection data), was important because the dataset was rich and provided relevant information although, the sample size of genome-sequenced isolates was small. In future studies, additional isolates will undergo genome sequencing. Ideally, published genomes of *H. pylori* from different geographic locations would be considered to provide more information about patients from different regions.

5.3 Future work

On the investigation of *tfs* ICE T4SS proteins that may be important for secretion and assembly, future work would involve bioinformatics analysis tools such as STRING to build a network of possible interactions between the proteins before wet laboratory experiments. Also, other methods such as co-purification assay (where the interacted proteins are cloned into each end of an expression vector such as pacycduet) and Co-Immunoprecipitation-Blotting assay (used to investigate complicated protein interactions or networks involving membrane and nuclear proteins) would be used as a secondary approach to analyse reported protein-protein interactions. Also, investigation of the cell growth (using a growth curve assay) would be considered as a first measure to solve the problem of protein toxicity. Protein toxicity may arise when the recombinant proteins perform an unnecessary and detrimental effect on the host cell (Rosano and Ceccarelli, 2014), which may interfere with the host organism's normal proliferation or homeostasis process thereby lowering the growth rate. The growth rate of the recombinant strain can

be compared with an empty vector and if the rate is slower than the empty vector, then it may be due to gene toxicity (Rosano and Ceccarelli, 2014).

To investigate the association of *tfs* T4SS genes with disease status in patients presenting with *H. pylori* infection, two areas are important. Firstly, understanding the role other factors such as environmental factors, lifestyle changes and nutritional factors play in individual patient disease status is important. Disease status in patients can be influenced by these factors, therefore, their contribution needs to be investigated in future studies by gathering enough data from individual patients. Another important area is to use a robust genome-wide study (GWAS) to detect variant genes especially in *tfs4* cluster of genes where hybrid or variant types are prominent. This future study will provide a robust method to detect fragmented genes and variant copies from *tfs* ICEs from clinical isolates and may give insight into their association with disease status in isolates. Therefore, I strongly recommend that future studies should include much larger sets of *H. pylori* genomic data with details including disease and histopathological severity of gastritis and premalignant changes.

REFERENCES

- ABADI, A. T. B. & PEREZ-PEREZ, G. 2016. Role of *dupA* in virulence of *Helicobacter pylori*. *World Journal of Gastroenterology*, 22, 10118.
- ALAM, J., SARKAR, A., KARMAKAR, B. C., GANGULY, M., PAUL, S. & MUKHOPADHYAY, A. K. 2020. Novel virulence factor *dupA* of *Helicobacter pylori* as an important risk determinant for disease manifestation: An overview. *World Journal of Gastroenterology*, 26, 4739.
- ALANDIYJANY, M. N. 2015. *Assembly and Functional Characterisation of the Tfs-encoded Type IV Secretion Systems of Helicobacter pylori*. University of Nottingham.
- ALANDIYJANY, M. N., CROXALL, N. J., GROVE, J. I. & DELAHAY, R. M. 2017. A role for the *tfs3* ICE-encoded type IV secretion system in pro-inflammatory signalling by the *Helicobacter pylori* Ser/Thr kinase, CtkA. *PLoS One*, 12, e0182144.
- ALEBIE, G. & KABA, D. 2016. Prevalence of *Helicobacter pylori* infection and associated factors among gastritis students in Jigjiga University, Jigjiga, Somali regional state of Ethiopia. *J Bacteriol Mycol*, 3, 00060.
- ALM, R. A., LING, L.-S. L., MOIR, D. T., KING, B. L., BROWN, E. D., DOIG, P. C., SMITH, D. R., NOONAN, B., GUILD, B. C. &

- DEJONGE, B. L. 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature*, 397, 176-180.
- ALTANBAYAR, O., AMGALANBAATAR, A., BATTOGTOKH, C., BAYARJARGAL, N., BELICK, D., VASCONCELOS, M. K., MACKENZIE, C. R., PFEFFER, K. & HENRICH, B. 2022. Characterization of the *cagA*-gene in *Helicobacter pylori* in Mongolia and detection of two EPIYA-A enriched CagA types. *International Journal of Medical Microbiology*, 312, 151552.
- ALVAREZ-MARTINEZ, C. E. & CHRISTIE, P. J. 2009. Biological diversity of prokaryotic type IV secretion systems. *Microbiology and Molecular Biology Reviews*, 73, 775-808.
- AMIEVA, M. & PEEK JR, R. M. 2016. Pathobiology of *Helicobacter pylori*-induced gastric cancer. *Gastroenterology*, 150, 64-78.
- ANDO, T., TSUZUKI, T., MIZUNO, T., MINAMI, M., INA, K., KUSUGAMI, K., TAKAMATSU, J., ADACHI, K., EL-OMAR, E. & OHTA, M. 2004. Characteristics of *Helicobacter pylori*-induced gastritis and the effect of *H. pylori* eradication in patients with chronic idiopathic thrombocytopenic purpura. *Helicobacter*, 9, 443-452.
- ANKOUANE, F., NOAH, D. N., ENYIME, F. N., NDJOLLÉ, C. M., DJAPA, R. N., NONGA, B. N., NJOYA, O. & NDAM, E. C. N. 2015. *Helicobacter pylori* and precancerous conditions of the stomach: the frequency of infection in a cross-sectional study of

- 79 consecutive patients with chronic antral gastritis in Yaoundé, Cameroon. *Pan African Medical Journal*, 20.
- ANYANWU, R. C. & JUKES, D. J. 1991. Food systems and food control in Nigeria. *Food policy*, 16, 112-126.
- ARGENT, R. H., KIDD, M., OWEN, R. J., THOMAS, R. J., LIMB, M. C. & ATHERTON, J. C. 2004. Determinants and consequences of different levels of CagA phosphorylation for clinical isolates of *Helicobacter pylori*. *Gastroenterology*, 127, 514-523.
- ARGENT, R. H., THOMAS, R. J., AVILES-JIMENEZ, F., LETLEY, D. P., LIMB, M. C., EL-OMAR, E. M. & ATHERTON, J. C. 2008. Toxigenic *Helicobacter pylori* infection precedes gastric hypochlorhydria in cancer relatives, and H. pylori virulence evolves in these families. *Clinical cancer research*, 14, 2227-2235.
- ATHERTON, J. C. 2006. The pathogenesis of *Helicobacter pylori*-induced gastro-duodenal diseases. *Annu. Rev. Pathol. Mech. Dis.*, 1, 63-96.
- ATHERTON, J. C., CAO, P., PEEK, R. M., TUMMURU, M. K., BLASER, M. J. & COVER, T. L. 1995. Mosaicism in Vacuolating Cytotoxin Alleles of *Helicobacter pylori*: ASSOCIATION OF SPECIFIC vacA TYPES WITH CYTOTOXIN PRODUCTION AND PEPTIC ULCERATION (*). *Journal of Biological Chemistry*, 270, 17771-17777.
- AUDIBERT, C., JANVIER, B., GRIGNON, B., SALAÜNA, L., BURUCOA, C., LECRON, J.-C. & FAUCHÈRE, J.-L. 2000.

- Correlation between IL-8 induction, *cagA* status and *vacA* genotypes in 153 French *Helicobacter pylori* isolates. *Research in microbiology*, 151, 191-200.
- AZUMA, T., YAMAKAWA, A., YAMAZAKI, S., FUKUTA, K., OHTANI, M., ITO, Y., DOJO, M., YAMAZAKI, Y. & KURIYAMA, M. 2002. Correlation between variation of the 3' region of the *cagA* gene in *Helicobacter pylori* and disease outcome in Japan. *The Journal of infectious diseases*, 186, 1621-1630.
- BACKERT, S., HAAS, R., GERHARD, M. & NAUMANN, M. 2017. The *Helicobacter pylori* Type IV Secretion System Encoded by the *cag* Pathogenicity Island: Architecture, Function, and Signaling. *Curr Top Microbiol Immunol*, 413, 187-220.
- BAJ, J., FORMA, A., SITARZ, M., PORTINCASA, P., GARRUTI, G., KRASOWSKA, D. & MACIEJEWSKI, R. 2021. *Helicobacter pylori* virulence factors—mechanisms of bacterial pathogenicity in the gastric microenvironment. *Cells*, 10, 27.
- BAKHTI, S. Z., LATIFI-NAVID, S., TOBNAGH, S. G., YAZDANBOD, K. & YAZDANBOD, A. 2020. Which genotype of *Helicobacter pylori*—*cagA* or *cagE*—Is better associated with gastric Cancer risk? Lessons from an extremely high-risk area in Iran. *Infection, Genetics and Evolution*, 85, 104431.
- BAKOUR, S., SANKAR, S. A., RATHORED, J., BIAGINI, P., RAOULT, D. & FOURNIER, P.-E. 2016. Identification of virulence factors and antibiotic resistance markers using bacterial genomics. *Future microbiology*, 11, 455-466.

- BALTRUS, D. A., AMIEVA, M. R., COVACCI, A., LOWE, T. M.,
MERRELL, D. S., OTTEMANN, K. M., STEIN, M., SALAMA, N.
R. & GUILLEMIN, K. 2009. The complete genome sequence of
Helicobacter pylori strain G27. *Journal of bacteriology*, 191, 447-
448.
- BELLANGER, X., PAYOT, S., LEBLOND-BOURGET, N. & GUÉDON,
G. 2014. Conjugative and mobilizable genomic islands in
bacteria: evolution and diversity. *FEMS Microbiology Reviews*,
38, 720-760.
- BLASER, M. J. 1999. Hypothesis: the changing relationships of
Helicobacter pylori and humans: implications for health and
disease. *The journal of infectious diseases*, 179, 1523-1530.
- BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a
flexible trimmer for Illumina sequence data. *Bioinformatics*, 30,
2114-2120.
- BOOTH, W. T., SCHLACHTER, C. R., POTE, S., USSIN, N., MANK, N.
J., KLAPPER, V., OFFERMANN, L. R., TANG, C., HURLBURT,
B. K. & CHRUSZCZ, M. 2018. Impact of an N-terminal
polyhistidine tag on protein thermal stability. *ACS omega*, 3, 760-
768.
- BOTELHO, J. & SCHULENBURG, H. 2021. The role of integrative and
conjugative elements in antibiotic resistance evolution. *Trends in
Microbiology*, 29, 8-18.
- BOUKHRIS, S. A., BENAIAH, D.-A., EL RHAZI, K., IBRAHIMI, S.,
NEJJARI, C., AMARTI, A., MAHMOUD, M., EL ABKARI, M.,

- SOULEIMANI, A. & BENNANI, B. 2012. Prevalence and distribution of *Helicobacter pylori* *cagA* and *vacA* genotypes in the Moroccan population with gastric disease. *European journal of clinical microbiology & infectious diseases*, 31, 1775-1781.
- BOURG, G., SUBE, R., O'CALLAGHAN, D. & PATEY, G. 2009. Interactions between *Brucella suis* VirB8 and its homolog TraJ from the plasmid pSB102 underline the dynamic nature of type IV secretion systems. *Journal of bacteriology*, 191, 2985-2992.
- BRICKER, B. J. 2002. PCR as a diagnostic tool for brucellosis. *Veterinary microbiology*, 90, 435-446.
- BRUCE, M. G. & MAAROOS, H. I. 2008. Epidemiology of *Helicobacter pylori* infection. *Helicobacter*, 13, 1-6.
- BUHRDORF, R., FÖRSTER, C., HAAS, R. & FISCHER, W. 2003. Topological analysis of a putative virB8 homologue essential for the *cag* type IV secretion system in *Helicobacter pylori*. *International journal of medical microbiology*, 293, 213-217.
- BURRUS, V. 2017. Mechanisms of stabilization of integrative and conjugative elements. *Current opinion in microbiology*, 38, 44-50.
- BURRUS, V. & WALDOR, M. K. 2004. Shaping bacterial genomes with integrative and conjugative elements. *Research in microbiology*, 155, 376-386.
- BUSLER, V. J., TORRES, V. J., MCCLAIN, M. S., TIRADO, O., FRIEDMAN, D. B. & COVER, T. L. 2006. Protein-protein

- interactions among *Helicobacter pylori* cag proteins. *Journal of bacteriology*, 188, 4787-4800.
- BUTI, L., SPOONER, E., VAN DER VEEN, A. G., RAPPUOLI, R., COVACCI, A. & PLOEGH, H. L. 2011. *Helicobacter pylori* cytotoxin-associated gene A (CagA) subverts the apoptosis-stimulating protein of p53 (ASPP2) tumor suppressor pathway of the host. *Proceedings of the National Academy of Sciences*, 108, 9238-9243.
- BYTZER, P. & O'MORAIN, C. 2005. Treatment of *Helicobacter pylori*. *Helicobacter*, 10, 40-46.
- CHAKRABORTY, A., MITRA, S., DE, D., PAL, A. J., GHAEMI, F., AHMADIAN, A. & FERRARA, M. 2021. Determining Protein–Protein Interaction Using Support Vector Machine: A Review. *IEEE Access*, 9, 12473-12490.
- CHEN, L., YANG, J., YU, J., YAO, Z., SUN, L., SHEN, Y. & JIN, Q. 2005. VFDB: a reference database for bacterial virulence factors. *Nucleic acids research*, 33, D325-D328.
- CHRISTIE, P. J. 2004. Type IV secretion: the *Agrobacterium* VirB/D4 and related conjugation systems. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1694, 219-234.
- CHRISTIE, P. J. 2016. The mosaic type IV secretion systems. *EcoSal Plus*, 7.
- CHRISTIE, P. J., ATMAKURI, K., KRISHNAMOORTHY, V., JAKUBOWSKI, S. & CASCALES, E. 2005. Biogenesis,

- architecture, and function of bacterial type IV secretion systems. *Annu. Rev. Microbiol.*, 59, 451-485.
- CHUNG, C., OLIVARES, A., TORRES, E., YILMAZ, O., COHEN, H. & PEREZ-PEREZ, G. 2010. Diversity of VacA intermediate region among *Helicobacter pylori* strains from several regions of the world. *Journal of clinical microbiology*, 48, 690-696.
- CHUNG, J., CHEN, G., WU, L., CHANG, H., LIN, J. G., YEH, C. & WANG, T. 1998. Effects of garlic compounds diallyl sulfide and diallyl disulfide on arylamine N-acetyltransferase activity in strains of *Helicobacter pylori* from peptic ulcer patients. *The American journal of Chinese medicine*, 26, 353-364.
- CHUNG, K. T. & SHELAT, V. G. 2017. Perforated peptic ulcer-an update. *World journal of gastrointestinal surgery*, 9, 1.
- CLUSTAL OMEGA. 2022. *Multiple Sequence Alignment*. [Online]. Available: <https://www.ebi.ac.uk/Tools/msa/clustalo/> [Accessed 23 June 2022].
- CORDOVA-MARKS, F. M., CARSON, W. O., MONETATHCHI, A., LITTLE, A. & ERDRICH, J. 2022. Native and Indigenous Populations and Gastric Cancer: A Worldwide Review. *International Journal of Environmental Research and Public Health*, 19, 5437.
- COSTA, T. R., HARB, L., KHARA, P., ZENG, L., HU, B. & CHRISTIE, P. J. 2021. Type IV secretion systems: Advances in structure, function, and activation. *Molecular microbiology*, 115, 436.

- COVACCI, A., CENSINI, S., BUGNOLI, M., PETRACCA, R.,
BURRONI, D., MACCHIA, G., MASSONE, A., PAPINI, E.,
XIANG, Z. & FIGURA, N. 1993. Molecular characterization of the
128-kDa immunodominant antigen of *Helicobacter pylori*
associated with cytotoxicity and duodenal ulcer. *Proceedings of
the National Academy of Sciences*, 90, 5791-5795.
- COVER, T. L. & BLASER, M. J. 2009. *Helicobacter pylori* in health and
disease. *Gastroenterology*, 136, 1863-1873.
- COVER, T. L., GLUPCZYNSKI, Y., LAGE, A. P., BURETTE, A.,
TUMMURU, M., PEREZ-PEREZ, G. I. & BLASER, M. J. 1995.
Serologic detection of infection with *cagA*+ *Helicobacter pylori*
strains. *Journal of clinical microbiology*, 33, 1496-1500.
- CRABTREE, J., COVACCI, A., FARMERY, S., XIANG, Z., TOMPKINS,
D., PERRY, S., LINDLEY, I. & RAPPUOLI, R. 1995.
Helicobacter pylori induced interleukin-8 expression in gastric
epithelial cells is associated with CagA positive phenotype.
Journal of clinical pathology, 48, 41-45.
- DARKO, R., YAWSON, A., OSEI, V., OWUSU-ANSAH, J. & ALUZE-
ELE, S. 2015. Changing patterns of the prevalence of
Helicobacter pylori among patients at a corporate hospital in
Ghana. *Ghana Medical Journal*, 49, 147-153.
- DE LIMA SILVA, L. L., OLIVEIRA, A. K. S., GAMA, A. R., RAMOS, A.
F. P. L., SILVA, A. M. T. C., BLANCO, A. J. V., VIEIRA, J. D. G.,
RASMUSSEM, L. T., CARNEIRO, L. C. & BARBOSA, M. S.
2021. *Helicobacter pylori* virulence *dupA* gene: risk factor or

- protective factor? *Brazilian Journal of Microbiology*, 52, 1921-1927.
- DELAHAY, R. M., CROXALL, N. J. & STEPHENS, A. D. 2018a. Phylogeographic diversity and mosaicism of the. *Mob DNA*, 9, 5.
- DELAHAY, R. M., CROXALL, N. J. & STEPHENS, A. D. 2018b. Phylogeographic diversity and mosaicism of the *Helicobacter pylori* *tfs* integrative and conjugative elements. *Mobile DNA*, 9, 1-20.
- DEMEKE, T. & JENKINS, G. R. 2010. Influence of DNA extraction methods, PCR inhibitors and quantification methods on real-time PCR assay of biotechnology-derived traits. *Analytical and bioanalytical chemistry*, 396, 1977-1990.
- DIXON, M. F., GENTA, R. M., YARDLEY, J. H. & CORREA, P. 1996. Classification and grading of gastritis: the updated Sydney system. *The American journal of surgical pathology*, 20, 1161-1181.
- DONKOR, E. S., DAYIE, N. T. & ADIKU, T. K. 2014. Bioinformatics with basic local alignment search tool (BLAST) and fast alignment (FASTA). *Journal of Bioinformatics and Sequence Analysis*, 6, 1-6.
- DORE, M. P. & PES, G. M. 2021. What is new in *Helicobacter pylori* diagnosis. An overview. *Journal of Clinical Medicine*, 10, 2091.
- DOTMATICS. 2022. *Geneious Prime: Comprehensive suite of molecular biology and sequence analysis tools*. [Online].

Available: <https://www.geneious.com/prime/> [Accessed 24 June 2022].

- EBULE, I., DJUNE FOKOU, A., SITEDJEYA MOKO, I., TANNI, B. & HEUGEU, C. 2017. Prevalence of *H. pylori* infection and atrophic gastritis among dyspeptic subjects in Cameroon using a panel of serum biomarkers (PGI, PGII, G17, Hplg). *Sch J App Med Sci*, 5, 1230-9.
- EL KHADIR, M., BOUKHRIS, S. A., ZAHIR, S. O., BENAIAH, D.-A., IBRAHIMI, S. A., CHBANI, L., EL ABKARI, M. & BENNANI, B. 2021. CagE, *cagA* and *cagA* 3' region polymorphism of *Helicobacter pylori* and their association with the intra-gastric diseases in Moroccan population. *Diagnostic Microbiology and Infectious Disease*, 100, 115372.
- ENTWISTLE, S., LI, X. & YIN, Y. 2019. Orphan genes shared by pathogenic genomes are more associated with bacterial pathogenicity. *Msystems*, 4, e00290-18.
- ERFFELINCK, M.-L., RIBEIRO, B., PERASSOLO, M., PAUWELS, L., POLLIER, J., STORME, V. & GOOSSENS, A. 2018. A user-friendly platform for yeast two-hybrid library screening using next generation sequencing. *PLoS One*, 13, e0201270.
- ERNST, P. B. & GOLD, B. D. 2000. The disease spectrum of *Helicobacter pylori*: the immunopathogenesis of gastroduodenal ulcer and gastric cancer. *Annual Reviews in Microbiology*, 54, 615-640.

- ERZIN, Y., KOKSAL, V., ALTUN, S., DOBRUCALI, A., ASLAN, M.,
ERDAMAR, S., DIRICAN, A. & KOCAZEYBEK, B. 2006.
Prevalence of *Helicobacter pylori* *vacA*, *cagA*, *cagE*, *iceA*, *babA2*
genotypes and correlation with clinical outcome in Turkish
patients with dyspepsia. *Helicobacter*, 11, 574-580.
- EVERHART, J. E. 2000. Recent developments in the epidemiology of
Helicobacter pylori. *Gastroenterology Clinics of North America*,
29, 559-578.
- EXPASY. 1 July 2020. *Translate*. [Online]. Available:
<https://web.expasy.org/translate/> [Accessed 23 June 2022].
- FARNBACHER, M., JAHNS, T., WILLRODT, D., DANIEL, R., HAAS,
R., GOESMANN, A., KURTZ, S. & RIEDER, G. 2010.
Sequencing, annotation, and comparative genome analysis of
the gerbil-adapted *Helicobacter pylori* strain B8. *BMC genomics*,
11, 1-22.
- FERREIRA, R. M., PINTO-RIBEIRO, I., WEN, X., MARCOS-PINTO, R.,
DINIS-RIBEIRO, M., CARNEIRO, F. & FIGUEIREDO, C. 2016.
Helicobacter pylori *cagA* promoter region sequences influence
CagA expression and interleukin 8 secretion. *The Journal of*
infectious diseases, 213, 669-673.
- FISCHER, W., TEGTMEYER, N., STINGL, K. & BACKERT, S. 2020.
Four chromosomal type IV secretion systems in *Helicobacter*
pylori: composition, structure and function. *Frontiers in*
Microbiology, 11, 1592.

- FISCHER, W., WINDHAGER, L., ROHRER, S., ZEILLER, M., KARNHOLZ, A., HOFFMANN, R., ZIMMER, R. & HAAS, R. 2010. Strain-specific genes of *Helicobacter pylori*: genome evolution driven by a novel type IV secretion system and genomic island transfer. *Nucleic acids research*, 38, 6089-6101.
- FOEGEDING, N. J., CASTON, R. R., MCCLAIN, M. S., OHI, M. D. & COVER, T. L. 2016. An overview of *Helicobacter pylori* VacA toxin biology. *Toxins*, 8, 173.
- FRANCESCHI, F., ANNALISA, T., DI RIENZO TERESA, D., IANIRO, G., FRANCO, S., VIVIANA, G., VALENTINA, T., RICCARDO, L. L. & ANTONIO, G. 2014. Role of *Helicobacter pylori* infection on nutrition and metabolism. *World Journal of Gastroenterology: WJG*, 20, 12809.
- FRAWLEY, E. R. & FANG, F. C. 2014. The ins and outs of bacterial iron metabolism. *Molecular microbiology*, 93, 609-616.
- FROMMER, W. B. & NINNEMANN, O. 1995. Heterologous expression of genes in bacterial, fungal, animal, and plant cells. *Annual review of plant biology*, 46, 419-444.
- FRONZES, R., CHRISTIE, P. J. & WAKSMAN, G. 2009a. The structural biology of type IV secretion systems. *Nature Reviews Microbiology*, 7, 703-714.
- FRONZES, R., SCHAFER, E., WANG, L., SAIBIL, H. R., ORLOVA, E. V. & WAKSMAN, G. 2009b. Structure of a type IV secretion system core complex. *science*, 323, 266-268.

- GADDY, J. A., RADIN, J. N., LOH, J. T., ZHANG, F., WASHINGTON, M. K., PEEK JR, R. M., ALGOOD, H. M. S. & COVER, T. L. 2013. High dietary salt intake exacerbates *Helicobacter pylori*-induced gastric carcinogenesis. *Infection and immunity*, 81, 2258-2267.
- GARG, A. & GUPTA, D. 2008. VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC bioinformatics*, 9, 1-12.
- GIETZ, R. D., SCHIESTL, R. H., WILLEMS, A. R. & WOODS, R. A. 1995. Studies on the transformation of intact yeast cells by the LiAc/SS-DNA/PEG procedure. *Yeast*, 11, 355-360.
- GOLDSTEIN, S., BEKA, L., GRAF, J. & KLASSEN, J. L. 2019. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC genomics*, 20, 1-17.
- GOMIS-RUÈTH, F. X., MONCALIAN, G., PÉREZ-LUQUE, R., GONZÁLEZ, A., CABEZÓN, E., DE LA CRUZ, F. & COLL, M. 2001. The bacterial conjugation protein TrwB resembles ring helicases and F1-ATPase. *Nature*, 409, 637-641.
- GOMIS-RUTH, F., SOLA, M. & COLL, M. 2004. Coupling factors in macromolecular type-IV secretion machineries. *Current pharmaceutical design*, 10, 1551-1565.
- GRABOWSKA, M., WAWRZYNIAK, D., ROLLE, K., CHOMCZYŃSKI, P., OZIEWICZ, S., JURGA, S. & BARCISZEWSKI, J. 2019. Let

- food be your medicine: nutraceutical properties of lycopene. *Food & function*, 10, 3090-3102.
- GRAHAM, D., OPEKUN, A., OSATO, M., EL-ZIMAITY, H., LEE, C., YAMAOKA, Y., QURESHI, W., CADOZ, M. & MONATH, T. 2004. Challenge model for *Helicobacter pylori* infection in human volunteers. *Gut*, 53, 1235-1243.
- GRAHAM, D. Y. & FISCHBACH, L. 2010. *Helicobacter pylori* treatment in the era of increasing antibiotic resistance. *Gut*, 59, 1143-1153.
- GRANSTROM, M., LEHOURS, P. & BENGTSSON, C. 2008. M? graud, F. Diagnosis of *Helicobacter pylori*. *Helicobacter*, 13, 7-12.
- GREEN, E. R. & MECSAS, J. 2016. Bacterial secretion systems: an overview. *Microbiology spectrum*, 4, 4.1. 13.
- GROHMANN, E., CHRISTIE, P. J., WAKSMAN, G. & BACKERT, S. 2018. Type IV secretion in Gram-negative and Gram-positive bacteria. *Molecular microbiology*, 107, 455-471.
- GROHMANN, E., MUTH, G. N. & ESPINOSA, M. 2003. Conjugative plasmid transfer in gram-positive bacteria. *Microbiology and molecular biology reviews*, 67, 277-301.
- GROVE, J. I., ALANDIYJANY, M. N. & DELAHAY, R. M. 2013. Site-specific relaxase activity of a VirD2-like protein encoded within the *tfs4* genomic island of *Helicobacter pylori*. *Journal of Biological Chemistry*, 288, 26385-26396.
- GUPTA, A., KAPIL, R., DHAKAN, D. B. & SHARMA, V. K. 2014. MP3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. *PloS one*, 9, e93907.

- HALL, T. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.*, 1999. 95-98.
- HARE, S., BAYLISS, R., BARON, C. & WAKSMAN, G. 2006. A large domain swap in the VirB11 ATPase of *Brucella suis* leaves the hexameric assembly intact. *Journal of molecular biology*, 360, 56-66.
- HARRISON, U., FOWORA, M. A., SERIKI, A. T., LOELL, E., MUELLER, S., UGO-IJEH, M., ONYEKWERE, C. A., LESI, O. A., OTEGBAYO, J. A. & AKERE, A. 2017. *Helicobacter pylori* strains from a Nigerian cohort show divergent antibiotic resistance rates and a uniform pathogenicity profile. *PloS one*, 12, e0176454.
- HATAKEYAMA, M. 2014. *Helicobacter pylori* CagA and gastric cancer: a paradigm for hit-and-run carcinogenesis. *Cell host & microbe*, 15, 306-316.
- HEIKEMA, A. P., HORST-KREFT, D., BOERS, S. A., JANSEN, R., HILTEMANN, S. D., DE KONING, W., KRAAIJ, R., DE RIDDER, M. A., VAN HOUTEN, C. B. & BONT, L. J. 2020. Comparison of Illumina versus nanopore 16S rRNA gene sequencing of the human nasal microbiota. *Genes*, 11, 1105.
- HIGASHI, H., TSUTSUMI, R., FUJITA, A., YAMAZAKI, S., ASAKA, M., AZUMA, T. & HATAKEYAMA, M. 2002. Biological activity of the *Helicobacter pylori* virulence factor CagA is determined by

- variation in the tyrosine phosphorylation sites. *Proceedings of the National Academy of Sciences*, 99, 14428-14433.
- HOOI, J. K., LAI, W. Y., NG, W. K., SUEN, M. M., UNDERWOOD, F. E., TANYINGOH, D., MALFERTHEINER, P., GRAHAM, D. Y., WONG, V. W. & WU, J. C. 2017. Global prevalence of *Helicobacter pylori* infection: systematic review and meta-analysis. *Gastroenterology*, 153, 420-429.
- HOSSEINI, E., POURSIANA, F., VAN DE WIELE, T., SAFAEI, H. G. & ADIBI, P. 2012. *Helicobacter pylori* in Iran: A systematic review on the association of genotypes and gastroduodenal diseases. *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, 17, 280.
- HUNT, R. 1996. The role of *Helicobacter pylori* in pathogenesis: the spectrum of clinical outcomes. *Scandinavian Journal of Gastroenterology*, 31, 3-9.
- HUSSEIN, N. R., ARGENT, R. H., MARX, C. K., PATEL, S. R., ROBINSON, K. & ATHERTON, J. C. 2010. *Helicobacter pylori dupA* is polymorphic, and its active form induces proinflammatory cytokine secretion by mononuclear cells. *The Journal of infectious diseases*, 202, 261-269.
- IMAGAWA, S., ITO, M., YOSHIHARA, M., EGUCHI, H., TANAKA, S. & CHAYAMA, K. 2010. *Helicobacter pylori dupA* and gastric acid secretion are negatively associated with gastric cancer development. *Journal of medical microbiology*, 59, 1484-1489.

- IQBAL, N., MUKHTAR, M. U., YANG, J., NIU, Q., LI, Z., ZHAO, S., ZHAO, Y., GUAN, G., LIU, Z. & YIN, H. 2021. Identification and evaluation of midgut protein RL12 of *Dermacentor silvarum* interacting with *Anaplasma ovis* VirD4. *Ticks and Tick-borne Diseases*, 12, 101677.
- JAKSZYN, P., BINGHAM, S., PERA, G., AGUDO, A., LUBEN, R., WELCH, A., BOEING, H., DEL GIUDICE, G., PALLI, D. & SAIEVA, C. 2006. Endogenous versus exogenous exposure to N-nitroso compounds and gastric cancer risk in the European Prospective Investigation into Cancer and Nutrition (EPIC-EURGAST) study. *Carcinogenesis*, 27, 1497-1501.
- JALAN, N., ARITUA, V., KUMAR, D., YU, F., JONES, J. B., GRAHAM, J. H., SETUBAL, J. C. & WANG, N. 2011. Comparative Genomic Analysis of *Xanthomonas Axonopodis* Pv. Citrumelo F1, Which Causes Citrus Bacterial Spot Disease, and Related Strains Provides Insights Into Virulence and Host Specificity. *Journal of bacteriology*, 193, 6342-6357.
- JAMES, P., HALLADAY, J. & CRAIG, E. A. 1996. Genomic libraries and a host strain designed for highly efficient two-hybrid selection in yeast. *Genetics*, 144, 1425-1436.
- JANG, S. H., LIM, J. W., MORIO, T. & KIM, H. 2012. Lycopene inhibits *Helicobacter pylori*-induced ATM/ATR-dependent DNA damage response in gastric epithelial AGS cells. *Free Radical Biology and Medicine*, 52, 607-615.

- JENSEN, P. J., FELDMAN, M., LAMONT, J. & GROVER, S. 2019. Acute and chronic gastritis due to *Helicobacter pylori*. *Up-to-date*.
- JOENSEN, K. G., SCHEUTZ, F., LUND, O., HASMAN, H., KAAS, R. S., NIELSEN, E. M. & AARESTRUP, F. M. 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *Journal of clinical microbiology*, 52, 1501-1510.
- JOHNSON, C. M. & GROSSMAN, A. D. 2015. Integrative and conjugative elements (ICEs): what they do and how they work. *Annual review of genetics*, 49, 577-601.
- JOSHI, N. & FASS, J. 2011. Sickie: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33)[Software].
- JUDAKI, A., RAHMANI, A., FEIZI, J., ASADOLLAHI, K. & HAFEZI AHMADI, M. R. 2017. Curcumin in combination with triple therapy regimes ameliorates oxidative stress and histopathologic changes in chronic gastritis-associated *Helicobacter pylori* infection. *Arquivos de gastroenterologia*, 54, 177-182.
- JUNG, S. W., SUGIMOTO, M., SHIOTA, S., GRAHAM, D. Y. & YAMAOKA, Y. 2012. The intact *dupA* cluster is a more reliable *Helicobacter pylori* virulence marker than *dupA* alone. *Infection and immunity*, 80, 381-387.
- KABAMBA, E. T., TUAN, V. P. & YAMAOKA, Y. 2018. Genetic populations and virulence factors of *Helicobacter pylori*. *Infection, Genetics and Evolution*, 60, 109-116.

- KANDULSKI, A., SELGRAD, M. & MALFERTHEINER, P. 2008.
Helicobacter pylori infection: a clinical overview. *Digestive and Liver Disease*, 40, 619-626.
- KARNHOLZ, A., HOEFLER, C., ODENBREIT, S., FISCHER, W.,
HOFREUTER, D. & HAAS, R. 2006. Functional and topological
characterization of novel components of the comB DNA
transformation competence system in *Helicobacter pylori*.
Journal of bacteriology, 188, 882-893.
- KERSULYTE, D., KALIA, A., GILMAN, R. & BERG, D. Genome
Sequence of *Helicobacter pylori* from the Remote Amazon:
traces of Asian ancestry of the First Americans. 2010.
Unpublished.[Google Scholar].
- KERSULYTE, D., VELAPATINO, B., MUKHOPADHYAY, A. K.,
CAHUAYME, L., BUSSALLEU, A., COMBE, J., GILMAN, R. H. &
BERG, D. E. 2003. Cluster of type IV secretion genes in
Helicobacter pylori's plasticity zone. *Journal of bacteriology*, 185,
3764-3772.
- KIDD, M., LASTOVICA, A., ATHERTON, J. & LOUW, J. 2001.
Conservation of the *cag* pathogenicity island is associated with
vacA alleles and gastroduodenal disease in South African
Helicobacter pylori isolates. *Gut*, 49, 11-17.
- KNORR, J., RICCI, V., HATAKEYAMA, M. & BACKERT, S. 2019.
Classification of *Helicobacter pylori* virulence factors: Is CagA a
toxin or not? *Trends in microbiology*, 27, 731-738.

- KRISCH, L. M., POSSELT, G., HAMMERL, P. & WESSLER, S. 2016. CagA phosphorylation in *Helicobacter pylori*-infected B cells is mediated by the nonreceptor tyrosine kinases of the Src and Abl families. *Infection and immunity*, 84, 2671-2680.
- KUIPERS, E., THIJS, J. & FESTEN, H. 1995. The prevalence of *Helicobacter pylori* in peptic ulcer disease. *Alimentary pharmacology & therapeutics*, 9, 59-69.
- KURTZ, S. 2003. The Vmatch large scale sequence analysis software. *Ref Type: Computer Program*, 412, 297.
- KUSTERS, J. G., VAN VLIET, A. H. & KUIPERS, E. J. 2006. Pathogenesis of *Helicobacter pylori* infection. *Clinical microbiology reviews*, 19, 449-490.
- KUTTER, S., BUHRDORF, R., HAAS, J., SCHNEIDER-BRACHERT, W., HAAS, R. & FISCHER, W. 2008. Protein subassemblies of the *Helicobacter pylori* Cag type IV secretion system revealed by localization and interaction studies. *Journal of bacteriology*, 190, 2161-2171.
- LAEMMLI, U. K. 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *nature*, 227, 680-685.
- LAWLEY, T., KLIMKE, W., GUBBINS, M. & FROST, L. 2003. F factor conjugation is a true type IV secretion system. *FEMS microbiology letters*, 224, 1-15.
- LEE, I. P. A. & ANDAM, C. P. 2022. Frequencies and characteristics of genome-wide recombination in *Streptococcus agalactiae*,

Streptococcus pyogenes, and *Streptococcus suis*. *Scientific Reports*, 12, 1-11.

LEE, Y.-C., DORE, M. P. & GRAHAM, D. Y. 2022. Diagnosis and treatment of *Helicobacter pylori* infection. *Annual Review of Medicine*, 73, 183-195.

LEONARDO, H., ROCCO, M. & ZAGARI, F. B. 2014. Epidemiology of *Helicobacter pylori* infection. *J Helicobacter*, 1, 1-5.

LI, X., XIE, Y., LIU, M., TAI, C., SUN, J., DENG, Z. & OU, H.-Y. 2018. oriTfinder: a web-based tool for the identification of origin of transfers in DNA sequences of bacterial mobile genetic elements. *Nucleic acids research*, 46, W229-W234.

LIU, H. Y., HOPPING, G. C., VAIDYANATHAN, U., RONQUILLO, Y. C., HOOPES, P. C. & MOSHIRFAR, M. 2019a. Polymerase chain reaction and its application in the diagnosis of infectious keratitis. *Medical Hypothesis, Discovery and Innovation in Ophthalmology*, 8, 152.

LIU, M., LI, X., XIE, Y., BI, D., SUN, J., LI, J., TAI, C., DENG, Z. & OU, H.-Y. 2019b. ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic acids research*, 47, D660-D665.

LOH, Y. H., JAKSZYN, P., LUBEN, R. N., MULLIGAN, A. A., MITROU, P. N. & KHAW, K.-T. 2011. N-nitroso compounds and cancer incidence: the European Prospective Investigation into Cancer and Nutrition (EPIC)–Norfolk Study. *The American journal of clinical nutrition*, 93, 1053-1061.

- LU, H., HSU, P.-I., GRAHAM, D. Y. & YAMAOKA, Y. 2005. Duodenal ulcer promoting gene of *Helicobacter pylori*. *Gastroenterology*, 128, 833-848.
- MAEDA, S., YOSHIDA, H., IKENOUE, T., OGURA, K., KANAI, F., KATO, N., SHIRATORI, Y. & OMATA, M. 1999. Structure of cag pathogenicity island in Japanese *Helicobacter pylori* isoates. *Gut*, 44, 336-341.
- MALATY, H. M. 2007. Epidemiology of *Helicobacter pylori* infection. *Best Practice & Research Clinical Gastroenterology*, 21, 205-214.
- MALFERTHEINER, P., CHAN, F. K. & MCCOLL, K. E. 2009. Peptic ulcer disease. *The lancet*, 374, 1449-1461.
- MAX-PLANCK INSTITUTE. 2022. *MPI Bioinformatics Toolkits*. [Online]. Available: <https://toolkit.tuebingen.mpg.de/> [Accessed 23 June 2022].
- MÉGRAUD, F., BESSÈDE, E. & VARON, C. 2015. *Helicobacter pylori* infection and gastric carcinoma. *Clinical Microbiology and Infection*, 21, 984-990.
- MEHLA, J., CAUFIELD, J. H. & UETZ, P. 2015a. Mapping protein–protein interactions using yeast two-hybrid assays. *Cold Spring Harbor Protocols*, 2015, pdb. prot086157.
- MEHLA, J., CAUFIELD, J. H. & UETZ, P. 2015b. The yeast two-hybrid system: a tool for mapping protein–protein interactions. *Cold Spring Harbor Protocols*, 2015, pdb. top083345.

- MENON, S., PIRAMANAYAKAM, S. & AGARWAL, G. 2021. Computational identification of promoter regions in prokaryotes and Eukaryotes. *EPRA International Journal of Agriculture and Rural Economic Research (ARER)*, 9, 21-28.
- MILLER, A. K. & WILLIAMS, S. M. 2021. *Helicobacter pylori* infection causes both protective and deleterious effects in human health and disease. *Genes & Immunity*, 22, 218-226.
- MUÑOZ-RAMÍREZ, Z. Y., MENDEZ-TENORIO, A., KATO, I., BRAVO, M. M., RIZZATO, C., THORELL, K., TORRES, R., AVILES-JIMENEZ, F., CAMORLINGA, M. & CANZIAN, F. 2017. Whole genome sequence and phylogenetic analysis show *Helicobacter pylori* strains from Latin America have followed a unique evolution pathway. *Frontiers in cellular and infection microbiology*, 7, 50.
- NAGAI, H. & ROY, C. R. 2003. Show me the substrates: modulation of host cell function by type IV secretion systems. *Cellular microbiology*, 5, 373-383.
- NAKAMURA, K., TAMURA, R., YASUI, Y. & OKAJIMA, H. 2021. Atypical presentation at acute gastritis: significant gastric wall thickening as a presentation of a primary *Helicobacter pylori* infection in children. *BMJ Case Reports CP*, 14, e243912.
- NAKAMURA, S., SUGIYAMA, T., MATSUMOTO, T., IJIMA, K., ONO, S., TAJIKA, M., TARI, A., KITADAI, Y., MATSUMOTO, H. & NAGAYA, T. 2012. Long-term clinical outcome of gastric MALT lymphoma after eradication of *Helicobacter pylori*: a multicentre

cohort follow-up study of 420 patients in Japan. *Gut*, 61, 507-513.

NATIONAL LIBRARY OF MEDICINE. 16 February 2022. *National Center for Biotechnology Information*. [Online]. National Center for Biotechnology Information. Available: <https://www.ncbi.nlm.nih.gov/> [Accessed 23 June 2022].

NEW ENGLAND BIOLABS. 2022. *Tm Calculator version 1.13.1*. [Online]. Available: <https://tmcaculator.neb.com/#!/main> [Accessed 23 June 2022].

NGUYEN, L., UCHIDA, T., TSUKAMOTO, Y., KURODA, A., OKIMOTO, T., KODAMA, M., MURAKAMI, K., FUJIOKA, T. & MORIYAMA, M. 2010. *Helicobacter pylori dupA* gene is not associated with clinical outcomes in the Japanese population. *Clinical microbiology and infection*, 16, 1264-1269.

NGUYEN, T. H., HO, T. T. M., NGUYEN-HOANG, T.-P., QUMAR, S., PHAM, T. T. D., BUI, Q. N., BULACH, D., NGUYEN, T.-V. & RAHMAN, M. 2021. The endemic *Helicobacter pylori* population in Southern Vietnam has both South East Asian and European origins. *Gut pathogens*, 13, 1-14.

NOTO, J. M., GADDY, J. A., LEE, J. Y., PIAZUELO, M. B., FRIEDMAN, D. B., COLVIN, D. C., ROMERO-GALLO, J., SUAREZ, G., LOH, J. & SLAUGHTER, J. C. 2012. Iron deficiency accelerates *Helicobacter pylori*-induced carcinogenesis in rodents and humans. *The Journal of clinical investigation*, 123.

- OGBODO, E. N. 2021. *Immunoregulatory properties of Helicobacter pylori derived molecules*. University of Nottingham.
- OLBERMANN, P., JOSENHANS, C., MOODLEY, Y., UHR, M., STAMER, C., VAUTERIN, M., SUERBAUM, S., ACHTMAN, M. & LINZ, B. 2010. A global overview of the genetic and functional diversity in the *Helicobacter pylori* cag pathogenicity island. *PLoS genetics*, 6, e1001069.
- OSEK, J., GALLIEN, P., TRUSZCZYŃSKI, M. & PROTZ, D. 1999. The use of polymerase chain reaction for determination of virulence factors of *Escherichia coli* strains isolated from pigs in Poland. *Comparative immunology, microbiology and infectious diseases*, 22, 163-174.
- OSMAN, E., ELNOSH, M. M., ALBASHA, A. M., FADL, A. A., MAROUF, L. O. I. & ALTAYB, H. N. 2021. Dissemination of *cagA* and *cagE* Virulence Genes Among *H. pylori* Strains From Sudanese Patients With Gastric Discomfort.
- ÖZTEKIN, M., YILMAZ, B., AĞAGÜNDÜZ, D. & CAPASSO, R. 2021. Overview of *Helicobacter pylori* Infection: Clinical Features, treatment, and nutritional aspects. *Diseases*, 9, 66.
- PAIANO, A., MARGIOTTA, A., DE LUCA, M. & BUCCI, C. 2019. Yeast Two-Hybrid Assay to Identify Interacting Proteins. *Current protocols in protein science*, 95, e70.
- PALAU, M., PIQUÉ, N., RAMÍREZ-LÁZARO, M. J., LARIO, S., CALVET, X. & MIÑANA-GALBIS, D. 2021. Whole-Genome Sequencing and Comparative Genomics of Three *Helicobacter*

- pylori* Strains Isolated from the Stomach of a Patient with Adenocarcinoma. *Pathogens*, 10, 331.
- PALFRAMAN, S. L., KWOK, T. & GABRIEL, K. 2012. Vacuolating cytotoxin A (VacA), a key toxin for *Helicobacter pylori* pathogenesis. *Frontiers in cellular and infection microbiology*, 2, 92.
- PAN, X. & VACHET, R. W. 2022. Membrane protein structures and interactions from covalent labeling coupled with mass spectrometry. *Mass Spectrometry Reviews*, 41, 51-69.
- PATEL, S., SMITH, K., LETLEY, D., COOK, K., MEMON, A., INGRAM, R., STAPLES, E., BACKERT, S., ZAITOUN, A. & ATHERTON, J. 2013. *Helicobacter pylori* downregulates expression of human β -defensin 1 in the gastric mucosa in a type IV secretion-dependent fashion. *Cellular microbiology*, 15, 2080-2092.
- PEEK JR, R. M., MILLER, G. G., THAM, K. T., PEREZ-PEREZ, G. I., ZHAO, X., ATHERTON, J. C. & BLASER, M. J. 1995. Heightened inflammatory response and cytokine expression in vivo to *cagA*⁺ *Helicobacter pylori* strains. *Laboratory investigation; a journal of technical methods and pathology*, 73, 760-770.
- PEREZ-PEREZ, G. I., ROTHENBACHER, D. & BRENNER, H. 2004. Epidemiology of *Helicobacter pylori* infection. *Helicobacter*, 9, 1-6.
- PHUC, B. H., TUAN, V. P., DUNG, H. D. Q., BINH, T. T., TUNG, P. H., TRI, T. D., THUAN, N. P. M., VAN KHIEU, V., TRANG, T. T. H.

- & AKADA, J. 2021a. *Helicobacter pylori* type 4 secretion systems as gastroduodenal disease markers. *Scientific reports*, 11, 1-12.
- PHUC, B. H., TUAN, V. P., DUNG, H. D. Q., BINH, T. T., TUNG, P. H., TRI, T. D., THUAN, N. P. M., VAN KHIEU, V., TRANG, T. T. H., AKADA, J., MATSUMOTO, T. & YAMAOKA, Y. 2021b. *Helicobacter pylori* type 4 secretion systems as gastroduodenal disease markers. *Sci Rep*, 11, 4584.
- POPLAWSKI, T., CHOJNACKI, C., CZUBATKA, A., KLUPINSKA, G., CHOJNACKI, J. & BLASIAK, J. 2013. *Helicobacter pylori* infection and antioxidants can modulate the genotoxic effects of heterocyclic amines in gastric mucosa cells. *Molecular biology reports*, 40, 5205-5212.
- PREMIER BIOSOFT. 1994 - 2022. *NET Primer*. [Online]. Available: <http://www.premierbiosoft.com/netprimer/index.html> [Accessed 23 June 2022].
- RAHMAN, M. T., UDDIN, M. S., SULTANA, R., MOUE, A. & SETU, M. 2013. Polymerase chain reaction (PCR): a short review. *Anwer Khan Modern Medical College Journal*, 4, 30-36.
- RAMAKRISHNAN, K. & SALINAS, R. C. 2007. Peptic ulcer disease. *American family physician*, 76, 1005-1012.
- RASHKOVA, S., ZHOU, X.-R., CHEN, J. & CHRISTIE, P. J. 2000. Self-assembly of the *Agrobacterium tumefaciens* VirB11 traffic ATPase. *Journal of Bacteriology*, 182, 4137-4145.

- REDDY CHICHILI, V. P., KUMAR, V. & SIVARAMAN, J. 2013. Linkers in the structural biology of protein–protein interactions. *Protein science*, 22, 153-167.
- RESHETNYAK, V. I. & RESHETNYAK, T. M. 2017. Significance of dormant forms of *Helicobacter pylori* in ulcerogenesis. *World journal of gastroenterology*, 23, 4867.
- REYES-LEON, A., ATHERTON, J. C., ARGENT, R. H., PUENTE, J. & TORRES, J. 2007. Heterogeneity in the activity of Mexican *Helicobacter pylori* strains in gastric epithelial cells and its association with diversity in the *cagA* gene. *Infection and immunity*, 75, 3445-3454.
- RHEAD, J. L., LETLEY, D. P., MOHAMMADI, M., HUSSEIN, N., MOHAGHEGHI, M. A., HOSSEINI, M. E. & ATHERTON, J. C. 2007. A new *Helicobacter pylori* vacuolating cytotoxin determinant, the intermediate region, is associated with gastric cancer. *Gastroenterology*, 133, 926-936.
- RIPOLL-ROZADA, J., ZUNZUNEGUI, S., DE LA CRUZ, F., ARECHAGA, I. & CABEZÓN, E. 2013. Functional interactions of VirB11 traffic ATPases with VirB4 and VirD4 molecular motors in type IV secretion systems. *Journal of bacteriology*, 195, 4195-4201.
- ROBINSON, K., ARGENT, R. H. & ATHERTON, J. C. 2007. The inflammatory and immune response to *Helicobacter pylori* infection. *Best practice & research Clinical gastroenterology*, 21, 237-259.

- ROBINSON, K. & ATHERTON, J. C. 2021. The Spectrum of *Helicobacter*-Mediated Diseases. *Annual Review of Pathology: Mechanisms of Disease*, 16, 123-144.
- ROMO-GONZÁLEZ, C., SALAMA, N. R., BURGEO-FERREIRA, J., PONCE-CASTANEDA, V., LAZCANO-PONCE, E., CAMORLINGA-PONCE, M. & TORRES, J. 2009. Differences in genome content among *Helicobacter pylori* isolates from patients with gastritis, duodenal ulcer, or gastric cancer reveal novel disease-associated genes. *Infection and immunity*, 77, 2201-2211.
- ROSANO, G. L. & CECCARELLI, E. A. 2014. Recombinant protein expression in *Escherichia coli*: advances and challenges. *Frontiers in microbiology*, 5, 172.
- ROUMPEKA, D., WALLACE, R., ESCALETES, F., FOTHERINGHAM, I. & WATSON, M. 2017. A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. *Front Genet.* 2017; 8: 23.
- SALIH, B. A., GUNER, A., KARADEMIR, A., USLU, M., OVALI, M. A., YAZICI, D., BOLEK, B. K. & ARIKAN, S. 2014. Evaluation of the effect of *cagPAI* genes of *Helicobacter pylori* on AGS epithelial cell morphology and IL-8 secretion. *Antonie Van Leeuwenhoek*, 105, 179-189.
- SAVVIDES, S. N., YEO, H.-J., BECK, M. R., BLAESING, F., LURZ, R., LANKA, E., BUHRDORF, R., FISCHER, W., HAAS, R. & WAKSMAN, G. 2003. VirB11 ATPases are dynamic hexameric

- assemblies: new insights into bacterial type IV secretion. *The EMBO journal*, 22, 1969-1980.
- SCHRAW, W., LI, Y., MCCLAIN, M. S., VAN DER GOOT, F. G. & COVER, T. L. 2002. Association of *Helicobacter pylori* vacuolating toxin (VacA) with lipid rafts. *Journal of Biological Chemistry*, 277, 34642-34650.
- SEEMANN, T. 2018. ABRicate: mass screening of contigs for antimicrobial and virulence genes. Department of Microbiology and Immunology, The University of Melbourne, Melbourne, Australia.
- SEYDEL, A., TASCA, E., BERTI, D., RAPPUOLI, R., DEL GIUDICE, G. & MONTECUCCO, C. 2002. Characterization and immunogenicity of the CagF protein of the cag pathogenicity island of *Helicobacter pylori*. *Infect Immun*, 70, 6468-70.
- SHARMA, S. A., TUMMURU, M. K., BLASER, M. J. & KERR, L. D. 1998. Activation of IL-8 gene expression by *Helicobacter pylori* is regulated by transcription factor nuclear factor- κ B in gastric epithelial cells. *The journal of immunology*, 160, 2401-2407.
- SHARNDAMA, H. C. & MBA, I. E. 2022. *Helicobacter pylori*: an up-to-date overview on the virulence and pathogenesis mechanisms. *Brazilian Journal of Microbiology*, 1-18.
- SHIOTA, S., MATSUNARI, O., WATADA, M., HANADA, K. & YAMAOKA, Y. 2010. Systematic review and meta-analysis: the relationship between the *Helicobacter pylori* *dupA* gene and clinical outcomes. *Gut pathogens*, 2, 1-6.

- SMITH, S., FOWORA, M. & PELLICANO, R. 2019. Infections with *Helicobacter pylori* and challenges encountered in Africa. *World journal of gastroenterology*, 25, 3183.
- SMITH, S. I., AJAYI, A., JOLAIYA, T., ONYEKWERE, C., SETSHEDI, M., SCHULZ, C., OTEGBAYO, J. A., NDIP, R., DIEYE, Y. & ALBORAIE, M. 2021. *Helicobacter pylori* infection in Africa: update of the current situation and challenges. *Digestive Diseases*.
- SMITH, S. M. 2017. Strategies for the purification of membrane proteins. *Protein Chromatography*. Springer.
- SONNENBERG, A. 2022. Epidemiology of *Helicobacter pylori*. *Alimentary pharmacology & therapeutics*, 55, S1-S13.
- STEIN, M., BAGNOLI, F., HALENBECK, R., RAPPUOLI, R., FANTL, W. J. & COVACCI, A. 2002. c-Src/Lyn kinases activate *Helicobacter pylori* CagA through tyrosine phosphorylation of the EPIYA motifs. *Molecular microbiology*, 43, 971-980.
- STINGL, K., MÜLLER, S., SCHEIDGEN-KLEYBOLDT, G., CLAUSEN, M. & MAIER, B. 2010. Composite system mediates two-step DNA uptake into *Helicobacter pylori*. *Proceedings of the National Academy of Sciences*, 107, 1184-1189.
- STOLER, N. & NEKRUTENKO, A. 2021. Sequencing error profiles of Illumina sequencing instruments. *NAR genomics and bioinformatics*, 3, lqab019.

- SUKRI, A., HANAFIAH, A., MOHAMAD ZIN, N. & KOSAI, N. R. 2020. Epidemiology and role of *Helicobacter pylori* virulence factors in gastric cancer carcinogenesis. *Apmis*, 128, 150-161.
- TATO, I., MATILLA, I., ARECHAGA, I., ZUNZUNEGUI, S., DE LA CRUZ, F. & CABEZON, E. 2007. The ATPase activity of the DNA transporter TrwB is modulated by protein TrwA: implications for a common assembly mechanism of DNA translocating motors. *Journal of Biological Chemistry*, 282, 25569-25576.
- TERRADOT, L. & WAKSMAN, G. 2011. Architecture of the *Helicobacter pylori* Cag-type IV secretion system. *The FEBS journal*, 278, 1213-1222.
- THORELL, K., LEHOURS, P. & VALE, F. F. 2017. Genomics of *Helicobacter pylori*. *Helicobacter*, 22, e12409.
- TOMB, J.-F., WHITE, O., KERLAVAGE, A. R., CLAYTON, R. A., SUTTON, G. G., FLEISCHMANN, R. D., KETCHUM, K. A., KLENK, H. P., GILL, S. & DOUGHERTY, B. A. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*, 388, 539-547.
- TSERENTOGTOKH, T., GANTUYA, B., SUBSOMWONG, P., OYUNTSETSEG, K., BOLOR, D., ERDENE-OCHIR, Y., AZZAYA, D., DAVAADORJ, D., UCHIDA, T. & MATSUHISA, T. 2019. Western-type *Helicobacter pylori* CagA are the most frequent type in Mongolian patients. *Cancers*, 11, 725.

- UNDERWOOD, A. P., MULDER, A., GHARBIA, S. & GREEN, J. 2005. Virulence Searcher: a tool for searching raw genome sequences from bacterial genomes for putative virulence factors. *Clinical microbiology and infection*, 11, 770-772.
- URS, A. N., NARULA, P. & THOMSON, M. 2014. Peptic ulcer disease. *Paediatrics and Child Health*, 24, 485-490.
- VALONES, M. A. A., GUIMARÃES, R. L., BRANDÃO, L. A. C., SOUZA, P. R. E. D., CARVALHO, A. D. A. T. & CROVELA, S. 2009. Principles and applications of polymerase chain reaction in medical diagnostic fields: a review. *Brazilian Journal of Microbiology*, 40, 1-11.
- VAN STRALEN, K. J., STEL, V. S., REITSMA, J. B., DEKKER, F. W., ZOCCALI, C. & JAGER, K. J. 2009. Diagnostic methods I: sensitivity, specificity, and other measures of accuracy. *Kidney international*, 75, 1257-1263.
- VELÁZQUEZ, M. & FEIRTAG, J. M. 1999. *Helicobacter pylori*: characteristics, pathogenicity, detection methods and mode of transmission implicating foods and water. *International journal of food microbiology*, 53, 95-104.
- VICARI, J. J., PEEK, R. M., FALK, G. W., GOLDBLUM, J. R., EASLEY, K. A., SCHNELL, J., PEREZ-PEREZ, G. I., HALTER, S. A., RICE, T. W. & BLASER, M. J. 1998. The seroprevalence of *cagA*-positive *Helicobacter pylori* strains in the spectrum of gastroesophageal reflux disease. *Gastroenterology*, 115, 50-57.

- WALLDEN, K., RIVERA-CALZADA, A. & WAKSMAN, G. 2010.
Microreview: Type IV secretion systems: versatility and diversity in function. *Cellular microbiology*, 12, 1203-1212.
- WANG, Y.-K., KUO, F.-C., LIU, C.-J., WU, M.-C., SHIH, H.-Y., WANG, S. S., WU, J.-Y., KUO, C.-H., HUANG, Y.-K. & WU, D.-C. 2015.
Diagnosis of *Helicobacter pylori* infection: Current options and developments. *World Journal of Gastroenterology: WJG*, 21, 11221.
- WARD, D. V., DRAPER, O., ZUPAN, J. R. & ZAMBRYSKI, P. C. 2002.
Peptide linkage mapping of the *Agrobacterium tumefaciens* vir-encoded type IV secretion system reveals protein subassemblies. *Proceedings of the National Academy of Sciences*, 99, 11493-11500.
- WARING, A., DRAKE, I., SCHORAH, C., WHITE, K., LYNCH, D., AXON, A. & DIXON, M. 1996. Ascorbic acid and total vitamin C concentrations in plasma, gastric juice, and gastrointestinal mucosa: effects of gastritis and oral supplementation. *Gut*, 38, 171-176.
- WASYLYK, B. 1985. Promoter elements of eukaryotic protein-coding genes. *Chromosomal Proteins and Gene Expression*. Springer.
- WEISS, B., JACQUEMIN-SABLON, A., LIVE, T. R., FAREED, G. C. & RICHARDSON, C. C. 1968. Enzymatic Breakage and Joining of Deoxyribonucleic Acid: VI. Further Purification And Properties Of Polynucleotide Ligase From *Escherichia Coli* Infected With

- Bacteriophage T4. *Journal of Biological Chemistry*, 243, 4543-4555.
- WEISS, E., SPICHER, C., HAAS, R. & FISCHER, W. 2019. Excision and transfer of an integrating and conjugative element in a bacterial species with high recombination efficiency. *Scientific reports*, 9, 1-14.
- WICK, R. R., SCHULTZ, M. B., ZOBEL, J. & HOLT, K. E. 2015. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, 31, 3350-3352.
- WILKINSON, D. J. 2019. *High throughput genomic analysis of Helicobacter pylori within-host diversity*. Nottingham Trent University.
- WISEMAN, M. 2008. The second world cancer research fund/american institute for cancer research expert report. food, nutrition, physical activity, and the prevention of cancer: a global perspective: nutrition society and BAPEN Medical Symposium on 'nutrition support in cancer therapy'. *Proceedings of the Nutrition Society*, 67, 253-256.
- WOLLE, K. & MALFERTHEINER, P. 2007. Treatment of *Helicobacter pylori*. *Best Practice & Research Clinical Gastroenterology*, 21, 315-324.
- WOODWARD, M., TUNSTALL-PEDOE, H. & MCCOLL, K. 2001. *Helicobacter pylori* infection reduces systemic availability of dietary vitamin C. *European journal of gastroenterology & hepatology*, 13, 233-237.

- WROBLEWSKI, L. E., PEEK JR, R. M. & WILSON, K. T. 2010.
Helicobacter pylori and gastric cancer: factors that modulate disease risk. *Clinical microbiology reviews*, 23, 713-739.
- YAMAOKA, Y. 2010. Mechanisms of disease: *Helicobacter pylori* virulence factors. *Nature reviews Gastroenterology & hepatology*, 7, 629-641.
- YAMAOKA, Y., EL-ZIMAITY, H. M., GUTIERREZ, O., FIGURA, N., KIM, J. K., KODAMA, T., KASHIMA, K. & GRAHAM, D. Y. 1999a. Relationship between the *cagA* 3'repeat region of *Helicobacter pylori*, gastric histology, and susceptibility to low pH. *Gastroenterology*, 117, 342-349.
- YAMAOKA, Y., KODAMA, T., GUTIERREZ, O., KIM, J. G., KASHIMA, K. & GRAHAM, D. Y. 1999b. Relationship between *Helicobacter pylori* *iceA*, *cagA*, and *vacA* status and clinical outcome: studies in four different countries. *Journal of clinical microbiology*, 37, 2274-2279.
- YAMAOKA, Y., KODAMA, T., KASHIMA, K., GRAHAM, D. Y. & SEPULVEDA, A. R. 1998. Variants of the 3' region of the *cagA* gene in *Helicobacter pylori* isolates from patients with different H. pylori-associated diseases. *Journal of clinical microbiology*, 36, 2258-2263.
- YAMAZAKI, S., YAMAKAWA, A., OKUDA, T., OHTANI, M., SUTO, H., ITO, Y., YAMAZAKI, Y., KEIDA, Y., HIGASHI, H. & HATAKEYAMA, M. 2005. Distinct diversity of *vacA*, *cagA*, and

- cagE* genes of *Helicobacter pylori* associated with peptic ulcer in Japan. *Journal of clinical microbiology*, 43, 3906-3916.
- YANG, S. & ROTHMAN, R. E. 2004. PCR-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings. *The Lancet infectious diseases*, 4, 337-348.
- YOUNG, C. L., BRITTON, Z. T. & ROBINSON, A. S. 2012. Recombinant protein expression and purification: a comprehensive review of affinity tags and microbial applications. *Biotechnology journal*, 7, 620-634.
- YOUSSEFI, M., GHAZVINI, K., FARSIANI, H., TAFAGHODI, M. & KEIKHA, M. 2020. A systematic review and meta-analysis of outcomes of infection with *Helicobacter pylori* *dupA*⁺ strains in Iranian patients. *Gene Reports*, 19, 100650.
- YUAN, Q., CARLE, A., GAO, C., SIVANESAN, D., ALY, K. A., HÖPPNER, C., KRALL, L., DOMKE, N. & BARON, C. 2005. Identification of the VirB4-VirB8-VirB5-VirB2 pilus assembly sequence of type IV secretion systems. *Journal of Biological Chemistry*, 280, 26349-26359.
- ZAIDI, S. F. 2016. *Helicobacter pylori* associated Asian enigma: Does diet deserve distinction? *World Journal of Gastrointestinal Oncology*, 8, 341.
- ZUBAY, G., MORSE, D. E., SCHRENK, W. J. & MILLER, J. 1972. Detection and isolation of the repressor protein for the tryptophan operon of *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 69, 1100-1103.

- ZULLO, A., HASSAN, C., ANDRIANI, A., CRISTOFARI, F., DE
FRANCESCO, V., IERARDI, E., TOMAO, S., MORINI, S. &
VAIRA, D. 2009. Eradication Therapy for *Helicobacter pylori* in
Patients With Gastric MALT Lymphoma: A Pooled Data Analysis
CME. *Official journal of the American College of
Gastroenterology*| ACG, 104, 1932-1937.
- ZUPAN, J., HACKWORTH, C. A., AGUILAR, J., WARD, D. &
ZAMBRYSKI, P. 2007. VirB1* promotes T-pilus formation in the
vir-Type IV secretion system of *Agrobacterium tumefaciens*.
Journal of bacteriology, 189, 6551-6563.

APPENDIX

Table 1. Summary of yeast-two hybrid assay results (3= *tfs3*, 4= *tfs4*).

MUA replica	pGBT9																
	3B4	3B8	3B10	3B11	3D2	3D4	4B4	4B8	4B9	4B10	4B11	4D2	4D4	4B4v	4B8v	4B11v	gB11
3B4	++			+++			NR	NR	NR	NR	NR		NR	NA	NA	NA	++
3B8	++	++/+++	++	++		NR	NR	+	NR	NR	NR		NR	NA	+	NA	NA
3B9	NR					NR	NR	NR	NR	NR	NR		NR	NA	NA	NA	NA
3B10		++	+++	++			NR	NR	NR	NR	NR		NR	+	NA	+	x
3B11	++	++	++	+++ / +++		++	NR	NR	NR	NR	x		NR	+	+++	+	+
3D2				++			+++				++			+++		++	
3D4		NR		++	+++		+++	NR	NR	NR	+++	+++	NR	++	NA	+++	NA
4B4	NR	NR	NR	NR		NR		++	NR	++			+ / x	NA	+++	NA	++
4B8	NR	NR	NR	NR		NR	++ / ++	++ / ++	++	+	+		+	++	+++ / +++	NA	NA
4B9	NR	NR	NR	NR		NR	NR	+++	+		NR		NR	++	NA	x	NA
4B10	NR	NR	NR	NR		NR			+	+				x	x	++	NA
4B11	+	NR	NR	+		+++	+++	+	NR	+	++ / +++			++	+	+	+
4D2	+	+	+++	+++		+++	+	+	+++	NA	+		+	++	++	+	+
4D4	NA	NA	+	+		+	++ / ++	NR	NR	+++				+	+++	+	+
4B4v	NA	NA	+	+	+++	++	+	x	+	+++	+		+	NA	++	+	+
4B8v	+++	+	NA	+		+	+++	+++ / +++	x	++	++		+	+	++	+	+
4B11v	NA	NA	NA	++		+++	+	++	x	+	+		+	++	++	+++	+
gB11	++	+	x	+		+	++	+	+	++	++		+	+	+	+	+++

MU replica	pGBT9																
	3B4	3B8	3B10	3B11	3D2	3D4	4B4	4B8	4B9	4B10	4B11	4D2	4D4	4B4v	4B8v	4B11v	gB11
3B4	x			+++			NR	NR	NR	NR	NR		NR	NA	NA	NA	++
3B8	x	++/+++	x	+++		NR	NR	x	NR	NR	NR		NR	NA	x	NA	NA
3B9	NR					NR	NR	NR	NR	NR	NR		NR	NA	NA	NA	NA
3B10		x	x	+++			NR	NR	NR	NR	NR		NR	x	NA	x	x
3B11	+++	+++	+++	+++ / +++			NR	NR	NR	NR	x		NR	+	+	x	x
3D2				x			+							+		x	
3D4		NR		+++	+		+	NR	NR	NR	+	+	NR	x	NA	x	NA
4B4	NR	NR	NR	NR		NR		+	NR	x			x	NA	++	NA	+
4B8	NR	NR	NR	NR		NR	+/+	+/+	x	x			x	x	++/+++	x	NA
4B9	NR	NR	NR	NR		NR	NR	+	x		NR		NR	+	NA	x	NA
4B10	NR	NR	NR	NR		NR			+	x				x	x	x	NA
4B11	x	NR	NR	+		x	x	x	NR	x	+/+++			+	x	+	+
4D2	+	x	x	+++		x	x	x	+	x	+		+	x	x	x	+
4D4	NA	NA	x	x		+	x	NR	NR	+				x	++	x	x
4B4v	NA	NA	x	x	+	x	x	x	x	x	x		x	NA	++	x	+
4B8v	x	+	NA	x		x	++	+/+++	x	x	+		x	+	++	+	x
4B11v	NA	NA	NA	+		+	+	+	x	x	+		x	+	++	+++	+
gB11	+	+	x	+		+	+	x	x	x	+		x	x	x	x	+++

Colour code

	Known homologous VirB interactions (Fronzes et al., 2009)
NR	Not reported interaction by Fronzes et al., 2009
	Possible interaction according to Fronzes et al., 2009
Plus/plus	This study/Alandiyan 2015 <i>tfs</i> interaction
Plus red only	Alandiyan 2015. additional interactions reported
Plus/plus	This study/ Alandiyan 2015 interaction positive with Fronzes et al., 2009

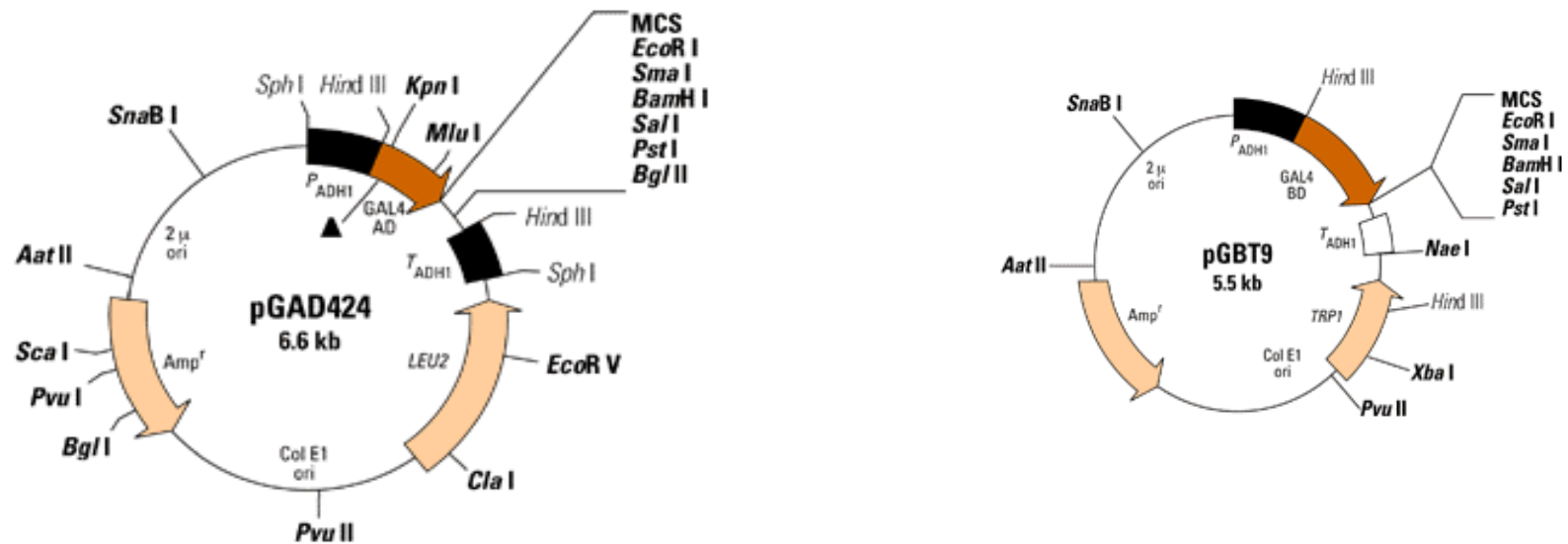


Fig 1: Maps of cloning vectors pGAD4242 (with activating domain) and pGBT9 (with binding domain) (addgene)

**Map of
pBAD/Myc-His**

The figure below summarizes the features of the pBAD/Myc-His vector. Complete sequences for all three pBAD/Myc-His vectors are available for downloading from our World Wide Web site (www.invitrogen.com) or by contacting Technical Service (see page 27). Details of each multiple cloning site are shown on pages 13-15.

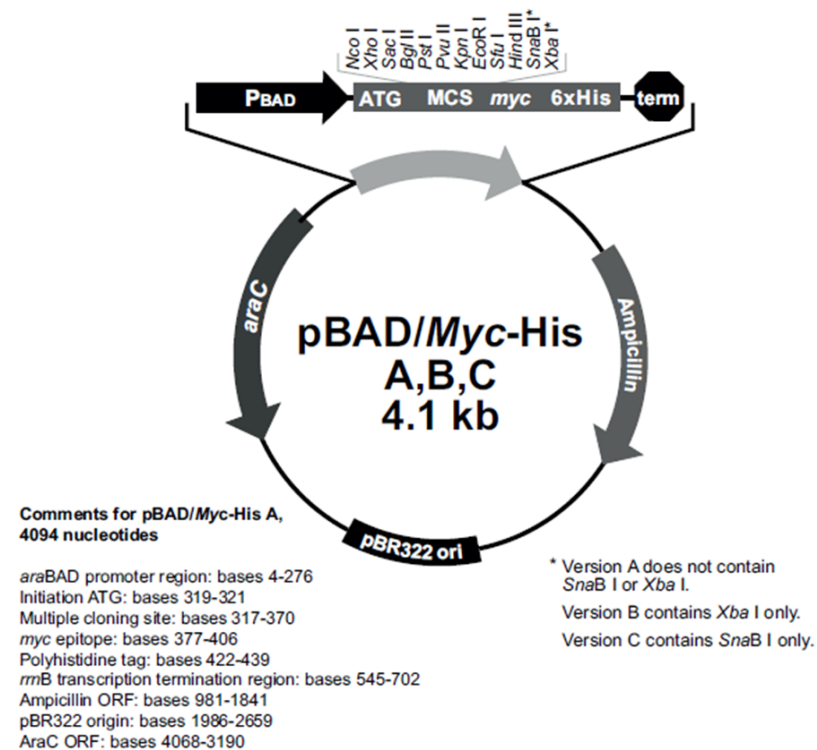


Fig 2: Map of pBAD expression vector (ThermoFisher Scientific)

<https://www.icloud.com/iclouddrive/06cY6OjxSKMpc4BYJPLFDjdmg#P12-results>

https://www.icloud.com/iclouddrive/07aC-Kg7E_rgTLzoN_k_62T_w#194A-result_ALL

<https://www.icloud.com/iclouddrive/07bX2fHK93CnUiLaYBqGXhi7g#results-P12>

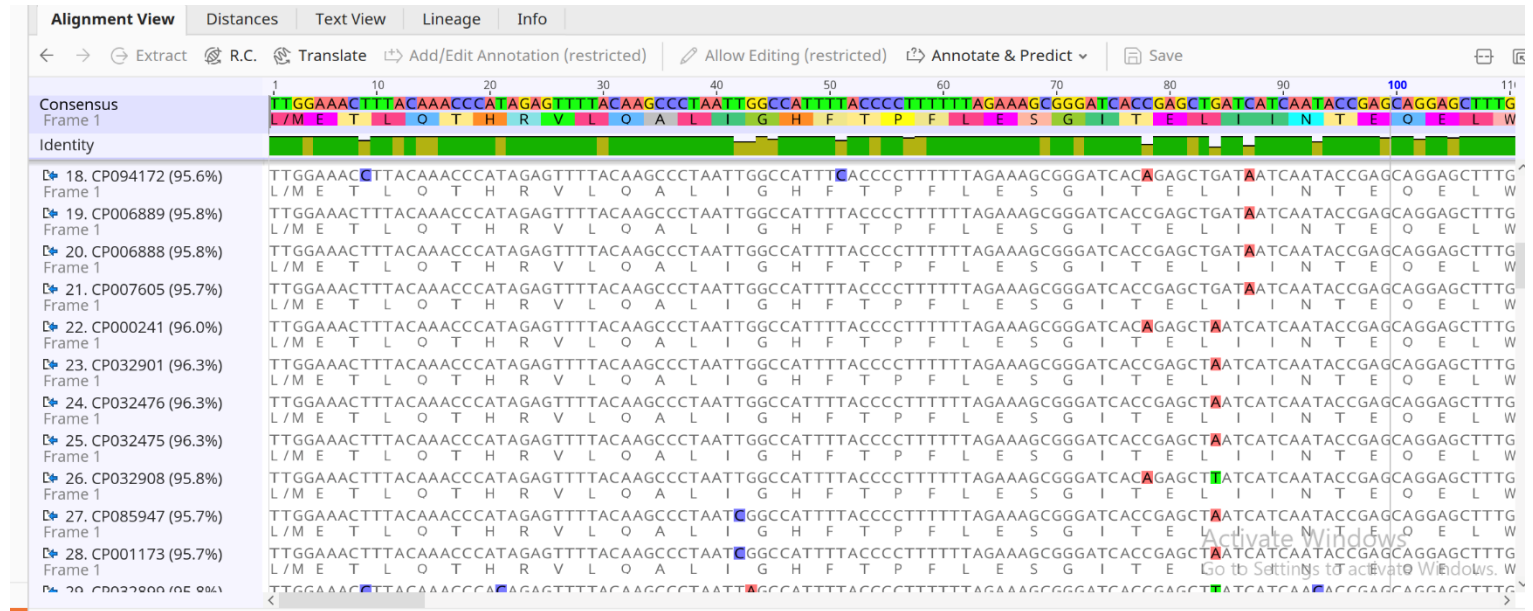
<https://www.icloud.com/iclouddrive/055SkLJrhyF54gWJwJxzjOUuw#results-Shi470>

Links to ABRicate results for the detection of *tfs4* cluster of genes in clinical isolates using P12 and Shi470 as reference strains.

A



B



C



Fig 3: A, B and C Showing sequence percentage identity (left) of aligned gB11 with homologous sequences from other strains of *H. pylori*. Image generated with Geneious Prime software (Restricted).