Stanford - South Africa

Biomedical Informatics Program

Identification of Bacterial Pathogenic Gene Classes Subject to Diversifying Selection

NAME : **Sumir Panji**

STUDENT NUMBER : **2355015**

SUPERVISOR : **Professor Winston Hide**

CO-SUPERVISOR : **Professor Vladimir Bajic**

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor Philosophiae in the Department of Biotechnology, University of the Western Cape.

2009

RESPICE PROSPICE

UNIVERSITY *of the* WESTERN CAPE

SANBI

Keywords

Helicobacter pylori

Neisseria meningitidis

Vibrio Cholerae

Positive Selection

Virulence Gene

Nucleotide Diversity

Statistical Enrichment

Functional Annotation

Biological Processes

Metabolic Pathways

Abstract

Availability of genome sequences for numerous bacterial species comprising of different bacterial strains allows elucidation of species and strain specific adaptations that facilitate their survival in widely fluctuating micro-environments and enhance their pathogenic potential. Different bacterial species use different strategies in their pathogenesis and the pathogenic potential of a bacterial species is dependent on its genomic complement of virulence factors. A bacterial virulence factor, within the context of this study, is defined as any endogenous protein product encoded by a gene that aids in the adhesion, invasion, colonization, persistence and pathogenesis of a bacterium within a host. Anecdotal evidence suggests that bacterial virulence genes are undergoing diversifying evolution to counteract the rapid adaptability of its host's immune defences. Genome sequences of pathogenic bacterial species and strains provide unique opportunities to study the action of diversifying selection operating on different classes of bacterial genes.

A computational pipeline was developed to assay for strain specific genes undergoing diversifying selection between *Helicobacter pylori* 26695 and J99, *Neisseria meningitidis* Z2491 and MC58 as well as *Vibrio cholerae* N16961 and O395. Existing databases housing collections of virulence genes curated from literature were mined for known bacterial virulence genes. Characterisation of functional biological processes for genes under positive selection using Gene Ontology (GO), Clusters of Orthologous Proteins (COG), sub-cellular localisation prediction and metabolic pathways was undertaken to determine common and unique processes under positive selection within the three bacterial species. Functional characterisation of *H. pylori*, *N. meningitidis* and *V. cholerae* known bacterial virulence genes was performed to elucidate shared and species unique modes of bacterial pathogenicity. Comparisons between genes under positive selection and known bacterial virulence genes were conducted to discover whether known virulence genes and the biological processes they are involved in are enriched for diversifying selection.

Apart from *N. meningitidis*, no statistically significant intersection could be established when genes under positive selection and known virulence genes are examined as separate entities, suggesting positive selection and bacterial virulence operate in a parallel, unrelated manner. However, a substantial overlap between the biological processes and functional annotations of genes under positive selection and known virulence genes was found. Four common bacterial virulence metabolic pathways were also found to contain genes under positive selection and known virulence genes within all three bacterial species, one of them (Lysine Biosynthesis I) is currently being investigated as an anti-bacterial target, thereby demonstrating a link between nucleotide diversity and bacterial pathogenesis.

Declaration

I declare that "*Identification of Bacterial Pathogenic Gene Classes Subject to Diversifying Selection*" is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Name: Sumir Panji **Date:** 24th November 2009.

Signed:.....

Acknowledgements

First and foremost I would like to thank Professor Winston Hide (Win), for his unwavering support, encouragement and personal dedication throughout my Doctoral studies. Win has constantly and unceasingly been a driving force in cultivating numerous collaborations and opportunities to ensure that I managed to acquire and develop inter-disciplinary skills that will stand me in good stead when pursuing my career in science.

I would like to gratefully acknowledge the Stanford South Africa Biomedical Informatics (SSABMI) program for enabling to me to pursue my PhD studies through funding and provision of numerous opportunities to broaden my inter-disciplinary scientific skill base. The SSABMI is a joint training program between Stanford University, University of the Western Cape, University of Cape Town and the National Institute of Communicable Diseases and is funded by the Fogarty Centre, a section of NIH, under award TW006993.

I would like to extend a heartfelt thanks to Dr. Mike Walker, consultant statistician at Stanford University for taking using his personal time to demystify and provide me with invaluable tutelage in statistics and R as part of the SSABMI programme.

I would like to thank Dr. Junaid Gamieldien for providing me with the initial positive selection detection pipeline PERL scripts he developed and taking time to explain how portions of the pipeline work.

I am grateful to both Dr. Konrad Scheffler for his detailed explanations of the CODEML control file parameters and how to conduct the Likelihood Ratio Tests (LRT) between CODEML's evolutionary models and Dr. Cathal Seoighe for his many insightful discussions on recombination and its affects upon positive selection assays.

I would like to say a big "Thank You" to Mario Jonas for his patience and invaluable help with all the PERL-ing and Allan Kamau for his assistance in logistics which was a big help towards completion of my research.

I would finally like to express my gratitude to my parents, Noordin and Usha Panji for their unconditional, unwavering and unequivocal support throughout my academic endeavours and pursuit of my higher education – "Thank you Mum and Dad".

List of Figures

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Appendix B1

Appendix B2

Appendix B4

Appendix C1

Appendix C2

Table of Contents

Chapter 1 : Microbial Genomes, Positive Selection and Bacterial Virulence

Introduction

Advances in genome sequencing technologies, assembly and annotation have driven the explosion of publicly available genomes for a myriad of microbial organisms. At present, there are 1,136 completed genome projects listed in the Genomes Online Database as of 12th November 2009, 83% (945 of 1,136 complete genomes) are from the Bacterial domain and are pre-dominantly human pathogens of medical importance[•]. The plethora of microbial genomes offers unprecedented opportunities to investigate bacterial genome evolution, identify virulent determinants, characterize microbial gene content and metabolic pathways while concurrently creating avenues for novel diagnostic and anti-microbial therapies [1-6]. The glut of genomic information has actuated development of numerous *in silico* tools for storage, annotation, dissemination and analysis of biological data and given rise to comparative genomics, an integral approach in the annotation, analysis and contextualisation of sequenced genomes [1,2,4-9].

A major goal in the study of microbial pathogens is the identification and characterization of microbial genes that facilitate invasion and colonization of a host species [2,5,6,9-12]. Systematic explorations of microbial genomic diversity and the dynamics shaping adaptations ensuing the survival and success of bacterial pathogens is now possible at the genome level [1,2,6,9,12-17]. Development of sophisticated computational tools and rigorous models of evolution coupled with the availability of microbial genomes provides unique opportunities to investigate the effect of evolutionary selective forces on microbial virulence determinants [2,6,9,12,15,17,18]. The cornucopia of available microbial genomes is being leveraged by researchers around the world to understand the impact of evolutionary selective pressures on microbial virulence determinants that enable the continued success of a microbial pathogen [13,15,17-19].

One of the fastest evolving gene classes in mammalian genomes are immune response related genes due to the dynamic inimical interplay between bacterial infection and host immunity [20-22]. Within this "biological arms race", microbial pathogens are under strong selective pressure to rapidly alter their genomic complement to ensure continued survival within their hosts and in widely fluctuating micro-environments [5,6,11]. Elucidating these signatures of rapid evolution from microbial genomes will provide glimpses to the origins of pathogenicity in response to selective pressures exerted by a bacteria's host environment, thereby establishing a relationship between microbial genome sequence diversity and pathogenesis [5-7,13].

[•] http://www.genomesonline.org/gold.cgi?want=Published+Complete+Genomes

Microbial Genomes – Knowledge, Applications, Limitations

Our understanding of microbial genomes has greatly advanced to a point where a bacterial species is no longer defined by its genome, but by its pan-genome [4,23,24]. Different strains of the same bacterial species have been sequenced to identify genotypic variations resulting in phenotypic traits that define those bacterial strains [4,23,24]. Intra-species bacterial genome sequencing and comparisons have identified genes unique to different strains that collectively form a bacterial species' genomic repertoire, this combined repertoire of bacterial genes from different bacterial strains represents a bacterial species' pan-genome [4,23,24]. Genes present in all bacterial strains within that bacterial species form its core genome, genes unique to two or more bacterial strains constitute its dispensable genome [4,23,24]. Bacterial genome sequencing and comparative genomics show the plasticity of microbial genomes is far greater then first envisioned with microbial gene pools and diversity being grossly underestimated [5,9,25]. The flexibility of bacterial genomes is underlined by three *Escherichia coli* strains (K12, O157:H7 and CFT073) having only 39.2% genes in common [26]. Compared to *E. coli* K12, *E. coli* O157:H7 was found to contain ~1,300 strain specific genes, many of which are associated with metabolic and virulence processes [26,27].

In addition to a revised definition of a bacterial species and diversity, microbial genomics is redefining our concept of microbial genome architecture [3,5,13]. Microbial genomes vary in size from 0.5 to 10 megabases with the majority of bacterial species containing a singular round chromosome [3,5,13]. Certain species such as *Borrelia burgdoferi* have linear chromosomes while pathogenic bacteria such as *Brucella* and *Vibrio* species contain two or more chromosomes [3,28]. Exogenous chromosomal genetic material carried by bacterial cells include plasmids that can vary both in size and number and encode factors such as antibiotic resistance genes [3,5,13]. In certain cases like in *Vibrio*, the second chromosome is postulated to be a mega-plasmid acquired by an ancestral *Vibrio* species causing the distinction between a second chromosome and a mega-plasmid to be blurred [3,28]. Similar to genome size, the guanine-cytosine percentage (%GC) content between microbial species also ranges widely from approximately 75% in various *Micrococcus* species to 25% in *Mycoplasma* species [3,5,13].

Genome analysis of a number of microbes has enabled the design of specific growth media for previously uncultivable bacteria and provided a basis for understanding resistance to anti-microbials by some bacterial species [3,10]. *In silico* metabolic pathway analysis suggested the inability of the causative agent of Whipple's disease, *Tropheryma whipplei,* to synthesis nine amino acids [10,29,30]. Incorporation of these nine amino acids to the growth media enabled *T. whipplei* to be axenically cultured for the very first time [10,29,30]. Similarly, resistance to fluoroquinolones by *T whipplei* was predicted by genome analysis which identified mutations

producing alanine instead of serine residues in the *parC* and *gyrA* genes which was subsequently confirmed experimentally [29,31]. Microbial genomic data is now being used in reverse vaccinology whereby the complete proteome of a micro-organism is screened for potential vaccine candidates opposed to the conventional empirical testing of antigens [10,32,33]. Potential vaccine targets using reverse vaccinology have been identified in species such as *Neisseria meningitidis* and *Streptococcus agalactiae* amongst others [10,32,33].

However, despite advancements in our understanding and leveraging of microbial genomes and the prodigious efforts of numerous research groups globally, functional knowledge for all genes within a given bacterial genome is far from complete. For example, only 66% of 4,460 genes from the laboratory work horse *E. coli* K-12, whose genome was sequenced in 1997, have experimental evidence for their function [34,35]. Compared to other sequenced genomes from all Kingdoms, *E. coli* K-12 represents the most complete genome annotation to date in both absolute and relative terms [35]. Clearly technological advances enable the rapid generation of genomic data at a greater rate then our ability to identify, characterise and decipher each individual component and how interactions between those components relate to an organism's phenotype.

In terms of genomic characterisation, Reed *et al* postulate four dimensions of genome annotation [36]. One-dimensional annotation comprises of genomic sequences demarcated with genes and their functions (where known) and is accomplished by gene finding algorithms, comparative genomics and controlled vocabularies (ontologies) [36]. Two-dimensional genome annotation utilizes literature mining and functional genomic studies to reconstruct metabolic and regulatory networks contained within a sequenced genome [35-37]. The spatial configurations of chromosomes and their genes during varying expression states within a cell constitutes the third-dimension of genome annotation [36]. Comparisons of bacterial genome sequences to decipher microbial genome diversity and architecture as well as identify evolutionary events that mould a genome represent the fourth-dimension of genome annotation [36].

Investigation of evolutionary events that shape genomic landscapes of bacterial pathogens will provide insights to the adaptations undergone by those pathogens that enable their successful infection and colonization of a host [5,6,11,13]. Such insights will facilitate a better understanding of virulence determinants encoded within a pathogen's genome and influence our strategies for combating bacterial pathogens using novel approaches, a fundamental aim of pathogen informatics.

Forces Shaping Microbial Genome Evolution

Gene content and genome size variation between microbial species is attributed to evolutionary processes driven by a bacterial species' interactions within widely fluctuating environments [13,38-40]. Gene loss and decay is observed when some micro-organisms shift from free living to being symbiotic or parasitic [13,38-40]. The host provides many of the metabolites and co-factors needed for growth and replication resulting in redundancy of some of the genetic machinery contained by the micro-organism [13,38-40]. Rather then expending energy maintaining and replicating the redundant genetic machinery, gene loss or inactivation is preferred resulting in reduced genome sizes or genomes with numerous pseudogenes [13,38-40]. The observed reduced genome sizes for endosymbionts such as *Buchnera aphidicola* and obligate intracellular pathogens like *Rickettsia prowazekii* are directly attributed to their niche adaptations [13,40,41]. Extensive genome decay is also observed in *Mycobacterium leprae* where only 49.5% of protein coding genes are functional and 27% of the genome is characterised as pseudogenes [16,42]. When compared to *Mycobacterium tuberculosis*, an estimated 2,000 genes have been lost in *M. leprae* since divergence from their last common mycobacterial ancestor [16,42]. Conversely, free living micro-organisms need to synthesis and scavenge essential metabolites and co-factors for their growth and replication in their fluctuating micro-environments resulting in large, diverse gene pools and consequently the maintenance of large genome sizes [13,38-40].

Just as gene loss may result in niche adaptation, gene gain enables colonization of new host species and altered pathogenesis by bacterial species [43]. Gene gain may occur via duplications of existing gene families or by the horizontal transfer of mobile genetic elements between bacterial species [5,13,43]. Horizontal (lateral) gene transfer is responsible for antibiotic resistance in previously susceptible bacteria and a shift from avirulence to virulence by some bacterial species within their host [5,13,43]. Genetic material transferred laterally between bacterial species include prophages, plasmids and larger mobile genetic elements such as pathogenicity islands [5,43]. Prophages are bacterial viruses whose genomes have been integrated into their bacterial host and can confer a virulence phenotype in some bacterial species like *Vibrio cholerae* or *Corynebacterium diphtheriae* where chromosomally integrated prophages CTXphi and corynephage beta encode for the cholera and diphtheria toxins respectively [44-46]. Integration of prophages into a micro-organism's genome propels sequence diversity between bacterial species and strains as prophages can constitute anywhere between 10-20% of a microbial genome [47].

Pathogenicity islands are larger horizontally transferred genomic elements which integrate into bacterial genomes and generally encode complete virulence systems such as type III secretion systems associated with pathogenic gram-negative bacteria like enterohaemorrhagic and enteropathogenic *E. coli* or the *Cag* pathogenicity island in *Helicobacter pylori* [5,43].

Increased virulence in Type I *H. pylori* strains is linked directly with the genomic integration of the *Cag* pathogenicity island and as such, marks an important evolutionary shift in the pathogenicity of *H. pylori* as avirulent Type II *H. pylori* do not contain the 40Kb *Cag* pathogenicity island [48,49]. Pathogenicity islands can be delineated from other mobile genetic elements as they have differing mean %GC contents in relation to the bacterial %GC genome mean and form clusters of adjoining genes which bestow distinctive virulent phenotypes in carrier bacterial strains compared to non-carrier bacterial strains [5,43,48,49].

Other forces driving microbial genome evolution include gene duplications, point mutations and recombination events. Gene duplication is essential for gene innovation and creation of novel functionality within an organism [50,51]. In bacteria, gene duplication mainly arises as an adaptable response to evolutionary selection pressures such as temperature fluctuations and nutrient unavailability [50,51]. Duplicated bacterial genes come under strong selective pressure and in most cases are expunged, in some cases they may undergo neo-functionalization to attain altered functions from their parent gene [50,51]. In *M. tuberculosis*, approximately 51% of genes contained within its genome are estimated to have arisen via gene duplication [16,52]. Point mutations can increase microbial diversity, promote survival and lead to an increased prevalence of some bacterial clones as in the case of *H. pylori* where point mutations within the 23SrRNA gene is responsible for resistance to the anti-microbial agent clarithromycin [53,54]. Recombination in certain bacteria like *Neisseria meningitidis* allows for the novel generation of antigenic variants for certain apparatus like pili which facilitates adhesion of *N. meningitidis* to host endothelial and epithelial cells [55]. Cell surface exposed pili that mediate host epithelial cell adhesion are subjected to host immune surveillance which exerts selection pressure for recombinant variants of pili genes like *PilE* from *N. meningitidis*, thus causing propagation of sequence diversity within a bacterial species [55].

Positive Selection: Theory and Implementation

The advent of complete genome sequence information facilitates empirical testing of orthologous aligned nucleotide sequences for elevated signs of sequence divergence based on Kimura's neutral evolutionary theory [56-58]. Generally, two closely related genomes will differ from their common ancestor due to evolutionary forces acting upon random mutations occurring in those genomes [56-58]. Most of these mutations occur due to random genetic drift which has little or no effect on an organism's fitness and as these changes do not alter any amino acids, they are classified as neutral [56-58]. A small subset of mutations that occur by random genetic drift may either be detrimental or beneficial to an organism within its given environmental niche. Mutations that prove detrimental to an organism's fitness come under strong purifying selection to be expunged from an organism's genome [56-58]. Mutations which confer a selective advantage to an organism within its environmental niche come under strong positive (diversifying) pressure to be fixed within an organism's genome [56-58].

Molecular evolutionary theory has been developed and constantly refined over the years using various methods and approaches to produce a rigorous, statistical framework for the detection of positive selection [58,59]. The common goal of all the various approaches involve determination of the parameter omega (ω) which is obtained by dividing the number of non-synonymous substitutions (amino acid altering; dN) by the number of synonymous substitutions (non amino acid altering; dS) [58-60]. As synonymous substitutions do not involve an amino acid change, their rate of substitution is considered to be neutral, non-synonymous substitutions are non-neutral as they involve an amino acid change and are considered to be a function of selective pressure acting on a gene product. Hence, $\omega = dN/dS$ and if $\omega > 1$ then positive selection is inferred, if $\omega = 1$ neutral evolution is inferred while $\omega < 1$ is indicative of purifying selection [58-61].

Methods that calculate ω from nucleotide sequence data can be classified as approximate (counting) and maximum likelihood methods [58,60,61]. Most approximate methods involve counting and calculating differences between synonymous and non-synonymous sites in an aligned nucleotide sequence making corrections for multiple substitutions at a site [58,60,61]. Many of the approximate methods use simplified mutational models of nucleotide substitutions with relatively few parameters that do not take into account unequal transition / transversion rates and differing organism specific nucleotide frequencies [58,61,62]. Depending on which nucleotide substitution model is employed, under-estimation and over estimation of synonymous and non-synonymous substitutions is possible as in the case of the JC69 model due to synonymous

transversions being less likely to occur at the third nucleotide position in a codon compared to transitions [58,61,62].

More biologically relevant models using codon substitutions were developed by Goldman and Yang incorporating a Markov process between codon substitutions within a Maximum Likelihood framework [61-63]. Codon based methods use the organism's genetic code to filter out stochastic noise and derive biologically realistic substitution models from the underlying sequence data enabling codon based models to fit sequence data better then nucleotide based models [61-63]. Codon based models also have more power to detect positive selection compared to approximate methods which in effect, averages ω over the whole length of a sequence alignment due to the assumption that all sites within an alignment are under equal selective pressure [60,63,64]. Usually, only a few amino acid sites within a protein might be under adaptive evolution as most sites are conserved to retain protein structure and function, codon based models take into account the paucity of sites under positive selection by allowing ω to vary amongst different codon sites thereby providing increased power in detecting positive selection [63-65]. The greater flexibility and accuracy of codon based models as well as their use of a rigorous statistical testing framework centred on Maximum Likelihood theory has made programs implementing such methods like PAML, the *de facto* standard when assaying for positive selection.

Maximum Likelihood is a statistical method which makes inferences about dataset parameters such as the transition / transversion rate and ω and from the underlying probability distribution of the dataset and assigns log-likelihood ratios to each codon model test used [58,64-67]. When assaying for positive selection using PAML, there are two models universally employed; M1a and M2a [15,18,58,61,65]. The M1a or nearly neutral model forms the null hypothesis within which near neutral evolution is assumed and permits variation of dN/dS (ω) amongst sites as opposed to the whole alignment [58,61,65,68]. Within the M1a model, ω is fixed with $\omega 1 = 1$ or $\omega 0 < 1$ for neutral evolution. The M2a or positive selection model enables for an additional proportion (p2) of sites where $\omega 2 > 1$ and is derived from the sequence data, hence allowing for a model of positive selection [58,61,65,68]. The log-likelihood ratio derived from the M1a model is compared to the log-likelihood ratio obtained from the M2a model. A log-likelihood ratio test (LRT) is preformed between M1a and M2a log-likelihood ratios and results are compared against a Chi-Square table of critical values using 2 degrees of freedom [58,61,65,68]. If the M1a model is rejected in favour of the M2a model, then positive selection can be inferred and *vice versa* [58,61,65,68].

Although methods for detecting positive selection have increased in sophistication, they still do have some drawbacks. Results obtained from any method assaying for positive selection are only as good as the input sequence alignment and even though a gene may be found to be under positive selection, no information as to the origin of the selection pressure is provided.

Bacterial Pathogens and Virulence

Multi-cellular organisms have been described as "a multitude of microbiomes" [69]. Varying degrees of relationships between microbial communities and their multi-cellular hosts exist ranging from mutualism, commensalism, parasitism to virulent [69]. Pathogenicity by a micro-organism is often associated with disease and morbidity occurring to a host caused by the presence as well as relative abundance of a micro-organism within a host [69-71]. However, in certain cases, disease and morbidity within a host may be caused by an absence of a microbial organism, hence the broad definition of a pathogen by Ehrlich *et al.* as a microbial organism "that results in a disruption of homeostasis" [69]. Historically, Robert Koch laid the basis for bacterial pathogens to be identified as the primary etiological factor for a disease only when four criteria can be met in what is termed Koch's postulates [5,72,73]. These criteria stipulate:

1) The bacterial organisms should be present in people with the disease and absent from people without the disease.
2) The bacterial organism should be isolated and maintained in a pure culture form.
3) The bacterial organism should be capable of causing an infection when re-inoculated into a host from its culture form.
4) The bacterial organism should be able to be re-isolated and cultured from its host after re-inoculation.

The advent of molecular biology resulted in a further refinement of Koch's postulates by Stanley Falkow to what are known as molecular Koch's postulates. Molecular Koch's postulates were developed to reflect our increased sophistication in isolating and manipulating bacterial genetic elements responsible for pathogenesis and attempt to provide a systematic framework for identification of a virulence factor [5,72,73]. These molecular Koch's postulates stipulate:

1) The proposed virulence determinant should be absent in non-pathogenic bacterial strains and only present in pathogenic bacterial strains.
2) Attenuation of a pathogenic bacterial strain should be observed with inactivation of the proposed virulence determinant / genes.
3) An attenuated bacterial strain should revert back to its wildtype pathogenic form upon re-introduction of the inactivated virulent determinant / genes.

Micro-organisms cause disease and morbidity within their host through a gamut of interactions mediated by virulence factors or determinants [70,71,74]. A virulence factor can describe bacterial products that directly cause detrimental damage to host cells e.g. toxins, or use of mechanisms for survival of the micro-organism within a host e.g. iron acquisition siderophores [70,71,74]. The term "virulence factor" is a biological concept and like many biological concepts, its use is context driven resulting in blurred definitional boundaries [71].

Attempts have been made to categorise known virulence factors into different classes based upon biological functions, however most of these groupings of bacterial virulence factors were determined in the pre-genomic era [6,12,19]. Many of the classically known virulence factors were derived using molecular biological approaches such as adhesion and colonization assays, *In vitro* Expression Technology (IVET) or transposon mutagenesis [6]. Such molecular techniques only allow certain classes of bacterial virulence determinants to be identified such as toxins or adhesive outer membrane proteins through quantification of what can be measured [6,70,74].

Application of high throughput novel genomic, transcriptomic and proteomic technologies has resulted in a further distortion of the definition of a bacterial virulence factor. Genes identified by homology, genes whose expression is up-regulated during specific infection periods, transcriptional regulators controlling the expression of virulence associated genes and post transcriptionally modified proteins are being added to the ever growing list of bacterial virulence factors [4,6,9,11,17]. With refinement and increased usage, high throughput technologies will no doubt contribute towards the debate in defining what constitutes a virulence factor and may possibly incorporate a systems based approach, whereby bacterial virulence will be defined on the basis of biological processes. Until such a time, the broad working definition of a virulence factor employed for the purpose of this study is; any endogenous bacterial gene product which aids in the adhesion, invasion, infection, colonization, persistence and pathogenesis of a microbial organism within its host.

Determination of Bacterial Genes Subject to Diversifying Selection

The plethora of bacterial species and strains sequenced coupled with the parallel development of molecular evolutionary theory facilitates the scanning of bacterial genomes for traces of genes undergoing rapid diversification [6,9,15,18,57,64]. There are number of reasons why one would scan closely related genomes to identify genes undergoing rapid evolution which include;

1) Discovery of nucleotide changes that may affect the fitness of an organism.
2) Identification of evolutionary changes which may provide insights into mechanisms contributing to the continued success of a pathogen.
3) Ascertain what biological processes are undergoing adaptation enabling a bacterial species to flourish within its environmental niche.
4) A better understanding of the dynamic interface between infection and immunity by identification of rapid evolutionary changes between bacterial strains.
5) Contribute to a better rationale in management and development of novel therapeutics to combat bacterial pathogens.

To date, several groups have focused on determining which genes are under positive selection within a single bacterial species and grouping these genes according to their known functional roles in bacterial virulence [15,18]. A limiting factor for many of these studies and indeed for this study, is the lack of functional information for the majority of genes within a bacterial genome. *E. coli* was used in previous studies partly due to the relatively high proportion of genes with functional annotation [15,18]. The approach undertaken in this study is to use a variety of microbial species opposed to a single bacterial species in order to:

1) Determine whether there are any universal classes of genes and common biological processes that maybe under positive selection across bacterial species.
2) Determine if there are any classes of genes and biological processes undergoing positive selection that are unique to each of the bacterial species under study.
3) Determine if there are any classes of genes under positive selection that have a substantial enrichment of known virulence genes characterised within those bacterial species.
4) Determine whether there are shared or unique biological processes between virulence genes within each bacterial species.

5) Determine whether there are shared or unique biological processes between genes under positive selection and virulence genes within each bacterial species.

The three bacterial species used in this study are *Helicobacter pylori*, *Neisseria meningitidis* and *Vibrio cholerae* with each bacterial species comprising of two strains summarised below in Table 1.1.

Bacterial Species (Strain)	**Accession**	**Genome Size (Mbp)**	**Nº of Genes**	**Morphology (Gram Stain)**	**Habitat**	**Pathogenesis**
H. pylori (26695)	NC_000915	1.66787	1630	Spiral, motile (Gram Negative)	Host associated mucosal layer of gastrointestinal tract	Gastric inflammation and peptic ulcer disease
H. pylori (J99)	NC_000921	1.6	1535	Spiral, motile (Gram Negative)	Host associated mucosal layer of gastrointestinal tract	Gastric inflammation and peptic ulcer disease
N. meningitidis (MC58)	NC_003112	2.3	2225	Cocci (Gram Negative)	Host associated – usually pharynx	Septicaemia, meningitis
N. meningitidis (Z2491)	NC_003116	2.2	2208	Cocci (Gram Negative)	Host associated – usually pharynx	Septicaemia, meningitis
V. cholerae (N16961)	NC_002505 NC_002506	4.03	4008	Curved bacillus (Gram Negative)	Aquatic	Cholera
V. cholerae (O395)	NC_009456 NC_009457	4.1	3998	Curved bacillus (Gram Negative)	Aquatic	Cholera

Table 1.1: GenBank genome accession numbers, genome size, number of genes, morphology, habitat and pathogenesis of bacterial strains used for determining classes of genes under positive selection.

H. pylori is estimated to colonize approximately half the world's population and is a leading global cause of peptic ulcers, stomach cancer and chronic active gastritis with infection durations ranging from a couple of years to decades [75-77]. *H. pylori* causes 5.5% of annual human cancer incidences and is one of the first bacterial pathogens to be recognized and classified as a Class I carcinogen by the World Health Organisation as it is a major risk factor in the development of gastric cancers [75-78]. *H. pylori* has no known natural reservoir outside of its

human host and because of a persistent chronic infection which if untreated results in a fatal outcome, *H. pylori* is considered an obligate pathogen. *H. pylori* is acquired orally from related family members and its origins have been traced to East Africa with its subsequent global spread linked to human population migrations out of Africa approximately 60,000 years ago [75,79]. Taxonomically, *H. pylori* belongs to the epsilon sub-division of proteobacteria and can further be broken down into gastric and non-gastric species [76,80]. Production of urease and high motility are characteristics of all known gastric *H. pylori* species with both factors considered essential for *H. pylori* survival and colonization within the hostile gastric environment [76,80]. Urease production helps neutralize the low pH of gastric acid and is vital for *H. pylori* motility towards more neutral pH gradients like the gastric mucosal layer within which the bacterium may replicate [76,80,81]. Approximately 6-7% of *H. pylori* genes are strain specific with 89 and 117 genes specific to *H. pylori* J99 and 26695 respectively, both *H. pylori* strains harbour the *Cag* pathogenicity island [14,77,82,83]. The pan-genome of *H. pylori* consists of the core genome calculated to comprise of 1,111 genes with an additional 400 genes making up the dispensable genome [78,84]. Variation between *H. pylori* strains operates primarily at the nucleotide level opposed to the amino acid level as many of the nucleotide differences are silent with respect to the protein product, indicating strong selective pressure to maintain functional conservation of the proteome [14,82,83].

N. meningitidis is usually part of the nasopharynx commensal flora and colonizes approximately 10% of the healthy human population, *N. meningitidis* is also an opportunistic pathogen which causes epidemic outbreaks of bacterial meningitis and septicaemia for reasons largely unknown [85-87]. Taxonomically, *N. meningitidis* belongs to the beta-proteobacterial subdivisions and different strains of *N. meningitidis* are divided into serogroups A, B, C, W135 and Y based on typing of their polysaccharide capsular structures [86,88,89]. *N. meningitidis* Z2491 belongs to serogroup A and is responsible for epidemic meningitis outbreaks within mainly developing countries while *N. meningitidis* MC58 belongs to serogroup B and is the etiological agent in sporadic meningitis outbreaks within developed countries that are characterized as being invasive [86]. The pan-genome of *N. meningitidis* is estimated to have 3,290 genes and is predicted to be open with the continual addition of 43 genes per a genome sequenced while the core genome is calculated to contain 1,337 genes at minimum [87,90].

V. cholerae, the causative agent of cholera, is a facultative pathogen mainly found in aquatic and estuarine environments associated with poor water quality and has been responsible for seven cholera pandemics [28,91,92]. *V. cholerae* is motile and belongs to the gamma-proteobacterial sub-division. Of the estimated 200 or so *V. cholerae* serotypes identified, *V. cholerae* serotype O1 is responsible for the seven cholera pandemics [28,91,92]. The O1 *V. cholerae*

serotype is further divided into two biotypes; the El Tor biotype which includes *V, cholerae* N16961 and the classical biotype which includes *V. cholerae* O395 [28,91,92]. The two O1 *V. cholerae* serotypes are differentiated based on biochemical properties with the El Tor biotype being Voges Proskaur reaction positive and haemolytic [28,91,92]. The El Tor biotype arose as the causative agent for the 7th cholera pandemic replacing the classical biotype responsible for the 6th cholera pandemic in prevalence [28,91,92]. *V. cholerae* N16961 was responsible for the 7th cholera pandemic in 1971 while strain O395 was isolated as the aetiological agent for the 6th cholera pandemic in 1965 [28,91,92]. There is conflicting evidence to the origin of the El Tor cholera strains with some studies showing both O1 classical and El Tor biotypes are part of a single epidemic clonal complex with intermediate strains between the two biotypes identified [93,94]. In contrast, some studies imply *V. cholerae* N16961 is more closely related to environmental isolates then to the *V. cholerae* O395 strain indicating that the 7th cholera pandemic originated independently from the 6th cholera pandemic [91,92]. The core genome of *V cholerae* is estimated to contain 2,787 genes while the dispensable genome comprises of 447 genes [95].

The three bacterial species comprising of two strains each (Table 1.1, page 16) were specifically chosen for this study due to their wide range of host-pathogen interactions and lifestyles which should exert different types of selection pressures upon their existing gene complement and the maturity of their sequenced genomes. *H. pylori* is an obligate pathogen with no known environmental reservoir other then humans and hence, is postulated to be undergoing host restriction [78]. *N. meningitidis* does not survive for long within the environment and is a commensal which in certain cases turns into an opportunistic pathogen within its host [90,96,97]. *V cholerae* is a free-living environmental bacterial species that is not part of the normal human microbial flora, but causes severe diarrheal disease when contracted and therefore is a facultative pathogen [6,71]. Although all three bacterial species have different lifestyles and interactions with a human host, they share a common element in that they are detrimental to their human host. Assaying for genes under positive selection within the three bacterial species should enable elucidation any common biological processes contributing to the pathogenesis of these micro-organisms and whether genes under positive selection within these bacterial species represent known virulence factors. Differences in gene classes under positive selection may provide insights into diversification undergone by these three bacterial species to facilitate their niche adaptations and possibly enhance their virulence potential which is reflected by their differing modes of pathogenicity.

Chapter 2 : Methods and Materials

Introduction

The following methods chapter is structured into twelve sections. Each section provides a description of the various aspects of data acquisition, handling, positive selection detection, functional annotation, metabolic pathway determination and statistical analysis conducted. The twelve sections presented in this method's chapter are:

1) Description of bacterial genome sequences used.
2) Pre-processing of bacterial GenBank genome sequence files for use in the pipeline approach employed to determine positive selection.
3) Description of the pipeline approach used to scan for positive selection which is divided into four subsections:
    a) Orthologue identification.
    b) Generation of orthologous in-frame codon alignments
    c) Assay for positive selection
    d) Annotation of bacterial genes under positive selection using GenBank submitted genome annotation files.
4) Curation of alignments for orthologous genes identified as being under positive selection.
5) Data mining of bacterial virulence genes.
6) Description of bacterial gene lists used in subsequent analyses (6 to 12).
7) Gene length and %GC content distribution analysis of the bacterial gene lists.
8) Gene Ontology annotation and enrichment analysis of the bacterial gene lists.
9) Clusters of Orthologous Proteins annotation and analysis of the bacterial gene lists.
10) Sub-cellular localization determination of protein products and analysis for the bacterial gene lists.
11) Metabolic pathway analysis of the bacterial gene lists.
12) Statistical methods and graphical packages used for data analyses and visualisation.

All the PERL programmes, UNIX commands, positive selection detection pipeline documentation, parsers used for data mining and R code for statistical analysis relevant to each of the above mentioned sections are presented in the appropriate Appendices referenced within the text.

Bacterial Sequences

RefSeq records for six bacterial strains belonging to three Proteobacterial species; *H. pylori* (strains 26695 and J99), *N. meningitidis* (strains Z2491 and MC58) and *V. cholerae* (strains N16961 and O395) together with their submitted annotation files were obtained from NCBI's bacterial genome repository via ftp in June 2007 [•]. Genome sequence information pertaining to the six bacterial strains are summarised below in Table 2.1.

Bacterial Strain	Accession	Genome Size (Mbp)	Nº of Genes	Sequencing Centre (Year Released)	PubMed ID
H. pylori 26695	NC_000915	1.66787	1630	TIGR (1997)	PMID : 9252185
H. pylori J99	NC_000921	1.6	1535	ASTRA (1999)	PMID : 9923682
N. meningitidis MC58	NC_003112	2.3	2225	TIGR (2000)	PMID : 10710307
N. meningitidis Z2491	NC_003116	2.2	2208	Sanger (2000)	PMID : 10761919
V. cholerae N16961	NC_002505 NC_002506	4.03	4008	TIGR (2000)	PMID : 10952301
V. cholerae O395	NC_009456 NC_009457	4.1	3998	TIGR (2007)	PMID : 19115014

Table 2.1: Bacterial strains, RefSeq accessions, sequencing centres, year of public data release and PubMed IDs of the primary genome project publications for the bacterial strains obtained from NCBI's ftp repository.

Pre-Processing of GenBank Bacterial Sequence Files

Nucleotide sequences of bacterial genes in Fasta format were attained using the European Molecular Biology Open Software Suite's (EMBOSS, Version 3.0.0) ExtractFeat utility [98]. In the instance of *V. cholerae*, extracted nucleotide gene sequences for chromosomes I and II were concatenated to form a single Fasta formatted file. Stream editor (sed) regular expressions were used to alter bacterial Fasta headers to remove special characters which interfere with downstream bioinformatics programmes, to contain less then 10 characters as required by the Phylip sequence format and facilitate accession parsing for subsequent analyses. Nucleotide gene sequences for each of the six bacterial strains (Table 2.1) were translated to their amino acid equivalents using GenBank's bacterial genetic code 11 by EMBOSS's TranSeq. Details of

[•] ftp://ftp.ncbi.nih.gov/genomes/Bacteria/

quality control checks to ensure fidelity of the extracted sequences after processing by EMBOSS tools and sed regular expressions are presented in Appendix A2 (Pipeline Documentation). Steps followed for the pre-processing of bacterial GenBank files are presented below in Figure 2.1.

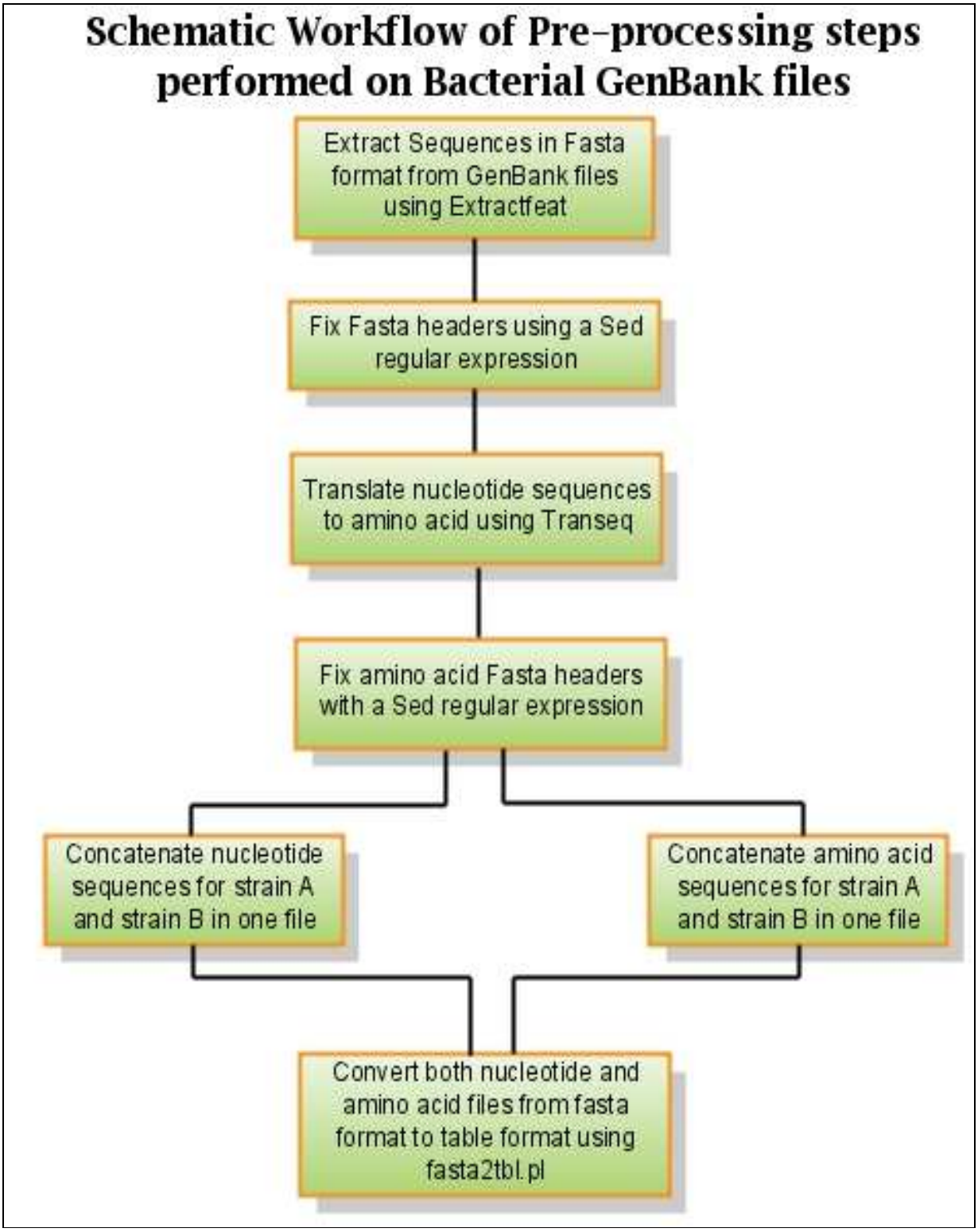


Figure 2.1: Schematic representation of the steps performed for the pre-processing of bacterial GenBank files before use in the positive selection detection pipeline.

Positive Selection Detection Pipeline

The pipeline approach used for detection of positive selection was originally conceived by Dr. Junaid Gamieldien. I have used the pipeline after making a number of changes that include pre-processing steps of bacterial GenBank files, incorporation of new bioinformatics tools, orthologous gene determination, better methods and models for detecting positive selection and a curation step to remove false positives. The pipeline comprises of PERL scripts (Appendix A1) that utilise established bioinformatics software modules in a sequential manner (Appendix A2). A workflow of the positive selection detection pipeline is presented overleaf in Figure 2.2.

Orthologue Identification

Translated amino acid sequences were used for putative orthologue identification as the amino acid alphabet consists of more characters (20 amino acids) then the nucleotide alphabet (4 deoxyribonucleic acids), thereby providing greater sensitivity [99-102]. Processed amino acid sequence files for each bacterial strain were formatted to create BLAST databases (see Appendix A2 for details) in separate directories. The Fasta_split.pl PERL script, originally written by Dr. Alan Christoffels, was used to split a multi-sequence Fasta formatted file into individual sequence Fasta files for bacterial strains A and B (Figure 2.2, page 23, Appendix A1). The individual amino acid sequence files form the query for BLASTP searches. BLASTP (Version 2.2.6) searches of the query against the BLAST databases were done by the ortho_BLAST.pl PERL script (Figure 2.2, page 23, Appendix A1). Parameters used for the BLASTP searches are $\geq$ 50 % identity and coverage, an expectation score of 1e-10 and a BLOSUM 62 matrix. Orthologous genes between bacterial strains A and B were identified using the Reciprocal Best Hits (RBH) method [99-102]. Strain A is used as the BLASTP query against strain B, generating a list of "best BLASTP hits [99-102]. The procedure is repeated with strain B forming the query against strain A, also generating a list of "best BLASTP hits" [99-102]. The intersection between the two lists of best BLASTP hits forms the reciprocal best BLASTP hit (RBH) of a given gene between strains A and B, allowing putative orthology to be inferred [99-102]. The PERL script, reciprocal_gene_ID.pl, was used to parse paired accession lists of top BLASTP hits for strain A against strain B and *vice versa* to obtain an intersection list of paired orthologous gene accessions between bacterial strains A and B (Figure 2.2, page 23, Appendix A1).

A major assumption of all positive selection detection methodologies is the input gene sequences are orthologous in nature and have arisen through vertical descent as opposed to duplication, horizontal gene transfer or recombination [58,103,104]. Therefore, identification of orthologous genes between bacterial strains is vital for any positive selection analysis because

paralogous genes arising from gene duplications are under different selection pressures from the parent gene and will provide erroneous results [51,104].

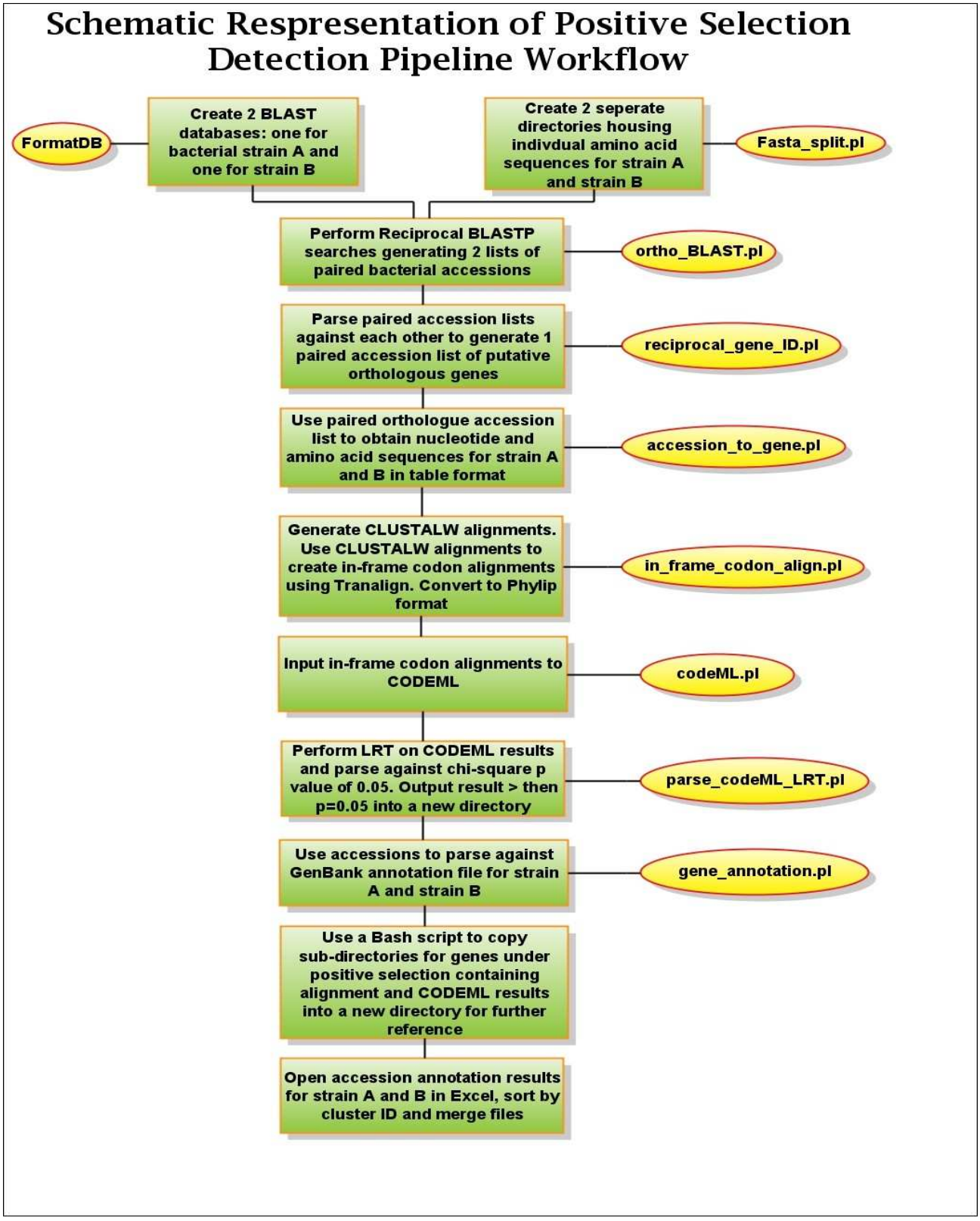


Figure 2.2: Graphical summary of workflow used for determination of positive selection. Green boxes represent each sequential pipeline step performed and yellow eclipses represent PERL scripts used for each sequential pipeline step.

Generation of In – Frame Codon Aligned Sequences

Paired accession gene lists obtained from RBH were used by the accession_to_gene.pl PERL script to extract orthologous amino acid and nucleotide sequences from the concatenated amino acid and nucleotide files created during the pre-processing of bacterial GenBank files (Figure 2.1, page 21, Appendix A1 and A2). Each identified orthologous gene pair between bacterial strains A and B will have two files, one containing amino acid sequences and the other nucleotide sequences in Table format (Figure 2.2, page 23). These two types of orthologous sequence files are used to generate in-frame codon alignments (Figure 2.2, page 23).

The PERL script In_frame_codon_align.pl uses the Table formatted orthologous amino acid and counterpart nucleotide sequence files as input to ClustalW (Version 1.83) creating global alignments for each sequence file type (Appendix A1) [105]. A ClustalW aligned amino acid file and an aligned nucleotide file are written out which can be viewed for quality control checking with any sequence viewing software such as BioEdit. Both the ClustalW aligned nucleotide and amino acid file are concatenated into a single file, from which in-frame codon alignments are created using EMBOSS's TranAlign utility (Figure 2.2, page 23, Appendix A1). The output is an in-frame, codon aligned, orthologous nucleotide gene pair which is converted to Phylip format for use by the CODEML module from the PAML package (Figure 2.2, page 23, Appendices A1 and A2) [58,61,106].

Assay for Positive Selection

The PERL script codeML.pl is used to call CODEML from the PAML package (Version 3.15) to screen each orthologous in-frame, codon aligned gene pair using a pre-specified control file (Appendix A2) designating the parameters and models to use by CODEML (Figure 2.2, page 23, Appendix A1). The codeML.pl PERL script writes out the CODEML results into a pre-specified file within the subdirectories those files are read from. On average, it takes 3 minutes for CODEML to calculate log likelihoods for the M1a and M2a specified models of neutral and positive selection. For *H. pylori*, *N. meningitidis* and *V. cholerae* which have 1,382, 1,707 and 3,431 putative orthologous gene pairs identified, it takes CODEML approximately 69.1, 85.4 and 171.6 hours respectively to generate results on a 2 x 2.4 GHz Intel Pentium 4 Xeon server with 4 GB of RAM.

CODEML results are parsed using the parse_codeML_LRT.pl PERL (Appendix A1) script which performs a Likelihood Ratio Test (LRT) between the log-likelihood results for the M1a and M2a tests [58,61,67,106]. The formula used for the LRT is presented below:

$$\mathbf{LRT = 2*(lnL1 - lnL2)}$$

lnL1 = Log likelihood ratio; M1a model **lnL2** = Log likelihood ratio; M2a model

Results of the LRT are compared to a Chi-Square critical value of 5.99 with two degrees of freedom which corresponds to P = 0.05, genes that reject the null hypothesis (M1a) in favour of the selection hypothesis (M2a) are written out into a new, user specified directory.

Annotation of Bacterial Genes under Positive Selection

Bacterial gene accessions of genes surpassing the LRT at P = 0.05 are used by the gene_annotation.pl PERL script to parse against their respective GenBank bacterial annotation files to create a file containing the cluster ID, gene accession and annotation for genes under positive selection in bacterial strain A (Figure 2.2, page 23). The procedure is repeated using bacterial strain B's annotation file to obtain annotations for genes under positive selection in bacterial strain B (Figure 2.2, page 23, Appendix A1). The two resulting files are opened in MicroSoft (MS) Excel, sorted by cluster ID number and merged to form a single MS Excel file containing the cluster ID number for tracking / reference, orthologous paired gene accessions for bacterial genes under positive selection between strains A and B and their respective annotations. The MS Excel file forms an easy, quick reference framework for further studies (Figure 2.2, page 23).

Curation of Alignments for Genes under Positive Selection

Putative orthologue identification is traditionally done using amino acid sequences as the amino acid alphabet contains more characters then the nucleotide alphabet and the latter is degenerate [101,102]. However, use of translated amino acid sequences for identification of putative orthologous genes, especially from closely related bacterial species, can introduce artefacts caused by sequencing errors resulting in a shift of the translational reading frame (frameshift errors), leading to false positives. Hence, a curation step of aligned orthologous sequences is required. Some groups generate orthologous nucleotide alignments together with amino acid alignments and screen for possible errors before running the sequences through a pipeline approach, others make no mention of any curation steps [15,18]. In this particular instance, all orthologous gene pairs identified as being under positive selection were checked using BLASTN nucleotide sequence alignments. The reasoning behind curating alignments after running the sequences through the pipeline is that one is only interested in how many gene alignments produce false positives, not how many putative orthologous gene pairs between the bacterial genomes have frameshift errors. Thus, one only examines the smaller sample space of genes identified as being under positive selection that is a subset of all orthologous gene pairs to gauge whether they are false positives caused by artefactual data. ClustalW alignments are not suitable for quality control checking of frameshifts or sequencing errors at the nucleotide level because it

is a global alignment programme that attempts to align the whole length of two orthologous sequences against each other, inevitably introducing gaps [105]. BLAST being a local alignment programme will only align regions between genes that have a maximal score [99]. Insertions or deletions within a gene in respect to its orthologous counterpart produces gaps which are multiples of 3 nucleotides as opposed to gaps of 1, 2, 4 and 5 nucleotides that result in a shift of the translational reading frame [18]. A combination of BLASTN nucleotide alignments and ClustalW amino acid alignments can be used to determine whether gaps of 1, 2, 4, and 5 nucleotides causes a shift in the translational reading frame within an amino acid alignment.

The following protocol involving manual curation was devised to examine orthologous BLASTN nucleotide alignments cross-referenced against their counterpart ClustalW amino acid alignments of all genes found to be under positive selection, and flag any potential false positives for exclusion from subsequent analyses:

1) Extract all Fasta nucleotide coding sequences for genes under positive selection from bacterial strain A.
2) Split the extracted nucleotide sequences from bacterial strain A into individual files.
3) Create a BLAST database of bacterial strain B's genome.
4) BLASTN the extracted nucleotide sequences for strain A against strain B.
5) Obtain the RBH for each nucleotide sequence of strain A against strain B.
6) Check each of strain A's nucleotide alignments against strain B's genome for presence of gaps within the alignment.
7) Count the number of gaps present in the nucleotide alignment, determine if the gaps are consecutive and note the results in an MS Excel workbook.
8) Check the ClustalW amino acid alignments of the corresponding nucleotide alignments to determine if the gaps are consistent with differences in protein / gene lengths.
9) Check the ClustalW amino acid alignments to determine the if corresponding position of the gap in the nucleotide alignment causes a change in reading frame by producing different amino acids at that nucleotide gap position within the amino acid alignment.
10) Exclude all genes found to be under positive selection for which a frameshift can be detected for *H. pylori, N. meningitidis* and *V. cholerae*.

Manual examination of aligned nucleotide sequences cross-referenced against their amino acid alignments takes about 5 days each for *H. pylori* and *N. meningitidis* and approximately 1 and half days for *V. cholerae*. The procedure can be automated by parsing the BLASTN results for the number of gaps contained in the alignment, however in this case manual curation of the alignments was preferred [18].

Virulence Gene Data Mining

The Microbial Virulence Database (MVirDB) and Virulence Factor Database (VFDB) containing virulence genes identified from database and literature sources were mined for *H. pylori*, *N. meningitidis* and *V. cholerae* known virulence genes [19,107]. MVirDB contains data from eight different publicly available databases as off November 2006 which are summarised below in Table 2.2 [107].

Data Base	Description
Tox-Prot	Subset of SwissProt database housing known protein toxins
SCORPION	Database of scorpion toxins
PRINTS	Database of virulence gene motifs
VFDB	Database of virulence genes compiled via literature mining
TVFac	Toxin and virulence factor database
Islander	Genomic islands database
ARGO	Database of antibiotic resistance genes
VIDA	Database of animal virus open reading frames

Table 2.2: Databases and a brief description of each database used by MVirDB to populate its "data warehouse" of known virulence genes.

MVirDB is not a curated resource at the sequence level as sequence entries obtained from the eight different databases (Table 2.2) are not merged resulting in a lot of sequence redundancy. MVirDB has multiple nucleotide and amino acid sequences of the same genes containing different accessions inherited from their parent databases and numerous motifs present within their virulence gene datasets. To deal with the high level of redundancy and numerous accessioning formats contained within MVirDB and obtain lists of unambiguous, unique virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae*, reciprocal BLASTs of bacterial genome genes against MVirDB organism specific datasets were conducted.

VFDB is highly curated of redundant entries and employs the same universal accession system adopted for use in this study [19]. VFDB datasets of known virulence genes identified by literature mining for *H. pylori*, *N. meningitidis* and *V. cholerae* were downloaded and converted to text files. The text files were parsed by means of a PERL script using bacterial gene accession IDs for *H. pylori*, *N. meningitidis* and *V. cholerae*.

Virulence gene lists for *H. pylori*, *N. meningitidis* and *V. cholerae* obtained from MVirDB and VFDB were consolidated and duplicate entries between both databases were removed to provide a single, combined list of unique, unambiguous virulence gene accessions for each of the three bacterial organisms.

Bacterial Gene Lists

There are three types of gene lists used for subsequent analyses which are constantly referred to within this study;

1) Orthologous gene pairs found to be under positive selection (genes under positive selection)
2) Known virulence genes obtained from data mining of MVirDB and VFDB (database identified virulence genes)
3) All the protein coding genes within a bacterial genome (background / genomic gene list).

Each gene list type comprises of the original, unique bacterial gene accessions submitted by their respective sequencing groups. Great care was taken to ensure bacterial genes run through the positive selection detection pipeline maintained their unique bacterial gene accessions submitted by the sequencing groups for unambiguous gene identification. Maintaining the original bacterial gene accessions facilitates data-mining due to the universal usage of these accessions by mostly all databases and enables accession parsing to obtain representative gene counts of the different gene lists for use in ensuing statistical analyses.

Bacterial Gene Features

Gene lengths and %GC content were obtained for genes under positive selection, database identified virulence genes and genomic genes to determine if differences between distributions in %GC content and gene length exist. The three different types of bacterial gene accession lists were used as input into the pickseq.pl* PERL programme to extract the relevant nucleotide sequences for each gene list type for *H. pylori*, *N. meningitidis* and *V. cholerae*.

The length of genes and %GC content of the nucleotide sequences extracted for bacterial genes under positive selection, database identified virulence genes and genomic genes belonging to *H. pylori*, *N. meningitidis* and *V. cholerae* were obtained using EMBOSS's InfoSeq utility [98]. InfoSeq was run on each of the three nucleotide sequence files for each bacterial organism resulting in a list of accessions with gene lengths and %GC content. The different gene list types were imported as text files into the R statistical environment where plots of the distributions of gene lengths and %GC content were created and statistical testing of the distributions between the different gene list types were performed [108]. In the case of the genomic gene lists, tRNA and rRNA features of gene length and %GC content were excluded so as not to bias the background distribution towards shorter gene lengths and varying %GC content.

* http://dendrome.ucdavis.edu/resources/scripts/

Gene Ontology Annotation of Bacterial Gene Lists

Functional annotation using Gene Ontology (GO) for genes under positive selection and database identified virulence genes were performed using the Database of Annotation, Visualisation and Integrated Discovery (DAVID, Version 6) [109-111]. DAVID is an integrative resource that provides amongst other annotations, GO mappings for genes belonging to different eukaryotic and prokaryotic organisms [111]. DAVID also executes a statistical enrichment analysis of GO terms mapped to an uploaded gene list with respect to the background distributions of GO terms mapped to the whole genome for the organism under study [111]. The goal was to determine what functional gene classes and biological processes are over-represented in the positively selected and database identified virulence gene lists with respect to their genomic GO distributions. GO annotations consist of three categories; biological process, cellular component and molecular function, each comprising of a five-tiered hierarchy with broad biological terms placed on top of the hierarchy and more specific biological terms placed at the bottom of the hierarchy [109-111].

Genes under positive selection and database identified virulence gene lists for *H. pylori*, *N. meningitidis* and *V. cholerae* were uploaded to DAVID's gene list manager to obtain functionally enriched GO terms with respect to each bacterial organism. To maintain consistency across all comparisons, gene list annotation and functional enrichment of GO terms for all gene lists was conducted at the 4th level of the GO hierarchy with an EASE score of 0.5 using DAVID [109-111]. The 4th level of the GO hierarchy was chosen for all gene list GO annotation and functional enrichment analyses after evaluation of GO annotations at all five different levels in an effort to maximise gene list coverage and obtain GO terms that are not too broad in their description as to be uninformative [109-111]. Four types of GO term functional enrichment comparisons were conducted:

1) Functional enrichment of GO terms for genes under positive selection with respect to the micro-organism's genomic GO distribution.
2) Functional enrichment of GO terms for database identified genes with respect to the micro-organism's genomic GO distribution.
3) Functional enrichment of GO terms for genes under positive selection with respect to database identified virulence genes for each micro-organism.
4) Manual comparison of GO terms between database identified virulence genes and genes under positive selection as DAVID does not allow cross-species GO term comparisons.

Clusters of Orthologous Proteins Functional Annotation

Clusters of orthologous proteins (COGs) was devised to facilitate rapid annotation and characterisation of newly sequenced genomes using functional annotations transferred by orthology of genes [102,112]. COGs comprises of twenty-five alphabetical codes, each one describing a particular high-level biological process [102,112]. COG annotations for the genome, genes under positive selection and database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae* were obtained from their respective annotation files from NCBI's ftp repository. The rationale was to determine if there are any biological processes common between genes under positive selection and database identified virulence genes and if these biological processes differ significantly from the genomic distribution of COG biological processes.

COG annotations for genes under positive selection, database identified virulence genes and genomic genes were parsed from their respective bacterial annotation files using a PERL script (COG_parser.pl-Appendix A3A) based on accession matching and sorted according to their alphabetical codes. Counts of genes for each gene list type belonging to each alphabetical COG category were obtained within MS Excel for further statistical analysis.

Sub-Cellular Localisation Determination

Sub-cellular localization of protein products for genes under positive selection, database identified virulence genes and genomic genes were obtained from the cPSORTb database (version 2.0) [113-116]. The cPSORTb database contains predictions for sub-cellular locations of protein products computed by six independent bioinformatics modules, each analysing a specific biological feature of an amino acid sequence for various signatures [113-116]. Each analytical module produces a score for a protein belonging to a certain sub-cellular site which is incorporated into a Bayesian probabilistic framework to produce a final score as to the likelihood that a protein localises to a particular sub-cellular site [113-116]. The aggregate scoring scheme ranges from 1 to 10 with a score of 7.5 indicative of very good evidence for the predicted sub-cellular localization site [113-116]. Depending on the type of bacterial gram stain, there are between five to six sub-cellular localization categories [113-116]. Sub-cellular localisation categories for gram negative bacteria like *H. pylori*, *N. meningitidis* and *V. cholerae* are Cytoplasmic, Cytoplasmic Membrane, Periplasmic, Outer Membrane, Extra Cellular and Unknown [113-116].

The cPSORTdb was accessed via the select organism page and display options set to contain bacterial gene accessions, predicted sub-cellular localization and the PSORT sub-cellular localization score fields for genomic genes belonging to *H. pylori*, *N. meningitidis* and

V. cholerae. A tab-delimited file of the following results was downloaded and a PERL parser was written (PSORT_parser.pl-Appendix A3B) which parses gene list accessions against gene accessions contained within the downloaded PSORT files of the respective bacterial organisms. The results are files containing PSORT scores and sub-cellular localizations for genes under positive selection and database identified virulence genes, both being a subset of the original genome wide PSORT annotation file. UNIX "grep" commands were used to obtain counts for genes belonging to each of the PSORT predicted sub-cellular categories for each gene list type for use in statistical analyses.

Bacterial Metabolic Pathways

H. pylori, *N. meningitidis* and *V. cholerae* accessions for genes under positive selection and database identified virulence genes were used to mine organism specific BioCyc databases (*H. pylori* and *V. cholerae* Database Version 13.5, *N. meningitidis* Database Version 13.1) [117,118]. The BioCyc database version for *N. meningitidis* differs from *H. pylori* and *V. cholerae* as the former is classified a Tier 3 database that has not undergone any form of curation while the latter two are Tier 2 databases whose contents have undergone some curation [117,118]. The Pathway Tools Omics browser was used to upload gene lists accessions and tabular results of BioCyc metabolic pathways for *H. pylori*, *N. meningitidis* and *V. cholerae* gene lists were obtained by selecting a pathway interaction threshold of 1 to provide as broad a coverage of gene lists as possible [117-119]. All metabolic pathways for *H. pylori*, *N. meningitidis* and *V. cholerae* gene lists in tabular format were downloaded and individual pathways of interest for each bacterial organism were further investigated by following BioCyc's hyperlinks to pathway information pages [117-119].

Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways for *H. pylori*, *N. meningitidis* and *V. cholerae* genes under positive selection and database identified virulence genes were obtained by uploading respective gene list accessions to DAVID's gene list manager and selecting KEGG pathway annotations [111,120,121]. DAVID was chosen for KEGG pathway annotation of bacterial gene lists due to its ability to provide pathway annotations for whole gene lists, statistical calculation of gene enrichment for a KEGG pathway and the option of using a foreground gene list to query a another gene list set as the background to obtain shared KEGG pathways between gene lists types. Parameters selected for individual gene list KEGG annotations were a pathway threshold of 1 and an EASE score of 0.5 to maintain consistency when comparing results across bacterial species. When using genes under positive selection as a foreground list and database identified virulence genes as a background gene list to determine

KEGG pathways shared by both gene list types, an EASE score of 1 was selected to increase KEGG pathway annotation coverage by DAVID.

Statistical Analyses and Graphics

All statistical analyses and summaries of datasets was conducted using the R statistical package (version 2.8.0) [108]. As most of the datasets involve gene counts and are therefore categorical (nominal / ordinal) data, non-parametric Chi-Square tests with Yate's continuity correction was mainly used for statistical testing [122].

Two types of Chi-Square tests were conducted, a one-way classification Chi-Square test to determine if any functional annotation categories have a statistically significant association at the P = 0.05 mark for either genes under positive selection or database identified virulence genes (Appendix A4) [122]. The second type of Chi-Square test was a two-way classification to determine if genes under positive selection and database identified virulence genes have a statistically significant association at the P = 0.05 level for a particular functional annotation category (Appendix A4) [122]. R functions were written to conduct both types of Chi-Square tests and where gene counts for a particular category were below 5, a Fisher's Exact test was used (Appendix A4) [122]. In the case of comparing gene lengths and %GC content distributions, non-parametric Kolmogorov-Smirnov tests were conducted within R.

All results from statistical testing were corrected for multiple hypothesis testing using the Holm's method as implemented within R [108,122]. Statistical enrichment analysis of GO terms and KEGG pathways for gene lists was done by DAVID which uses a Fisher's Exact test and utilises a variety of methods like Bonferroni to correct for multiple hypothesis testing [111].

Graphs and images were plotted using R's graphical capabilities and the Lattice Package in R [108,123]. Bacterial chromosomal wheels were constructed using the Microbial Genome Viewer online application [124]. Cellular overviews of metabolic pathways for *H. pylori*, *N. meningitidis* and *V. cholerae* gene lists were generated using the Pathway Tools and Omics platform which is part of the BioCyc collection of tools [117,119]. KEGG pathway maps containing genes from bacterial gene lists were obtained from DAVID [111,125].

Chapter 3 : Characterisation of Positive Selection and Bacterial Virulence Genes

Introduction

The following chapter provides results and discussions for a variety of analyses conducted to obtain gene counts for orthologous gene pairs, genes under positive selection and database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae*. The number of orthologous gene pairs indentified for *H. pylori*, *N. meningitidis* and *V. cholerae* are examined in relation to gene content of the different bacterial species. Counts of genes under positive selection after manual curation of orthologous sequence alignments, as well as an appraisal of how robust the results for genes under positive selection are by continuing to reject the null hypothesis at differing P value thresholds is explored. A review of published literature was undertaken to determine what factors may contribute towards nucleotide diversity within *H. pylori*, *N. meningitidis* and *V. cholerae* in order to understand differences in gene numbers observed to be under positive selection within each of the bacterial species. Recombination and its potential effects upon the results of a positive selection assay in general and with regards to the positive selection results obtained in this study is discussed.

Details of virulence gene data mining and gene counts obtained from each virulence database together with results of statistical tests to evaluate if database identified virulence genes are enriched for genes under positive selection is presented. The possibility of a database bias caused by differing numbers of virulence gene records contained within each virulence database and the effect it may have upon the statistical results when testing for an association between databases identified virulence genes and genes under positive selection is investigated.

Physical properties of genes under positive selection and database identified virulence genes were inspected to establish whether these gene lists follow similar distributions with respect to each other and the genome. The physical properties *H. pylori*, *N. meningitidis* and *V. cholerae* gene lists appraised are gene lengths and %GC content. Differences in distributions of physical properties for each gene list type were tested for within each individual bacterial organism and cross species comparisons of their distribution characteristics were conducted.

Identification of Bacterial Orthologous Genes

Putative orthologous genes between *H. pylori* (strains 26695 and J99), *N. meningitidis* (strains Z2491 and MC58) and *V. cholerae* (strains N16961 and O395) were identified by the reciprocal best hit (RBH) BLASTP method [101,102,126]. If the top BLASTP hit of a gene in genome A is the top BLASTP hit in genome B and *vice versa*, putative orthology between gene pairs was inferred [101,126,127]. The RBH BLASTP method enables one to one orthology between gene sets from bacterial strains to be established as opposed to one to many homologous relationships that arise by conducting a one way BLASTP search [101,126,128]. Counts of putative orthologues identified by the RBH BLASTP are summarized below in Table 3.1.

Bacterial Species	Bacterial Strain	N^{o} of Genes in Genome	N^{o} of Putative Orthologues
H. pylori	26695	1630	1382
	J99	1535	
N. meningitidis	MC58	2225	1707
	Z2491	2208	
V. cholerae	N16961	4008	3431
	O395	3998	

Table 3.1 : Number of genes annotated within each bacterial genome and the number of putative orthologues identified by RBH BLASTP searches.

The number of orthologous gene pairs identified mirrors the bacterial gene content of the bacterial organisms (Table 3.1). For instance, the difference between orthologous gene counts for *H. pylori* and *V. cholerae* is 2,049 genes, similar to the difference of 2,378 genes between *H. pylori* 26695 and *V. cholerae* N16961 (Table 3.1). Similarly, *N. meningitidis* and *V. cholerae* orthologous gene counts differ by 1,724 genes, comparable to the gene content difference of 1,783 genes between *N. meningitidis* MC58 and *V. cholerae* N16961 (Table 3.1). Differences in gene content between the three bacterial species (Table 3.1) may possibly reflect the various lifestyles adopted by those bacterial species. Host dependent, obligate pathogens such as *H. pylori* are known to have smaller gene contents compared to free living, environmental bacterial organisms, like *V. cholerae* [13,38,39].

Intra-species differences in gene content may arise from gene duplications, expansions, decay or inactivation as well as an element of error introduced by gene finding algorithms and the constant refinement of genome annotations for each of the bacterial strains [36,129,130].

Orthologous Bacterial Genes subject to Positive Selection

In-frame, codon aligned orthologous gene pairs for *H. pylori*, *N. meningitidis* and *V. cholerae* were assayed for positive selection using CODEML from the PAML package [58,106]. CODEML utilises a rigorous statistical framework derived from Maximum Likelihood theory and codon based models of evolution [58,66]. Maximum Likelihood methods simultaneously calculate likelihoods for the occurrence of neutral evolution and positive selection based on explicit codon substitution models that are parameterised from the underlying, input sequence data [58,61,65,66]. Codon based models of evolution whose parameters are derived from the underlying data have increased power for detecting positive selection as the dN / dS ratio is not calculated as an average over the length of a sequence alignment [58,61,63]. Greater power by Maximum Likelihood methods is achieved by enabling ω (ratio of dN/dS) to vary amongst sites (codons) as opposed to the entire length of a sequence alignment [58,61,65,106]. The null hypothesis, (M1a) fixes ω to 1 while the positive selection model, (M2a) estimates ω as a free parameter from the underlying data [58,61,65,106]. Results from the two models, (M1a and M2a), are subjected to a likelihood ratio test (LRT) and compared to a Chi-Square distribution, enabling either rejection the positive selection hypothesis in favour of the null hypothesis of no selection or *vice versa* [58,61,65,106]. Maximum Likelihood methods have been meticulously tested using simulated and real data and found to be consistent in their output, leading to their widespread use in many positive selection analyses [15,18,131,132]. However, like most computational methods, results of positive selection assays are only as good as the quality of the aligned input sequences.

Orthologous gene pair alignments initially found to be under positive selection for *H. pylori*, *N. meningitidis* and *V. cholerae* were manually curated to ensure the fidelity of the positively selected gene datasets by removal of false positives which may introduce an element of error upon further analyses. Manual curation of orthologous alignments was conducted by examining nucleotide alignments for gaps of 1, 2, 4 and 5 basepairs (bp) [18]. The position of the gaps was recorded and cross checked against their counter-part amino acid alignments to determine if a shift in reading frame has occurred from the start of a gap [18]. An example of an orthologous gene pair found to be under positive selection due to a gap in the nucleotide alignment resulting in a frameshift translation of the amino acid sequence is presented overleaf in Figures 3.1 and 3.2.

```
>H_pylori_26695_genome
          Length = 1667867

 Score =  414 bits (209), Expect = e-117
 Identities = 279/301 (92%), Gaps = 1/301 (0%)
 Strand = Plus / Plus


Query: 1       atgaaaaaggttgtttttttattattattcatgctaggggttagaagcgcacagtgat 60
                |||||||||||||||||||||| ||| | || |||||||||||||||||||| |||   |
Sbjct: 207932  atgaaaaaggttgtttttttattgttagttatactaggggttagaagcgcaaagtact 207991


Query: 61      tattgtagtgatcattgtgaaggcacgccagatagccgtatccctcctatggggtttcat 120
                ||||| |||||||||| ||||||||||||||||||||||||||||||||||||||||||
Sbjct: 207992  tattgcagtgatcattgcgaaggcacgccagatagccgtatccctcctatggggtttcat 208051


Query: 121     ttcagttttgtgcattcagtgaaatattacttgcaagacccacaaaagcgtgatcacaag 180
                |||||||||||||||||||||||||||||||||| || ||| |||| |||||||||
Sbjct: 208052  ttcagttttgtgcattcagtgaaatattacttgcaagatccgcaagagcgcgatcacaag 208111


Query: 181     cttgaaaaatgccacaaagcctttgactcgacgcttaaggttaattttatcacgaagtct 240
                |||||||||||||||  |||||||||| ||||| |||||||||||||||| ||||| |||
Sbjct: 208112  cttgaaaaatgccatcaagcctttgattcgactcttaaggttaattttattacgaa-tct 208170


Query: 241     tttaaaaaggattgcaagcatgcgcaaatggctttagagcaagctcaaaaagaaactcca 300
                |||||||||||||||||||||||||||||||||||||||||||||| |||||||  ||||||
Sbjct: 208171  tttaaaaaggattgcaagcatgcgcaaatggctttagagcaagcccaaaaagggactcca 208230


Query: 301     t 301
                |
Sbjct: 208231  t 208231
```

Position 237

Figure 3.1 : A single gap in the BLASTN nucleotide alignment of *H. pylori* J99 jhp0189 against the genome of *H. pylori* 26695. A 1 bp gap at nucleotide position 237 was recorded and cross checked with its counterpart amino acid alignment (Figure 3.2).

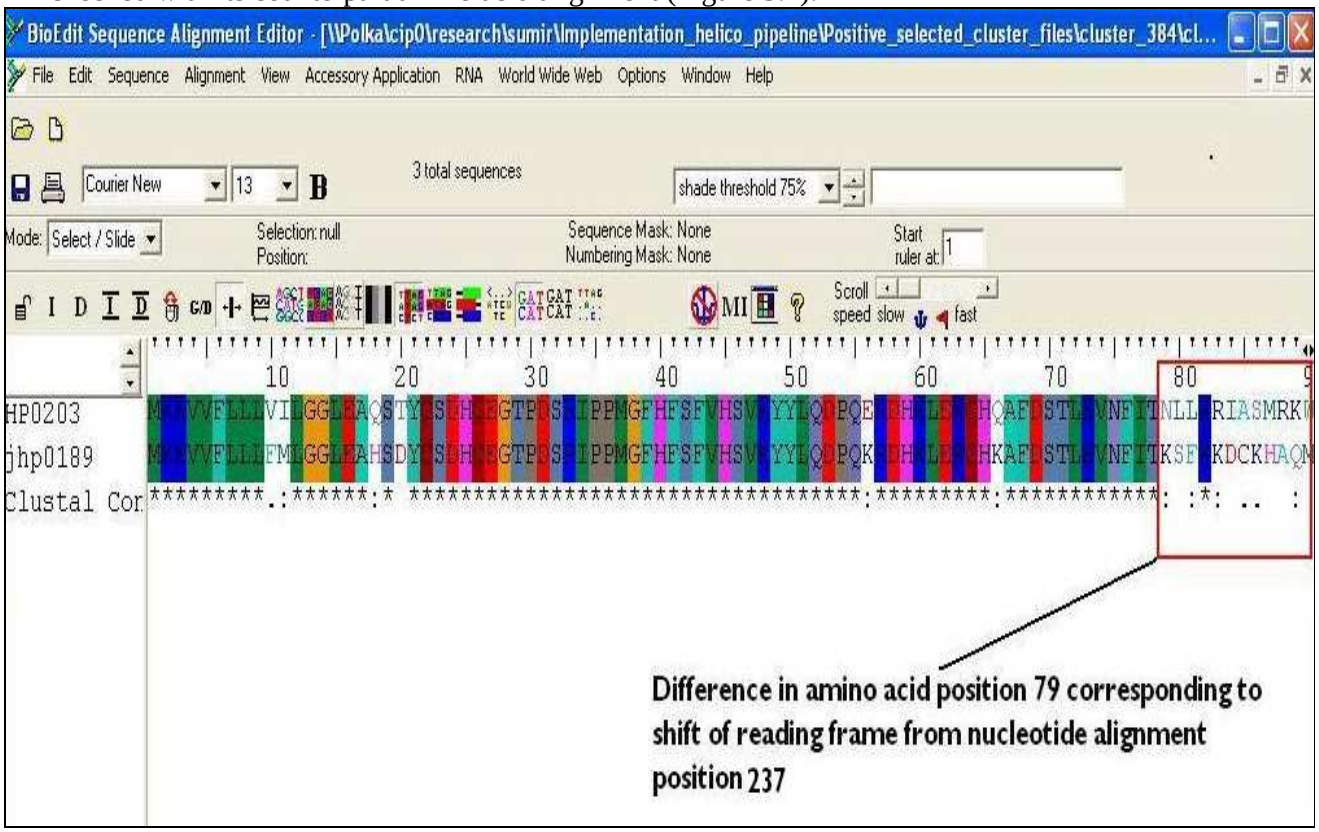


Figure 3.2 : Examination of the counterpart ClustalW amino acid alignment within BioEdit indicates a change of amino acid consensus alignment at position 79 corresponding to a shift of the translational reading frame from nucleotide position 237 (237/3 = 79th amino acid).

The number of orthologous gene pairs found to be under positive selection before and after manual curation are summarised below in Table 3.2. A full list of orthologous genes pairs found to be under positive selection after manual curation for *H. pylori*, *N. meningitidis* and *V. cholerae* with their annotations obtained from their submitted GenBank annotation files can be found in Appendix B1.

Bacterial Species	**Bacterial Strain**	**N^{o} of Putative Orthologues**	**N^{o} of Orthologous gene pairs under Positive Selection**	
			Before Manual Curation	**After Manual Curation**
H. pylori	26695	1382	311	230
	J99			
N. meningitidis	MC58	1707	259	218
	Z2491			
V. cholerae	N16961	3431	79	23
	O395			

Table 3.2 : Orthologous gene pair numbers found to be under positive selection following a LRT test using a Chi-Square critical value of P = 0.05 before and after manual curation of alignment datasets for *H. pylori*, *N. meningitidis* and *V. cholerae*.

A characteristic reduction in gene numbers is observed from the number of orthologous gene pairs identified, to the number of genes under positive selection after manual curation (Table 3.2). The manual curation approach employed is an *in silico* method and as such, is conservative. Changes in reading frames caused by single base pair insertions or deletions within a gene in respect to its orthologous counterpart may be a *bona fide* biological events that has occurred within the lineage of one of the bacterial strains. The only way to be absolutely certain that single base pair insertions or deletions are artefactual or *bona fide* biological occurrences is to re-sequence the genetic loci in question (Figures 3.1 and 3.2, page 36) [82,129].

Robustness of Positive Selection Results

LRT scores of genes under positive selection after manual curation (Table 3.2, page 37) were parsed under increasingly stringent P values to determine how many positively selected genes continue to reject the null hypothesis of no selection. The number of orthologous gene pairs found to be under selection at varying P-values is presented below in Figure 3.3.

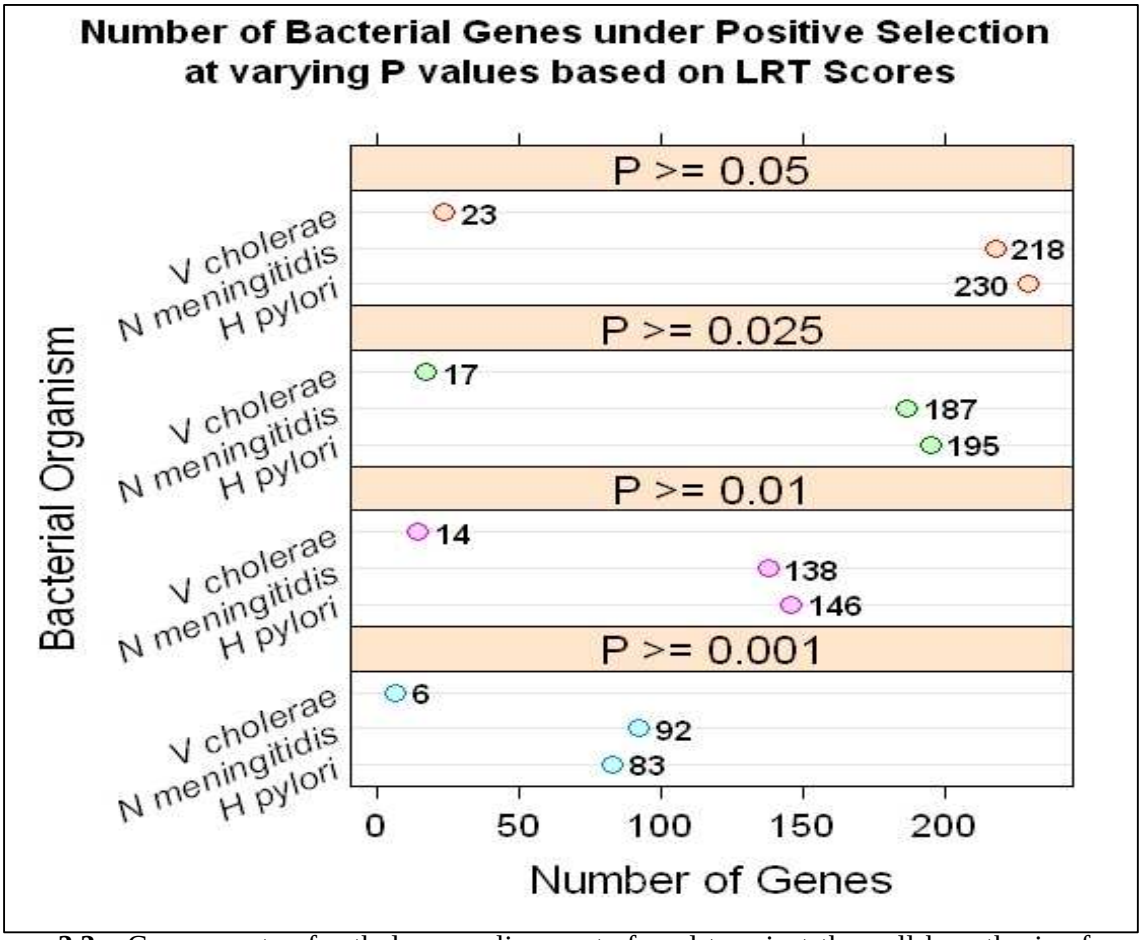


Figure 3.3 : Gene counts of orthologous alignments found to reject the null hypothesis of no selection under increasingly stringent P-values for *H. pylori*, *N. meningitidis* and *V. cholerae*.

The number of orthologous gene pairs which continue to reject the null hypothesis of no selection decreases with increasingly stringent P-value thresholds (Figure 3.3). The ability of an orthologous gene pair to reject the null hypothesis of no selection at increasingly stringent P values is not a reflection of the strength of selective pressure occurring on that particular gene. Instead, orthologous gene pairs found to be under positive selection at increasingly stringent P values are indicative of how closely differences between an orthologous gene sequence pair follows a particular model distribution, thereby making the alternative hypothesis harder to reject. As the accepted standard for reporting statistical significance in biomedical literature is placed at the P = 0.05 level, orthologous gene pairs that equal to or surpass the P = 0.05 threshold after manual curation were retained for subsequent analysis [122,133].

Differences in Positive Selection Gene Numbers between Bacterial Species

The number of genes under positive selection appears to follow an inverse trend in comparison to gene content and number of orthologues identified within *H. pylori*, *N. meningitidis* and *V. cholerae* (Table 3.2, page 37). *H. pylori* has the least number of genes and orthologues identified, yet has more genes under positive selection in comparison to *N. meningitidis* and *V. cholerae* (Table 3.2 page 37). If all forces driving genome diversification for *H. pylori*, *N. meningitidis* and *V. cholerae* were equal, one would expect the number of genes under positive selection to increase with gene content and genome size, as observed with the number of orthologous gene pairs identified (Table 3.1, page 34).

The proportionally large number of genes under positive selection suggests high levels of nucleotide diversity between *H. pylori* and *N. meningitidis* strains, in comparison to *V. cholerae* strains (Table 3.2, page 37). To determine why an inverse relationship between genome size and genes under positive selection is observed for *H. pylori*, *N. meningitidis* and *V. cholerae*, factors contributing to nucleotide diversity within the bacterial species are examined in detail.

H. pylori Nucleotide Diversity

The high level of nucleotide diversity observed between the two *H. pylori* strains is attributable to a high mutation rate and frequent recombination [134-141]. *H. pylori* is described as a panmictic species, whose frequency of recombination events exceeds those of other bacterial species such *N. meningitidis* or *E. coli* [136,140,142]. High allelic diversity generated by recombination events in *H. pylori* suggests colonization of a host occurs by mixed populations of *H. pylori* strains [78,137,139,140]. Paradoxically, clonal lineages of *H. pylori* exhibiting little variation in genome diversity have been observed over durations as great as nine years within individuals [78,139,143,144]. Clonal population structures of *H. pylori* are also observed in different ethnic groups partitioned geographically, as well as within related family units [145,146]. These observations imply that although recombination events are frequent in *H. pylori*, they operate over extended evolutionary timescales and require an initial infection by a mixed population of *H. pylori* strains [137,139,146].

Although recombination is a major driver of genome diversity in *H. pylori*, it does not adequately explain the large proportion of genes under positive selection within *H. pylori* for a number of reasons. Firstly, recombination between *H. pylori* strains would require a multiple strains of *H. pylori* to be present [78,136,137,139,140,146]. *H. pylori* 26695 and *H. pylori* J99 were isolated from different patients at different geographical regions with little information to

suggest a mixed or single strain *H. pylori* infection, and therefore any possible contact or recombination occurring between *H. pylori* 26695 and J99 strains within those two patients [14,77]. Secondly, analysis of core genes show *H. pylori* 26695 and J99 can be partitioned into separate geographical population groups, the former strain placed in an hpEurope group (common to Europe) and the latter strain into an hpAfrica1 group (common to Africa) [75,84]. Thirdly, chronic *H. pylori* infections resolve to form a single, stable clonal population over time after an initial mixed strain infection [137,139,143,146]. Fourthly, *H. pylori* strain 26695 and J99 would have been clonal in nature as they were repeatedly sub-cultured before being sequenced [14,77].

The frequency of mutations undergone in *H. pylori* is much greater in comparison to other bacterial species such as *E. coli* [78,135,136,141]. The high mutational rate observed within *H. pylori* is attributed to an absence of defined homologues for well characterised DNA repair pathways such as base excision repair (BER) or the mismatch repair (MMR) systems [78,141]. Higher mutational rates within *H. pylori* compared to certain *E. coli* mutator strains has led to growing speculation *H. pylori* implements *de novo* mutational tactics as an adaptable response to changing environments [53,78,135,141]. Though the majority of mutations that occur in *H. pylori* are synonymous substitutions, *H. pylori* does have a higher rate of divergence at non-synonymous sites when compared to other bacterial species like *E. coli* [14,83,140,141]. Another contributing factor towards the large nucleotide diversity observed within *H. pylori* is a lack of any known genetic bottlenecks as there is no known, naturally occurring reservoir for *H. pylori* [39,78,139,141,147,148]. The vertical mode of *H. pylori* transmission would produce a reduced effective population size whose fitness is not governed by their ability to scavenge nutrients within a free-living environment, therefore decreasing the intensity and efficacy of purifying selection to expunge mutations [39,78,139,141,147,149]. Hence, *H. pylori*'s vertical mode of transmission from family members to offspring may result in a possible absence of purifying selection [139,145,146].

Interestingly, variability between *H. pylori* strains that colonize different compartments of the stomach has been observed [14,77,134,139,146,147]. As *H. pylori* J99 and 26695 were isolated from different patients displaying different pathologies, it is conceivable the two *H. pylori* strains under study have undergone adaptations to facilitate colonization of their respective gastric niches [14,77,134,139,146,147]. The high adaptability and plasticity of *H. pylori*'s genome, coupled with its sometimes clonal population structure as indicated by chronic infections in differing gastric compartments indicates the proportionally large number of genes under positive selection between *H. pylori* 26695 and J99 is to be expected.

***N. meningitidis* Nucleotide Diversity**

Similar to *H. pylori*, *N. meningitidis* genome fluidity is generated by a mixture of recombination and mutation which needs to be counter-balanced by genome stabilizing mechanisms, to ensure survival and prevent de-speciation [96,150-153]. *N. meningitidis* posses DNA repair pathways, although the complexity and efficacy of these DNA repair pathways are much reduced in comparison to *E. coli* with genes like *mutS* and *mutL* undergoing repeated mutational and recombinational deterioration [96,151,153].

Neisserial species are characterised as naturally competent throughout their life-cycle with homologous recombination occurring by exogenous DNA recognition of a ten basepair DNA uptake sequence (DUS, (5'-GCCGTCTGAA-3')), transportation / uptake of the exogenous DNA via the type IV pilus apparatus and RecA mediated recombination [96,154-156]. While intra-species recombination is frequent, transformation studies show gene flow between neisserial and non-genus related bacterial species is rare, with so far a single recorded event [87,96,153,157,158]. Recombination is usually regarded as a source for generating genetic diversity, a less well reported aspect is recombination's homogenizing effect on microbial populations like *N. meningitidis* [96,153,159]. Neisserial DUSs are concentrated in loci responsible for DNA fidelity such mis-match repair genes, leading to a growing hypothesis that recombination events within *N. meningitidis* are mainly involved with genome repair and stability through their homogenizing effect [96,153,154,160]. Promotion of genome stability by DUSs opposed to genome diversification caused by their recombinational effect, has recently been verified with lower densities of DUSs found within regions of *N. meningitidis*'s dispensable genome opposed to its core genome, thus preserving genome homogeneity [160].

Although *N. meningitidis* is a naturally competent bacterium with high rates of recombination, most disease causing strains termed as "hyper-invasive" are clonal as determined by multi-locus sequence typing (MLST) [96,150-153]. Many of these hyper-invasive meningococcal strains are robust over time and geographical ranges with their clonal structure intact, indicating they are immune to diversification caused by recombination [90,96,153,161]. Reduced efficacy of DNA fidelity maintaining pathways within neisserial species may provide a mechanism whereby numerous, genotypic variants are produced, with the most successful variants resulting in host colonization [90,96,153,161]. Consistent with this hypothesis are the small regulatory networks found in *N. meningitidis* suggesting that meningococcal adaptation is punctuated by mutational and not regulated responses[96,153].

In addition to an elevated global mutational rate, mutational hotspots within *N. meningitidis*' genome that drive variability exist [90,96,97,153]. Localised hypermutational rates are

caused by short repeat sequences such as micro-satellites which are prevalent within the *N. meningitidis* genome, resulting in a phenomenon known as phase variation [90,96,97,153]. Phase variation is an altered state of gene expression at the transcriptional and / or translational level causing either "on" or "off" gene expression and is caused by tandem repeat sequence expansion / contraction within a gene's promoter region or coding sequence [90,96,153]. The expansion / contraction of repeat sequences within *N. meningitidis* is largely caused by slipped strand mispairing, which on the synthesis DNA strand results in expansions while deletions occur by mispairing on the DNA template strand during DNA replication and repair of these repeat sequences [90,96,153]. Hypermutable loci which result in phase variation modulate expression of outer membrane proteins, adhesions and pilli and are dubbed as "contingency loci" [96,153,162]. An advantage of these hypermutable loci is the creation of numerous, phenotypic variants allowing for selection of advantageous phenotypes in varying environments and enhanced transmissibility as many hyper-invasive meningococcal strains have elevated rates of phase variation [90,96,153,161].

***V. cholerae* Nucleotide Diversity**

Vibrio choleare strains O395 and N16961 are closely related and belong to the same O1 serogroup, the former strain is categorized as a classical biotype, while the latter strain is classified as an El Tor biotype [91,93,94,163]. Both the N16961 and O395 *V. cholerae* genomes display remarkable conservation in terms of gene content differing by approximately 1% (between 36 and 77 genes), although this number is likely to be higher due to the asymmetric usage of N16961 micro-arrays [17,163]. In terms of genome stability, from 126 bacterial genomes surveyed, *V. cholerae* ranks 26th where as *N. meningitidis* ranks 121st (*H. pylori* not included in the study), indicating *V. cholerae* gene content and order is highly conserved [164]. DNA repair, replication and fidelity mechanisms within *V. cholerae* are intact with well defined homologues present which may contribute to a low level of nucleotide diversity [28].

The small number of genes found to be under positive selection within *V. cholerae* is consistent with comparative nucleotide sequence analysis of the *mdh* housekeeping gene locus within 23 strains belonging to the O1 serotype, including N16961 and O395, whereby only nine polymorphic sites were identified, eight being synonymous substitutions [93]. A comparative genomics study of *V. cholerae* N16961 and O395 also supports the low level of nucleotide diversity observed by identifying a total of 45 polymorphic sites between the lineages leading to N16961 and O395, and a further 519 occurring along the branches leading to *V. cholerae* O395 and N16961 [91]. Both sets of authors suggest sequence conservation amongst sites between O1 *V. cholerae* biotypes implies a recent evolutionary divergence for N16961 and O395 strains, with Feng *et al* dating the time of divergence between these two biotypes at approximately 1880 [91,93].

As both *V. cholerae* N16961 and O395 strains occupy similar environmental niches and geographical locations, the recent time of divergence would suggest insufficient time and separation between the two strains has not occurred for an accumulation of adaptive mutations to be fixed. Additionally, the episodic waves of cholera pandemics suggest that in between such pandemics, *V. cholerae* residing within its natural aquatic reservoir would be subject to strong purifying selection in order to expunge any mutations arising during the founder flush period [93]. Hence, the recent time of divergence, intact DNA repair and replication mechanisms and strong purifying selection within their aquatic environments may all contribute to the small number of genes observed to be under positive selection between *V. cholerae* N16961 and O395.

Recombination and Positive Selection

In terms of phylogenetic studies and assaying for positive selection, recombination between organisms may produce erroneous results due to the assumption of vertical transmission of genetic material being violated [103,131,165]. When assaying for positive selection using gene pairs, the effect of recombination is less severe in producing false positives compared to using multiple sequences from more then two bacterial strains♦. Within a multiple sequence alignment, both phylogenetic tree topology and branch lengths are assumed to be constant along the length of the alignment during positive selection analyses [103,131,165]. In the presence of recombination, both phylogenetic tree topology and branch lengths will differ amongst sites in the multiple sequence alignment, thereby providing erroneous results due to different sites having different phylogenetic histories [103,131,165].

In contrast, pairwise sequence comparisons consist of a single branched phylogenetic tree whose topology and branch length remains constant along the sequence alignment and therefore is unaffected by recombination events. A disadvantage of using pairwise sequence comparisons though is reduced power for detection of positive selection as more sequences from more bacterial strains provide more data increasing the power to detect positive selection [58,60,66]. Efforts were made to conduct a recombination assay for the three bacterial organisms under study, but all recombination software require at minimum four to five aligned sequences in order to test for recombination breakpoints [103,165]. At the time this study was conducted, there were only two bacterial strains whose genomes had been sequenced for each of the species and placed in the public domain. The pipeline approach utilized in this study can easily be scaled to incorporate those new genomes with the addition of a recombinational assay.

♦ Personal Communication; Dr. Cathal Seoighe

Bacterial Virulence Genes and Positive Selection

Bacterial virulence genes for *H. pylori, N. meningitidis* and *V. cholerae* were mined from the Virulence Factor Database and Microbial Virulence Database (VFDB and MVirDB respectively) [19,107]. MVirDB and VFDB aim to provide compilations of known bacterial virulence genes collated from numerous scattered literature and database sources [19,107]. Data obtained from MVirDB was curated for sequence redundancy to obtain a list of unambiguous, unique virulence genes for *H. pylori, N. meningitidis* and *V. cholerae*. VFDB differs from MVirDB in that it is highly curated, thereby providing non-redundant lists of virulence genes and their accessions [19,107]. Lists of virulence genes obtained from MVirDB and VFDB were merged and duplicate entries between the two virulence gene lists were removed to prevent double counting during statistical and functional analyses. The result is a single list of known virulence genes for each of the bacterial organisms under study (Appendix B2). The number of unique and common genes obtained from MVirDB and VFDB is presented below in Table 3.3.

Bacterial Organism	***H. pylori***		***N. meningitidis***		***V. cholerae***	
Virulence Database	**MVirDB**	**VFDB**	**MVirDB**	**VFDB**	**MVirDB**	**VFDB**
Nº of Virulence Genes	138	87	37	65	33	165
Nº of Virulence Genes in Common	72		34		29	
Nº of Database Unique Virulence Genes	66	15	3	31	4	136
Nº of Unique Virulence Genes in Merged Gene List	153		68		169	

Table 3.3: Gene counts of database identified virulence genes after curation for redundancy from MVirDB and VFDB. Gene counts include the total number of virulence genes found in each database, number of virulence genes common to and unique to each database as well as the number of merged, unique virulence genes.

The number of virulence gene records within MVirDB and VFDB for *H. pylori, N. meningitidis* and *V. cholerae* differ, VFDB contains more unique virulence genes for *V. cholerae* and *N. meningitidis* while MVirDB has more unique virulence genes for *H. pylori* compared to VFDB (Table 3.3). Differences in database records between VFDB and MVirDB may occur due to differing definitions of the term "virulence factor" and data acquisition methods used to

construct both the databases [19,71,107]. MVirDB uses a broad data mining strategy based on keyword searches of other databases, VFDB utilizes a focalized curation approach based on primary literature and categorization of identified virulence genes into classes [19,107]. In terms of database bias, MVirDB contain more unique virulence genes for *H. pylori* while VFDB has more unique records for *N. meningitidis* and *V. cholerae* (Table 3.3, page 44). The intersections between positively selected genes and database identified virulence genes from VFDB and MVirDB for *H. pylori*, *N. meningitidis* and *V. cholerae* are presented as a series of Venn diagrams overleaf in Figure 3.4.

To determine if there is a statistically significant enrichment of virulence genes undergoing positive selection, one way classification Chi-Square tests were conducted and the results are presented below in Table 3.4, the full output of the Chi-Square tests are contained within Appendix B3.

Bacterial Species	**N^{o} of Genes under Positive Selection**	**N^{o} of Virulence Genes**	**N^{o} of Virulence Genes under Positive Selection**	**P - value**
H. pylori	230	153	22	0.9192
N. meningitidis	218	68	14	0.005074
V. cholerae	23	169	1	1

Table 3.4 : Results of Chi-Square tests conducted to determine if there is a statistically significant intersection between database identified virulence genes and genes undergoing positive selection within *H. pylori*, *N. meningitidis* and *V. cholerae*.

For *H. pylori* and *V. cholerae* there is no statistically significant enrichment of known virulence genes undergoing positive selection at the P = 0.05 mark (Table 3.4). *N. meningitidis* does have a statistically significant association between known virulence genes and genes undergoing positive selection at the P = 0.05 threshold (Table 3.4). Genes under positive selection, database identified virulence genes and virulence genes under positive selection for *H. pylori*, *N. meningitidis* and *V cholerae* have been mapped to the genomes of their respective bacterial species and are presented as chromosomal wheels in Figures 3.5, 3.6 and 3.7 on pages 47, 48 and 49 respectively.

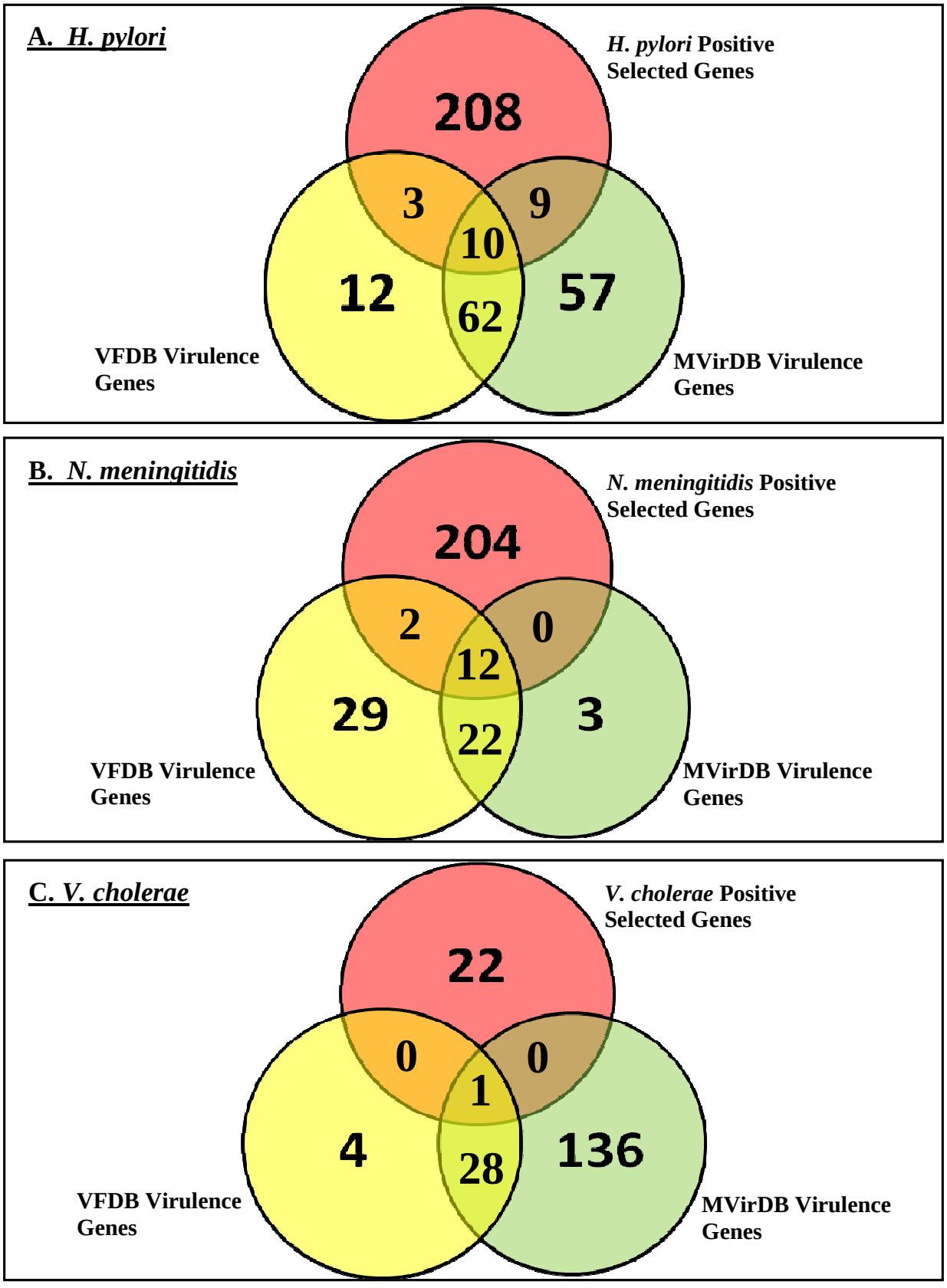


Figure 3.4 : Gene counts for the intersections between positively selected genes and unique virulence genes mined from MVirDB and VFDB for : A. *H. pylori* ; B. *N. meningitidis* ; C. *V. cholerae*.

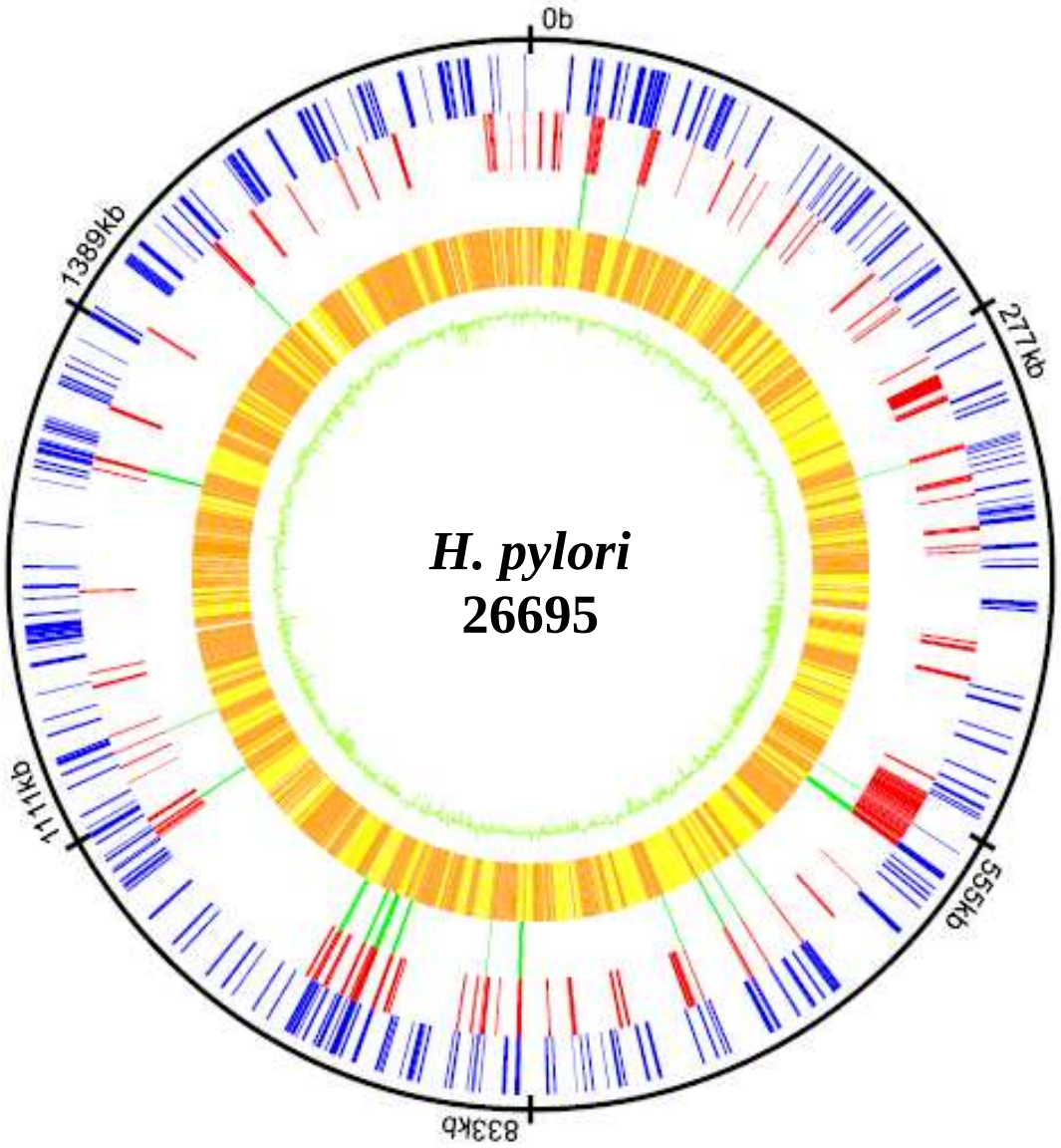
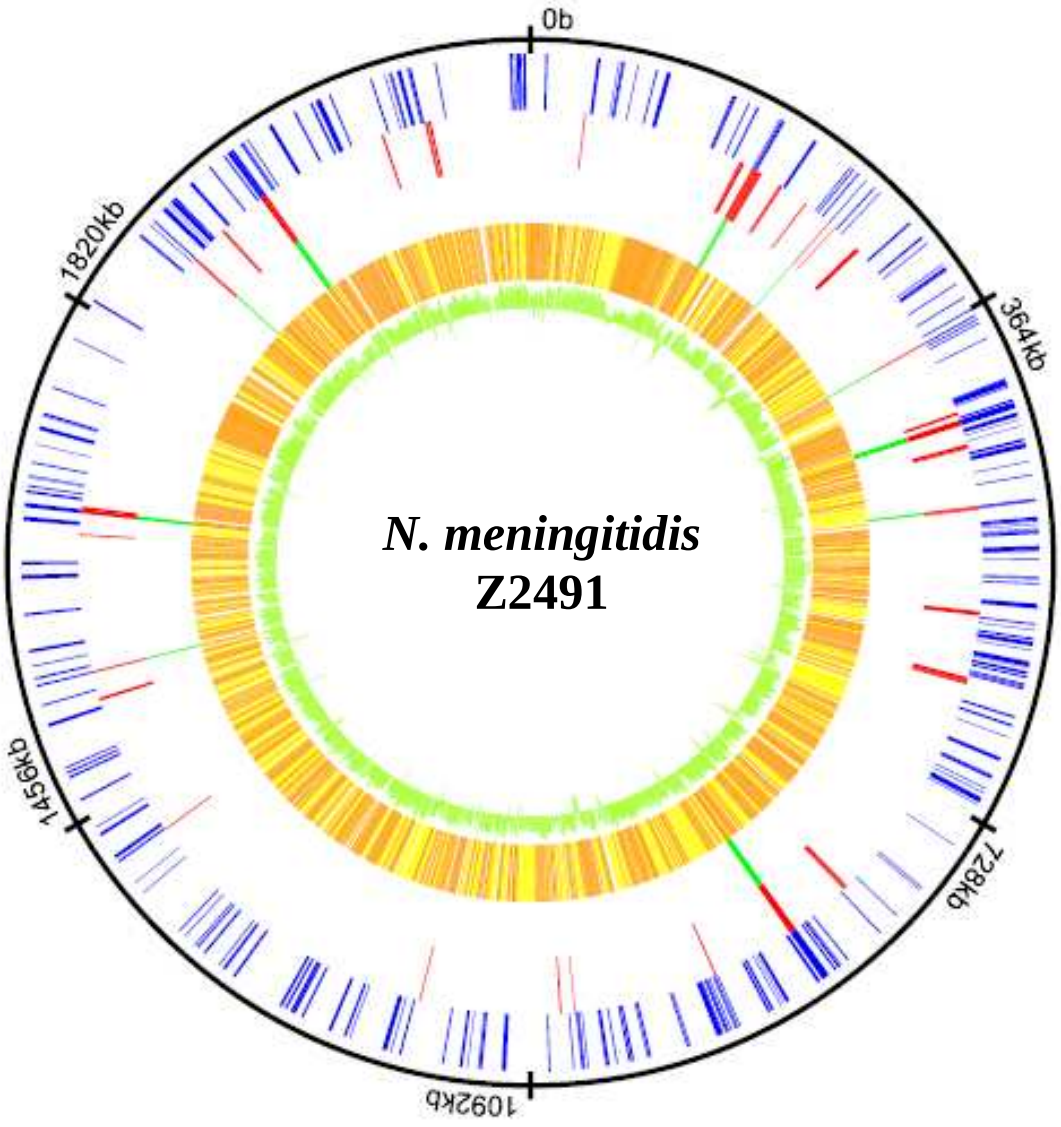


Figure 3.5 : Chromosomal wheel of *H. pylori* 26695 genome displaying the location of genes under positive selection between *H. pylori* 26695 and J99, database identified virulence genes, virulence genes under positive selection above genes contained on the plus and minus strand and their %GC content.



Key

- Genes under Positive Selection
- Database identified Virulence Genes
- Virulence genes under Positive Selection
- *N. meningitidis* Z2491genes on the plus strand
- *N. meningitidis* Z2491 genes on the minus strand
- % GC content

Figure 3.6 : Chromosomal wheel depicting the location of genes found to be under positive selection between *N. meningitidis* Z2491 and MC58, database identified virulence genes, virulence genes under positive selection, genes on the plus and minus strand as well as %GC content for *N. meningitidis* Z2491's genome.

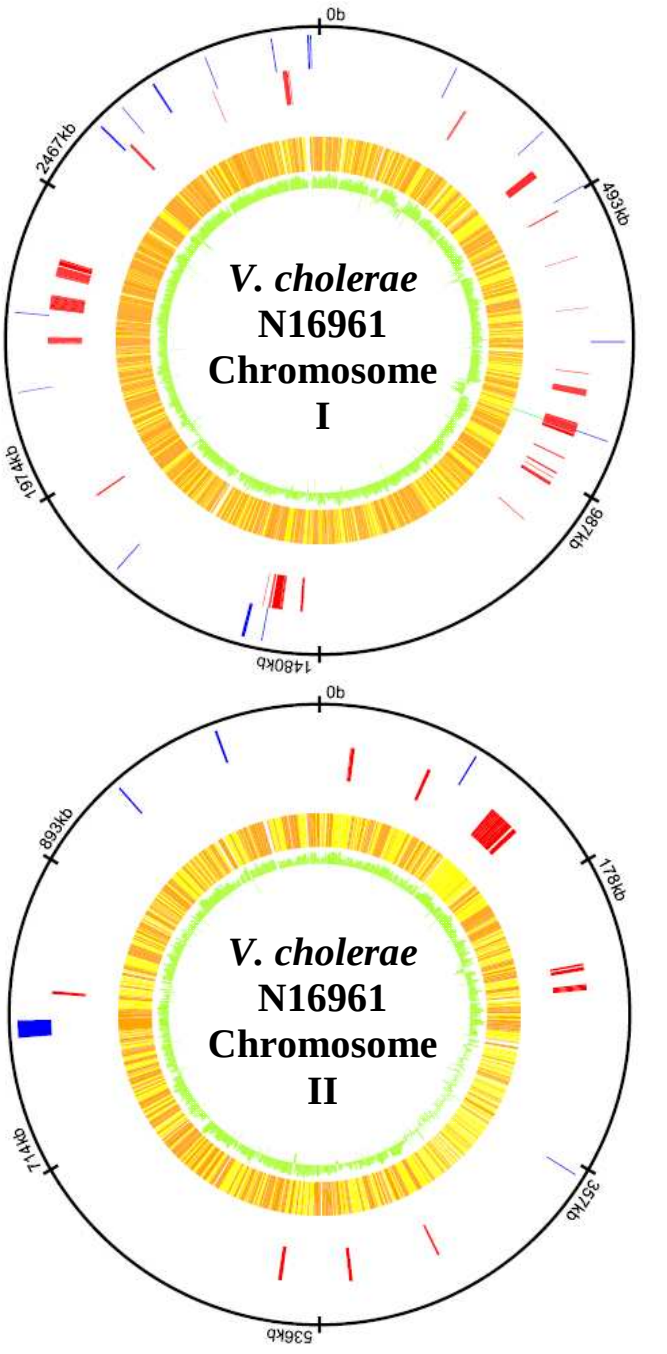


Figure 3.7 : Chromosomal wheels of *V. cholerae* N16961 chromosome I and chromosome II depicting the genomic location of genes found to be under positive selection between *V. cholerae* N16961 and O395, database identified virulence genes, virulence genes under positive selection, genes on the plus and minus strand as well as %GC content.

Influence of Databases on Virulence Genes Under Positive Selection

Both MvirDB and VFDB contain gene records of different virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae* (Table 3.3, page 44; Figure 3.4 page 46). To investigate if the differing virulence gene numbers mined from MVirDB and VFDB may possibly skew the P-values when testing for an enrichment of virulence genes undergoing positive selection (Table 3.4, page 45), the number of virulence genes identified from each database and the overlap between genes under positive selection was tested individually. Results of one-way classification Chi-Square tests to determine if virulence gene records within either MVirDB or VFDB individually may have a statistically significant intersection with genes under positive selection within *H. pylori*, *N. meningitidis* and *V. cholerae* are presented below in Table 3.5.

	Bacterial Organism					
	H. pylori		***N. meningitidis***		***V. cholerae***	
N^{o} of Genes under Positive Selection	230		218		23	
Virulence Gene Database	**MVirDB**	**VFDB**	**MVirDB**	**VFDB**	**MVirDB**	**VFDB**
N^{o} of Virulence Genes	138	87	37	68	33	165
N^{o} of Virulence Genes under Positive Selection	19	13	12	14	1	1
P - value	0.7685	0.8858	0.0001258	0.005074	0.4735	0.6367

Table 3.5 : One-way classification Chi-Square tests were conducted to determine if either MVirDB or VFDB identified virulence genes are statistically enriched for genes undergoing positive selection. Full output of the Chi-Square test results are contained in Appendix B3.

For *H. pylori* and *V. cholerae*, neither VFDB nor MVirDB database identified virulence genes are statistically enriched for genes undergoing positive selection at the P = 0.05 threshold (Table 3.5). There is no database bias in terms of a skewed P value for *H. pylori* and *V. cholerae* with regards to the number of virulence gene records present within MVirDB or VFDB and their association with genes under positive selection as their P values are similar (Table 3.5). The P value for the one-way classification Chi-Square test for genes under positive selection and virulence genes for *N. meningitidis* is much smaller for MVirDB compared to VFDB (Table 3.5). It is possible the larger identified virulence gene dataset from VFDB could result in a slightly higher overall P value when testing for a significant overlap between VFDB and

MVirDB virulence genes and genes undergoing positive selection within *N. meningitidis* (Table 3.4 page 45). The skewing of the P value caused by VFDB is supported by the overall P value for the intersection between the merged virulence gene list and genes under positive selection within *N. meningitidis* (P = 0.005074; Table 3.4 page 45) being the same as the P value obtained when testing between VFDB and genes under positive selection in *N. meningitidis* (P = 0.005074; Table 3.5, page 50), compared to the P value obtained when testing between MVirDB and *N. meningitidis* genes under positive selection (P = 0.0001258; Table 3.5, page 50).

Positive Selection and Virulence Gene Features

Apart from *N. meningitidis*, database identified virulence genes for *H. pylori* and *V. cholerae* do not display a statistically significant enrichment for genes under positive selection. As there is no significant association between genes under positive selection and database identified virulence genes, are there other features that these gene sets may or may not share? To investigate any commonalities or differences between genes under positive selection and database identified virulence genes, traits such as gene lengths and Guanine-Cytosine percentage content (%GC content) of the two gene type subsets were examined in detail, and compared to the overall genome gene length and %GC content distributions.

Gene lengths in basepairs and %GC content of genes under positive selection, database identified virulence genes and all genes contained within the genome (excluding tRNAs and rRNAs) for *H. pylori*, *N. meningitidis* and *V. cholerae* were obtained using EMBOSS's InfoSeq utility [98]. An example of an outputted EMBOSS InfoSeq file is presented below in Figure 3.8.

Name	Accession	Type	Length	%GC
HP0022	-	N	1566	39.97
HP0033	-	N	2226	39.44
HP0037	-	N	1056	38.16
HP0038	-	N	738	39.97
HP0048	-	N	2310	41.77
HP0052	-	N	993	34.74
HP0056	-	N	3558	40.84
HP0060	-	N	2442	39.03
HP0062	-	N	261	39.46
HP0063	-	N	1491	38.83
HP0065	-	N	354	35.59
HP0066	-	N	2496	40.06
HP0069	-	N	765	40.92

Figure 3.8: Example of an output list from EMBOSS's InfoSeq utility containing bacterial gene features for genes under positive selection in *H. pylori* 26695. The name column contains bacterial gene accessions used through out this study. The type field indicates sequence type (nucleotide in this instance), the length field is the gene length in basepairs and the %GC column the %GC content calculated over the length of a gene.

The workflow used to obtain gene lengths and %GC content for all gene list types for *H. pylori*, *N. meningitidis* and *V. cholerae* is presented below in Figure 3.9.

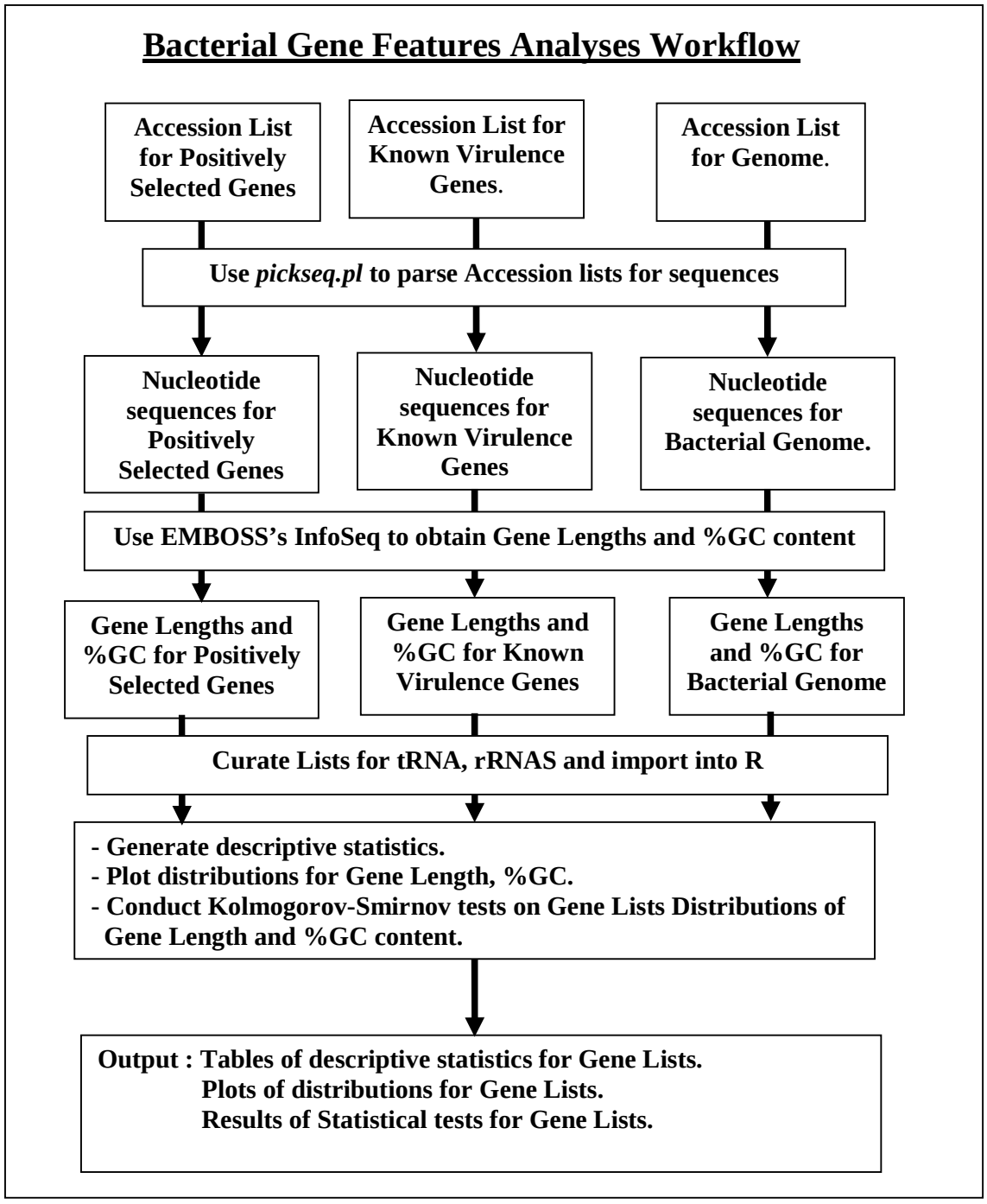


Figure 3.9 : Workflow schema employed in order to obtain the gene lengths of nucleotide sequences and %GC content for each of the genes under positive selection, database identified virulence genes and genomic genes for *H. pylori*, *N. meningitidis* and *V. cholerae* to statistically test the distributions of genes belonging to each of the three gene list types with respect to each other.

Comparisons of Gene Length Distributions

Distributions of the three gene list types (genes under positive selection, database identified virulence genes and genomic genes) for *H. pylori*, *N. meningitidis* and *V. cholerae* are presented as series of histograms below and overleaf in Figures 3.10, 3.11 and 3.12. Descriptive statistics of gene length datasets for genes under positive selection, database identified virulence genes and genome genes belonging to *H. pylori*, *N. meningitidis* and *V. cholerae* are presented in Appendix B4, Table B4A.

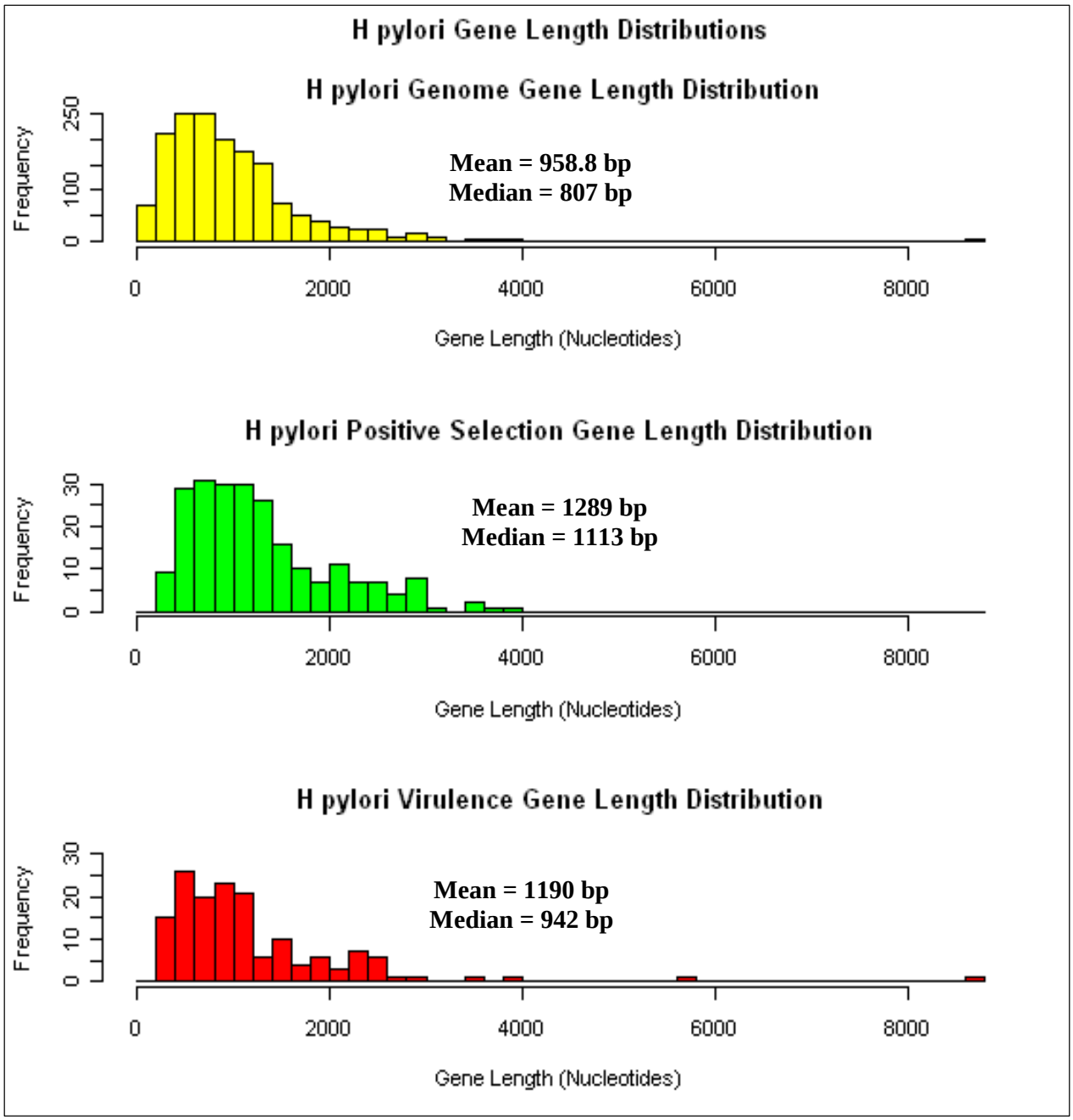


Figure 3.10 : Histograms providing distributions, mean and median of gene lengths obtained for all genes in the genome (Yellow), genes under positive selection (Green) and database identified virulence genes (Red) for *H. pylori*.

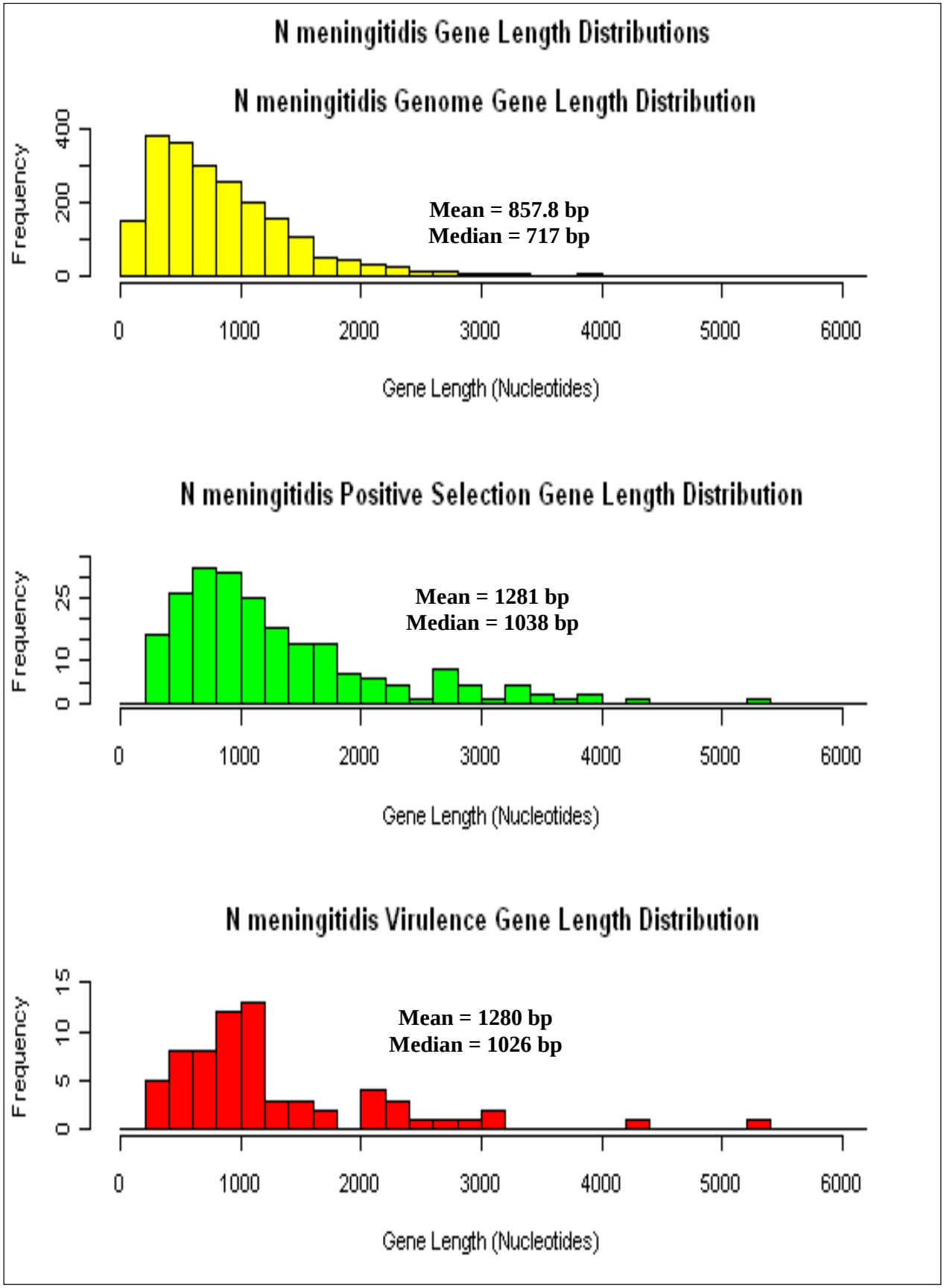


Figure 3.11 : Distributions, mean and median gene lengths for *N. meningitidis* genes in the genome (Yellow), genes under positive selection (Green) and database identified virulence genes (Red).

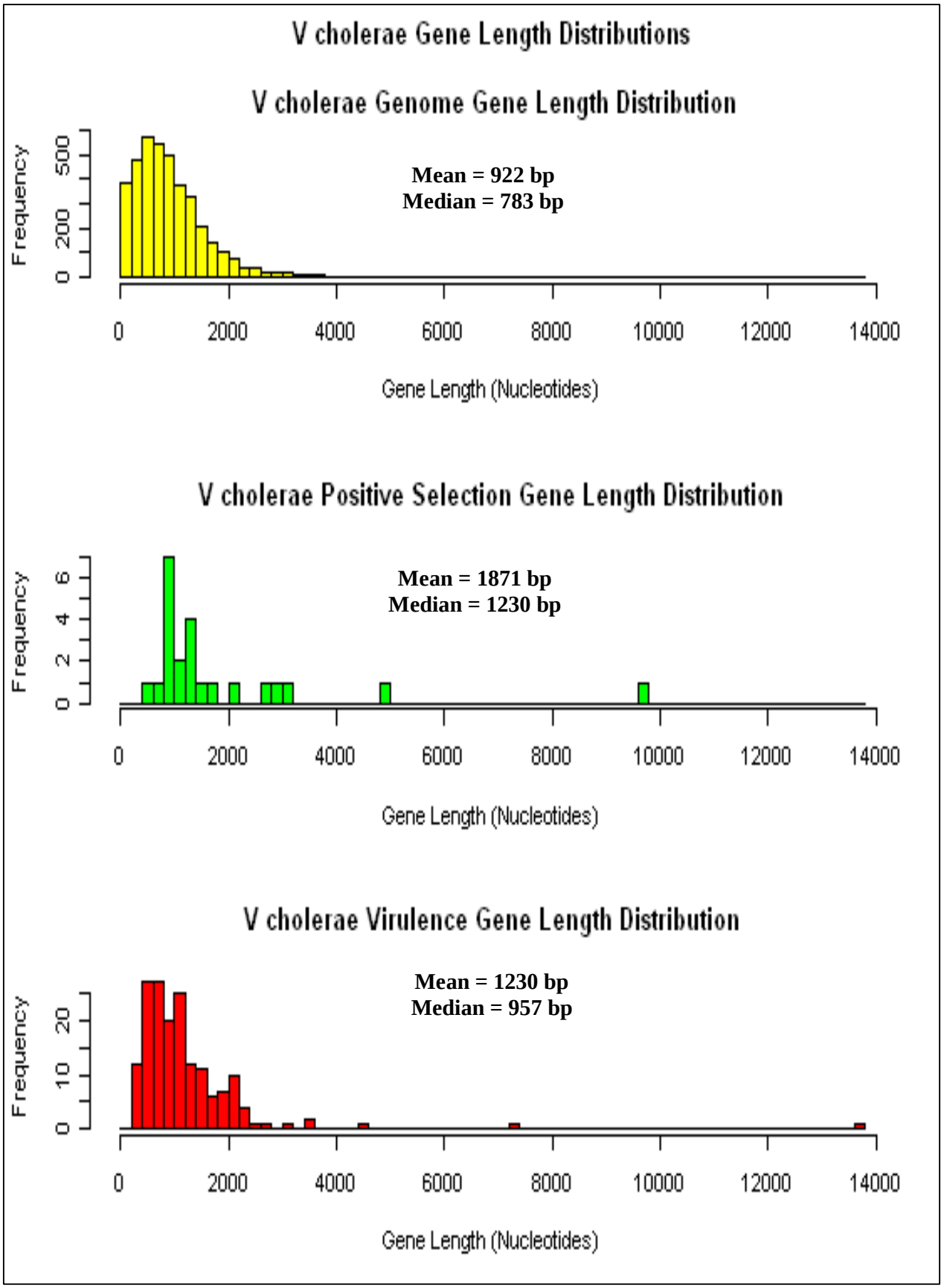


Figure 3.12 : Histograms providing the distribution, mean and median gene lengths for all genes present *V. cholerae* chromosome I and II (Yellow), genes under positive selection (Green) and database identified virulence genes (Red).

The distributions of gene lengths between genomic genes, genes under positive selection and database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae* does not follow a normal distribution (Figures 3.10, 3.11, and 3.12; pages 53, 54, and 55 respectively). To test if the distributions of gene lengths significantly differ between genes under positive selection, database identified virulence genes and genomic genes, non-parametric Kolmogorov-Smirnov tests were conducted and P-values were adjusted using the Holm's method to correct for multiple hypotheses testing as implemented within R [108].

Bacterial Organism	Gene Lists Tested	P – Value	Adjusted P – Value
H. pylori	Positive Selection vs Virulence Genes	0.01218	0.02436
	Positive Selection vs Genomic Genes	1.395e – 07	4.170e – 07
	Virulence Genes vs Genomic Genes	0.02079	0.02436
N. meningitidis	Positive Selection vs Virulence Genes	0.5515	0.5515
	Positive Selection vs Genomic Genes	7.664e – 11	2.2992e – 10
	Virulence Genes vs Genomic Genes	4.386e – 05	8.772e – 05
V. cholerae	Positive Selection vs Virulence Genes	0.04792	0.04792
	Positive Selection vs Genomic Genes	0.0005018	0.0010036
	Virulence Genes vs Genomic Genes	0.0001114	0.0003342

Table 3.6 : Distributions of gene lengths between genes under positive selection, database identified virulence genes and genomic genes for *H. pylori*, *N. meningitidis* and *V. cholerae* were tested against each other using a Kolmogorov-Smirnov test with P values being adjusted for multiple hypotheses testing using the Holm's method.

Gene lengths of *H. pylori* genes under positive selection, database identified virulence genes and genomic genes follow different, statistically significant distributions at the P = 0.05 interval (Table 3.6). Genes under positive selection within *H. pylori* have a distribution skewed towards longer genes compared to database identified virulence genes and genomic genes

(Figure 3.10, page 53). On average, *H. pylori* genes under positive selection are 99bp and ~330bp longer then database identified virulence genes and genomic genes respectively (Figure 3.10, page 53). Database identified *H. pylori* virulence genes are also biased towards a longer gene length distribution compared to genomic genes (adjusted P = 0.02436, Table 3.6, page 56) and on average, are ~ 231bp longer in length (Figure 3.10, page 53). Genes under positive selection and database identified virulence gene length distributions differ significantly enough from genomic genes for them to be distinguished as two separate subsets within the *H. pylori* genome (adjusted P = 0.02436 , adjusted P = 4.170e – 7, Table 3.6, page 56).

Gene length distributions of genes under positive selection and database identified virulence genes for *N. meningitidis* do not differ statistically at the P = 0.05 mark (adjusted P = 0.5515). Mean gene lengths for *N. meningitidis* genes under positive selection and database identified virulence genes are quite similar at 1,280bp (Figure 3.11, page 54). Both genes under positive selection and database identified virulence genes for *N. meningitidis* are biased towards a longer gene length distribution, compared to the genomic mean of 858bp (Figure 3.11, page 54 and Table 3.6, page 56). Hence, *N. meningitidis* database identified virulence genes and genes under positive selection are on average, longer then the normal gene length distribution within the genome (adjusted P = 8.772e – 05 and P = 2.2992e – 10 respectively, Table 3.6, page 56) and share similar gene length distributions. As there is a statistically significant enrichment of virulence genes undergoing positive selection in *N. meningitidis* (Table 3.5, page 50), one would expect similar gene length distributions between the two gene lists.

V. cholerae differences in gene length distributions are similar to those observed in *H. pylori* with genes under positive selection being on average, longer then database identified virulence genes and genomic genes (Figure 3.12, page 55). *V. cholerae* gene length distribution between genes under positive selection and database virulence genes differs statistically at the P = 0.05 value (adjusted P = 0.04792, Table 3.6, page 56). Gene length differences between genes under positive selection and database identified virulence genes are smaller, compared to differences in gene lengths by both sets of gene list types to genomic genes (adjusted P = 0.0010036 and P = 0.0003342, respectively, Table 3.6, page 56).

A similarity shared by *H. pylori*, *N. meningitidis* and *V. cholerae* is that on average, genes under positive selection are longer then database identified genes which in turn, are longer then mean genomic gene lengths (Figures 3.10, 3.11 and 3.12, pages 53, 54 and 55). A possible reason for why genes under positive selection may be biased towards longer gene lengths is a longer sequence will contain a larger number of codons within the alignment, thereby providing

more data making the null hypothesis of no selection easier to reject by CODEML[♣]. The observation that longer sequences are more likely to reject the null hypothesis of no selection is validated by examining the mean lengths of genes undergoing positive selection with an LRT score significant at the P = 0.001 level, and comparing them with the mean gene lengths of all genes undergoing positive selection with an LRT score significant at P = 0.05 mark (Figure 3.3, page 38). For *H. pylori*, the mean length of the 83 genes whose LRT score is significant at the P = 0.001 level (Figure 3.3, page 38) is 1,605bp compared 1,289bp for all 230 genes whose LRT score is significant at the P = 0.05 level (Figure 3.3, page 38, Figure 3.10 page 53). Similarly, for *N. meningitidis* and *V. cholerae* the mean lengths of the 92 and 6 genes equal to or surpassing the significance test at P = 0.001 are 1,511bp and 2,558bp respectively, compared to 1,281bp and 1,871bp for all genes making the P = 0.05 threshold (Figure 3.3, page 38, Figure 3.11, page 54 and Figure 3.12, page 55).

The fact that CODEML is more likely to reject the null hypothesis of no selection for genes whose lengths are greater then the genomic mean could be an artefact of longer genes containing more codons thereby increasing the probability of non-synonymous substitutions occurring, rather then a selective bias by CODEML for genes of a certain length. The reason that CODEML is not selectively biased towards identifying genes of a certain length to be under positive selection is supported by examining the longest genes within *H. pylori*, *N. meningitidis* and *V. cholerae* genomes which are HP0289 (8,682bp), NMA0688 (6,048bp) and VC1451 (13,677bp) respectively which were found not to be under positive selection by CODEML. In comparison, the maximum gene lengths of genes under positive selection within *H. pylori*, *N. meningitidis* and *V. cholerae* are HP0887 (3,873bp), NMA0905 (5,322bp) and VCA0849 (9,792bp) respectively.

The finding that virulence genes on average tend to be longer then the genomic gene mean for *H. pylori*, *N. meningitidis* and *V. cholerae* is surprising (Figures 3.10, 3.11 and 3.12, pages 53, 54 and 55). There does not appear to be any literature which would suggest *a priori* that genes involved in virulence processes should be on average, longer then the genomic gene length average.

♣ Personal Communication with Dr. Cathal Seoighe

Comparisons of %GC Content Distributions

The %GC content for the three gene list types (genes under positive selection, database identified virulence genes and genomic genes) were obtained in a similar fashion to the gene lengths using EMBOSS's InfoSeq utility for *H. pylori*, *N. meningitidis* and *V. cholerae* (Figure 3.9, page 52) [98]. Distributions of %GC for the three gene list types for each bacterial organism under study were visualized as a series of histograms presented in Figures 3.13, 3.14 and 3.15 (below and pages 60, 61). Descriptive statistics of the %GC datasets for *H. pylori*, *N. meningitidis* and *V. cholerae* are presented in Table B4B, Appendix B4.

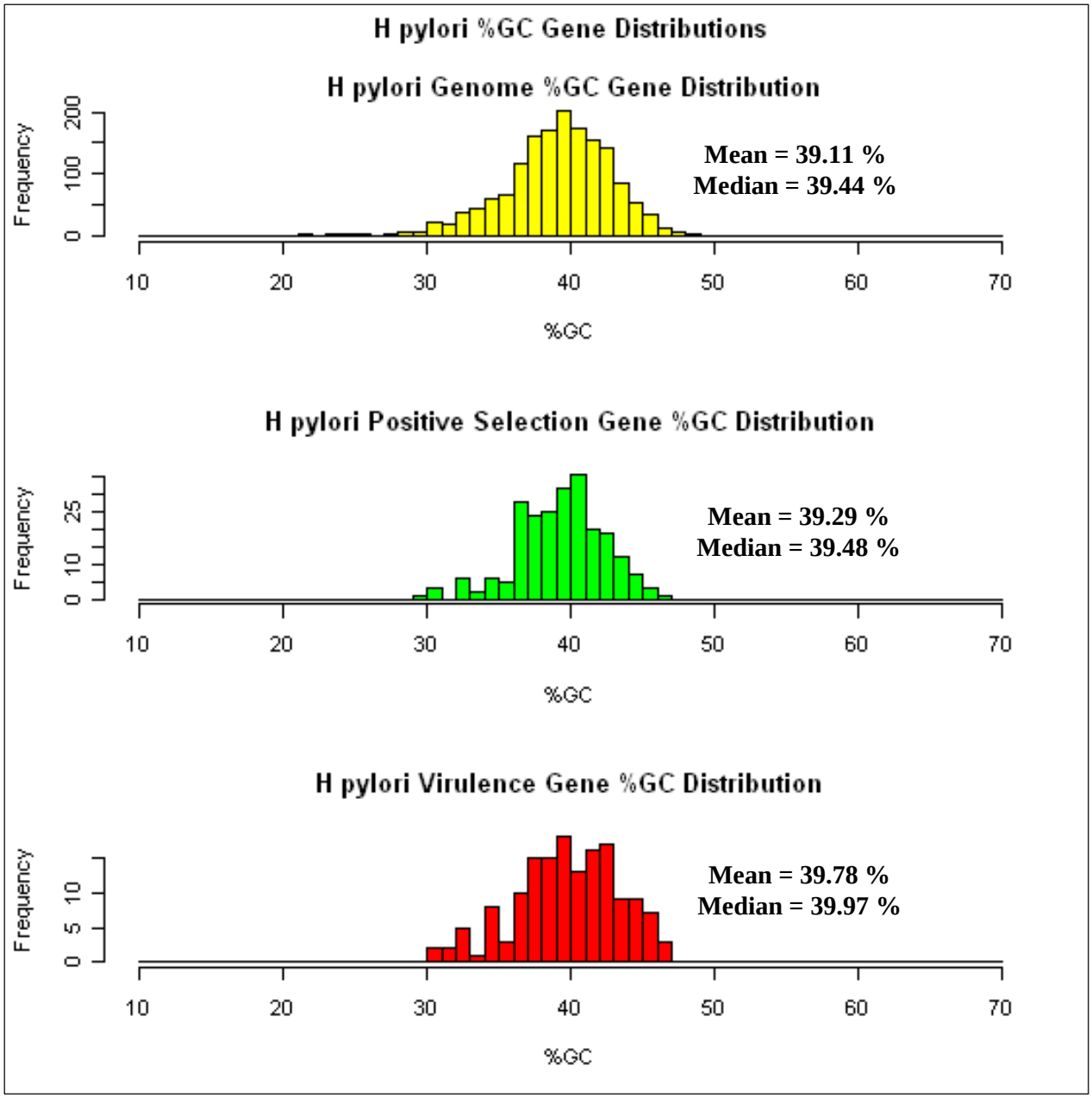


Figure 3.13 : Distribution of %GC content for all genes within *H. pylori*'s genome (Yellow), genes under positive selection (Green), database identified virulence genes (Red) as well as the mean and median for each gene list type.

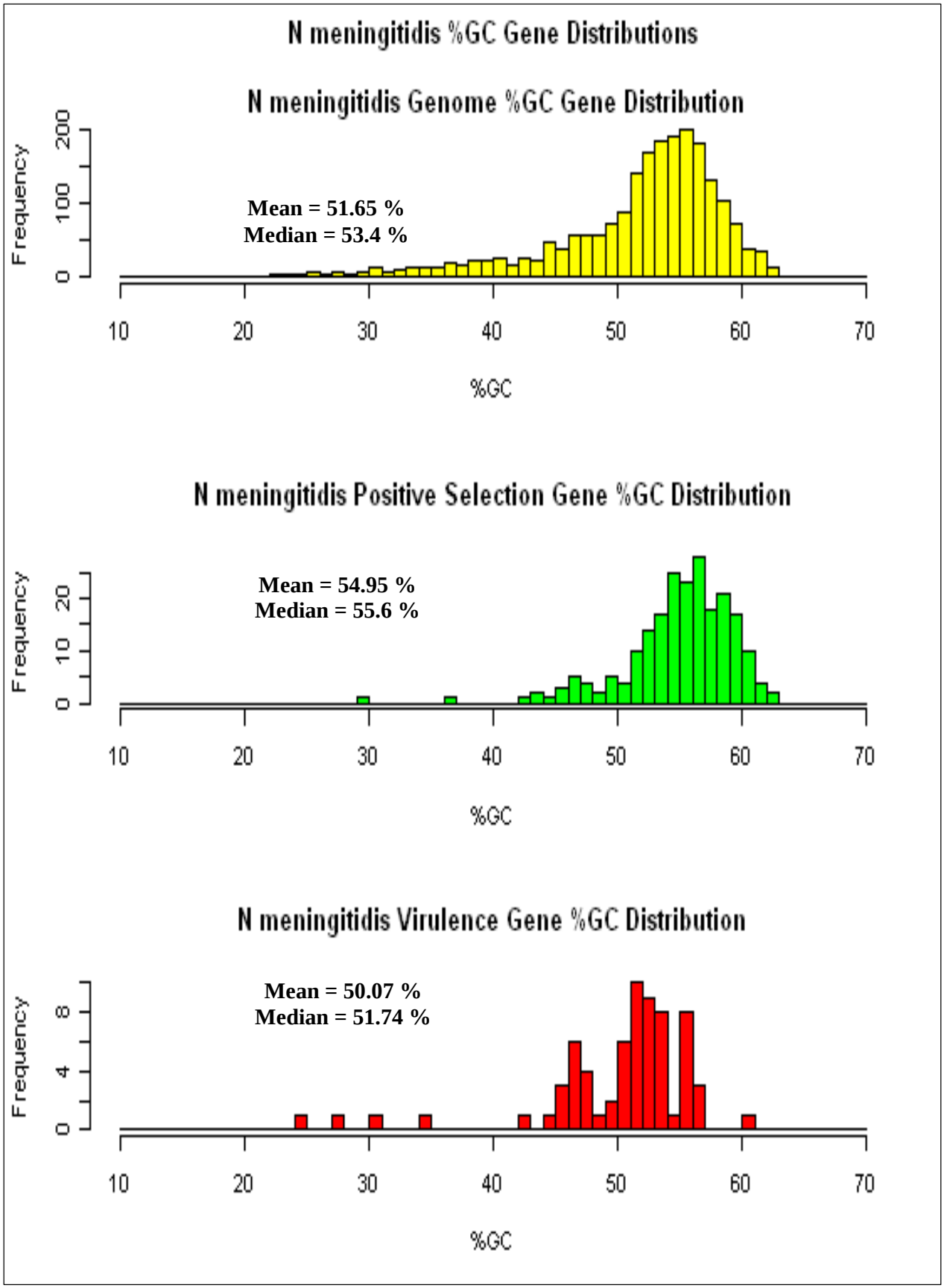


Figure 3.14 : *N. meningitidis* %GC content for all the genomic genes (Yellow), genes under positive selection (Green) and database identified virulence genes (Red) as well as the mean and median %GC content for each gene list type.

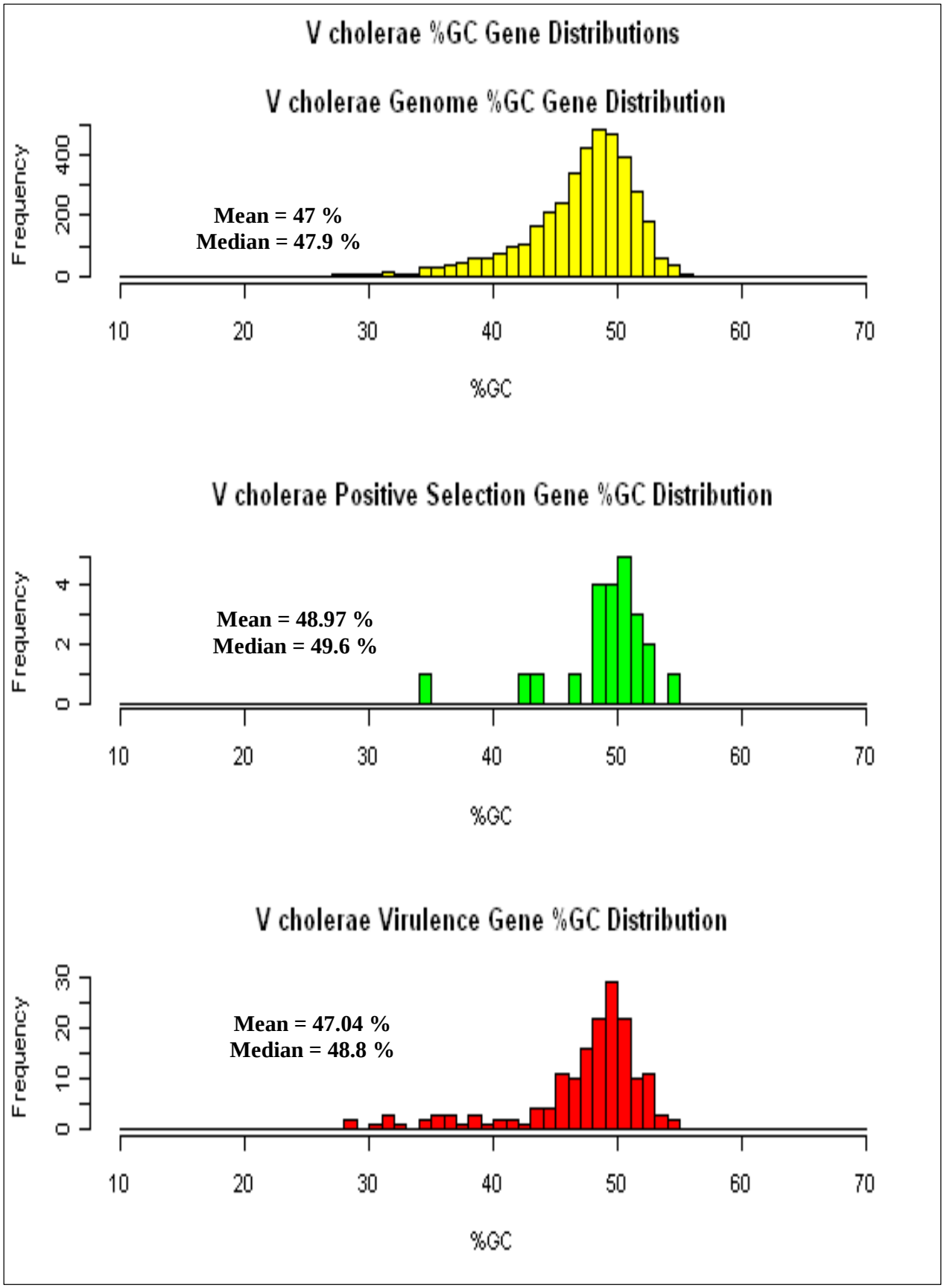


Figure 3.15 : Histograms showing %GC distributions, mean and medians for *V. cholerae* genomic genes (Yellow), genes under positive selection (Green) and database identified virulence genes (Red).

Distributions of %GC content for genes under positive selection and database identified virulence genes were tested against each other and against the genomic distribution to determine if there are any differences between the three gene list types for *H. pylori*, *N. meningitidis* and *V. cholerae* using Kolmorogov – Smirnov tests and Holm's correction for multiple testing, the results are presented below in Table 3.7.

Bacterial Organism	Gene Lists Tested	P – Value	Adjusted P – Value
H. pylori	Positive Selection vs Virulence Genes	0.02705	0.08115
	Positive Selection vs Genomic Genes	0.2814	0.28140
	Virulence Genes vs Genomic Genes	0.08131	0.16262
N. meningitidis	Positive Selection vs Virulence Genes	8.634e – 12	2.5902e – 11
	Positive Selection vs Genomic Genes	3.398e – 10	6.796e – 10
	Virulence Genes vs Genomic Genes	0.0001082	0.0001082
V. cholerae	Positive Selection vs Virulence Genes	0.1096	0.1096
	Positive Selection vs Genomic Genes	0.004702	0.014106
	Virulence Genes vs Genomic Genes	0.04080	0.0612

Table 3.7 : Results of Kolmorogov – Smirnov tests conducted to determine if the %GC content distributions for genes under positive selection, database identified virulence genes and genomic genes statistically differ from each other with P values corrected for multiple hypotheses testing using Holm's method.

No statistically significant differences in %GC distributions could be found between *H. pylori* genes under positive selection, database identified virulence genes and genomic genes at the P = 0.05 value (Table 3.7). The mean %GC value for *H. pylori* genomic genes, genes under positive selection and database identified virulence genes exhibits slight variations at the 39% interval, with virulence genes being slightly more %GC rich with a mean of 39.78% compared to the genomic mean of 39.11% (Figure 3.13, page 59). The similarity of mean %GC content of

database identified virulence genes compared to the genomic mean for *H. pylori* is surprising as one would expect the *Cag* pathogenicity island (*Cag* PAI) contained within the virulence gene dataset to skew the %GC content away from the genomic mean. The mean %GC content of genes present within the *Cag* PAI (HP0520 – HP0548; excluding HP0533), was calculated to be 35% which is significantly different from the genomic distribution mean of 39.11% (P = 2.047e-07, Kolmogorov – Smirnov test). Hence, although there are differences between specific virulence gene subsets with regards to %GC distribution, these differences are smoothened out when the %GC content of the whole virulence dataset, as opposed to subsets, are compared to the %GC content of genomic genes and genes under positive selection in *H. pylori*.

For *N. meningitidis*, all three types of gene lists have statistically significant differences in their %GC content distributions (Table 3.7, page 62). Genes under positive selection are skewed towards a more %GC rich content compared to genomic genes (adjusted P = 6.796e-10) and database identified virulence genes (adjusted P = 2.5902e-11; Table 3.7, page 62, Figure 3.14, page 60). Database identified virulence genes for *N. meningitidis* are skewed towards a smaller mean %GC content compared to the mean genomic distribution (Figure 3.14, page 60). Compared to genes under positive selection, the %GC content distribution of database identified virulence genes is closer to the genomic %GC content distribution as seen by the larger P-values (adjusted P = 0.0001082 compared to adjusted P = 2.5902e-11, Table 3.7, page 62). The closeness of *N. meningitidis* %GC content distribution of database identified virulence genes towards genomic genes and not genes under positive selection is unexpected. *N. meningitidis* database identified virulence genes are enriched for genes under positive selection and thus, a similar %GC content distribution would be expected, as in the case of gene lengths (Table 3.6, page 56). However, more database identified virulence genes for *N. meningitidis* below the 50% GC content interval can be seen compared to genes under positive selection (Figure 3.14, page 60). Hence, although database identified virulence genes are enriched for genes undergoing positive selection, the two gene list types can be differentiated based upon their %GC content distributions (Table 3.7, page 62, Figure 3.14, page 60). *N. meningitidis* database identified virulence genes can also be differentiated from genomic genes by having a statistically significant %GC content distribution that is less then the genome, while genes under positive selection can be differentiated from the genome by having higher %GC contents (Table 3.7, page 62, Figure 3.14, page 60).

Compared to genomic genes, genes under positive selection within *V. cholerae* have a different statistically significant %GC distribution (adjusted P = 0.014106, Table 3.7, page 62) and are on average, more GC rich (Figure 3.15, page 61). Genes under positive selection and database identified virulence genes for *V. cholerae* do not have a statistically significantly

different %GC distribution (adjusted P = 0.1096, Table 3.7, page 62). Genomic genes and database identified virulence genes also do not have a statistically significantly different %GC distribution at the P = 0.05 threshold (adjusted P = 0.0612, Table 3.7, page 62).

N. meningitidis and *V. cholerae* genes under positive selection tend to have a slightly higher mean %GC content compared to database identified virulence genes and the genome as a whole (Figures 3.14 and 3.15, pages 60 and 61). *H. pylori* database identified virulence genes have a higher mean %GC content as compared to genomic genes (Figure 3.13, page 59) while *N. meningitidis* and *V. cholerae* database identified virulence genes have lower %GC means compared to genomic genes with the differences in %GC content being more pronounced within *N. meningitidis* (Figures 3.14 and 3.15, pages 60 and 61).

Chapter 4 : Functional Annotation

Introduction

Genes under positive selection and database identified virulence genes for *H. pylori, N. meningitidis* and *V. cholerae* have so far been examined based on gene counts and physical properties, such as gene lengths and %GC content. Within this chapter, biological properties of genes under positive selection and database identified virulence genes for *H. pylori, N. meningitidis* and *V. cholerae* are investigated to elucidate the biological properties of each gene list type. Different functional annotation systems were employed to scrutinize genes under positive selection to establish what biological processes are under positive selection and to what extent do any functional similarities or differences exist across the three bacterial species. Additionally, functional characterization of database identified virulence genes was undertaken to determine whether virulence processes for an obligatory pathogen, commensal / opportunistic pathogen and free living facultative pathogen (*H. pylori, N. meningitidis* and *V. cholerae* respectively) have any similarities.

Functional annotation of gene lists occurs at a 1 – dimensional level and routine genome sequencing has propelled the proliferation of annotation systems, controlled vocabularies and tools that seek to provide meaningful, biological contextualization of sequence based information [36]. Different systems of functional classification either use a controlled vocabulary of terms (ontologies), for which putative function is assigned to a gene sequence based on homology or where possible, experimental evidence. Other systems utilize algorithmic prediction methods based upon sequence features to classify gene sequences into biologically relevant categories. Regardless of the approach used, the coverage and depth of biological annotations for an organism present within that functional classification system will have a direct impact upon functional analyses of experimentally generated gene lists.

Functional annotation of genes under positive selection and database identified virulence genes was conducted using controlled vocabularies from Gene Ontology (GO) and Clusters of Orthologous Groups of proteins (COGs) to determine if there are any enriched biological processes under positive selection and to characterize database identified virulence gene processes [102,110,112]. Sub-cellular localization predictions of protein products for genes under positive selection and database identified virulence genes for *H. pylori, N. meningitidis* and *V. cholerae* was investigated [113-115].

Gene Ontology (GO) Functional Annotation of Bacterial Gene Lists

A major aim of the collaborative GO effort is to describe the purpose of a gene and its products within an organism through the use of a universal, unambiguous organized vocabulary of biological terms [110]. GO annotations comprise of three descriptive facets; biological process, molecular function and cellular component to describe the biological role a gene's product contributes to, the biochemical reaction a gene may be involved in and the position within a cell where gene's product is active [110]. Each of the three GO descriptive facets encompass a five-tiered hierarchy which have broad biological terms placed on top of the hierarchy and more specific biological terms placed at the bottom of the hierarchy [109-111].

The biological community has widely adopted GO annotations with over 2,940 publication citations♣ and has given rise to numerous bioinformatics tools that seek to annotate and find GO term associations amongst gene lists generated from high throughput experiments♠. A popular tool providing GO annotations coupled with statistical interpretation of GO term significance for a gene list of interest is the Database of Annotation, Visualization and Integrated Discovery (DAVID) [111,166]. DAVID aims to expedite the biological discovery process from experimentation by providing a centralized, integrated resource of numerous annotation types for characterizing a gene list of interest's biological associations and to date, has been used in over 1,000 publications [111,166].

Lists of genes under positive selection for *H. pylori*, *N. meningitidis* and *V. cholerae* were uploaded to DAVID's web – interface and an enrichment of GO terms were assayed for using genome wide GO annotations of the respective bacterial species as background lists. Identified GO annotations for genes under positive selection were compared across the three bacterial species to determine what biological processes are undergoing positive selection and if those processes are similar or discrete. Similarly, database identified virulence genes were also uploaded to DAVID's web – interface and an over representation of GO terms relative to the bacterial organism's genome was obtained. GO annotations of database identified virulence gene lists were obtained to determine if the three bacterial species use similar or differing biological processes in their pathogenesis. Finally, comparisons of GO annotations of genes under positive selection and database identified virulence genes lists was conducted manually and through the use of DAVID's gene list manager to ascertain whether any commonalities between biological processes under positive selection and bacterial virulence exist.

♣ http://www.geneontology.org/cgi-bin/biblio.cgi - As of November 2009

♠ http://www.geneontology.org/GO.tools.shtml#out_house

GO Processes under Positive Selection

Bacterial gene list annotation using GO terms was done at the 4th level of the GO hierarchy with an EASE score of 0.5 within DAVID for all bacterial gene lists so as to maintain consistency when comparing GO terms across all three bacterial species. After evaluation of all GO terms at the five different levels of the GO hierarchy, selection of GO annotation terms at the 4th level of the GO hierarchy was settled upon in an attempt to maximize gene list coverage and at the same time, obtain GO terms that are not too diffuse in their definitions. DAVID utilizes a Fisher's exact test when assaying for GO category functional enrichment and makes corrections for multiple hypothesis testing using a number of methods, thereby providing a good statistical metric that can be compared across bacterial systems [111,166]. Coverage of the number of genes mapped to GO terms for the uploaded lists of genes under positive selection are summarized below, in Table 4.1.

Bacterial Organism	**N^{o} of Genes under Positive Selection**	**N^{o} of Genes under Positive Selection with GO Terms Mapped**	**Percentage (%) of Gene List GO Term Coverage**
H. pylori	230	97	42%
N. meningitidis	218	64	29%
V. cholerae	23	7	30%

Table 4.1 : Coverage of gene lists for genes under positive selection that could be mapped to GO terms using the DAVID system in numbers and percentage.

GO term annotation coverage of *H. pylori* genes under positive selection is higher compared to coverage of GO term annotations for genes under positive selection within the *N. meningitidis* and *V. cholerae* bacterial systems (Table 4.1). In terms of functional enrichment of GO categories for genes under positive selection, no statistically significant enrichment for a particular GO category could be found at the P = 0.05 mark after correcting for multiple hypothesis testing.

The GO terms for genes under positive selection within *H. pylori*, *N. meningitidis* and *V. cholerae* and the number of genes mapped to each GO term is presented overleaf, in Figure 4.1.

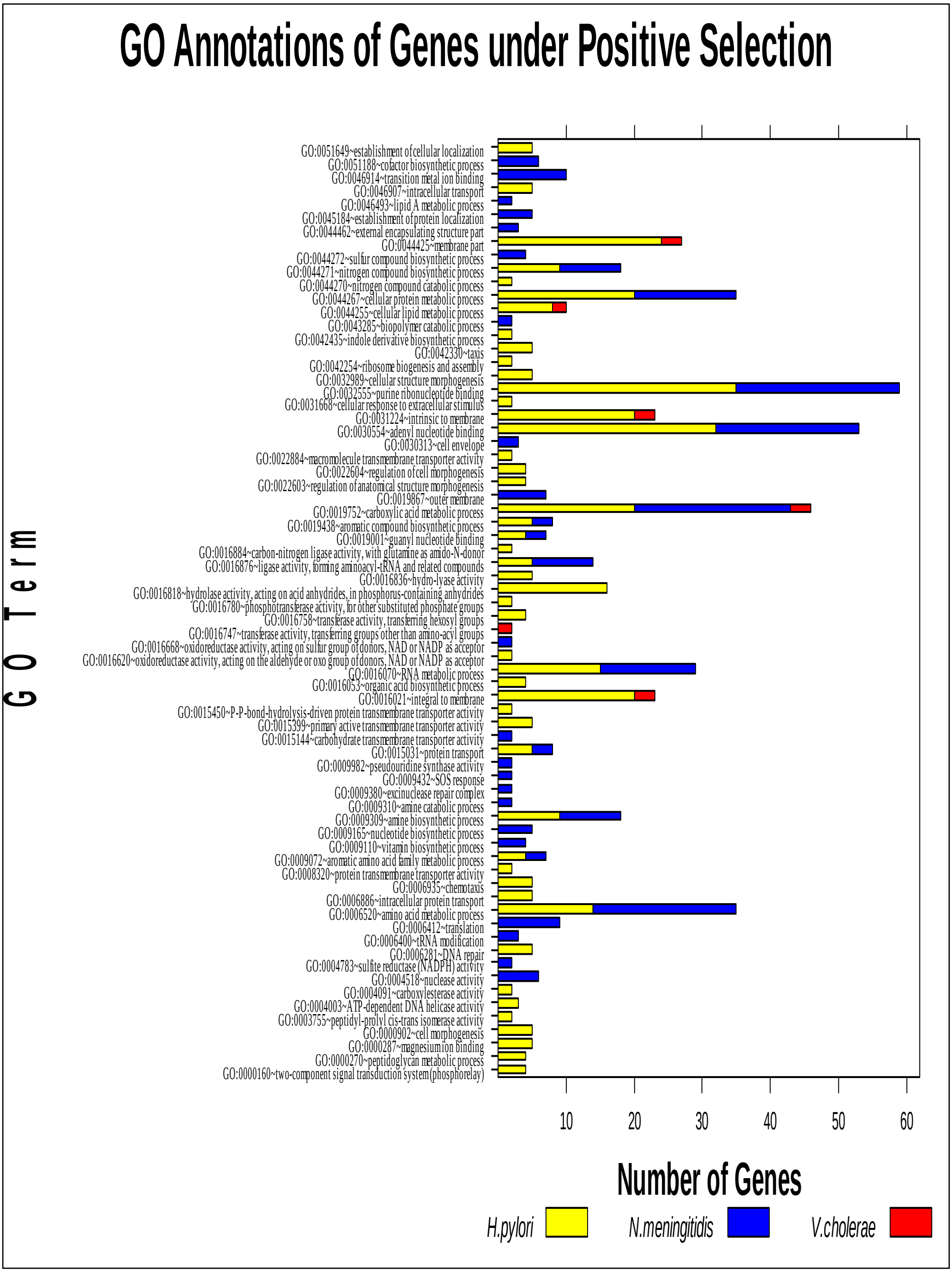


Figure 4.1 : Level four GO annotations that map to genes under positive selection within *H pylori*, *N meningitidis* and *V cholerae* using the DAVID system.

Nine common GO terms are shared between *H. pylori* and *N. meningitidis* genes under positive selection with purine ribonucleotide binding (GO:0032555), adenyl nucleotide binding (GO:0030554) and cellular metabolic process (GO:0044267) sharing the most genes between the two organisms (Figure 4.1, page 68). *H. pylori* and *N. meningitidis* share other DNA and RNA metabolism GO categories such as guanyl nucleotide binding (GO:0019001) and RNA metabolic processes (GO:0016070) that are under positive selection within the two bacterial systems (Figure 4.1, page 68). A second class of common biological processes under positive selection within *H. pylori* and *N. meningitidis* involves formation of biosynthetic compounds such as amino acids (GO:0006520), nitrogenous (GO:0044271) and aromatic bio-molecules (GO:0019438). The large number of genes involved with DNA metabolism that are under positive selection within *H. pylori* and *N. meningitidis* may possibly reflect the active maintenance of DNA uptake and incorporation apparatuses. Most pathogenic bacteria operate within a framework that favours gene loss, yet numerous pathogenic bacteria tend to actively maintain systems for exogenous DNA uptake which are energy intensive [5,38,69,84]. Both *H. pylori* and *N. meningitidis* have high rates of recombination and are naturally competent species, with *N. meningitidis* requiring DNA uptake to maintain its genomic integrity [78,153]. An interesting extension to this hypothesis is selection for DNA metabolism and continual recombination between strains may enhance the survival of a species and increase its virulence potential by producing new phenotypes with novel antigens to counter adaptive immune responses, as is the case with hyper-invasive *N. meningitidis* strains [69,78,90,153]. For *H. pylori*, DNA metabolism terms such as adenyl nucleotide binding (GO:0030554) may reflect a selective pressure on its growth rate as defined media culture studies reveal a reduction in the growth rate of *H. pylori* by as much as 50% in low concentration adenine and purine media [167,168].

Shared GO terms for *H. pylori* and *V. cholerae* genes under positive selection mainly relate to genes which are intrinsic (GO:0031224) and integral for cell membrane structure (GO:0016021) (Figure 4.1, page 68). *N. meningitidis* does have some genes under positive selection which are part of the cell envelope (GO:0030313), though these GO terms do not overlap with membrane associated GO terms for *H. pylori* and *V. cholerae* (Figure 4.1, page 68). A common GO term under positive selection across all three bacterial species is carboxylic acid metabolic processes (GO:0019752) which involves pathways and chemical reactions utilizing carboxylic acids, mainly pertaining to the energy producing glycolysis pathway that operates under anaerobic conditions [169].

Unique GO terms for *H. pylori* genes under positive selection involve protein transmembrane transport (GO:0008320), intracellular protein transport (GO:0006886), chemotaxis (GO:0006935), cell morphogenesis (GO:0000902) and establishment of cellular

localization (GO:0051649), amongst others (Figure 4.1, page 68). *N. meningitidis* unique GO terms for genes under positive selection involve processes such as DNA repair like the SOS response (GO:0009432) and excinuclease repair complex (GO:0009380), translation (GO:0006412) and the outer membrane (GO:00019867) (Figure 4.1, page 68). Overall, *V. cholerae* genes under positive selection share more GO terms with *H. pylori* genes under positive selection as compared to *N. meningitidis* (Figure 4.1, page 68).

GO Processes of Bacterial Virulence Genes

Enterez Gene IDs of database identified virulence gene lists for *H. pylori*, *N. meningitidis* and *V. cholerae* were uploaded to the DAVID system and the 4th level of the GO hierarchy was chosen for annotation. The number of database identified virulence genes and their coverage by GO annotation terms is presented below in Table 4.2.

Bacterial Organism	N^{o} of Database Identified Virulence Genes	N^{o} of Virulence Genes with GO Terms Mapped	Percentage (%) of Gene List GO Term Coverage
H. pylori	153	90	59%
N. meningitidis	68	24	35%
V. cholerae	168	98	58%

Table 4.2 : The number of database identified virulence genes with 4th level GO terms and the percentage of gene list coverage.

In all instances, database identified virulence genes have a higher percentage of genes with GO terms mapped to them compared to genes under positive selection (Table 4.2, Table 4.1, page 67). Higher coverage of database identified virulence genes with 4th level GO annotations may represent a bias in terms of these gene sets being intensively studied. The majority of database identified virulence genes are mined from the literature and hence, have been well characterised. GO annotation curators also use published literature where possible to assign GO terms to genes thereby resulting in a higher coverage of GO annotations for database identified virulence genes. GO terms for database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae* are presented overleaf, in Figure 4.2.

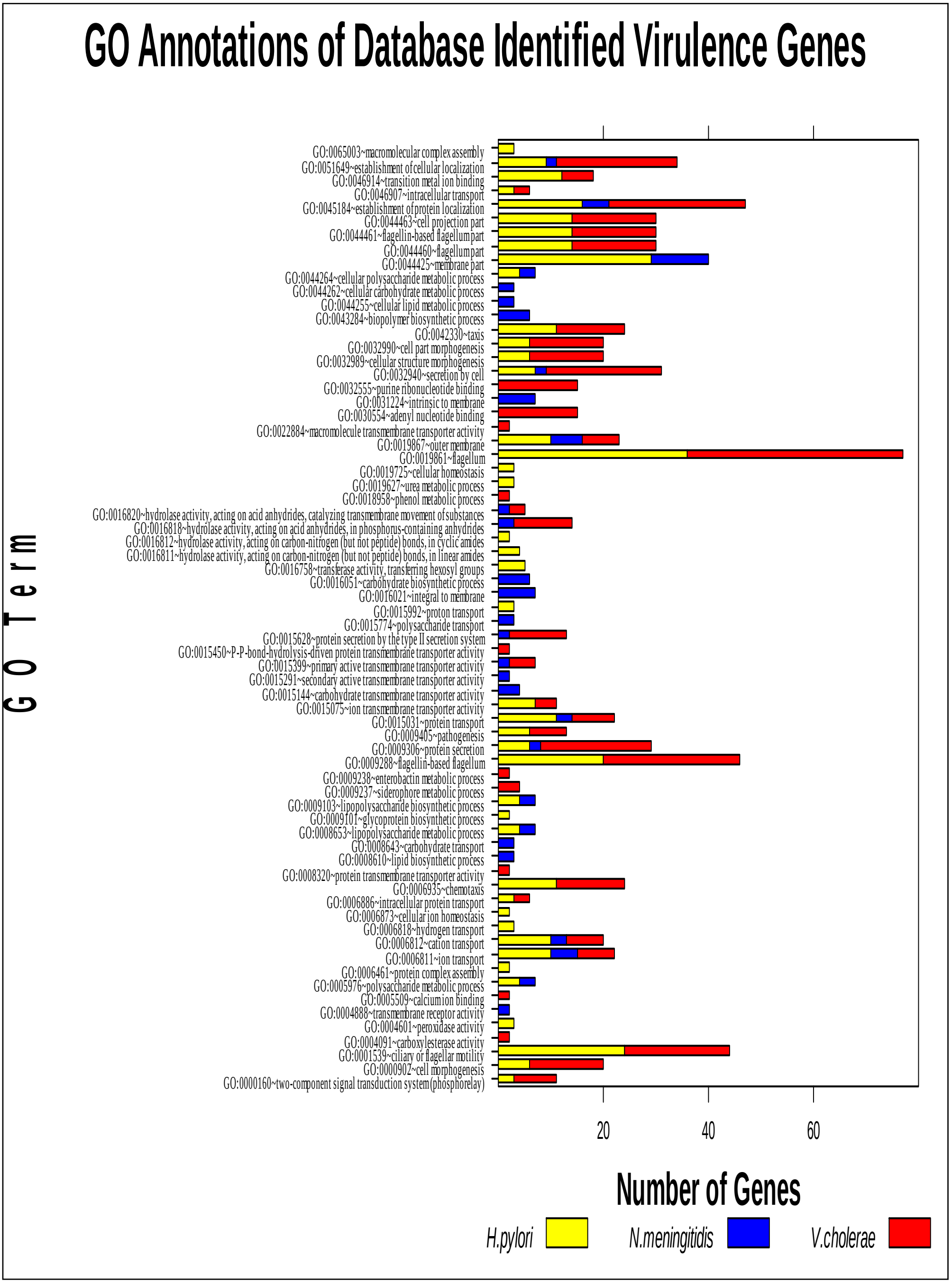


Figure 4.2 : GO terms at the 4th level of the GO hierarchy that map to database identified virulence genes within *H pylori*, *N meningitidis* and *V cholerae* using the DAVID system.

Database identified virulence genes show greater congruence between the bacterial species with eight common GO terms shared by all three bacterial species (Figure 4.2, page 71), compared to a single GO term mapped to all three bacterial species for genes under positive selection (Figure 4.1, page 68). A number of GO terms for database identified virulence genes were found to statistically significantly enriched at the P = 0.05 level, even after stringent Bonferroni corrections for multiple hypothesis testing, as implemented by DAVID. Enriched GO terms found to statistically significant for *H. pylori*, *N. meningitidis* and *V. cholerae* together with their Bonferroni adjusted P-values are presented below, in Table 4.3.

GO ID ~ Term	**Bacterial Species and Adjusted P – value**		
	H. pylori	***N. meningitidis***	***V. cholerae***
GO:0006935~chemotaxis	0.002287881	NIL	0.039493473
GO:0044463~cell projection part	9.13e - 05	NIL	1.35e - 10
GO:0045184~establishment of protein localization	1.27e - 05	NIL	8.94e - 11
GO:0051649~establishment of cellular localization	0.043026936	NIL	7.30e - 13
GO:0019861~flagellum	9.50e - 17	NIL	9.70e - 35
GO:0001539~ciliary or flagellar motility	1.01e - 15	NIL	1.16e - 18
GO:0009288~flagellin-based flagellum	7.21e - 08	NIL	1.63e - 19
GO:0044461~flagellin-based flagellum part	9.13e - 05	NIL	1.35e - 10
GO:0015031~protein transport	0.023301077	NIL	NIL
GO:0015144~carbohydratetransmembrane transporter activity	NIL	0.038649114	NIL
GO:0000902~cell morphogenesis	NIL	NIL	6.29e - 06
GO:0032989~cellular structure morphogenesis	NIL	NIL	6.29e - 06
GO:0032990~cell part morphogenesis	NIL	NIL	9.22e - 11
GO:0032940~secretion by cell	NIL	NIL	1.96e - 12
GO:0009306~protein secretion	NIL	NIL	7.83e - 12
GO:0015628~protein secretion by the type II secretion system	NIL	NIL	2.67e - 07

Table 4.3 : Enriched GO Terms for database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae* at the P = 0.05 level and their Bonferroni adjusted P values.

H. pylori and *V. cholerae* database identified virulence gene lists share more GO terms that are statistically enriched at the P = 0.05 level, then either organism shares with *N. meningitidis* (Table 4.3). Majority of the GO terms for database identified virulence genes shared by *H. pylori* and *V. cholerae* involve the flagellum (GO:0019725), chemotaxis (GO:0006935) and motility (GO:0001539) (Figure 4.2, page 71, Table 4.3). Common GO terms for database identified virulence genes shared by all three bacterial species are transport of molecules (GO:0006812,

GO:0015301), protein secretion (GO:0009306) and outer membrane proteins (GO:0019867) (Figure 4.2, page 71). Unique GO terms for *N. meningitidis* database identified virulence genes include lipid and carbohydrate metabolism (GO:0044255, GO:0044262) and genes intrinsic to the cell membrane (GO:0031224) (Figure 4.2, page 71). Unique GO terms for database identified virulence genes for *H. pylori* and *V. cholerae* mainly pertain to metabolism and transporter activity (Figure 4.2, page 71).

Genes under positive selection for *H. pylori*, *N. meningitidis* and *V. cholerae* have 70 GO terms which could be assigned (Figure 4.1, page 68). Database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae* have a total of 68 GO terms which could be assigned (Figure 4.2, page 71). The intersection between the 70 GO terms for genes under positive selection (Figure 4.1, page 68) and the 68 GO terms for database identified virulence genes (Figure 4.2, page 71) is 27 (~39% of 68 and 70 GO terms), based on manual cross-checking of GO terms. Manual cross-checking of GO terms for genes under positive selection and database identified virulence genes reveal that some GO terms under positive selection within one bacterial organism, are the same as GO terms for database identified virulence genes within another organism. For example, the GO terms purine ribonucleotide binding (GO:0032555) and adenyl nucleotide binding (GO:0030554) are represented in genes under positive selection for *H. pylori* and *N. meningitidis* (Figure 4.1, page 68), the same GO terms are also present within database identified virulence genes for *V. cholerae* (Figure 4.2, page 71). Other GO terms for genes under positive selection within *H. pylori* include protein transmembrane transporter activity (GO:0008320) which is present on the list of *V. cholerae* database identified virulence genes (Figures 4.1, 4.2, pages 68, 71). Hence, it is possible that not all virulence processes have been characterized uniformly across all three bacterial species as a substantial overlap exists between genes under positive selection in one bacterial species that have the same GO terms as database identified virulence genes for another bacterial species. Alternatively, virulence processes for a certain bacterial species like *V. cholerae* are unique to that bacterial species and are not considered to be virulence processes within another bacterial species like *H. pylori*, thereby identifying species unique modes of pathogenicity.

Common GO Processes between Genes under Positive Selection and Database Identified Virulence Genes

To determine what GO terms are represented between genes under positive selection and database identified virulence genes, Entrez Gene IDs of database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae* were combined into a single file and uploaded to DAVID's web-interface as a background gene list. Similarly, a combined file of Entrez Gene IDs for genes under positive selection within *H. pylori*, *N. meningitidis* and *V. cholerae* was uploaded to DAVID's list manager as the foreground list. The foreground list (*H. pylori*, *N. meningitidis* and *V. cholerae* genes under positive selection) was compared to the background list (*H. pylori*, *N. meningitidis* and *V. cholerae* database identified virulence genes) to establish what GO terms are shared between the two gene list types. This form of gene list type comparison differs from the manual cross-checking of GO terms obtained for genes under positive selection and database identified virulence genes by use of a specific gene list type as the background instead of the bacterial species' genome. Results of the comparison for genes under positive selection and database identified virulence genes lists are presented overleaf, in Figure 4.3.

No statistically significant enrichment of GO terms at the P = 0.05 level could be found when comparing genes under positive selection and database identified virulence gene lists for *H. pylori*, *N. meningitidis* and *V. cholerae*. However, there is a high overlap between GO Terms for the two types of gene lists. The number of common GO terms between the gene lists under positive selection and database identified virulence genes, as determined by DAVID, is 22 (Figure 4.3, page 75). Of those 22 common GO terms (Figure 4.3, page 75), 15 are the same as the 27 Go terms obtained by manual cross-checking of GO terms from Figure 4.1 (page 68) and Figure 4.2 (page 71) while 7 and 12 GO terms are unique to each type of comparison. Hence, of the 27 common GO terms obtained by manual cross-checking of GO terms, a further 7 shared GO terms can be determined between genes under positive selection and database identified virulence genes.

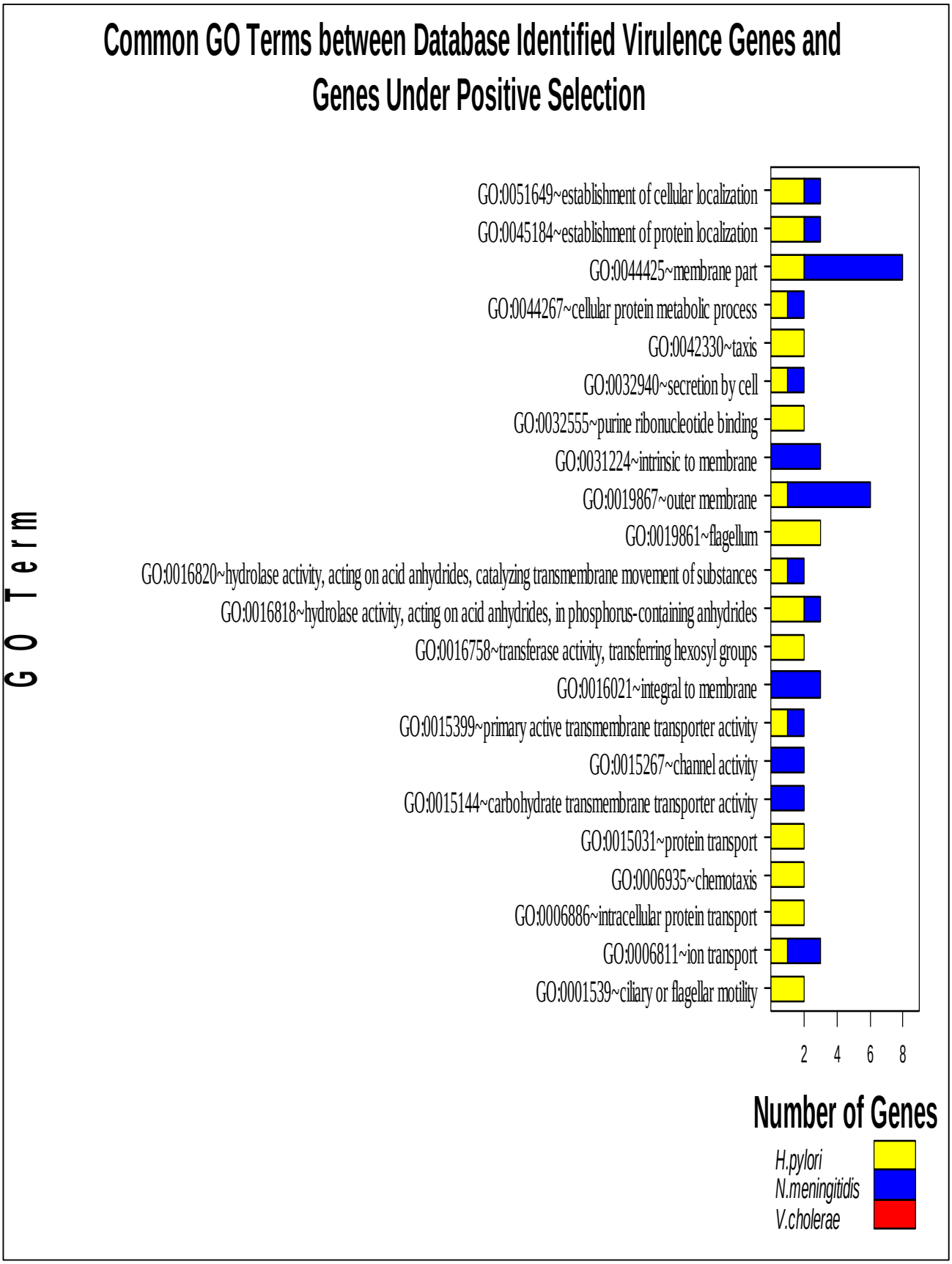


Figure 4.3 : GO terms that were found to be common between database identified virulence genes (background gene list) and genes under positive selection within *H. pylori*, *N. meningitidis* and *V. cholerae* (foreground gene list) using the DAVID system.

Differences in the number of GO terms obtained by manual cross-checking of GO terms across bacterial species and gene list comparisons performed using DAVID's list manager are mainly due to two factors. Firstly, the background gene lists differ, in the case of manual cross-checking of GO terms, the respective bacterial organism's genome was used as the background gene list, thereby providing more GO terms mapped to the bacterial organism's genome. In the instance of gene list comparisons done within DAVID, only the GO terms mapped to a bacterial organism's gene list are used, providing a much smaller subset of mapped GO terms. Secondly, manual cross-checking of GO term annotations is comparative in nature resulting in a transfer of annotations from a gene list type of one bacterial organism, to another gene list type of another bacterial organism [1,3,4,7].

The number of genes identified belonging to a specific GO term such as purine ribonucleotide binding (GO:0032555) also differ between Figures 4.3 (page 75), 4.1 (page 68) and 4.2 (page 71). When assaying for common GO terms between all genes under positive selection (foreground gene list) and all database identified virulence genes (background gene list), the combined gene lists submitted to DAVID are automatically partitioned into species specific gene lists, and used for GO term analysis that is species defined. Thus, similar GO terms for one gene list type can not obtained for another gene list belonging to a different bacterial species within the DAVID system, resulting in absence of GO terms for *V. cholerae* in Figure 4.3 (page 75), although GO annotations clearly exist (Figures 4.1, 4.2, pages 68 and 71).

Although there are no functionally enriched GO terms between genes under positive selection and database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae*, the two gene list types share a total of 34 GO terms based on comparative and direct gene list characterisation. Hence, genes under positive selection and database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae* do share a number of common biological processes based on GO annotations. In terms of genes under positive selection, *H. pylori* has more shared GO terms with *N. meningitidis* which is not surprising given the low number of genes under positive selection within *V. cholerae*. However, database identified virulence genes for *H. pylori* have more shared and statistically enriched GO terms in common with *V. cholerae* then either organism has compared to *N. meningitidis*, possibly due to *H. pylori* and *V. cholerae* adopting similar lifestyles in that they are enteric pathogens while *N. meningitidis* is not [14,28,85].

Clusters of Orthologous Proteins (COGs) Functional Annotation

Clusters of Orthologous Proteins (COGs) is an annotation strategy based on phylogeny and evolutionary orthology comprising of twenty-five alphabetical codes describing high level biological processes within a sequenced genome [102,112]. COGs was devised to characterize newly sequenced genomes, mostly prokaryotic, by identification of evolutionary orthologous gene sets between genomes representative of key phylogenetic lineages [102,112]. Genes from sequenced bacterial organisms are used to conduct reciprocal BLASTs to delineate orthologous relationships from paralogous ones [102,112]. Identified orthologous genes are functionally grouped into COG alphabetical categories based upon transfer of functional information available from well characterized genes [102,112]. COGs' stringent identification of orthologues provides phylogenetically and taxonomically well defined gene sets for use in annotating newly sequenced prokaryotic genomes and studying functional processes specific or common to different taxonomic lineages [15,102,112].

COG annotations differ from GO annotations in that not all genes within a given genome will be annotated as only orthologous gene sets from distinct phylogenetic lineages are classified by COGs whereas GO seeks to annotate all gene products within a given genome [102,109,110,112]. COGs also uses evolutionary orthology based on phylogeny for annotation purposes in contrast to functional homology and literature based methods GO annotation pipelines use [102,109,110,112]. However, GO has a significant element of ongoing human curation of annotated gene products while COGs is mostly automated in its annotation strategy [102,109,110,112]. GO was initially designed for annotation of eukaryotic genomes while COGs initial focus was prokaryotic genome annotation, although that is changing [112,170].

COG annotations for all genes in the genome, genes under positive selection and database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae* were obtained from their respective GenBank submitted genome annotation files. Statistical testing was conducted to determine if any alphabetical COG categories display a significant association for genes under positive selection and database identified virulence genes belonging to *H. pylori*, *N. meningitidis* and *V. cholerae*. Statistical testing was also performed to determine whether any alphabetical COG categories share an abundance of database identified virulence genes which are under positive selection for *H. pylori*, *N. meningitidis* and *V. cholerae*.

COG Functional Annotation of Bacterial Gene Lists

COG identifiers for genome genes, genes under positive selection and database identified virulence genes were parsed from their respective genome annotation files and sorted according to their alphabetical COG category codes. The total number of COG annotations for each microbial genome, genes under positive selection and database identified virulence genes placed in a COG category and the percentage coverage of each gene list type for *H. pylori*, *N. meningitidis* and *V. cholerae* are summarized below in Table 4.4.

Bacterial Organism	N^{o} of Genes in Genome placed in a COG Category (%)	N^{o} of Genes under Positive Selection placed in a COG Category (%)	N^{o} of Database Identified Virulence Genes placed in a COG Category (%)
H. pylori	978 (62 % of 1,576 genes)	178 (77 % of 230 genes)	118 (77 % of 153 genes)
N. meningitidis	1356 (66 % of 2,065 genes)	199 (91 % of 218 genes)	47 (69 % of 68 genes)
V. cholerae	2507 (65 % of 3,835 genes)	18 (78 % of 23 genes)	142 (85 % of 168 genes)

Table 4.4 : Number of genes for each gene list type which could be placed in a COG alphabetical functional category and percentage coverage of each gene list type for *H. pylori*, *N. meningitidis* and *V. cholerae*.

COG annotation coverage for genes under positive selection and database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae* (Table 4.4) is greater compared to GO annotation coverage of the same gene lists (Table 4.1, page 67, Table 4.2, page 70). GO annotation coverage does not surpass 50% of genes under positive selection (Table 4.1, page 67) while COG annotation coverage for the same gene lists is greater then 75% (Table 4.4). Similarly, higher coverage of database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae* is achieved by COGs compared to GO annotations (Table 4.4, Table 4.2, page 70). Apart from *V. cholerae*, COG annotation coverage for genes under positive selection is equal to or greater then COG annotation coverage for database identified virulence genes (Table 4.4). Counts of genome genes, genes under positive selection and database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae* belonging to each COG functional category are presented overleaf in Figure 4.4.

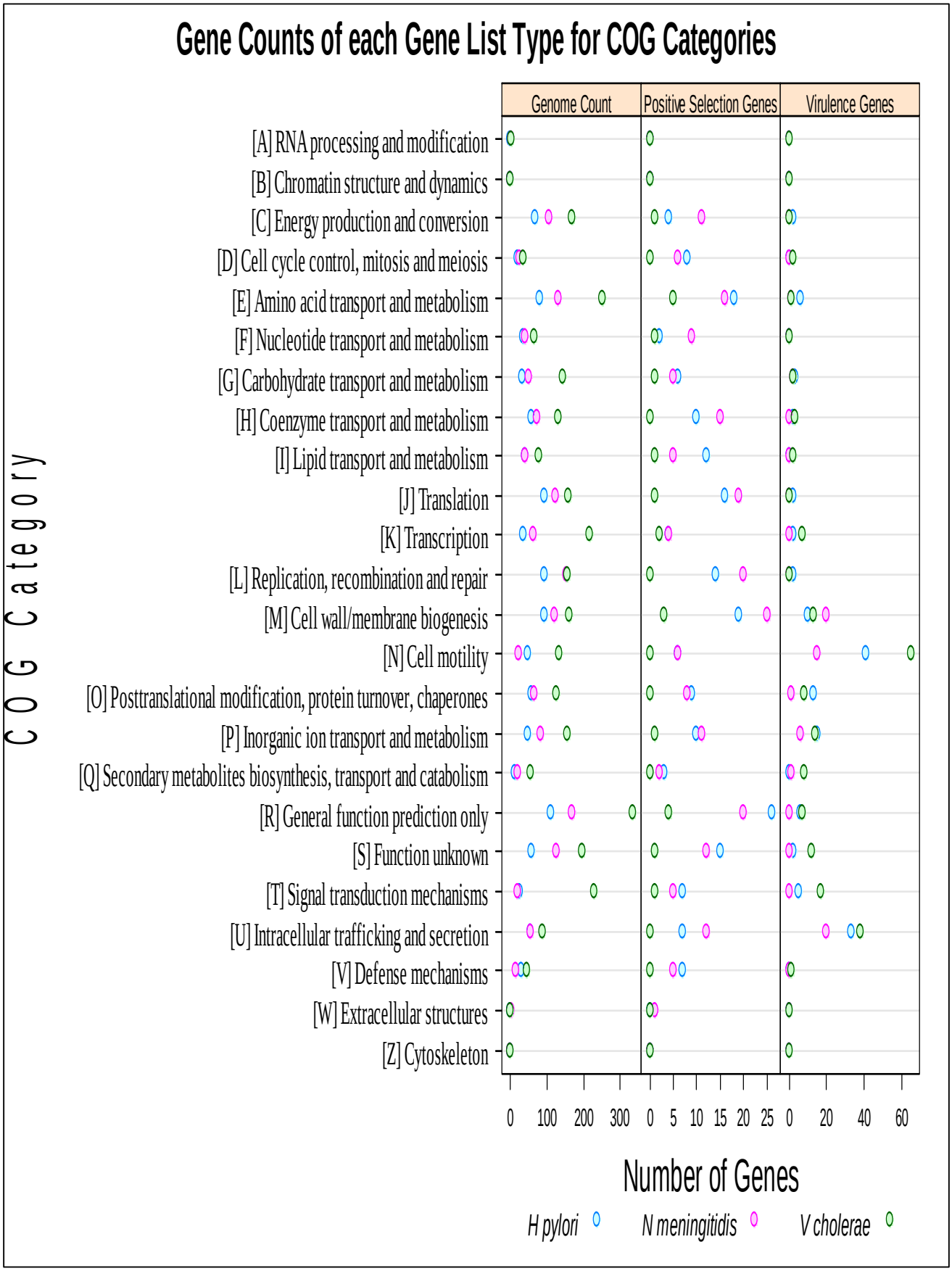


Figure 4.4 : Gene counts of genome genes, genes under positive selection and database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae* belonging to each alphabetical COG category.

Similar to GO annotations, genes under positive selection for *H. pylori*, *N. meningitidis* and *V. cholerae* are less well annotated compared to database identified virulence genes, although the percentage coverage of the two gene list types by COG annotations are similar (Table 4.4, page 78). The mean percentage coverage of genes under positive selection for *H. pylori*, *N. meningitidis* and *V. cholerae* by COG annotations is 82% (Table 4.4, page 78). The mean percentage coverage of database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae* by COG annotations is 77% (Table 4.4, page 78). If uninformative COG categories such as [R] general function prediction and [S] Function unknown are excluded (Figure 4.4, page 79), mean percentage COG annotation coverage for *H. pylori*, *N. meningitidis* and *V. cholerae* genes under positive selection drops to 64%. Similar calculations of mean percentage coverage by COG annotations for *H. pylori*, *N. meningitidis* and *V. cholerae* database identified virulence genes will result in 71% being annotated by COG categories other then [R] general function prediction and [S] Function unknown (Figure 4.4, page 79). Mean percentage coverage by COG annotations for genes under positive selection for *H. pylori*, *N. meningitidis* and *V. cholerae* (64%) is almost double compared to GO annotations of the same gene lists which is 34% (Table 4.1, page 67). The higher coverage of gene lists by COG annotations in comparison to GO annotations is due to COGs being established longer then the GO system and COGs' initial focus on prokaryotic genome annotation [102,109,110,112].

From the genome gene list, *V. cholerae* have the largest count for almost all of the COG alphabetical categories followed by *N. meningitidis* and *H. pylori*, possibly reflecting the order of gene content between the three bacterial organisms (Figure 4.4, page 79). *V. cholerae* has the smallest count of genes under positive selection belonging to each COG alphabetical category due to the small proportion of genes under positive selection compared to *N. meningitidis* and *H. pylori* (Figure 4.4, page 79). One way-classification Chi-Square tests with Holm's correction for multiple testing to determine if any COG alphabetical categories (Figure 4.4, page 79) have a statistically significant abundance of genes under positive selection for *H. pylori*, *N. meningitidis* and *V. cholerae* show no statistically significant abundance for any COG alphabetical category at the P = 0.05 level (Appendix C1). *N. meningitidis* genes under positive selection have the highest count for COG categories; [F] Nucleotide transport and metabolism, [H] Co-enzyme transport and metabolism, [J] Translation, [M] Cell wall/membrane biogenesis and [U] Intracellular trafficking and secretion compared to *H. pylori* and *V. cholerae* (Figure 4.4, page 79). The above mentioned COG categories for *N. meningitidis* genes under positive selection roughly mirror the GO terms found to be specific to *N. meningitidis* genes under positive selection (Figure 4.1, page 68). Analogous to GO specific terms for *H. pylori* genes under positive selection (Figure 4.1, page 68), COG categories [E] Amino acid transport and

metabolism, [G] Carbohydrate transport and metabolism, [I] Lipid transport and metabolism as well as [T] Signal transduction mechanisms have the highest counts of *H. pylori* genes under positive selection (Figure 4.4, page 79).

Database identified virulence genes show a statistically significant association for a number of alphabetical COG categories at the P = 0.05 level, after correction for multiple hypothesis testing. Alphabetical COG categories statistically enriched for *H. pylori, N. meningitidis* and *V. cholerae* database identified virulence genes are presented below in Table 4.5 and in Appendix C1.

COG Category	*H. pylori*	*N. meningitidis*	*V. cholerae*
[E] Amino acid transport and metabolism	NIL	NIL	0.003138
[J] Translation	0.03517	NIL	NIL
[K] Transcription	NIL	NIL	0.029953
[L] Replication, recombination and repair	0.03517	NIL	NIL
[M] Cell wall/membrane biogenesis	NIL	4.3056e – 15	NIL
[N] Cell motility	2.86e – 15	1.76e – 15	3.08e – 15
[P] Inorganic ion transport and metabolism	5.5748e – 4	NIL	NIL
[U] Intracellular trafficking and secretion	2.86e – 15	1.76e – 15	3.08e – 15

Table 4.5 : P values for Chi-Square tests conducted to determine which alphabetical COG categories have a statistical enrichment of database identified virulence genes for *H. pylori, N. meningitidis* and *V. cholerae*. The P values presented have been adjusted for multiple hypotheses testing using Holm's correction method.

COG categories enriched for database identified virulence genes common to all three bacterial species are [N] Cell motility and [U] Intracellular trafficking and secretion (Table 4.5), *V. cholerae* has the highest gene count for both the above mentioned COG categories followed by *H. pylori* and *N. meningitidis* respectively (Figure 4.4, page 79). Cell motility is generally considered to be a virulence factor in aiding movement towards hospitable environmental niches for colonization and procurement of nutrients [6,11,12,72]. The main form of contact and interaction a bacterial cell has with its environment is through the uptake of nutrients crossing through the cell membrane and secretion of waste products and colonisation enhancing factors such as toxins into its surroundings [6,11,12,72]. Species specific COG categories statistically enriched for *N. meningitidis* database identified virulence genes are [M] Cell wall/membrane biogenesis (Table 4.5). For *H. pylori* database identified virulence genes, COG categories [J] Translation, [L] Replication, recombination and repair and [P] Inorganic ion transport and metabolism are unique to this micro-organism and are statistically significantly enriched (Table 4.5). Species specific

COG categories statistically enriched for *V. cholerae* database identified virulence genes are [E] Amino acid transport and metabolism and [K] Transcription (Table 4.5, page 81).

Two way classification Chi-Square tests to determine whether any alphabetical COG categories have a statistically significant association between genes under positive selection and database identified virulence genes identify two COG categories for *N. meningitidis* at the P = 0.05 level (Appendix C1). The two COG categories enriched for *N. meningitidis* genes under positive selection and database identified virulence genes are [M] Cell wall/membrane biogenesis (adjusted P = 0.007115) and [N] Cell motility (adjusted P = 0.007115). *H. pylori* genes under positive selection and database identified virulence genes shows some statistically significant association for alphabetical COG categories [N] Cell motility and [J] Translation at the P = 0.05 mark, although this statistically significant association does not hold after correcting for multiple hypotheses testing (Appendix C1). For *V. cholerae*, the only gene pair that is under positive selection and is a database identified virulence gene (VC0837: VC0395_A0362; Figure 3.4, page 46) does not belong to any COG category.

Similar to GO annotations, no statistically significant functional grouping for genes under positive selection in *H. pylori, N. meningitidis* and *V. cholerae* could be found using COG alphabetical categories, even though the gene list coverage is almost double that of GO annotations. Common and species unique COG alphabetical categories are found to be enriched for database identified virulence genes indicating some commonalities as well as differences between characterised virulence processes for *H. pylori, N. meningitidis* and *V. cholerae* exist (Table 4.5, page 81).

Subcellular Localization of Bacterial Proteins

Bacterial contact, colonization and infection of a host requires interactions between a bacterial cell and a host which may occur through use of surface exposed proteins such as adhesins, as in the case of *N. meningitidis*, secretion of bio-molecules like cholera toxin or urease from *V. cholerae* and *H. pylori* respectively, transport of survival factors or movement to hospitable environmental niches through chemotaxis and flagellum motility [6,11,70,71,74,81]. These pathogen-host interactions are mediated by proteins and factors that have well defined sub-cellular localisation sites. Elucidation of subcellular sites for an active gene product compliments functional annotation of bacterial gene lists as both the location and function of an active gene product are intertwined, evolutionary conserved attributes [113,115,171]. Knowledge of subcellular sites for an active gene product provides insights as to what evolutionary pressures may operate on a particular gene, e.g. surface exposed proteins are known to be under diversifying selection to avoid recognition by the immune system [6,55,162].

GO and COG functional annotation of genes under positive selection and database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae* indicate a number of genes involved in transport, metabolism, secretion, cell membrane structure and motility amongst others (Figures 4.1, 4.2, 4.3 and 4.4, pages 68, 71, 75 and 79 respectively). GO and COGs are inherently biologically process driven functional annotation systems providing little information as to the precise subcellular site proteins of genes under positive selection and database identified virulence genes operate within a cell [102,109,110,112].

Subcellular localization prediction of proteins is amenable to bioinformatic analysis as cues governing the subcellular compartments where a protein is active are usually embedded within the synthesized protein [113,115,172]. Possible encoded signals within a synthesized protein include distinctive amino acid profiles, alpha-helical secondary structures integral to cytoplasmic membrane proteins, beta-barrel secondary structures for gram negative bacterial cells that extend across the outer membrane, or N-terminal hydrophobic signal peptides which chauffer the export of a protein from the cytoplasm [115]. Numerous bioinformatics systems predicting protein subcellular localization have been developed by different groups concentrating on various organisms and proteins using an assortment of approaches ranging from Hidden Markov Models to support vector machine learning methods [113-115]. A popular bioinformatics system developed for sub-cellular localization prediction within bacterial organisms is PSORTb which relies on multiple sequence features as classifiers and is based upon the PSORT I method, originally implemented by Nakai and Kanehisa [113-115,172].

PSORTb Subcellular Localization Prediction of Bacterial Proteins

PSORTb (version 2) uses six independent bioinformatics software modules for subcellular localization prediction and integrates scores from these six modules using a Bayesian network statistical approach to generate a final subcellular site prediction for a given protein [114-116]. PSORTb uses SCL-BLAST for homology based prediction, an outer membrane motif identification module, Hidden Markov Model for Topology prediction (HMMTOP) for transmembrane helix prediction and a signal peptide identification module also based on a Hidden Markov Model [113-116]. A motif and a profile matching module both trained on PROSITE data are also used, the latter differing from the former by inclusion of a position specific weighted matrix [113-116]. PSORTb version 2 differs from version 1.1 by implementation of a support vector machine learning algorithm which uses recurring sub sequences within a protein to delineate between pre-defined subcellular localization sites, resulting in increased precision (specificity) and bacterial proteome coverage [113-115]. Each module outputs a probability value that is normalised by a factor 10 multiplication, and passed onto a Bayesian network to generate a final prediction score for a particular subcellular site [113-115]. The Bayesian network appraises score distributions for each subcellular site and predicts a site depending on the score distributions' biasness for a particular site [113-115]. The scoring scheme for PSORTb ranges from 0 to 10, scores greater then a recommended threshold of 7.5 are used to generate the final subcellular localization prediction site with a high precision (96%) [114,115].

The subcellular localization prediction sites defined by PSORT for gram negative bacteria like *H. pylori*, *N. meningitidis* and *V. cholerae* are Cytoplasmic, Cytoplasmic Membrane, Periplasmic, Outer Membrane, Extra Cellular and Unknown. Cytoplasmic Membrane is the region between the cytoplasm and the inner cell wall and is usually associated with the movement of vesicles for the export of cell products to the external environment (exocytosis) or intake of products from the external environment into the bacterial cell (endocytosis). The Periplasmic interval (interval between the Cytoplasmic membrane and Outer Membrane for gram negative bacteria) is of great interest as proteins within this region form components of Type II and III secretion systems and two-component sensors e.g. histadine-kinase sensors which facilitate interaction and movement of bio-molecules between internal and external environments of a bacterial cell. Outer Membrane proteins are located on the outer bacterial cell membrane like adhesins while Extra Cellular proteins are secreted by the bacterial cell into the environment e.g. siderophores. A protein may have multiple localization sites within a bacterial cell, an equal distribution of scores for a number subcellular localization sites or not reach the 7.5 confidence score threshold, in such instances they are classified as Unknown [113-116].

PSORTb Subcellular Localization Prediction of Bacterial Gene Lists

Subcellular localization predictions of database identified virulence genes, genes under positive selection and all genes within the genome for *H. pylori*, *N. meningitidis* and *V. cholerae* were obtained from cPSORTdb (version 2) [114,116]. Bacterial gene products that did not make the recommended PSORTb threshold score of greater then 7.5 were excluded from subsequent analysis. In all instances, bacterial proteins with a final score below the 7.5 threshold were classified by PSORTb as Unknown. In terms of coverage, PSORTb will place every protein for a gram negative bacterial proteome into one of six subcellular localization categories and if the category Unknown is included, gene list coverage would reach 100%. However, as the category Unknown is uninformative and all genes not making the 7.5 cut-off for a subcellular localization site fall into this category, gene counts for Unknown were excluded from calculating percentage coverage of gene lists and statistical analysis. Gene list coverage by PSORTb for database identified virulence genes, genes under positive selection and all genomic genes for *H. pylori*, *N. meningitidis* and *V. cholerae* are presented below in Table 4.6.

Bacterial Organism	Nº of Genome Genes belonging to a PSORTb subcellular localization site (%)	Nº of Genes under Positive Selection belonging to a PSORTb subcellular localization site (%)	Nº of Database Identified Virulence Genes belonging to a PSORTb subcellular localization site (%)
H. pylori	861 (55 % of 1,576 genes)	151 (66 % of 230 genes)	94 (61 % of 153 genes)
N. meningitidis	1103 (53 % of 2,065 genes)	131 (60 % of 218 genes)	33 (49 % of 68 genes)
V. cholerae	2162 (56 % of 3,835 genes)	14 (61 % of 23 genes)	95 (57 % of 168 genes)

Table 4.6 : Number of genes for *H. pylori*, *N. meningitidis* and *V. cholerae* genome, database identified virulence genes and genes under positive selection that pass the recommended 7.5 PSORTb score threshold belonging to a subcellular localization site other then Unknown and percentage coverage of those gene list types.

Percentage coverage of genes under positive selection by PSORTb is higher then percentage coverage of database identified virulence genes in all instances (Table 4.6). The number of *H. pylori*, *N. meningitidis* and *V. cholerae* genes belonging to each PSORTb subcellular localization site for all gene list types is presented overleaf in Figure 4.5.

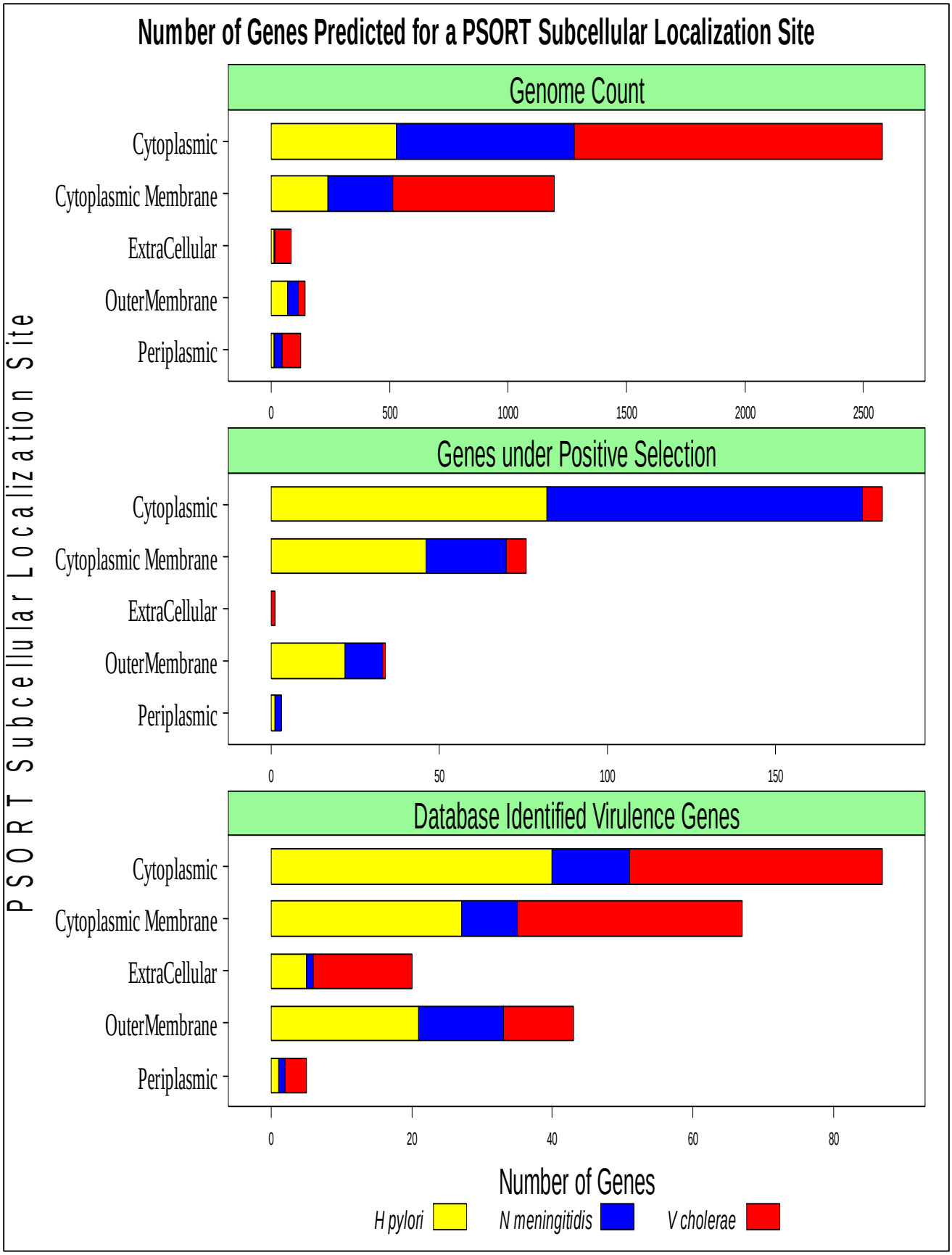


Figure 4.5 : Gene counts for genome genes, genes under positive selection and database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae* belonging to a specific PSORTb subcellular localization site.

The majority of genomic genes localize within the Cytoplasmic and the Cytoplasmic Membrane regions for *H. pylori*, *N. meningitidis* and *V. cholerae* (Figure 4.5, page 86). *H. pylori* has a larger proportion of genomic genes present on the Outer Membrane compared to *N. meningitidis* and *V. cholerae*, *V. cholerae* has more Extra Cellular proteins then either *H. pylori* or *N. meningitidis* (Figure 4.5, page 86). The preponderance of genes under positive selection for *H. pylori*, *N. meningitidis* and *V. cholerae* localize to the Cytoplasmic and Cytoplasmic Membrane locations (Figure 4.5, page 86). Compared to *N. meningitidis* and *V. cholerae*, *H. pylori* has more genes under positive selection localized to the Outer Membrane and only *V. cholerae* does not have any genes under positive selection predicted to localize within the Periplasmic interval (Figure 4.5, page 86). One-way classification Chi-Square tests reveal a statistically significant association for the PSORTb Outer Membrane subcellular localization site for *H. pylori* (Adjusted P = 0.00768) and *N. meningitidis* (Adjusted P = 0.04582) genes under positive selection. The full results of the Chi-Square tests conducted are available in Table C2A, Appendix C2.

A large proportion of database identified virulence genes for *V. cholerae* are predicted to be Extra Cellular and Periplasmic in comparison to *H. pylori* and *N. meningitidis* (Figure 4.5, page 86). Both *H. pylori* and *N. meningitidis* have more database identified virulence genes associated with the Outer Membrane compared to *V. cholerae* (Figure 4.5, page 86). A number of PSORTb subcellular localization sites that are statistically enriched for database identified virulence genes together with their adjusted P values for *H. pylori*, *N. meningitidis* and *V. cholerae* are presented below in Table 4.7.

PSORTb Subcellular Localization Site	**Bacterial Organism and Adjusted P Values**		
	H. pylori	***N. meningitidis***	***V. cholerae***
Cytoplasmic	0.000428	0.000145	2.571e – 05
Cytoplasmic Membrane	1	1	1
Periplasmic	1	1	1
Outer Membrane	8.99e – 07	3.285e – 09	5.695e – 07
Extra Cellular	0.0178	0.3438	2.6784e – 06

Table 4.7 : P values of Chi-Square tests with Holm's correction for multiple hypothesis testing for database identified virulence genes association with specific subcellular localization sites for *H. pylori*, *N. meningitidis* and *V. cholerae*. Full results of statistical testing for all PSORTb categories for database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae* is contained in Table C2B, Appendix C2.

Database identified virulence gene products show a statistically significant association at the P = 0.05 level for subcellular localization sites Cytoplasmic and Outer Membrane (Table 4.7, page 87). Across all three bacterial species, database identified virulence gene products have the strongest association, based on P values, for being located on the Outer Membrane (Table 4.7, page 87) compared to all other PSORTb subcellular localization sites. Gene products for database identified virulence genes do not have a statistical association at the P = 0.05 level for being localized at the Cytoplasmic membrane or Periplasmic interval across all three bacterial species (Table 4.7, page 87).

Two-way classification Chi-Square tests to determine if there is an association between database identified virulence genes that are under positive selection for a PSORTb site indicate that no particular PSORTb subcellular localization site has a statistically significant association at the P = 0.05 level (Table C2C, Appendix C2, page 87). Hence, although gene products for genes under positive selection and database identified virulence genes show a statistical enrichment at the P = 0.05 level for localization to the Outer Membrane when tested individually, the overlap between the two gene lists does not appear to be statistically significant. The absence of strong statistical association between database identified virulence genes and genes under positive selection may indicate that not all Outer Membrane genes for *H. pylori* or *N. meningitidis* are virulence genes, under which a large proportion of genes under positive selection belong to. Alternatively, the absence of strong statistical association between Outer Membrane genes undergoing positive selection and database identified virulence genes may reflect an incomplete virulence gene characterization repertoire for the three bacterial species as functional gene annotation is an ongoing process [6,82,129].

Chapter 5 : Metabolic Pathways

Introduction

Biological cells are complex systems whose phenotypic behaviour is governed by expression of its constituent genotype. As in all complex systems, components of the system (in this instance genes and their products) do not function in isolation. Instead, genes and their products function in an organised, concerted manner creating biological networks of various complexities based upon interconnectivity [173-176]. Biological networks form a natural extension for leveraging high throughput genomics data to decipher the organisation and information flow within complex, biological systems [36,173-179]. Biological networks can be defined at a variety of levels ranging from physical gene clustering in a genome, gene expression networks, signalling networks, transcriptional regulatory networks, neurological networks, biochemical metabolic networks and protein-protein interaction maps [36,173-178]. Each type of biological network differs in the experimental datasets used to establish the network, the topologies and interconnectivity of the resulting networks and the species used to derive the network [36,173-178]. A common goal of most biological network studies is the contextualisation of biological information within a systematic framework allowing the grouping and characterisation of similar, interacting components under pre-specified biological conditions to provide an integrative view of how a complex biological system functions [36,173-179].

Network modelling of biological systems demonstrate they share similarities with other types of networks from non-biological complex systems in that they are non-random and are scale-free [36,173-179]. Networks are modelled using graph theory within which members of the network are termed as “nodes” and links between nodes are termed as “edges” [36,173-179]. Edges between nodes in a network may be directed whereby there is a unidirectional process flow from one node to another or they may be undirected whereby a bi-directional process flow between nodes occurs [36,173-179]. Random network models assume the number of edges a node has to other nodes within the network follows a Poisson distribution with most nodes having a similar number of edges to the average number of edges for all nodes within the network [173-175,177]. Empirical studies of networks in complex systems show these networks do not follow a Poisson distribution but instead are scale free whereby the distribution of edges for a given node can be described with a power law and in contrast to random network models, are non-uniform [36,173-179]. Scale free networks are characterised by having a few central nodes known as hubs that connect to numerous other nodes within the network and unlike random network models based on a Poisson distribution, a single node and the number of edges it contains can not be used to typify other nodes within the network [36,173-179].

Using protein-protein interaction data, Fraser *et al* showed interacting proteins within eukaryotic organisms co-evolve and genes involved in the same pathways have similar rates of evolution [173,180]. If interacting genes have similar rates of evolution, can genes under positive selection for *H. pylori*, *N. meningitidis* and *V. cholerae* be segregated into similar metabolic pathways? In addition, do database identified virulence genes also function as modules within any metabolic pathways and do these pathways overlap with genes under positive selection for *H. pylori*, *N. meningitidis* and *V. cholerae*?

Metabolic Pathways and Databases

Analyses of metabolic networks which places metabolites as nodes and edges as reactions catalysed by enzymes have established they are scale free networks [173,175,177-179]. The majority of metabolic substrates are engaged in a few biochemical reactions while the minority form hubs connected to numerous biochemical reactions [173,175,177-179]. Within metabolic networks, there are modules or pathways of interacting nodes which are coupled together to accomplish specific biochemical reactions [173,175,177-179]. Edges between nodes in metabolic networks and pathways are generally directed as the end product of one metabolic reaction forms the substrate for another [173,175,177-179]. Construction of metabolic pathways and networks is based upon four levels of varying detail [36,178,179]. Within the first level of detail, the metabolic pathway, gene products involved within the pathway, metabolites, substrates and the type of reaction catalysed are defined [36,178,179]. The second level of detail defines the charged chemical formulae of metabolites catalysed by enzymatic reactions within the pathway [36,178,179]. Balancing of charges within the chemical equation formulae of enzymatic reactions provides the exact molecular proportions of metabolites that flow through the metabolic pathway providing the third level of detail while the fourth level of detail is represented by the directionality of substrate flow through the metabolic pathway [36,178,179].

Two popular, well documented pathway databases which provide collections of metabolic pathways are the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the BioCyc collection of individual organism specific databases [36,117,118,120,125,181]. BioCyc and KEGG aim to connect genomic information with molecular interaction and functional information of gene products within the context of biological pathways [117,118,120,125,181]. Both KEGG and BioCyc have genome sequence information and gene product annotations linked to substrate specificity, metabolite formulae, stoichiometry, enzymatic pathway directionality as well as incorporate literature and experimental evidence for curation of their metabolic pathways [36,117,118,120,125,181]. Methods used to construct the KEGG and BioCyc databases differ although both initially begin with well defined maps of central metabolism common to all organisms [36,117,118,120,125,181].

KEGG contains manually constructed maps of metabolic networks upon which enzymatic gene products with known functions are overlaid to form a reference map [120,121,125]. Homology searches of reference map genes are conducted to identify orthologous genes from other organisms which are assigned their relevant Enzyme Commission (EC) numbers [120,121,125]. The EC numbers are matched to EC numbers in the reference maps and used to computationally generate metabolic pathway maps that are organism specific [120,121,125]. BioCyc is a collection of databases each housing individual organism specific metabolic pathways and a heavily curated reference database of metabolic pathways that is organism independent (MetaCyc) [117,118]. Organism genome information for all databases in BioCyc is obtained from publically available sources and metabolic pathway information is incorporated through the use of the PathoLogic programme [117,118,182]. The PathoLogic programme identifies enzymes and their EC numbers from the genome under annotation, determines their corresponding EC counterparts from the reference database of metabolic pathways (MetaCyc) and assigns appropriate pathway information to that genome [117,118,182]. In the current state, the organism specific pathway database created is termed as "Tier 3" whose information is solely based on metabolic pathways obtained from MetaCyc and have not undergone any review [117,118]. Tier 1 databases within the BioCyc collection, like MetaCyc and EcoCyc, have undergone years of manual curation and are constantly being updated to add experimentally determined pathways [117,118]. Tier 2 databases are similar to Tier 3 databases in that they are computationally generated, but differ by having some level of manual curation to establish the fidelity of the pathway predictions contained within them [117,118].

Metabolic pathways of genes under positive selection and database identified virulence genes for *H. pylori*, *N. meningitidis* and *V. cholerae* were investigated using KEGG and BioCyc's organism specific databases and web-based Pathway Tools [117,118,120,125,182]. BioCyc databases for *H. pylori* (strain 26695) and *V. cholerae* (strain N16961) are classified as Tier 2 databases while *N. meningitidis* (strains Z2491 and MC58) are categorised as Tier 3 databases[•]. Metabolic pathway annotations from KEGG for genes under positive selection and database identified virulence genes were obtained using DAVID. DAVID's gene list manager enables one to interrogate KEGG pathway annotations to obtain pathways that are common between genes under positive selection and database identified virulence genes.

[•] http://biocyc.org/biocyc-pgdb-list.shtml

H. pylori BioCyc Metabolic Pathways

H. pylori accessions for genes under positive selection and database identified virulence genes were used to query BioCyc's (Version 13.5) *H. pylori* (strain 26695) database for metabolic pathway annotations selecting a threshold of 1 for individual metabolic pathways. In terms of coverage, 54 (23% of 230) *H. pylori* genes under positive selection and 22 (14% of 153) *H. pylori* database identified virulence genes can be placed in a metabolic pathway by BioCyc. BioCyc coverage of *H. pylori* gene lists may not be an accurate measure of annotation depth as not all genes within the *H. pylori* genome are involved in metabolic pathways e.g. adhesins or flagellum. However, there are "pathway holes" in some metabolic pathways for which the corresponding genes remain to be elucidated, it is feasible that some genes on the gene lists will be assigned to these pathway holes once experimental evidence for their function becomes available [183,184]. Metabolic pathways obtained from BioCyc for *H. pylori* genes under positive selection and database identified virulence genes are presented overleaf in Table 5.1.

The 54 *H. pylori* genes under positive selection belong to thirty-nine different metabolic pathways, the 22 *H. pylori* database identified virulence genes map to fifteen different metabolic pathways within BioCyc (Table 5.1, page 93). The combined number of metabolic pathways represented by *H. pylori* genes under positive selection and database identified virulence genes is fifty-four. Of these fifty-four metabolic pathways, six are shared between genes under positive selection and database identified virulence genes (Table 5.1, page 93). As there are 22 genes in common between *H. pylori* positive selection and database identified virulence gene lists (Figure 3.4, page 46), some overlap in metabolic pathways between the two gene lists is not unexpected. Of the 22 *H. pylori* database identified genes under positive selection, 4 genes map to four different BioCyc metabolic pathways; Arginine Degradation VI (Arginase 2 pathway), Arginine Degradation VII, Enterobacterial Common Antigen Biosynthesis and Urea Degradation II (Table 5.1, page 93). A further two BioCyc metabolic pathways for *H. pylori* genes under positive selection and database identified virulence genes that is not based on the overlap of the two gene lists are Colanic Acid Building Blocks Biosynthesis and GDP-Mannose Metabolism (Table 5.1, page 93).

Row N^{O}	BioCyc Metabolic Pathways of *H. pylori* Genes under Positive Selection	BioCyc Metabolic Pathways of *H. pylori* Database Identified Virulence Genes	BioCyc Metabolic Pathways Common to *H. pylori* Genes under Positive Selection and Database Identified Virulence Genes
1	Arginine Degradation VI (Arginase 2 pathway)	Arginine Degradation VI (Arginase 2 pathway)	Arginine Degradation VI (Arginase 2 pathway)
2	Arginine Degradation VII	Arginine Degradation VII	Arginine Degradation VII
3	Colanic Acid Building Blocks Biosynthesis	Colanic Acid Building Blocks Biosynthesis	Colanic Acid Building Blocks Biosynthesis
4	Enterobacterial Common Antigen Biosynthesis	Enterobacterial Common Antigen Biosynthesis	Enterobacterial Common Antigen Biosynthesis
5	GDP-Mannose Metabolism	GDP-Mannose Metabolism	GDP-Mannose Metabolism
6	Urea Degradation II	Urea Degradation II	Urea Degradation II
7	Cardiolipin Biosynthesis I	Glutamate Biosynthesis II	
8	CDP Diacylglycerol Biosynthesis II	Glutamate Biosynthesis III	
9	Chorismate Biosynthesis	Glutamate Degradation I	
10	Fatty Acid Biosynthesis Initiation II	Glutamine Biosynthesis I	
11	Fatty Acid Biosynthesis Initiation III	Glutamine Biosynthesis II	
12	Fatty Acid Elongation (saturated)	Lipid-A-precursor Biosynthesis	
13	Fatty Fcid ß	Oxidation I Proline Biosynthesis II	
14	FormylTHF Biosynthesis I	Thioredoxin Pathway	
15	Gluconeogenesis	UDP Glucose Conversion	
16	Glyceraldehyde 3-Phosphate Degradation		
17	Glycolysis I		
18	Glycolysis IV (plant cytosol)		
19	Homoserine Biosynthesis		
20	Lysine Biosynthesis I		
21	NAD Biosynthesis I (from Aspartate)		
22	Peptidoglycan Biosynthesis I		
23	Proline Degradation II		
24	Purine nucleotides *de novo* Biosynthesis I		
25	Quinate Degradation		
26	Respiration (anaerobic)		
27	Salvage Pathways of Purine and Pyrimidine nucleotides		
28	Serine Biosynthesis		
29	Superpathway of Fatty acid Biosynthesis Initiation (E. coli)		
30	TCA Cycle Variation II		
31	Tetrahydrofolate Biosynthesis I		
32	Tetrapyrrole Biosynthesis I		
33	tRNA Charging Pathway		
34	Tryptophan Biosynthesis		
35	UDP-N-acetyl-D-glucosamine Biosynthesis I		
36	Thiamine Biosynthesis		

Table 5.1 : BioCyc metabolic pathways determined for *H. pylori* genes under positive selection, database identified virulence genes and metabolic pathways common to both gene lists.

A cellular overview of *H. pylori*'s BioCyc metabolic pathways for genes under positive selection, database identified virulence genes and metabolic pathways shared between the two gene list types is presented overleaf in Figure 5.1. The Urea Degradation II pathway containing database identified virulence genes under positive selection (Figure 5.1, Box 1, page 95) is catalysed by the enzyme urease within *H. pylori* and is vital for the organism's survival and colonisation within the acidic gastric environment and maintenance of pH homeostasis within the bacterial cell [76,81,185,186]. *H. pylori* produces ammonia by the urea degradation pathway to buffer the pH of its micro-environment within the stomach and urease enzyme synthesis within *H. pylori* is prolific contributing to approximately 10% of the total protein content [76,81,185-187].

Arginine Degradation VI and VII (Figure 5.1, Boxes 2 and 3 respectively, page 95) pathways are variant pathways of the arginine degradation super pathway producing different end product metabolites. Arginine degradation pathways are used by bacteria which have limited sources of free nitrogen, L-arginine is hydrolysed by the enzyme arginase (HP1399 - rocF) to create urea and L-ornithine [117,118]. *H. pylori* has altered its physiology to exploit amino acids like arginine as a carbon source instead of carbohydrates due to the unavailability of free nitrogen and the constant pressure on the bacterium to raise the pH of its micro-environment within its gastric niche [176]. Two of four Arginine Degradation VII pathway genes within *H. pylori* are under positive selection, the first gene in the Arginine Degradation VII pathway is a database identified virulence gene under positive selection (Figure 5.1, Box 3, page 95).

Enterobacterial Common Antigen (ECA) (Figure 5.1, Box 4, page 95) is an outer membrane glycolipid found in enterobacterial members from the proteobacterial lineage [188-190]. Deficiencies in the ECA pathway are known to reduce motility within *H. pylori*, *E. coli* and *Serratia marcescens*, motility is an essential factor for *H. pylori* chemo-taxis and colonisation of the gastric niche [76,188-190].

The GDP-Mannose metabolic pathway (Figure 5.1, Box 5, page 95) is a common pathway present in eukaryotic and prokaryotic organisms [191]. GDP-α-D-mannose is converted to GDP-L-fucose which is a component of the lipopolysaccharide (LPS) O-antigen of *H. pylori* [191,192]. Fucosylation of LPS O-chains is important in mediating bacterial cell adhesion to host cells and enables Lewis antigen mimicry which is directly linked to gastritis [193-195].

Colanic acid or M antigen (Figure 5.1, Box 6, page 95) is an exopolysaccharide which also contains GDP-L-fucose and is common to enterobacterial species [196-198]. Within *E. coli*, colonic acid synthesis has no known role in bacterial virulence but is postulated to prevent desiccation and enable survival of the bacterium when not host-associated [196-198]. Within *H pylori*, genes under positive selection and database identified virulence genes are part of the colanic acid biosynthesis pathway (Figure 5.1, Box 6, page 95).

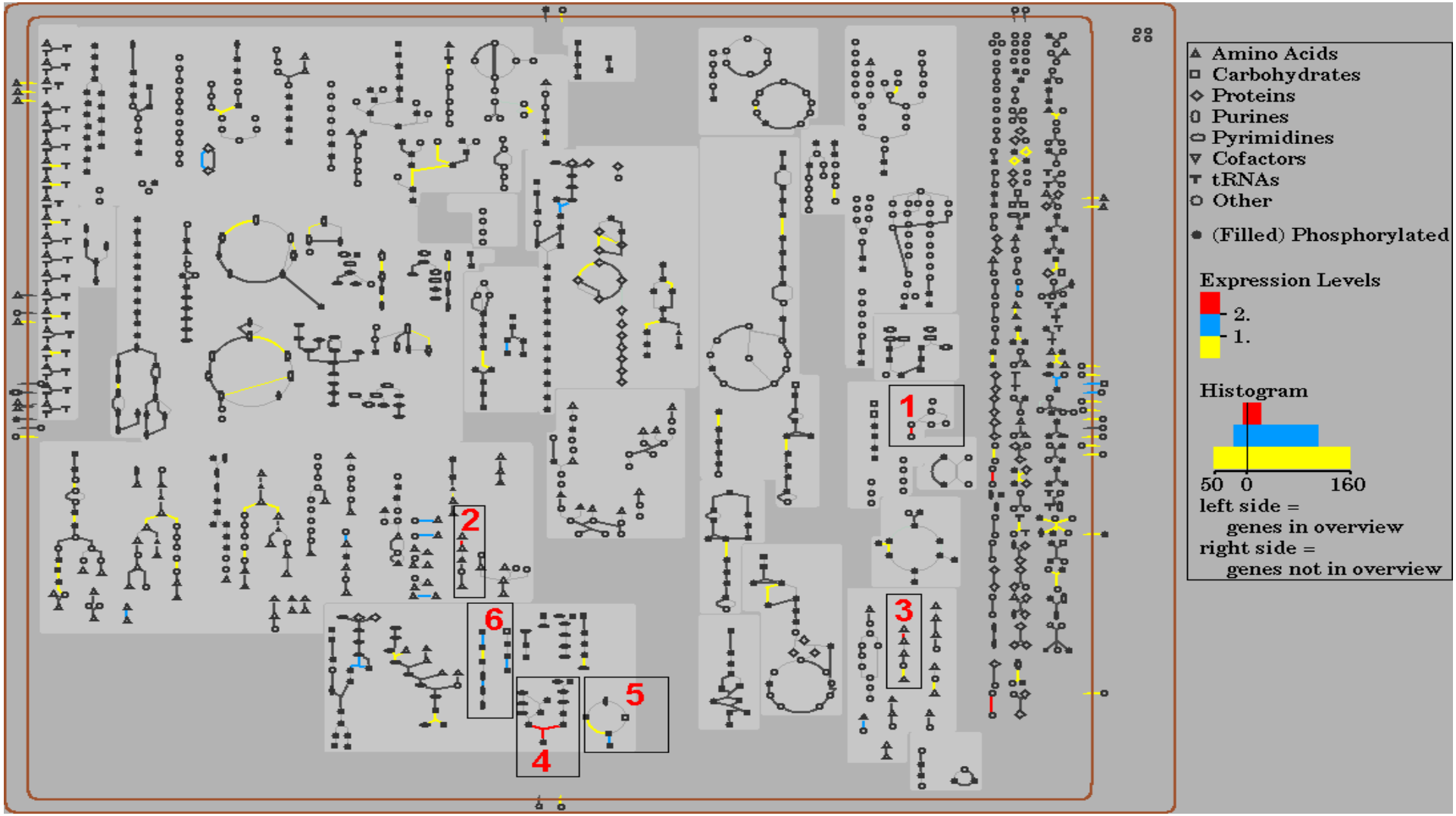


Figure 5.1 : BioCyc metabolic pathways for *H. pylori* genes under positive selection (yellow lines), database identified virulence genes (blue lines) and pathways common to both gene lists (red lines). Six common metabolic pathways between *H. pylori* genes under positive selection and database identified virulence genes are highlighted in numbered boxes; 1 – Urea Degradation II, 2 and 3 – Arginine Degradation VI and VII respectively, 4 – Enterobacterial Common Antigen Biosynthesis, 5 – GDP-Mannonse Metabolism, 6 – Colanic Acid Building Blocks Biosynthesis (Table 5.1, page 93)

KEGG Pathways of H. pylori Genes

KEGG pathways for *H. pylori* genes under positive selection and database identified virulence genes were obtained using DAVID [111,121,125]. DAVID was chosen for KEGG pathway determination for reasons including ease of acquiring KEGG pathway annotations for complete gene lists and statistical enrichment analyses for genes belonging to a KEGG pathway. Additionally, DAVID provides shared KEGG pathways between gene lists by using a specific gene list as a foreground and another as a background gene list. *H. pylori* accessions for genes under positive selection and database identified virulence genes were uploaded to DAVID's gene list manager and an EASE score of 0.5 was selected to maximise gene list coverage and maintain consistency across comparisons.

From 230 *H. pylori* genes identified as being under positive selection, 19% (44 of 230 genes) could be placed into fifteen different KEGG pathways. None of the KEGG pathways show a statistical enrichment for genes under positive selection at the P = 0.05 level after correction for multiple hypotheses testing. The fifteen KEGG pathways *H. pylori* genes under positive selection are mapped to is presented below in Table 5.2.

KEGG Pathway Reference	KEGG Pathway Description
hpy00330	Arginine and proline metabolism
hpy00300	Lysine biosynthesis
hpy00970	Aminoacyl-tRNA biosynthesis
hpy00550	Peptidoglycan biosynthesis
hpy05120	Epithelial cell signaling in Helicobacter pylori infection
hpy00790	Folate biosynthesis
hpy00450	Selenoamino acid metabolism
hpy00770	Pantothenate and CoA biosynthesis
hpy03060	Protein export
hpy00061	Fatty acid biosynthesis
hpy03030	DNA polymerase
hpy00400	Phenylalanine, tyrosine and tryptophan biosynthesis
hpy02030	Bacterial chemotaxis – General
hpy02010	ABC transporters – General
hpy02020	Two-component system – General

Table 5.2 : Description of KEGG pathways and pathway identifiers mapped to *H. pylori* genes under positive selection obtained via DAVID.

KEGG pathway descriptions and maps are more general and have less detail compared to BioCyc's metabolic pathways which are discrete and have well defined organism specific metabolic pathways [118]

Of 153 *H. pylori* database identified virulence genes, 54% (83 of 153 genes) could be placed into thirteen different KEGG pathways. Of these thirteen KEGG pathways, five show a statistical enrichment for database identified virulence genes at the P = 0.05 interval after stringent Bonferroni corrections for multiple hypothesis testing. The fourteen KEGG pathways for *H. pylori* database identified virulence genes and P values obtained from DAVID are summarised below in Table 5.3.

KEGG Pathway Reference	**KEGG Pathway Description**	**Adjusted P Value**
hpy05120	Epithelial cell signalling in *Helicobacter pylori* infection	6.59e-30
hpy02040	Flagellar assembly	1.15e-27
hpy03080	Type IV secretion system	8.81e-10
hpy03070	Type III secretion system	1.28e-06
hpy02030	Bacterial chemotaxis – General	2.53e-05
hpy02020	Two-component system – General	0.0918
hpy00220	Urea cycle and metabolism of amino groups	0.987
hpy00643	Styrene degradation	1
hpy00632	Benzoate degradation via CoA ligation	1
hpy00791	Atrazine degradation	1
hpy00460	Cyanoamino acid metabolism	1
hpy00051	Fructose and mannose metabolism	1
hpy00360	Phenylalanine metabolism	1

Table 5.3 : KEGG pathway identifiers and descriptions for *H. pylori* database identified virulence genes. Bonferroni adjusted P values show statistical enrichment of *H. pylori* database identified virulence genes for five out of thirteen KEGG pathways at the P = 0.05 level.

Epithelial cell signalling during *H. pylori* infection, flagellar assembly, chemotaxis and secretion systems (type III and IV) are KEGG pathways statistically enriched for *H. pylori* database identified virulence genes at the P = 0.05 level (Table 5.3). Examination of the five KEGG pathway maps enriched for *H. pylori* database identified virulence genes to determine where those genes are located reveal in some cases, they constitute all genes present within that specific KEGG pathway map. *H. pylori* database identified virulence genes and their locations on KEGG pathway maps for Epithelial cell signalling in *H. pylori* infection and Flagellar assembly are presented in Figures 5.2 and 5.3 overleaf.

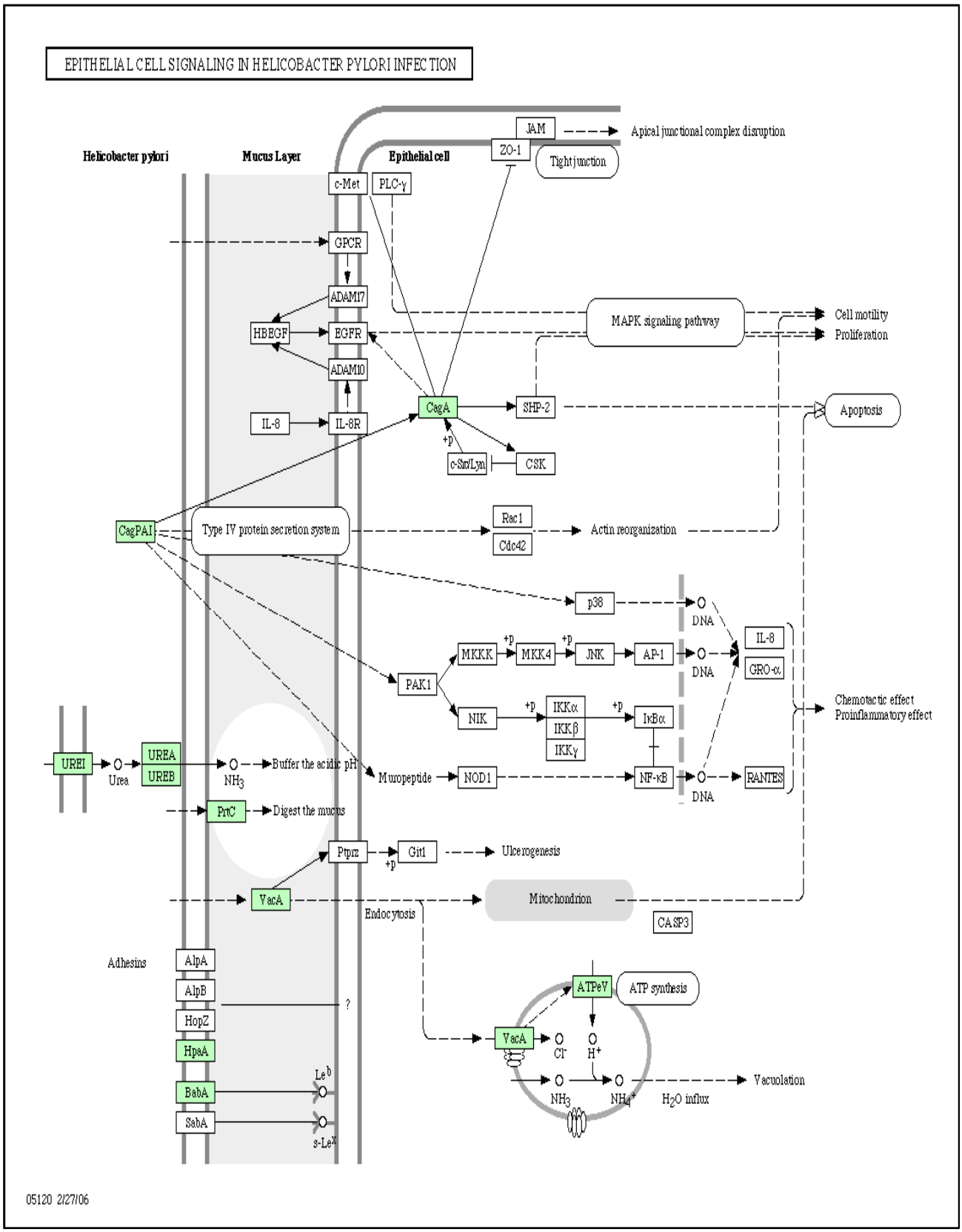


Figure 5.2 : KEGG pathway map (hpy05120) for Epithelial cell signalling in *H. pylori* infection, all genes in green boxes except ATPeV and PrtC are *H. pylori* database identified virulence genes that are part of this KEGG pathway map. The KEGG pathway depicts the insertion of the vacuolating cytotoxin (VacA) into the plasma membrane of an epithelial cell and translocation of CagA into the epithelial cell by the Cag type IV secretion system.

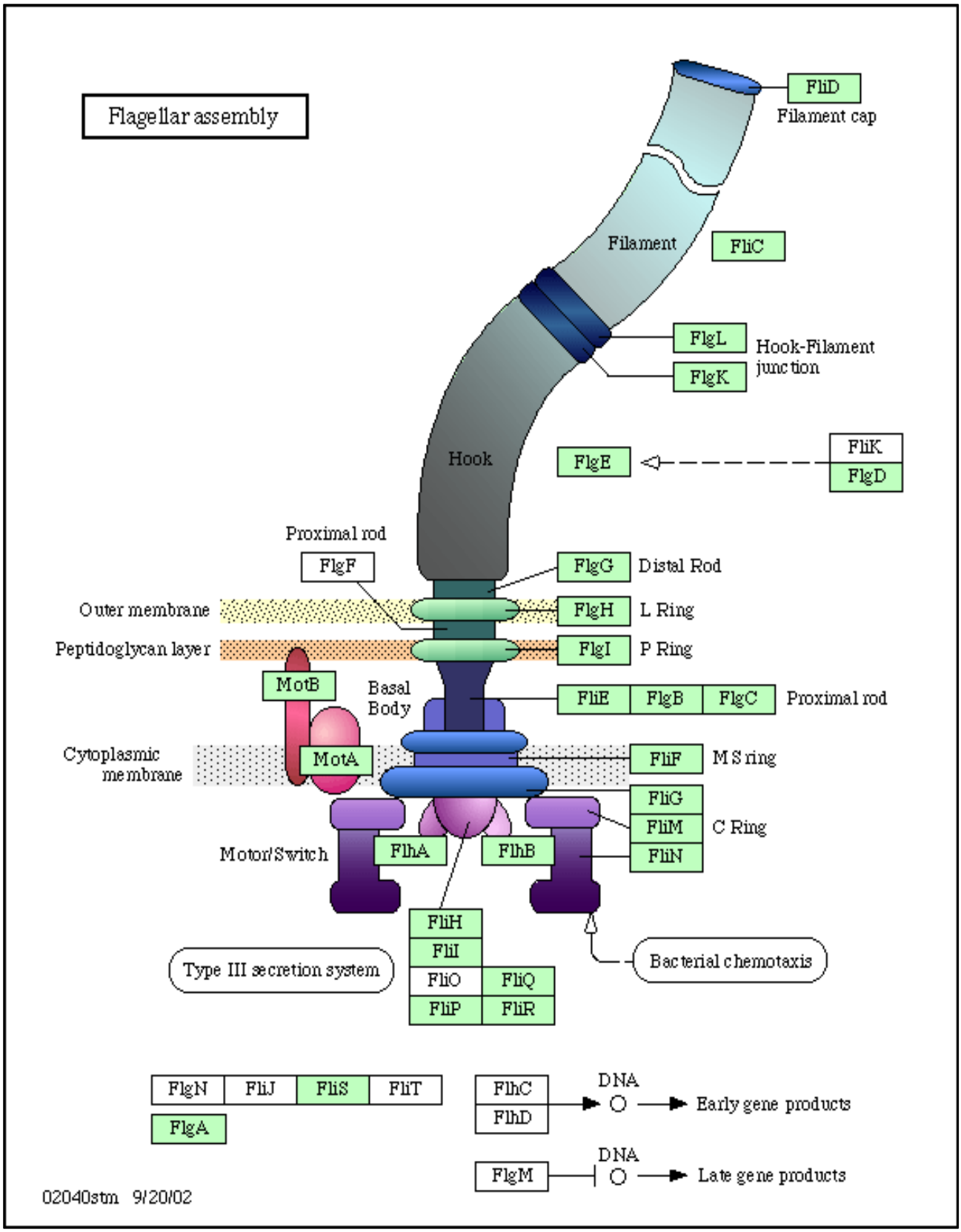


Figure 5.3 : KEGG pathway map for Flagellar assembly (hpy02040) in *H. pylori*. All genes in green boxes are database identified virulence genes for *H. pylori* that can be placed onto this KEGG pathway map. The pathway depicts components of gene products used to construct the flagellum within *H. pylori*. The flagellum within *H. pylori* is essential for motility of the organism and enables movement to high pH gradients within the stomach.

H. pylori genes under positive selection were set as the foreground list and database identified virulence genes as the background gene list in DAVID to ascertain if the two gene lists have any shared KEGG pathways. An EASE score of 1 was selected as no KEGG pathways between the two gene list types were found to be statistically significant at an EASE threshold of 0.5, and to provide as many shared KEGG pathways as possible. Five KEGG pathways are shared by *H. pylori* genes under positive selection and database identified virulence genes; Protein export (hpy03060), Bacterial chemotaxis (hpy02030-general pathway), Two-component system (hpy02020-general pathway), Type IV secretion system (hpy03080) and Epithelial cell signaling in *H. pylori* infection (hpy05120). The *VacA*, *BabA* and *CagA* genes are shared by *H. pylori* database identified virulence genes and genes under positive selection for the Epithelial cell signaling in *H. pylori* infection KEGG pathway (hpy05120), (Figure 5.2, page 98). The general bacterial chemotaxis KEGG pathway (hpy02030), presented below in Figure 5.4, has three *H. pylori* database identified virulence genes under positive selection; *MCP*, *CheA* and *CheV*.

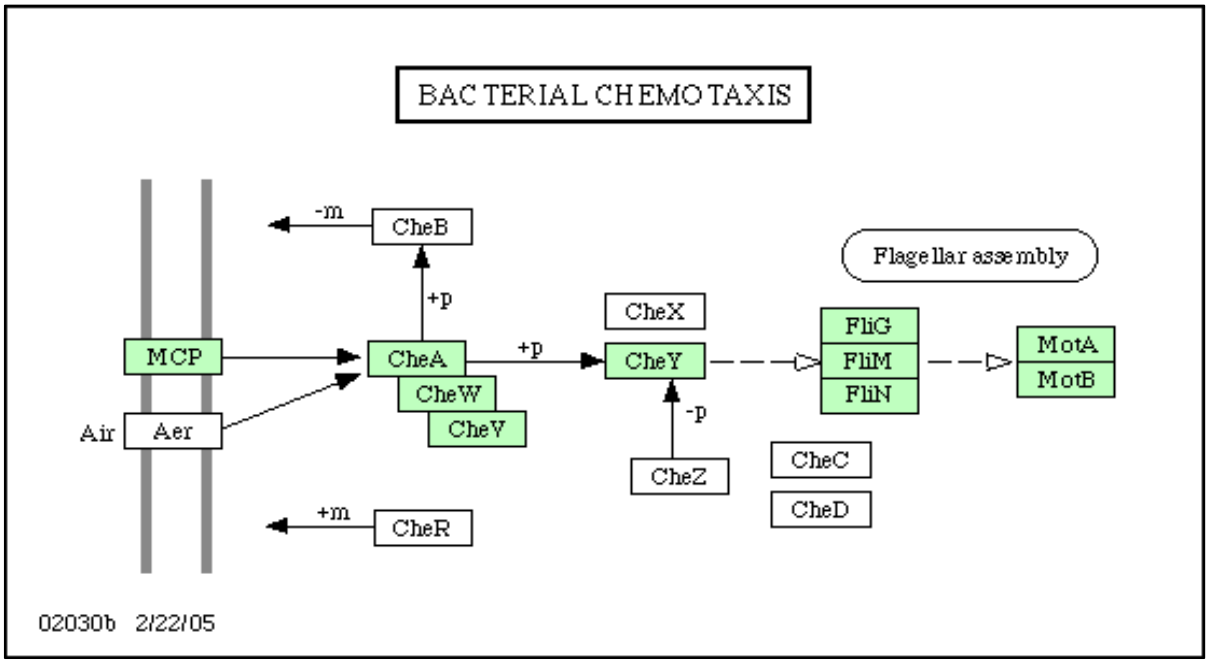


Figure 5.4 : The general bacterial chemotaxis KEGG pathway map (hpy02030) for which *H. pylori* genes under positive selection and database identified virulence can be placed (*MCP*, *CheA* and *CheV*). The CheA histadine kinase is autophosphorylated providing the substrate for the CheY response regulator which modulates the flagellum motor (Figure 5.3, page 99), resulting in bacterial motility.

Although no statistical enrichment of KEGG pathways for *H. pylori* genes under positive selection and database identified virulence genes could be found, these two gene lists have common elements that function in two characterised pathways that are involved in *H. pylori* virulence processes of epithelial cell signalling and chemotaxis / motility.

N. meningitidis **BioCyc Metabolic Pathways**

N. meningitidis BioCyc databases are classified as Tier 3 as their contents have not been reviewed or manually curated like Tier 2 *H. pylori* and *V. cholerae* BioCyc databases [117,118]. BioCyc's *N. meningitidis* database (strain Z2491, Version 13.1) was queried using *N. meningitidis* accessions for genes under positive selection and database identified virulence genes with a metabolic pathway threshold of 1. Of the 218 *N. meningitidis* genes under positive selection, 34% (75 of 218 genes) belong to sixty-four BioCyc metabolic pathways which are summarised below in Table 5.4.

BioCyc Metabolic Pathways of *N. meningitidis* Genes under Positive Selection (Part 1)	**BioCyc Metabolic Pathways of *N. meningitidis* Genes under Positive Selection (Part 2)**
5-aminoimidazole ribonucleotide biosynthesis I	heme biosynthesis from uroporphyrinogen II
acrylonitrile degradation	homolactic fermentation
aerobic respiration -- electron donor II	lipid-A-precursor biosynthesis
arginine biosynthesis I	lysine biosynthesis I
arginine biosynthesis II (acetyl cycle)	methylerythritol phosphate pathway
arginine degradation I (arginase pathway)	methylglyoxal degradation I
asparagine degradation I	mixed acid fermentation
autoinducer AI-2 biosynthesis	NAD biosynthesis I (from aspartate)
biotin biosynthesis I	NAD salvage pathway I
biotin-carboxyl carrier protein	NADH to cytochrome *bd* oxidase electron transfer
CDP-diacylglycerol biosynthesis I	NADH to cytochrome *bo* oxidase electron transfer
CDP-diacylglycerol biosynthesis II	ornithine biosynthesis
CDP-diacylglycerol biosynthesis III	peptidoglycan biosynthesis I
chorismate biosynthesis I	peptidoglycan biosynthesis III
citrulline degradation	ppGpp biosynthesis
CMP-KDO biosynthesis I	proline degradation I
colanic acid building blocks biosynthesis	respiration (anaerobic)
fatty acid activation	S-adenosyl-L-methionine cycle I (prokaryotic)
fatty acid ß-oxidation I	serine biosynthesis
flavin biosynthesis I (bacteria)	siroheme biosynthesis
folate transformations	sulfate reduction I (assimilatory)
formaldehyde oxidation I	TCA cycle variation I
formylTHF biosynthesis I	TCA cycle variation II
formylTHF biosynthesis II	TCA cycle variation IV
GDP-mannose biosynthesis I	tetrahydrofolate biosynthesis
GDP-mannose biosynthesis II	thiamin biosynthesis I
gluconeogenesis I	threonine biosynthesis from homoserine
glutathione biosynthesis	tRNA charging pathway
glycine betaine degradation	tryptophan biosynthesis
glycine biosynthesis I	ubiquinone-8 biosynthesis (prokaryotic)
glycolysis I	uridine-5'-phosphate biosynthesis
glycolysis III (Thermotoga)	UDP-galactose biosynthesis (salvage pathway from galactose using UDP-glucose)

Table 5.4 : Sixty-four BioCyc metabolic pathways that 75 *N. meningitidis* genes under positive selection are mapped to. The above table is divided into two columns (Part 1 and Part 2) providing thirty-two BioCyc metabolic pathways descriptions each.

N. meningitidis BioCyc coverage of genes under positive selection (34% of 218 genes) is greater then BioCyc's coverage of *H. pylori* genes under positive selection (23% of 230 genes). *N. meningitidis* genes under positive selection map to twenty-eight more BioCyc metabolic pathways compared to *H. pylori* genes under positive selection (Table 5.4, page 101, Table 5.1, page 93). *N. meningitidis* and *H. pylori* BioCyc databases have 155 and 143 BioCyc metabolic pathways respectively. Hence, the greater number of metabolic pathways for *N. meningitidis* genes under positive selection compared to *H. pylori* is not necessarily a reflection of *N. meningitidis* having more metabolic pathways then *H. pylori*. Instead, increased BioCyc coverage of *N. meningitidis* genes under positive selection compared to BioCyc's *H. pylori* coverage might be due to a higher false positive rate [118,182]. The PathoLogic prediction scoring scheme is deliberately designed to be as expansive as possible by favouring recall (sensitivity) over precision (specificity), resulting in potential false positive pathway predictions [118,182]. Creators of BioCyc reason the provision of an expansive overview of metabolic pathway predictions for the user to scrutinise and investigate further is better then the total omission of ambiguous metabolic pathways [118,182]. Also, the *H. pylori* BioCyc database is a Tier 2 database which has been manually reviewed and curated compared to *N. meningitidis*' Tier 3 database status. Tier 3 databases have neither been reviewed nor curated and are therefore likely to contain a higher number of metabolic pathway predictions compared to Tier 2 databases [118,182].

Of the 68 *N. meningitidis* database identified virulence genes, 13% (9 of 68 genes) map to three BioCyc metabolic pathways. Even though database identified virulence gene list coverage by BioCyc for *N. meningitidis* (13% of 68 genes) is similar to that of *H. pylori* (14% of 153 genes), *N. meningitidis* database identified virulence genes belong to twelve less BioCyc metabolic pathways compared to *H. pylori*. Lipid A-core biosynthesis, CMP-*N*-acetylneuraminate biosynthesis II (bacteria) and enterobacterial common antigen biosynthesis are the three BioCyc metabolic pathways *N. meningitidis* database indentified virulence genes belong to and are mutually exclusive from the sixty-four metabolic pathways identified for *N. meningitidis* genes under positive selection (Table 5.4, page 101). Absence of shared metabolic pathways between *N. meningitidis* genes under positive selection and database identified virulence genes is surprising given the statistically significant intersection between both gene list types. It is possible common elements between the two *N. meningitidis* gene list types are involved in processes other then metabolic interactions e.g. transport, adhesion, cell membrane integrity. A cellular overview of metabolic pathways which *N. meningitidis* genes under positive selection and database identified virulence genes can be mapped to is presented overleaf in Figure 5.5.

Figure 5.5 : Cellular overview of *N. meningitidis*' BioCyc metabolic pathways for genes under positive selection (yellow lines), database identified virulence genes (blue lines) and genes common to both gene lists (red lines). Numbered boxes are 1: Capsule polysaccharide import / export membrane proteins, 2: IgA1 / Fibrin peptide, 3: Siroheme Biosynthesis, 4: Proline degradation I, 5: lipid A-core biosynthesis, 6: CMP-*N*-acetylneuraminate biosynthesis II (bacteria) and 7: Enterobacterial Common Antigen Biosynthesis.

The observation *N. meningitidis* genes under positive selection and database identified virulence genes do not share any metabolic pathways because genes common to both gene lists are involved in processes other then metabolism is confirmed by *N. meningitidis*' cellular overview map (Figure 5.5, page 103). Four *N. meningitidis* database identified virulence genes under positive selection are capsular polysaccharide transporters (Box 1, Figure 5.5, page 103) and a fifth gene is an IgA1 protease (Box 2, Figure 5.5, page 103).

The capsular polysaccharide transporters (Box1 – red lines, Figure 5.5, page 103) are part of the *ctr* capsular polysaccharide transport operon in *N. meningitidis*. Capsular polysaccharides, mainly composed of sialic acid polymers, are virulence factors linked to invasive *N. meningitidis* meningococcal disease and confer resistance to the bacterium against host bactericidal antibodies and innate immune responses like phagocytosis [199-201]. The *ctr* operon encodes for four proteins; ctrB, ctrD, ctrC and ctrA (Box 1 - red lines from left to right, Figure 5.5, page 103) which function as an ABC transporter exporting synthesised capsular polysaccharides from the bacterial cell [199-201]. Interestingly, other genes close to the ctr transport cluster are also ABC transporters which have yet to be characterised and similar to the *ctr* genes, are under positive selection although they are not deemed as virulence genes by any of the virulence databases (Box1 – yellow lines, Figure 5.5, page 103).

The IgA1 protease (Box 2, Figure 5.5, page 103) cleaves peptide bonds within the human IgA hinge region, is an *N. meningitidis* database identified virulence gene undergoing positive selection [202-204]. Human IgA is predominantly found in the upper respiratory tract and mucosal surfaces where it limits bacterial infection by adhering to the bacterial cell's exterior causing agglutination and preventing epithelial cell receptor-adhesin mediated synergy [202-204]. IgA1 protease is an important virulence factor as non-pathogenic neisserial species do not posses the enzyme and invasive *N. meningitidis* strains have higher levels of IgA1 protease expression compared to commensal *N. meningitidis* strains [204]. Within the BioCyc metabolic map (Box 2, Figure 5.5, page 103), IgA1 is predicted to cleave fibrin into soluble products like plasmin recruited from the bacterium's host [205,206].

N. meningitidis genes under positive selection form discrete metabolic pathway modules whereby other genes within the same pathway are also under positive selection (Figure 5.5, page 103). Two highlighted metabolic pathways within which all genes are under positive selection in *N. meningitidis* are siroheme biosynthesis and proline degradation I (Figure 5.5, page 103, Boxes 3 and 4 respectively).

Siroheme, synthesised by the siroheme pathway (Box 3, Figure 5.5, page 103), is a prosthetic heme like group of the bacterial and yeast enzyme sulphite reductase which catalyses the reduction of sulphite to sulphide, sulphide is used as a substrate for synthesis of sulphur

containing amino acids like cysteine [207-209]. Microarray gene expression characterisation of *N. meningitidis* upon contact with epithelial cells *in vitrio* show genes involved in amino acid synthesis like siroheme synthase and sulphate uptake genes are up-regulated, leading the authors to conclude sulphur metabolism and acquisition have an important role in *N. meningitidis* adhesion [210]. In addition to sulphite reduction, an intermediate compound of the siroeheme biosynthesis pathway known as sirohydrochlorin, a metal free precursor of siroheme, is used for vitamin B12 synthesis [207-209].

The proline degradation I pathway (Box 4, Figure 5.5, page 103) is conserved amongst many bacterial species with the PutA enzyme catabolising L-proline to L-glutamate, L-glutamate is used as an energy and carbon source or a source of nitrogen for growth in bacterial species like *E. coli* [211,212]. Within *E. coli*, the PutA enzyme is found to have a variety of roles depending on proline concentrations; during proline availability PutA functions as an inner cytoplasmic membrane linked enzyme catabolising proline to glutamate, during proline deficient conditions PutA functions as a DNA binding transcriptional repressor of genes involved in the proline pathway degradation after moving to the cytoplasm [211,212]. For a fastidious micro-organism like *N. meningitidis* that utilises limited carbon sources for growth, energy producing pathways like the proline degradation pathway are important because establishment of micro-colonies in the respiratory tract mucosa by founder micro-organisms requires rapid growth and short generation times [97].

Of the three BioCyc metabolic pathways determined for *N. meningitidis* database identified virulence genes (Lipid A-core biosynthesis, CMP-*N*-acetylneuraminate biosynthesis II-bacteria and enterobacterial common antigen biosynthesis), the enterobacterial common antigen biosynthesis pathway (Box 7, Figure 5.5, page 103) is shared by *H. pylori* and *N. meningitidis* database identified virulence gene lists. The lipid A-core biosynthesis pathway (Figure 5.5, Box 5, page 103) produces lipooligosaccharides (LOS) which are pro-inflammatory and abets in meningococcal adherence and invasion of host epithelial cells [213,214]. The lipid A component of LOS acts as an hydrophobic anchor for LOS just below the outer membrane and directly induces inflammatory responses within the host [213,214].

The CMP-*N*-acetylneuraminate biosynthesis II (bacteria) pathway (Figure 5.5, Box 6, page 103) produces sialic acid, which in mammals facilitates cell to cell adhesion and recognition while in bacterial species, like *N. meningitidis*, sialic acid enables bacteria to circumvent host immune responses [199-201,215,216]. Sialic acid constitutes a major component of the *N. meningitidis* capsule and differences in sialic acid polymers are used for capsular serotyping, vaccine targets and epidemiological markers [199-201,215,216]. The synthesised sialic acid polymers are exported to the bacterial cell surface by the ctr ABC transporters (Box 1, Figure 5.5, page 103) [199-201].

KEGG Pathways of *N. meningitidis* Genes.

KEGG pathway maps for *N. meningitidis* genes under positive selection and database identified virulence genes were mined from DAVID (Version 6) using a pathway threshold of 1 and an EASE score of 0.5. KEGG pathway coverage of *N. meningitidis* genes under positive selection is 28% (62 genes of 218) which belong to nineteen different KEGG pathways. None of the nineteen KEGG pathways are statistically enriched for *N. meningitidis* genes under positive selection at the P = 0.05 interval after correction for multiple hypotheses testing. The nineteen KEGG pathways which *N. meningitidis* genes under positive selection belong to are summarised below in Table 5.5.

KEGG Pathway Reference	KEGG Pathway Description
nma02010	ABC transporters – General
nma00970	Aminoacyl tRNA biosynthesis
nma00330	Arginine and proline metabolism
nma00780	Biotin metabolism
nma00460	Cyanoamino acid metabolism
nma03030	DNA polymerase
nma00251	Glutamate metabolism
nma00260	Glycine, serine and threonine metabolism
nma00630	Glyoxylate and dicarboxylate metabolism
nma00540	Lipopolysaccharide biosynthesis
nma00760	Nicotinate and nicotinamide metabolism
nma00550	Peptidoglycan biosynthesis
nma00400	Phenylalanine, tyrosine and tryptophan biosynthesis
nma00860	Porphyrin and chlorophyll metabolism
nma00230	Purine metabolism
nma00240	Pyrimidine metabolism
nma03090	Type II secretion system
nma00220	Urea cycle and metabolism of amino groups
nma00290	Valine, leucine and isoleucine biosynthesis

Table 5.5 : Nineteen KEGG pathway map identifiers and descriptions *N. meningitidis* genes under positive selection belong to.

N. meningitidis and *H. pylori* genes under positive selection have six KEGG pathways in common from a total of thirty four; ABC transporters-General, aminoacyl tRNA

biosynthesis,arginine and proline metabolism, DNA polymerase, peptidoglycan biosynthesis and phenylalanine, tyrosine and tryptophan biosynthesis (Table 5.5, page 106, Table 5.2, page 96).

KEGG pathway coverage of *N. meningitidis* database identified virulence genes is 35% (24 of 68 genes) that belong to four separate KEGG pathways which are summarised below in Table 5.6.

KEGG Pathway Reference	KEGG Pathway Description	Adjusted P Value
nma03090	Type II secretion system	6.14e-19
nma02010	ABC transporters – General	0.193231417
nma01031	Glycan structures - biosynthesis 2	0.999998372
nma00540	Lipopolysaccharide biosynthesis	1

Table 5.6 : KEGG pathway identifiers, descriptions and Bonferroni adjusted P values for KEGG pathway enrichment of *N. meningitidis* database identified virulence genes.

Of the four KEGG pathways, only the Type II secretion system (nma03090) displays a statistical enrichment for *N. meningitidis* database identified virulence genes at the P = 0.05 level (Table 5.6). *N. meningitidis* genes under positive selection and database identified virulence genes share three common KEGG pathways; Type II secretion system, ABC transporters-general and Lipopolysaccharide biosynthesis (Table 5.6 and Table 5.5, page 106). Comparison of KEGG pathway map descriptions show there are no common KEGG pathways between *N. meningitidis* and *H. pylori* database identified virulence genes (Table 5.6 and Table 5.3 page 97).

The KEGG *N. meningitidis* Type II secretion system pathway map (nma03090) contains both genes under positive selection and database identified virulence genes and is presented overleaf in Figure 5.6. Type II secretion systems enable movement of proteins from the cytoplasm across the inner and outer membranes to the cell surface [217-219]. Genes mapped within the *N. meningitidis* Type II secretion KEGG pathway are mainly involved with Type IV pilus assembly and secretion (Figure 5.6, page 108) and are evolutionary interrelated sharing many of the protein secretion components [217-219]. Pili are ancillary filamentous glycosylated protein adhesins which mediate bacterial attachment to host cells and are associated with neisserial competence [220,221]. The type IV pili of *N. meningitidis* comprises of repetitive PilE protein subunits capped with a potential adhesin (PilC) on the pili tip [220,221]. Pili are synthesised within the cytoplasm and transported to the periplasm were they are folded and exported across the outer cell membrane by a multi-protein complex comprising of general secretion pathway proteins that form a structure termed the secreton (Figure 5.6, page 108) [217-219].

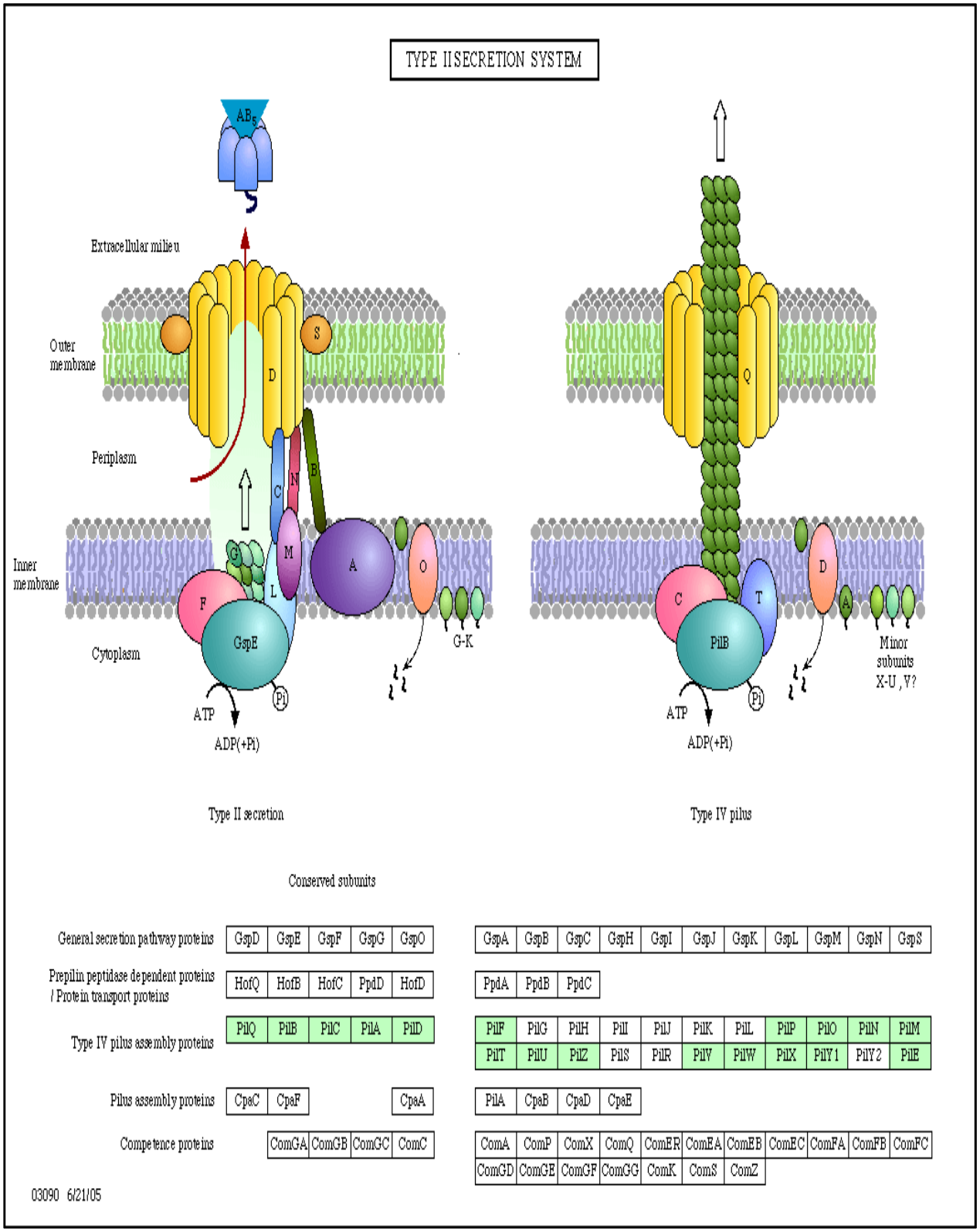


Figure 5.6 : KEGG Type II secretion and Type IV pilus *N. meningitidis* pathway map (nma03090). The *pilA, pilV, pilW and pilE* genes are under positive selection and the *pilQ, pilB, pilC, pilA, pilD, pilF, pilP, pilQ, pilN, pilM, pilT, pilU, pilY1* and *pilE* are database identified virulence genes. The KEGG pathway map depicts conserved proteins used for the synthesis, assembly and export of the Type IV pilus across the bacterial periplasm to the bacterial cell surface.

V. cholerae **BioCyc Metabolic Pathways**

The BioCyc *V. cholerae* Tier 2 database (strain N16961, Version 13.5) was mined for metabolic pathways containing genes under positive selection and database identified virulence genes using accessions for each gene list and a pathway threshold setting of 1. BioCyc coverage of the 23 *V. cholerae* genes under positive selection is 26% (6 of 23 genes). The 6 *V. cholerae* genes under positive selection map to seventeen BioCyc metabolic pathways presented below in Table 5.7. Four of the seventeen metabolic pathways determined within *V. cholera* are also represented by *H. pylori* and *N. meningitidis* genes under positive selection and are boldly italicised in Table 5.7 below.

BioCyc Metabolic Pathways of *V. cholerae* Genes Under Positive Selection
ectoine biosynthesis
pyruvate fermentation to acetate II
valine degradation I
superpathway of KDO2-lipid A biosynthesis
KDO2-lipid A biosynthesis I
superpathway of glycolysis, pyruvate dehydrogenase, TCA, and glyoxylate bypass
lactate oxidation
valine degradation II
acetyl-CoA biosynthesis (from pyruvate)
proline biosynthesis I
lysine biosynthesis VI
citrate fermentation to diacetyl
homoserine biosynthesis
respiration (anaerobic)
fatty acid β-oxidation I
lysine biosynthesis I
tRNA charging pathway

Table 5.7 : Seventeen BioCyc metabolic pathways *V. cholerae* genes under positive selection belong to. Four metabolic pathways that contain genes under positive selection within *V. cholerae, H. pylori* and *N. meningitidis* are marked in bold italic font.

Manual examination of the four BioCyc pathways containing genes under positive selection from all three bacterial species (Table 5.7) reveal different genes within different steps of the metabolic pathways are under positive selection as opposed to functional homologues.

BioCyc coverage of *V. cholerae* database identified virulence genes is 6% (10 of 169 genes) and is comparatively lower then *H. pylori* (22% of 153 genes) and *N. meningitidis* (13% of 68 genes) coverage of database identified virulence genes. Low coverage of *V. cholerae* database identified virulence genes may indicate they are not involved in any metabolic processes and therefore have not been mapped to BioCyc's metabolic pathways or alternatively, the metabolic processes those virulence genes are involved in are yet to be elucidated. The 10 *V. cholerae* database identified virulence genes map to three BioCyc metabolic pathways; 2,3-dihydroxybenzoate biosynthesis, enterobacterial common antigen biosynthesis and 1,4-dihydroxy-2-naphthoate biosynthesis I. There are no shared BioCyc metabolic pathways between *V. cholerae* genes under positive selection and database identified virulence genes. The enterobacterial common antigen biosynthesis pathway is the only pathway common to database identified virulence genes across all three bacterial organisms and only within *H. pylori*, are genes belonging to this pathway under positive selection. A cellular overview of BioCyc metabolic pathways for *V. cholerae* genes under positive selection and database identified virulence genes is presented in Figure 5.7 overleaf.

Ectoine, produced from aspartate by the ectoine biosynthesis pathway (Box 1, Figure 5.7, page 111), is an osmoprotectant facilitating survival of halophilic bacteria like *V. cholerae* in conditions of variable salinity [222,223]. *V. cholerae* constantly encounters fluctuations in salinity within it native environmental niche and during passage through the digestive tract, differences in osmolarity within the digestive tract are postulated to activate virulence mechanisms like production of cholera toxin [222,223]. Salinity also plays an important role in *V. cholerae*'s survival, optimal growth by the bacteria is reached in conditions that have similar salinity levels as its estuarine habitat [222,223]. *V. cholerae* ectoine synthesis is upregulated during growth in highly saline conditions and deletion of genes encoding for ectoine synthesis is shown to impair *V. cholerae*'s growth in those conditions [222,223].

The anaerobic respiration pathway within *V. cholerae* (Box 2, Figure 5.7, page 111) is activated in response to low oxygen conditions such as those encountered by the bacterium in the intestine [224,225]. Exposure to anaerobic conditions are known to trigger the expression of virulence genes for a number of enteric pathogens like *Salmonella typhi* [226,227]. However, in *V. cholerae* growth under *in vitro* anaerobic conditions was found to repress expression of the cholera toxin although expression of the positive regulator of the cholera toxin genes, *toxT*, was found to increase [226]. *In vivo* transcriptional profiling studies also support the upregulation of other *V. cholerae* virulence gene regulators like *toxR*, *tcpP* and *tcpH* [224,225]. Hence, *V. cholerae* virulence gene regulation and expression under anaerobic growth conditions and their relation to pathogenesis has yet to be fully deciphered within *V. cholerae* [224,225].

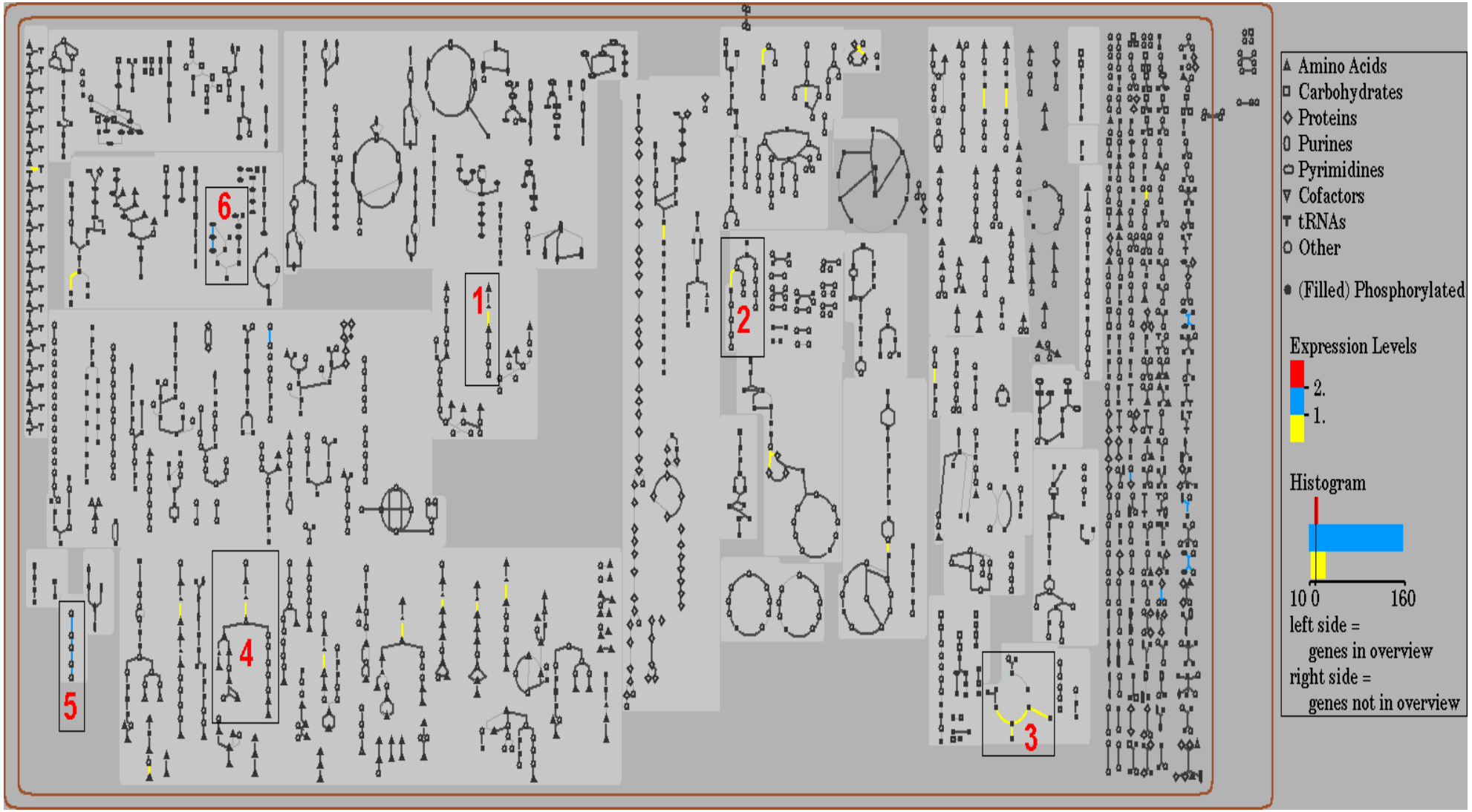


Figure 5.7 : Cellular overview of *V. cholerae* metabolic pathways for genes under positive selection (yellow lines) and database identified virulence genes (blue lines). Metabolic pathways highlighted in numbered boxes are: Box 1 – Ectoine Biosynthesis, Box 2 – Anaerobic Respiration, Box 3 – Fatty acid β-oxidation I, Box 4 – Lysine Biosynthesis I, 5 - 2,3-dihydroxybenzoate Biosynthesis and 6 - Enterobacterial Common Antigen Biosynthesis.

The fatty acid β-oxidation I metabolic pathway has genes under positive selection in all three bacterial organisms (Table 5.7, page 109). The fatty acid β-oxidation pathway (Box 3, Figure 5.7, page 111) is used by bacterial species to generate energy for growth and as a carbon source, two versions of this pathway exist that function either under aerobic or anaerobic conditions [228-230]. Studies in *E. coli* indicate that enteric bacteria utilise the anaerobic fatty acid β-oxidation pathway as the only source of carbon for growth and respiration within the oxygen poor intestinal environment [228,230]. Although most studies of the fatty acid β-oxidation pathway have been carried out within *E. coli*, it is possible the anaerobic version of this pathway applies to *H. pylori* and *V. cholerae* as both are enteric pathogens that encounter low oxygen concentrations in the human digestive system [228,230]. The anaerobic version of the fatty acid β-oxidation pathway may not apply to *N. meningitidis* as the bacterium is mainly aerobic and not starved of oxygen within the nasopharynx.

The lysine biosynthesis pathway I which has genes under positive selection in *V. cholerae* (Box 4, Figure 5.7, page 111), uses the diaminopimelate (DAP) pathway to anabolically convert aspartate to lysine, an essential amino acid for protein synthesis and has genes under positive selection in all three bacterial organisms (Table 5.7, page 109) [231-233]. Both DAP and lysine form important components of the bacterial peptidoglycan cell wall and is essential for bacterial cell growth which is why the lysine biosynthesis pathway I is currently being investigated as an antibacterial drug target due the absence of this pathway in mammals as lysine is acquired through dietary intake [231-233].

Iron is an essential element required for *V. cholerae* growth, respiration, DNA metabolism, electron transport and a co-factor for enzymes [234-238]. The bulk of iron contained within a mammalian host is firmly attached to proteins like haemoglobin, hence bacterial species like *V. cholerae* have evolved high affinity iron-sequestering molecules like siderophores that bind and transport extracellular iron into the bacterial cell [234-238]. All genes contained within the 2,3-dihydroxybenzoate biosynthesis pathway synthesising vibriobactin are database identified virulence genes (Box 5, Figure 5.7, page 111). The 2,3-dihydroxybenzoate biosynthesis pathway (Box 5, Figure 5.7, page 111) is part of the enterobactin biosynthesis super-pathway within BioCyc which synthesises the catecholate siderophore vibriobactin during iron deficient conditions [235,236]. *V. cholerae* does not produce enterobactin, a structurally similar catecholate siderophore produced by *E. coli*, although the bacterium does have receptors to bind and use enterobactin synthesised by other bacteria [234,237,238]. Vibriobactin comprises of three residues of 2,3-dihydroxybenzoyl attached to polyamine backbone of norspermindine and is used to scavenge iron from the bacterial's host and environment [237,238].

KEGG Pathways of *V. cholerae* Genes

V. cholerae accessions of genes under positive selection and database identified virulence genes were used to query DAVID for KEGG pathways using an EASE score of 0.5. Coverage of the 23 *V. cholerae* genes under positive selection for KEGG pathways within DAVID is 15% (3 of 23 genes) which belong to two KEGG pathways; Butanoate metabolism (vch00650) and Valine, leucine and isoleucine biosynthesis (vch00290), none of which are statistically enriched for genes under positive selection at the P = 0.05 level. The valine, leucine and isoleucine biosynthesis KEGG pathway (vch00290) is the only pathway in common between *V. cholerae* and *N. meningitidis* genes under positive selection (Table 5.5, page 106), *V. cholerae* and *H. pylori* genes under positive selection do not have any shared KEGG pathways (Table 5.2, page 96).

Of the 169 *V. cholerae* database identified virulence genes, 54% (92 of 169 genes) map to eleven KEGG pathways, presented below in Table 5.8.

KEGG Pathway Reference	KEGG Pathway Description	Adjusted P Value
vch02040	Flagellar assembly	5.90e-37
vch03090	Type II secretion system	3.53e-10
vch05114	Cholera - Life cycle	6.03e-10
vch05111	Cholera – Colonization	1.19e-07
vch03070	Type III secretion system	2.26e-06
vch05110	Cholera – Infection	8.23e-04
vch01053	Biosynthesis of siderophore group nonribosomal peptides	0.009017124
vch05113	Cholera – Environment	0.011914327
vch02030	Bacterial chemotaxis – General	0.892170334
vch02020	Two-component system - General	0.995410557
vch05112	Cholera – Diarrhoea	1

Table 5.8 : KEGG pathway identifiers and description for *V. cholerae* genes database identified virulence genes as well as Bonferroni corrected results for pathway enrichment analysis.

Eight of the eleven KEGG pathways mapped to *V. cholerae* database identified virulence genes show a statistical enrichment at the P = 0.05 interval after stringent Bonferroni corrections for multiple hypothesis testing (Table 5.8). Four of eleven KEGG pathways are common between *V. cholerae* and *H. pylori*; flagellar assembly, type III secretion system, bacterial chemotaxis and two-component system (Table 5.8 and Table 5.3, page 97). *V. cholerae* and *N. meningitidis* database identified virulence genes have only one common KEGG pathway; type II

secretion system (Table 5.8 and Table 5.6, pages 113 and 107). The higher number of shared KEGG pathways between *V. cholerae* and *H. pylori* database identified virulence genes may indicate a gross similarity in their lifestyles as they are both enteric pathogens using similar strategies such as motility and chemo-taxis while *N. meningitidis* is not an enteric pathogen and utilises different strategies during colonisation like pili mediated adhesion.

Comparison of *V. cholerae* genes under positive selection and database identified virulence genes reveal there are no common KEGG pathways mapped to both gene list types. Similar to *H. pylori* KEGG pathway maps, *V. cholerae* database identified virulence genes constitute all the genes mapped to some KEGG pathways. An example of two KEGG pathway maps comprising exclusively of *V. cholerae* database identified virulence genes are presented below and overleaf in Figures 5.8 and 5.9 respectively.

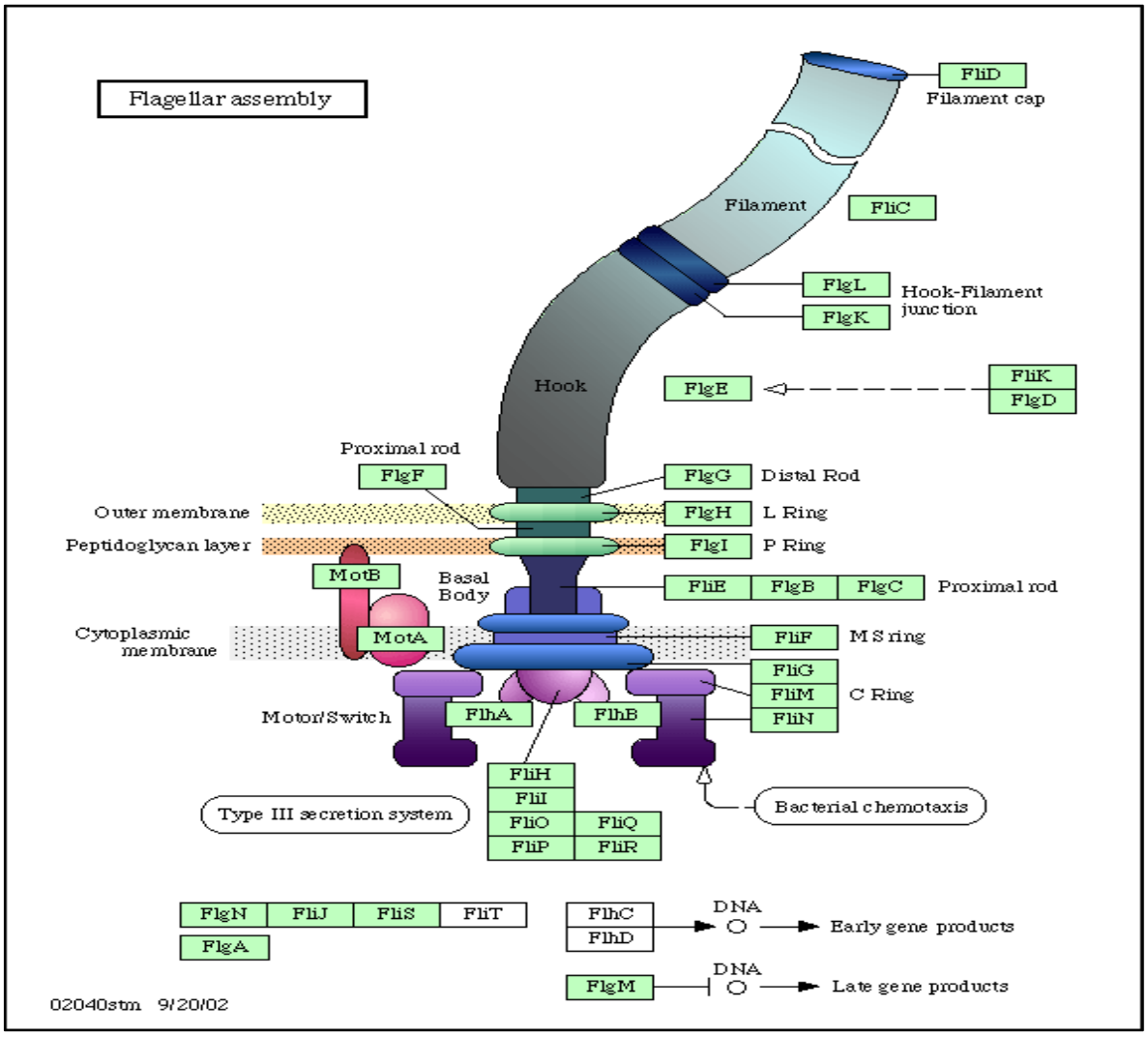


Figure 5.8 : KEGG pathway map for *V. cholerae* Flagellar assembly (vch02040). Genes in green boxes encode protein products used in the assembly of the *V. cholerae* flagellum that is exported to the bacterial outer membrane using a Type III secretion system. Similar to *H. pylori* (Figure 5.3, page 99), all genes contained within this KEGG pathway map highlighted in green boxes are database identified virulence genes.

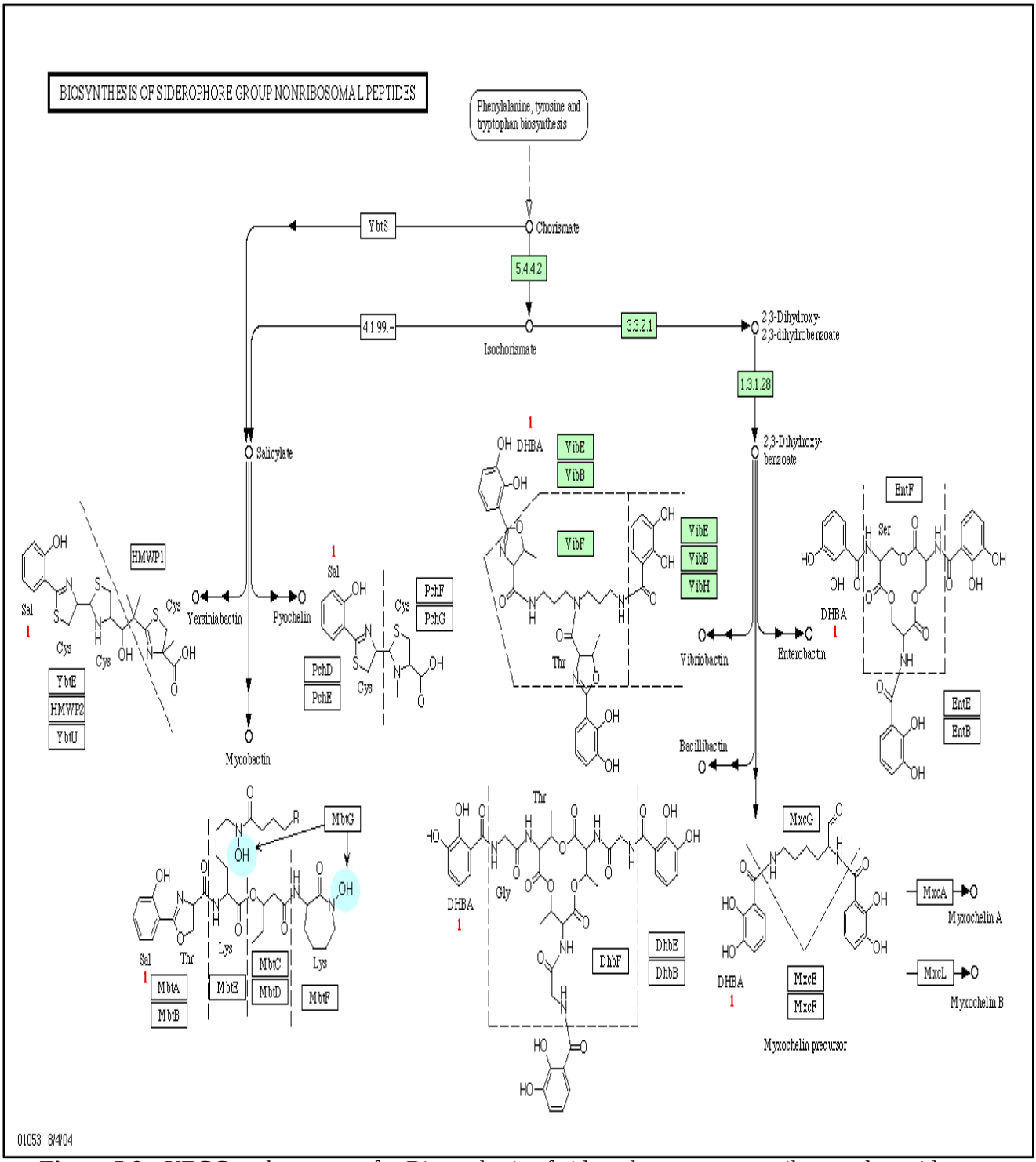


Figure 5.9 : KEGG pathway map for Biosynthesis of siderophore group nonribosomal peptides (vch01053) depicting the synthesis of the *V. cholerae* vibriobactin siderophore, all genes and enzymes highlighted in green boxes are *V cholerae* database identified virulence genes. Three molecules of chorismate are converted to isochorismate by vibriobactin-specific isochorismate synthase (EC 5.4.4.2). The enzyme vibriobactin-specific isochorismatase (EC 3.3.2.1) converts isochorismate to three molecules of 2,3-dihydroxybenzoate (DHB). The VibB and VibE enzymes catalyse DHB to add the catechol moiety. VibF activates L-threonine converting it to a soluble product for transfer and attachment of DHB to the norspermindine backbone by amide bond formation catalysed by VibH. VibB, VibE , VibH and VibF are all used as a multi-protein complex to assemble the vibriobactin with VibE activating DHB to 2,3-dihydroxybenzoic acid (DBHA) to produce mature vibriobactin [239,240].

Chapter 6 : Conclusions

The number of orthologous gene pairs identified between bacterial strains mirrors the bacterial gene content across all three bacterial species. The number of genes under positive selection within each bacterial species does not mirror the gene content of the three bacterial species. The discord in gene numbers under positive selection in relation to bacterial gene content suggests that forces driving genome diversification between the three bacterial species are non-random. Factors that may influence genome diversification between *H. pylori*, *N. meningitidis* and *V. cholerae* include the integrity of DNA repair mechanisms, usage of *de novo* mutational strategies or contingency loci, environmental reservoirs that exert strong purifying selection to expunge mutations and sufficient time between divergence of bacterial strains for an accumulation of mutations to occur [78,93,95,96].

The relationship between positive selection and bacterial virulence is not well defined as not all genes under positive selection are known bacterial virulence genes and *vice versa*. Of the three bacterial species, only *N. meningitidis* has a statistically significant enrichment of virulence genes undergoing positive selection. Although *H. pylori* and *V. cholerae* known virulence genes are not statistically enriched for positive selection, both bacterial species do have virulence genes that are undergoing positive selection (Appendix B2). Both *H. pylori* and *V. cholerae* have more then twice the number of database identified virulence genes compared to *N. meningitidis* (Appendix B2). A larger sample size of virulence genes would ideally result in a larger statistically significant intersection between *H. pylori* and *V. cholerae* genes under positive selection and known virulence genes.

Statistical tests based on categorical data like gene counts are directly affected by the number of elements contained within each class such as the number of virulence genes and genes under positive selection and their resulting intersection. Virulence gene identification within a bacterial species is still incomplete and various definitions of bacterial virulence results in different virulence gene lists like those obtained from VFDB and MVirDB. In order for a bacterial gene to be categorised as a virulence gene, it must have some functional information linking that gene product to a virulence process. All database identified virulence genes for *H. pylori* and *N. meningitidis* have functional annotations while only 13 of 169 *V. cholera* database identified virulence genes do not have any functional annotations (Appendix B2). In contrast, a large number of genes under positive selection within *H. pylori*, *N. meningitidis* and *V. cholerae* are classified as hypothetical (81, 47 and 9 genes respectively, Appendix B1) because their biological functions are unknown. Application of high throughput genomic technologies like micro-arrays are identifying numerous genes with upregulated expressions under simulated

infection conditions whose biological functions are unknown and thus, are also annotated as hypothetical genes [224,225]. As functional annotation of microbial genomes improve, gene counts of bacterial virulence genes identified are liable to increase and possibly result in a larger intersection with genes under positive selection. Hence, an absence of a statistically significant enrichment for *H. pylori* and *V. cholerae* virulence genes undergoing positive selection at the current level of gene functional annotation does not necessarily preclude the existence of a synergistic correlation between genes under positive selection and virulence genes.

In terms of physical properties, genes under positive selection in all three bacterial species are on average, longer then genomic and bacterial virulence genes and have different statistically significant distributions with the exception of *N. meningitidis*. Genes under positive selection have longer gene lengths due to the presence of more codons resulting in a higher probability of non-synonymous substitutions occurring. Virulence genes in all three bacterial species are on average, longer then the genomic mean gene length and follow different distributions that are statistically significant from the genomic gene length distribution. There is little literature to suggest *a priori* why virulence genes are on average, more likely to be longer then genomic mean length. In this regard, virulence genes and genes under positive selection from all three bacterial species have a similar physical property that differentiates them from the genomic gene length distribution. The %GC content distribution between gene lists for the three bacterial species is heterogeneous. *N. meningitidis* and *V. cholerae* genes under positive selection have different %GC distributions compared to their genomic distribution while both gene lists types in *H. pylori* follow a similar %GC distribution to that of its genome. *N. meningitidis* virulence genes are %GC poor compared to genes under positive selection and the genome mean (Table B4B, Appendix B4). *V. cholerae* genes under positive selection have a slightly higher %GC content compared to database identified virulence genes and the genomic mean. Hence, %GC content of the bacterial genomes and their respective gene lists are properties unique to each of the bacterial species.

An annotation bias for bacterial virulence genes exists as bacterial virulence genes are more likely to have a functional annotation then bacterial genes not known to be involved in virulence processes. The annotation bias of bacterial virulence genes is propagated through most functional annotation systems resulting in greater functional annotation coverage of bacterial virulence genes in comparison to genes under positive selection for all three bacterial species.

H. pylori, *N. meningitidis* and *V. cholerae* virulence genes and genes under positive selection have 34 GO terms in common at the current level of bacterial gene functional annotation. No statistical enrichment of common GO terms could be found for bacterial genes under positive selection indicating they could be involved in heterogeneous biological processes.

Alternatively, the lack of functional annotations for genes under positive selection in all three bacterial species may prevent those genes from forming functionally related groups resulting in a lack of statistical enrichment for any biological functional process. *H. pylori* and *V. cholerae* virulence genes and genes under positive selection share more GO terms pertaining to biological processes like chemo-taxis, cell motility and cell membrane structure then either bacterial species shares with *N. meningitidis*. The congruence of GO terms between *H. pylori* and *V. cholerae* gene lists may reflect gross similarities in lifestyles, both bacterial species are motile enteric pathogens which may use similar mechanisms to determine environmental gradients like pH, temperature, salinity and nutrient availability [6,71,76,78]. *N. meningitidis* genes under positive selection and virulence genes have GO terms relating to the cell wall envelope, transport and secretion. *H. pylori* and *N. meningitidis* genes under positive selection share GO terms involved in DNA metabolism possibly due to the micro-organisms' unique growth requirements and maintenance of genome homogeneity [5,38,69,153,168]. GO annotation of bacterial gene lists indicates unequal virulence gene characterisation across all three bacterial species. GO terms present for virulence genes in one bacterial species are also present within another bacterial species, but those GO terms are not associated with any virulence genes in the latter bacterial species. Instead those bacterial virulence GO terms in one bacterial species are associated with genes under positive selection in another bacterial species. Hence, although they may be more biological processes shared by bacterial genes under positive selection and virulence genes then the 34 common GO terms suggests, unequal characterisation of virulence within each bacterial species means those common GO terms will not be found. Conversely, as the biological processes described by those GO terms exist in all three bacterial species, those biological processes might only be considered virulence associated in one of the bacterial species and not another, thereby providing a means to discriminate between species unique modes of pathogenicity.

Similar to GO annotations, COG annotation coverage of database identified virulence genes are greater compared to coverage of genes under positive selection. Alphabetical COG categories statistically enriched for virulence genes common to all three bacterial species include [N] cell motility and [U] intracellular trafficking and secretion indicating that some virulence processes are shared by the three bacterial species. A number of species unique alphabetical COG categories are also enriched for virulence genes such as [P] inorganic ion and transport for *H. pylori*, [M] cell wall / membrane biogenesis for *N. meningitidis* and [E] amino acid transport and metabolism for *V. cholerae*. No statistical enrichment for any alphabetical COG category could be found for *H. pylori*, *N. meningitidis* and *V. cholerae* genes under positive selection. Of the three bacterial species, only *N. meningitidis* database identified virulence genes and genes

under positive selection show a statistically significant association for the COG categories [M] cell wall / membrane biogenesis and [N] cell motility.

PSORTb is the only functional annotation system employed in this study that provides similar coverage for genes under positive selection and database identified virulence genes. PSORTb is able to achieve similar coverage for known virulence genes and genes under positive selection because it is a predictive classification system that utilises biological sequence features encoded within a protein sequence to predict a sub-cellular localisation sites and does not rely on pre-existing functional annotations. The cytoplasm and outer membrane subcellular localisation sites show a statistically significant enrichment for bacterial virulence genes in all three bacterial species. The extra-cellular PSORTb category shows a statistically significant enrichment for *H. pylori* and *V. cholerae* virulence genes indicating the secretion of factors like the *H. pylori* vacuolating cytotoxin or the cholera toxin into their surrounding environment plays a major roles in these bacterial species' pathogenicity [6,71,76,78,81,92]. The outer membrane subcellular localisation site is enriched for *H. pylori* and *N. meningitidis* virulence genes and genes under positive selection. Although the outer membrane subcellular localisation site is enriched for *H. pylori* and *N. meningitidis* virulence genes and genes under positive selection, no statistically significant association can be found between the two classes of genes when tested in unison. The lack of statistical association between *H. pylori* and *N. meningitidis* virulence genes and genes under positive selection might be because some of the outer membrane genes under positive selection in *H. pylori* and *N. meningitidis* are not virulence genes. Or alternatively, due to a lack of functional annotations for outer membrane genes under positive selection, they are not classified as virulence genes resulting in a lack of statistical association between the two gene classes for the outer membrane. In terms of species similarity, both *H. pylori* and *N. meningitidis* colonise and interact with their host environment through the use of cell surface proteins that mediate adhesion and epithelial cell interactions that are under host immune surveillance which exerts selective pressure for those outer membrane proteins to diversify [6,15,76,78,96]. *V. cholerae* does not colonise human hosts and therefore may not have well developed factors that facilitate interactions as well as host immune evasion like *H. pylori* and *N. meningitidis* [6].

KEGG pathway maps show good concordance of virulence genes with some KEGG pathway maps like flagellar assembly and Type II and III secretion systems enriched for *H. pylori*, *N. meningitidis* and *V. cholerae* virulence genes. In some instances, KEGG pathway maps comprise exclusively of database identified virulence genes. No KEGG pathway maps were found to be statistically enriched for genes under positive selection in all three bacterial species which is unsurprising as only genes with known functions are placed in pathway maps. Species unique KEGG pathway maps such as epithelial cell signalling in *H. pylori* infection and

Type II secretion systems in *N. meningitidis* are virulence processes that contain both known virulence genes and genes under positive selection. Hence, genes under positive selection are involved in the same virulence processes as known virulence genes.

Metabolic pathway analysis of genes under positive selection and database identified virulence genes by BioCyc places more genes under positive selection in metabolic pathways compared to known virulence genes. The majority of *N. meningitidis* virulence genes appear to be involved in transport and adhesion as opposed to any known metabolic processes. Within *H. pylori* and *N. meningitidis*, known virulence genes and genes under positive selection form part of the same metabolic pathways, in some cases virulence genes involved in those metabolic pathways are under positive selection. Metabolic pathway analysis of *H. pylori* genes indentified two metabolic pathways common to genes under positive selection and known virulence genes that are not based on genes shared by both gene lists. One of the metabolic pathways, GDP-Mannose Metabolism is involved in the fucosylation of LPS-O chains which enables Lewis antigen mimicry facilitating *H. pylori* adhesion to host epithelial cells, a biological process directly linked to gastritis [6,193-195]. Metabolic pathway analysis reveals *N. meningitidis* known virulence genes under positive selection and linked to the cell membrane are surrounded by other cell membrane linked proteins that are also under positive selection, but due to a lack of functional annotation for these cell membrane linked proteins under positive selection, they are not identified as virulence genes.

Comparison of two metabolic pathways between *H. pylori* and *V. cholerae* reveals discrepancies between defined virulence processes between bacterial species. The urea degradation pathway within *H. pylori* produces ammonia to preserve cell homeostasis in low pH conditions [76,185,186]. The ectoine biosynthesis pathway within *V. cholerae* produces ectoine from aspartate to also preserve cell homeostasis in highly saline conditions [222,223]. Both metabolic pathways and end products are vital for *H. pylori*'s and *V cholerae*'s survival in their respective environmental niches and both metabolic pathways contain genes that are under positive selection [76,185,186,222,223]. The only difference between the two analogous metabolic pathways is genes involved in the *H. pylori* urea degradation pathway are characterised as virulence genes while genes involved in the *V. cholerae* ectoine synthesis pathway are not characterised as virulence genes.

Species specific metabolic pathways for known virulence genes exists e.g. the urea degradation pathway in *H. pylori*, the sialic synthesising pathway in *N. meningitidis* and the vibriobactin synthesis pathway in *V. cholerae*. Metabolic pathway analysis of genes under positive selection show there are four metabolic pathways shared by all three bacterial species, of those four metabolic pathways, one (lysine biosynthesis I) is currently being investigated as

an antibacterial drug target [231]. The substantial number of metabolic pathways containing *H. pylori* and *N. meningitidis* genes under positive selection and known virulence genes indicates bacterial genes under positive selection and known bacterial virulence genes can be grouped on a biological process level.

Assaying for positive selection using a variety of bacterial species demonstrates there are similar genes undergoing positive selection between bacterial species which are part of well characterised virulence processes. There are also classes of genes undergoing positive selection that are involved in species unique virulence processes. At the current level of functional annotation, statistical significance between genes under positive selection and bacterial virulence genes based on an intersection of common gene elements between both gene classes can not be determined, with the exception of *N. meningitidis*. The absence of any statistically significant intersection between genes under positive selection and known bacterial virulence genes when examined as separate entities does not preclude the existence of a relationship between positive selection and bacterial virulence. On a biological process level, genes under positive selection and known bacterial virulence genes have more in common with each other then if both gene classes are examined as separate, discrete entities. Numerous, well characterised bacterial virulence processes contain genes which are under positive selection as determined by their annotations and pathways they operate within. Genes and their products do not function in isolation within complex biological systems and well characterised virulence processes comprise of genes under positive selection. Therefore a relationship between positive selection and bacterial virulence can be determined on a biological process level, thereby establishing a link between nucleotide sequence diversity and bacterial virulence.

Reference List

1. Bansal,A.K. Bioinformatics in microbial biotechnology--a mini review. *Microb. Cell Fact.* **4**, 19 (2005).

2. Burrack,L.S. & Higgins,D.E. Genomic approaches to understanding bacterial virulence. *Curr. Opin. Microbiol* **10**, 4-9 (2007).

3. Guzman,E., Romeu,A., & Garcia-Vallve,S. Completely sequenced genomes of pathogenic bacteria: a review. *Enferm. Infecc. Microbiol Clin.* **26**, 88-98 (2008).

4. Medini,D., Serruto,D., Parkhill,J., Relman,D.A., Donati,C., Moxon,R., Falkow,S., & Rappuoli,R. Microbiology in the post-genomic era. *Nat. Rev. Microbiol* **6**, 419-430 (2008).

5. Pallen,M.J. & Wren,B.W. Bacterial pathogenomics. *Nature* **449**, 835-842 (2007).

6. Raskin,D.M., Seshadri,R., Pukatzki,S.U., & Mekalanos,J.J. Bacterial genomics and pathogen evolution. *Cell* **124**, 703-714 (2006).

7. Furlong,R.F. & Yang,Z. Comparative genomics coming of age. *Heredity* **91**, 533-534 (2003).

8. Galperin,M.Y. & Cochrane,G.R. Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009. *Nucleic Acids Res* **37**, D1-D4 (2009).

9. Zhang,R. & Zhang,C.T. The impact of comparative genomics on infectious disease research. *Microbes. Infect.* **8**, 1613-1622 (2006).

10. Fournier,P.E., Drancourt,M., & Raoult,D. Bacterial genome sequencing and its use in infectious diseases. *Lancet Infect. Dis.* **7**, 711-723 (2007).

11. Merrell,D.S. & Falkow,S. Frontal and stealth attack strategies in microbial pathogenesis. *Nature* **430**, 250-256 (2004).

12. Wu,H.J., Wang,A.H., & Jennings,M.P. Discovery of virulence factors of pathogenic bacteria. *Curr. Opin. Chem. Biol* **12**, 93-101 (2008).

13. Wren,B.W. Microbial genome analysis: insights into virulence, host adaptation and evolution. *Nat. Rev. Genet.* **1**, 30-39 (2000).

14. Alm,R.A., Ling,L.S., Moir,D.T., King,B.L., Brown,E.D., Doig,P.C., Smith,D.R., Noonan,B., Guild,B.C., deJonge,B.L., Carmel,G., Tummino,P.J., Caruso,A., Uria-Nickelsen,M., Mills,D.M., Ives,C., Gibson,R., Merberg,D., Mills,S.D., Jiang,Q., Taylor,D.E., Vovis,G.F., & Trust,T.J. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397**, 176-180 (1999).

15. Chen,S.L., Hung,C.S., Xu,J., Reigstad,C.S., Magrini,V., Sabo,A., Blasiar,D., Bieri,T., Meyer,R.R., Ozersky,P., Armstrong,J.R., Fulton,R.S., Latreille,J.P., Spieth,J., Hooton,T.M., Mardis,E.R., Hultgren,S.J., & Gordon,J.I. Identification of genes subject

to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. *Proc. Natl. Acad. Sci U. S. A* **103**, 5977-5982 (2006).

16. Cole,S.T. Comparative mycobacterial genomics. *Curr. Opin. Microbiol* **1**, 567-571 (1998).

17. Dziejman,M., Balon,E., Boyd,D., Fraser,C.M., Heidelberg,J.F., & Mekalanos,J.J. Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease. *Proc. Natl. Acad. Sci U. S. A* **99**, 1556-1561 (2002).

18. Petersen,L., Bollback,J.P., Dimmic,M., Hubisz,M., & Nielsen,R. Genes under positive selection in *Escherichia coli*. *Genome Res* **17**, 1336-1343 (2007).

19. Chen,L., Yang,J., Yu,J., Yao,Z., Sun,L., Shen,Y., & Jin,Q. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res* **33**, D325-D328 (2005).

20. Gibbs,R.A., Rogers,J., Katze,M.G., Bumgarner,R., Weinstock,G.M., Mardis,E.R., Remington,K.A., Strausberg,R.L., Venter,J.C., Wilson,R.K., Batzer,M.A., Bustamante,C.D., Eichler,E.E., Hahn,M.W., Hardison,R.C., Makova,K.D., Miller,W., Milosavljevic,A., Palermo,R.E., Siepel,A., Sikela,J.M., Attaway,T., Bell,S., Bernard,K.E., Buhay,C.J., Chandrabose,M.N., Dao,M., Davis,C., Delehaunty,K.D., Ding,Y., Dinh,H.H., Dugan-Rocha,S., Fulton,L.A., Gabisi,R.A., Garner,T.T., Godfrey,J., Hawes,A.C., Hernandez,J., Hines,S., Holder,M., Hume,J., Jhangiani,S.N., Joshi,V., Khan,Z.M., Kirkness,E.F., Cree,A., Fowler,R.G., Lee,S., Lewis,L.R., Li,Z., Liu,Y.S., Moore,S.M., Muzny,D., Nazareth,L.V., Ngo,D.N., Okwuonu,G.O., Pai,G., Parker,D., Paul,H.A., Pfannkoch,C., Pohl,C.S., Rogers,Y.H., Ruiz,S.J., Sabo,A., Santibanez,J., Schneider,B.W., Smith,S.M., Sodergren,E., Svatek,A.F., Utterback,T.R., Vattathil,S., Warren,W., White,C.S., Chinwalla,A.T., Feng,Y., Halpern,A.L., Hillier,L.W., Huang,X., Minx,P., Nelson,J.O., Pepin,K.H., Qin,X., Sutton,G.G., Venter,E., Walenz,B.P., Wallis,J.W., Worley,K.C., Yang,S.P., Jones,S.M., Marra,M.A., Rocchi,M., Schein,J.E., Baertsch,R., Clarke,L., Csuros,M., Glasscock,J., Harris,R.A., Havlak,P., Jackson,A.R., Jiang,H., Liu,Y., Messina,D.N., Shen,Y., Song,H.X., Wylie,T., Zhang,L., Birney,E., Han,K., Konkel,M.K., Lee,J., Smit,A.F., Ullmer,B., Wang,H., Xing,J., Burhans,R., Cheng,Z., Karro,J.E., Ma,J., Raney,B., She,X., Cox,M.J., Demuth,J.P., Dumas,L.J., Han,S.G., Hopkins,J., Karimpour-Fard,A., Kim,Y.H., Pollack,J.R., Vinar,T., Addo-Quaye,C., Degenhardt,J., Denby,A., Hubisz,M.J., Indap,A., Kosiol,C., Lahn,B.T., Lawson,H.A., Marklein,A., Nielsen,R., Vallender,E.J., Clark,A.G., Ferguson,B., Hernandez,R.D., Hirani,K., Kehrer-Sawatzki,H., Kolb,J., Patil,S., Pu,L.L., Ren,Y., Smith,D.G., Wheeler,D.A., Schenck,I., Ball,E.V., Chen,R., Cooper,D.N., Giardine,B., Hsu,F., Kent,W.J., Lesk,A., Nelson,D.L., O'brien,W.E., Prufer,K., Stenson,P.D., Wallace,J.C., Ke,H., Liu,X.M., Wang,P., Xiang,A.P., Yang,F., Barber,G.P., Haussler,D., Karolchik,D., Kern,A.D., Kuhn,R.M., Smith,K.E., & Zwieg,A.S. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222-234 (2007).

21. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., Funke,R., Gage,D., Harris,K., Heaford,A., Howland,J., Kann,L., Lehoczky,J., LeVine,R., McEwan,P., McKernan,K., Meldrim,J., Mesirov,J.P., Miranda,C., Morris,W., Naylor,J., Raymond,C., Rosetti,M., Santos,R., Sheridan,A., Sougnez,C., Stange-Thomann,N., Stojanovic,N., Subramanian,A., Wyman,D., Rogers,J., Sulston,J., Ainscough,R., Beck,S., Bentley,D., Burton,J., Clee,C., Carter,N., Coulson,A., Deadman,R., Deloukas,P., Dunham,A., Dunham,I., Durbin,R., French,L., Grafham,D., Gregory,S., Hubbard,T., Humphray,S., Hunt,A.,

Jones,M., Lloyd,C., McMurray,A., Matthews,L., Mercer,S., Milne,S., Mullikin,J.C., Mungall,A., Plumb,R., Ross,M., Shownkeen,R., Sims,S., Waterston,R.H., Wilson,R.K., Hillier,L.W., McPherson,J.D., Marra,M.A., Mardis,E.R., Fulton,L.A., Chinwalla,A.T., Pepin,K.H., Gish,W.R., Chissoe,S.L., Wendl,M.C., Delehaunty,K.D., Miner,T.L., Delehaunty,A., Kramer,J.B., Cook,L.L., Fulton,R.S., Johnson,D.L., Minx,P.J., Clifton,S.W., Hawkins,T., Branscomb,E., Predki,P., Richardson,P., Wenning,S., Slezak,T., Doggett,N., Cheng,J.F., Olsen,A., Lucas,S., Elkin,C., Uberbacher,E., Frazier,M., Gibbs,R.A., Muzny,D.M., Scherer,S.E., Bouck,J.B., Sodergren,E.J., Worley,K.C., Rives,C.M., Gorrell,J.H., Metzker,M.L., Naylor,S.L., Kucherlapati,R.S., Nelson,D.L., Weinstock,G.M., Sakaki,Y., Fujiyama,A., Hattori,M., Yada,T., Toyoda,A., Itoh,T., Kawagoe,C., Watanabe,H., Totoki,Y., Taylor,T., Weissenbach,J., Heilig,R., Saurin,W., Artiguenave,F., Brottier,P., Bruls,T., Pelletier,E., Robert,C., Wincker,P., Smith,D.R., Doucette-Stamm,L., Rubenfield,M., Weinstock,K., Lee,H.M., Dubois,J., Rosenthal,A., Platzer,M., Nyakatura,G., Taudien,S., Rump,A., Yang,H., Yu,J., Wang,J., Huang,G., Gu,J., Hood,L., Rowen,L., Madan,A., Qin,S., Davis,R.W., Federspiel,N.A., Abola,A.P., Proctor,M.J., Myers,R.M., Schmutz,J., Dickson,M., Grimwood,J., Cox,D.R., Olson,M.V., Kaul,R., Raymond,C., Shimizu,N., Kawasaki,K., Minoshima,S., Evans,G.A., Athanasiou,M., Schultz,R., Roe,B.A., Chen,F., Pan,H., Ramser,J., Lehrach,H., Reinhardt,R., McCombie,W.R., de la,B.M., Dedhia,N., Blocker,H., Hornischer,K., Nordsiek,G., Agarwala,R., Aravind,L., Bailey,J.A., Bateman,A., Batzoglou,S., Birney,E., Bork,P., Brown,D.G., Burge,C.B., Cerutti,L., Chen,H.C., Church,D., Clamp,M., Copley,R.R., Doerks,T., Eddy,S.R., Eichler,E.E., Furey,T.S., Galagan,J., Gilbert,J.G., Harmon,C., Hayashizaki,Y., Haussler,D., Hermjakob,H., Hokamp,K., Jang,W., Johnson,L.S., Jones,T.A., Kasif,S., Kaspryzk,A., Kennedy,S., Kent,W.J., Kitts,P., Koonin,E.V., Korf,I., Kulp,D., Lancet,D., Lowe,T.M., McLysaght,A., Mikkelsen,T., Moran,J.V., Mulder,N., Pollara,V.J., Ponting,C.P., Schuler,G., Schultz,J., Slater,G., Smit,A.F., Stupka,E., Szustakowski,J., Thierry-Mieg,D., Thierry-Mieg,J., Wagner,L., Wallis,J., Wheeler,R., Williams,A., Wolf,Y.I., Wolfe,K.H., Yang,S.P., Yeh,R.F., Collins,F., Guyer,M.S., Peterson,J., Felsenfeld,A., Wetterstrand,K.A., Patrinos,A., Morgan,M.J., de Jong,P., Catanese,J.J., Osoegawa,K., Shizuya,H., Choi,S., & Chen,Y.J. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

22. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P., Antonarakis,S.E., Attwood,J., Baertsch,R., Bailey,J., Barlow,K., Beck,S., Berry,E., Birren,B., Bloom,T., Bork,P., Botcherby,M., Bray,N., Brent,M.R., Brown,D.G., Brown,S.D., Bult,C., Burton,J., Butler,J., Campbell,R.D., Carninci,P., Cawley,S., Chiaromonte,F., Chinwalla,A.T., Church,D.M., Clamp,M., Clee,C., Collins,F.S., Cook,L.L., Copley,R.R., Coulson,A., Couronne,O., Cuff,J., Curwen,V., Cutts,T., Daly,M., David,R., Davies,J., Delehaunty,K.D., Deri,J., Dermitzakis,E.T., Dewey,C., Dickens,N.J., Diekhans,M., Dodge,S., Dubchak,I., Dunn,D.M., Eddy,S.R., Elnitski,L., Emes,R.D., Eswara,P., Eyras,E., Felsenfeld,A., Fewell,G.A., Flicek,P., Foley,K., Frankel,W.N., Fulton,L.A., Fulton,R.S., Furey,T.S., Gage,D., Gibbs,R.A., Glusman,G., Gnerre,S., Goldman,N., Goodstadt,L., Grafham,D., Graves,T.A., Green,E.D., Gregory,S., Guigo,R., Guyer,M., Hardison,R.C., Haussler,D., Hayashizaki,Y., Hillier,L.W., Hinrichs,A., Hlavina,W., Holzer,T., Hsu,F., Hua,A., Hubbard,T., Hunt,A., Jackson,I., Jaffe,D.B., Johnson,L.S., Jones,M., Jones,T.A., Joy,A., Kamal,M., Karlsson,E.K., Karolchik,D., Kasprzyk,A., Kawai,J., Keibler,E., Kells,C., Kent,W.J., Kirby,A., Kolbe,D.L., Korf,I., Kucherlapati,R.S., Kulbokas,E.J., Kulp,D., Landers,T., Leger,J.P., Leonard,S., Letunic,I., LeVine,R., Li,J., Li,M., Lloyd,C., Lucas,S., Ma,B., Maglott,D.R., Mardis,E.R., Matthews,L., Mauceli,E., Mayer,J.H., McCarthy,M., McCombie,W.R.,

McLaren,S., McLay,K., McPherson,J.D., Meldrim,J., Meredith,B., Mesirov,J.P., Miller,W., Miner,T.L., Mongin,E., Montgomery,K.T., Morgan,M., Mott,R., Mullikin,J.C., Muzny,D.M., Nash,W.E., Nelson,J.O., Nhan,M.N., Nicol,R., Ning,Z., Nusbaum,C., O'Connor,M.J., Okazaki,Y., Oliver,K., Overton-Larty,E., Pachter,L., Parra,G., Pepin,K.H., Peterson,J., Pevzner,P., Plumb,R., Pohl,C.S., Poliakov,A., Ponce,T.C., Ponting,C.P., Potter,S., Quail,M., Reymond,A., Roe,B.A., Roskin,K.M., Rubin,E.M., Rust,A.G., Santos,R., Sapojnikov,V., Schultz,B., Schultz,J., Schwartz,M.S., Schwartz,S., Scott,C., Seaman,S., Searle,S., Sharpe,T., Sheridan,A., Shownkeen,R., Sims,S., Singer,J.B., Slater,G., Smit,A., Smith,D.R., Spencer,B., Stabenau,A., Stange-Thomann,N., Sugnet,C., Suyama,M., Tesler,G., Thompson,J., Torrents,D., Trevaskis,E., Tromp,J., Ucla,C., Ureta-Vidal,A., Vinson,J.P., Von Niederhausern,A.C., Wade,C.M., Wall,M., Weber,R.J., Weiss,R.B., Wendl,M.C., West,A.P., Wetterstrand,K., Wheeler,R., Whelan,S., Wierzbowski,J., Willey,D., Williams,S., Wilson,R.K., Winter,E., Worley,K.C., Wyman,D., Yang,S., Yang,S.P., Zdobnov,E.M., Zody,M.C., & Lander,E.S. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).

23. Tettelin,H., Masignani,V., Cieslewicz,M.J., Donati,C., Medini,D., Ward,N.L., Angiuoli,S.V., Crabtree,J., Jones,A.L., Durkin,A.S., Deboy,R.T., Davidsen,T.M., Mora,M., Scarselli,M., Ros,I., Peterson,J.D., Hauser,C.R., Sundaram,J.P., Nelson,W.C., Madupu,R., Brinkac,L.M., Dodson,R.J., Rosovitz,M.J., Sullivan,S.A., Daugherty,S.C., Haft,D.H., Selengut,J., Gwinn,M.L., Zhou,L., Zafar,N., Khouri,H., Radune,D., Dimitrov,G., Watkins,K., O'Connor,K.J., Smith,S., Utterback,T.R., White,O., Rubens,C.E., Grandi,G., Madoff,L.C., Kasper,D.L., Telford,J.L., Wessels,M.R., Rappuoli,R., & Fraser,C.M. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci U. S. A* **102**, 13950-13955 (2005).

24. Tettelin,H., Riley,D., Cattuto,C., & Medini,D. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol* **11**, 472-477 (2008).

25. Koonin,E.V. & Wolf,Y.I. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res* **36**, 6688-6719 (2008).

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

27. Perna,N.T., Plunkett,G., III, Burland,V., Mau,B., Glasner,J.D., Rose,D.J., Mayhew,G.F., Evans,P.S., Gregor,J., Kirkpatrick,H.A., Posfai,G., Hackett,J., Klink,S., Boutin,A., Shao,Y., Miller,L., Grotbeck,E.J., Davis,N.W., Lim,A., Dimalanta,E.T., Potamousis,K.D., Apodaca,J., Anantharaman,T.S., Lin,J., Yen,G., Schwartz,D.C., Welch,R.A., & Blattner,F.R. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**, 529-533 (2001).

28. Heidelberg,J.F., Eisen,J.A., Nelson,W.C., Clayton,R.A., Gwinn,M.L., Dodson,R.J., Haft,D.H., Hickey,E.K., Peterson,J.D., Umayam,L., Gill,S.R., Nelson,K.E., Read,T.D., Tettelin,H., Richardson,D., Ermolaeva,M.D., Vamathevan,J., Bass,S., Qin,H., Dragoi,I., Sellers,P., McDonald,L., Utterback,T., Fleishmann,R.D., Nierman,W.C., White,O., Salzberg,S.L., Smith,H.O., Colwell,R.R., Mekalanos,J.J., Venter,J.C., &

Fraser,C.M. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406**, 477-483 (2000).

29. Raoult,D., Ogata,H., Audic,S., Robert,C., Suhre,K., Drancourt,M., & Claverie,J.M. *Tropheryma whipplei* Twist: a human pathogenic Actinobacteria with a reduced genome. *Genome Res* **13**, 1800-1809 (2003).

30. Renesto,P., Crapoulet,N., Ogata,H., La Scola,B., Vestris,G., Claverie,J.M., & Raoult,D. Genome-based design of a cell-free culture medium for *Tropheryma whipplei*. *Lancet* **362**, 447-449 (2003).

31. Masselot,F., Boulos,A., Maurin,M., Rolain,J.M., & Raoult,D. Molecular evaluation of antibiotic susceptibility: *Tropheryma whipplei* paradigm. *Antimicrob. Agents Chemother.* **47**, 1658-1664 (2003).

32. Maione,D., Margarit,I., Rinaudo,C.D., Masignani,V., Mora,M., Scarselli,M., Tettelin,H., Brettoni,C., Iacobini,E.T., Rosini,R., D'Agostino,N., Miorin,L., Buccato,S., Mariani,M., Galli,G., Nogarotto,R., Nardi,D., V, Vegni,F., Fraser,C., Mancuso,G., Teti,G., Madoff,L.C., Paoletti,L.C., Rappuoli,R., Kasper,D.L., Telford,J.L., & Grandi,G. Identification of a universal Group B streptococcus vaccine by multiple genome screen. *Science* **309**, 148-150 (2005).

33. Pizza,M., Scarlato,V., Masignani,V., Giuliani,M.M., Arico,B., Comanducci,M., Jennings,G.T., Baldi,L., Bartolini,E., Capecchi,B., Galeotti,C.L., Luzzi,E., Manetti,R., Marchetti,E., Mora,M., Nuti,S., Ratti,G., Santini,L., Savino,S., Scarselli,M., Storni,E., Zuo,P., Broeker,M., Hundt,E., Knapp,B., Blair,E., Mason,T., Tettelin,H., Hood,D.W., Jeffries,A.C., Saunders,N.J., Granoff,D.M., Venter,J.C., Moxon,E.R., Grandi,G., & Rappuoli,R. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* **287**, 1816-1820 (2000).

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

35. Karp,P.D., Keseler,I.M., Shearer,A., Latendresse,M., Krummenacker,M., Paley,S.M., Paulsen,I., Collado-Vides,J., Gama-Castro,S., Peralta-Gil,M., Santos-Zavaleta,A., Penaloza-Spinola,M.I., Bonavides-Martinez,C., & Ingraham,J. Multidimensional annotation of the *Escherichia coli* K-12 genome. *Nucleic Acids Res* **35**, 7577-7590 (2007).

36. Reed,J.L., Famili,I., Thiele,I., & Palsson,B.O. Towards multidimensional genome annotation. *Nat. Rev. Genet.* **7**, 130-141 (2006).

37. Schilling,C.H., Covert,M.W., Famili,I., Church,G.M., Edwards,J.S., & Palsson,B.O. Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol.* **184**, 4582-4593 (2002).

38. Merhej,V., Royer-Carenzi,M., Pontarotti,P., & Raoult,D. Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biol Direct.* **4**, 13 (2009).

39. Moran,N.A. & Plague,G.R. Genomic changes following host restriction in bacteria. *Curr. Opin. Genet. Dev.* **14**, 627-633 (2004).

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

41. Andersson,S.G., Zomorodipour,A., Andersson,J.O., Sicheritz-Ponten,T., Alsmark,U.C., Podowski,R.M., Naslund,A.K., Eriksson,A.S., Winkler,H.H., & Kurland,C.G. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133-140 (1998).

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

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

44. Brussow,H., Canchaya,C., & Hardt,W.D. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol. Biol Rev.* **68**, 560-602 (2004).

45. FREEMAN,V.J. Studies on the virulence of bacteriophage-infected strains of *Corynebacterium diphtheriae*. *J. Bacteriol.* **61**, 675-688 (1951).

46. Waldor,M.K. & Mekalanos,J.J. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* **272**, 1910-1914 (1996).

47. Hendrix,R.W., Smith,M.C., Burns,R.N., Ford,M.E., & Hatfull,G.F. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc. Natl. Acad. Sci U. S. A* **96**, 2192-2197 (1999).

48. Censini,S., Lange,C., Xiang,Z., Crabtree,J.E., Ghiara,P., Borodovsky,M., Rappuoli,R., & Covacci,A. cag, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. *Proc. Natl. Acad. Sci U. S. A* **93**, 14648-14653 (1996).

49. Covacci,A., Falkow,S., Berg,D.E., & Rappuoli,R. Did the inheritance of a pathogenicity island modify the virulence of *Helicobacter pylori*? *Trends Microbiol* **5**, 205-208 (1997).

50. Gevers,D., Vandepoele,K., Simillon,C., & Van de,P.Y. Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol* **12**, 148-154 (2004).

51. He,X. & Zhang,J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**, 1157-1164 (2005).

52. Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S.V., Eiglmeier,K., Gas,S., Barry,C.E., III, Tekaia,F., Badcock,K., Basham,D., Brown,D., Chillingworth,T., Connor,R., Davies,R., Devlin,K., Feltwell,T., Gentles,S., Hamlin,N., Holroyd,S., Hornsby,T., Jagels,K., Krogh,A., McLean,J., Moule,S., Murphy,L., Oliver,K., Osborne,J., Quail,M.A., Rajandream,M.A., Rogers,J., Rutter,S., Seeger,K., Skelton,J., Squares,R., Squares,S., Sulston,J.E., Taylor,K., Whitehead,S., & Barrell,B.G. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537-544 (1998).

53. Pina,M., Occhialini,A., Monteiro,L., Doermann,H.P., & Megraud,F. Detection of point mutations associated with resistance of *Helicobacter pylori* to clarithromycin by hybridization in liquid phase. *J. Clin. Microbiol* **36**, 3285-3290 (1998).

54. Versalovic,J., Shortridge,D., Kibler,K., Griffy,M.V., Beyer,J., Flamm,R.K., Tanaka,S.K., Graham,D.Y., & Go,M.F. Mutations in 23S rRNA are associated with clarithromycin resistance in *Helicobacter pylori*. *Antimicrob. Agents Chemother.* **40**, 477-480 (1996).

55. Andrews,T.D. & Gojobori,T. Strong positive selection and recombination drive the antigenic variation of the PilE protein of the human pathogen *Neisseria meningitidis*. *Genetics* **166**, 25-32 (2004).

56. Kimura,M. Evolutionary rate at the molecular level. *Nature* **217**, 624-626 (1968).

57. Ureta-Vidal,A., Ettwiller,L., & Birney,E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* **4**, 251-262 (2003).

58. Yang,Z. & Bielawski,J.P. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15**, 496-503 (2000).

59. Miyata,T. & Yasunaga,T. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol.* **16**, 23-36 (1980).

60. Yang,Z. Inference of selection from multiple species alignments. *Curr. Opin. Genet. Dev.* **12**, 688-694 (2002).

61. Yang,Z. & Nielsen,R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol Evol.* **17**, 32-43 (2000).

62. Lio,P. & Goldman,N. Models of molecular evolution and phylogeny. *Genome Res* **8**, 1233-1244 (1998).

63. Goldman,N. & Yang,Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol Evol.* **11**, 725-736 (1994).

64. Nielsen,R. & Yang,Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929-936 (1998).

65. Yang,Z. & Swanson,W.J. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol. Biol Evol.* **19**, 49-57 (2002).

66. Yang,Z., Nielsen,R., Goldman,N., & Pedersen,A.M. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431-449 (2000).

67. Yang,Z. & Nielsen,R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol Evol.* **19**, 908-917 (2002).

68. Zhang,J., Nielsen,R., & Yang,Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol Evol.* **22**, 2472-2479 (2005).

69. Ehrlich,G.D., Hiller,N.L., & Hu,F.Z. What makes pathogens pathogenic. *Genome Biol* **9**, 225 (2008).

70. Finlay,B.B. & Falkow,S. Common themes in microbial pathogenicity. *Microbiol Rev.* **53**, 210-230 (1989).

71. Mekalanos,J.J. Environmental signals controlling expression of virulence determinants in bacteria. *J. Bacteriol.* **174**, 1-7 (1992).

72. Falkow,S. Molecular Koch's postulates applied to microbial pathogenicity. *Rev. Infect. Dis.* **10 Suppl 2**, S274-S276 (1988).

73. Falkow,S. Molecular Koch's postulates applied to bacterial pathogenicity--a personal recollection 15 years later. *Nat. Rev. Microbiol* **2**, 67-72 (2004).

74. Finlay,B.B. & Falkow,S. Common themes in microbial pathogenicity revisited. *Microbiol Mol. Biol Rev.* **61**, 136-169 (1997).

75. Falush,D., Wirth,T., Linz,B., Pritchard,J.K., Stephens,M., Kidd,M., Blaser,M.J., Graham,D.Y., Vacher,S., Perez-Perez,G.I., Yamaoka,Y., Megraud,F., Otto,K., Reichard,U., Katzowitsch,E., Wang,X., Achtman,M., & Suerbaum,S. Traces of human migrations in *Helicobacter pylori* populations. *Science* **299**, 1582-1585 (2003).

76. Kusters,J.G., van Vliet,A.H., & Kuipers,E.J. Pathogenesis of *Helicobacter pylori* infection. *Clin. Microbiol Rev.* **19**, 449-490 (2006).

77. 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., Nelson,K., Quackenbush,J., Zhou,L., Kirkness,E.F., Peterson,S., Loftus,B., Richardson,D., Dodson,R., Khalak,H.G., Glodek,A., McKenney,K., Fitzegerald,L.M., Lee,N., Adams,M.D., Hickey,E.K., Berg,D.E., Gocayne,J.D., Utterback,T.R., Peterson,J.D., Kelley,J.M., Cotton,M.D., Weidman,J.M., Fujii,C., Bowman,C., Watthey,L., Wallin,E., Hayes,W.S., Borodovsky,M., Karp,P.D., Smith,H.O., Fraser,C.M., & Venter,J.C. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539-547 (1997).

78. Suerbaum,S. & Josenhans,C. *Helicobacter pylori* evolution and phenotypic diversification in a changing host. *Nat. Rev. Microbiol* **5**, 441-452 (2007).

79. Spratt,B.G. Microbiology. Stomachs out of Africa. *Science* **299**, 1528-1529 (2003).

80. Solnick,J.V. & Schauer,D.B. Emergence of diverse Helicobacter species in the pathogenesis of gastric and enterohepatic diseases. *Clin. Microbiol Rev.* **14**, 59-97 (2001).

81. Nakamura,H., Yoshiyama,H., Takeuchi,H., Mizote,T., Okita,K., & Nakazawa,T. Urease plays an important role in the chemotactic motility of *Helicobacter pylori* in a viscous environment. *Infect. Immun.* **66**, 4832-4837 (1998).

82. Boneca,I.G., de Reuse,H., Epinat,J.C., Pupin,M., Labigne,A., & Moszer,I. A revised annotation and comparative analysis of *Helicobacter pylori* genomes. *Nucleic Acids Res* **31**, 1704-1714 (2003).

83. Doig,P., de Jonge,B.L., Alm,R.A., Brown,E.D., Uria-Nickelsen,M., Noonan,B., Mills,S.D., Tummino,P., Carmel,G., Guild,B.C., Moir,D.T., Vovis,G.F., & Trust,T.J. *Helicobacter pylori* physiology predicted from genomic comparison of two strains. *Microbiol Mol. Biol Rev.* **63**, 675-707 (1999).

84. Gressmann,H., Linz,B., Ghai,R., Pleissner,K.P., Schlapbach,R., Yamaoka,Y., Kraft,C., Suerbaum,S., Meyer,T.F., & Achtman,M. Gain and loss of multiple genes during the evolution of *Helicobacter pylori*. *PLoS. Genet.* **1**, e43 (2005).

85. Parkhill,J., Achtman,M., James,K.D., Bentley,S.D., Churcher,C., Klee,S.R., Morelli,G., Basham,D., Brown,D., Chillingworth,T., Davies,R.M., Davis,P., Devlin,K., Feltwell,T., Hamlin,N., Holroyd,S., Jagels,K., Leather,S., Moule,S., Mungall,K., Quail,M.A., Rajandream,M.A., Rutherford,K.M., Simmonds,M., Skelton,J., Whitehead,S., Spratt,B.G., & Barrell,B.G. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404**, 502-506 (2000).

86. Tettelin,H., Saunders,N.J., Heidelberg,J., Jeffries,A.C., Nelson,K.E., Eisen,J.A., Ketchum,K.A., Hood,D.W., Peden,J.F., Dodson,R.J., Nelson,W.C., Gwinn,M.L., DeBoy,R., Peterson,J.D., Hickey,E.K., Haft,D.H., Salzberg,S.L., White,O., Fleischmann,R.D., Dougherty,B.A., Mason,T., Ciecko,A., Parksey,D.S., Blair,E., Cittone,H., Clark,E.B., Cotton,M.D., Utterback,T.R., Khouri,H., Qin,H., Vamathevan,J., Gill,J., Scarlato,V., Masignani,V., Pizza,M., Grandi,G., Sun,L., Smith,H.O., Fraser,C.M., Moxon,E.R., Rappuoli,R., & Venter,J.C. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **287**, 1809-1815 (2000).

87. Schoen,C., Blom,J., Claus,H., Schramm-Gluck,A., Brandt,P., Muller,T., Goesmann,A., Joseph,B., Konietzny,S., Kurzai,O., Schmitt,C., Friedrich,T., Linke,B., Vogel,U., & Frosch,M. Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in *Neisseria meningitidis*. *Proc. Natl. Acad. Sci U. S. A* **105**, 3473-3478 (2008).

88. Gotschlich,E.C., Goldschneider,I., & Artenstein,M.S. Human immunity to the meningococcus. IV. Immunogenicity of group A and group C meningococcal polysaccharides in human volunteers. *J. Exp. Med.* **129**, 1367-1384 (1969).

89. Hotopp,J.C., Grifantini,R., Kumar,N., Tzeng,Y.L., Fouts,D., Frigimelica,E., Draghi,M., Giuliani,M.M., Rappuoli,R., Stephens,D.S., Grandi,G., & Tettelin,H. Comparative genomics of *Neisseria meningitidis*: core genome, islands of horizontal transfer and pathogen-specific genes. *Microbiology* **152**, 3733-3749 (2006).

90. Schoen,C., Tettelin,H., Parkhill,J., & Frosch,M. Genome flexibility in *Neisseria meningitidis*. *Vaccine* **27 Suppl 2**, B103-B111 (2009).

91. Feng, Lu, Reeves, Peter R, Lan, Ruiting, Ren, Yi, Gao, Chunxu, Zhou, Zhemin, Ren, Yan, Cheng, Jiansong, Wang, Wei, Wang, Jianmei, Qian, Wubin, Li, Dan, and Wang, Lei. A Recalibrated Molecular Clock and Independent Origins for the Cholera Pandemic Clones. PLoS ONE 3[12]. 2008. Public Library of Science.
Ref Type: Journal (Full)

92. Lin,W., Fullner,K.J., Clayton,R., Sexton,J.A., Rogers,M.B., Calia,K.E., Calderwood,S.B., Fraser,C., & Mekalanos,J.J. Identification of a *Vibrio cholerae* RTX toxin gene cluster that is tightly linked to the cholera toxin prophage. *Proc. Natl. Acad. Sci U. S. A* **96**, 1071-1076 (1999).

93. O'Shea,Y.A., Reen,F.J., Quirke,A.M., & Boyd,E.F. Evolutionary genetic analysis of the emergence of epidemic *Vibrio cholerae* isolates on the basis of comparative nucleotide sequence analysis and multilocus virulence gene profiles. *J. Clin. Microbiol* **42**, 4657-4671 (2004).

94. Safa,A., Bhuiyan,N.A., Alam,M., Sack,D.A., & Nair,G.B. Genomic relatedness of the new Matlab variants of *Vibrio cholerae* O1 to the classical and El Tor biotypes as determined by pulsed-field gel electrophoresis. *J. Clin. Microbiol* **43**, 1401-1404 (2005).

95. Keymer,D.P., Miller,M.C., Schoolnik,G.K., & Boehm,A.B. Genomic and phenotypic diversity of coastal *Vibrio cholerae* strains is linked to environmental factors. *Appl. Environ. Microbiol.* **73**, 3705-3714 (2007).

96. Davidsen,T. & Tonjum,T. Meningococcal genome dynamics. *Nat. Rev. Microbiol* **4**, 11-22 (2006).

97. Schoen,C., Joseph,B., Claus,H., Vogel,U., & Frosch,M. Living in a changing environment: insights into host adaptation in *Neisseria meningitidis* from comparative genomics. *Int. J. Med. Microbiol* **297**, 601-613 (2007).

98. Rice,P., Longden,I., & Bleasby,A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276-277 (2000).

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

100. Jordan,I.K., Rogozin,I.B., Wolf,Y.I., & Koonin,E.V. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res* **12**, 962-968 (2002).

101. Moreno-Hagelsieb,G. & Latimer,K. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* **24**, 319-324 (2008).

102. Tatusov,R.L., Koonin,E.V., & Lipman,D.J. A genomic perspective on protein families. *Science* **278**, 631-637 (1997).

103. Scheffler,K., Martin,D.P., & Seoighe,C. Robust inference of positive selection from recombining coding sequences. *Bioinformatics.* **22**, 2493-2499 (2006).

104. Studer,R.A. & Robinson-Rechavi,M. How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet.* **25**, 210-216 (2009).

105. Thompson,J.D., Higgins,D.G., & Gibson,T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-4680 (1994).

106. Yang,Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol Evol.* **24**, 1586-1591 (2007).

107. Zhou,C.E., Smith,J., Lam,M., Zemla,A., Dyer,M.D., & Slezak,T. MvirDB--a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res* **35**, D391-D394 (2007).

108. R Foundation for Statistical Computing, Vienna Austria. R Development Core Team (2005). R: A language and environment for statistical computing. 2009.
Ref Type: Computer Program

109. Creating the gene ontology resource: design and implementation. *Genome Res.* **11**, 1425-1433 (2001).

110. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., Harris,M.A., Hill,D.P., Issel-Tarver,L., Kasarskis,A., Lewis,S., Matese,J.C., Richardson,J.E., Ringwald,M., Rubin,G.M., & Sherlock,G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25-29 (2000).

111. Dennis,G., Jr., Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C., & Lempicki,R.A. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4**, 3 (2003).

112. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N., Rao,B.S., Smirnov,S., Sverdlov,A.V., Vasudevan,S., Wolf,Y.I., Yin,J.J., & Natale,D.A. The COG database: an updated version includes eukaryotes. *BMC. Bioinformatics.* **4**, 41 (2003).

113. Gardy,J.L., Spencer,C., Wang,K., Ester,M., Tusnady,G.E., Simon,I., Hua,S., deFays,K., Lambert,C., Nakai,K., & Brinkman,F.S. PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res* **31**, 3613-3617 (2003).

114. Gardy,J.L., Laird,M.R., Chen,F., Rey,S., Walsh,C.J., Ester,M., & Brinkman,F.S. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics.* **21**, 617-623 (2005).

115. Gardy,J.L. & Brinkman,F.S. Methods for predicting bacterial protein subcellular localization. *Nat. Rev. Microbiol* **4**, 741-751 (2006).

116. Rey,S., Acab,M., Gardy,J.L., Laird,M.R., deFays,K., Lambert,C., & Brinkman,F.S. PSORTdb: a protein subcellular localization database for bacteria. *Nucleic Acids Res.* **33**, D164-D168 (2005).

117. Caspi,R., Foerster,H., Fulcher,C.A., Kaipa,P., Krummenacker,M., Latendresse,M., Paley,S., Rhee,S.Y., Shearer,A.G., Tissier,C., Walk,T.C., Zhang,P., & Karp,P.D. The

MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* **36**, D623-D631 (2008).

118. Karp,P.D., Ouzounis,C.A., Moore-Kochlacs,C., Goldovsky,L., Kaipa,P., Ahren,D., Tsoka,S., Darzentas,N., Kunin,V., & Lopez-Bigas,N. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* **33**, 6083-6089 (2005).

119. Paley,S.M. & Karp,P.D. The Pathway Tools cellular overview diagram and Omics Viewer. *Nucleic Acids Res* **34**, 3771-3778 (2006).

120. Kanehisa,M. & Goto,S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).

121. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H., & Kanehisa,M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **27**, 29-34 (1999).

122. Glantz,S.A. Primer of biostatistics. (McGraw-Hill Medical Pub. Division, New York; 2005).

123. Sarakar, D. Lattice. 2(2), 19-23. 2002. R News.
Ref Type: Report

124. Kerkhoven,R., van Enckevort,F.H., Boekhorst,J., Molenaar,D., & Siezen,R.J. Visualization for genomics: the Microbial Genome Viewer. *Bioinformatics* **20**, 1812-1814 (2004).

125. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y., & Hattori,M. The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32**, D277-D280 (2004).

126. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T., Durbin,R., Eyras,E., Gilbert,J., Hammond,M., Hubbard,T., Kasprzyk,A., Keefe,D., Lehvaslaiho,H., Iyer,V., Melsopp,C., Mongin,E., Pettett,R., Potter,S., Rust,A., Schmidt,E., Searle,S., Slater,G., Smith,J., Spooner,W., Stabenau,A., Stalker,J., Stupka,E., Ureta-Vidal,A., Vastrik,I., & Birney,E. Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res* **31**, 38-42 (2003).

127. Tatusov,R.L., Koonin,E.V., & Lipman,D.J. A genomic perspective on protein families. *Science* **278**, 631-637 (1997).

128. Tatusov,R.L., Koonin,E.V., & Lipman,D.J. A genomic perspective on protein families. *Science* **278**, 631-637 (1997).

129. Bocs,S., Danchin,A., & Medigue,C. Re-annotation of genome microbial coding-sequences: finding new genes and inaccurately annotated genes. *BMC. Bioinformatics.* **3**, 5 (2002).

130. Nielsen,P. & Krogh,A. Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics.* **21**, 4322-4329 (2005).

131. Shriner,D., Nickle,D.C., Jensen,M.A., & Mullins,J.I. Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet. Res.* **81**, 115-121 (2003).

132. Yang,W., Bielawski,J.P., & Yang,Z. Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. *J. Mol. Evol.* **57**, 212-221 (2003).

133. Sterne,J.A. & Davey,S.G. Sifting the evidence-what's wrong with significance tests? *BMJ* **322**, 226-231 (2001).

134. Bjorkholm,B., Lundin,A., Sillen,A., Guillemin,K., Salama,N., Rubio,C., Gordon,J.I., Falk,P., & Engstrand,L. Comparison of genetic divergence and fitness between two subclones of *Helicobacter pylori*. *Infect. Immun.* **69**, 7832-7838 (2001).

135. Bjorkholm,B., Sjolund,M., Falk,P.G., Berg,O.G., Engstrand,L., & Andersson,D.I. Mutation frequency and biological cost of antibiotic resistance in *Helicobacter pylori*. *Proc. Natl. Acad. Sci U. S. A* **98**, 14607-14612 (2001).

136. Falush,D., Kraft,C., Taylor,N.S., Correa,P., Fox,J.G., Achtman,M., & Suerbaum,S. Recombination and mutation during long-term gastric colonization by *Helicobacter pylori:* estimates of clock rates, recombination size, and minimal age. *Proc. Natl. Acad. Sci U. S. A* **98**, 15056-15061 (2001).

137. Israel,D.A., Salama,N., Krishna,U., Rieger,U.M., Atherton,J.C., Falkow,S., & Peek,R.M., Jr. *Helicobacter pylori* genetic diversity within the gastric niche of a single human host. *Proc. Natl. Acad. Sci U. S. A* **98**, 14625-14630 (2001).

138. Kraft,C., Stack,A., Josenhans,C., Niehus,E., Dietrich,G., Correa,P., Fox,J.G., Falush,D., & Suerbaum,S. Genomic changes during chronic *Helicobacter pylori* infection. *J. Bacteriol.* **188**, 249-254 (2006).

139. Salama,N.R., Gonzalez-Valencia,G., Deatherage,B., Aviles-Jimenez,F., Atherton,J.C., Graham,D.Y., & Torres,J. Genetic analysis of *Helicobacter pylori* strain populations colonizing the stomach at different times postinfection. *J. Bacteriol.* **189**, 3834-3845 (2007).

140. Suerbaum,S., Smith,J.M., Bapumia,K., Morelli,G., Smith,N.H., Kunstmann,E., Dyrek,I., & Achtman,M. Free recombination within *Helicobacter pylori*. *Proc. Natl. Acad. Sci U. S. A* **95**, 12619-12624 (1998).

141. Wang,G., Humayun,M.Z., & Taylor,D.E. Mutation as an origin of genetic variability in *Helicobacter pylori*. *Trends Microbiol* **7**, 488-493 (1999).

142. Kersulyte,D., Chalkauskas,H., & Berg,D.E. Emergence of recombinant strains of *Helicobacter pylori* during human infection. *Mol. Microbiol* **31**, 31-43 (1999).

143. Lundin,A., Bjorkholm,B., Kupershmidt,I., Unemo,M., Nilsson,P., Andersson,D.I., & Engstrand,L. Slow genetic divergence of *Helicobacter pylori* strains during long-term colonization. *Infect. Immun.* **73**, 4818-4822 (2005).

144. Prouzet-Mauleon,V., Hussain,M.A., Lamouliatte,H., Kauser,F., Megraud,F., & Ahmed,N. Pathogen evolution in vivo: genome dynamics of two isolates obtained 9 years apart from a duodenal ulcer patient infected with a single *Helicobacter pylori* strain. *J. Clin. Microbiol* **43**, 4237-4241 (2005).

145. Achtman,M., Azuma,T., Berg,D.E., Ito,Y., Morelli,G., Pan,Z.J., Suerbaum,S., Thompson,S.A., van der,E.A., & van Doorn,L.J. Recombination and clonal groupings

within *Helicobacter pylori* from different geographical regions. *Mol. Microbiol* **32**, 459-470 (1999).

146. Han,S.R., Zschausch,H.C., Meyer,H.G., Schneider,T., Loos,M., Bhakdi,S., & Maeurer,M.J. *Helicobacter pylori*: clonal population structure and restricted transmission within families revealed by molecular typing. *J. Clin. Microbiol* **38**, 3646-3651 (2000).

147. Achtman,M. & Suerbaum,S. Sequence variation in *Helicobacter pylori*. *Trends Microbiol* **8**, 57-58 (2000).

148. Wang,I., I, Taylor,D.E., & Humayun,M.Z. Response from wang, humayun and taylor. *Trends Microbiol* **8**, 58 (2000).

149. Wang,I., I, Taylor,D.E., & Humayun,M.Z. Response from wang, humayun and taylor. *Trends Microbiol* **8**, 58 (2000).

150. Feil,E.J., Maiden,M.C., Achtman,M., & Spratt,B.G. The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Mol. Biol Evol.* **16**, 1496-1502 (1999).

151. Jolley,K.A., Sun,L., Moxon,E.R., & Maiden,M.C. Dam inactivation in *Neisseria meningitidis*: prevalence among diverse hyperinvasive lineages. *BMC Microbiol* **4**, 34 (2004).

152. Jolley,K.A., Wilson,D.J., Kriz,P., McVean,G., & Maiden,M.C. The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. *Mol. Biol Evol.* **22**, 562-569 (2005).

153. Maiden,M.C. Population genomics: diversity and virulence in the Neisseria. *Curr. Opin. Microbiol* **11**, 467-471 (2008).

154. Davidsen,T., Rodland,E.A., Lagesen,K., Seeberg,E., Rognes,T., & Tonjum,T. Biased distribution of DNA uptake sequences towards genome maintenance genes. *Nucleic Acids Res* **32**, 1050-1058 (2004).

155. Elkins,C., Thomas,C.E., Seifert,H.S., & Sparling,P.F. Species-specific uptake of DNA by gonococci is mediated by a 10-base-pair sequence. *J. Bacteriol.* **173**, 3911-3913 (1991).

156. Goodman,S.D. & Scocca,J.J. Identification and arrangement of the DNA sequence recognized in specific transformation of *Neisseria gonorrhoeae*. *Proc. Natl. Acad. Sci U. S. A* **85**, 6982-6986 (1988).

157. Bennett,J.S., Jolley,K.A., Sparling,P.F., Saunders,N.J., Hart,C.A., Feavers,I.M., & Maiden,M.C. Species status of Neisseria gonorrhoeae: evolutionary and epidemiological inferences from multilocus sequence typing. *BMC Biol* **5**, 35 (2007).

158. Kroll,J.S., Wilks,K.E., Farrant,J.L., & Langford,P.R. Natural genetic exchange between Haemophilus and Neisseria: intergeneric transfer of chromosomal genes between major human pathogens. *Proc. Natl. Acad. Sci U. S. A* **95**, 12381-12385 (1998).

159. Hanage,W.P., Fraser,C., & Spratt,B.G. The impact of homologous recombination on the generation of diversity in bacteria. *J. Theor. Biol* **239**, 210-219 (2006).

160. Treangen,T.J., Ambur,O.H., Tonjum,T., & Rocha,E.P. The impact of the neisserial DNA uptake sequences on genome evolution and stability. *Genome Biol* **9**, R60 (2008).

161. Caugant,D.A. Genetics and evolution of *Neisseria meningitidis*: importance for the epidemiology of meningococcal disease. *Infect. Genet. Evol.* **8**, 558-565 (2008).

162. Moxon,E.R., Rainey,P.B., Nowak,M.A., & Lenski,R.E. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol* **4**, 24-33 (1994).

163. Pang,B., Yan,M., Cui,Z., Ye,X., Diao,B., Ren,Y., Gao,S., Zhang,L., & Kan,B. Genetic diversity of toxigenic and nontoxigenic *Vibrio cholerae* serogroups O1 and O139 revealed by array-based comparative genomic hybridization. *J. Bacteriol.* **189**, 4837-4849 (2007).

164. Rocha,E.P. Inference and analysis of the relative stability of bacterial chromosomes. *Mol. Biol Evol.* **23**, 513-522 (2006).

165. Posada,D. & Crandall,K.A. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl. Acad. Sci. U. S. A* **98**, 13757-13762 (2001).

166. Huang,d.W., Sherman,B.T., & Lempicki,R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44-57 (2009).

167. Mobley,H.L.T., Mendz,G.L., & Hazell,S.L. Helicobacter pylori physiology and genetics. (ASM Press, Washington, DC; 2001).

168. Reynolds,D.J. & Penn,C.W. Characteristics of *Helicobacter pylori* growth in a defined medium and determination of its amino acid requirements. *Microbiology* **140 (Pt 10)**, 2649-2656 (1994).

169. Berg,J.M., Tymoczko,J.L., & Stryer,L. Biochemistry. (W.H. Freeman, New York; 2002).

170. Giglio,M.G., Collmer,C.W., Lomax,J., & Ireland,A. Applying the Gene Ontology in microbial annotation. *Trends Microbiol.* **17**, 262-268 (2009).

171. Nair,R. & Rost,B. Sequence conserved for subcellular localization. *Protein Sci* **11**, 2836-2847 (2002).

172. Nakai,K. & Kanehisa,M. Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins* **11**, 95-110 (1991).

173. Alm,E. & Arkin,A.P. Biological networks. *Curr. Opin. Struct. Biol* **13**, 193-202 (2003).

174. Barabasi,A.L. & Albert,R. Emergence of scaling in random networks. *Science* **286**, 509-512 (1999).

175. Barabasi,A.L. & Oltvai,Z.N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101-113 (2004).

176. Price,N.D., Reed,J.L., & Palsson,B.O. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol* **2**, 886-897 (2004).

177. Barabasi,A.L. Scale-free networks: a decade and beyond. *Science* **325**, 412-413 (2009).

178. Jeong,H., Tombor,B., Albert,R., Oltvai,Z.N., & Barabasi,A.L. The large-scale organization of metabolic networks. *Nature* **407**, 651-654 (2000).

179. Ravasz,E., Somera,A.L., Mongru,D.A., Oltvai,Z.N., & Barabasi,A.L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551-1555 (2002).

180. Fraser,H.B., Hirsh,A.E., Steinmetz,L.M., Scharfe,C., & Feldman,M.W. Evolutionary rate in the protein interaction network. *Science* **296**, 750-752 (2002).

181. Wittig,U. & De Beuckelaer,A. Analysis and comparison of metabolic pathway databases. *Brief. Bioinform.* **2**, 126-142 (2001).

182. Karp,P.D., Paley,S., & Romero,P. The Pathway Tools software. *Bioinformatics* **18 Suppl 1**, S225-S232 (2002).

183. Green,M.L. & Karp,P.D. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* **5**, 76 (2004).

184. Osterman,A. & Overbeek,R. Missing genes in metabolic pathways: a comparative genomics approach. *Curr. Opin. Chem. Biol* **7**, 238-251 (2003).

185. Benoit,S.L., Mehta,N., Weinberg,M.V., Maier,C., & Maier,R.J. Interaction between the *Helicobacter pylori* accessory proteins HypA and UreE is needed for urease maturation. *Microbiology* **153**, 1474-1482 (2007).

186. Sachs,G., Scott,D., Weeks,D., & Melchers,K. The compartment buffered by the urease of *Helicobacter pylori*: cytoplasm or periplasm? *Trends Microbiol* **10**, 217-218 (2002).

187. Bauerfeind,P., Garner,R., Dunn,B.E., & Mobley,H.L. Synthesis and activity of *Helicobacter pylori* urease and catalase at low pH. *Gut* **40**, 25-30 (1997).

188. Castelli,M.E., Fedrigo,G.V., Clementin,A.L., Ielmini,M.V., Feldman,M.F., & Garcia,V.E. Enterobacterial common antigen integrity is a checkpoint for flagellar biogenesis in *Serratia marcescens*. *J Bacteriol.* **190**, 213-220 (2008).

189. Inoue,T., Shingaki,R., Hirose,S., Waki,K., Mori,H., & Fukui,K. Genome-wide screening of genes required for swarming motility in *Escherichia coli* K-12. *J Bacteriol.* **189**, 950-957 (2007).

190. Merkx-Jacques,A., Obhi,R.K., Bethune,G., & Creuzenet,C. The *Helicobacter pylori flaA1* and *wbpB* genes control lipopolysaccharide and flagellum synthesis and function. *J Bacteriol.* **186**, 2253-2265 (2004).

191. Wu,B., Zhang,Y., & Wang,P.G. Identification and characterization of GDP-d-mannose 4,6-dehydratase and GDP-l-fucose snthetase in a GDP-l-fucose biosynthetic gene

cluster from *Helicobacter pylori*. *Biochem. Biophys. Res Commun.* **285**, 364-371 (2001).

192. Wu,B., Zhang,Y., Zheng,R., Guo,C., & Wang,P.G. Bifunctional phosphomannose isomerase/GDP-D-mannose pyrophosphorylase is the point of control for GDP-D-mannose biosynthesis in *Helicobacter pylori*. *FEBS Lett.* **519**, 87-92 (2002).

193. Eaton,K.A., Logan,S.M., Baker,P.E., Peterson,R.A., Monteiro,M.A., & Altman,E. *Helicobacter pylori* with a truncated lipopolysaccharide O chain fails to induce gastritis in SCID mice injected with splenocytes from wild-type C57BL/6J mice. *Infect. Immun.* **72**, 3925-3931 (2004).

194. Karlsson,K.A. The human gastric colonizer *Helicobacter pylori*: a challenge for host-parasite glycobiology. *Glycobiology* **10**, 761-771 (2000).

195. Moran,A.P. Relevance of fucosylation and Lewis antigen expression in the bacterial gastroduodenal pathogen *Helicobacter pylori*. *Carbohydr. Res* **343**, 1952-1965 (2008).

196. Allen,P.M., Fisher,D., Saunders,J.R., & Hart,C.A. The role of capsular polysaccharide K21b of Klebsiella and of the structurally related colanic-acid polysaccharide of *Escherichia coli* in resistance to phagocytosis and serum killing. *J Med. Microbiol* **24**, 363-370 (1987).

197. Majdalani,N. & Gottesman,S. The Rcs phosphorelay: a complex signal transduction system. *Annu. Rev. Microbiol* **59**, 379-405 (2005).

198. Stevenson,G., Andrianopoulos,K., Hobbs,M., & Reeves,P.R. Organization of the *Escherichia coli* K-12 gene cluster responsible for production of the extracellular polysaccharide colanic acid. *J Bacteriol.* **178**, 4885-4893 (1996).

199. Frosch,M., Muller,D., Bousset,K., & Muller,A. Conserved outer membrane protein of *Neisseria meningitidis* involved in capsule expression. *Infect. Immun.* **60**, 798-803 (1992).

200. Tzeng,Y.L., Swartley,J.S., Miller,Y.K., Nisbet,R.E., Liu,L.J., Ahn,J.H., & Stephens,D.S. Transcriptional regulation of divergent capsule biosynthesis and transport operon promoters in serogroup B *Neisseria meningitidis*. *Infect. Immun.* **69**, 2502-2511 (2001).

201. Uria,M.J., Zhang,Q., Li,Y., Chan,A., Exley,R.M., Gollan,B., Chan,H., Feavers,I., Yarwood,A., Abad,R., Borrow,R., Fleck,R.A., Mulloy,B., Vazquez,J.A., & Tang,C.M. A generic mechanism in *Neisseria meningitidis* for enhanced resistance against bactericidal antibodies. *J Exp. Med.* **205**, 1423-1434 (2008).

202. Lamm,M.E., Nedrud,J.G., Kaetzel,C.S., & Mazanec,M.B. IgA and mucosal defense. *APMIS* **103**, 241-246 (1995).

203. Lomholt,H., Poulsen,K., Caugant,D.A., & Kilian,M. Molecular polymorphism and epidemiology of *Neisseria meningitidis* immunoglobulin A1 proteases. *Proc. Natl. Acad. Sci U. S. A* **89**, 2120-2124 (1992).

204. Vitovski,S., Read,R.C., & Sayers,J.R. Invasive isolates of *Neisseria meningitidis* possess enhanced immunoglobulin A1 protease activity compared to colonizing strains. *FASEB J* **13**, 331-337 (1999).

205. Knaust,A., Weber,M.V., Hammerschmidt,S., Bergmann,S., Frosch,M., & Kurzai,O. Cytosolic proteins contribute to surface plasminogen recruitment of *Neisseria meningitidis*. *J Bacteriol.* **189**, 3246-3255 (2007).

206. Ullberg,M., Kuusela,P., Kristiansen,B.E., & Kronvall,G. Binding of plasminogen to *Neisseria meningitidis* and *Neisseria gonorrhoeae* and formation of surface-associated plasmin. *J Infect. Dis.* **166**, 1329-1334 (1992).

207. Hansen,J., Muldbjerg,M., Cherest,H., & Surdin-Kerjan,Y. Siroheme biosynthesis in *Saccharomyces cerevisiae* requires the products of both the MET1 and MET8 genes. *FEBS Lett.* **401**, 20-24 (1997).

208. Spencer,J.B., Stolowich,N.J., Roessner,C.A., & Scott,A.I. The *Escherichia coli cysG* gene encodes the multifunctional protein, siroheme synthase. *FEBS Lett.* **335**, 57-60 (1993).

209. Kolko,M.M., Kapetanovich,L.A., & Lawrence,J.G. Alternative pathways for siroheme synthesis in *Klebsiella aerogenes*. *J Bacteriol.* **183**, 328-335 (2001).

210. R.GRIFANTINI, E.BARTOLINI, A.MUZZI, M.DRAGHI, E.FRIGIMELICA, J.BERGER, F.RANDAZZO, & G.GRANDI Gene Expression Profile in *Neisseria meningitidis* and *Neisseria lactamica* upon Host-Cell Contact. *MICROARRAYS, IMMUNE RESPONSES, AND VACCINES* **975**, 202-216 (2002).

211. Menzel,R. & Roth,J. Purification of the putA gene product. A bifunctional membrane-bound protein from *Salmonella typhimurium* responsible for the two-step oxidation of proline to glutamate. *J Biol Chem.* **256**, 9755-9761 (1981).

212. Zhou,Y., Zhu,W., Bellur,P.S., Rewinkel,D., & Becker,D.F. Direct linking of metabolism and gene expression in the proline utilization A protein from *Escherichia coli*. *Amino. Acids* **35**, 711-718 (2008).

213. Genevrois,S., Steeghs,L., Roholl,P., Letesson,J.J., & van der,L.P. The Omp85 protein of *Neisseria meningitidis* is required for lipid export to the outer membrane. *EMBO J* **22**, 1780-1789 (2003).

214. Plant,L., Sundqvist,J., Zughaier,S., Lovkvist,L., Stephens,D.S., & Jonsson,A.B. Lipooligosaccharide structure contributes to multiple steps in the virulence of *Neisseria meningitidis*. *Infect. Immun.* **74**, 1360-1367 (2006).

215. Swartley,J.S., Marfin,A.A., Edupuganti,S., Liu,L.J., Cieslak,P., Perkins,B., Wenger,J.D., & Stephens,D.S. Capsule switching of *Neisseria meningitidis*. *Proc. Natl. Acad. Sci U. S. A* **94**, 271-276 (1997).

216. Tanner,M.E. The enzymes of sialic acid biosynthesis. *Bioorg. Chem.* **33**, 216-228 (2005).

217. Peabody,C.R., Chung,Y.J., Yen,M.R., Vidal-Ingigliardi,D., Pugsley,A.P., & Saier,M.H., Jr. Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella. *Microbiology* **149**, 3051-3072 (2003).

218. Sandkvist,M. Biology of type II secretion. *Mol. Microbiol* **40**, 271-283 (2001).

219. Wooldridge,K. Bacterial secreted proteins secretory mechanisms and role in pathogenesis. (Caister Academic Press, Wymondham; 2009).

220. Kahler,C.M., Martin,L.E., Tzeng,Y.L., Miller,Y.K., Sharkey,K., Stephens,D.S., & Davies,J.K. Polymorphisms in pilin glycosylation Locus of *Neisseria meningitidis* expressing class II pili. *Infect. Immun.* **69**, 3597-3604 (2001).

221. van Deuren,M., Brandtzaeg,P., & van der Meer,J.W. Update on meningococcal disease with emphasis on pathogenesis and clinical management. *Clin. Microbiol Rev.* **13**, 144-66, table (2000).

222. Pflughoeft,K.J., Kierek,K., & Watnick,P.I. Role of ectoine in *Vibrio cholerae* osmoadaptation. *Appl Environ Microbiol* **69**, 5919-5927 (2003).

223. Shikuma,N.J. & Yildiz,F.H. Identification and characterization of OscR, a transcriptional regulator involved in osmolarity adaptation in *Vibrio cholerae*. *J Bacteriol.* **191**, 4082-4096 (2009).

224. Bina,J., Zhu,J., Dziejman,M., Faruque,S., Calderwood,S., & Mekalanos,J. ToxR regulon of *Vibrio cholerae* and its expression in vibrios shed by cholera patients. *Proc. Natl. Acad. Sci U. S. A* **100**, 2801-2806 (2003).

225. Xu,Q., Dziejman,M., & Mekalanos,J.J. Determination of the transcriptome of *Vibrio cholerae* during intraintestinal growth and midexponential phase in vitro. *Proc. Natl. Acad. Sci U. S. A* **100**, 1286-1291 (2003).

226. Krishnan,H.H., Ghosh,A., Paul,K., & Chowdhury,R. Effect of anaerobiosis on expression of virulence factors in *Vibrio cholerae*. *Infect. Immun.* **72**, 3961-3967 (2004).

227. Leclerc,G.J., Tartera,C., & Metcalf,E.S. Environmental regulation of *Salmonella typhi* invasion-defective mutants. *Infect. Immun.* **66**, 682-691 (1998).

228. Campbell,J.W., Morgan-Kiss,R.M., & Cronan,J.E., Jr. A new *Escherichia coli* metabolic competency: growth on fatty acids by a novel anaerobic beta-oxidation pathway. *Mol. Microbiol* **47**, 793-805 (2003).

229. Clark,D. Regulation of fatty acid degradation in *Escherichia coli*: analysis by operon fusion. *J Bacteriol.* **148**, 521-526 (1981).

230. Morgan-Kiss,R.M. & Cronan,J.E. The *Escherichia coli fadK* (*ydiD*) gene encodes an anerobically regulated short chain acyl-CoA synthetase. *J Biol Chem.* **279**, 37324-37333 (2004).

231. Hutton,C.A., Perugini,M.A., & Gerrard,J.A. Inhibition of lysine biosynthesis: an evolving antibiotic strategy. *Mol. Biosyst.* **3**, 458-465 (2007).

232. Rodionov,D.A., Vitreschak,A.G., Mironov,A.A., & Gelfand,M.S. Regulation of lysine biosynthesis and transport genes in bacteria: yet another RNA riboswitch? *Nucleic Acids Res* **31**, 6748-6757 (2003).

233. Velasco,A.M., Leguina,J.I., & Lazcano,A. Molecular evolution of the lysine biosynthetic pathways. *J Mol. Evol.* **55**, 445-459 (2002).

234. Mey,A.R., Wyckoff,E.E., Oglesby,A.G., Rab,E., Taylor,R.K., & Payne,S.M. Identification of the *Vibrio cholerae* enterobactin receptors VctA and IrgA: IrgA is not required for virulence. *Infect. Immun.* **70**, 3419-3426 (2002).

235. Mey,A.R., Wyckoff,E.E., Kanukurthy,V., Fisher,C.R., & Payne,S.M. Iron and fur regulation in *Vibrio cholerae* and the role of fur in virulence. *Infect. Immun.* **73**, 8167-8178 (2005).

236. Miethke,M. & Marahiel,M.A. Siderophore-based iron acquisition and pathogen control. *Microbiol Mol. Biol Rev.* **71**, 413-451 (2007).

237. Wyckoff,E.E., Stoebner,J.A., Reed,K.E., & Payne,S.M. Cloning of a *Vibrio cholerae* vibriobactin gene cluster: identification of genes required for early steps in siderophore biosynthesis. *J Bacteriol.* **179**, 7055-7062 (1997).

238. Wyckoff,E.E., Valle,A.M., Smith,S.L., & Payne,S.M. A multifunctional ATP-binding cassette transporter system from *Vibrio cholerae* transports vibriobactin and enterobactin. *J Bacteriol.* **181**, 7588-7596 (1999).

239. Crosa,J.H. & Walsh,C.T. Genetics and assembly line enzymology of siderophore biosynthesis in bacteria. *Microbiol Mol. Biol Rev.* **66**, 223-249 (2002).

240. Wyckoff,E.E., Smith,S.L., & Payne,S.M. VibD and VibH are required for late steps in vibriobactin biosynthesis in *Vibrio cholerae*. *J Bacteriol.* **183**, 1830-1834 (2001).

Appendix A1

Positive Selection Pipeline PERL Scripts

Fasta_split.pl

```
#!/usr/bin/perl
# Alan Christoffels- 14.01.99
# fasta_split.pl - script to read in a fasta formatted file of sequences and
# print out each sequence in a separate file where the file names are
# sequencename.fasta
use warnings;

$/="\n>";        # set my deliminator to be a new line with a > fatsa header
$file = $ARGV[0];      # specify the input file as the first file o the command line
$dir = $ARGV[1];      # specify the output directory to write out the sequences
$i= 0;                 # initialise my counter to 0
system "mkdir $dir";   # create my output directory specified on the command line
open (FILE, "$file"); #open the input file
while (<FILE>) {       # while in the input file
  $i++;                # keep reading all the sequences
  chop $_;              # remove stop codon * from end of sequence

  if ($i > 1) {          # as soon as my sequence count reaches 1, add fasta headers
    $_ = ">".$_;
  }

  if ($_=~/^>/) {              # pattern match for fasta headers
    $_ =~/^>(\S+)\s+/;         # place gene accession into $_variable
    $name = $1;                # name my sequence using the accession ID
    $name =~ s/\|/_/g;         # substitute "pipes" for underscores
    open (F, ">$dir/$name.fasta");              #open my command line specified output file
    print F "$_\n";            # print my individual sequence file into a file named
    close(F);                  #after its accession ID with a .fasta file extension
  }
}
close(FILE);
```

Ortho_BLAST.pl

```
#! /usr/bin/perl
# subroutine to parse BLAST results
use warnings;

if ($#ARGV < 2) {              #print warnings  on command line if I forget an input file
  print "usage: ortho_BLAST.pl [input dir] [database1] [output dir]\n";
  exit;
}

$dir = $ARGV[0];               # intialise my input directory, database and output directory
$dbase1 = $ARGV[1];
$outdir = $ARGV[2];

system "mkdir $outdir";                # unix system command to make my output directory
open (CLUSTERS, ">clusters");                  # print paired accessions to file called clusters

opendir(DIR, "$dir");          # open my input directory to read sequence files
@files = readdir(DIR);         # read sequence files
foreach $file(@files) {        # for each sequence file read, run a BLAST using the
  next unless (-f "$dir/$file");
  # unix system call and BLAST commands below
  system "blastall -p blastp -d ./$dbase1 -i $dir/$file -v 1 -b 1 -F F > $outdir/$file.blast";
  parseBlast(); # call on my parseBlast sub-routine above to parse the BLAST results
}

sub parseBlast(){
  $count = 0;                                  # initialise counter to 0
  open (BLAST, "$outdir/$file.blast");         # open the directory containing BLAST results
  while (<BLAST>){
    s/\be\-/1e\-/g;
    @array =split;               # place my BLAST output results into an array
    if ($_ =~ /^Query\= /) {
      $query = $array[1];                # place the accession of my query gene into an array
    }
    elsif ($_ =~ /^>(.+)/) {             # if my BLAST hit has more then 1 sequence
      ++$count;                          #then continue counting
      if ($count == 1){
           $hit1 = $1;
      }
    }
    elsif ($_ =~ /Expect = (\S+)/){
      if ($count == 1){
           $E1 = $1;
      }
    }

    if ($_ =~ /Identities \= \d+\/\d+ \((\d+)\%.*\((\d+)\%\)/) {
      if ($count == 1){
           $percent1 = $1;
      }
```

```
    }
    elsif ($_ =~ /No hits/)  {
      $percent1 = 0;
    }
  }
  close BLAST;
  # print "$query $hit1 $E1 $percent1\n";
  # parse my BLAST results for and expectation score greeter then or equal to 1e10
  # and greater then or equal to 50% coverage and similarity
  # also create paired accession file clusters for reciprocal BLASTs
  if ($E1 < 1E-10 && $percent1 >= 50){
    print CLUSTERS "$query $hit1\n";
  }else{
    system "rm -fr $outdir/$file.blast";
  }
}
```

reciprocal_gene_ID.pl

```
#!/usr/bin/perl
use warnings;
# initialise my paired accessions into variables
my $acc1 = ();
my $acc2 = ();
my $acc3 = ();
my $acc4 = ();
# open my first paired accession file
open(FILE1,$ARGV[0]) or die "Couldn't open file 1\n";

while (<FILE1>) {
  chomp;
  @array1 = split;              # split on white space to get two accessions per line
  my $acc1 = $array1[0];        #Strain A accession ID
  my $acc2 = $array1[1];        #Strain B accession ID

  open(FILE2,$ARGV[1]) or die "Couldn't open file 1\n";
  while (<FILE2>) {
    chomp;
    @array2 = split;
    my $acc3 = $array2[0];              # Strain B accession ID
    my $acc4 = $array2[1];              # Strain A accession ID

    if(($acc1 eq $acc4) && ($acc2 eq $acc3)){       # If the paired accession ID
      print "$acc1 $acc3\n";                        # from list 1 match list 2 then print orthologous
    }                                               # paired accessions for use
  }
  close FILE2;
}

close FILE1;
```

accession_to_gene.pl

```
#!/usr/bin/perl
use warnings;

if ($#ARGV < 3){
  die "\n\nusage: makeClusters.pl <blastclust_outfile> <DNA-table format> <PROTEIN-TABLE
FORMAT> <output dir> \n\n\t\tNote: Sequence files in TABLE FORMAT eg. \nseq1
AGTCCATTCCAGG\nseq2  AGTGGGGGCAGG\n";
}

# initialise my variables
$clusterfile = $ARGV[0];       # paired orthlogous accession ID list
$dna = $ARGV[1];               # DNA sequences strain A and B in Table format
$protein = $ARGV[2];           # Protein sequences strain A and B in Table format
$outdir = $ARGV[3];            # Output directory

system "mkdir $outdir";        # Unix system call to create output directory

#Read DNA sequence and accession IDs into a hash table for quick retrieval
open(DNA, "$dna");
%dna_seqs = ();
while(<DNA>){
  @array1 = split;
  $dna_seqs{$array1[0]} = $array1[1];
}
close DNA;

#Read protein sequence and accession IDs into a hash table for quick retrieval
open(PROT, "$protein");
%prot_seqs = ();
while(<PROT>){
  @array2 = split;
  $prot_seqs{$array2[0]} = $array2[1];
}
close PROT;

#Grab each accession ID from paired accession cluster file and fetch
#sequences from hash table using accession IDs as a key from hash tables
#Write to a modified fasta file with each sequence on one line
open(CLUSTS, "$clusterfile");
while(<CLUSTS>){
  @array = split;
  $name1 = $array[0];
  $name2 = $array[1];
  ++$num;
  $outfile = "cluster_" . "$num";
  # write out .dna amd .prot files into outfile in directory
  mkdir("$outdir/$outfile",0777);
  open(DNA_OUT, ">$outdir/$outfile/$outfile.dna");
  open(PROT_OUT, ">$outdir/$outfile/$outfile.prot");
  # pull out my DNA and amino acid sequences for the given accession ID key for strain A
  foreach $name(@array){
```

```
  if ($name =~ /$name1/){
    $dnaseq1 = $dna_seqs{$name};
    for ($i = 1, $i <= 3, ++$i) {
        chop($dnaseq1);                 # remove stop codons for CODEML
    }
    $dna1 = ">$name\n$dnaseq1\n";
    $prot1 = ">$name\n$prot_seqs{$name}\n";
  }
  # pull out my DNA and amino acid sequences for the given accession ID key for
  # strain B
  elsif ($name =~ /$name2/){
    $dnaseq2 = $dna_seqs{$name};
    for ($i = 1, $i <= 3, ++$i) {
        chop($dnaseq2);
    }
    $dna2 = ">$name\n$dnaseq2\n";
    $prot2 = ">$name\n$prot_seqs{$name}\n";
  }
 }
 #print my orthologous DNA sequences into a .dna file
 #print my orthologous amino acid sequences in a .prot file
 print DNA_OUT "$dna1$dna2";
 print PROT_OUT "$prot1$prot2";
 close DNA_OUT;
 close PROT_OUT;
}
```

In_frame_codon_align.pl

```
#!/usr/bin/perl
use warnings;

$maindir = getcwd();
$dir = $ARGV[0];
# open my working directory with all the orthologous sequences identified by
# reciporcal BLAST and pattern match for directories called cluster_
opendir(DIR, "$maindir/$dir") or die "can't opendir";
while (defined($nextdir = readdir(DIR))) {
  if ($nextdir =~ /cluster_.+/)
    {
      print $nextdir;
      in_frame_codon_align("$maindir/$dir/$nextdir");
    }
}

closedir(DIR);

exit;
#sub-routine to generate ClustalW alignments and in-frame codon alignments
# and convert inframe codon aligned sequences into Phylip format
sub in_frame_codon_align{
  my ($nextdir) = $_[0];

  print "********** $nextdir\n";
  @dir_array = split(/\//, $nextdir);
  $name = $dir_array[$#dir_array];

  #Run ClustalW for alignment of protein sequences using a unix system call
  system("clustalw  -gapopen=40 -output=gde -outorder=input -outfile=$nextdir/$name.gde -
infile=$nextdir/$name.prot");

  #Run clustal for viewing alignment quality control checking
  system ("clustalw -gapopen=40 -outorder=input -infile=$nextdir/$name.prot");
  # Run sed expressions to replace : and % characters from ClustalW alignmnets
  # with and underscore and a fasta header
  system ("sed 's/:/_/g' $nextdir/$name.dna > $nextdir/dna_file");
  system("sed 's/%/>/g' $nextdir/$name.gde > $nextdir/protalign; sed 's/:/_/g' $nextdir/$name.dna
> $nextdir/dna_file; cat $nextdir/protalign $nextdir/dna_file > $nextdir/catted");

  #Use the ClustalW amino acid alignment as a template for EMBOSS's tranalign
  #to place nucleotide codons onto amino acids generating in-frame codon alignments
  #for CODEML.
  system ("tranalign -asequence $nextdir/$name.dna -bsequence $nextdir/protalign -table 11 -
outseq $nextdir/pre_infile");
  #convert DNA alignment to Phylip format for CODEML using readseq
  system("readseq -a -f=11 $nextdir/pre_infile > $nextdir/infile");
}
```

codeML.pl

```
#!/usr/bin/perl
use warnings;
use Cwd;

$maindir = getcwd();
# initialise my directory to be specified from the command line
$dir = $ARGV[0];
# create a progress log to determine if all directories have been processed by
# codeML
open (LOG, ">progress.log_validation");
opendir(DIR, "$maindir/$dir") or die "can't opendir";
# Look for directories called cluster_ containing in-frame codon aligned files
while (defined($nextdir = readdir(DIR))) {
  if ($nextdir =~ /cluster_.+/)
    {
      print $nextdir;
      run_codeML("$maindir/$dir/$nextdir");
    }
}

closedir(DIR);

exit;
# define my subroutine run_codeML to feed my infile to codeML and a pre-specifed
# codeML control file specifying models M1a and M2a
sub run_codeML{
  my ($nextdir) = $_[0];

  print "********** $nextdir\n";
  @dir_array = split(/\//, $nextdir);

  #temporarily copy Phylip format infile from working directory
  system ("cp $nextdir/infile $maindir");

  #Run codeML
  system ("codeml pipe.ctl");

  #move codeML results file to working directory
  system ("mv outfile_results 2ML.ds 2ML.dn $nextdir");
  print LOG "$nextdir processed\n";
}
```

parse_codeML.pl

```
#!/usr/bin/perl
use warnings;

$maindir = getcwd();
#Initialise my working directories and make an ouput directory containing parsed
# CODEML results
$dir = $ARGV[0];
$outdir = $ARGV[1];
system "mkdir $outdir";
opendir(DIR, "$maindir/$dir") or die "can't opendir";
while (defined($nextdir = readdir(DIR))) {
  if ($nextdir =~ /cluster_.+/)
    {
      print $nextdir, "\n" ;
      parse_codeml("$maindir/$dir/$nextdir", "$maindir/$outdir/$nextdir");
    }
}

closedir(DIR);

exit;
# Define my sub-routine to parse the CODEML results
sub parse_codeml{
  # Create CODEML results in output directory maintaining the same directory
  # structure as they were read in so I can track the cluster ID number
  my ($nextdir) = $_[0];
  my ($outdir) = $_[1];
  my @accessions = () ;
  # Open the CODEML results file
  open(CODEML, "$nextdir/outfile") || next;

  # Pattern match the CODEML results files for M1a and M2a reuslts and place into
  # a hash for look up
  while(<CODEML>) {
    if(/^#\d:\s(.+)\n/){
      push(@accessions, $1) ;
    }
    @array = split;
    if ($_ =~ /^Model\s[1|2]/){
      $w2 = $array[2] ;
    }
    elsif ($_ =~ /lnL/){
      chomp;
      $ratio = $array[4];
      if(defined $w2){
           $hash{$w2} = $ratio ;
      }
    }
  }
  close CODEML;
  # Determine if value of the LRT test is greater then 5.99
```

```
  # (P = 0.05 Chi-Square critical value with two degrees of freedom)
  if(log_likelihood(\%hash) > 5.99){
    open(RESULT_FILE, ">>$outdir") || die "Can't create file $outdir";
    print RESULT_FILE "@accessions\n" ;
    foreach (keys %hash){
      print RESULT_FILE "$_ $hash{$_}\n" ;
    }
    # If result of LRT greater then 5.99 then print paired accessions, Log liklihoods
    # for M1a and M2a and LRT value to a reults file
    print RESULT_FILE log_likelihood(\%hash), "\n" ;
    close RESULT_FILE;
  }
}

# Sub-routine to print paired accession IDs obtained from CODEML results file
sub print_hash{
  my %hash = $_[0] ;
  foreach (keys %hash){
    print "$_ $hash{$_}\n" ;
  }
}

# Sub-routine to perform the LRT test on log liklihoods of M1a and M2a models.
sub log_likelihood{
  my $hashref = $_[0] ;
  my %hash = %$hashref ;
  my $diff = 2*($hash{"PositiveSelection"} - ($hash{"NearlyNeutral"})) ;
  return($diff) ;
}
```

gene_annotation.pl

```
#!/usr/bin/env perl
use warnings;
use File::Find;
$annotation = shift;              # take in the genbank .ptt annotation file from the command line
find(\&wanted, ".");              # sub-routine for File::Find that specifies to look in the directory
                                  #being worked

# Define my sub routine to open parsed codeML files and pattern
# match accession IDs for strain A and B
sub wanted {
  my $file  = $_;
  # "-d" option below confirms that file is NOT a directory
  # if "-d" is true, then my "$_" is a directory, exit
  return if -d $_;
  return unless /^cluster_/;      # exit if no files named  "*cluster_*" exist
  open FH, $file or die "Can not open file :-(", $!;
  my $str = <FH>;
  my ($acc) = $str =~ /(HP\w+)/; # finds the query accession number for H. pylori 26695
# my ($acc) = $str =~ /(jhp\w+)/; # finds the query accession number for H. pylori J99
# my ($acc) = $str =~ /(NMA\w+)/; # finds the query accession number for N. meningitidis
Z2491
# my ($acc) = $str =~ /(NMB\w+)/; # finds the query accession number for N. meningitidis
MC58
# my ($acc) = $str =~ /(VC\w+)/; # finds the query accession number for V. cholerae N16961
# my ($acc) = $str =~ /(0395_\w+)/; # finds the query accession number for V. cholerae O395
  close FH;
  parse_annot_file($acc);
}

sub parse_annot_file {
  # Find the query accession number in the annotation file specified on the command line
  my ($acc) = @_;
  open ANN, $annotation or die "Can not open file :-(", $!;
  while ($line = <ANN>) {
    print $line if $line =~ /$acc/;         # if my accession ID matches, print annotation file line
  }
}
```

Appendix A2

Positive Selection Pipeline Documentation

Pre-processing of GenBank Files

Nucleotide sequences of the bacterial genes are extracted from their GenBank file using EMBOSS's (Version 3.0.0) ExtractFeat utility with following commands:

> extractfeat [input GenBank file] -type gene –describe locus_tag –
> osformat2 fasta –supper1 > [Output file name]

The resulting output file will contain gene sequences for the bacterial organism with fasta headers that look similar to the *H pylori* 26695 example shown below:

```
>NC_000915_217_633 [gene] (locus_tag="HP0001") Helicobacter pylori 26695,
complete genome.
ATGGCGACACGAACTCAAGCCAGGGGGGCTGTGGTTGAATTGTTGTATGCGTTTGAGAGC
GGTAATGAAGAAATTAAAAAAATCGCTTCTAGCATGTTAGAAGAAAAAAAGATTAAAAAC
AACCAACTCGCTTTCGCTTTAAGCCTTTTTAATGGCGTGTTAGAAAAAATCAATGAAATT
GACGCCCTCATCGAGCCGCATTTAAAAGACTGGGATTTCAAGCGATTAGGGAGCATGGAA
AAGGCGATTTTACGCTTAGGAGCGTATGAAATTGGCTTCACGCCCACGCAAAACCCTATC
ATCATCAATGAATGCATAGAGCTTGGCAAACTCTACGCTGAGCCTAACACCCCTAAATTT
TTAAACGCTATCTTGGATTCTTTGAGCAAAAAGCTCACTCAAAAACCCTTGAATTGA
```

The fasta format headings for each gene need to be replaced for a number of reasons:

1) To remove any special characters that may be wrongly interpreted by any of the applications in the pipeline e.g. formatDB, PAML, transeq, ClustalW.
2) Truncate the header information to less then or equal to 10 characters as prerequisite for generating the Pyllip formatted files as input for PAML.
3) Allow a standardized accessioning system that can be used for accession matching and parsing for subsequent analyses.

The fasta format headings are modified using the following sed regular expression:

```
sed 's/NC_[0-9]*_[0-9]*_[0-9]* //g;s/\[gene]//g;s/locus_tag//g;s/Helicobacter
pylori 26695, complete genome.//g;s/[(]//g;s/[)]//g;s/=//g;s/"//g' [infile] >
[outfile]
```

The above sed regular expression results in the following output:

```
>HP0001
ATGGCGACACGAACTCAAGCCAGGGGGGCTGTGGTTGAATTGTTGTATGCGTTTGAGAGC
GGTAATGAAGAAATTAAAAAAATCGCTTCTAGCATGTTAGAAGAAAAAAAGATTAAAAAC
AACCAACTCGCTTTCGCTTTAAGCCTTTTTAATGGCGTGTTAGAAAAAATCAATGAAATT
GACGCCCTCATCGAGCCGCATTTAAAAGACTGGGATTTCAAGCGATTAGGGAGCATGGAA
AAGGCGATTTTACGCTTAGGAGCGTATGAAATTGGCTTCACGCCCACGCAAAACCCTATC
ATCATCAATGAATGCATAGAGCTTGGCAAACTCTACGCTGAGCCTAACACCCCTAAATTT
TTAAACGCTATCTTGGATTCTTTGAGCAAAAAGCTCACTCAAAAACCCTTGAATTGA
```

Quality control checks preformed to determine if EMBOSS's ExtractFeat utility functions correctly include selecting 15 randomly chosen gene sequences from the unprocessed GenBank file and aligning those gene sequences with the corresponding gene sequences obtained by ExtracFeat to determine if there are any differences in nucleotide sequence length and characters. The two sets of sequences showed a 100% match.

Quality control checks using alignment methods to determine if the sed regular expression affected portions other then the fasta header were preformed, sequences pre and post sed fasta header modification displayed a 100% sequence match.

The result of the extraction and sed processing is a file of nucleotide sequences for each bacterial gene which contains the only gene's accession ID as the fasta header. This gene accession ID is universally recognized by all databases and can be used for querying different resources such as NCBI's Gene, PSORT, BioCyc etc.

The next step in pre-processing of the sequence files is to translate the extracted nucleotide fasta files into amino acid sequences to perform reciprocal BLASTs for orthologue identification and also use for generating in-frame codon alignments. EMBOSS's TranSeq utility is used with the following commands:

```
transeq [input_file] –table =11 > [output_file]
```

This will produce amino acid fasta sequences of the nucleotide sequences as shown below:

```
>HP0001_1
MATRTQARGAVVELLYAFESGNEEIKKIASSMLEEKKIKNNQLAFALSLFNGVLEKINEI
DALIEPHLKDWDFKRLGSMEKAILRLGAYEIGFTPTQNPIIINECIELGKLYAEPNTPKF
LNAILDSLSKKLTQKPLN*
```

The following sed regular expression is used to remove the "_1" suffix added by EMBOSS's TranSeq utility as subsequent steps rely on pattern matching of the accession headers to obtain orthologous nucleotide sequences and align them to their amino acid counterparts for positive selection analysis.

```
s'/_1 $//g' [infile] > [outfile]
```

The following out-file will contain fasta sequences without the "_1" suffix as shown below:

```
>HP0001
MATRTQARGAVVELLYAFESGNEEIKKIASSMLEEKKIKNNQLAFALSLFNGVLEKINEI
DALIEPHLKDWDFKRLGSMEKAILRLGAYEIGFTPTQNPIIINECIELGKLYAEPNTPKF
LNAILDSLSKKLTQKPLN*
```

Both the processed nucleotide and amino acid files contain all the genes for the bacterial strains' genome in fasta format and are ready for use in the pipeline.

Creation of Sequence and BLAST Databases.

After pre-processing and obtaining the nucleotide and amino acid files for each of the bacterial strains, a sub-directory should be created which houses the sequences.

Two blast sub-directories should be created, each one housing a bacterial strain's amino acid file. Two separate sub-directories should also be created that will each contain individual amino acid sequences in separate files. The **Fasta_split.pl** perl script written by Dr. Alan Christoffels takes the multi-sequence amino acid file created in the pre-processing step as input and writes out each amino acid fasta sequence into an individual file within a sub-directory created as output. The perl script is run as follows:

```
perl Fasta_split.pl [Input_file] [Output_directory]
```

To check if the **Fasta_split.pl** perl script functions correctly, the number of sequences in the original input file is counted using the following unix command:

```
grep ">" [Input_file] | wc -l
```

The number of individual sequence files within the sub-directory is also counted using the following unix command;

```
ls | wc -l
```

In both instances, the number of fasta headers counted and the number of individual amino acid fasta files in the sub-directory created are the same. At the end of this step, they should be 2 sub-directories, one for each bacterial strain which contains individual files for each amino acid sequence of a bacterial strain.

The next step is to create the BLAST databases of the sequences for each of the 2 bacterial strains. The amino acid file containing all the amino acid sequences should be copied into the separate BLAST sub-directories, 1 in each BLAST sub-directory. The amino acid sequences are formatted to create the BLAST databases using the following command:

```
formatdb –i [Input_File] –p T
```

After running fomatDB, the formatdb_log file should be checked to see how many sequences were formatted and if this number corresponds to the number of sequences contained with the amino acid file as a control check.

Identification of Orthologous Sequences

The **ortho_BLAST.pl** perl script was written which takes in as input each amino acid individual file in the sub-directory of amino acid sequences and BLASTs each sequence identifying the best hit and as output places the alignment in an individual file in another sub-directory created specified by the user. The perl script is run as follows:

```
perl orthoBLAST.pl [input_directory] [blast_directory1/file_name]
[Output_directory]
```

The BLASTP parameters specified in the **ortho_BLAST.pl** perl script are ≥ 50 % identity and coverage, an expectation score of 1E-10 and a BLOSUM 62 matrix contained in the same directory the BLASTP searches are being conducted. In addition, **ortho_BLAST.pl** outputs a file called "clusters" which contains paired accessions of the best hit from genome A and genome B as shown below:

```
jhp0049 HP0057
jhp1311 HP1416
jhp1368 HP1475
jhp0151 HP0164
jhp0602 HP0657
jhp1414 HP1525
jhp1470 HP1562
jhp0254 HP0269
jhp0705 HP0768
jhp1066 HP1138
jhp0300 HP0613
```

At this point, multiple genes from strain A can have its best hit to a single database gene from strain B. Hence, the reciprocal best BLAST hit (RBH) method is used to identify orthologous genes. **ortho_BLAST.pl** should be opened in a text editor and the term "clusters" should be changed to "clusters_recip" or any other name other then clusters to ensure the first clusters file is not over-written.

ortho_BLAST.pl is re-run, this time reversing the input query sequences and the BLAST database, e.g if strain A formed the query and strain B formed the BLAST database, then strain B should know be the query and strain A the BLAST database.

This should provide a second list of paired accessions similar to the one shown below:

```
HP0367 jhp1014
HP0818 jhp0757
HP1179 jhp1105
HP0413 jhp0826
HP0874 jhp0808
HP1225 jhp1146
HP0009 jhp1164
HP0920 jhp0854
HP1281 jhp1202
HP0065 jhp0060
```

The two sets of paired accession lists are to be parsed against each other using the **reciprocal_gene_ID.pl** perl script as follows:

```
perl reciprocal_gene_ID.pl [cluster_file_1] [cluster_file_2] > [output_file]
```

The output file specified by the user will contain paired accessions of the best reciprocal BLASTP hit for each gene of strain A and strain B which are orthologous.

Creation of Sequence Files for Generating In-Frame Codon Alignments

The **accession_to_gene.pl** perl script is used to take as input the reciprocal paired accession file generated from the identification of orthologous sequences and matches them to the accessions for nucleotide and corresponding amino acid sequences from the Table formatted files and outputs the sequences to a pre-specified output directory. In the output directory, a number of subdirectories are created called cluster_No, one for each cluster of accessions from the input cluster file. In each cluster directory, there is a .dna and a .prot file, each containing 2 orthologous sequences, one bacterial strain A and B. The **accession_to_gene.pl** perl script used to generate these files is run as follows:

```
perl accession_to_gene.pl [clusterfile] [DNA_table_format_file] [Protein_table
_format_file] [Output_directory]
```

Generation of In-Frame Codon Alignments

The .prot and .dna files in table format are used to generate in-frame codon alignments by the **In_frame_codon_align.pl** perl script. The perl script uses CLUSTALW to generate alignments between the nucleotide sequences in the .dna files and amino acid sequences in the .prot file and outputs them in GDE format alignment viewing. The aligned amino acid and nucleotide files are concatenated and EMBOSS's TranAlign utility is used on these concatenated files to generate in-frame codon alignments based on the CLUSTALW (Version 1.83) alignments. The outputted in-frame codon aligned file is further processed by the ReadSeq programme called by **In_frame_codon_align.pl** to convert the out-putted in-frame codon aligned files into Phyllip format which is required as input to the PAML (Version 3.15) programme. The usage of the perl script is as follows:

```
perl In_frame_codon_align.pl [input_dir]
```

Detection of Positive Selection

The **codeML.pl** perl script is run on the directory housing the files generated from the previous step. **codeML.pl** script opens the final input file for PAML, runs the CODEML module on each alignment using a pre-specified control file. The control file for PAML specifies that both the M1a and M2a model should be run on the in-frame codon alignments and the tree file to use. The control file should be located within the directory the codeML.pl script is being run in and have the same name as it is given within the perl script. The input file as well as the ouput results files for parsing CODEML results must be specified in the actual CODEML control file and in the perl script. The usage of the perl script is as follows:

codeML.pl `[input_directory]`

The control file containing the parameters used and tree structure file are presented overleaf. The codeml run using the M1a (null model) and the M2a (selection model) takes on average 3 minutes to process a file: **That is 3mins x 1382 infiles = 4146 minutes / 60 = 69.1 hours / 24 = 2.8 days to run for *H pylori*.**

The script will write the results in the file specified which in this case is the outfile with some ancillary result files called rst, rst1 and rub. The file called outfile is the main results files. Progress log for validation. The CODEML control file and tree file used are presented overleaf.

CODEML Control File

```
      seqfile = infile * sequence data filename
     treefile = stewart.trees      * tree structure file name
     outfile = outfile          * main result file name

       noisy = 9  * 0,1,2,3,9: how much rubbish on the screen
     verbose = 2  * 0: concise; 1: detailed, 2: too much
     runmode = 0  * 0: user tree;  1: semi-automatic;  2: automatic
               * 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise

     seqtype = 1  * 1:codons; 2:AAs; 3:codons-->AAs
   CodonFreq = 0  * 0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table
*       ndata = 10
       clock = 0  * 0:no clock, 1:clock; 2:local clock; 3:CombinedAnalysis
      aaDist = 0  * 0:equal, +:geometric; -:linear, 1-6:G1974,Miyata,c,p,v,a
  aaRatefile = dat/jones.dat  * only used for aa seqs with model=empirical(_F)
               * dayhoff.dat, jones.dat, wag.dat, mtmam.dat, or your own

       model = 0
               * models for codons:
                   * 0:one, 1:b, 2:2 or more dN/dS ratios for branches
               * models for AAs or codon-translated AAs:
                   * 0:poisson, 1:proportional, 2:Empirical, 3:Empirical+F
                   * 6:FromCodon, 7:AAClasses, 8:REVaa_0, 9:REVaa(nr=189)

     NSsites = 1 2  * 0:one w;1:neutral;2:selection; 3:discrete;4:freqs;
               * 5:gamma;6:2gamma;7:beta;8:beta&w;9:beta&gamma;
               * 10:beta&gamma+1; 11:beta&normal>1; 12:0&2normal>1;
               * 13:3normal>0

       icode = 0  * 0:universal code; 1:mammalian mt; 2-10:see below
       Mgene = 0
               * codon: 0:rates, 1:separate; 2:diff pi, 3:diff kapa, 4:all diff
               * AA: 0:rates, 1:separate

   * fix_kappa = 0  * 1: kappa fixed, 0: kappa to be estimated
    *    kappa = 2  * initial or fixed kappa
   * fix_omega = 0  * 1: omega or omega_1 fixed, 0: estimate
    *    omega = .4 * initial or fixed omega, for codons or codon-based AAs

   * fix_alpha = 1  * 0: estimate gamma shape parameter; 1: fix it at alpha
      * alpha = 0. * initial or fixed alpha, 0:infinity (constant rate)
     * Malpha = 0  * different alphas for genes
      * ncatG = 8  * # of categories in dG of NSsites models

      * getSE = 0  * 0: don't want them, 1: want S.E.s of estimates
* RateAncestor = 1  * (0,1,2): rates (alpha>0) or ancestral states (1 or 2)

   Small_Diff = .5e-6
    cleandata = 0  * remove sites with ambiguity data (1:yes, 0:no)?
*  fix_blength = -1  * 0: ignore, -1: random, 1: initial, 2: fixed
       method = 0   * 0: simultaneous; 1: one branch at a time

* Genetic codes: 0:universal, 1:mammalian mt., 2:yeast mt., 3:mold mt.,
* 4: invertebrate mt., 5: ciliate nuclear, 6: echinoderm mt.,
* 7: euplotid mt., 8: alternative yeast nu. 9: ascidian mt.,
* 10: blepharisma nu.
* These codes correspond to transl_table 1 to 11 of GENEBANK.
```

Tree file

```
2  1     (1,2);
```

Parsing Positive Selection Analysis Results

Once the PAML results have been generated, **parse_codeML_LRT.pl** is used to parse the results. **codeML.pl** parses the results outfile for the log likelihood scores of the M1a and M2a models and use those values to perform an Logliklihood ratio test which is a statistical test determining the goodness of fit for the sequence data to the CODEML specified models (the neutral and selection model in this case). The formula for the test is :

$$LRT = 2*(lnL1-lnL2)$$

1nL1 being the log-likelihood for M1a and 1nL2 being the log-likelihood for the M2a model. **codeML.pl** automatically calculates this value and will only output the results if the value is greater then 5.99 which is the value obtained from a Chi-Square table with 2 degrees of freedom at the 0.05 probability level. These parameters can be changed within **codeML.pl**. The **codeML.pl** perl script is run as follows:

```
perl codeML.pl [input directory] [output directory]
```

Obtaining Bacterial Gene Annotations for genes under Positive Selection

In order to obtain the annotations for the result sequences, the .ptt files from the NCBI ftp directory was downloaded and parsed using **gene_annotation.pl** perl script. The **gene_annotation.pl** script and the .ptt file should be in the same directory as the CODEML results. The script will need to be run twice – once for each of the bacterial strains and for each run, the previous .ptt file and results file needs to be moved up one directory so as not to "confuse" the perl script, **gene_annotation.pl** is run as follows:

```
perl gene_annotation.pl [GenBank annotation file] > [output file]
```

Alternatively, the following unix command written in a bash shell can also be used to obtain bacterial gene annotations:

```
for s in `grep START results_parse/cluster_* | cut -f2 -d" " | sed
"s/[[:alpha:]]*//g"`; do grep $s hp_26695.ptt >>hhp_genes; done
```

The above unix command greps for the START position in a particular results file, splits the array on the second field based on space, a sed regular expression is used to remove any alphabetic characters before the start positions from the results file. The .ptt file saved as a .txt file is parsed against and the results are appended to file with the done command allowing the recursive processing of all the cluster files.

The results are a series of gene list that have been identified as having sites under positive selection and can be used for analysis.

To copy the directories with all the orthlogous alignments and full CODEML results etc of genes found to be under positive selection for further analysis, the following perl “one-liner” is used:

```
## Create a sequence_results directory in the Paml parsed results directory, copy and paste the one liner in the Paml parsed results directory.
perl -e 'foreach(@ARGV){system("cp $_ $_.dat");system("mv *.dat sequence_results/");}' cluster_*
```

External Dependencies / Applications.

- Bacterial GenBank sequences and annotation files
- EMBOSS
- Formatdb
- BLASTP version 2.2.4 [Aug-26-2002] or higher.
- ClustalW version 1.83
- Readseq (Feb 1993)
- PAML version 3.15, a control file and tree file.
- Perl
- Unix

Appendix A3A

COG_parser.pl

```
#!/usr/bin/perl
use warnings;
#Remove the first two line of the .ptt file or suppress warnings to avoid error messages which
#does not affect the results of the parser
#Open my file of paired accession IDs
open(FILE1,$ARGV[0]) or die "Couldn’t open file 1\n";

while (<FILE1>) {
        chomp;
        @array1 = split;
        my $acc1 = $array1[0];
#Open my GenBAnk COG annotation .ptt file
open(FILE2,$ARGV[1]) or die "Couldn’t open file 2\n";
        while (<FILE2>) {
        chomp;
        @array2 = split /\t/;              #Split columns on tabs
        my $acc_match = $array2[5]; # match my accession IDs to the sixth column in the table
        my $cog_annot = $array2[7];  # place the eighth column of the array containing my COG
annotations
        if($acc1 eq $acc_match){     # provide a condition for matching accession IDs
        print "$acc1\t";                  # Print my accession ID
        print "$acc_match\t";       # Print the accession ID matched in the .ptt COG file
        print "$cog_annot\n";       # Print my matching COG annotation
      }
         }
      close FILE2;
         }
close FILE1;
```

Appendix A3B

PSORT_parser.pl

```
#!/usr/bin/perl
use warnings;

# Open my file of paired Accession IDs
open(FILE1,$ARGV[0]) or die "Couldn't open file 1\n";

#Place my accession IDs into an array for matching
while (<FILE1>) {
chomp;
@array1 = split;
  my $acc1 = $array1[0];

#Open my PSORT annotation file
open(FILE2,$ARGV[1]) or die "Couldn't open file 2\n";
while (<FILE2>) {
chomp;
@array2 = split /\t/;                    #split columns of tab delimniator
        my $acc_match = $array2[3]; #define my accession IDs in the fourth column of the
annotation file
        my $protein_annot = $array2[1]; #define the protein annotation column in the array
        my $localisation = $array2[5];  # define the subcellular localisation site column in the
array
        my $prediction_score = $array2[6]; # define the coumn with the PSORT score column in
the array
        if($acc1 eq $acc_match){      # condition, if accession IDs match then print the
following fields.
        print "$acc1\t";
        print "$acc_match\t";
        print "$protein_annot\t";
        print "$localisation\t";
        print "$prediction_score\n";
  }
 }
  close FILE2;
 }
 close FILE1;
```

Appendix A4

Chi_Square R Functions

One-Way Classification Chi-Square Test R Function

An R function written to conduct one way classification Chi-Square tests to determine if genes for a particular gene list (genes under positive selection or database identified virulence genes) are enriched for a certain functional annotation category is presented below:

```
> one_way_classification_chi_function = function(gene_list_category_count,
genome_gene_count_for_category, whole_gene_list_count,
genome_count_of_all_genes_annot){
a1 = (gene_list_category_count)
a2 = (genome_gene_count_for_category - gene_list_category_count)
b1 = (whole_gene_list_count - gene_list_category_count)
b2 = genome_count_of_all_genes_annot - (a1 + a2 + b1)
row1 = c(a1,a2)
row2 = c(b1,b2)
contingency_table = rbind(row1,row2)
print(contingency_table)
print(chisq.test(contingency_table))
return(fisher.test(contingency_table))
}
```

Input to the one-way classification Chi-Square R function consists of gene counts for genes belonging to either the positive selection gene list or the database identified virulence genes, number of annotated gene list genes in that category, total number of gene list genes annotated by that system and the total number of genome genes annotated by that system. A listing of the input numbers is summarised below and overleaf.

1) gene_list_category_count = count of gene list genes present in that particular functional annotation category.

2) genome_gene_count_for_category = the total number of genome genes for a bacterial organism placed within that particular functional annotation category.

3) whole_gene_list_count = total number of gene list genes assigned to all functional annotation categories within that annotation system.

4) genome_count_of_all_genes_annot = number of genome genes for a species that have been classified into functional annotation categories within that annotation system.

The one-way Chi-Square classification R function uses gene counts provided by the user for annotation categories and systems to generate a 2 x 2 contingency table populated with the correct values for each cell, manual 2 x 2 contingency tables were constructed to double check the fidelity of the 2 x 2 contingency tables generated in R. A Chi-Square test with Yates continuity correction is performed on the 2 x 2 contingency tables and in cases where gene counts are below the recommended count of 5, a Fisher's Exact Test is also conducted.

A worked example using the one-way classification Chi-Square R function is presented below for *N. meningitidis* genes under positive selection belonging to a COG category;

Input Gene Counts:

Total Number of *N. meningitidis* genes under positive selection belonging to COG category C (Energy production and conversion) = **11 genes**

Total Number of *N. meningitidis* genome genes placed in COG category C = **105 genes**

Total number of *N. meningitidis* genes under positive selection with COG annotations = **199 genes**

Total number of *N. meningitidis* genome genes with COG annotations = **1356 genes**

```
>one_way_classification_chi_function(11,105,199,1356)
     [,1] [,2]
row1   11   94      # 2 x 2 contingency table generated for Chi-Square Test
row2  188 1063

        Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_table
X-squared = 1.2599, df = 1, p-value = 0.2617   # Results of Chi-Square Test


        Fisher's Exact Test for Count Data

data:  contingency_table
p-value = 0.2505        # Results for Fisher's Exact Test
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.3133919 1.2698440
sample estimates:
odds ratio
 0.661851
```

Two-Way Classification Chi-Square Test R Function

A two-way classification Chi-Square test was used to determine is database identified virulence genes and database identified virulence genes have a statistically significant intersection and association for a particular annotation category. The two-way classification Chi-Square R function written is presented below:

```
> two_way_classification_chi_function  = function(positive_selection_gene_count,
virulence_gene_count, gene_list_intersection, genome_number){
a1 = (gene_list_intersection)
a2 = (positive_selection_gene_count - gene_list_intersection)
b1 = (virulence_gene_count - gene_list_intersection)
b2 = genome_number-(a1+a2+b1)
row1 = c(a1,a2)
row2 = c(b1,b2)
contingency_table = rbind(row1,row2)
print(contingency_table)
print(chisq.test(contingency_table))
return(fisher.test(contingency_table))
```

Input gene numbers for the two-way classification Chi-Sqaure R function are as follows:

1) positive_selection_gene_count = number of genes under positive selection in bacterial species
2) virulence_gene_count = number of database identified virulence genes in bacterial species
3) gene_list_intersection = number of database identified virulence genes that are under positive selection and belong an annotation category
4) genome_number = number of genes in the genome that are annotated within a functional system.

The two-way classification Chi-Square R function uses the input gene numbers to generate a 2 x 2 contingency table which was double checked manually to ensure its correctness. A Chi-Square test is conducted on that 2 x 2 contingency table and in cases where gene counts are below 5, a Fisher's Exact Test is conducted. A worked example using the two-way classification Chi-Square R function is presented overleaf.

Input Gene Numbers

Number of *N. meningitidis* genes under positive selection = **218 genes**

Number of *N. meningitidis* database identified virulence genes = **68 genes**

Number of database identified virulence genes under positive selection (intersection between gene lists) = **14 genes**

Number of *N. meningitidis* genome genes = **2208 genes**

```
> two_way_classification_chi_function(218, 68, 14, 2208)
    [,1] [,2]
row1   14  204
row2   54 1936    # 2 x 2 contingency table

        Pearson's Chi-squared test with Yates' continuity correction
data:  contingency_table
X-squared = 7.8528, df = 1, p-value = 0.005074  # Results of Chi-Square Test


        Fisher's Exact Test for Count Data
data:  contingency_table
p-value = 0.006048        # Results for Fisher's Exact Test
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.238984 4.585353
sample estimates:
odds ratio
  2.459031
```

Appendix B1

Genes Under Positive Selection

H. pylori Genes Under Positive Selection

Table B1A (Part 1 of 7) : *H. pylori* Genes Under Positive Selection				
Cluster Identity	***H. pylori* 26695 Accessions**	***H. pylori* J99 Accessions**	**Enterez Gene Name**	**GenBank Annotation**
cluster_267	HP0022	jhp0020	eptA	hypothetical protein
cluster_844	***HP0069***	***jhp0064***	***ureF***	***urease accessory protein (ureF)***
cluster_688	HP0080	jhp0074		hypothetical protein
cluster_919	HP0088	jhp0081	rpoD	RNA polymerase sigma factor
cluster_792	HP0102	jhp0094		hypothetical protein
cluster_668	HP0120	jhp0110		hypothetical protein
cluster_1007	HP0206	jhp0192		hypothetical protein
cluster_1131	HP0238	jhp0223	proS	prolyl-tRNA synthetase
cluster_154	HP0252	jhp0237	omp7	outer membrane protein (omp7)
cluster_1401	HP0258	jhp0242		hypothetical protein
cluster_530	HP0298	jhp0283	dppA	"dipeptide ABC transporter, periplasmic dipeptide-binding protein (dppA)"
cluster_1155	HP0347	jhp0321		hypothetical protein
cluster_1	HP0367	jhp1014		hypothetical protein
cluster_32	HP0373	jhp1008		putative outer membrane protein
cluster_105	HP0392	jhp0989	cheA	histidine kinase (cheA)
cluster_1343	HP0398	jhp0983		hypothetical protein
cluster_1205	HP0462	jhp0414	hsdS	type I restriction enzyme S protein (hsdS)
cluster_936	HP0499	jhp0451	DR-phospholipase A	phospholipase A1 precursor (DR-phospholipase A)
cluster_817	HP0513	jhp0462		hypothetical protein
cluster_640	HP0519	jhp0468		hypothetical protein
cluster_1348	***HP0547***	***jhp0495***	***cag26 / cagA***	***cag pathogenicity island protein (cag26)***
cluster_344	***HP0651***	***jhp0596***	***fucT (j99)***	***"ALPHA (1,3)-FUCOSYLTRANSFERASE"***
cluster_22	HP0675	jhp0617	xerC	integrase/recombinase (xerC)
cluster_414	HP0717	jhp0655	dnaX (j99)	DNA polymerase III subunits gamma and tau
cluster_53	HP0728	jhp0665		hypothetical protein
cluster_884	HP0731	jhp0668		hypothetical protein
cluster_207	HP0772	jhp0709	amiA	N-acetylmuramoyl-L-alanine amidase (amiA)
cluster_267	HP0022	jhp0020	eptA	hypothetical protein

Table B1A (Part 2 of 7) : *H. pylori* Genes Under Positive Selection				
Cluster Identity	***H. pylori* 26695 Accessions**	***H. pylori* J99 Accessions**	**Enterez Gene Name**	**GenBank Annotation**
cluster_61	***HP0887***	***jhp0819***	***vacA***	***vacuolating cytotoxin***
cluster_974	HP0906	jhp0842		hypothetical protein
cluster_1365	HP0958	jhp0892		hypothetical protein
cluster_612	HP0973	jhp0907		hypothetical protein
cluster_93	HP0996	jhp0942		hypothetical protein
cluster_799	HP1017	jhp0406	rocE	amino acid permease (rocE)
cluster_439	HP1028	jhp0396		hypothetical protein
cluster_720	HP1048	jhp0377	infB	translation initiation factor IF-2
cluster_528	HP1053	jhp0372	minC	septum formation inhibitor
cluster_466	HP1090	jhp0335	ftsK	cell division protein (ftsK)
cluster_349	HP1105	jhp1032		LPS biosynthesis protein
cluster_287	HP1142	jhp1070		hypothetical protein
cluster_533	HP1156	jhp1083	omp25	outer membrane protein (omp25)
cluster_743	HP1157	jhp1084	omp26	outer membrane protein (omp26)
cluster_1019	HP1177	jhp1103	omp27	outer membrane protein (omp27)
cluster_1306	***HP1243***	***jhp0833***	***omp28 / babA***	***outer membrane protein (omp28)***
cluster_107	HP1250	jhp1171		hypothetical protein
cluster_515	HP1252	jhp1173	oppA	"oligopeptide ABC transporter, periplasmic oligopeptide-binding protein (oppA)"
cluster_638	HP1284	jhp1204		hypothetical protein
cluster_221	HP1329	jhp1249	czcA	cation efflux system protein (czcA)
cluster_750	HP1362	jhp1280	dnaB	replicative DNA helicase
cluster_1165	HP1364	jhp1282		"signal-transducing protein, histidine kinase"
cluster_647	HP1433	jhp1326		hypothetical protein
cluster_565	HP1470	jhp1363	polA	DNA polymerase I (polA)
cluster_1113	HP1550	jhp1449	SecD	protein export protein SecD
cluster_760	HP1568	jhp1476		hypothetical protein
cluster_833	HP1587	jhp1493		hypothetical protein
cluster_1327	HP0033	jhp0029	clpA	ATP-dependent C1p protease (clpA)
cluster_728	***HP0037***	***jhp0033***		***NADH-ubiquinone oxidoreductase subunit***
cluster_935	***HP0038***	***jhp0034***	***comB1***	***hypothetical protein***
cluster_356	HP0048	jhp0041	hypF	transcriptional regulator (hypF)
cluster_1389	HP0052	jhp0044		putative TYPE II DNA MODIFICATION ENZYME (METHYLTRANSFERASE)
cluster_404	HP0060	jhp0054		hypothetical protein
cluster_816	HP0062	jhp0057		hypothetical protein
cluster_1025	HP0063	jhp0058		hypothetical protein
cluster_11	HP0065	jhp0060		hypothetical protein
cluster_789	HP0056	jhp0048	putA	delta-1-pyrroline-5-carboxylate dehydrogenase

Table B1A (Part 3 of 7) : *H. pylori* Genes Under Positive Selection				
Cluster Identity	***H. pylori* 26695 Accessions**	***H. pylori* J99 Accessions**	**Enterez Gene Name**	**GenBank Annotation**
cluster_218	HP0066	jhp0061		conserved hypothetical ATP-binding protein
cluster_519	HP0092	jhp0085	hsdM	type II restriction enzyme M protein (hsdM)
cluster_540	HP0099	jhp0091	tlpA	methyl-accepting chemotaxis protein (tlpA)
cluster_379	HP0100	jhp0092		hypothetical protein
cluster_1003	HP0103	jhp0095	tlpB	methyl-accepting chemotaxis protein (tlpB)
cluster_1238	HP0110	jhp0102	grpE	co-chaperone and heat shock protein (grpE)
cluster_1369	HP0149	jhp0137		hypothetical protein
cluster_795	***HP0159***	***jhp0147***	***rfaJ***	***"lipopolysaccharide 1,2-glucosyltransferase (rfaJ)"***
cluster_1240	HP0167	jhp0153		hypothetical protein
cluster_645	HP0170	jhp0156		hypothetical protein
cluster_250	HP0175	jhp0161		cell binding factor 2
cluster_73	HP0180	jhp0168	lnt	apolipoprotein N-acyltransferase
cluster_1314	HP0186	jhp0174		hypothetical protein
cluster_924	HP0190	jhp0176		conserved hypothetical secreted protein
cluster_1128	HP0191	jhp0177	frdB (j99)	succinate dehydrogenase
cluster_1397	HP0201	jhp0187	plsX	fatty acid/phospholipid synthesis protein
cluster_1034	HP0212	jhp0198	dapE (j99)	succinyl-diaminopimelate desuccinylase
cluster_21	HP0214	jhp0200	huNaDC-1	sodium-dependent transporter (huNaDC-1)
cluster_1293	HP0226	jhp0211		hypothetical protein
cluster_302	HP0234	jhp0219		hypothetical protein
cluster_1336	HP0239	jhp0224	hemA	glutamyl-tRNA reductase
cluster_1063	HP0275	jhp0260	addB	ATP-dependent nuclease (addB)
cluster_488	HP0285	jhp0270		hypothetical protein
cluster_1403	HP0304	jhp0289		hypothetical protein
cluster_804	HP0308	jhp0293		hypothetical protein
cluster_1015	HP0309	jhp0294		hypothetical protein
cluster_887	***HP0327***	***jhp0310***	***flaG***	***flagellar protein G (flaG)***
cluster_1300	HP0329	jhp0312	nadE	NH(3)-dependent NAD+ synthetase (nadE)
cluster_491	HP0331	jhp0314	minD	cell division inhibitor (minD)
cluster_511	HP0338	jhp0320		hypothetical protein
cluster_551	HP0350	jhp0324		hypothetical protein
cluster_782	HP0358	jhp1022		hypothetical protein
cluster_1043	HP0371	jhp1010	fabE / accB	biotin carboxyl carrier protein (fabE)

Table B1A (Part 4 of 7) : *H. pylori* Genes Under Positive Selection				
Cluster Identity	***H. pylori* 26695 Accessions**	***H. pylori* J99 Accessions**	**Enterez Gene Name**	**GenBank Annotation**
cluster_446	HP0375	jhp1006		hypothetical protein
cluster_1070	HP0378	jhp1003	ycf5	cytochrome c biogenesis protein (ycf5)
cluster_162	HP0401	jhp0980	aroA	3-phosphoshikimate 1-carboxyvinyltransferase
cluster_837	HP0417	jhp0967	metG	methionine--tRNA ligase
cluster_1046	HP0418	jhp0966		hypothetical protein
cluster_449	HP0421	jhp0963	capJ	type 1 capsular polysaccharide biosynthesis protein J (capJ)
cluster_400	HP0465	jhp0417		hypothetical protein
cluster_663	HP0479	jhp0431		hypothetical protein
cluster_896	HP0486	jhp0438		hypothetical protein
cluster_1001	HP0508	jhp0458		hypothetical protein
cluster_219	HP0517	jhp0466	era	GTP-binding protein Era
cluster_92	***HP0535***	***jhp0483***	***cag14***	***cag pathogenicity island protein (cag14)***
cluster_147	HP0554	jhp0501		hypothetical protein
cluster_980	HP0558	jhp0505	fabB	3-oxoacyl-(acyl carrier protein) synthase
cluster_793	HP0563	jhp0510		hypothetical protein
cluster_615	HP0569	jhp0516	gtp1	GTP-binding protein (gtp1)
cluster_275	HP0586	jhp0534		hypothetical protein
cluster_618	HP0615	jhp0558	lig	DNA ligase
cluster_822	***HP0616***	***jhp0559***	***cheV***	***chemotaxis protein (cheV)***
cluster_1032	HP0617	jhp0560	aspS	aspartyl-tRNA synthetase
cluster_1241	HP0618	jhp0561	adk	adenylate kinase
cluster_1058	HP0623	jhp0567	murC	UDP-N-acetylmuramate--L-alanine ligase
cluster_878	HP0629	jhp0572		hypothetical protein
cluster_1291	HP0630	jhp0573	mda66	modulator of drug activity (mda66)
cluster_98	***HP0638***	***jhp0581***	***omp13***	***outer membrane protein (omp13)***
cluster_300	HP0639	jhp0582		hypothetical protein
cluster_544	HP0652	jhp0597	serB	phosphoserine phosphatase (serB)
cluster_1168	HP0655	jhp0600		protective surface antigen D15
cluster_853	***HP0679***	***jhp0620***	***wbpB***	***lipopolysaccharide biosynthesis protein (wbpB)***
cluster_462	HP0683	jhp0624	glmU	UDP-N-acetylglucosamine pyrophosphorylase (glmU)
cluster_1134	HP0745	jhp0682		hypothetical protein
cluster_531	HP0749	jhp0686	ftsX	cell division membrane protein (ftsX)
cluster_348	HP0754	jhp0691		hypothetical protein
cluster_600	HP0768	jhp0705	moaA	molybdenum cofactor biosynthesis protein A
cluster_469	***HP0786***	***jhp0723***	***secA***	***translocase***
cluster_286	HP0791	jhp0727	cadA / hmcT	"cadmium-transporting ATPase, P-type (cadA)"
cluster_696	HP0690	jhp0638	fadA / thl	acetyl coenzyme A acetyltransferase (thiolase) (fadA)
cluster_388	HP0710	jhp0649		putative Outer membrane protein
cluster_699	HP0737	jhp0674	pgpA	PHOSPHATIDYLGLYCEROPHOSPHATASE A

Table B1A (Part 5 of 7) : *H. pylori* Genes Under Positive Selection				
Cluster Identity	***H. pylori* 26695 Accessions**	***H. pylori* J99 Accessions**	**Enterez Gene Name**	**GenBank Annotation**
cluster_160	HP0806	jhp0742		hypothetical protein
cluster_575	HP0808	jhp0744	acpS	4'-phosphopantetheinyl transferase
cluster_783	***HP0809***	***jhp0745***	***fliL***	***flagellar basal body-associated protein***
cluster_1199	HP0810	jhp0746		hypothetical protein
cluster_395	HP0813	jhp0749		hypothetical protein
cluster_1044	HP0822	jhp0761	hom	homoserine dehydrogenase
cluster_447	HP0826	jhp0765	lex2B	lipooligosaccharide 5G8 epitope biosynthesis-associated protein (lex2B)
cluster_1118	HP0841	jhp0779	dfp	phosphopantothenoylcysteine synthase/decarboxylase
cluster_514	HP0845	jhp0783	thiM (j99)	hydroxyethylthiazole kinase
cluster_726	HP0846	jhp0784	hsdR	type I restriction enzyme R protein (hsdR)
cluster_534	HP0851	jhp0787		hypothetical protein
cluster_764	HP0859	jhp0793	rfaD / gmhD	ADP-L-glycero-D-mannoheptose-6-epimerase (rfaD)
cluster_1384	HP0861	jhp0795		hypothetical protein
cluster_373	HP0863	jhp0797		hypothetical protein
cluster_424	***HP0876***	***jhp0810***	***frpB***	***iron-regulated outer membrane protein (frpB)***
cluster_1073	HP0885	jhp0817	mviN	virulence factor mviN protein (mviN)
cluster_893	HP0890	jhp0823		hypothetical protein
cluster_706	***HP0896***	***jhp1164***	***omp19 / babB***	***outer membrane protein (omp19)***
cluster_1119	HP0898	jhp0835	hypD	hydrogenase expression/formation protein (hypD)
cluster_165	HP0909	jhp0845		hypothetical protein
cluster_787	HP0911	jhp0847	rep	"rep helicase, single-stranded DNA-dependent ATPase (rep)"
cluster_998	***HP0912***	***jhp0848***	***omp20 / hopC***	***outer membrane protein (omp20)***
cluster_1206	***HP0913***	***jhp0849***	***omp21 / hopB***	***outer membrane protein (omp21)***
cluster_216	HP0921	jhp0855	gap	glyceraldehyde-3-phosphate dehydrogenase (gap)
cluster_637	HP0923	jhp0857	omp22	outer membrane protein (omp22)
cluster_294	HP0940	jhp0875	yckK	"amino acid ABC transporter, periplasmic binding protein (yckK)"
cluster_110	HP0946	jhp0880		hypothetical protein
cluster_146	HP0959	jhp0893		hypothetical protein
cluster_220	HP0978	jhp0912	ftsA	cell division protein (ftsA) protein
cluster_1399	HP1013	jhp0410	dapA	dihydrodipicolinate synthase
cluster_1319	HP1044	jhp0380		hypothetical protein
cluster_756	HP1060	jhp0365		sec-independent translocase
cluster_593	HP1016	jhp0407	pgsA	phosphatidylglycerophosphate synthase (pgsA)

Table B1A (Part 6 of 7) : *H. pylori* Genes Under Positive Selection				
Cluster Identity	***H. pylori* 26695 Accessions**	***H. pylori* J99 Accessions**	**Enterez Gene Name**	**GenBank Annotation**
cluster_200	HP1020	jhp0404	ispDF	"bifunctional 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase/2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase protein"
cluster_23	HP1026	jhp0398		conserved hypothetical helicase-like protein
cluster_674	***HP1035***	***jhp0389***	***flhF***	***flagellar biosynthesis protein***
cluster_1295	HP1038	jhp0386	aroD	3-dehydroquinate dehydratase
cluster_964	HP1061	jhp0364		hypothetical protein
cluster_626	HP1079	jhp0346		hypothetical protein
cluster_856	***HP1086***	***jhp0339***	***tly / hlyA***	***hemolysin (tly)***
cluster_1065	HP1087	jhp0338	ribF (j99)	hypothetical protein
cluster_677	HP1091	jhp0334	kgtP	alpha-ketoglutarate permease (kgtP)
cluster_780	HP1113	jhp1040	omp24	outer membrane protein (omp24)
cluster_208	HP1123	jhp1052	slyD	"peptidyl-prolyl cis-trans isomerase, FKBP-type rotamase (slyD)"
cluster_311	HP1149	jhp1076		hypothetical protein
cluster_1342	HP1153	jhp1080		valyl-tRNA synthetase
cluster_332	HP1155	jhp1082	murG	N-acetylglucosaminyl transferase
cluster_1178	HP1165	jhp1092	tetA	"tetracycline resistance protein tetA(P), putative"
cluster_1200	HP1171	jhp1098	glnQ	"glutamine ABC transporter, ATP-binding protein (glnQ)"
cluster_34	HP1185	jhp1111		sugar efflux transporter
cluster_125	HP1200	jhp1123	rplJ	50S ribosomal protein L10
cluster_1021	HP1223	jhp1144		hypothetical protein
cluster_240	HP1232	jhp1153	folP	dihydropteroate synthase (folP)
cluster_865	HP1235	jhp1156		hypothetical protein
cluster_684	HP1240	jhp1161	maf	Maf-like protein
cluster_894	HP1241	jhp1162	alaS	alanyl-tRNA synthetase
cluster_292	HP1245	jhp1166	ssb	single-strand DNA-binding protein
cluster_934	HP1254	jhp1175	bioC	biotin synthesis protein (bioC)
cluster_1139	HP1255	jhp1176	secG	protein-export membrane protein
cluster_335	HP1258	jhp1179		conserved hypothetical mitochondrial protein 4
cluster_1414	HP1275	jhp1196		phosphomannomutase
cluster_403	HP1277	jhp1198	trpA	"tryptophan synthase, alpha subunit (trpA)"
cluster_815	HP1279	jhp1200	trpC	bifunctional indole-3-glycerol phosphate synthase/phosphoribosylanthranilate isomerase
cluster_1262	HP1333	jhp1253		hypothetical protein
cluster_959	HP1363	jhp1281		hypothetical protein
cluster_1368	HP1365	jhp1443		response regulator
cluster_664	HP1290	jhp1210	pnuC	nicotinamide mononucleotide transporter (pnuC)
cluster_1366	HP1309	jhp1229	rplN	50S ribosomal protein L14

Table B1A (Part 7 of 7) : *H. pylori* Genes Under Positive Selection				
Cluster Identity	***H. pylori* 26695 Accessions**	***H. pylori* J99 Accessions**	**Enterez Gene Name**	**GenBank Annotation**
cluster_541	HP1361	jhp1279	comE3 / comEC	competence locus E (comE3)
cluster_1393	HP1371	jhp1285		type III restriction enzyme R protein
cluster_17	HP1384	jhp1441		hypothetical protein
cluster_458	HP1392	jhp1435		fibronectin/fibrinogen-binding protein
cluster_481	***HP1399***	***jhp1427***	***rocF***	***arginase (rocF)***
cluster_733	HP1402	jhp1424	hsdR	type I restriction enzyme R protein (hsdR)
cluster_343	HP1407	jhp1299	rbn (j99)	hypothetical protein
cluster_1422	HP1424	jhp1319		hypothetical protein
cluster_823	HP1428	jhp1325		hypothetical protein
cluster_1033	HP1429	jhp1324	kpsF	polysialic acid capsule expression protein (kpsF)
cluster_20	HP1430	jhp1323		conserved hypothetical ATP-binding protein
cluster_926	HP1453	jhp1346		hypothetical protein
cluster_986	HP1472	jhp1365	mod	type IIS restriction enzyme M protein (mod)
cluster_800	HP1478	jhp1371	uvrD / rep	DNA helicase II (uvrD)
cluster_827	HP1484	jhp1377		hypothetical protein
cluster_882	HP1497	jhp1390	pth	peptidyl-tRNA hydrolase
cluster_329	HP1503	jhp1396	copA / fixI	"cation-transporting ATPase, P-type (copA)"
cluster_947	HP1506	jhp1399	gltS	glutamate permease (gltS)
cluster_1357	HP1508	jhp1401		ferrodoxin-like protein
cluster_179	HP1521	jhp1410	res	type III restriction enzyme R protein (res)
cluster_830	HP1530	jhp1418	punB	purine nucleoside phosphorylase (punB)
cluster_467	HP1541	jhp1458	trcF / mfd	transcription-repair coupling factor (trcF)
cluster_1091	HP1544	jhp1455	tagE	toxR-activated gene (tagE)
cluster_1299	HP1545	jhp1454	folC	folylpolyglutamate synthase (folC)
cluster_308	HP1553	jhp1446	pcrA (j99)	helicase
cluster_1360	HP1564	jhp1472		outer membrane protein

Table B1A : *H. pylori* strains 26695 (HP) and J99 (jhp) orthologous genes pairs found to be under positive selection after curation of alignments for frameshifts together with their GenBank annotations. Genes highlighted in bold italics are database identified virulence genes under positive selection.

N. meningitidis Genes Under Positive Selection

Table B1B (Part 1 of 7) : *N. meningitidis* Genes Under Positive Selection				
Cluster Identity	***N. meningitidis* MC58 Accessions**	***N. meningitidis* Z2491 Accessions**	**Enterez Gene Name**	**GenBank Annotation**
cluster_1218	NMB0004	NMA0251		EpiH/GdmH-related protein
cluster_1793	NMB0012	NMA0259		putative transmembrane transport protein
cluster_1556	***NMB0018***	***NMA0264***	***pilE***	***pilin PilE***
cluster_305	NMB0032	NMA0277		putative lipoprotein
cluster_617	NMB0199	NMA0069	lpxB	lipid-A-disaccharide synthase
cluster_754	NMB0270	NMA2216		putative bioH protein
cluster_1009	NMB0271	NMA2215		hypothetical protein
cluster_599	NMB0341	NMA2146		"Neisseria-specific antigen protein, TspA"
cluster_928	NMB0401	NMA2084	putA	proline dehydrogenase
cluster_1720	***NMB0460***	***NMA2025***	***tbp2 / tbpB***	***transferrin-binding protein 2***
cluster_414	NMB0462	NMA2023	potD-1 (mc58)	"spermidine/putrescine ABC transporter, periplasmic spermidine/putrescine-binding protein"
cluster_1461	NMB0506	NMA0692		hypothetical protein
cluster_862	NMB0598	NMA0802		Maf/YceF/YhdE family protein
cluster_1107	NMB0741	NMA0954		hypothetical protein
cluster_796	NMB0887	NMA1107		putative type IV pilus assembly protein PilV
cluster_1053	NMB0888	NMA1108		putative membrane protein
cluster_1060	NMB0990	NMA1191		hypothetical protein
cluster_1591	NMB0992	NMA1200	hsf (mc58)	adhesin
cluster_1599	NMB1041	NMA1445		GTP-binding protein
cluster_1580	NMB1194	NMA1367	cysG-2 / cysG	siroheme synthase
cluster_523	NMB1293	NMA1503		hypothetical protein
cluster_385	NMB1314	NMA1527	ftsK-2 (mc58)	cell division protein FtsK
cluster_1466	NMB1374	NMA1588	truB	tRNA pseudouridine synthase B
cluster_870	NMB1559	NMA1747	gshB	glutathione synthetase
cluster_627	NMB1564	NMA1753		hypothetical protein
cluster_1246	NMB1585	NMA1774		"transcriptional regulator, MarR family"
cluster_1250	NMB1631	NMA1795		hypothetical protein
cluster_992	NMB1687	NMA1946		hypothetical protein
cluster_1254	NMB1688	NMA1947	ansA / ans	L-asparaginase I
cluster_1527	NMB1735	NMA1991	relA	GTP pyrophosphokinase
cluster_1529	NMB1791	NMA0672	cafA	cytoplasmic axial filament protein
cluster_1467	NMB1825	NMA0634		hypothetical protein
cluster_798	NMB1699	NMA1955		hypothetical protein

Table B1B (Part 2 of 7) : *N. meningitidis* Genes Under Positive Selection				
Cluster Identity	***N. meningitidis* MC58 Accessions**	***N. meningitidis* Z2491 Accessions**	**Enterez Gene Name**	**GenBank Annotation**
cluster_1231	NMB1830	NMA0625		putative phosphoglycolate phosphatase
cluster_260	NMB1846	NMA0611		Mrp/NBP35 family protein
cluster_232	NMB1896	NMA0560	dpnC / drg	type II restriction enzyme DpnI
cluster_1418	NMB1915	NMA0540		hypothetical protein
cluster_1739	***NMB1929***	***NMA0524***	***lgtA***	***lacto-N-neotetraose biosynthesis glycosyl transferase LgtA***
cluster_1105	NMB1958	NMA0493		putative thioredoxin
cluster_1156	NMB1970	NMA0477		putative para-aminobenzoate synthetase component I/4-amino-4-deoxychorismate lyase
cluster_1744	***NMB1985***	***NMA0457***	***hap / iga2***	***IgA-specific serine endopeptidase***
cluster_1547	NMB1997	NMA0444	gloB (mc58)	hydroxyacylglutathione hydrolase
cluster_1317	NMB2051	NMA0385	petC	"ubiquinol--cytochrome c reductase, cytochrome c1"
cluster_1614	NMB2105	NMA0324	mafB	mafB protein
cluster_421	NMB2132	NMA0299		transferrin-binding protein-related protein
cluster_449	NMB0008	NMA0255		putative cell division protein FtsX
cluster_1088	NMB0029	NMA0274	hprA (mc58)	glycerate dehydrogenase
cluster_116	NMB0050	NMA0220		putative integral membrane protein
cluster_770	***NMB0071***	***NMA0198***	***ctrA***	***capsule polysaccharide export outer membrane protein CtrA***
cluster_1027	***NMB0072***	***NMA0197***	***ctrB***	***capsule polysaccharide export inner-membrane protein CtrB***
cluster_1559	***NMB0074***	***NMA0195***	***ctrD***	***capsule polysaccharide export ATP-binding protein CtrD***
cluster_37	NMB0088	NMA0178		putative outer membrane protein P1
cluster_705	NMB0105	NMA0169		PhnO-related protein
cluster_1802	NMB0115	NMA0159	ntrX (mc58)	nitrogen assimilation regulatory protein NtrX
cluster_226	NMB0116	NMA0158	dprA	DNA processing chain A
cluster_777	NMB0174	NMA0094	valS	valyl-tRNA synthetase
cluster_809	NMB0180	NMA0087	lpxD	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase
cluster_589	NMB0192	NMA0075	rnhB	ribonuclease HII
cluster_341	NMB0198	NMA0070	rluC	ribosomal large subunit pseudouridine synthase C
cluster_1472	NMB0204	NMA0065		putative lipoprotein
cluster_749	NMB0214	NMA0054	prlC	oligopeptidase A
cluster_893	NMB0246	NMA0014	nuoF	"NADH dehydrogenase I, F subunit"

Table B1B (Part 3 of 7) : *N. meningitidis* Genes Under Positive Selection				
Cluster Identity	***N. meningitidis* MC58 Accessions**	***N. meningitidis* Z2491 Accessions**	**Enterez Gene Name**	**GenBank Annotation**
cluster_946	NMB0259	NMA2228	nuoN	NADH dehydrogenase subunit N
cluster_163	NMB0262	NMA2225	xseB	"exodeoxyribonuclease, small subunit"
cluster_433	NMB0263	NMA2224		ribosome-associated GTPase
cluster_716	NMB0264	NMA2223		"ABC transporter, ATP-binding protein"
cluster_1509	NMB0267	NMA2219		putative periplasmic protein
cluster_1780	NMB0314	NMA2173		hypothetical protein
cluster_295	NMB0334	NMA2154	pgi-1 / pgi2	glucose-6-phosphate isomerase
cluster_571	NMB0335	NMA2153	dapD	"2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase"
cluster_628	NMB0348	NMA2139		hypothetical protein
cluster_897	NMB0349	NMA2138		glutamyl-tRNA synthetase
cluster_372	NMB0353	NMA2134		hypothetical protein
cluster_440	NMB0366	NMA2122		hypothetical protein
cluster_1373	NMB0394	NMA2090	nadA	quinolinate synthetase
cluster_139	NMB0405	NMA2080	comM (mc58)	competence protein ComM
cluster_1519	NMB0416	NMA2068	murF	"UDP-N-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate--D-alanyl-D-alanyl ligase"
cluster_1084	NMB0433	NMA2052	acnA	aconitate hydratase
cluster_1404	NMB0447	NMA2038	recO	DNA repair protein RecO
cluster_1429	NMB0453	NMA2032	mutT	mutT protein
cluster_112	NMB0455	NMA2030		hypothetical protein
cluster_144	***NMB0461***	***NMA2024***	***tbp1 / tbpA***	***transferrin-binding protein 1***
cluster_1524	NMB0472	NMA2013	bioF	8-amino-7-oxononanoate synthase
cluster_1323	NMB0530	NMA0708		beta-hexosaminidase
cluster_1590	NMB0531	NMA0709		putative integral membrane protein
cluster_1616	NMB0538	NMA0717		hypothetical protein
cluster_858	NMB0541	NMA0720		hypothetical protein
cluster_67	NMB0545	NMA0724		hypothetical protein
cluster_612	NMB0547	NMA0726		type IV pilin protein
cluster_1532	NMB0575	NMA0759	gcvH	glycine cleavage system protein H
cluster_1439	NMB0602	NMA0806	hitA (mc58)	hitA protein
cluster_838	NMB0638	NMA0848	galU	UTP--glucose-1-phosphate uridylyltransferase
cluster_1040	NMB0682	NMA0884	pyrC	dihydroorotase
cluster_813	NMB0688	NMA0890	trpF	N-(5'-phosphoribosyl)anthranilate isomerase
cluster_1598	NMB0690	NMA0892	purF	amidophosphoribosyltransferase
cluster_1807	NMB0622	NMA0830	lolA	outer membrane lipoprotein carrier protein
cluster_163	NMB0262	NMA2225	xseB	"exodeoxyribonuclease, small subunit"

Table B1B (Part 4 of 7) : *N. meningitidis* Genes Under Positive Selection				
Cluster Identity	***N. meningitidis* MC58 Accessions**	***N. meningitidis* Z2491 Accessions**	**Enterez Gene Name**	**GenBank Annotation**
cluster_566	NMB0693	NMA0896	folC	folylpolyglutamate synthase/dihydrofolate synthase
cluster_100	***NMB0700***	***NMA0905***	***iga***	***IgA-specific serine endopeptidase***
cluster_649	NMB0702	NMA0906	comA	competence protein ComA
cluster_787	NMB0728	NMA0937	pheT	phenylalanyl-tRNA synthetase beta subunit
cluster_1848	NMB0731	NMA0941		hypothetical protein
cluster_817	NMB0734	NMA0944		hypothetical protein
cluster_843	NMB0740	NMA0952	recN	DNA repair protein RecN
cluster_1211	NMB0767	NMA0978	pfs	5-methylthioadenosine nucleosidase/S-adenosylhomocysteine nucleosidase
cluster_1746	NMB0769	NMA0980		DNA polymerase III subunit delta
cluster_721	NMB0771	NMA0982		hypothetical protein
cluster_476	NMB0777	NMA0988		uroporphyrinogen-III synthetase
cluster_1014	NMB0779	NMA0990		hypothetical protein
cluster_1816	NMB0781	NMA0991	hemE	uroporphyrinogen decarboxylase
cluster_509	NMB0783	NMA0993		hypothetical protein
cluster_790	NMB0784	NMA0994		putative phage shock protein E precursor
cluster_1048	NMB0785	NMA0995	recB	exodeoxyribonuclease V 135 KD polypeptide
cluster_1713	NMB0809	NMA1019		hypothetical protein
cluster_690	NMB0811	NMA1021	murB	UDP-N-acetylpyruvoylglucosamine reductase
cluster_1	NMB0834	NMA1814		"transposase, IS30 family"
cluster_1344	NMB0839	NMA1048	pmbA	pmbA protein
cluster_299	NMB0841	NMA1051		putative membrane protein
cluster_1658	NMB0852	NMA1063		GTP-binding protein EngA
cluster_351	NMB0854	NMA1065	hisS	histidyl-tRNA synthetase
cluster_140	NMB0866	NMA1084		putative periplasmic protein
cluster_175	NMB0872	NMA1090		putative periplasmic protein
cluster_446	NMB0873	NMA1091	lolB	outer membrane lipoprotein LolB precursor
cluster_1823	NMB0884	NMA1104	sodB	superoxide dismutase
cluster_517	NMB0886	NMA1106	fimT	fimbrial protein FimT
cluster_117	NMB0962	NMA1159	uvrA	excinuclease ABC subunit A
cluster_1798	NMB0980	NMA1177	pntA	NAD(P) transhydrogenase subunit alpha
cluster_222	NMB0981	NMA1179	serB	phosphoserine phosphatase
cluster_1832	NMB0987	NMA1303		putative N-acetylmuramoyl-L-alanine amidase
cluster_1617	NMB0999	NMA1207		NifR3/SMM1 family protein
cluster_196	NMB1024	NMA1464		hypothetical protein
cluster_1348	NMB0895	NMA1114		hypothetical protein
cluster_1320	NMB0935	NMA1130	miaA	tRNA delta(2)-isopentenylpyrophosphate transferase

Table B1B (Part 5 of 7) : *N. meningitidis* Genes Under Positive Selection				
Cluster Identity	***N. meningitidis* MC58 Accessions**	***N. meningitidis* Z2491 Accessions**	**Enterez Gene Name**	**GenBank Annotation**
cluster_234	NMB1030	NMA1457		hypothetical protein
cluster_814	NMB1039	NMA1447		hypothetical protein
cluster_1104	NMB1046	NMA1440	thrC	threonine synthase
cluster_75	NMB1055	NMA1254	glyA	serine hydroxymethyltransferase
cluster_368	NMB1062	NMA1261		putative integral membrane protein
cluster_1706	NMB1067	NMA1266	ftsK-1 (mc58)	cell division protein FtsK
cluster_1545	NMB1085	NMA1864		putative N-acetylmuramoyl-L-alanine amidase
cluster_541	NMB1094	NMA1313		hypothetical protein
cluster_404	NMB1115	NMA1325		putative tail fibre protein
cluster_209	NMB1183	NMA1356	mpl-2 / mpl	UDP-N-acetylmuramate:L-alanyl-gamma-D-glutamyl-meso-diaminopimelate ligase
cluster_1821	NMB1189	NMA1362	cysI-2 / cysI	"sulfite reductase hemoprotein, beta-component"
cluster_514	NMB1190	NMA1363	cysJ-2 / cysJ	"sulfite reductase (NADPH) flavoprotein, alpha component"
cluster_1082	NMB1199	NMA1370	typA	GTP-binding protein TypA
cluster_1582	NMB1240	NMA1409		"ABC transporter, ATP-binding protein"
cluster_303	NMB1249	NMA1418		putative nitrate/nitrite sensory protein NarX
cluster_1637	NMB1253	NMA1424		hypothetical protein
cluster_1526	NMB1284	NMA1494		hypothetical protein
cluster_1164	NMB1310	NMA1524	gcpE	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase
cluster_1799	NMB1331	NMA1545	uvrB	excinuclease ABC subunit B
cluster_1293	NMB1336	NMA1550		hypothetical protein
cluster_859	NMB1353	NMA1565		aldehyde dehydrogenase family protein
cluster_1437	NMB1368	NMA1580		putative ATP-dependent RNA helicase
cluster_672	NMB1371	NMA1584	argD	bifunctional N-succinyldiaminopimelate-aminotransferase/acetylornithine transaminase protein
cluster_1125	NMB1400	NMA1620		ABC transporter family protein
cluster_1733	NMB1421	NMA1633		nifR3 protein
cluster_193	***NMB1429***	***NMA1642***	***porA***	***outer membrane protein PorA***
cluster_1269	NMB1432	NMA1644	aroA	3-phosphoshikimate 1-carboxyvinyltransferase
cluster_1841	NMB1440	NMA1652		hypothetical protein
cluster_20	NMB1447	NMA1660	rep	ATP-dependent DNA helicase
cluster_365	NMB1467	NMA1679	ppx (gppA-z2491)	exopolyphosphatase

Table B1B (Part 6 of 7) : *N. meningitidis* Genes Under Positive Selection				
Cluster Identity	***N. meningitidis* MC58 Accessions**	***N. meningitidis* Z2491 Accessions**	**Enterez Gene Name**	**GenBank Annotation**
cluster_1845	NMB1497	NMA1700		TonB-dependent receptor
cluster_76	NMB1506	NMA1707	argS	arginyl-tRNA synthetase
cluster_1546	NMB1536	NMA1735	secA	translocase
cluster_1046	***NMB1540***	***NMA1739***	***lbpA***	***lactoferrin-binding protein A***
cluster_1339	NMB1548	NMA1797		putative tspB protein
cluster_1629	NMB1555	NMA1743	fadD-2 / fadD	long-chain-fatty-acid--CoA ligase
cluster_1157	NMB1566	NMA1755	purN	phosphoribosylglycinamide formyltransferase
cluster_1451	NMB1573	NMA1762	argF	"ornithine carbamoyltransferase, catabolic"
cluster_511	NMB1595	NMA1788	alaS	alanyl-tRNA synthetase
cluster_1314	NMB1644	NMA1898		hypothetical protein
cluster_61	***NMB1705***	***NMA1959***	***rfaK***	***"alpha-1,2-N-acetylglucosamine transferase"***
cluster_1162	NMB1715	NMA1969	mtrD	multiple transferable resistance system protein MtrD
cluster_114	NMB1718	NMA1972		hypothetical protein
cluster_932	NMB1720	NMA1974	recC	exodeoxyribonuclease V 125 kD polypeptide
cluster_1192	NMB1721	NMA1975		hypothetical protein
cluster_1227	NMB1784	NMA0681		hypothetical protein
cluster_1294	NMB1797	NMA0665		penicillin-binding protein 3
cluster_95	NMB1814	NMA0647	aroB	3-dehydroquinate synthase
cluster_361	NMB1815	NMA0646		hypothetical protein
cluster_913	NMB1817	NMA0644	ribD	riboflavin biosynthesis protein RibD
cluster_120	NMB1820	NMA0639	pglB	pilin glycosylation protein PglB
cluster_673	NMB1822	NMA0637	pglD	pilin glycosylation protein PglD
cluster_156	NMB1827	NMA0632	dnaE	"DNA polymerase III, alpha subunit"
cluster_190	NMB1833	NMA0622	ileS	isoleucyl-tRNA synthetase
cluster_744	NMB1835	NMA0620	tyrS	tyrosyl-tRNA synthetase
cluster_497	NMB1840	NMA0616		hypothetical protein
cluster_1098	NMB1855	NMA0602	carB	"carbamoyl-phosphate synthase, large subunit"
cluster_1149	NMB1868	NMA0588	xerC	integrase/recombinase XerC
cluster_711	NMB1885	NMA0572	pcm	protein-L-isoaspartate O-methyltransferase
cluster_501	NMB1897	NMA0559	leuS	leucyl-tRNA synthetase
cluster_865	NMB1907	NMA0548	yidC	putative inner membrane protein translocase component YidC
cluster_1129	NMB1908	NMA0547		hypothetical protein
cluster_73	NMB1910	NMA0545		hypothetical protein
cluster_363	NMB1871	NMA0585		hypothetical protein
cluster_916	NMB1873	NMA0583		"putative DNA polymerase, bacteriophage-type"

Table B1B (Part 7 of 7) : *N. meningitidis* Genes Under Positive Selection				
Cluster Identity	***N. meningitidis* MC58 Accessions**	***N. meningitidis* Z2491 Accessions**	**Enterez Gene Name**	**GenBank Annotation**
cluster_366	NMB1918	NMA0536		acyl-carrier-protein S-malonyltransferase
cluster_647	NMB1919	NMA0535		"ABC transporter, ATP-binding protein"
cluster_1475	***NMB1928***	***NMA0525***	***lgtB***	***lacto-N-neotetraose biosynthesis glycosyl transferase LgtB***
cluster_1846	NMB1948	NMA0504		"ABC transporter, ATP-binding protein"
cluster_624	NMB1969	NMA0478		putative serotype-1-specific antigen
cluster_103	NMB1973	NMA0472	groES	co-chaperonin GroES
cluster_1449	NMB1978	NMA0466	cyaY	frataxin-like protein
cluster_949	NMB1982	NMA0462	polA	DNA polymerase I
cluster_436	NMB1987	NMA0454	thdF	tRNA modification GTPase
cluster_474	NMB1993	NMA0448		"iron(III) ABC transporter, ATP-binding protein"
cluster_1274	NMB1996	NMA0445	purL	phosphoribosylformylglycinamidine synthase
cluster_662	NMB2024	NMA0416		hypothetical protein
cluster_1756	NMB2034	NMA0404		putative 1-acyl-sn-glycerol-3-phosphate acyltransferase
cluster_448	NMB2036	NMA0402	truA	tRNA pseudouridine synthase A
cluster_1255	***NMB2039***	***NMA0398***	***porB***	***major outer membrane protein PIB***
cluster_214	NMB2041	NMA0394		thiamin pyrophosphokinase-related protein
cluster_853	NMB2062	NMA0373	thiF	thiF protein
cluster_636	NMB2074	NMA0358		hypothetical protein
cluster_906	NMB2075	NMA0357		BirA protein/Bvg accessory factor
cluster_1693	NMB2078	NMA0353		hypothetical protein
cluster_490	NMB2099	NMA0330		hypothetical protein

Table B1B : Genes under positive selection between *N. meningitidis* strains MC58 (NMB) and Z2491 (NMA) after curation of alignments for frameshifts. Genes highlighted in bold italics are database identified virulence genes under positive selection.

V. cholerae **Genes Under Positive Selection**

Table B1C (Part 1 of 1) : *V. cholerae* Genes Under Positive Selection				
Cluster Identity	***V. cholerae* N16961 Accessions**	***V. cholerae* O395 Accessions**	**Enterez Gene Name**	**GenBank Annotation**
cluster_1096	VC0212	VC0395_A2591	msbB (O395)	lipid A biosynthesis lauroyl acyltransferase
cluster_606	VC0370	VC0395_A2781		hypothetical protein
cluster_492	VC0460	VC0395_A0012	proC (O395)	pyrroline-5-carboxylate reductase
cluster_2309	VC0694	VC0395_A0225		hypothetical protein
cluster_2274	***VC0837***	***VC0395_A0362***	***tcpF (O395)***	***toxin co-regulated pilus biosynthesis protein F***
cluster_1375	VC1460	VC0395_A1063		hypothetical protein
cluster_1683	VC1492	VC0395_A1099		hypothetical protein
cluster_3151	VC1688	VC0395_A1293		hypothetical protein
cluster_1570	VC1987	VC0395_A1572		"outer membrane lipoprotein Slp, putative"
cluster_1299	VC2107	VC0395_A1691		aspartate-semialdehyde dehydrogenase
cluster_826	VC2414	VC0395_A1990	aceE	"pyruvate dehydrogenase, E1 component"
cluster_155	VC2450	VC0395_A2028	mazG (O395)	mazG protein
cluster_217	VC2503	VC0395_A2085	valS	valyl-tRNA synthetase
cluster_2810	VC2610	VC0395_A2188		hypothetical protein
cluster_2306	VC2712	VC0395_A2284		xanthine/uracil permease family protein
cluster_3244	VC2758	VC0395_A2534	fadB	"fatty oxidation complex, alpha subunit"
cluster_1734	VC2761	VC0395_A2531		multidrug resistance protein
cluster_580	VCA0082	VC0395_0058		"transcriptional regulator, LysR family"
cluster_3098	VCA0382	VC0395_0855		hypothetical protein
cluster_1192	VCA0849	VC0395_0388		hypothetical protein
cluster_582	VCA0994	VC0395_0243		hypothetical protein
cluster_771	VCA1062	VC0395_0180	potE	putrescine-ornithine antiporter
cluster_3419	VCA1073	VC0395_0169	putA	"proline dehydrogenase/delta-1-pyrroline-5-carboxylate dehydrogenase, authentic point mutation"

Table B1C : Genes under positive selection between *V. cholerae* strains N16961 (VC and VCA) and O395 (VC_0395) after curation of alignments for frameshifts. Database identified virulence genes under positive selection are highlighted in bold italics.

Appendix B2

Database Identified Virulence Genes

H. pylori Database Identified Virulence Genes

Table B2A (Part 1 of 4) *H. pylori* Database Identified Virulence Genes		
***H. pylori* Gene Accession**	**Gene Name**	**GenBank Annotation**
HP0009	omp1	outer membrane protein
HP1129	-	biopolymer transport protein (exbD)
HP0824	-	thioredoxin (trxA)
HP0071	-	urease accessory protein (ureI)
HP0522	-	cag pathogenicity island protein (cag3)
HP0220	-	cysteine desulfurase
HP0887	**-**	***vacuolating cytotoxin***
HP1238	-	aliphatic amidase (aimE)
HP0529	-	cag pathogenicity island protein (cag9)
HP1340	-	biopolymer transport protein (exbD)
HP0535	**-**	***cag pathogenicity island protein (cag14)***
HP0638	**-**	***outer membrane protein (omp13)***
HP0797	-	flagellar sheath adhesin hpaA
HP0439	-	hypothetical protein
HP0541	-	cag pathogenicity island protein (cag20)
HP0246	-	flagellar P-ring protein precursor
HP0753	-	flagellar protein FliS
HP1419	-	flagellar biosynthesis protein
HP1475	-	phosphopantetheine adenylyltransferase
HP0915	-	iron-regulated outer membrane protein (frpB)
HP1077	-	nickel transport protein (nixA)
HP1027	-	ferric uptake regulation protein (fur)
HP0825	-	thioredoxin reductase (trxB)
HP0072	-	urease beta subunit (urea amidohydrolase) (ureB)
HP0523	-	cag pathogenicity island protein (cag4)
HP0221	-	nifU-like protein
HP0380	-	glutamate dehydrogenase
HP0888	-	"iron(III) dicitrate ABC transporter, ATP-binding protein (fecE)"
HP0039m	-	Component of conjugal plasmid transfer system
HP0393	-	chemotaxis protein (cheV)
HP0542	-	cag pathogenicity island protein (cag21)2
HP1400	-	iron(III) dicitrate transport protein (fecA)
HP0651	**-**	***fucosyltransferase***
HP1341	-	siderophore-mediated iron transport protein (tonB)
HP0536	-	cag pathogenicity island protein (cag15)

Table B2A (Part 2 of 4) *H. pylori* Database Identified Virulence Genes		
***H. pylori* Gene Accession**	**Gene Name**	**GenBank Annotation**
HP0601	-	flagellin
HP0807	-	iron(III) dicitrate transport protein (fecA)
HP0459	-	virB4 homolog (virB4)
HP1174	-	glucose/galactose transporter (gluP)
HP1585	-	flagellar basal-body rod protein (flgG)
HP0876	**-**	***iron-regulated outer membrane protein (frpB)***
HP0067	-	urease accessory protein (ureH)
HP1130	-	biopolymer transport protein (exbB)
HP0017	-	virB4 homolog (virB4)
HP0073	-	urease alpha subunit (ureA) (urea amidohydrolase)
HP0524	-	cag pathogenicity island protein (cag5)
HP0279	-	lipopolysaccharide heptosyltransferase-1 (rfaC)
HP0325	-	flagellar L-ring protein precursor
HP0786	**-**	***translocase***
HP0889	-	"iron(III) dicitrate ABC transporter, permease protein (fecD)"
HP0530	-	cag pathogenicity island protein (cag10)
HP1399	**-**	***arginase (rocF)***
HP1445	-	biopolymer transport protein (exbB)
HP0537	-	cag pathogenicity island protein (cag16)
HP0235	-	conserved hypothetical secreted protein
HP0543	-	cag pathogenicity island protein (cag22)
HP0042	-	trbI protein
HP1420	-	flagellum-specific ATP synthase
HP1119	-	flagellar hook-associated protein
HP0870	-	flagellar hook protein
HP0512	-	glutamine synthetase (glnA)
HP0068	-	urease accessory protein (ureG)
HP0525	-	virB11 homolog
HP0684	fliP	flagellar biosynthesis protein
HP0326	-	CMP-N-acetylneuraminic acid synthetase (neuA)
HP0531	-	cag pathogenicity island protein (cag11)
HP1446	-	biopolymer transport protein (exbD)
HP1041	-	flagellar biosynthesis protein
HP0538	-	cag pathogenicity island protein (cag17)
HP0441	-	VirB4 homolog
HP0037	**-**	***NADH ubiquinone oxidoreductase subunit***
HP0544	-	cag pathogenicity island protein (cag23)
HP1561	-	"iron(III) ABC transporter, periplasmic iron-binding protein (ceuE)"
HP0043	-	mannose-6-phosphate isomerase (pmi) or (algA)
HP1035	**-**	***flagellar biosynthesis protein***

Table B2A (Part 3 of 4) *H. pylori* Database Identified Virulence Genes		
***H. pylori* Gene Accession**	**Gene Name**	**GenBank Annotation**
HP0653	-	nonheme iron-containing ferritin (pfr)
HP0351	-	flagellar M-ring protein
HP1067	-	chemotaxis protein (cheY)
HP0159	**-**	***"lipopolysaccharide 1,2-glucosyltransferase (rfaJ)"***
HP1421	-	conjugative transfer regulon protein (trbB)
HP0815	-	flagellar motor protein
HP0616	**-**	***chemotaxis protein (cheV)***
HP0069	**-**	***urease accessory protein (ureF)***
HP0115	-	Flagellin
HP0679	**-**	***lipopolysaccharide biosynthesis protein (wbpB)***
HP1490	-	hypothetical protein
HP0725	omp17	outer membrane protein
HP1086	**-**	***hemolysin (tly)***
HP0019	-	chemotaxis protein (cheV)
HP0526	-	cag pathogenicity island protein (cag6)
HP0178	-	spore coat polysaccharide biosynthesis protein E
HP0685	-	flagellar biosynthetic protein (fliP)
HP1092	-	flagellar basal-body rod protein (flgG)
HP0532	-	cag pathogenicity island protein (cag12)
HP0539	-	cag pathogenicity island protein (cag18)
HP0038	**-**	***DNA transformation competancy*** (J99 annotation)
HP0545	-	cag pathogenicity island protein (cag24)
HP0243	-	neutrophil activating protein (napA) (bacterioferritin)
HP1562	-	"iron(III) ABC transporter, periplasmic iron-binding protein (ceuE)"
HP0044	-	GDP-D-mannose dehydratase (rfbD)
HP1512	-	iron-regulated outer membrane protein (frpB)
HP0352	-	flagellar motor protein
HP0912	**-**	***outer membrane protein (omp20)***
HP0816	-	flagellar motor protein
HP0520	-	cag pathogenicity island protein (cag1)
HP1030	-	flagellar motor switch protein
HP0527	-	cag pathogenicity island protein (cag7)
HP1339	-	biopolymer transport protein (exbB)
HP0686	-	iron(III) dicitrate transport protein (fecA)
HP0390	-	adhesin-thiol peroxidase (tagD))
HP0294	-	aliphatic amidase (aimE)
HP1557	-	flagellar basal body protein
HP0546	-	cag pathogenicity island protein (cag25)

Table B2A (Part 4 of 4) *H. pylori* Database Identified Virulence Genes		
***H. pylori* Gene Accession**	**Gene Name**	**GenBank Annotation**
HP1563	-	alkyl hydroperoxide reductase (tsaA)
HP0045	-	nodulation protein (nolK)
HP0353	-	flagellar assembly protein
HP0907	-	flagellar biosynthesis protein
HP0867	lpxB	lipid-A-disaccharide synthase
HP0913	**-**	***outer membrane protein (omp21)***
HP0160	-	conserved hypothetical secreted protein
HP0770	-	flagellar biosynthesis protein
HP0070	-	urease accessory protein (ureE)
HP0173	-	flagellar biosynthesis protein
HP1031	-	flagellar motor switch protein
HP0379	-	fucosyltransferase
HP0528	-	cag pathogenicity island protein (cag8)
HP0584	-	flagellar motor switch protein
HP0687	-	iron(II) transport protein (feoB)
HP1243	**-**	***outer membrane protein (omp28)***
HP0534	-	cag pathogenicity island protein (cag13)
HP0289	-	toxin-like outer membrane protein
HP0391	-	purine-binding chemotaxis protein (cheW)
HP0540	-	cag pathogenicity island protein (cag19)
HP0136	-	bacterioferritin comigratory protein (bcp)
HP1558	-	flagellar basal body rod protein
HP0547	**-**	***cag pathogenicity island protein (cag26)***
HP0752	-	flagellar hook-associated protein
HP0360	-	UDP-glucose 4-epimerase
HP0896	**-**	***outer membrane protein (omp19)***
HP0127	-	outer membrane protein (omp4)
HP0722	omp16	outer membrane protein
HP0093	-	hypothetical protein
HP0751	-	flagellar protein FlaG
HP0295	-	flagellar hook-associated protein
HP0327	**-**	***flagellar protein G (flaG)***
HP1032	-	flagellar biosynthesis sigma factor FliA
HP0809	**-**	***flagellar basal body-associated protein***
HP0908	-	flagellar hook protein
HP1274	-	paralysed flagella protein (pflA)
HP1477	-	flagellar basal body P-ring biosynthesis protein
HP1559	-	flagellar basal body rod protein
HP1575	-	"ABC transporter, putative"

Table B2A : *H. pylori* database identified virulence genes and their GenBank annotations. *H. pylori* database identified virulence genes under positive selection are highlighted in bold italics.

N. meningitidis Database Identified Virulence Genes

Table B2B (Part 1 of 2) *N. meningitidis* Database Identified Virulence Genes		
N. meningitidis **Gene Accession**	**Gene Name**	**GenBank Annotation**
NMA0049	-	RNA polymerase sigma factor
NMA0185	-	capsule polysaccharide modification protein
NMA0186	-	capsule polysaccharide modification protein
NMA0195	**-**	***capsule polysaccharide export ATP-binding protein***
NMA0196	-	capsule polysaccharide export inner-membrane protein
NMA0197	**-**	***capsule polysaccharide export inner-membrane protein***
NMA0198	**-**	***capsule polysaccharide export outer membrane protein***
NMA0199	-	putative UDP-N-acetyl-D-glucosamine 2-epimerase
NMA0200	-	putative capsule biosynthesis protein
NMA0201	-	putative capsule biosynthesis protein
NMA0202	-	putative capsule biosynthesis protein
NMA0218	-	putative pilus retraction protein
NMA0219	-	pilT-like protein
NMA0243	-	lipopolysaccharide heptosyltransferase I
NMA0264	**-**	***fimbrial protein precursor (pilin)***
NMA0265	-	pilS1
NMA0266	-	pilS2
NMA0267	-	pilS3
NMA0268	-	pilS4
NMA0269	-	pilS5
NMA0270	-	pilS6
NMA0271	-	pilS7
NMA0272	-	truncated pilin
NMA0293	-	pseudogene (pilus-associated protein)
NMA0398	**-**	***"porin, major outer membrane protein P.I"***
NMA0453	-	putative iron-regulated outer membrane protein
NMA0457	**-**	***IgA-specific serine endopeptidase***
NMA0474	-	haemoglobin-haptoglobin-utilization protein
NMA0475	-	haemoglobin-haptoglobin-utilization protein
NMA0524	**-**	***lacto-N-neotetraose biosynthesis glycosyl tranferase***
NMA0525	**-**	***lacto-N-neotetraose biosynthesis glycosyl tranferase***
NMA0527	-	lacto-N-neotetraose biosynthesis glycosyl tranferase
NMA0609	-	pilus-associated protein
NMA0650	-	pilus secretin
NMA0652	-	pilus assembly protein
NMA0653	-	putative pilus assembly protein
NMA0654	-	putative pilus assembly protein
NMA0842	-	putative iron-uptake permease ATP-binding protein
NMA0651	-	pilus assembly protein

Table B2B (Part 2 of 2) *N. meningitidis* Database Identified Virulence Genes		
***N. meningitidis* Gene Accession**	**Gene Name**	***N. meningitidis* Gene Accession**
NMA0843	-	putative iron-uptake permease inner membrane protein
NMA0844	-	major ferric iron binding protein
NMA0905	**-**	***IgA1 protease***
NMA0979	-	putative pilus retraction protein
NMA1110	-	putative pilin
NMA1118	-	"CMP-N-acetylneuraminate-beta-galactosamide-alpha- 2,3-sialyltransferase"
NMA1251	-	outer membrane protein precursor
NMA1523	-	putative lipoprotein
NMA1626	-	putative RTX-family exoprotein
NMA1642	**-**	***"porin, class I outer membrane protein"***
NMA1676	-	opacity protein
NMA1727	-	ADP-heptose:LPS heptosyltransferase II
NMA1739	**-**	***lactoferrin binding protein A***
NMA1740	-	lactoferrin-binding protein
NMA1890	-	opacity protein
NMA1925	-	haemoglobin recepto
NMA1958	-	"beta-1,4-glucosyltransferase"
NMA1959	**-**	***"alpha 1,2 N-acetylglucosamine transferase"***
NMA1983	-	biopolymer transport protein
NMA1985	-	TonB protein
NMA1995	-	periplasmic type I secretion system protein
NMA2024	**-**	***transferrin-binding protein A***
NMA2025	**-**	***transferrin-binding protein B***
NMA2043	-	opacity protein
NMA2113	-	adhesin MafB2
NMA2155	-	pilus-assembly protein
NMA2156	-	type IV prepilin leader peptidase
NMA2159	-	type IV pilus assembly protein

Table B2B : *N. meningitidis* database identified virulence genes and their GenBank annotations. *N. meningitidis* database identified virulence genes under positive selection are highlighted in bold italics.

V. cholerae Database Identified Virulence Genes

Table B2C (Part 1 of 4) *V. cholerae* Database Identified Virulence Genes		
***V. cholerae* Gene Accession**	**Gene Name**	**GenBank Annotation**
VC0398	-	regulatory protein CsrD
VC0399	-	MSHA biogenesis protein MshI
VC0400	-	MSHA biogenesis protein MshJ
VC0401	-	MSHA biogenesis protein MshK
VC0402	-	MSHA biogenesis protein MshL
VC0403	-	MSHA biogenesis protein MshM
VC0404	-	MSHA biogenesis protein MshN
VC0405	-	MSHA biogenesis protein MshE
VC0406	-	MSHA biogenesis protein MshG
VC0407	-	MSHA biogenesis protein MshF
VC0408	-	MSHA pilin protein MshB
VC0409	-	MSHA pilin protein MshA
VC0410	-	MSHA pilin protein MshC
VC0411	-	MSHA pilin protein MshD
VC2423	-	fimbrial protein
VC2424	-	type IV pilus assembly protein PilB
VC2425	-	type IV pilin biogenesis protein PilC
VC2426	-	leader peptidase PilD
VC0918	wecC	UDP-N-acetyl-D-mannosamine dehydrogenase
VC0917	-	UDP-N-acetylglucosamine 2-epimerase
VC0263	-	"galactosyl-transferase, putative"
VC0260	-	mannosyl-transferase
VC0934	-	"capsular polysaccharide biosynthesis glycosyltransferase, putative"
VC0935	-	hypothetical protein
VC0936	-	polysaccharide export-related protein
VC0937	-	"exopolysaccharide biosynthesis protein, putative"
VC0927	-	UDP-N-acetyl-D-mannosamine transferase
VC2187	-	flagellin
VC2190	flgL	flagellar hook-associated protein FlgL
VC2191	flgK	flagellar hook-associated protein FlgK
VC2192	flgJ	flagellar rod assembly protein/muramidase FlgJ
VC2194	flgH	flagellar basal body L-ring protein
VC2195	flgG	flagellar basal body rod protein FlgG
VC2196	flgF	flagellar basal body rod protein FlgF
VC2197	flgE	flagellar hook protein FlgE
VC2198	flgD	flagellar basal body rod modification protein
VC2199	flgC	flagellar basal body rod protein FlgC
VC2200	flgB	flagellar basal body rod protein FlgB
VC2201	-	chemotaxis protein methyltransferase CheR
VC2202	-	chemotaxis protein CheV

Table B2C (Part 2 of 4) *V. cholerae* Database Identified Virulence Genes		
***V. cholerae* Gene Accession**	**Gene Name**	**GenBank Annotation**
VC2203	flgA	flagellar basal body P-ring biosynthesis protein FlgA
VC2204	-	"negative regulator of flagellin synthesis FlgM, putative"
VC2205	-	hypothetical protein
VC2120	flhB	flagellar biosynthesis protein FlhB
VC2121	fliR	flagellar biosynthesis protein FliR
VC2122	fliQ	flagellar biosynthesis protein FliQ
VC2123	fliP	flagellar biosynthesis protein FliP
VC2124	-	flagellar protein FliO
VC2125	fliN	flagellar motor switch protein
VC2126	fliM	flagellar motor switch protein FliM
VC2127	fliL	flagellar basal body-associated protein FliL
VC2128	-	"flagellar hook-length control protein FliK, putative"
VC2129	fliJ	flagellar biosynthesis chaperone
VC2130	fliI	flagellum-specific ATP synthase
VC2131	fliH	flagellar assembly protein H
VC2132	fliG	flagellar motor switch protein G
VC2133	fliF	flagellar MS-ring protein
VC2134	fliE	flagellar hook-basal body protein FliE
VC2135	-	sigma-54 dependent response regulator
VC2136	-	sensory box sensor histidine kinase
VC2137	-	sigma-54 dependent transcriptional activator
VC2138	fliS	flagellar protein FliS
VC2139	-	"flagellar rod protein FlaI, putative"
VC2140	fliD	flagellar capping protein
VC2141	-	flagellar protein FlaG
VC2142	-	flagellin
VC2143	-	flagellin
VC2144	-	flagellin
VC2059	-	purine-binding chemotaxis protein CheW
VC2062	-	chemotaxis-specific methylesterase
VC2063	-	chemotaxis protein CheA
VC2064	-	chemotaxis protein CheZ
VC2065	-	chemotaxis protein CheY
VC2066	fliA	flagellar biosynthesis sigma factor
VC2067	-	MinD-related protein
VC2068	flhF	flagellar biosynthesis regulator FlhF
VC2069	flhA	flagellar biosynthesis protein FlhA
VC0892	-	flagellar motor protein PomA
VC0893	motB	flagellar motor protein MotB
VC1008	-	sodium-type flagellar protein MotY
VC2601	-	sodium-type flagellar protein MotX
VCA0112	-	hypothetical protein
VCA0865	-	hemagglutinin/protease

Table B2C (Part 3 of 4) *V. cholerae* Database Identified Virulence Genes		
***V. cholerae* Gene Accession**	**Gene Name**	**GenBank Annotation**
VC1784	-	neuraminidase
VC0475	-	enterobactin receptor protein
VCA0576	-	heme transport protein HutA
VCA0064	-	"TonB system receptor, putative"
VCA0625	-	TonB receptor-related protein
VC0776	-	iron-enterobactin transporter periplasmic binding protein
VC0777	-	"ferric vibriobactin ABC transporter, permease protein"
VC0778	-	"ferric vibriobactin ABC transporter, permease protein"
VC0779	-	"ferric vibriobactin ABC transporter, ATP-binding protein"
VCA0227	-	"iron(III) ABC transporter, periplasmic iron-compound-binding protein"
VCA0228	-	"iron(III) ABC transporter, permease protein"
VCA0229	-	"iron(III) ABC transporter, permease protein"
VCA0230	-	"iron(III) ABC transporter, ATP-binding protein"
VC0771	-	vibriobactin-specific isochorismatase
VC0772	-	"vibriobactin-specific 2,3-dihydroxybenzoate-AMP ligase"
VC0773	-	vibriobactin-specific isochorismate synthase
VC0774	-	"2,3-dihydroxybenzoate-2,3-dehydrogenase"
VC0775	-	"vibriobactin synthesis protein, putative"
VC0780	-	vibriobactin synthase component D
VC2209	vibF	nonribosomal peptide synthetase VibF
VC2210	-	vibriobactin utilization protein ViuB
VC2211	-	ferric vibriobactin receptor
VC0557	-	S-ribosylhomocysteinase
VCA0523	-	hypothetical protein
VC2723	-	general secretion pathway protein N
VC2724	-	cholera toxin secretion protein EpsM
VC2725	-	general secretion pathway protein L
VC2726	-	general secretion pathway protein K
VC2727	-	general secretion pathway protein J
VC2728	-	general secretion pathway protein I
VC2729	-	general secretion pathway protein H
VC2730	-	general secretion pathway protein G
VC2731	-	general secretion pathway protein F
VC2732	-	general secretion pathway protein E
VC2734	-	general secretion pathway protein C
VC1415	-	hcp protein
VCA0017	-	hcp protein
VC1416	-	vgrG protein
VCA0018	-	vgrG protein
VCA0123	-	vgrG protein
VCA0110	-	hypothetical protein
VCA0111	-	hypothetical protein

Table B2C (Part 4 of 4) *V. cholerae* Database Identified Virulence Genes		
***V. cholerae* Gene Accession**	**Gene Name**	**GenBank Annotation**
VCA0113	-	hypothetical protein
VCA0114	-	hypothetical protein
VCA0115	-	hypothetical protein
VCA0116	-	clpB protein
VCA0118	-	hypothetical protein
VCA0119	-	hypothetical protein
VCA0120	-	IcmF-related protein
VCA0218	-	thermolabile hemolysin
VCA0219	-	haemolysin
VC0827	-	toxin co-regulated pilus biosynthesis protein H
VC0833	-	toxin co-regulated pilus biosynthesis protein D
VC0640	secG	preprotein translocase subunit SecG
VC1459	-	accessory cholera enterotoxin
VC2193	flgI	flagellar basal body P-ring protein
VC0828	-	toxin co-regulated pilin
VC0834	-	toxin co-regulated pilus biosynthesis protein S
VC1447	-	RTX toxin transporter
VC0840	-	accessory colonization factor AcfB
VC0744	secF	preprotein translocase subunit SecF
VCA0117	-	sigma-54 dependent transcriptional regulator
VC2188	-	flagellin
VC0829	-	toxin co-regulated pilus biosynthesis protein B
VC0835	-	toxin co-regulated pilus biosynthesis protein T
VC1448	-	RTX toxin transporter
VC0841	-	accessory colonization factor AcfC
VC1473	-	hypothetical protein
VC0836	-	toxin co-regulated pilus biosynthesis protein E
VC0830	-	toxin co-regulated pilus biosynthesis protein Q
VC0837	-	***toxin co-regulated pilus biosynthesis protein F***
VC1456	-	"cholera enterotoxin, B subunit"
VC0825	-	toxin co-regulated pilus biosynthesis protein I
VC0831	-	toxin co-regulated pilus biosynthesis outer membrane protein C
VC0838	-	TCP pilus virulence regulatory protein
VC1450	-	RTX toxin activating protein
VC0844	-	accessory colonization factor AcfA
VC1457	-	"cholera enterotoxin, A subunit"
VC0826	-	toxin co-regulated pilus biosynthesis protein P
VC0832	-	toxin co-regulated pilus biosynthesis protein R
VC0839	-	leader peptidase TcpJ
VC1451	-	RTX toxin RtxA
VC0845	-	hypothetical protein
VC1458	-	zona occludens toxin

Table B2C : *V. cholerae* database identified virulence genes and their GenBank annotations. *V. cholerae* database identified virulence genes under positive selection are highlighted in bold italics.

Appendix B3

Statistical Testing for Association between Genes Under Positive Selection and Database Identified Virulence Genes

H pylori

```
> two_way_classification_chi_function (230,153,22,1535)
    row1 row2
[1,]   22  131
[2,]  208 1174

        Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_table
X-squared = 0.0103, df = 1, p-value = 0.9192


        Fisher's Exact Test for Count Data

data:  contingency_table
p-value = 0.9052
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.560765 1.539217
sample estimates:
odds ratio
 0.9478948
```

N meningitidis

```
> two_way_classification_chi_function(218, 68, 14, 2208)
    [,1] [,2]
row1   14  204
row2   54 1936

        Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_table
X-squared = 7.8528, df = 1, p-value = 0.005074


        Fisher's Exact Test for Count Data

data:  contingency_table
p-value = 0.006048
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.238984 4.585353
sample estimates:
```

odds ratio
2.459031

V cholerae

> two_way_classification_chi_function (23,169,1,3998)
row1 row2
[1,] 1 168
[2,] 22 3807

Pearson's Chi-squared test with Yates' continuity correction

data: contingency_table
X-squared = 0.2409, df = 1, p-value = 0.6236

Fisher's Exact Test for Count Data

data: contingency_table
p-value = 1
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.02481221 6.45051469
sample estimates:
odds ratio
1.030033

Warning message:
In chisq.test(contingency_table) :
Chi-squared approximation may be incorrect

Statistical Testing For Virulence Gene Database Bias

***H pylori* MVirDB**

two_way_classification_chi_function (230,138,19,1535)
row1 row2
[1,] 19 119
[2,] 211 1186

Pearson's Chi-squared test with Yates' continuity correction

data: contingency_table
X-squared = 0.0867, df = 1, **p-value = 0.7685**

***H pylori* VFDB**

two_way_classification_chi_function (230,87,13,1535)
row1 row2
[1,] 13 74
[2,] 217 1231

Pearson's Chi-squared test with Yates' continuity correction

data: contingency_table
X-squared = 0.0206, df = 1, **p-value = 0.8858**

***N meningitidis* MVirDB**

two_way_classification_chi_function (218,37,12,2208)
    row1 row2
[1,]   12   25
[2,] 206 1965

Pearson's Chi-squared test with Yates' continuity correction

data: contingency_table
X-squared = 19.0206, df = 1, **p-value = 1.293e-05**

***N meningitidis* VFDB**

two_way_classification_chi_function (218,68,14,2208)
    row1 row2
[1,]   14   54
[2,] 204 1936

Pearson's Chi-squared test with Yates' continuity correction

data: contingency_table
X-squared = 7.8528, df = 1, **p-value = 0.005074**

***V cholerae* MVirDB**

two_way_classification_chi_function (23,33,1,3998)
    [,1] [,2]
row1   1   22
row2   32 3943

Pearson's Chi-squared test with Yates' continuity correction

data: contingency_table
X-squared = 0.5139, df = 1, p-value = 0.4735

Fisher's Exact Test for Count Data

data: contingency_table
p-value = 0.174
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.1316469 36.8349874
sample estimates:
odds ratio
  5.594729

Warning message:
In chisq.test(contingency_table) :
  Chi-squared approximation may be incorrect

***V cholerae* VFDB**

two_way_classification_chi_function (23,165,1,3998)
    [,1] [,2]
row1   1   22
row2 164 3811

        Pearson's Chi-squared test with Yates' continuity correction

data: contingency_table
X-squared = 0.223, df = 1, p-value = 0.6367

        Fisher's Exact Test for Count Data

data: contingency_table
p-value = 0.6217
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.02544047 6.61662864
sample estimates:
odds ratio
  1.056276

Warning message:
In chisq.test(contingency_table) :
  Chi-squared approximation may be incorrect

Appendix B4

Gene Feature Summary Statistics

Summary Statistics For Gene Lengths

Bacterial Organism	Gene List Type	Descriptive Statistic (Units = Base Pair)				
		Mean	Median	Maximum	Minimum	Standard Deviation
H pylori	**Positive Selection**	1289	1113	3873	207	744
	Virulence	1190	942	8682	243	998
	Genome	958.8	807	8682	39	717
N meningitidis	**Positive Selection**	1281	1038	5322	216	856.5
	Virulence	1280	1026	5322	300	941.3
	Genome	857.8	717	6048	72	636.9
V cholerae	**Positive Selection**	1871	1230	9792	426	2003
	Virulence	1230	957	13680	270	1268.5
	Genome	922	783	13680	81	715.93

Table B4A : Descriptive statistical summary of gene lengths obtained for genes under positive selection, database identified virulence genes and the genomic genes for *H pylori*, *N meningitidis* and *V cholerae*.

Summary Statistics For %GC Content

Bacterial Organism	Gene List Type	Descriptive Statistic (Units = % GC)				
		Mean	Median	Maximum	Minimum	Standard Deviation
H pylori	Positive Selection	39.29	39.48	46.26	29.74	3.01
	Virulence	39.78	39.97	46.89	30.43	3.61
	Genome	39.11	39.44	60.61	21.51	3.75
N meningitidis	Positive Selection	54.95	55.6	62.24	29.46	4.48
	Virulence	50.07	51.74	60.53	24.54	6.33
	Genome	51.65	53.4	67.86	17.78	7.09
V cholerae	Positive Selection	48.97	47.04	54.6	34.9	4.05
	Virulence	47.04	48.8	54.39	28.07	5.38
	Genome	47	47.9	56.5	25.6	4.46

Table B4B : Descriptive statistical summary of %GC content obtained for genes under positive selection, database identified virulence genes and the genomic genes for *H pylori*, *N meningitidis* and *V cholerae*.

Appendix C1

COG Functional Category Enrichment Statistical Tests

H. pylori COG Functional Category Enrichment Statistical Tests

Total number of *H. pylori* genome genes with COG annotations = 978
Total number of *H. pylori* positively selected genes with COG annotations = 178 of 230 genes
Total number of *H. pylori* virulence genes with COG annotations = 118 of 153 genes

COG Category	Number of *H. pylori* Database Identified Virulence Genes in COG Category	Number of *H. pylori* genome genes in COG Category	P value	Adjusted P value
[C] Energy production and conversion	2	67	0.01762	0.4096
[E] Amino acid transport and metabolism	6	80	0.2588	1
[G] Carbohydrate transport and metabolism	3	31	0.8929	1
[H] Coenzyme transport and metabolism	2	58	0.03633	0.21798
[J] Translation	2	93	0.003517	0.03517
[K] Transcription	2	33	0.4154	1
[L] Replication, recombination and repair	2	92	0.003827	0.03517
[M] Cell wall/membrane biogenesis	10	92	0.84	1
[N] Cell motility	41	48	2.2e-16	2.86e – 15
[O]Posttranslational modification, protein turnover, chaperones	13	58	0.02221	0.15547
[P] Inorganic ion transport and metabolism	15	47	5.068e-05	0.00055748
[T] Signal transduction mechanisms	5	25	0.2122	1
[U] Intracellular trafficking and secretion	33	55	2.2e-16	2.86e – 15

Table C1A : Results of a one-way classification Chi-Square tests to determine which COG categories are enriched for *H. pylori* database identified virulence genes. Holm's method was used to correct P values for multiple hypothesis testing.

COG Category	Number of *H. pylori* Positively Selected Genes in COG Category	Number of *H. pylori* genome genes in COG Category	P value	Adjusted P value
[C] Energy production and conversion	4	67	0.004867	0.087606
[D] Cell cycle control, mitosis and meiosis	8	20	0.0178	0.3026
[E] Amino acid transport and metabolism	18	80	0.3740	1
[F] Nucleotide transport and metabolism	2	34	0.06832	1
[G] Carbohydrate transport and metabolism	6	31	0.9464	1
[H] Coenzyme transport and metabolism	10	58	0.9843	1
[I] Lipid transport and metabolism	12	38	0.04933	0.7892
[J] Translation	16	93	0.9041	1
[K] Transcription	4	33	0.4894	1
[L] Replication, recombination and repair	14	92	0.524	1
[M] Cell wall/membrane biogenesis	19	92	0.6182	1
[N] Cell motility	6	48	0.391	1
[O]Posttranslational modification, protein turnover, chaperones	9	58	0.711	1
[P] Inorganic ion transport and metabolism	10	47	0.714	1
[Q] Secondary metabolites biosynthesis, transport and catabolism	3	12	0.4658	1
[T] Signal transduction mechanisms	7	25	0.1939	1
[U] Intracellular trafficking and secretion	7	55	0.3665	1
[V] Defense mechanisms	7	29	0.5505	1

Table C1B : Results of a one-way classification Chi-Square tests to determine which COG categories are enriched for *H. pylori* genes under positive selection. Holm's method was used to correct P values for multiple hypothesis testing.

COG Category	Number of *H. pylori* database identified virulence genes under positive selection belonging to COG Category	P value	Adjusted P value
[E] Amino acid transport and metabolism	1	0.8788	1
[J] Translation	2	0.02805	0.25245
[L] Replication, recombination and repair	1	0.2826	1
[M] Cell wall/membrane biogenesis	1	0.6812	1
[N] Cell motility	3	0.03281	0.262
[O] Posttranslational modification, protein turnover, chaperones	1	0.6676	1
[P] Inorganic ion transport and metabolism	1	0.1353	0.81180
[T] Signal transduction mechanisms	1	0.9113	1
[U] Intracellular trafficking and secretion	2	0.1025	0.71750

Table C1C : Results of a two-way classification Chi-Square tests to determine which COG categories are enriched for *H. pylori* database identified virulence genes that are under positive selection. Holm's method was used to correct P values for multiple hypothesis testing.

N. meningitidis **COG Functional Category Enrichment Statistical Tests**

Number of *N. meningitidis* genome genes with COG annotations = 1356
Number of *N. meningitidis* positively selected genes with COG annotations = 199 of 218 genes
Number of *N. meningitidis* virulence genes with COG annotations = 47 of 68 genes

COG Category	Number of *N. meningitidis* Database Identified Virulence Genes in COG Category	Number of *N. meningitidis* genome genes in COG Category	P value	Adjusted P value
[E] Amino acid transport and metabolism	1	129	0.1327	0.5308
[G] Carbohydrate transport and metabolism	2	49	0.8746	1
[M] Cell wall/membrane biogenesis	20	119	7.176e-16	4.3056e – 15
[N] Cell motility	15	21	2.2e-16	1.76e – 15
[O]Posttranslational modification, protein turnover, chaperones	1	64	0.615	1
[P] Inorganic ion transport and metabolism	6	81	0.09167	0.4583
[Q] Secondary metabolites biosynthesis, transport and catabolism	1	18	0.8723	1
[U] Intracellular trafficking and secretion	20	55	2.2e-16	1.76e – 15

Table C1D : Results of a one-way classification Chi-Square tests to determine which COG categories are enriched for *N. meningitidis* database identified virulence genes. Holm's method was used to correct P values for multiple hypothesis testing.

COG Category	Number of *N. meningitidis* Positively Selected Genes in COG Category	Number of *N. meningitidis* genome genes in COG Category	P value	Adjusted P value
[C] Energy production and conversion	11	105	0.2617	1
[D] Cell cycle control, mitosis and meiosis	6	24	0.2497	1
[E] Amino acid transport and metabolism	16	129	0.5248	1
[F] Nucleotide transport and metabolism	9	39	0.2023	1
[G] Carbohydrate transport and metabolism	5	49	0.4868	1
[H] Coenzyme transport and metabolism	15	71	0.1598	1
[I] Lipid transport and metabolism	5	40	0.8667	1
[J] Translation	19	122	0.873	1
[K] Transcription	4	63	0.08358	1
[L] Replication, recombination and repair	20	152	0.6603	1
[M] Cell wall/membrane biogenesis	25	119	0.05634	1
[N] Cell motility	6	21	0.1329	1
[O]Posttranslational modification, protein turnover, chaperones	8	64	0.7468	1
[P] Inorganic ion transport and metabolism	11	81	0.9002	1
[Q] Secondary metabolites biosynthesis, transport and catabolism	2	18	0.9244	1
[R] General function prediction only	20	167		
[S] Function unknown	12	126		
[T] Signal transduction mechanisms	5	20	0.3191	1
[U] Intracellular trafficking and secretion	12	55	0.1823	1
[V] Defense mechanisms	5	15	0.09168	1
[W] Extracellular structures	1	1	0.1468	1

Table C1E : Results of a one-way classification Chi-Square tests to determine which COG categories are enriched for *N. meningitidis* genes under positive selection. Holm's method was used to correct P values for multiple hypothesis testing.

COG Category	Number of *N. meningitidis* database identified virulence genes under positive selection belonging to COG Category	P value	Adjusted P value
[G] Carbohydrate transport and metabolism	1	0.1956	0.5574
[M] Cell wall/membrane biogenesis	10	0.001423	0.007115
[N] Cell motility	1	0.001677	0.007115
[P] Inorganic ion transport and metabolism	2	0.1858	0.5574
[U] Intracellular trafficking and secretion	3	0.5025	0.5574

Table C1F : Results of a two-way classification Chi-Square tests to determine which COG categories are enriched for *N. meningitidis* database identified virulence genes that are under positive selection. Holm's method was used to correct P values for multiple hypothesis testing.

***V. cholerae* COG Functional Category Enrichment Statistical Tests**

Total number of *V. cholerae* genome genes with COG annotations = 2507
Total number of *V. cholerae* positively selected genes with COG annotations = 18 of 23 genes
Total number of *V. cholerae* virulence genes with COG annotations = 142 of 169 genes

COG Category	Number of *V. cholerae* Database Identified Virulence Genes in COG Category	Number of *V. cholerae* genome genes in COG Category	P value	Adjusted P value
[D] Cell cycle control, mitosis and meiosis	2	35	0.7224	1
[E] Amino acid transport and metabolism	1	250	0.0002615	0.003138
[G] Carbohydrate transport and metabolism	2	143	0.03696	0.3264
[H] Coenzyme transport and metabolism	3	130	0.1322	0.9254
[I] Lipid transport and metabolism	2	78	0.3398	1
[K] Transcription	2	216	0.002723	0.029953
[M] Cell wall/membrane biogenesis	13	160	0.2244	1
[N] Cell motility	65	133	2.2e-16	3.08e – 15
[O]Posttranslational modification, protein turnover, chaperones	8	124	0.8494	1
[P] Inorganic ion transport and metabolism	14	156	0.09529	0.76232
[Q] Secondary metabolites biosynthesis, transport and catabolism	8	54	0.00983	0.0983
[T] Signal transduction mechanisms	17	227	0.2728	1
[U] Intracellular trafficking and secretion	38	88	2.2e-16	3.08e – 15
[V] Defense mechanisms	1	45	0.4949	1

Table C1G : Results of a one-way classification Chi-Square tests to determine which COG categories are enriched for *V. cholerae* database identified virulence genes. Holm's method was used to correct P values for multiple hypothesis testing.

COG Category	Number of *V. cholerae* Positively Selected Genes in COG Category	Number of *V. cholerae* genome genes in COG Category	P value	Adjusted P value
[C] Energy production and conversion	1	168	1	1
[E] Amino acid transport and metabolism	5	250	0.02739	0.2739
[F] Nucleotide transport and metabolism	1	64	0.3732	1
[G] Carbohydrate transport and metabolism	1	143	0.6293	1
[I] Lipid transport and metabolism	1	78	0.435	1
[J] Translation	1	158	0.722	1
[K] Transcription	2	216	0.6642	1
[M] Cell wall/membrane biogenesis	3	160	0.1029	0.9261
[P] Inorganic ion transport and metabolism	1	156	0.7099	1
[T] Signal transduction mechanisms	1	227	0.9148	1

Table C1H : Results of a one-way classification Chi-Square tests to determine which COG categories are enriched for *V. cholerae* genes under positive selection. Holm's method was used to correct P values for multiple hypothesis testing.

VC0837, the only *V. cholerae* database identifed virulence gene under positive selection does not belong to a COG's alphabetical biological process category.

Appendix C2

PSORT Sub-Cellular Localisation Enrichment Statistical Tests

Chi-Square Tests for Association between PSORT Categories and Genes under Positive Selection

Total Number of *H pylori* genes annotated by PSORT cutoff > 7.5 = **861**

Total Number of *H pylori* genes under positive selection annotated by PSORT cutoff > 7.5 = **151**

Total Number of *N meningitidis* genes annotated by PSORT cutoff > 7.5 = **1103**

Total Number of *N meningitidis* genes under positive selection annotated by PSORT cutoff > 7.5 = **131**

Total Number of *V cholerae* genes annotated by PSORT cutoff > 7.5 = **2162**

Total Number of *V cholerae* genes under positive selection annotated by PSORT cutoff > 7.5 = **14**

PSORT Category	**Bacterial Organism : *H pylori***			
	Genome Number	**N^{o} of genes under Positive Selection**	**P – Value**	**Adjusted P – Value**
Cytoplasmic	529	82	0.05853	0.17559
Cytoplasmic Membrane	238	46	0.4512	0.90240
Periplasmic	13	1	0.4845	0.90240
OuterMembrane	69	22	0.001920	0.00768
ExtraCellular	12	0	Not Tested	Not Tested
PSORT Category	**Bacterial Organism : *N meningitidis***			
	Genome Number	**N^{o} of genes under Positive Selection**	**P – Value**	**Adjusted P – Value**
Cytoplasmic	749	94	0.365	0.73
Cytoplasmic Membrane	274	24	0.08323	0.24969
Periplasmic	32	2	0.4163	0.73
OuterMembrane	44	11	0.01213	0.04582
ExtraCellular	4	0	Not Tested	Not Tested
PSORT Category	**Bacterial Organism : *V cholerae***			
	Genome Number	**N^{o} of genes under Positive Selection**	**P – Value**	**Adjusted P – Value**
Cytoplasmic	1303	6	0.2884	0.8652
Cytoplasmic Membrane	683	6	0.3919	0.8652
Periplasmic	80	0	Not Tested	Not Tested
OuterMembrane	28	1	0.1673	0.6692
ExtraCellular	68	1	0.3616	0.8652

Table C2A : Results of one-way classification Chi-square tests to determine if genes under positive selection in *H pylori*, *N meningitidis* and *V cholerae* have statistically significant association with a predicted PSORT sub-cellular localisation site. Holm's method was used to adjust P values for multiple hypotheses testing.

Chi-Square Tests for Association between PSORT Categories and Database Identified Virulence Genes

Total Number of *H pylori* genes annotated by PSORT cutoff > 7.5 = **861**

Total Number of *H pylori* virulence genes annotated by PSORT cutoff > 7.5 = **94**

Total Number of *N meningitidis* genes annotated by PSORT cutoff > 7.5 = **1103**

Total Number of *N meningitidis* virulence genes annotated by PSORT cutoff > 7.5 = **33**

Total Number of *V cholerae* genes annotated by PSORT cutoff > 7.5 = **2162**

Total Number of *V cholerae* virulence genes annotated by PSORT cutoff > 7.5 = **95**

PSORT Category	Bacterial Organism : *H pylori*			
	Genome Number	**N^0 of Database Identified virulence genes**	**P – Value**	**Adjusted P – Value**
Cytoplasmic	529	40	0.0001072	0.0004288
Cytoplasmic Membrane	238	27	0.8996	1
Periplasmic	13	1	1	1
OuterMembrane	69	21	1.798e-07	8.9900e-07
ExtraCellular	12	5	0.005936	0.0178
PSORT Category	**Bacterial Organism : *N meningitidis***			
	Genome Number	**N^0 of Database Identified virulence genes**	**P – Value**	**Adjusted P – Value**
Cytoplasmic	749	11	3.627e-05	0.00014508
Cytoplasmic Membrane	274	8	0.9016	1
Periplasmic	32	1	1	1
OuterMembrane	44	12	6.57e-10	3.2850e-09
ExtraCellular	4	1	0.1146	0.3438
PSORT Category	**Bacterial Organism : *V cholerae***			
	Genome Number	**N^0 of Database Identified virulence genes**	**P – Value**	**Adjusted P – Value**
Cytoplasmic	1303	36	8.57e-06	2.5710e-05
Cytoplasmic Membrane	683	32	0.7369	1
Periplasmic	80	3	1	1
OuterMembrane	28	10	1.139e-07	5.6950e-07
ExtraCellular	68	14	6.696e-07	2.6784e-06

Table C2B : Results of one-way classification Chi-square tests to determine if database identified virulence genes for *H pylori*, *N meningitidis* and *V cholerae* have statistically significant association with a predicted PSORT sub-cellular localisation site. Holm's method was used to adjust P values for multiple hypotheses testing.

Chi-Square Tests for Association between PSORT Categories for Database Identified Virulence Genes and Genes under Positive Selection

Total Number of *H pylori* genes under positive selection annotated by PSORT cutoff > 7.5 = **151**

Total Number of *H pylori* virulence genes annotated by PSORT cutoff > 7.5 = **94**

Total Number of *N meningitidis* genes under positive selection annotated by PSORT cutoff > 7.5 = **131**

Total Number of *N meningitidis* virulence genes annotated by PSORT cutoff > 7.5 = **33**

Total Number of *V cholerae* genes annotated by PSORT cutoff > 7.5 = **14**

Total Number of *V cholerae* virulence genes annotated by PSORT cutoff > 7.5 = **95**

PSORT Category	**Bacterial Organism : *H pylori***			
	N^{o} of genes under Positive Selection	**N^{o} of Database Identified virulence genes**	**P – Value**	**Adjusted P – Value**
Cytoplasmic	82	40 (6 genes in common)	0.8917	1
Cytoplasmic Membrane	46	27 (2 genes in common)	0.1594	0.4782
Periplasmic	1	1 (0 genes in common)		
OuterMembrane	22	21 (8 genes in common)	0.6516	1
ExtraCellular	0	5		
PSORT Category	**Bacterial Organism : *N meningitidis***			
	N^{o} of genes under Positive Selection	**N^{o} of Database Identified virulence genes**	**P – Value**	**Adjusted P – Value**
Cytoplasmic	94	11 (2 genes in common)	0.6372	1
Cytoplasmic Membrane	24	8 (1 gene in common)	0.5245	1
Periplasmic	2	1 (0 genes in common)		
OuterMembrane	11	12 (5 genes in common)	0.1387	0.4161
ExtraCellular	0	1		

Table C2C : Results of a two-way classification Chi-Square tests to determine which PSORT sub-cellular localisation sites have a statistically significant association at the P = 0.05 level for *H. pylori* and *N. meningitidis* database identified virulence genes that are under positive selection. Holm's method was used to correct P values for multiple hypothesis testing.

The only *V cholerae* database identified virulence gene that is under positive selection (VC0837) has a PSORT score of 2 and belongs to the sub-cellular localisation category unknown.

}