QUANTIFYING TRENDS IN BACTERIAL VIRULENCE AND PATHOGEN-ASSOCIATED GENES THROUGH LARGE SCALE BIOINFORMATICS ANALYSIS

by

Anastasia Amber Fedynak
B.Comp., University of Guelph, 2003
B.Sc., University of Alberta, 2002

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

In the
Department of Molecular Biology and Biochemistry

SIMON FRASER UNIVERSITY

2007

APPROVAL

Name: **Anastasia Amber Fedynak**

Degree: **Doctor of Philosophy**

Title of Thesis: **Quantifying trends in bacterial virulence and pathogen-associated genes through large scale bioinformatics analysis**

Examining Committee:

Chair: **Dr. Bruce P. Brandhorst**
Professor, Department of Molecular Biology and Biochemistry

Dr. Fiona S.L. Brinkman
Senior Supervisor
Associate Professor, Department of Molecular Biology and Biochemistry

---

Dr. Lisa Craig
Supervisor
Assistant Professor, Department of Molecular Biology and Biochemistry

---

Dr. Kay C. Wiese
Supervisor
Associate Professor, School of Computing Science

Dr. Jack Chen
Internal Examiner
Associate Professor, Department of Molecular Biology and Biochemistry

---

Dr. Gee W. Lau
External Examiner
Assistant Professor, College of Veterinary Medicine, University of Illinois at Urbana-Champaign

Date Defended: Monday December 10, 2007

ABSTRACT

With an increasing number of bacterial genomes becoming available, we are now able to investigate and quantify selected general trends in pathogenicity shared across diverse pathogens which have been previously anecdotally reported but have not yet been quantified on a larger scale. In addition, we can perform more high-throughput approaches for the identification of virulence-associated genes that represent possible therapeutic or prophylactic targets.

In this study, I systematically examined up to 267 pathogen and non-pathogen genomes from diverse genera, and identified trends associated with a curated data set of known bacterial virulence factors (VFs). I show, in support of previous anecdotal statements, that genomic islands (clusters of genes of probable horizontal origin) disproportionately do contain more VFs than the rest of a given genome ($p < 2.20E-16$), supporting their important role in pathogen evolution.

To gain insights into the types of genes that may play a more virulence-specific role in pathogens, I also performed an analysis to identify pathogen-associated genes (genes found predominately in pathogens across multiple genera, but not found in non-pathogens). I found that disproportionately high numbers of pathogen-associated VFs are "offensive" (involved in active invasion of the host), such as certain types of toxins, as well as Type III and Type IV

secretion systems. Some of the pathogen-specific genes identified have apparently not yet been examined for their potential as vaccine components or drug targets and merit further study.

As the first step in the initiation of more sophisticated analyses of trends in virulence, I also developed a Virulence Gene Experiment Database (VGEDB) that incorporates contextual information about virulence. This database is unique in that entries are centered around describing a particular virulence gene experiment, rather than a virulence gene. I used this database in part to investigate a common BLAST-based approach for computationally identifying VFs in genomic sequences. My analysis suggests that this common VF-prediction method is very inaccurate.

This work in general provides the first large-scale, multi-genera, quantitative data describing selected trends in bacterial virulence and provides global insights regarding pathogen evolution and pathogen-associated traits of primary importance in a pathogenic lifestyle.

Keywords: Bioinformatics; prokaryotes; genomics; virulence factors; pathogen-associated; genomic islands; pathogenicity; bacteria; virulence;

DEDICATION

A very special dedication to the two people who mean the most to me... mom and dad. Thank you so much for all your spirit, support, and love. You have been and will always be the inspiration in my life and career.

To my brother, sister, and extended family... thank you for your generous love and support over the years. You always believed in me and supported me, and I am very grateful for that. I am also grateful (and always look forward to) the delicious Christmas dinner and visiting with everyone each year...

To my friend Pascal... thank you for always being there for me, and providing advice and encouragement when I needed it most. I am deeply grateful to you for making this chapter of my life such a happy and memorable one.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to everyone in the Brinkman Lab, first and foremost - Dr. Fiona Brinkman, you have been an exceptional role model and mentor for me over the years. It has been such a privilege to be part of this incredible team of people, and I sincerely thank all of you. I would also like to thank my supervisory committee, Dr. Lisa Craig, and Dr. Kay Wiese, as well as members of the MBB department for their guidance and support throughout my thesis studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

ADP	Adenosine diphosphate
ATP	Adenosine triphosphate
BCG	Bacillus Calmette-Guerin
BLAST	Basic Local Alignment Search Tool
bp	Base pair
cAMP	Cyclic adenosine monophosphate
CDC	Cholesterol-dependent cytolysin
COG	Clusters of Orthologous Groups
DIMOB	IslandPath-DIMOB method for GI prediction
DINUC	IslandPath-DINUC method for GI prediction
FN	False negative
FP	False positive
GI	Genomic island
GTP	Guanosine triphosphate
HGT	Horizontal gene transfer
IVET	*In vivo* expression technology
LEE	Locus of enterocyte effacement
LLO	Listeriolysin O
MHC	Major histocompatibility complex
NCBI	National Center for Biotechnology Information
ORF	Open reading frame
PAI	Pathogenicity island
PCR	Polymerase chain reaction
Precision	TP/(TP+FP); usually used equivalently to specificity
Recall	TP/(TP+FN); usually used equivalently to sensitivity
RPS-BLAST	Reverse PSI-BLAST

RTX	Repeat in toxin
Sensitivity	TP/(TP+FN)
Specificity	TP/(TP+FP)
SPI	*Salmonella* pathogenicity Island
STM	Signature-tagged mutagenesis
T3S	Type III secretion
T3SS	Type III secretion system
T4S	Type IV secretion
T4SS	Type IV secretion system
TIGR	The Institute for Genomic Research
TN	True negative
TP	True positive
TvFac	Toxin and Virulence Factor Database
VF	Virulence factor
VFDB	Virulence Factor Database
VGEDB	Virulence Gene Experiment Database

CHAPTER 1 INTRODUCTION

1.1 Infectious disease: a global burden

Infectious diseases are among the leading causes of mortality with an estimated 15 million annual deaths worldwide (Morens et al. 2004). The discovery and use of antibiotics in the 1940s and 1950s posed a milestone in history, with major declines in mortality rates associated with certain bacterial infections. However, along with the increasing use of antibiotics came the rapid emergence of pathogens that are resistant to antibiotics (Binder et al. 1999), which posed a major problem for disease treatment and therapy. Continuous overuse of antibiotics over the decades has now led to a serious crisis, with an increasing incidence of pathogens that are resistant to multiple antibiotics (Morens et al. 2004), including multiply-resistant *S. aureus* strains resistant to penicillin, methicillin, and vancomycin (Centers for Disease Control and Prevention 2002) – one of the last remaining treatments available. Furthermore, recent bioterrorist attacks, such as the anthrax attack in 2001 (Jernigan et al. 2002), has heightened public alarm about our vulnerability and susceptibility to these diseases.

Further study of pathogens and their complex interactions with the host, pathogen and environment can provide fundamental insights of pathogenic mechanisms and traits that may aid in the development of new vaccines and

antimicrobials essential for combating the re-emerging threat of infectious diseases of bacterial origin.

1.2 The concept of bacterial virulence

In 1890 Robert Koch devised a set of four scientific criteria, now known as Koch's postulates, used to establish that a particular pathogen was the causative agent of an infectious disease (Table 1.1). The first of these postulates states that the pathogen should be isolated from all cases of the disease and not be associated with healthy individuals. Secondly, the pathogen must be isolated from the infected individual, and it should be able to be grown in pure culture. Thirdly, the disease state should reoccur when the pathogen is used to infect a healthy individual, and finally, the pathogen can again be isolated from this newly infected individual. However, it was soon noted that these postulates cannot be universally applied to all pathogens. For example, the virulence of some pathogens can range in severity and therefore may not result in a similar disease state in different individuals. Furthermore, some bacteria cannot easily be grown in pure culture as they are difficult to grow under normal laboratory conditions.

Table 1.1 Koch's Postulates to determine if a pathogen is the causative agent of a particular disease

1. **The pathogen should be isolated from all cases of the disease, but absent from healthy individuals**
2. **The pathogen must be isolated from the infected individual and grown in pure culture**
3. **The disease state should reoccur when the pathogen is used to infect a healthy individual**
4. **The pathogen can again be isolated from this newly infected individual**

Stanley Falkow attempted to add more rigor in the identification of molecular factors involved in virulence, and in the late 1980's proposed "Molecular Koch's Postulates" (Falkow 1988) to define a virulence factor (Table 1.2). There are many versions of these postulates, but they are basically composed of the following three criteria: First, the virulence factor should be present in all pathogenic strains and absent from any close non-pathogenic relatives. Second, inactivation of the virulence factor gene should result in attenuated virulence in an animal infection model. Lastly, complementation of the inactivated gene with the functional one should re-establish virulence in the animal model. Although these postulates serve well as guidelines, Falkow noted that identifying virulence factors is becoming more complex, and that these postulates have certain limitations and should not always be strictly followed (Falkow 2004). For example, a particular virulence factor gene may be responsible for virulence in some hosts but not others. Additionally, it is increasingly difficult to clearly distinguish pathogenic and non-pathogenic species.

Table 1.2 Molecular Koch's Postulates to determine that a given virulence factor gene contributes to disease

1. The gene or phenotype should be associated with pathogenic strains and absent from non-pathogenic strains
2. Inactivation of the gene should result in attenuated virulence in an appropriate infection model
3. Complementation of the inactivated gene with the functional one should re-establish virulence in the animal model

Through the years, we have gained a better understanding of the complex nature of host-pathogen interactions. There is now an increasing appreciation that infectious disease is a much more complex phenomenon, and is the result of an interplay between the host, pathogen and environment. The virulence outcome can depend on multiple factors including specific traits inherent to the host and pathogen as well as the environmental niche. For example, the ability of different hosts to eliminate the pathogen can vary significantly, from complete clearance, to asymptomatic carrier states, to a more severe onset of the disease. Some bacteria can colonize specific privileged sites in the body where conditions are favourable for their growth and proliferation, but not other sites. Furthermore, establishing infection is greatly mediated by specific factors, commonly called virulence factors, which are utilized by a particular pathogen to cause disease. The following section will discuss the definition of a virulence factor in more detail.

1.3 The definition of a virulence factor

Virulence factors (VFs) have previously been defined as a "microbial product that permits the pathogen to cause disease" (Casadevall et al. 1999). In this context, the term VF is quite generic, and can be used to describe any factor involved in virulence. VFs have a wide variety of roles in the disease process, and their degree of virulence or damage to the host can vary widely. Some of the major roles of VFs include: facilitate attachment or invasion of host cells, causing direct damage to the host cell or surrounding tissue, evasion of host defences,

and gathering nutrients from the environment. Some VFs also play a role in transport or regulation of other VFs.

It has also been noted that researchers often tend to distinguish between 'true virulence genes', 'virulence-associated genes', and 'virulence life-style genes' to some degree (Wassenaar et al. 2001). In this case, 'true virulence genes' are pathogen-specific (i.e. absent from nonpathogens) and are defined as 'gene products directly involved in interactions with the host and are directly responsible for pathological damage'. There is also an increasing appreciation that many classic VFs, originally thought to be pathogen-specific, are being found in non-pathogens (Zhang et al. 2003), which re-iterates the fact that virulence and VFs are quite complex. In fact, some have proposed that the term VF should be used less and that they should instead be referred to instead as "host interaction factors" (Holden et al. 2004).

In summary, VFs are complex factors, and their role and ability to cause damage to the host varies widely. Nonetheless, certain VFs have been well established, as we have a good understanding of the mechanistic role of some of them in the disease process. Some of the major VF categories that are relevant to this dissertation work are discussed in greater detail in the next section.

1.4 Virulence factors

Several stages in the infection process are mediated by bacterial VFs that allow the pathogen to establish infection. When the pathogen first comes into contact with its host, it requires specialized factors to infect and colonize its host.

Initial attachment usually involves adhesion factors that bind to the host cell surface (section 1.4.1). Following attachment, some pathogens can invade host cells or the underlying tissues in the body using invasins (section 1.4.2). Most pathogenic bacteria also produce and secrete toxins (section 1.4.3), which can cause damage to the host tissue, but whose actions are somehow beneficial to the bacteria. They must also possess mechanisms to evade host immune defenses once they have been detected (section 1.4.4), and obtain nutrients from the environment that are essential for their growth and survival (section 1.4.5). Additionally, bacteria possess various VF transport systems (section 1.4.6), including specialized systems that directly inject factors into the host cells. Finally, bacteria have evolved strategies to regulate the production of particular VFs at different stages of infection (section 1.4.7).

For the purposes of this thesis, all factors that facilitate infection of the bacteria will be collectively referred to as VFs. Table 1 provides a summary of the major types of VFs, as well as some examples from well established pathogens. In the following sections, I will describe these types of VFs in more detail, as well as discuss the mechanism and function of selected VFs during infection. This is not meant to represent a comprehensive list of all bacterial VFs, but only to provide the basic concept of how these types of VFs can be utilized by a pathogen and their relative contributions to the disease process.

Table 1.3 Types of bacterial virulence factors

Type of virulence factor	Example virulence factor and species
Adhesins	Type 1 fimbriae (*Salmonella typhimurium*) Flagella (*Pseudomonas aeruginosa*) Type IV bundle-forming pili (enteropathogenic *Escherichia coli)* Intimin and Tir (enteropathogenic *Escherichia coli*)
Invasins	SPI-1 encoded genes (*Salmonella typhimurium*) Invasin (*Yersinia enterolitica*)
Toxins	Cholera toxin (*Vibrio cholerae*) Pertussis toxin (*Bordetella pertussis*) Anthrax toxin (*Bacillus anthracis*) Adenylate cyclase toxin (*Bordetella pertussis*) Listeriolysin O (*Listeria monocytogenes*) Hemolysin (*Escherichia coli*)
Evasion of host defences	Alginate (*Pseudomonas aeruginosa)* IgA1 protease (*Neisseria gonorrhoeae*)
Iron Uptake	Enterobactin (*Salmonella typhimurium*) Yersiniabactin (*Yersinia pestis*)
Transport of VFs	Type III secretion system (*Yersinia pestis*) Type IV secretion system (*Agrobacterium tumefaciens*)
Regulation of VFs	Quorum sensing (*Pseudomonas aeruginosa*)

1.4.1 Adhesins

Adhesions mediate the initial interaction between a pathogen and its host. Many structures on the surface of a bacterial cell have been shown to function in adhesion in some way, although their primary role may not necessarily be in adhesion. Different adhesions can preferentially bind to selected tissue cell types

and cellular molecules. For example, enteropathogenic *E. coli* utilizes type IV bundle-forming pili to mediate initial attachment to host epithelial cells (Tobe et al. 2002). Additional adhesion proteins, intimin and Tir (translocated intimin receptor), facilitate more intimate attachment (Jores et al. 2004). These two proteins are encoded on the LEE (Locus of Enterocyte Effacement) pathogenicity island, in addition to numerous effectors, and genes encoding a Type III secretion system (T3SS; section 1.4.6). Tir is secreted into the host cell via the T3SS, and mediates intimate attachment by binding to intimin on the bacterial cell surface. At the same time, translocated effector proteins mediate actin cytoskeletal rearrangements leading to formation of the pedestal-like structures characteristic of *E. coli* infections (Jerse et al. 1990; Jerse et al. 1991).

1.4.2 Invasins

Invasins are factors that facilitate invasion of the pathogen into the host cell or underlying tissues. They typically act by disrupting the host cell cytoskeleton or signalling pathways, allowing entry of the bacteria into host cells or dissemination throughout the body. For example, at least 13 secreted effectors encoded in SPI-1 (*Salmonella* Pathogenicity Island-1) of *Salmonella typhimurium* are involved in invasion, most of which function by disrupting host cell actin cytoskeleton leading to membrane ruffling (Zhou et al. 2001). Effective invasion of the pathogen typically allows for more suitable conditions for their growth and survival.

1.4.3 Toxins

Many bacterial pathogens produce and secrete exotoxins that are critical for pathogenesis. The symptoms associated with an infectious disease can be a direct result of the activity of the toxins. Such toxins are generally unique to different pathogens and their toxicity can range in severity. For example, some are cytotoxic on their own, and can directly cause death of the target cell. Others may allow the pathogen to escape from host immune cells and enter a more favorable environment.

Toxins (exotoxins) can be generally classified into three categories: Type I: Bacterial superantigens, Type II: Membrane damaging toxins, and Type III: Intracellular acting toxins. Some toxins, however, do have multiple functions and are therefore difficult to classify into one of these categories. A good example of this is the anthrax toxin from *Bacillus anthracis.* The anthrax toxin is composed of three components. Two of these components, the lethal factor and edema factor, both act intracellularly like Type III toxins, while the third component, the protective antigen, is more similar to Type II toxins, as it forms a pore that facilitates delivery of the lethal factor and edema factor into the cell cytosol (reviewed in (Ascenzi et al. 2002)). Nonetheless, this classification system represents a good general system for classifying toxins that is loosely based on their mechanism of action.

Since toxins became one focus of this thesis, in the sections below I will give a brief summary of these three categories of toxins, as well as describe in

more detail selected toxins, their mechanism of action, and their role in pathogenesis.

Type I toxins: Bacterial superantigens

Bacterial superantigens are an unusual type of bacterial toxin produced by *Staphylococcus aureus* and *Streptococcus pyogenes* (Marrack et al. 1990). These toxins are unique as they can directly bind to Major Histocompatibility Complex class II (MHC class II) and T-cell receptors, stimulate a large number of T-cells, and induce a massive inflammatory response (reviewed in (Herman et al. 1991)). In normal antigen presentation, the antigen is first digested by macrophages into peptides, which are then presented in a complex with MHC class II at the cell surface. These complexes are recognized by a small number T-cells and stimulate their proliferation. Superantigens, however, skip the digestion step as they can directly interact with MHC class II and T-cell receptors and activate T-cells. The number of stimulated T-cells is much greater, resulting in a much more profound inflammatory response and damage to surrounding epithelial cells, which can lead to disease manifestations such as toxic shock in the case of *S. aureus* infection (Kotzin et al. 1993).

Type II toxins: Membrane damaging toxins

There are generally two types of membrane damaging toxins: 1) toxins that disrupt the integrity of host cell membranes, and 2) pore-forming toxins, toxins that "punch holes" in the host cell membrane. The first class of toxins generally exhibit an enzymatic activity that damages phospholipids in the host

cell membrane, leading to cell lysis. One example is Phospholipase C in *Listeria monocytogenes*. The second type of membrane-damaging toxin, the pore-forming toxins, comprises the majority of Type II toxins. They function by forming a pore or channel in the host cell membrane.

The Cholesterol Dependent Cytolysins (CDCs) comprise one family of pore-forming toxins. They preferentially bind to cholesterol-rich membranes and oligomerize to form large pores. These pores generally consist of 30-50 subunits and range in diameter from 250-300Å. To date, CDCs have been identified in 5 genera of pathogens: *Clostridium*, *Streptococcus*, *Listeria*, *Bacillus*, and *Arcanobacterium.* Although the mechanism of pore-formation in these CDCs is relatively similar, they are structurally distinct and seem to contribute to different aspects of pathogenesis (reviewed in (Tweten 2005)). Listeriolysin O (LLO) is one CDC found in the human pathogen *Listeria monocytogenes.* LLO has a unique property that is not found in other non-listerial CDCs: its activity is pH-dependent. This property enables *Listeria* to escape the host immune system and establish infection. Geoffroy, *et al* showed that purified LLO is highly active in acidic pH, and exhibits low activity in neutral pH (Geoffroy et al. 1987). This is important for Listerial pathogenesis since during the course of infection, *Listeria* are engulfed by phagosomes. The acidic environment of the phagosome triggers LLO activity, which perforates the phagosome and allows the bacteria to escape into the more neutral environment of the host cell cytosol where they can then thrive and proliferate (Tweten 2005).

Type III toxins: Intracellular toxins

Many bacterial pathogens possess toxins that act intracellularly in the host cell. Some toxins function by altering or regulating synthesis of cyclic AMP (cyclic adenosine monophosphate). cAMP is an important messenger required for many metabolic and cellular processes, and its production is strictly regulated. Toxins that alter cAMP production are therefore usually critical for the progression of a bacterial infection.

Adenylate cyclase toxins are able to themselves catalyze the production of cAMP. For example, the invasive adenylate cyclase toxin from *Bordetella pertussis* is able to significantly disrupt the immune defense mechanism by specifically targeting immune cells. cAMP accumulation in these cells significantly reduces or halts their normal cellular function thereby weakening immune defenses. A study by Confer *et al.,* showed that phagocytic cells incubated with *Bordetella* extracts exhibited increased accumulation of cAMP, as well as reduced superoxide generation, chemotaxis, particle ingestion, and killing capacity (Confer et al. 1982). In addition to adenylate cyclase from *Bordetella*, three additional adenylate cyclase toxins have been identified: the edema factor of *B. anthracis*, exoenzyme Y of *Pseudomonas aeruginosa*, and adenylate cyclase in *Yersinia pestis* (reviewed in (Ahuja et al. 2004)).

Some bacterial toxins utilize an ADP-ribosyltransferase (adenosine diphosphate-ribosyltransferase) mechanism to increase intracellular cAMP levels. For example, both pertussis toxin and cholera toxin ADP-ribosylate G protein subunits leading to continual synthesis of cAMP from ATP (adenosine

triphosphate) by cellular adenylyl cyclase (Cassel et al. 1978; Katada et al. 1982). This ultimately leads to symptoms like the massive diarrhea associated with *V. cholera* infection and anaphylaxis associated with *B. pertussis* infection. Additionally, diphtheria toxin from *Corynebacterium diphtheria* and exotoxin A from *P. aeruginosa* utilize an ADP-ribosylation mechanism to inactivate eukaryotic elongation factor 2 (functions in elongation of polypeptide chains) and thereby causing death of the target cell (Collier 2001). Finally, exoenzymes S and T from *P. aeruginosa* are bifunctional toxins that contain both ADP-ribosyltransferase and GTPase-activitating domains (Barbieri et al. 2004). These toxins act on different cellular targets and so likely contribute to different aspects of pathogenesis. The specific cellular substrates inactivated by ADP-ribosylation are not entirely known, however reports have found that the ADP-ribosyltransferase domains are involved in antiphagocytosis, disruption of the actin cytoskeleton, and cause apoptosis of host cells (Aktories et al. 2005; Barbieri et al. 2004)

1.4.4 Evasion of host defences

Many bacteria have evolved mechanisms that evade the host's immune system defences. Various mechanisms are used by pathogens to block different stages in the host immune response. One strategy involves inactivation of antibodies using antibody-specific proteases. The IgA1 protease of *Neisseria gonorrhoeae* is one representative example of this (Pohlner et al. 1987). Certain bacteria can also produce extracellular capsules or biofilms that prevent

phagocytosis, such as the alginate biofilm produced by *P. aeruginosa* for example (Simpson et al. 1988).

1.4.5 Iron uptake

Many bacteria have evolved mechanisms to uptake iron from their host (for a review, see (Wooldridge et al. 1993). One method is by the production of siderophores. Siderophores are compounds with a high affinity for iron, which scavenge and remove iron from host proteins. Enterobactin, is a classic example of one such siderophore that binds iron with a very high affinity (Pollack et al. 1970). Another strategy used is to secrete toxins that kill host cells and release cellular contents, making iron more easily accessible and hence significantly enhancing bacterial growth (Waalwijk et al. 1983).

1.4.6 Transport of virulence factors

To date, 7 different types of secretion systems (referred to as Type I-VII or T1SS-T6SS) have been described that are associated with the transport of VFs (Gerlach et al. 2007). These transport mechanisms can be generally divided into two classes, those that secrete factors into extracellular space (for example, Type I, II, and V), and those that directly inject VFs, or "effectors", into the target cell upon contact (for example, Type III and IV).

T1SSs incorporate a Sec-independent process that delivers the protein directly into the extracellular space in a single step. This is though to be the simplest of systems, consisting of a multi-protein apparatus that includes an ATPase-binding cassette transporters (Holland et al. 2005).

T2SSs incorporate a two-step process: firstly, proteins are secreted into the periplasm by the general secretory pathway, and secondly, they are further transported into extracellular space via a protein complex containing secretin (Johnson et al. 2006). The biogenesis of the T2SSs is thought to function in a similar manner to that of type IV pili.

In T3SSs and T4SSs, effector proteins are injected directly into the target cell cytosol through a needle-like complex referred to as an 'injectisome'. T4SSs can also mediate conjugal transfer between bacteria. These systems are discussed further in the sections below.

In T5SSs, also referred to as autotransporters or self-transporters, proteins are secreted in two steps: first they are secreted across the inner membrane into the periplasm via the Sec-dependent pathway, and second, they utilize a self-transport mechanism that releases the protein outside of the cell (Henderson et al. 2004).

The mechanisms of T6SSs have not yet been fully studied, however they have been shown to secrete potential effector proteins via a host-cell contact dependent mechanism (Pukatzki et al. 2006), Additionally, identified substrates lack an N-terminal signal sequence suggesting their secretion is Sec-independent (Gerlach et al. 2007). T6SS have been identified in both *V. cholera* and *P. aeruginosa* (Mougous et al. 2006; Pukatzki et al. 2006).

T7SSs have been relatively recently identified and its mechanism of secretion is not fully understood. It appears to be distinct from other secretion systems in that the secreted proteins seem to be dependent upon one another

for effective secretion (Abdallah et al. 2007). T7SSs have been identified in a variety of Gram-positive bacteria (reviewed in (Abdallah et al. 2007)), although certain systems, such as that found in *L. monocytogenes* for example, have not been shown to play a role in virulence (Way et al. 2005).

In the sections below, I will discuss in more detail the structure and components of T3SSs and T4SSs, since these systems became a topic of focus during this thesis study.

Type III secretion systems

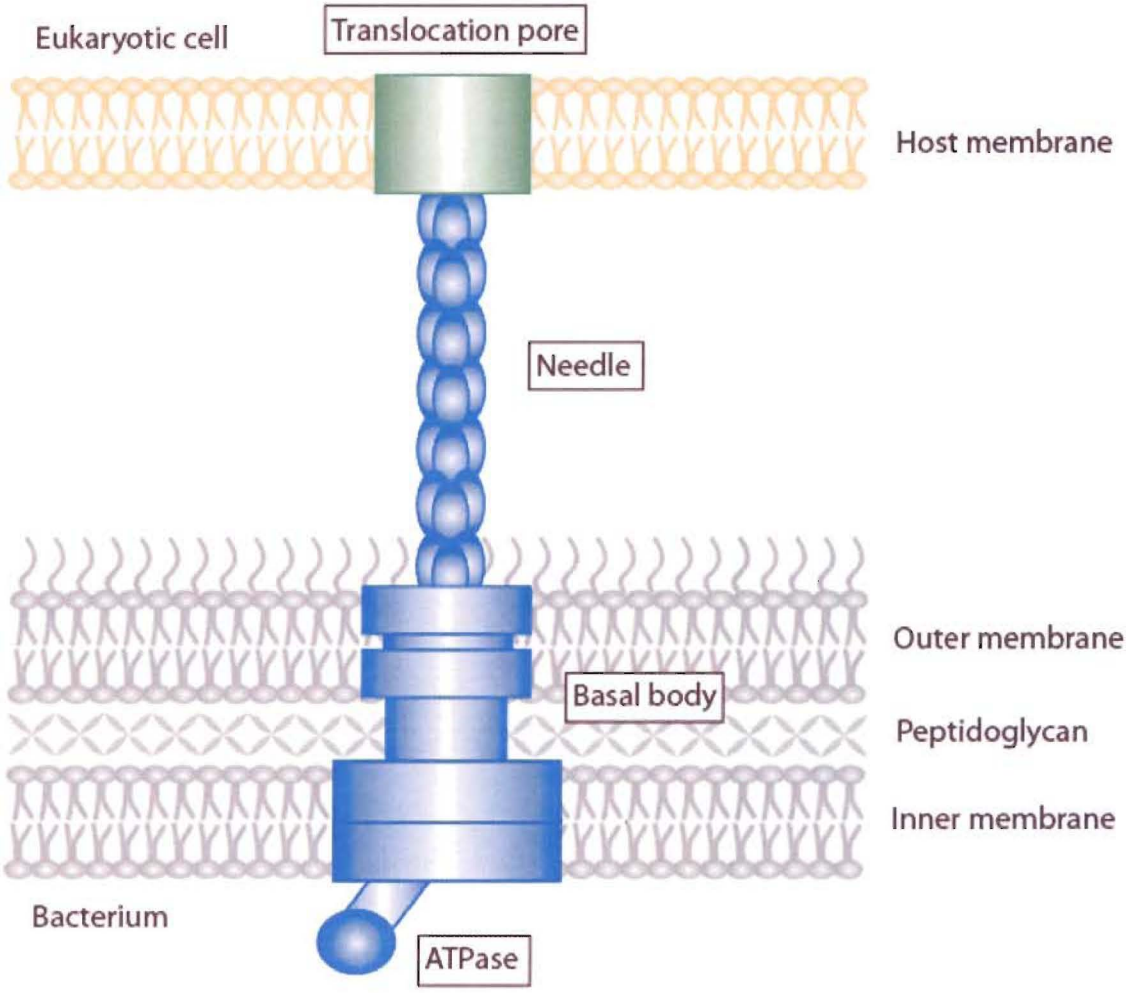
A T3SS is a secretion apparatus utilized by many pathogenic bacteria to inject VFs, called "effectors", directly into the cytosol of eukaryotic host cells (Galan et al. 1999). These effectors can mimic or interfere with host cellular signaling pathways, which is of some benefit for the pathogen. T3SSs have been discovered in diverse Gram-negative pathogens of plants and animals (Hueck 1998), as well as commensal and symbiotic bacteria where they serve to promote invasion or establish mutualistic association with their hosts (Dale et al. 2001; Dale et al. 2002; Pallen et al. 2007). Seven different families of Type III secretion (T3S) injectisomes have been identified to date based on phylogenetic analysis of their conserved proteins (Cornelis 2006). T3SS are distributed among bacteria through mechanisms of horizontal gene transfer as they are often associated with pathogenicity islands (Groisman et al. 1996).

A typical T3SS consists of a needle-like structure, the injectisome, and a translocation pore (Figure 1.1). The injectisome needle has been described as a "molecular syringe", where effector proteins are injected directly into the

cytoplasm of the target cell in one step. Secretion of certain effector proteins are assisted by cytosolic chaperones (Mota et al. 2005). The injectisome needle protrudes from the bacterial cell and is anchored to a complex of proteins forming the basal body, a series of rings spanning the inner and outer membranes. Several basal body proteins are homologous to flagella, suggesting an evolutionary relationship between the two systems (Saier 2004). Finally at the tip of the injectisome needle is a translocation pore that is inserted into the target cell membrane, allowing VFs, or "effectors" to be injected directly into the cytosol of the target cell.

Figure 1.1 Structure of a T3SS.

A typical T3SS is composed of an injectisome and a translocation pore. The injectisome consists of a needle-like structure anchored to a complex of proteins forming the basal body. At the tip of the injectisome is the translocation pore, which penetrates the host membrane allowing effectors to be directly secreted into the cytoplasm.



The Ysc-Yop T3SS in *Yersinia pestis*, the causative agent of the black plague, was the first system to be identified in the late 1980's. Researchers noticed that when *Yersinia* are incubated in low calcium concentrations, an array of proteins were secreted, referred to as Yops, or *Yersinia* outer proteins (Heesemann et al. 1986). These Yops were later identified as the virulence-associated effector proteins that function by blocking phagocytosis and the host pro-inflammatory response, and therefore allowing survival and rapid dissemination of invading *Yersinia,* and for the extreme pathogenicity and fatality rate associated with *Yersinia* infection (Cornelis 2002b).

The Ysc-Yop secretion system is now the most well studied T3SS, with homologous systems also present in *Vibrio* spp., *P. aeruginosa*, *Photorhabdus*, *Bordetella* and *Aeromonas* spp. (He et al. 2004). It encompasses more than 20 genes which make up the Ysc (Yop secretion) injectisome and effector Yops. To date, six Yop effectors have been identified: YopH, YopE, YopT, YpkA/YopO, YopP/YopJ and YopM (Cornelis 2002a). Three genes (*lcrV*, *yopB*, and *yopD*) form the translocation pore and are required for effective translocation of effector proteins (Marenne et al. 2003; Neyt et al. 1999). Genes that encode parts of the basal body include: YscC encoding secretin (the outer ring component), YscN an ATPase, YscR-V (proteins in the basal body in contact with the cytoplasmic membrane), YscJ and YscQ. Three genes (TyeA, LcrG, YopN/LcrE) have been previously shown to play a regulatory role by blocking the secretion channel in the absence of host cell contact (Cornelis et al. 2000), Finally, YopK/YopQ regulates the size of the pore in the target membrane (Cornelis et al. 1998).

Type IV secretion systems

T4SS are versatile systems used by a variety of bacteria. They can be used for either conjugal transfer, i.e. to mediate the exchange of DNA between cells, or similar to T3SSs, they can secrete VFs directly into the cytoplasm of the target cell (Cascales et al. 2003). One of the most well studied systems is the VirB/D4 system in *Agrobacterium tumefaciens* (Christie 1997). It is composed of at least 12 proteins, VirB1-11 and VirD4, which mediates the transfer of T-DNA and effector proteins (VirE2, VirE3, VirF, and VirD5) into host plant cells (Backert et al. 2006). There are currently two working models for substrate translocation: the "Channel model" and the "Piston model" (reviewed in (Cascales et al. 2003)). T4SSs differ from T3SSs in that substrates may be first translocated into the periplasm either by using the coupling protein (VirD4), the general secretory pathway, or another pathway, and then subsequently secreted across the outer membrane via the T4SS (Cascales et al. 2003). In T3SS, however, effector translocation often occurs in a single step ((Mota et al. 2005); see above section on T3SSs). The VF substrates function by disrupting normal cellular growth, leading to production of tumors and crown gall disease. Similar T4SSs have been described in *B. pertussis, Helicobacter pylori, Brucella* spp., *Bartonella* spp., and *Legionella pneumophila* (Cascales et al. 2003).

1.4.7 Regulation of virulence factors

Bacteria possess multiple mechanisms to regulate the production of VFs (for a review, see (Cotter et al. 2000)). Certain VFs may be required at particular stages of infection, and so bacteria must possess mechanisms for coordinating

expression of genes at certain times. Some regulatory systems function in response to environmental stimuli: such as a change in pH, nutrient availability, or population density. Quorum sensing is one example of a regulatory mechanism that allows the bacteria to detect the density of the surrounding population, and respond by regulating the expression of various genes. One such example is the *las* and *rhl* systems in *P. aeruginosa,* which regulates the production of multiple VFs and is critical for development of biofilms (Davies et al. 1998; Whiteley et al. 1999).

1.5 Genomic islands and virulence

Genomic islands (GIs) are defined as clusters of genes of potential horizontal origin in a prokaryotic genome. They are commonly associated with genes that provide some adaptive advantage for a microbe's particular lifestyle (Hacker et al. 2000; Hentschel et al. 2001). GIs were first disovered in uropathogenic *Escherichia coli* in late 1980s as large, unstable regions containing virulence-associated genes, and hence coining of the term "pathogenicity island" (PAI) (Hacker et al. 1990; Knapp et al. 1986). Since then, PAIs have been shown to encode numerous virulence-associated genes that are important for the pathogen to survive and cause disease in its host: for example, toxins, iron uptake, adhesions, and T3SSs (Boyd et al. 2002; Dobrindt et al. 2004; Finlay et al. 1997; Gal-Mor et al. 2006; Groisman et al. 1996; Hacker et al. 1997; Hacker et al. 2000; Schmidt et al. 2004; Wilson et al. 2003).

In addition to PAIs, several other similar elements with horizontal origins exist that encode functions other than virulence that allow the microbe to adapt to

and explore new niches. For instance, genes involved in nitrogen fixation and interaction with plant hosts are encoded on "symbiosis islands" in *Rhizobiaceae* spp., (Sullivan et al. 1998). Collectively, these genetic elements are referred to as GIs. Although GIs seem to be identified more frequently in Gram-negative bacteria, they are also present in Gram-positives and share similar structural features of a typical GI discussed below (Dobrindt et al. 2004; Hacker et al. 2000).

All types of GIs are commonly associated with particular structural features that distinguish them from the rest of the genome (for a review, see (Hacker et al. 1997)). These features typically include the following: 1) sporadic distribution among closely related species or strains; 2) abnormal G+C content and codon usage compared to the rest of the genome; 3) associated with tRNA loci; 4) large in size, ranging from 10-200kb; 5) flanked by direct repeats; 6) associated with mobility genes; and 7) relatively unstable regions that are spontaneously excised from the chromosome. A GI usually contains at least one or a combination of these features suggesting the region has horizontal origins.

Although PAIs have been noted on several occasions for their association with VFs, no analysis has yet been reported that examines whether this trend is systematically true across diverse lineages of pathogens, Such an analysis is now possible, as we have access to high quality predictions of GIs (discussed more in section 1.8.3) and high quality datasets of VFs (section 1.8.1). Furthermore, a more global analysis quantifying the function of genes encoded in islands can provide important insights into the evolution of pathogens.

1.6 Vaccines

The term vaccination was first used by Edward Jenner in 1796 when he performed his now classic experiments involving the use of cowpox to immunize against smallpox (Jenner 1800). Early advances in vaccine development were also strongly influenced by the work of Louis Pasteur. In his early research of chicken cholera, Pasteur showed that inoculation of the inactivated pathogen, later known as *P. multocida,* into chickens rendered them immune to future infection (Pasteur 1880). This principle of complete inactivation conferring disease resistance permitted the development of many vaccines. For example, some vaccines consist of inactivated toxins (toxoids), such as diphtheria and tetanus vaccines (Plotkin 2005).

Early vaccine design generally involved conventional laboratory approaches, where individual antigens were identified and tested for their ability to induce an immune response. This process can require a significant amount of labour and time, and cannot be easily applied to organisms that are difficult to grow in the lab. Technological advances in molecular biology and sequencing of complete genomes facilitated the development of variety of large-scale laboratory and bioinformatics approaches to identify novel drug targets; *in-vivo* expression technology, signature-tagged mutagenesis, and comparative genomics are only a few examples of available technology (sections 1.7.2 and 1.8.2). These approaches can simultaneously screen entire genomes for hundreds of potential candidates.

There is renewed interest in capitalizing on VFs that are pathogen-specific as potential vaccine candidates (Russmann 2004). Additionally, VFs are also of interest as they may be targets for antimicrobials or "anti-virulence" drugs, where the microbe is essentially "disarmed" and evolves towards a less virulent form (Gandon et al. 2003). It is also becoming clear that the traditional approach of broad-spectrum antimicrobial development can select for antimicrobial resistance. Therefore, there is a growing interest in complementing more conventional drug discovery approaches with new approaches such as utilizing pathogen-specific mechanisms and anti-virulence-based approaches.

1.7 Laboratory-based identification of virulence factors

Several experimental approaches are now used to discover VFs. Such approaches are increasingly being implemented on a larger, genome-wide scale. The following section provides a brief introduction to low-throughput techniques used for VF discovery, as well as a closer look at increasingly popular high-throughput approaches like signature-tagged mutagenesis (STM). Limitations associated with these approaches are also discussed.

1.7.1 Low-throughput approaches

Cloning

One method involves cloning a virulence gene from a known pathogen into an otherwise non-pathogen, which then expresses the virulence phenotype. For example, *E. coli* K12, a non-pathogenic laboratory strain, does not typically invade tissue culture cells. However, cloning of the *inv* locus of *Yersinia*

pseudotuberculosis into *E. coli* was sufficient for *E. coli* to invade cultured HEp-2 cells (Isberg et al. 1987). However, this approach is limited because not all cloned genes can be easily expressed in an avirulent host.

Transposon mutagenesis

In transposon mutagenesis, a given VF gene is inactivated by a transposon insertion leading to reduced virulence in the mutant bacteria compared to wild-type. Such experiments can be performed either *in vitro* or *in vivo*. For example, mutants harbouring a transposon insertion mutation can exhibit attenuated virulence in a mouse infection model (Portnoy et al. 1988), or decreased survival in macrophages (Fields et al. 1986). However, there are two limitations of this method where particular genes can be falsely identified as VFs. Firstly, transposon insertion can sometimes result in polar effects of downstream genes, and second, the transposon insertion may disrupt genes that are essential for growth resulting in "auxotrophic" mutants. These limitations are described in more detail in the limitations of laboratory based approaches section below (section 1.7.3).

Transcriptional fusions

Transcriptional fusion is a method commonly used to identify VFs based on their regulatory properties (Salyers et al. 2002). In one variation of this approach, the target VF gene is fused to the gene, like *phoA*, encoding an easily detectable enzyme (alkaline phosphatase, in this example). PhoA has a unique property in that its activity is localization dependent, i.e. it is active only when

localized to the membrane or is secreted. One can then use this technique to screen for surface or secreted VF mutants with attenuated virulence *in vivo*. This technique has been used to identify a toxin co-regulated pilus subunit, *tcpA,* in *V. cholera* that is necessary for colonization in mice (Taylor et al. 1987).

1.7.2 High-throughput approaches

A variety of laboratory methods are used that enable high-throughput, parallel screening for VFs. Some of the more common methods used include *in vivo* expression technology (IVET), STM, and DNA microarrays. In this section, I will describe a brief overview of these methods with particular focus on STM. I will also discuss some recent technical adaptations and advances in STM technology that have enhanced the overall versatility of this technique.

Signature tagged mutagenesis

STM is a negative selection method that identifies particular mutants with attenuated virulence *in vivo* in a mixed mutant population. This is possible because each mutant harbours its own unique DNA signature that allows for identification of a particular VF gene. Hensel and colleagues were the first to use STM to identify VFs in *S. typhimurium* in a murine model of typhoid fever (Hensel et al. 1995). In this study, the signature tags are linked to transposons, and consist of a unique 40 bp variable region flanked by 20 bp invariable regions (used for amplification and labelling of tags by PCR). Transposition of signature tags into bacteria leads to a mixed pool of mutants referred to as the input pool. Following infection of the mice with the input pool, the input pool is compared to

the output pool (mutants recovered after infection), and those mutants absent from the output pool but present in the input pool are varients that have a mutation in a gene that likely contributes virulence under these particular infection conditions.

Since its initial use, STM has become increasingly popular and has been used to identify over 1900 virulence and colonization factors in almost all major human pathogenic bacteria (Saenz et al. 2005). Many innovative technical adaptations have been described including comparative-based STM studies, where *P. aeruginosa* mutants were used to infect both wild type and genetically manipulated mice deficient in SP-A (surfactant protein A). Two mutants exhibited differential clearance in these mice, indicating these genes may be involved in resistance to SP-A mediated clearance in the mouse lung (Zhang et al. 2005). STM is also extremely versatile, and in addition to bacteria, has been applied to yeast, fungi, viruses, parasites, and mammalian cells (Mazurkiewicz et al. 2006).

In vivo expression technology

IVET as the name implies, identifies potential genes that must be expressed for survival of the bacteria in an appropriate *in vivo* infection model. This technique was first used by Mahan *et al.,* to identify *S. typhimurium* genes expressed in a mouse infection model (Mahan et al. 1993). This approach is based on the premise that *purA,* a purine biosynthesis gene, is essential for *S. typhimurium* mutants to infect mice. In this study, they first used DNA fragments from the *Salmonella* genome, and then ligated them to a promoterless *purA-lacZ* gene combination. A fraction of the DNA fragments should hybridize such that a

potential gene of interest and its associated promoter would ensure expression of *purA-lacZ* genes. These fragments are then randomly inserted into a mutant *Salmonella* strain that lacks a functional *purA* gene, and are subsequently used to infect the mice. Any surviving strains isolated after infection must therefore contain an active *purA* gene. These surviving strains are then plated and *lacZ* expression is detected. Those colonies with positive LacZ expression *in vivo,* but not *in vitro*, likely contain a DNA fragment required for *in vivo* infection and are subsequently isolated for further study.

DNA Microarrays

Comparative microarray based approaches is another technique widely used to identify virulence genes (Behr et al. 1999; Champion et al. 2005; Hotopp et al. 2006; Snyder et al. 2006; Stabler et al. 2005). It is based on the principle that genomic DNA unique to the pathogenic reference strain and absent from closely related non-pathogens, likely contain genes associated with virulence. One study compared virulent *M. tuberculosis* H37Rv with attenuated *M. bovis* BCG (Bacillus Calmette-Guerin) strains (Behr et al. 1999). BCG strains are used as live attenuated vaccines against *M. tuberculosis* infection. However, it is not clear why these strains are attenuated. In this study, they found one region encoding 9 open reading frames (ORFs) present in H37Rv but absent from all tested BCG strains suggesting that deletion of this region in BCG strains may be responsible for attenuated virulence. In addition to comparative analyses, several methods exist utilizing microarray technology to identify VF genes, such as

detection of genes expressed during *in vivo* infection for example (Boyce et al. 2004).

1.7.3 Strengths and limitations of laboratory based approaches

An advantage of low-throughput gene knockout experiments is that a VF gene can be directly targeted for inactivation and tested on an individual basis. Data from such low-throughput experiments is thought to be of the highest quality, and such methods are the most reliable for identification of a given VF gene. An advantage of high-througput approaches such as STM, IVET, and identification *in vivo* expressed genes through DNA microarrays, is that because they are screening for genes necessary for the bacteria to survive in a particular host, potential VF genes are identified regardless of their function or role in disease. An additional advantage it that many potential VFs are screened and/or identified in parallel and hence these methods provide a rapid and in theory, a more comprehensive set of VFs involved in infection of a particular host. Additionally, with STM, a pool of individual mutants are screened simultaneously in a single host and thus significantly reducing the cost and number of animals involved in such laboratory experiments.

A disadvantage of the approaches that utilize transposon mutagenesis, such as STM, is that in certain circumstances, genes can be falsely identified as VFs. For example, bacteria with mutations in genes that are essential for growth and survival, will likely exhibit attenuated virulence *in vivo* and therefore be falsely identified. However, usually such auxotrophic mutants can be detected early if tested for normal growth *in vitro*. Additionally, transposition in a gene may

affect expression of downstream genes leading to polar effects. In other words, if we consider two genes, gene A and B, gene A containing a transposon mutation that has been identified as the VF, and gene B being downstream of gene A. Gene B may in fact be the true VF, but transposon insertion upstream has affected is normal function, leading to false identification of gene A. However, these polar effects can be detected if all genes downstream of gene A (in the same coding direction) are individually screened for attenuated virulence. In other words, if individual knock-out mutants of genes downstream of gene A do not exhibit reduced virulence compared to wild-type, but a knock-out of gene A does, then it can be reasonably assumed that gene A is in fact the true VF.

IVET technology has two additional limitations: the first is that only genes that are highly expressed during infection can be easily detected. If a gene is not highly expressed, then it may not produce sufficient quantities of PurA required for the pathogen to survive. Secondly, not all genes identified are essential for infection, and when mutated and tested individually, do not show attenuated virulence on their own possibly reflecting the co-operative and complex nature of virulence.

Although several methods have been described that reduce the number of falsely identified VFs in these screens, they are not always performed in practice. It is especially difficult to test each individual gene in the high-throughput screens, simply because hundreds of potential genes are identified in parallel. In these cases, usually tests for data quality are either not performed at all or only for a small subset of genes, which can be misleading for downstream analyses.

Another limitation of laboratory approaches in general is that some pathogens are difficult to grow under normal laboratory conditions. For example, one metagenomic study revealed that many unculturable pathogens appear to be involved in gum disease (Pennisi 2004). With the continuing increase of data from metagenomic studies, it can be expected that an increasingly notable share of the genomic data will come from unculturable organisms that cannot be easily studied in the lab, making common molecular laboratory techniques impractical or impossible to perform.

Finally, the laboratory methods discussed here all involve a significant amount of labour and cost, especially if animals are involved in the experiment. Therefore, at minimum we can significantly reduce the amount of time and resources used, by complementing these laboratory approaches with additional bioinformatics and computational approaches.

1.8 Bioinformatics analysis of virulence

1.8.1 Virulence factor databases

There are four databases currently available that contain information about VFs (Table 1.4): 1) Virulence Factor Database (VFDB), 2) MvirDB from the Lawrence Livermore National Laboratory, 3) PRINTS database of virulence factors, and 4) Toxin and Virulence Factor database (TvFac).

VFDB (www.mgc.ac.cn/VFs/; (Chen et al. 2005)) is specifically focused on information about bacterial VFs and contains high quality, manually curated data. The VFDB currently contains 402 VFs, 24 PAIs, and over 2345 VF related genes

from 24 different pathogen genera, including the most well known medically important pathogens (statistics acquired in October 2007). Each VF entry is accompanied by relevant primary literature articles, detailed descriptions of their function and pathogenic mechanism, structural features, and links to protein sequence information through NCBI. The VFDB also provides an intuitive classification scheme that divides the VFs into categories based on their function. Each VF is classified as "Offensive", "Defensive", "Nonspecific", or "Regulatory", as well as additional sub-classifications like "Type III secretion system", or "Toxin" for example (Appendix A). The VFDB has also recently incorporated additional comparative genomics tools, such as the comparison of VF homologs in multiple strains of the same genera to facilitate further comparative pathogenomic studies.

MvirDB (http://mvirdb.llnl.gov/; (Zhou et al. 2007)) is the most comprehensive of the databases. It combines protein, sequence, and annotation data from 8 different publicly available databases including the following: 1) Tox-Prot subset of toxins from the Swiss-Prot protein database (Jungo et al. 2005), 2) SCORPION database of scorpion toxins (Srinivasan et al. 2002), 3) PRINTS database of virulence factors (subset of the PRINTS protein fingerprint database (Attwood et al. 2003)), 4) VFDB (Chen et al. 2005), 5) TVFac Toxin and Virulence Factor database (unpublished), 6) Islander database of genomic islands (Mantri et al. 2004), 7) ARGO database of antibiotic resistance genes (Scaria et al. 2005), and 8) VIDA database of animal viruses (Alba et al. 2001). The database contains a total of 9095 genes from 1220 organisms and provides

a user friendly web interface for easily browsing and searching a gene of interest. In addition to VFs, MvirDB also contains information on antibiotic resistance genes, genomic islands, viral proteins, and proteins from other organisms, such as scorpion toxins. Although MvirDB provides the most comprehensive datasource, it is not specifically focused on bacterial VFs. Furthermore, the data is not manually checked for quality, making it difficult for more intricate, qualitative studies of bacterial virulence.

PRINTS database of virulence factors (www.jenner.ac.uk/BacBix3/PPprints.htm) provides a simple list of bacterial and non-bacterial VFs on a single webpage. VFs are classified into one of 8 categories: "Adherence/Colonization factors", "Invasions", "Cell surface factors", "Exotoxins", "Transporters", "Siderophores", "Miscellaneous", and "Non-bacterial virulence factors". The database currently contains over 170 VFs from 38 organisms and provides links to the protein fingerprint data through PRINTS (Attwood et al. 2003) and sequence information for each.

TvFac (www.tvfac.lanl.gov) from the Los Alamos National Laboratory contains data on bacterial VFs and phage-related genes. VFs are classified into one of 13 different functional categories and the user can browse or search through these categories through the web interface. Information about a given gene's role in virulence is limited and in an unstructured free-text format. The database contains 278 VFs and VF homologs from 15 different bacterial pathogens. Links to protein, sequence, BLAST reports, and additional functional

classification tools such as COG (Clusters of Orthologous Groups; (Tatusov et al. 1997)), are provided with each TvFac entry.

- Of the four available databases, the VFDB seems to contain the most high-quality dataset of VFs and related genes, based on a manual review of each database. The VFDB has the biggest focus on manually curated data. Additionally, their hierarchical VF classification schemes provides both a structured means to functionally classify VFs and a valuable resource for further bioinformatics analysis of trends in virulence, such as the characterization of VFs in GIs (Chapter 2), and analysis of pathogen-associated genes (Chapter 3).

One common limitation of the databases listed above is that they do not contain relevant information that appropriately reflects the contextual nature of a VF. They are mainly focused on providing lists of VFs and do not provide structured, contextual information about the experimental conditions under which the genes in this database play a role in virulence. I therefore propose that new databases are needed that are centered around a given experiment that demonstrates that a gene is involved in virulence under particular conditions, such as the development of the Virulence Gene Experiment Database (VGEDB; Chapter 6). Only through the development of a database that contains such detailed information can we really start to address fundamental questions regarding more sophisticated patterns associated with virulence in pathogens.

Table 1.4 Overview of the four virulence factor databases developed to date

Database Name	Website	No. of virulence factor entries	No. of organisms	Published	Comments
BACBIX/ PRINTS database of virulence factors	www.jenner.ac.uk/ BacBix3/PPprints.htm	170	38	No	Simple list of VFs on one webpage. Links to PRINTS protein fingerprint data and sequence information.
TVFac (Los Alamos National Laboratory Toxins and Virulence Factor Database)	www.tvfac.lanl.gov	278	15	No	Entries include genes predicted to be VFs, based on sequence similarity. Information about a gene's role in virulence is limited and in free-text format difficult to analyze. Focus is on providing gene and protein sequence information.
VFDB: The Virulence Factor Database	www.mgc.ac.cn/VFs/	402	24	Yes, by Chen et al (Chen et al. 2005) in January 2005	Focus is on listing genes that are VFs. Information about their role in virulence is in a free-text, unstructured format that is difficult to analyze.
MvirDB	http://mvirdb.llnl.gov/	9059	1220	Yes, by Zhou et al (Zhou et al. 2007) in November 2006	Most comprehensive of the four databases. Contains toxins, VFs and antibiotic resistance genes from 8 different publicly available databases, but data not analyzed for quality.

1.8.2 Computational identification of virulence factors

The currently available methods for computationally identifying VFs in a genomic sequence are limited, and there is not one generally adopted approach used. This is likely due to the considerable complexity of VFs, and the various roles they play in pathogenesis for different pathogens.

One common method for detecting VFs in genomic sequence involves a BLAST similarity search to identify homologs of well-known VFs (usually a dataset of known VFs from a single closely related pathogen species). This approach has been previously used to identify a conserved T3SS in the *Chlamydia pneumonia* genome (Kalman et al. 1999). Additonally, similarity search has been combined with multiple other bioinformatics methods: for example, subcellular localization prediction, presence of particular motifs associated with VFs, and identification of sequence features associated with antigenically varying proteins (Pizza et al. 2000; Wizemann et al. 2001).

Another method involves comparative-based genomics, where basically a pathogen genome is compared to closely related non-pathogens, to identify genes or regions that are unique to the pathogen and hence likely contribute to virulence (Tang et al. 1998; Whittam et al. 2002; Zhang et al. 2006). This is based on the fact that virulence genes are often found clustered together corresponding to putative pathogenicity islands. Similar to the first method, these comparative studies use either a BLAST or similar search tool to detect islands specific to the pathogen. In these cases, genes in putative islands are identified by their lack of a significant homolog in the non-pathogenic strain (Huynen et al. 1997; Perna et al. 2001). In one such study, they compared the genomes of *E. coli* O157:H7, a food-borne pathogenic strain isolated in Sakai, to the non-pathogenic *E. coli* K12 strain. They reported 1632 genes specific to pathogenic O157:H7, of which 131 were associated with virulence. These virulence-associated genes included many well-known VFs, such as Intimin and genes

encoding a T3SS located on the LEE (Locus of Enterocyte Effacement) pathogenicity island (Hayashi et al. 2001).

Despite the fact that the methods mentioned above have been successful in identifying known VFs, none have been fully developed into tools that can be re-used or are generally applicable across multiple, diverse genera of pathogens. There is not one method that is generally accepted for identifying VFs in a newly sequenced genome. Genome centres such as the Sanger Centre and The Institute for Genomic Research (TIGR) do not have any standard methodology for the identification or annotation of VFs and their results are largely based on more general protein functional categories (H. Tettelin, TIGR, personal communication with F.S.L. Brinkman; Julian Parkhill, Sanger Centre, personal communication with F.S.L. Brinkman). A more global analysis of genes that are significantly pathogen-associated across multiple genera, such as described in Chapter 3, is one approach that could be used to identify particular types of genes that may be more virulence-associated.

Furthermore, to my knowledge, there has been no report to date that measures the accuracy of these approaches. Such ad-hoc approaches threaten to undermine the confidence of informatics to identify candidate VFs in newly sequenced genomes. As the amount of genomic data continues to increase exponentially, from metagenomic studies for example, propagation of errors in accuracy will inevitably increase as well. I therefore propose that more investigation of the accuracy of these approaches is needed (Chapter 7), and

suggest improvements can be made through the development of high quality VF datasets, such as the VGEDB (described in Chapter 6).

1.8.3 Computational prediction of genomic islands

As mentioned above, genomic islands (GIs) have been anecdotally noted to disproportionately contain VFs (termed pathogenicity islands) and so GI prediction is an important focus of most efforts to analyze the sequence of pathogen genomes (reviewed in (Dobrindt et al. 2004; Finlay et al. 1997; Hentschel et al. 2001; Schmidt et al. 2004)). There are two types of methods commonly used for predicting GIs in genomic sequence: comparative genomics-based approaches and sequence composition based approaches (Langille et al. 2008).

Comparative genomics approaches are in essence similar to those used to identify VFs (section 1.8.2). The basic concept is that the genomic content of one or more closely related genomes is compared, and clusters of genes unique to one strain and absent from the other(s) likely correspond to putative GIs. Some of the most popular methods available to compare genomic content are through homology search tools, such as BLAST, or through DNA microarray technology. Some examples of GIs detected with these methods are discussed in sections 1.7.2 and 1.8.2. Currently, there is no published software tool available that automatically predicts GIs based on a comparative analysis. This is likely due to the difficulties involved in automatically choosing an appropriate reference genome for comparison.

Sequence composition based approaches are based on the premise that phylogenetically related strains exhibit similar sequence composition that constitutes their own genetic signature. Therefore, if a segment of the genome has a signature that is different from the rest of the genome, this region may have been horizontally acquired. Some of the more common measures of sequence compositional bias are the measure of G+C content (%G+C), or dinucleotide bias. Several software tools have been developed that use these, or slight variations of these measurements to predict GIs (Hsiao et al. 2003; Merkl 2004; Yoon et al. 2005; Zhang et al. 2004). Additionally, a combination of atypical composition with other structural features commonly associated with GIs (described in section 1.5), such as mobility genes (transposases or integrases, for example), is also used (Hsiao et al. 2005).

A disadvantage of comparative approaches is that at minimum two genomes are required in the comparison, and often choosing an appropriate reference genome can be quite difficult. For sequence compositional approaches, however, no reference strain is required. On the other hand, one advantage of comparative methods is that both recent and ancient GIs can be detected depending on the genomes chosen for comparison. However, for sequence-based methods, in some cases ancient GIs may not be easily detected as they usually evolve overtime (ameliorate) to have a similar sequence composition as the rest of the genome.

The accuracy of these methods to computationally predict GIs is difficult to assess, since (as with all evolutonary features) GIs are inferred and so there are

no definitively true positive or true negative GI datasets. Despite this, there are datasets of very probable GIs, based on manual literature research, that have been well characterized and these have been used to test the accuracy of GI predictors (Hsiao et al. 2005). In addition, recent research in the Brinkman Lab has led to the development of datasets of probable GIs and non-GIs based on comparative genomics analysis, which can then be used to evaluated sequence composition based GI predictors (since the two GI identification methods are independent; (Langille et al. unpublished)). Through these datasets, all available GI prediction methods have recently had their accuracy tested. IslandPath-DINUC (Hsiao et al. 2005) is the method that was found to have the highest overall accuracy (Langille et al. unpublished). For this method, GIs are defined as 8 or more ORFs with dinucleotide bias. Alternatively, for the more specific/precise analysis, GIs are defined as 8 or more ORFs with dinucleotide bias plus the presence of one or more mobility genes (IslandPath-DIMOB method). The availability of the GI predictions from these methods permits further robust analyses of virulence trends, such as confirming anecdotal reports that VFs are indeed associated with GIs (Chapter 2). Also, a more global analysis quantifying the function of genes encoded in islands can also provide important insights into the evolution of pathogens (Chapter 2).

1.9 Goal of the present research

At the onset of my project, there was an increasing appreciation that virulence is a much more complex phenomenon than previously thought. As more bacterial genomes became available, it was observed that many of the

classic VFs, many of which were thought to be solely in pathogens, were also found in non-pathogenic bacteria. In addition, there was increased interest in the study of GIs (regions with potential horizontal origins), as it was frequently observed that VFs are often associated with these islands. However, these trends, and others, have not yet been quantified on a large scale encompassing multiple pathogens from diverse genera.

In this study, I used datasets of VFs from the most well known medically important human pathogens as well as datasets of computationally predicted GIs, and confirmed that in fact VFs are associated with these regions (Chapter 2). In addition, in a large-scale global analysis, I identified and investigated types of genes that are solely pathogen-associated (i.e. genes that are only found in pathogens and not found in non-pathogens), and in multiple pathogen genera, with the hypothesis that their role in the disease process is more virulence-specific (Chapter 3). I found that particular types of VFs, such as toxins and those involved in secretion of VFs are more pathogen-associated, and discuss their current and potential use in vaccine development (Chapter 4 and Chapter 5). In addition, to initiate an attempt to deal with the complex nature of virulence, I have developed the VGEDB, a resource that incorporates detailed information about experimental conditions used to identify a given VF (Chapter 6). This database differs from other VF databases in that it contains more contextual information relevant to virulence conditions, and a given database entry is centered around a virulence gene experiment, rather than a virulence gene. This database can potentially enable more sophisticated analyses of virulence, and in

this study I use it to investigate the accuracy of common BLAST-based methodology for the identification of VFs in sequenced genomes (Chapter 7). Overall, in my thesis work, I confirm and quantify previous anecdotally reported observations in virulence, as well as provide more fundamental insights regarding pathogenesis and virulence-associated genes that could aid in development of new vaccines and therapeutics.

CHAPTER 2
ESTIMATING THE PREVALENCE OF VIRULENCE FACTORS AND PATHOGEN-ASSOCIATED GENES INSIDE AND OUTSIDE OF GENOMIC ISLANDS

2.1 Introduction

With the number and diversity of bacterial genomes sequenced, we are now able to investigate selected anecdotally reported observations in pathogenicity and quantify them on a more global scale. For example, it has been frequently noted that many virulence genes are associated with genomic islands (GIs; clusters of genes of probable horizontal origin) (Boyd et al. 2002; Dobrindt et al. 2004; Finlay et al. 1997; Groisman et al. 1996; Hentschel et al. 2001; Ochman et al. 2001; Schmidt et al. 2004; Shankar et al. 2002). In fact, the first GIs identified harboured genes involved in virulence, and hence were called "pathogenicity islands" (PAIs) (Hacker et al. 1990). Since then many others have frequently noted this apparent association (reviewed in (Boyd et al. 2002; Dobrindt et al. 2004; Finlay et al. 1997; Groisman et al. 1996; Hacker et al. 1997; Hacker et al. 2000; Ochman et al. 2001; Pallen et al. 2007; Schmidt et al. 2004)). However, no analysis has yet been reported that examines whether this trend is systematically true across diverse lineages of pathogens, using a method for predicting GIs that is suitably accurate.

Such an analysis is now possible, as we have developed methods for high quality predictions of GIs that have had their accuracy tested, plus we have

access to additional datasets of known GIs (Hsiao et al. 2005; Langille et al. unpublished). There is also a curated dataset of VFs available through the VFDB (Chen et al. 2005) which may be cross-referenced with current bacterial genome datasets. Finally, availability of VF classification schemes, as well as my own method for computationally identifying pathogen-associated VFs, allows us to further characterize features of VFs that may be associated with GIs.

For this study, I used datasets of known VFs from the VFDB and predicted GI datasets to quantify the occurance of VFs in GIs and non-GIs. Consistent with previous anecdotal reports, I found that GIs do contain a significantly higher proportion of VFs and pathogen-associated genes (genes found only in pathogens; Chapter 3) compared to non-GIs ($p < 2.20E-16$). In addition, this study provides quantitative evidence that certain types VFs are strongly associated with GIs, including VFs that are more offensive, such as T3S, and T4S system genes. The implications of these results on therapeutic development and the evolution of pathogenicity are discussed.

2.2 Materials and methods

I obtained a dataset of 1819 VFs (28 well-known pathogens) from the VFDB in 2005. The other available VF databases, such as the PRINTS database (www.jenner.ac.uk/BacBix3/PPprints.htm), TvFac (www.tvfac.lanl.gov/), and MvirDB (Zhou et al. 2007), (www.tvfac.lanl.gov/) were examined as well, but the curated VFDB was found to be of the highest quality (see Table 1.4 and section 1.8.1). In collaboration with William Hsiao (Brinkman Lab), we quantified the occurrence of VFs in GIs. We used a subset of 1227 VFs from 26 pathogens

from the VFDB (representing 18 species and 15 genera) that had complete genome sequences available and GIs predicted through IslandPath. Additionally, we quantified the occurrence of both pathogen-associated genes (genes found predominately in pathogens; described in Chapter 3) and "common" genes (genes found in both pathogens and non-pathogens) in GIs.

To prevent circular logic where known PAIs are defined by the presence of VFs and VFs are therefore found predominately in PAIs, GIs were defined based on attributes that are independent of their VF gene content. I used two GI prediction methods previously used for other analyses of GIs that were determined to be effective methods for identifying GIs on a high-throughput scale, and can be uniformly applied to all the pathogens studied (Hsiao et al. 2005). For my first analysis, a GI was defined as a region consisting of 8 or more ORFs with dinucleotide bias (DINUC dataset; calculated as the frequency of dinucleotides in a cluster of ORFs compared to the entire genome) as predicted by IslandPath (Hsiao et al. 2003). This GI prediction method is noted for having higher sensitivity. I also used a more stringent definition of a GI that requires the GIs to contain both dinucleotide bias and one or more mobility genes (DIMOB dataset), as this method is more precise/specific (Hsiao et al. 2005). Note that since there are many more genes in general outside of GIs, than in GIs, for any genome, it is important to examine proportions of VF genes inside and outside islands, as a function of the total number of genes inside and outside of such GI regions.

I obtained the VFDB classification scheme along with the VF dataset from the VFDB. Classification of VFs as "offensive", "defensive", "regulation" and "nonspecific", were retrieved through the VFDB website and is shown in Appendix A.

I calculated all statistics for over-representation of VFs in GIs by first tabulating the number of VFs in GIs, total GIs, number of VFs in non-GIs, and total non-GIs in a 2x2 contingency table and then using Chi-squared test with Yates' correction for continuity (correction used on 2x2 contingency tables where there is only one degree of freedom). For those categories with small values (< 5), the Fisher's Exact Test was used instead. Similar statistical analyses were done for functional classification of genes in islands, where the number of genes in each VFDB category was used in the calculation. Since multiple categories are examined in parallel, the Benjamini and Hochberg False Discovery Rate correction for multiple testing was performed for all functional category analyses. I considered p-values smaller than 0.05 to be significant. All statistics were performed using the R statistics package.

2.3 Results

2.3.1 Genomic islands disproportionately contain more virulence factors and pathogen-associated genes

Consistent with previous anecdotal reports, this analysis indicated that GIs, as predicted using the IslandPath-DINUC method, do contain a significantly higher proportion of VFs compared to non-GIs ($p < 2.20E-16$; Table 2.1). On average for all pathogens studied, 4.5% of genes in predicted islands encode

VFs, compared with 1.2% of genes outside of islands. This method of GI prediction used has the highest overall accuracy of any GI prediction method currently available (Hsiao et al. 2005; Langille et al. unpublished). I obtained similar results using a more stringent definition of a GI that requires the GIs to contain both dinucleotide bias and one or more mobility genes (IslandPath-DIMOB dataset; Table 2.2). This method is more precise/specific (Hsiao et al. 2005; Langille et al. unpublished). Regardless of which approach was used there was clearly a significant bias in terms of more VFs being located in such predicted GI regions.

Table 2.1 Proportions of VFs in GIs vs. outside of GIs – DINUC dataset (more sensitive method)

VF Dataset	GIs[a]		Outside of GIs		*p*-value[b]
	Number of VFs/Total number of genes in GIs[c]	Proportion of genes in GIs that are VFs (%)	Number of VFs/Total number of genes in non-GIs[c]	Proportion of genes in non-GIs that are VFs (%)	
VFDB	443/9801	4.5	784/67690	1.2	< 2.20E-16*
Pathogen-associated VFs[d]	157/9301	1.7	163/63783	0.3	< 2.20E-16*
"Common" VFs[e]	286/9357	3.1	621/64055	1.0	< 2.20E-16*

[a] GIs are defined as 8 or more consecutive ORFs with dinucleotide bias as predicted with IslandPath (DINUC dataset).

[b] Pearson's Chi-squared test with Yates' continuity correction. Asterisks indicate statistical significance (p-value < 0.05).

[c] Total number of genes in GIs varies according to the number of genomes used that contain pathogen-associated, "Common", or both types of VFs.

[d] Pathogen-associated VFs have homologs only in other pathogen genomes, at the similarity cut-off used.

[e] "Common" VFs have homologs in both pathogens and non-pathogens, at the similarity cutoff used.

Table 2.2 Proportions of VFs in GIs vs. outside of GIs – DIMOB dataset (more specific method).

VF Dataset	GIs[a]		Outside of GIs		*p*-value[b]
	Number of VFs/Total number of genes in GIs[c]	Proportion of genes in GIs that are VFs (%)	Number of VFs/Total number of genes in non-GIs[c]	Proportion of genes in non-GIs that are VFs (%)	
VFDB	203/3395	6.0	1024/74096	1.4	< 2.20E-16*
Pathogen-associated VFs[d]	65/3287	2.0	255/69797	0.4	< 2.20E-16*
"Common" VFs[e]	138/3240	4.3	769/70172	1.1	< 2.20E-16*

[a] GIs are defined as 8 or more consecutive ORFs with dinucleotide bias plus one or more mobility genes as predicted by IslandPath (DIMOB dataset).

[b] Pearson's Chi-squared test with Yates' continuity correction. Asterisks indicate statistical significance (p-value < 0.05).

[c] Total number of genes in GIs varies according to the number of genomes used that contain pathogen-associated, "Common", or both types of VFs.

[d] Pathogen-associated VFs have homologs only in other pathogen genomes, at the similarity cut-off used.

[e] "Common" VFs have homologs in both pathogens and non-pathogens, at the similarity cutoff used.

I also investigated the relationship between GIs and pathogen-associated VFs, defined as VFs found predominately or only in pathogens (see Chapter 3). I quantified the occurrence, in GIs, of pathogen-associated VFs and "common" VFs (the latter are found in both pathogens and non-pathogens) and found that regardless of the GI prediction criteria used (DINUC or DIMOB), both pathogen-associated and "common" VFs are present in higher proportions in GIs than non-GIs (DINUC dataset Table 2.1; DIMOB dataset Table 2.2).

In addition to the above analyses involving datasets of known VFs, I also examined trends in the distribution of VF homologs in an expanded dataset of all completely sequenced bacterial genomes available as of February 2006 (total

267 genomes) so that I could investigate this trend globally, and in non-pathogens as well (in particular host-associated non-pathogens that contain VF homologs). While this homology-based approach is not as robust as the above analysis involving known VFs, VF homologs in pathogens were found to be slightly more common inside islands versus non-islands (17.7% and 14.7% respectively; p-value < 2.20E-16). However, notably, VF homologs in non-pathogens (limited to non-pathogens that are associated with a host by commensal or mutualistic associations), are equally found in islands and non-islands (14.3% and 14.2% of island and non-island genes, respectively). Such VF homologs in non-pathogens likely comprise genes not involved in virulence, per se, but rather involved in "host interactions", while the VF homologs in pathogens will include a mix of both host interaction factors as well as virulence genes more directly involved in pathogenicity.

I also extended the above analysis of pathogen-associated VFs in GIs by using the expanded genome dataset of all completely sequenced bacterial genomes available as of February 2006 (total 267 genomes). Each gene in a given genome was identified as pathogen-associated, non-pathogen-associated (found predominately or only in non-pathogens), or "common", according to our approach (see Chapter 3). I found that both pathogen-associated and non-pathogen-associated genes occur more frequently in GIs than non-GIs ($p < 2.20E-16$ for both). This supports previous observations that species or family-specific genes tend to be more commonly found in GIs reflecting a proposed large, novel gene pool that is associated with GIs (Hsiao et al. 2005). Still, this

does not detract from the earlier observation that VFs in general, including those common to both pathogens and non-pathogens, are clearly disproportionately associated with GIs.

2.3.2 Genomic islands disproportionately contain offensive virulence factors, such as Type III secretion genes

To study whether specific types of VFs are more likely to be associated with such probable horizontally transferred regions, I divided the VFs into classes based on the VFDB classification scheme and examined the functional categories of VFs in GIs versus non-GIs (with statistical corrections for multiple testing). I found that genes over-represented in GIs are classified as T3SS and T4SS - including their corresponding secreted effector proteins, as well as toxins, proteases, adherence factors, iron uptake, antiphagocytosis factors, and "Unclassified" genes (DINUC dataset; Table 2.3), where the "Unclassified" class mostly contains VF-associated genes that have not been functionally characterized according to the VFDB classification scheme. These results are consistent with previous reports that T3SS and T4SS genes are closely associated with PAIs (Hacker et al. 2000). With the more precise/specific IslandPath-DIMOB based GI detection method, the T3SS and T4SS genes are not more significantly associated with GIs (Table 2.4). However, it should be noted that such secretion systems may not have the types of mobile genes near them that the DIMOB-based method detects.

It is also notable that, regardless of the GI detection method used, VFs classified as "offensive" by the VFDB (i.e. VFs involved in active invasion of the

host) are very significantly associated with islands ($p < 2.20E-16$) while most of the "defensive" VFs (involved in passive defense/evasion of the host) have no preferential association with GIs (DINUC dataset Table 2.3; DIMOB dataset Table 2.4). There are no classes of VFs, according to the VFDB classification system, which are more prevalent outside of GIs at a statistically significant level.

Table 2.3 VFDB classification of VFs in GIs and non-GIs (DINUC dataset)

VFDB Classification[a]	GIs		non-GIs		*p*-value[b]
	VFs (#)	Proportion of genes in GIs(%)	VFs (#)	Proportion of genes in non-GIs (%)	
Type III secretion system[c] (O)	61	0.62	109	0.16	2.42E-15*
Type IV secretion system[d] (O)	35	0.36	15	0.02	3.63E-15*
Unclassified (NA)	185	1.89	158	0.23	7.26E-15*
Adherence (O)	59	0.60	138	0.20	4.69E-12*
Iron uptake (NS)	33	0.34	59	0.09	3.85E-10*
Antiphagocytosis (D)	23	0.23	66	0.10	1.84E-03*
Toxin (O)	18	0.18	53	0.08	9.65E-03*
Protease (D)	5	0.05	5	0.01	9.82E-03*
Type II secretion system (O)	6	0.06	15	0.02	2.27E-01
Magnesium uptake (NS)	1	0.01	0	0.00	4.17E-01
Invasion (O)	2	0.02	4	0.01	5.09E-01
Actin-based motility (O)	1	0.01	1	0.00	6.52E-01
IgA1 Protease (D)	1	0.01	2	0.00	8.47E-01
Mangenese uptake (NA)	0	0.00	1	0.00	1
Heat-shock protein (NA)	0	0.00	1	0.00	1
Complement resistance (NA)	0	0.00	1	0.00	1
Anti-proteolysis (D)	0	0.00	1	0.00	1
Plasminogen activator (NA)	0	0.00	3	0.00	1
Serum resistance (D)	0	0.00	3	0.00	1
Proinflammatory effect (NA)	0	0.00	2	0.00	1
Pigment (NA)	0	0.00	2	0.00	1
Immune evasion (NA)	0	0.00	2	0.00	1
Cellular metabolism (D)	0	0.00	9	0.01	1
Biosurfactant (NA)	0	0.00	2	0.00	1
Enzyme (NS)	0	0.00	8	0.01	1
Complement protease (D)	0	0.00	2	0.00	1
Cell wall (NA)	1	0.01	6	0.01	1
Motility (O)	3	0.03	31	0.05	1
Molecular mimicry (NA)	0	0.00	4	0.01	1
Endotoxin (NA)	3	0.03	29	0.04	1
Stress protein (D)	1	0.01	11	0.02	1

VFDB Classification[a]	GIs		non-GIs		*p*-value[b]
	VFs (#)	Proportion of genes in GIs(%)	VFs (#)	Proportion of genes in non-GIs (%)	
Exoenzyme (NS)	2	0.02	17	0.03	1
Regulation (R)	3	0.03	24	0.04	1
TOTAL	443		784		

[a] VFs are defined as those genes curated as being VFs according to the VFDB. Only those VFs in the VFDB where GI predictions were available from IslandPath were included in the analysis. O = Offensive; D = Defensive; NS = Nonspecific; R = Regulation; NA = Not available.

[b] Pearson's Chi-squared test with Yates' continuity correction. Asterisks indicate statistical significance (p-value < 0.05).

[c] Includes Type III secretion system genes and Type III translocated proteins

[d] Includes Type IV secretion system genes and Type IV secretory proteins.

Table 2.4 VFDB classification of VFs in GIs and non-GIs (DIMOB dataset)

VFDB Classification[a]	GIs		non-GIs		*p*-value[b]
	VFs (#)	Proportion of genes in GIs(%)	VFs (#)	Proportion of genes in non-GIs (%)	
Unclassified (NA)	120	3.7	223	0.3	3.63E-15*
Adherence (O)	46	1.4	151	0.2	7.26E-15*
Toxin (O)	12	0.4	59	0.1	1.25E-05*
Protease (D)	4	0.1	6	0.0	5.15E-03*
Iron uptake (NS)	10	0.3	82	0.1	3.50E-02*
Actin-based motility (O)	1	0.0	1	0.0	4.71E-01
Type IV secretion[d] (O)	0	0.0	50	0.1	7.15E-01
Endotoxin (NA)	3	0.1	29	0.0	7.71E-01
Type III secretion[c] (O)	5	0.1	164	0.2	1
Type II secretion (O)	0	0.0	21	0.0	1
Stress protein (D)	0	0.0	12	0.0	1
Serum resistance (D)	0	0.0	3	0.0	1
Regulation (R)	1	0.0	26	0.0	1
Proinflammatory effect (NA)	0	0.0	2	0.0	1
Plasminogen activator (NA)	0	0.0	3	0.0	1
Pigment (NA)	0	0.0	2	0.0	1
Molecular mimicry (NA)	0	0.0	4	0.0	1
Mangenese uptake (NA)	0	0.0	1	0.0	1
Magnesium uptake (NS)	0	0.0	1	0.0	1
Invasion (O)	0	0.0	6	0.0	1
Motility (O)	0	0.0	34	0.0	1
Immune evasion (NA)	0	0.0	2	0.0	1
IgA1 Protease (D)	0	0.0	3	0.0	1
Heat-shock protein (NA)	0	0.0	1	0.0	1
Exoenzyme (NS)	0	0.0	19	0.0	1
Enzyme (NS)	0	0.0	8	0.0	1
Complement resistance (NA)	0	0.0	1	0.0	1
Complement protease (D)	0	0.0	2	0.0	1
Cellular metabolism (D)	0	0.0	9	0.0	1
Cell wall (NA)	0	0.0	7	0.0	1
Antiphagocytosis (D)	4	0.1	85	0.1	1

Biosurfactant (NA)	0	0.0	2	0.0	1
Anti-proteolysis (D)	0	0.0	1	0.0	1
Total	40		646		

[a] VFs are defined as those genes curated as being VFs according to the VFDB. Only those VFs in the VFDB where GI predictions were available from IslandPath were included in the analysis. O = Offensive; D = Defensive; NS = Nonspecific; R = Regulation; NA = Not available.

[b] Pearson's Chi-squared test with Yates' continuity correction. Asterisks indicate statistical significance (p-value < 0.05).

[c] Includes Type III secretion system genes and Type III translocated proteins.

[d] Includes Type IV secretion system genes and Type IV secretory proteins.

2.4 Discussion

These results confirm previous anecdotal reports that VFs are in fact more common in GIs than outside of GIs, which supports the important role of GIs in pathogen evolution. I also present quantitative evidence that "offensive" VFs are significantly associated with GIs (such as genes involved in T3S and T4S), as well as 'Unclassified' genes. These results are consistent with previous reports that T3SSs and T4SSs are closely associated with PAIs (Hacker et al. 2000), as well as previous reports that more novel genes are associated with GIs (Hsiao et al. 2005). Furthermore, the majority of these associations hold true regardless of whether we use a more sensitive or specific method for GI identification (with the exception of T3S and T4S genes which were found to be significantly associated with GIs according to the DINUC criteria of GI prediction but not the DIMOB criteria) . Even though our method will not detect some GIs (i.e. those with the same sequence composition) and so will tend to under-predict GIs, we never observe a statistically significant association of VFs with regions outside of GIs for any class of VFs.

Also of note, VF homologs form a higher proportion of genes in GIs for pathogens, while in host-associated non-pathogens VF homologs (commonly "host interaction factors") are notably more equally distributed in GIs versus non-GI genomic regions. These observations suggest that pathogenicity, as opposed to host interaction, is often a fairly recently developed phenomenon in a species (on an evolutionary time scale detected by GI analysis). That is, we propose that VFs that are more directly involved in virulence, with more "offensive" rather than "defensive" actions, may be more associated with GIs (and therefore more recently acquired) versus "host interaction factors" that are not pathogen-specific. These observations support proposals that pathogenicity is often a fairly recently developed phenomenon in a species and is frequently an evolutionary dead end due to the difficulty of balancing the benefits of increased virulence with the negatives associated with killing the host (Maurelli 2007).

GIs appear to provide a critical flexible mechanism for allowing increased, invasive infection of their host. Several evolutionary models have been proposed to explain how VFs are maintained on GIs, and these models are consistent with the importance of GIs in pathogen evolution. Jeff Smith (Smith 2001) proposed that in a pathogen population, there are a small number of "cheaters" that themselves do not possess certain extracellularly-acting VFs but benefit from the effect of these VFs released by the non-cheater strains. Without the VFs, the cheater strains are metabolically more fit than the non-cheaters, and therefore their number would increase in the population over time. However, cheaters, due to the lack of VFs have decreased infectiousness, and Smith proposed that

horizontal gene transfer is a possible mechanism to minimize "cheater" strains and restore infectiousness in the pathogen population. As a result, certain VFs are maintained on mobile elements, including GIs. It is worth noting that in our PSORTb study, predicted extracellular proteins are over-represented in pathogen-associated genes (see Chapter 3). In a second proposed model, Sokurenko and colleagues adopted the classical source-sink model of population genetics to describe virulence evolution (Sokurenko et al. 2006). For opportunistic pathogens, the environmental reservoir represents a self-sustainable source whereas the opportunistic infection represents a venture into a sink. They proposed that acquisition of PAIs as a mechanism to facilitate adaptation of the source to sink transition whereas the loss of PAIs accompanies the sink to source transition. However, since possessing a PAI can significantly increase the pathogen's fitness in the sink, which in term increases the back flow of PAI-possessing strains into the source population, VFs in PAIs can be maintained despite VFs negative fitness value in the environment. It is notable that, in our study, many of the over-represented VFs in GIs are involved in active invasion that harm the host in some way and there is no obvious functionality for these VFs outside of the host environment. Maintaining these VFs on GIs, and likely other horizontally acquired elements like phage and plasmids, therefore appears to provide important evolutionary flexibility for these pathogens.

CHAPTER 3
CHARACTERIZATION OF PATHOGEN-ASSOCIATED GENES

3.1 Introduction

Most bacterial VFs were originally thought to be associated only with pathogens. However, as the number of genome sequences began to increase, it became clear that many of the "classic" VFs were also encoded in the genomes of non-pathogenic, commensal bacteria (Pallen et al. 2007; Snyder et al. 2006; Zhang et al. 2003). Microarray analyses also supported this; for example, many of the known virulence associated genes in pathogenic *Neisseria* spp., were also found to be present in the closely-related non-pathogen *Neisseria lactamica* (Snyder et al. 2006). There was an increasing understanding that virulence is a much more complex phonomenon than previously thought, and reflects an interplay between the host, pathogen, and environmental factors. It was also suggested that the term "virulence factor" should be used less and instead they should be more appropriately referred to as "host interaction factors" (Holden et al. 2004). However, it is evident that certain types of genes, such as botulinum toxin, are both necessary and sufficient to cause disease on their own (Shukla et al. 2005). I therefore wished to examine to what degree there may be VFs that are so critical for disease processes that their very presence is strongly associated with disease, rather than simply host colonization/interaction.

Now that there are many genome sequences available from both pathogenic and non-pathogenic strains of diverse bacterial genera, I investigated the degree in which there are classes of genes that may be pathogen-specific or notably pathogen-associated. Previous analyses of pathogen-specific genes have been limited to certain species or genera (for example, (Anisimova et al. 2007; Champion et al. 2005; Dozois et al. 2003; Hotopp et al. 2006; Snyder et al. 2006; Stabler et al. 2005)), but a more global analysis is now possible. While such an analysis is still limited by the scope of bacterial genome sequences and VFs currently available, any VFs observed to be present in pathogenic strains from diverse bacterial genera, with no detectable homologs in non-pathogenic strains of the same genera, are considered good candidates for being classified as pathogen-associated. I set out to examine whether such genes could be identified within a diverse bacterial genome dataset, and examined common features of such genes, with the hypothesis that they may play a more virulence-specific role in pathogens. Such genes also represent targets for possible novel therapeutic strategies that interfere with pathogen-specific traits as previously shown before (Hung et al. 2005; Russmann 2004).

In this study, I used whole genome datasets from diverse pathogens and non-pathogens, to identify genes that are pathogen-specific or significantly pathogen-associated, and then used various functional classification tools to examine common features associated with such genes. I found that pathogen-associated VFs are disproportionately "offensive", such as toxins and T3S and T4S systems. This suggests that these types of genes may serve more

virulence-specific roles in pathogens. Such genes also warrant further investigation as they may represent targets for possible novel therapeutics.

3.2 Materials and methods

Each VF from the VFDB dataset (described in section 2.2) was identified as pathogen-associated (found predominately in pathogens), or "common" (found in both pathogens and non-pathogens) through a BLAST similarity search against the deduced proteomes of 166 pathogenic and 101 non-pathogenic sequenced prokaryotic genomes downloaded from the National Center for Biotechnology Information (NCBI) FTP site in February 2006. An e-value cut-off of 10^{-7} was used to exclude distant homologs. In an initial investigation, I examined more and less stringent cut-offs of 10^{-12} and 10^{-5}, and found that the vast majority of trends analyzed still hold when these other cut-offs were examined.

Pathogen, non-pathogen, or host-associated status for each genome was obtained through TIGRs Microbial Genome Properties table (Haft et al. 2005) (some manual curation for data quality and overall completeness was performed on this dataset). I also identified each gene in the 267 sequenced genomes as pathogen-associated, "common", or non-pathogen-associated (genes found predominately in non-pathogens), in a similar manner as described above. Complete lists of the pathogen-associated and "Common" genes identified for each genome are available at the following website: http://www.pathogenomics.sfu.ca/pathogen-associated/index.html. For the purposes of these analyses, the term 'pathogen-associated' versus pathogen-

specific is used to refer to those genes that may potentially have homologs in nonpathogens, but the genome sequences of these particular nonpathogens are not yet available and hence the genes was identified in pathogens only according to this BLAST analysis. Conversely, pathogen-specific genes denote those 'true virulence genes' that are specific to pathogens and absent from nonpathogens. Additionally, to reduce redundancy and bias in this whole genome dataset (multiple genome sequences from a particular genera or species), this analysis was repeated using a subset of genomes with a minimum evolutionary distance (substitutions/site) of 0.05 (based on phylogenetic analysis by (Ciccarelli et al. 2006)). For the analysis of pathogen-associated and "common" genes in multiple genera, a gene is defined as in multiple genera if it was found in a minimum of 2 different genera according to the cut-off used. Genus information for each organism was obtained from the NCBI taxonomy database (Wheeler et al. 2000). Using genus information, I was able to identify genes found only in pathogens and in multiple genera that had no homologs (according to our similarity cutoff) in non-pathogens of the same genera.

VFDB functional classification of VFs as well as "offensive", "defensive", etc, classifications are described in section 2.2. COG (Clusters of Orthologous Groups) (Tatusov et al. 1997) assignments for each VF and predicted gene product from each complete genome were performed using RPS-BLAST (Reverse Position Specific Iterative-BLAST) against the COG database (obtained from NCBI FTP); and the top BLAST hit below an e-value of 0.01 was chosen (Altschul et al. 1990). If no hit was found with an e-value below 0.01, the gene

was labelled 'Unclassified'. For those genes that had multiple COG classifications, each COG would be counted to reflect the multiplicity nature of COG classification. PSORTb version 2 (Gardy et al. 2005) was used to predict protein subcellular localization for all VFs and complete deduced proteomes used in this analysis.

Over- or under-representation of particular functional classifications (VFDB classification, COG, and PSORTb subcellular localization) of pathogen-associated and "common" VFs from the VFDB were calculated by comparing the number of pathogen-associated genes in a given category against "common" genes in the same category. Statistics were calculated using the Chi-squared Test with Yates' correction with corrections for multiple testing (described in section 2.2). For statistics involving all sequenced genomes, I looked for over or under-representation of a given functional category by first calculating the percent of pathogen-associated or "common" genes in a given category for a particular genome and then calculated the two-tailed paired *t*-test (pathogen-associated and "common") using the values across all organisms.

3.3 Results

3.3.1 Pathogen-associated genes are disproportionately offensive virulence factors, such as toxins and Type III and Type IV secretion systems

VFs in the VFDB were identified as either pathogen-associated or "common" to both pathogens and non-pathogens, and then used selected protein functional classification tools to determine the distribution of functional classes for

each. VFs were first classified into 33 different virulence-related categories using the VFDB classification scheme (described in section 2.2). I found that pathogen-associated VFs are disproportionately toxins and involved in T3S and T4S (Table 3.1). Conversely, "common" VFs are disproportionately involved in 'Iron uptake', 'Antiphagocytosis', 'Endotoxin', 'Motility', 'Regulation', and 'Protease' (Table 3.1). The VFDB also classifies VFs as either 'offensive' or 'defensive'. VFs classified as 'offensive' were found to be significantly disproportionately pathogen-associated ($p < 2.20E-16$); while, "defensive" VFs were found to be common to both pathogens and non-pathogens ($p = 1.13E-08$).

Table 3.1 VFDB classification of pathogen-associated and "common" VFs from the VFDB

VFDB Classification	Pathogen-associated VFs		"Common" VFs		*p*-value[b]
	#	%[a]	#	%[a]	
Categories with a higher percentage of Pathogen-associated VFs					
Toxin	66	12.67	34	2.62	7.26E-15*
Type III secretion system	106	20.35	121	9.32	3.48E-09*
Type IV secretion system	32	6.14	18	1.39	5.56E-07*
Plasminogen activator	2	0.38	1	0.08	4.69E-01
Anti-proteolysis	1	0.19	0	0.00	5.91E-01
Actin-based motility	1	0.19	1	0.08	8.53E-01
Proinflammatory effect	1	0.19	1	0.08	9.00E-01
Exoenzyme	7	1.34	13	1.00	9.64E-01
Categories with a higher percentage of "Common" VFs					
Iron uptake	5	0.96	93	7.16	1.79E-06*
Antiphagocytosis	7	1.34	87	6.70	3.53E-05*
Motility	1	0.19	33	2.54	1.03E-03*
Endotoxin	1	0.19	31	2.39	2.37E-03*
Regulation	2	0.38	29	2.23	1.72E-02*
Protease	0	0.00	15	1.16	3.15E-02*
Stress protein	0	0.00	12	0.92	7.92E-02
Cell wall	0	0.00	11	0.85	1.21E-01
Cellular metabolism	0	0.00	10	0.77	1.97E-01
Enzyme	0	0.00	8	0.62	2.91E-01
Invasion	2	0.38	14	1.08	5.80E-01
Type II secretion system	4	0.77	18	1.39	6.75E-01
Molecular mimicry	0	0.00	4	0.31	8.75E-01
Serum resistance	0	0.00	3	0.23	8.83E-01
IgA1 Protease	0	0.00	3	0.23	9.27E-01
Adherence	77	14.78	204	15.72	9.59E-01
Magnesium uptake	0	0.00	1	0.08	1
Pigment	0	0.00	2	0.15	1
Mangenese uptake	0	0.00	1	0.08	1
Unclassified	206	39.54	522	40.22	1
Immune evasion	0	0.00	2	0.15	1

VFDB Classification	Pathogen-associated VFs		"Common" VFs		*p*-value[b]
	#	%[a]	#	%[a]	
Heat-shock protein	0	0.00	1	0.08	1
Complement resistance	0	0.00	1	0.08	1
Biosurfactant	0	0.00	2	0.15	1
Complement protease	0	0.00	2	0.15	1
TOTAL	521		1298		

[a] Based on the percentage of pathogen-associated or "Common" VFs in a given functional category.

[b] Pearson's Chi-squared test with Yates' continuity correction. Asterisks indicate statistical significance (p-value < 0.05).

I further classified pathogen-associated and "common" VFs using COG (Clusters of Orthologous Groups; (Tatusov et al. 1997)). I included an additional 'Unclassified' COG category for those genes that do not belong to any COG – a category that generally represents relatively novel genes that lack homologs between species. The most notable significant difference in the distribution of COG categories between pathogen-associated and "common" VFs was seen in this 'Unclassified' category, where pathogen-associated VFs had a significantly higher proportion of COG 'Unclassified' proteins compared to "common" VFs (82.4% and 14.1% respectively; p = 5.72E-15; Table 3.2). I further confirmed that significantly higher proportions of pathogen-associated genes belong to the 'Unclassified' class when compared to the "common" genes in the expanded dataset containing 267 bacterial genomes (p < 5.50E-15; Table 3.3). Also, pathogen-associated genes in a single genus are less well characterized than those in multiple genera – 93% and 71.9% respectively are 'Unclassified' (p < 2.20E-16, two-tailed paired *t*-test). Based on these observations I propose that

the lack of characterization of pathogen-associated genes may reflect the lack of sequenced homologs available for such genes and the limitation of our dependence on homologous annotation transfer to classify new genes.

Based on COG, I found that there is no significant difference between pathogen-associated and non-pathogen-associated genes in terms of the number of genes with no functional classification, suggesting that generic protein functional classification schemes such as COG do not provide adequate coverage for species-specific genes.

Table 3.2 COG classification of pathogen-associated and "common" VFs (VFDB dataset)

COG Functional Category	Pathogen-associated VFs		"Common" VFs		*p*-value[b]
	#	%[a]	#	%[a]	
Categories with a higher percentage of Pathogen-associated VFs					
Unclassified	453	82.36	225	14.10	5.72E-15*
Categories with a higher or same percentage of "Common" VFs					
Cell wall/membrane/envelope biogenesis	2	0.36	171	10.71	7.02E-07*
Inorganic ion transport and metabolism	1	0.18	71	4.45	9.82E-03*
Replication, recombination and repair	4	0.73	108	6.77	1.03E-02*
Carbohydrate transport and metabolism	0	0.00	43	2.69	2.54E-02*
Secondary metabolites biosynthesis, transport and catabolism	1	0.18	56	3.51	3.53E-02*
Posttranslational modification, protein turnover, chaperones	0	0.00	33	2.07	6.54E-02
Defense mechanisms	0	0.00	26	1.63	1.77E-01
Intracellular trafficking, secretion, and vesicular transport	29	5.27	295	18.48	1.77E-01
Energy production and conversion	0	0.00	24	1.50	1.86E-01
Amino acid transport and metabolism	1	0.18	30	1.88	4.05E-01
Cell motility	26	4.73	230	14.41	5.81E-01
Transcription	9	1.64	94	5.89	5.91E-01
Coenzyme transport and metabolism	0	0.00	11	0.69	7.05E-01
Translation, ribosomal structure and biogenesis	0	0.00	4	0.25	1
Lipid transport and metabolism	0	0.00	10	0.63	1
Signal transduction mechanisms	6	1.09	54	3.38	1
RNA processing and modification	0	0.00	0	0.00	1
Function unknown	8	1.45	42	2.63	1
Nucleotide transport and metabolism	0	0.00	1	0.06	1
Nuclear structure	0	0.00	0	0.00	1
Extracellular structures	0	0.00	2	0.13	1
Cytoskeleton	0	0.00	0	0.00	1
Chromatin structure and dynamics	0	0.00	0	0.00	1
Cell cycle control, cell division, chromosome partitioning	1	0.18	9	0.56	1
General function prediction only	9	1.64	57	3.57	1
TOTAL	550		1596		

[a] Based on the percentage of pathogen-associated or "Common" VFs in a given functional category.

[b] Pearson's Chi-squared test with Yates' continuity correction. Asterisks indicate statistical significance (p-value < 0.05)

Table 3.3 COG classification of pathogen-associated and "common" genes (expanded genome dataset – 267 complete genomes)

COG Functional Category	Pathogen-associated VFs (%)[a]	"Common" VFs (%)[a]	*p*-value[b]
Categories with a higher percentage of Pathogen-associated VFs			
Unclassified	89.3	8.5	5.50E-15*
Categories with a higher percentage of "Common" VFs			
Secondary metabolites biosynthesis, transport and catabolism	0.1	1.6	2.75E-16*
Intracellular trafficking, secretion, and vesicular transport	0.9	2.4	2.89E-16*
Translation, ribosomal structure and biogenesis	0.4	8.2	3.06E-16*
Defense mechanisms	0.1	1.5	3.24E-16*
Cell cycle control, cell division, chromosome partitioning	0.3	1.1	3.44E-16*
Signal transduction mechanisms	0.4	3.6	3.67E-16*
Replication, recombination and repair	0.7	6.4	3.93E-16*
Function unknown	2.2	7.2	4.23E-16*
Coenzyme transport and metabolism	0.2	4.1	4.58E-16*
Nucleotide transport and metabolism	0.1	2.8	5.00E-16*
Lipid transport and metabolism	0.2	3.2	5.50E-16*
Carbohydrate transport and metabolism	0.3	6.0	6.11E-16*
Cell wall/membrane/envelope biogenesis	0.7	5.2	6.88E-16*
Transcription	0.7	5.8	7.86E-16*
Posttranslational modification, protein turnover, chaperones	0.3	3.8	9.17E-16*
Energy production and conversion	0.4	5.2	1.10E-15*
Amino acid transport and metabolism	0.3	7.5	1.38E-15*
Inorganic ion transport and metabolism	0.3	4.6	1.83E-15*
General function prediction only	1.3	9.8	2.75E-15*
Cell motility	0.5	1.5	6.14E-14*
Extracellular structures	0.0	0.0	1.55E-05*
RNA processing and modification	0.0	0.0	1.15E-02*
Chromatin structure and dynamics	0.0	0.0	1.17E-01
Cytoskeleton	0.0	0.0	8.34E-01

[a] Based on the percentage of pathogen-associated or "Common" VFs in a given functional category.

[b] Two-tailed paired *t*-test. Asterisks indicate statistical significance (p-value < 0.05)

PSORTb v.2.0 (Gardy et al. 2005) was used to predict protein subcellular localization of pathogen-associated and "common" gene products. I first examined subcellular localization for VFs from the VFDB. I found that compared to "common" VFs, pathogen-associated VFs in Gram-negative bacteria have disproportionately higher "Unknown" localization ($p = 1.32E-15$; Table 3.4), and Gram-positive bacteria have disproportionately higher "Extracellular" localization ($p = 8.25E-07$; Table 3.5), where these extracellular proteins may correspond to VFs secreted by the pathogen to act on the host cell. These similar trends were also observed in our expanded genome dataset (Table 3.6 for Gram-negative genomes; Table 3.7 for Gram-positive genomes). This observation is notable as it fits well with the "cheater hypothesis" proposed by Jeff Smith (Smith 2001) (see conclusions in Chapter 2).

Table 3.4 PSORTb-predicted protein subcellular localization of pathogen-associated and "Common" VFs for Gram-negative bacteria (VFDB dataset)

Subcellular Localization	Pathogen-associated VFs		"Common" VFs		*p*-value[b]
	#	%[a]	#	%[a]	
Categories with a higher percentage of Pathogen-associated VFs					
Unknown	310	67.83	458	38.78	1.32E-15*
Extracellular	24	5.25	59	5.00	9.31E-01
Categories with a higher percentage of "Common" VFs					
CytoplasmicMembrane	19	4.16	186	15.75	1.04E-09*
OuterMembrane	14	3.06	124	10.50	3.86E-06*
Periplasmic	2	0.44	41	3.47	2.87E-04*
Cytoplasmic	88	19.26	313	26.50	3.29E-03*
TOTAL	457		1181		

[a] Based on the percentage of pathogen-associated or "Common" VFs in a given functional category.

[b] Pearson's Chi-squared test with Yates' continuity correction. Asterisks indicate statistical significance (p-value < 0.05)

Table 3.5 PSORTb-predicted protein subcellular localization of pathogen-associated and "Common" VFs for Gram-positive bacteria (VFDB dataset)

Subcellular Localization	Pathogen-associated VFs		"Common" VFs		*p*-value[b]
	#	%[a]	#	%[a]	
Categories with a higher percentage of Pathogen-associated VFs					
Extracellular	35	54.69	19	16.24	8.25E-07*
Unknown	14	21.88	22	18.80	7.64E-01
Categories with a higher percentage of Pathogen-associated VFs					
Cytoplasmic	5	7.81	31	26.50	1.22E-02*
CytoplasmicMembrane	4	6.25	20	17.09	6.95E-02
Cellwall	6	9.38	25	21.37	8.20E-02
TOTAL	64		117		

[a] Based on the percentage of pathogen-associated or "Common" VFs in a given functional category.

[b] Pearson's Chi-squared test with Yates' continuity correction. Asterisks indicate statistical significance (p-value < 0.05)

Table 3.6 PSORTb-predicted protein subcellular localization of pathogen-associated and "Common" VFs for Gram-negative bacteria (expanded genome dataset – 267 complete genomes)

Localization	Pathogen-associated (%)[a]	Common (%)[a]	*p*-value[b]
Categories with a higher percentage of Pathogen-associated VFs			
Unknown	70.80	40.11	3.30E-16*
Outer Membrane	3.24	2.41	3.42E-02*
Categories with a higher percentage of "Common" VFs			
Periplasmic	0.09	2.01	4.40E-16*
Cytoplasmic Membrane	9.25	18.53	6.60E-16*
Cytoplasmic	16.65	36.47	1.32E-15*
Extracellular	0.21	0.47	1.62E-04*

[a] Based on the percentage of pathogen-associated or "Common" VFs in a given functional category.

[b] Two-tailed paired *t*-test. Asterisks indicate statistical significance (p-value < 0.05)

Table 3.7 PSORTb-predicted protein subcellular localization of pathogen-associated and "Common" VFs for Gram-positive bacteria (expanded genome dataset – 267 complete genomes)

Localization	Pathogen-associated (%)[a]	Common (%)[a]	*p*-value[b]
Categories with a higher percentage of Pathogen-associated VFs			
Unknown	50.09	20.38	5.50E-16*
Extracellular	10.49	1.72	5.78E-12*
Cell Wall	1.07	0.90	1.06E-01
Categories with a higher percentage of "Common" VFs			
Cytoplasmic	25.29	54.75	1.10E-15*
Cytoplasmic Membrane	14.12	20.75	1.83E-09*

[a] Based on the percentage of pathogen-associated or "Common" VFs in a given functional category.

[b] Two-tailed paired *t*-test. Asterisks indicate statistical significance (p-value < 0.05)

3.3.2 Reducing sampling bias of sequenced bacterial genomes

One potential source of bias with these functional category analyses is that the taxonomical distribution of the genomes sequenced to date is uneven. In particular some pathogens are over-represented by multiple strains while certain,

predominately non-pathogenic, taxa are sparsely represented. To reduce redundancy and bias in the whole genome dataset, we selected a subset of pathogen and non-pathogen genomes with a minimum evolutionary distance (substitutions/site) of 0.05 (adapted from a comprehensive phylogenetic analysis (Ciccarelli et al. 2006)). This essentially reduced the number of pathogen genomes that were highly similar (e.g., multiple strains of a pathogen) and thus reduced sampling bias. When this less-biased genome dataset was analyzed again using the same classification schemes and methods as described above, no major differences in results were observed ruling out sampling bias as a major contributing factor to our observations (data not shown).

3.3.3 Limitations of this study

This study of pathogen-associated genes of course has several limitations. Firstly, it is limited by the number, and diversity, of genome sequences, and known VFs, currently available. However, I felt that the diversity of species whose genome sequences were available was sufficient to provide an early sense of the degree in which certain gene types were pathogen-associated since multiple well-studied pathogens, with closely related non-pathogenic relatives, had complete genomes available from diverse phyla. I also repeated these analyses using hundreds of genomes, taking into account the phylogenetic distance between species to reduce the redundancy of the genomes dataset in order to reduce potential biases due to sampling. Similar results were obtained, with the same statistically significant observations, with this pared down dataset. Regardless, clearly this analysis, or a similar type of analysis, bears repeating as

the number of genome sequences available increases. Future analyses will need to account for non-pathogens that may have recently evolved from pathogens. The contextual nature of pathogenicity (for example, how an organism can be a pathogen in one species and not in another) complicates analysis and will need to be further considered. This analysis was also limited by the cutoffs used to measure the similarity between sequences. I chose cutoffs that did not produce a notably different result from cutoffs slightly above or below it. However, any hard cutoff is not perfect and so I performed further manual inspection of results for a given gene identified as pathogen-associated before pursuing further in depth analysis of the gene of interest. It should also be taken into consideration that some proteins, like T3SS effectors, may appear to be more pathogen-associated simply because there are less constraints on their sequence and they have diverged in sequence more rapidly. However, by focusing most of this analysis on those genes that are found in multiple genera, I have been identifying genes that do share a certain degree of similarity. Finally, I also investigated the utility of different gene function classification systems in this analysis, like COG, SUPERFAMILY, PRINTS, and the VFDB. It became clear over the course of this study that general classification systems like COG do not perform well in detecting trends in virulence since the classification system does not include most VFs. The VFDB, with its curated dataset and virulence-guided classification system, was the most effective. There are still some VFDB classifications that could benefit from more curation – for example the T3SS component classification could be improved further – but this more virulence-specific VFDB

classification was of the most utility. More effort should be made to build upon such efforts and develop a high quality ontology that is relevant to virulence, to complement other ontology efforts.

Even with all of the limitations in this analysis described above, including genome sequences available, VFs known, and classification systems available, the criterion used clearly identifies genes and gene categories that have a notable pathogen-association.

3.4 Discussion

Through the identification and analysis of pathogen-associated VFs, I found that toxins, T3S, and T4S system genes may be disproportionately associated with pathogens, suggesting that these types of genes may serve more specific roles in pathogenesis. Furthermore, I show that pathogen-associated genes are not well classified with common, general protein function classification systems. I therefore propose there is an overall need to improve coverage of current classification systems, because even programs with high precision and recall, such as PSORTb (Gardy et al. 2005), have significantly lower predictive capabilities for species-specific genes as shown above.

This investigation of putative pathogen-associated genes reveals several universal strategies adopted by pathogens that can be used to gain access to and colonize privileged sites in hosts. These strategies appear to be absent in non-pathogenic strains which typically do not colonize privileged sites and therefore do not elicit a strong inflammatory response (Brown et al. 2006).

Overall, these results suggest that while we have made substantial progress towards understanding pathogenicity mechanisms, there are still many virulence-associated factors and mechanisms that we don't understand. Additionally, systematic screening for genes that are predominately or exclusively found in pathogens, such as the one carried out here, may provide an alternative strategy to identify potential VFs that more directly responsible for virulence, rather than host-interaction factors.

This study also provides the beginnings of a list of pathogen-associated genes that could be used as targets for novel therapeutic strategies that specifically target pathogen-specific traits or mechanisms. I also provide whole genome datasets of pathogen-associated, nonpathogen-associated and "Common" genes identified in this study that are available for downloading at the following website: www.pathogenomics.sfu.ca/pathogen-associated/index.html. In Chapter 4 and Chapter 5, I expand my analyses of selected classes of pathogen-associated VFs, and discuss their potential as candidates for vaccine development.

CHAPTER 4
PATHOGEN-ASSOCIATED GENES ENCODE SPECIALIZED COMPONENTS OF TYPE III SECRETION

4.1 Introduction

Many pathogenic bacteria use T3SSs to deliver VFs, called effectors, directly into the cytosol of host cells. T3SSs have also been discovered in commensals and symbiotic bacteria, as a mechanism of interacting with their hosts (Tampakaki et al. 2004). A more detailed description of T3SSs is presented above in section 1.4.6.

In this study, I investigated and identified components of T3SSs that may be more critical for virulence in pathogens by identifying components that are predominately associated with pathogens and lack homologs with significant sequence similarity in non-pathogens. I therefore used the BLAST-based approach (described in Chapter 3) to identify pathogen-associated components of the Ysc-Yop T3SS in *Yersinia* spp. I found particular components of this system are predominately pathogen-associated, specifically, the effectors, translocation pore, and regulatory genes involved in response to host cell contact. Investigation of these components as potential vaccine candidates is also discussed.

4.2 Materials and methods

A list of genes and components of the *Yersinia* Ysc-Yop T3SS were assembled by utilizing the subset of VFs classified as "T3SS" and "Type III secreted proteins", according to the VFDB classification scheme (section 1.8.1). Additional genes and annotation information were curated from the literature for overall completeness of the dataset. T3SSs from enteropathogenic *E. coli*, *Salmonella* spp., as well as the T4SS in *Agrobacterium* were also retrieved and investigated in a similar manner. In the case of the T4SS, genes from the VFDB classified as "T4SS" and "Type IV secretory protein" were used. The components involved in the Ysc-Yop system seemed to be the most comprehensively studied and so in depth analysis was done for this system.

Each gene was labeled either pathogen-associated or "common" using the BLAST-based approach described Chapter 3. They were further manually inspected to identify any that may have been falsely labeled as pathogen-associated due to the cutoff used. For example, in some cases a gene may have a significant homolog in a non-pathogen, but the e-value score is relatively close, but slightly above, the chosen cutoff of 10-7. In these cases, the BLAST reports are further manually examined for homology over the entire protein or only over a small region or domain for example, and any questionable cases are reported. I also used a less stringent criteria, by allowing a given gene to have a maximum of 1 homolog in a non-pathogen. This would account for genes that should be labelled pathogen-associated, but were not according to the BLAST analsysis. For example, homology over only a small region or domain with a gene in a non-

pathogen, may falsely label a gene as "common" when in fact it should be pathogen-associated. Again, overall patterns in pathogen-associated components of the entire T3SS were re-analyzed and compared to original data.

4.3 Results

4.3.1 Effectors, translocation pore and genes involved in host-contact regulation are strongly pathogen-associated

According to my BLAST-based analysis, I found certain components of the Ysc-Yop T3SS tend to be strongly associated with pathogens. These include the effector proteins, genes involved in formation of the translocation pore, genes regulated by host-dependent contact, and additional external components. All pathogen-associated genes are shown in Figure 4.1.

Effector Yops have remarkable ability to evade the host immune system by blocking phagocytosis and the host pro-inflammatory response (Cornelis 2002b). Four out of 6 known Yop effectors were identified as pathogen-associated according to this BLAST analysis, including: YopE, YopH, YpkA/YopO, and YopP/YopJ, as well as the known effector chaperones SycT, SycH, SycE/YerA (chaperones for YopT, YopH, and YopE respectively). Effector YopT was identified as "common" to both pathogens and non-pathogens as it had significant similarity to a gene found in nonpathogen *Hahella chejuensis*, a marine microbe. Recent genomic sequencing of *H. chejuensis* led to the discovery of two T3SSs and other virulence-associated genes, suggesting it may be pathogenic to some marine eukaryotes (Jeong et al. 2005), although this is not confirmed. The final effector YopM was also identified as "common", however

the precise function of this protein remains unclear. The *yopM* gene was found in multiple pathogen species and the plant-symbiont *Bradyrhizobium japonicum*, in which a conserved T3SS was identified (Mazurier et al. 2006). It has been suggested, but not confirmed, that the T3SS gene cluster identified is involved in early interactions for establishing symbiosis with its host (Gottfert et al. 2001).

Formation of a pore in the host cell membrane is required for effective translocation of Yop effectors into the host cell cytosol. Pore formation requires YopB, YopD, and LcrV (Marenne et al. 2003; Neyt et al. 1999), which were all identified as pathogen-associated and present in 4 genera (*Yersinia*, *Vibrio*, *Pseudomonas*, and *Photorhabdus*).

YscW, a gene previously shown to be required for secretion of YopB, YopD, and LcrV (Allaoui et al. 1995), is pathogen-associated and found in 3 different genera. TyeA and LcrG were identified as pathogen-associated and found in the same 4 genera as above. These proteins play a role in blocking the secretion channel in the absence of host cell contact (Cornelis et al. 2000). A third gene with similar function, *yopN/lcrE*, is "common" and found in 6 different genera including the non-pathogens *Hahella chejuensis* and *Desulfovibrio vulgaris* Hildenborough, a sulfate-reducing bacterium, whose genome also encodes essential T3SS genes (Heidelberg et al. 2004). Finally, *yopK/yopQ* which regulates the size of the pore in the target membrane (Cornelis et al. 1998) is pathogen-associated and in a single genus.

Genes which encode parts of the external injectisome, including *yscO* and *yscP* (functions as a ruler, regulating needle length), components whose

products are required for Yop secretion, and YscX, another secreted component whose function is not well understood, are all pathogen-associated and found in 3 to 4 genera. YscF, the major subunit of the needle complex, is found in 6 different pathogen species (of the generas *Pseudomonas*, *Vibrio*, *Photorhabdus*, and *Yersinia*) and one insect endosymbiont, *Sodalis glossinidius* str 'morsitans'.

It was recently observed that *S. glossinidius* does contain genes homologous to T3SS genes in *Yersinia* and *Salmonella*. Several lines of evidence suggest that adaptation of *S. glossinidius* to a symbiotic lifestyle from free-living is fairly recent, and massive genome erosion has removed or inactivated certain T3SS components and effectors that are likely to harm the host (Dale et al. 2005; Toh et al. 2006). This provides an interesting example where T3SS has adapted to mutualistic interaction. This also reflects the need for methods that identify pathogen-associated genes to allow for such recent evolutionary events.

Similar pathogen-associated genes/components were observed for the enteropathogenic *E. coli* T3SS encoded by the LEE PAI (Figure 4.2), the *Salmonella* SPI-1 (Figure 4.3), as well as the *A. tumefaciens* T4SS (Figure 4.4). The majority of external or secreted componts in these systems were identified as pathogen-associated as well and in notably less genera than "common" genes. Additionally, when the more or less stringent criteria (see materials and methods) for identifying pathogen-associated genes were re-examined, I found the overall trend that these particular components are still strongly pathogen-associated similar to the original dataset.

Figure 4.1 Pathogen-associated and "Common" genes involved in *Yersinia* Ysc-Yop T3SS.

Genes that are pathogen-associated are shown in red and genes "common" to pathogens and non-pathogens are in black. The numbers in parentheses represent the number of different genera this gene is found in according to the BLAST analysis.

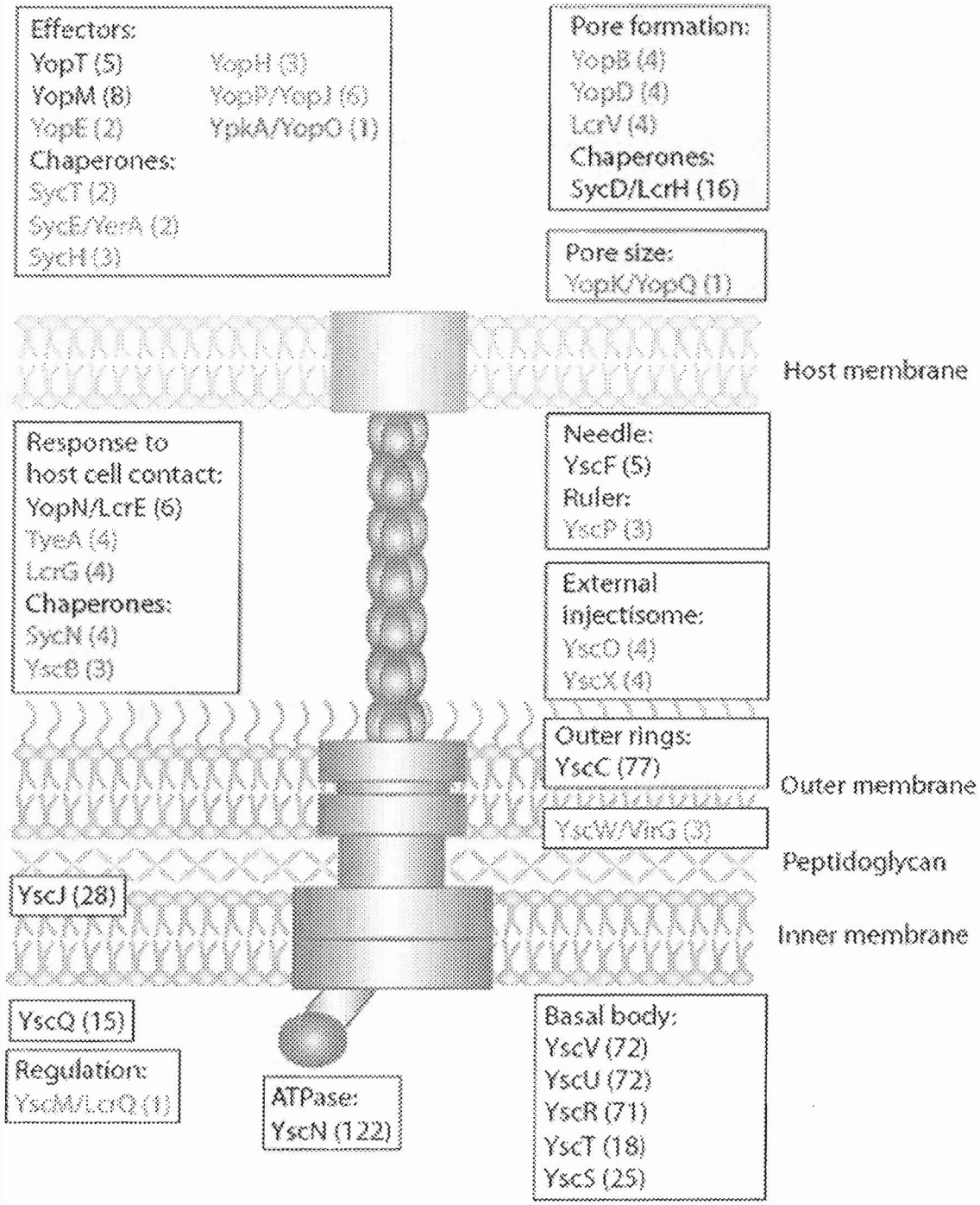


Figure 4.2 Pathogen-associated and "Common" genes involved in the enteropathogenic *E. coli* T3SS.

Genes that are pathogen-associated are shown in red and genes "common" to pathogens and non-pathogens are in black. The numbers in parentheses represent the number of different genera this gene is found in according to the BLAST analysis. Those genes with homologs in a maximum on 1 nonpathogen were classified as pathogen-associated.

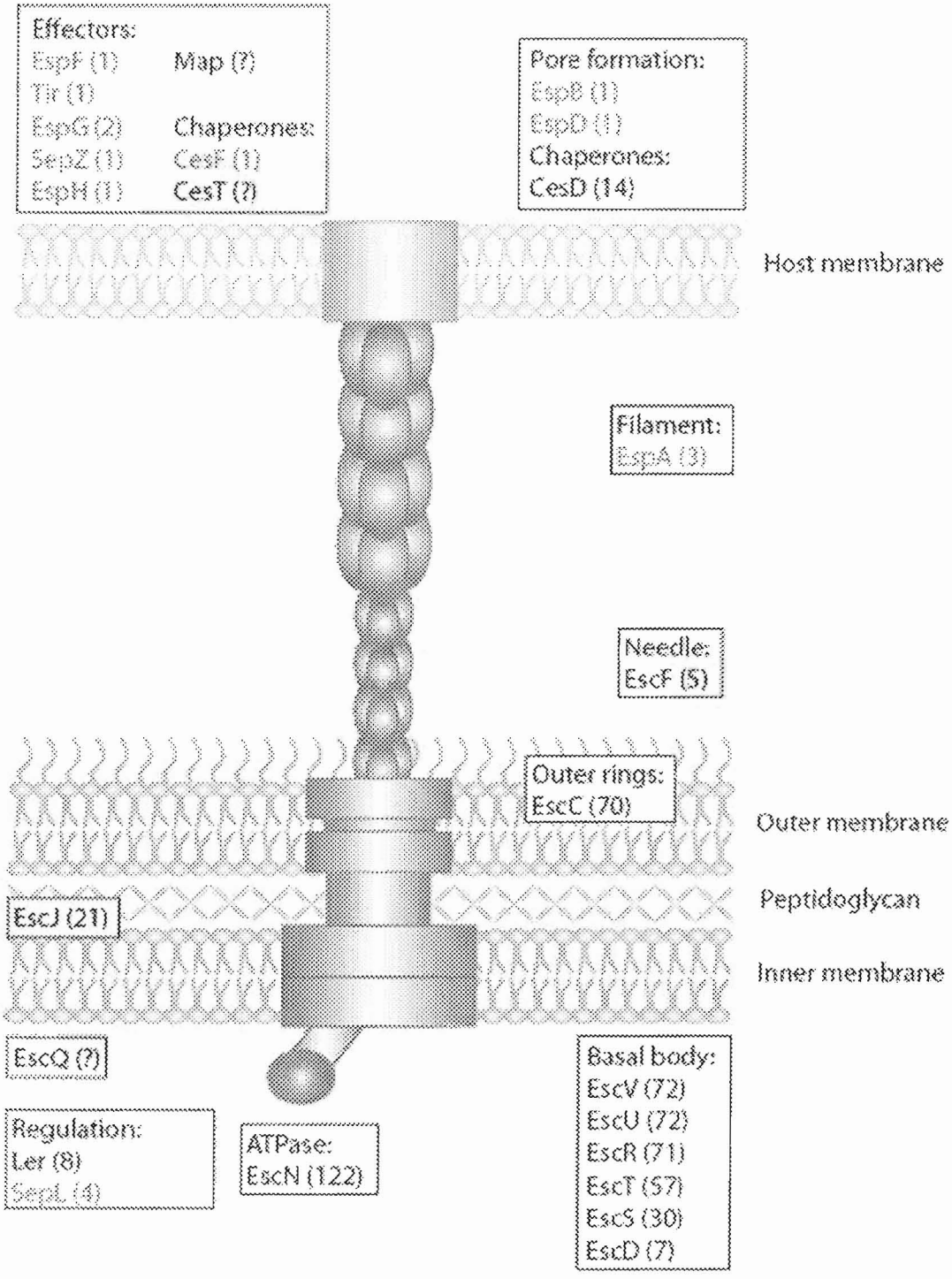


Figure 4.3 Pathogen-associated and "Common" genes involved in the *Salmonella* SPI-1 T3SS.

Genes that are pathogen-associated are shown in red and genes "common" to pathogens and non-pathogens are in black. The numbers in parentheses represent the number of different genera this gene is found in according to the BLAST analysis. Those genes with homologs in a maximum on 1 nonpathogen were classified as pathogen-associated.

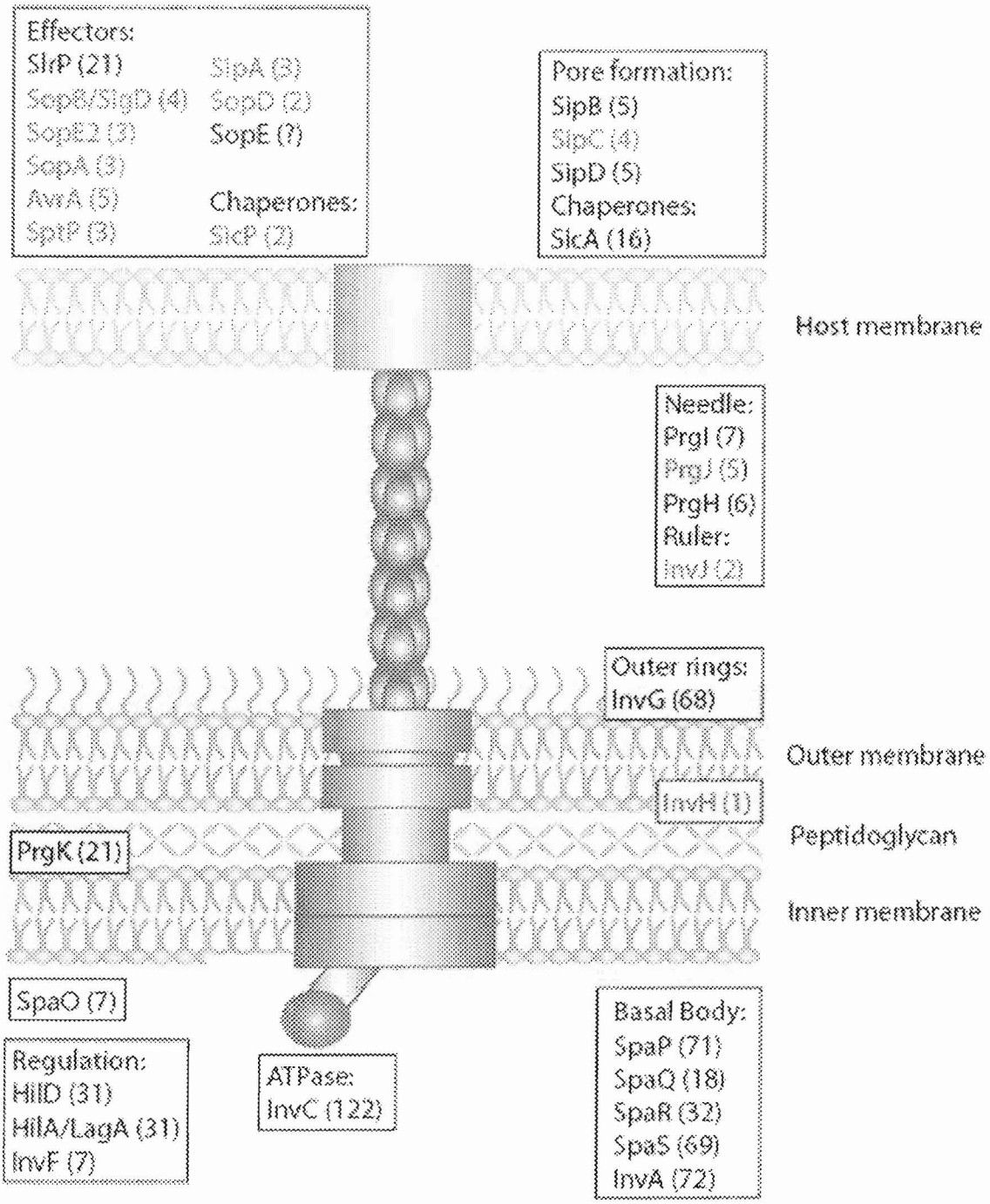
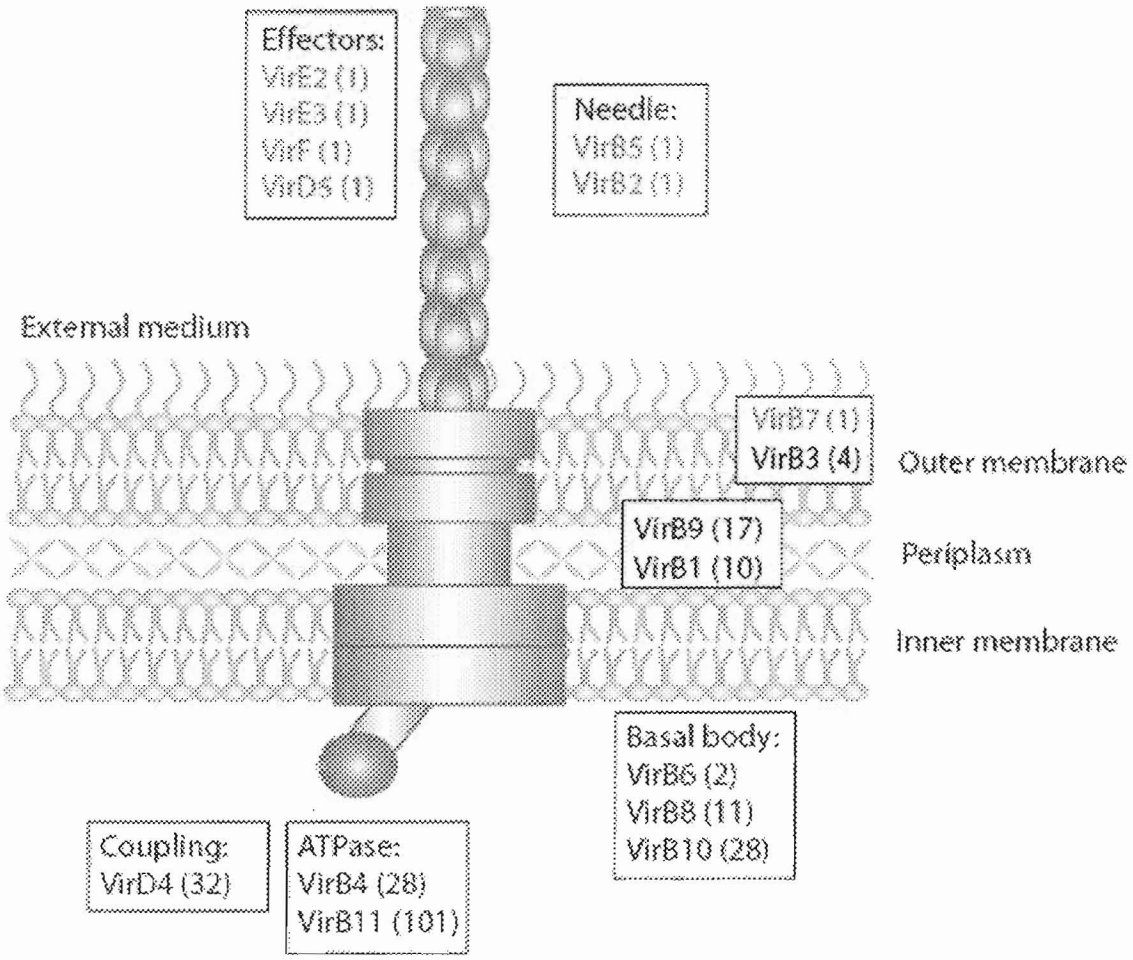


Figure 4.4 Pathogen-associated and "Common" genes involved in the *Agrobacterium tumefaciens* T4SS.

Genes that are pathogen-associated are shown in red and genes "common" to pathogens and non-pathogens are in black. The numbers in parentheses represent the number of different genera this gene is found in according to the BLAST analysis.



4.3.2 Basal body genes are "common" to pathogens and non-pathogens

T3SS genes which encode parts of the basal body, are known to be evolutionarily related to some flagella genes (reviewed in (Saier 2004)). In the Ysc-Yop T3SS, these include YscC encoding secretin (the outer ring component), YscN an ATPase, YscR-V (proteins in the basal body in contact with the cytoplasmic membrane), YscJ and YscQ. It is, therefore, not surprising that all basal body genes in this system were identified as "common" to both pathogens and non-pathogens, and are found in a minimum of 15 different genera with the analysis cutoffs used (Figure 4.1). Similar trends were observed for the *E. coli* T3SS (Figure 4.2), *Salmonella* T3SS (Figure 4.3), and the *Agrobacterium* T4SS (Figure 4.4).

4.4 Discussion

While it is clear that T3SS can be adapted to non-pathogenic purposes, this analysis suggests that certain specialized components are unique in, or strongly associated with pathogens. In summary, T3SS genes involved in formation of a pore in the host membrane, effectors, genes encoding part of the external injectisome, and those involved in host-cell contact dependent regulation are disproportionately associated with pathogens, whereas genes involved in the basal body complex are more "common" to both pathogens and non-pathogens. These specialized pathogen-associated components may therefore be important for pathogens that utilize T3SSs to interact with their hosts.

However, it should also be mentioned that some of these components, like the effectors for example, have notably low sequence identity between species. Such proteins, if they indeed are subject to less sequence constraints, may diverge more rapidly to a point that they are not detectable as similar across species at the BLAST cutoff used. However, even if some of these proteins are simply diverging faster, and therefore appearing more pathogen-specific, versus being selectively lost or gained in non-pathogens versus pathogens, the fact remains that the pathogen-associated genes identified do not have close homologs in the non-pathogens examined.

It has been suggested that targeting virulence-antigens may force pathogens to evolve toward less virulence (Gandon et al. 2003). Therefore, pathogen-associated components of the T3SSs may be better candidates for vaccine development than the T3SS as a whole as currently envisaged (Russmann 2004). To our knowledge, many of these more pathogen-associated components of the T3SS have not yet been specifically investigated for their utility as vaccine components. However, notably, particular pathogen-specific components (of the T3SS systems studied) investigated to date for their potential as vaccine components have been found to be immunogenic and protective. One such example is LcrV in *Yersinia pestis* which has been previously shown to be protective in mice (Anderson et al. 1996; Leary et al. 1995).

CHAPTER 5
CERTAIN CLASSES OF TOXINS ARE PATHOGEN-ASSOCIATED

5.1 Introduction

Genes that are conserved across multiple genera of pathogens and rarely or never found in non-pathogens in these same genera may play a more universal role in pathogenesis and virulence. Through my analysis of pathogen-associated genes (Chapter 3), I found that pathogen-associated genes are disproportionately toxins. However, there are several classes of toxins (see section 1.4.3 for a review) and I wished to determine if there were certain types of toxins that were disproportionately associated with pathogens.

Through a more indepth, semi-manual analysis, I have now found that several particular classes of toxins are pathogen-associated and found in multiple genera of pathogens. These classes include toxins with pore-forming, adenylate cyclase, and ADP-ribosyltransferase activities. Additionally, particular pathogen-associated toxins identified in this study have been successfully used as vaccine components, and so I propose that others that have not yet been investigated may be used as well and warrant further study.

5.2 Materials and methods

A dataset of VFs from the most well-known medically important pathogens was obtained from the VFDB (section 1.8.1). A subset of these VFs were

classified as "Toxin" according to the VFDB classification scheme, as well as additional subclassifications according to their function: for example, "pore-forming" or "ADP-ribosyltransferase". Each toxin gene was identified as pathogen-associated or "common" according to a BLAST-based analysis described in Chapter 3. Particular genes that may have been falsely identified as pathogen-associated or "common" were further manually inspected (see materials and methods in Chapter 3 for more details). I performed some manual curation of the functional classifications for each gene for overall completeness. Toxins gene were also subdivided into categories based on their COG (Clusters of Orthologous Groups; (Tatusov et al. 1997)) functional category. The major focus of this analysis was on toxins genes or categories present in multiple genera of pathogens, as determined by the BLAST analysis (Chapter 3), so mainly those particular genes/categories are reported.

5.3 Results

5.3.1 Pore-forming toxins, including cholesterol-dependent cytolysins, are strongly pathogen-associated

Through my analysis of pathogen-associated genes, I found that several pore-forming toxins, proteins that "punch holes" in the host cell membrane, are pathogen-associated. In particular one family of these toxins, the CDCs were found to be pathogen-associated and in multiple genera of pathogens (Table 5.1).

Table 5.1 Pathogen-associated toxins in multiple genera of diverse pathogens

Toxin category	Example toxins and species	Number of pathogen genera[a]
Pore-forming toxins (Cholesterol-dependent cytolysins)	Listeriolysin O (*Listeria monocytogenes*)	5
	Pneumolysin (*Streptococcus pneumoniae*)	5
	Streptolysin O (*Streptococcus pyogenes*)	5
Adenylate cyclase	Exoenzyme Y (*Pseudomonas aeruginosa*)	4
	Anthrax edema factor (*Bacillus anthracis*)	4
ADP-ribosyltransferase and/or GTPase activating	Exoenzyme S (*Pseudomonas aeruginosa*)	3
	Exoenzyme T (*Pseudomonas aeruginosa*)	4
	Pertussis toxin (*Bordetella pertussis*)	3

[a] Number of different pathogen genera this toxin is in according to the BLAST based analysis.

According to this analysis, CDCs from the VFDB dataset (which includes listeriolysin O, pneumolysin, and streptolysin O) were all found to be pathogen-associated according to the BLAST cutoff used (see materials and methods). CDCs have previously been shown to be present in the following 5 genera: *Clostridium*, *Streptococcus*, *Listeria*, *Bacillus*, and *Arcanobacterium.* This analysis did not extend to include CDCs in *Arcanobacterium*, because its genome sequence is not yet available. However, they were present in the other 4 pathogen genera: *Clostridium*, *Streptococcus*, *Listeria*, and *Bacillus*. They also show significant homology to putative hemolysins in the human pathogen *Bacteroides fragilis.* A relatively significant homolog to listeriolysin O was found in a non-pathogen *Lactobacillus acidiophilus*, which although below the chosen cutoff value, did show fairly high homology (e-value = 2.40E-06). However closer examination shows that the homology is not over the entire length of the listeriolysin O protein, and is only limited to small regions of the protein.

Other types of pore-forming toxins were also identified as pathogen-associated and in multiple genera through our analysis, such as *hla* encoding α-hemolysin, leukocidins (*lukF* and *lukS*) from *Staphylococcus aureus*, and the protective antigen in anthrax toxin (*pagA*) which forms the pore that delivers the lethal factor and edema factor into the host cytosol. All pathogen-associated toxins found in single or multiple genera are listed in Appendix B, and all "common" toxins are listed in Appendix C.

5.3.2 Adenylate cyclase toxins are pathogen-associated

Toxins with adenylate cyclase activity act on target cells by regulating the concentration of intracellular cAMP. Four such secreted toxins have been identified to date that increase cAMP concentration thereby modulating or halting cellular function (reviewed in (Ahuja et al. 2004)). Two of the four toxins were identified as pathogen-associated and present in multiple genera in my analysis: the anthrax toxin edema factor in *Bacillus anthracis* (encoded by *cya*) and exoenzyme Y from *Pseudomonas aeruginosa*, both of which are found in 4 genera of pathogens (Table 5.1). The third toxin with adenylate cyclase activity, encoded by the *cyaA* gene in *Bordetella pertussis*, was identified as "common" and found in 38 different genera. However, this toxin is bifunctional and contains both adenylate cyclase and hemolytic properties. The hemolytic domain is linked to a glycine-rich repeat motif that is found in all toxins in the RTX (repeat in toxin) family (Ladant et al. 1999), and this motif has been previously been identified in both pathogenic and some non-pathogenic Gram-negative bacteria (Kuhnert et al. 1997). Therefore, it is possible that the domain with adenylate cyclase activity

is pathogen-associated. Further manual inspection of the BLAST reports show that in fact selected non-pathogens analyzed only show homology to the hemolysin-encoded portion in the C-terminal region of CyaA, and not the N-terminal adenylate cyclase domain suggesting that the adenylate-cyclase domain may be pathogen-associated. The fourth adenylate cyclase toxin of *Yersinia pestis* was not included in our analysis as it was not obtained with the original VFDB dataset.

5.3.3 Toxins with ADP-ribosyltransferase activity are pathogen-associated

Toxins with ADP-ribosyltransferase activity were also identified as pathogen-associated, including pertussis toxin, cholera toxin, and *P. aeruginosa* toxins exotoxin A, exoenzyme S and T (Table 5.1). Some of these toxins were found in multiple pathogen genera: *exoS*, *exoT*, and *ptxA*, the active subunit of pertussis toxin, were found in 4,4, and 3 genera respectively, whereas the cholera toxin active subunit (*ctxA*) and exotoxin A were found in a single genus.

5.3.4 Additional pathogen-associated toxins

Some of the toxins identified as pathogen-associated in our analysis are either species-specific or genera-specific (Appendix B). Additionally, some toxins are present in 2 closely related genera. For example, the bacterial superantigens present in *Staphylococcus* and *Streptococcus* sp., are known to be structurally homologous (Baker et al. 2004). In my analysis the majority of superantigens were found to be in the above 2 genera, including staphylococcal enterotoxins (*entD*, *entE*, *sea*, *seb*, *sec1*, *sec3*, *sed*, *seg2*, *seh*, *sek2*) and almost

streptococcal pyrogenic exotoxins (*sme2*, *speA*, *speC*, *speG*, *speH*, *speI*, *speJ*, *speL*, *speM*, and *ssa*), with the exception of *speK,* which was identified in *Streptococcus* only. However, it is notable that particular toxins such as those with pore-forming, adenylate cyclase, and ADP-ribosyltransferase activity are still found within diverse genera and clearly are pathogen-associated.

5.4 Discussion

In this study, I identify particular toxin classes such as those with pore-forming, adenylate cyclase, and ADP-ribosyltransferase activity, that are found within diverse genera and are pathogen-associated. Several of the toxins identified here have been used as vaccine components, or have shown potential through *in vivo* immunization studies. For example, pertussis toxin mutants deficient in key enzymatic residues, were shown to be protective against *B. pertussis* infection in mice (Pizza et al. 1989). Inactivated toxoids, like this one, are still being used in vaccines (Plotkin 2005). Other studies report the efficacy of the adenylate cyclase toxin from *B. pertussis* as a vaccine component in combination with other antigens (Macdonald-Fyall et al. 2004; Orr et al. 2007). Also, several lines of evidence suggest that LLO peptides induce protective immunity against Listerial infection in mice (Bouwer et al. 1996; Harty et al. 1992)

These results, combined with the observation that many of these VFs are not part of the core pathogen genomes, suggest that if we put selection pressure on virulence specific antigens, we may be able to effectively reduce the number of pathogens carrying these genes, and hence provide selection for pathogens to evolve into less virulent forms. Our study confirms that several VFs used in

successful vaccinations are indeed specific to pathogens (based on the current pathogen and non-pathogen data available). Additional literature review of the pathogen-associated VFs we have identified in this analysis shows that some are protective either on their own, or in combination with other VFs. However, not all have been tested and clearly it would be prudent to examine the efficacy of other pathogen-associated associated genes that have not yet been investigated for their effectiveness in vaccines. Antigens that are common to both pathogens and commensals are presumably less likely to elicit strong immunogenic responses. This study provides the beginnings of a list of toxin genes (along with other pathogen-associated proteins, both of known and unknown function) that may encode good candidates for vaccine development.

CHAPTER 6
THE VIRULENCE GENE EXPERIMENT DATABASE (VGEDB)

6.1 Introduction

In light of increasing appreciation that defining a gene as a VF is a very contextual phenomenon, there is a need to develop highly-structured database resources that contain detailed information of the particular conditions in which a given gene is involved in virulence. Currently, there are four databases specifically focused on information about VFs (summarized in Table 1.4), of which two are published, the VFDB (Chen et al. 2005) and MvirDB (Zhou et al. 2007). These databases are further reviewed in section 1.8.1.

All of the available VF databases are centered around lists of VFs and do not provide structured, contextual information about the experimental conditions under which the genes appear to play a role in virulence. I have therefore developed the VGEDB, where each entry in the database contains information about a given virulence gene experiment, rather than a virulence gene.

Currently, the majority of VGEDB entries consist of STM experiments (section 1.7.2), a high-throughput approach for VF identification, as well as additional gene knockout experiments (and associated complementation experiments) that satisfy Molecular Koch's Postulates (section 1.2; (Falkow 1988)) for bacterial virulence gene identification (discussed in section1.2). I then

utilize the VGEDB to examine the accuracy of a BLAST-based computational approach for identifying VFs in genomic sequence (described in Chapter 7), and briefly discuss how continued expansion of the VGEDB can potentially improve the accuracy of such approaches. Additionally, I discuss how the rich contextual information in VGEDB can be used to gain insights into future questions regarding trends in bacterial virulence that could not be easily examined before.

6.2 Development of the VGEDB

The VGEDB currently contains over 960 virulence gene experiments from 16 different pathogens. The majority of these entries (710/960) are from STM experiments. The other 250 entries involve genes that have been inactivated individually or "knock-out" genes, where a measurable decrease in virulence on its host is observed and measured. For a given VF experiment, all experimental results, methods, gene, sequence, bacteria and host information, are manually curated from the literature and compiled into the database. The VGEDB database schema (Figure 6.1) consists of 8 tables: Experiment, Organism, Result, Literature, Gene, Nucleotide, Protein, and Name.

The Experiment table contains details of the type of mutant, host and infection conditions for a given experiment. An example of an annotated experiment is illustrated in Figure 6.2. Both *in vitro* and *in vivo* experiments are included. Details of an *in vivo* experiment typically include the number of host organisms used, infection dose, time required for infection, method of inoculation, etc. Records for *in vitro* experiments generally include the number of replications, multiplicity of infection, and the number of host cells and bacteria used in the

infection. I also note the type of "knock-out" method used: insertion or deletion mutation for example, and if available, if the mutation had a polar or non-polar effect.

The Organism table stores information on bacterial strain and host used in the infection. A host can be a whole organism (e.g. mouse) or a type of cell (e.g. HELA cell). If available, taxonomy information for all bacterial strains and hosts used in the experiment are retrieved from the National Center for Biotechnology Information (NCBI) website and included in the database.

The Result table contains all results for a given experiment. To accommodate different result formats, a code 'EX' identifies exact numerical/statistical results, 'AP' is noted for results presented in graphical format in a publication, which are approximated from the original data by the curator, and 'FI' is indicated for figures such as electron microscopy images. Additionally, auxotrophic mutants (mutants with reduced growth rate compared to wild-type *in vitro*) are flagged so the user can choose not to retrieve these particular genes.

The Literature table contains a reference to the original published journal article through NCBI Pubmed. The Gene table incorporates various functional classifications for a given gene, provided from the following sources: 1) VFDB (Chen et al. 2005), 2) PRINTS database of virulence factors, 3) TvFac from the Los Alamos National Laboratory, 4) our own VGEDB classification, 5) COG (Tatusov et al. 1997), and 6) PSORTb (Gardy et al. 2005). Additionally, the Gene table is linked to the Protein and Nucelotide tables, which store sequence information retrieved from GenPept and GenBank respectively. Finally, the Name

table stores all gene, protein, and ORF names, as well as any alternative names, such that experiments on a similar gene but with different name can be tracked and retrieved over time. The VGEDB is a relational database, implemented in Perl, and developed with the open source software mySQL.

Figure 6.1 VGEDB Database schema

Each table is represented by a box with data centered around one infection experiment for a given gene, involving a particular pathogen and host organism/tissue. For brevity, some fields are not included in this diagram. The relationships between the tables are shown (1=one; M=many). For example: The Literature table is linked to the Experiment table by the 'Pubmed_id' in a 1 to Many relationship, since there may be many experiments in one publication. PK: Primary key, FK: Foreign key.

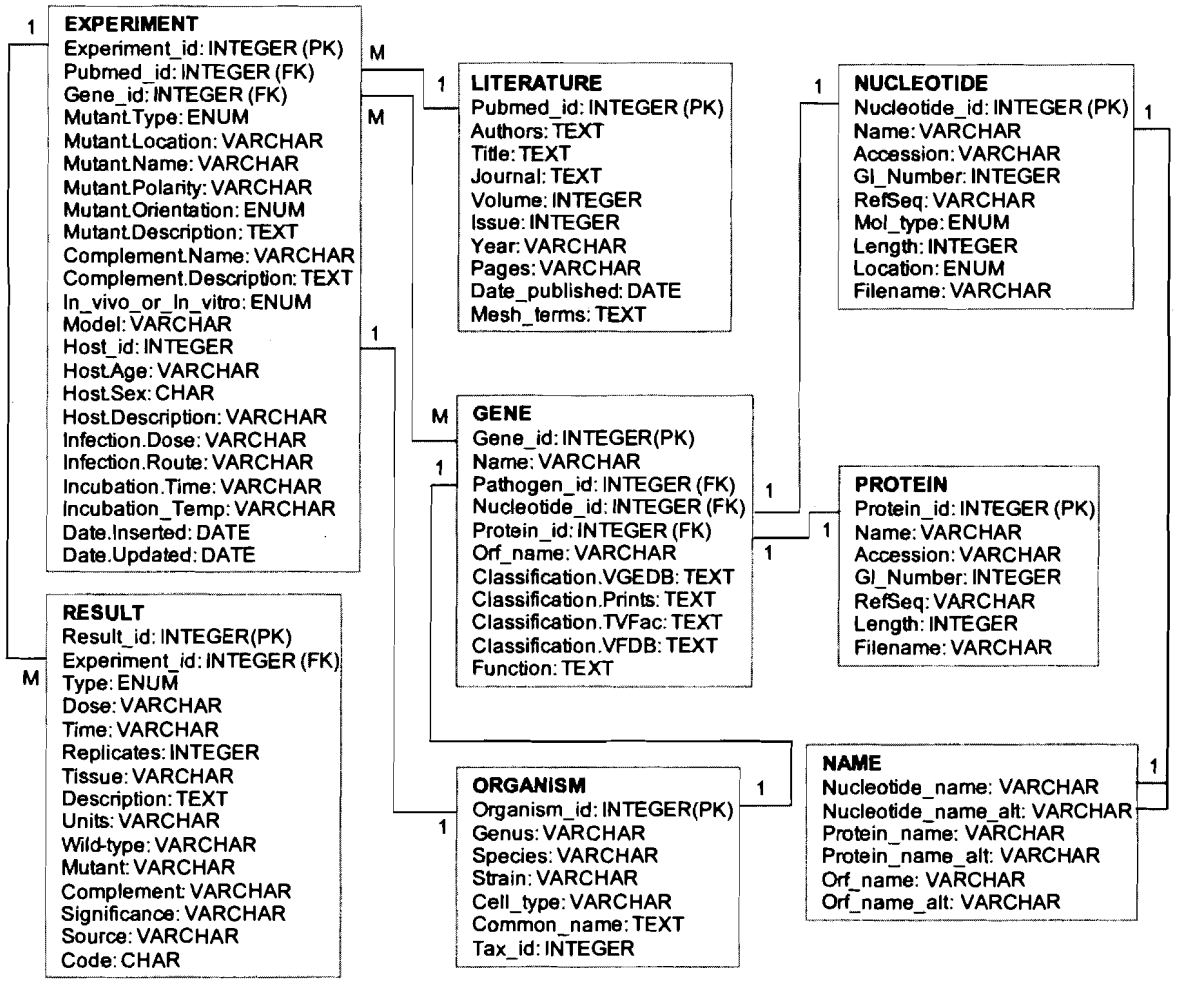
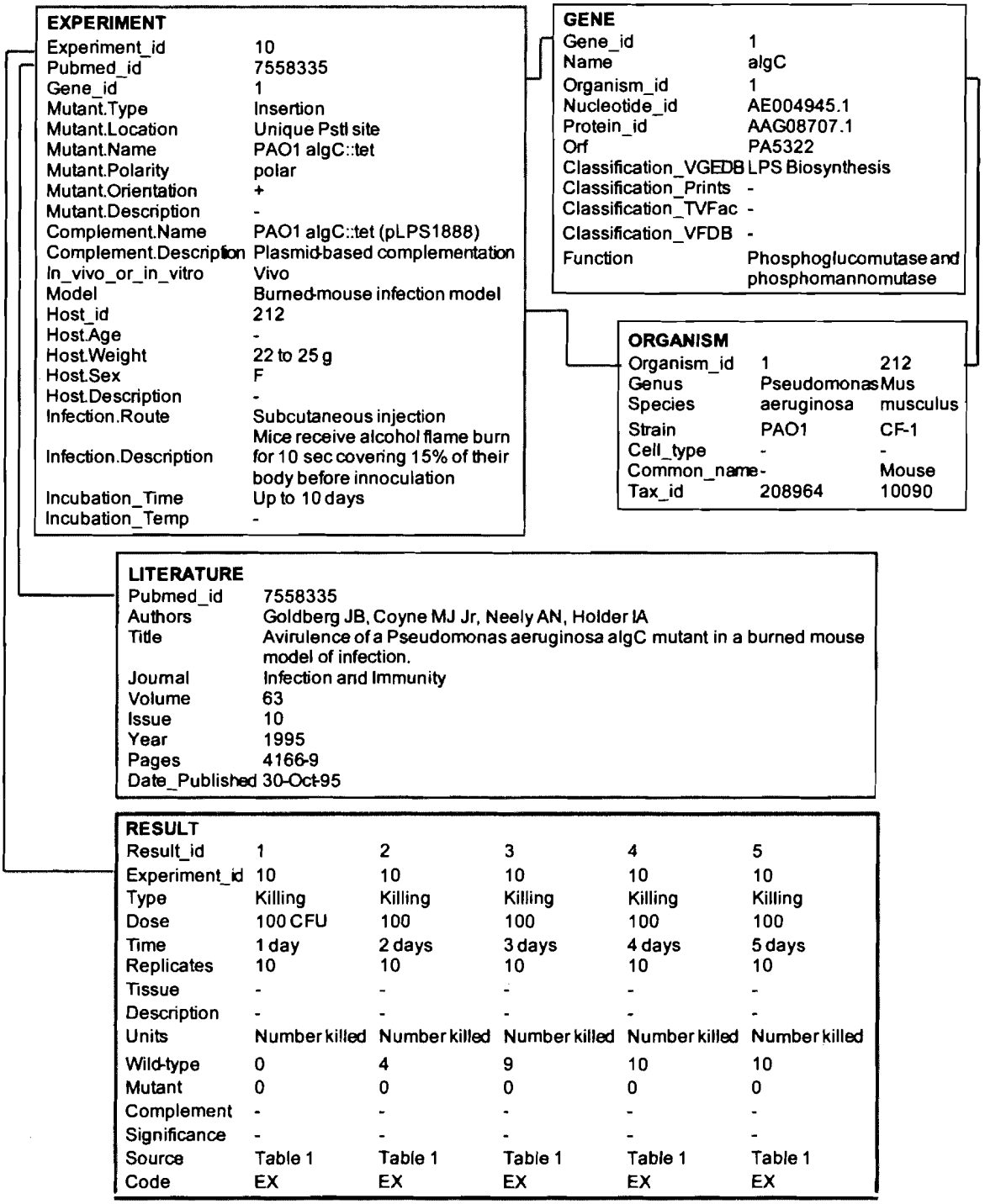


Figure 6.2 Example VGEDB entry

This example annotates an experiment in which the virulence of wild-type *Pseudomonas aeruginosa* PAO1, and that of a genetically defined algC mutant, PAO1 algC::tet, were compared in a burned-mouse model of infection (Goldberg et al. 1995). **For brevity, some tables and fields are not included in this diagram.**



EXPERIMENT

Experiment_id	10
Pubmed_id	7558335
Gene_id	1
Mutant.Type	Insertion
Mutant.Location	Unique PstI site
Mutant.Name	PAO1 algC::tet
Mutant.Polarity	polar
Mutant.Orientation	+
Mutant.Description	-
Complement.Name	PAO1 algC::tet (pLPS1888)
Complement.Description	Plasmid-based complementation
In_vivo_or_in_vitro	Vivo
Model	Burned-mouse infection model
Host_id	212
Host.Age	-
Host.Weight	22 to 25 g
Host.Sex	F
Host.Description	-
Infection.Route	Subcutaneous injection
Infection.Description	Mice receive alcohol flame burn for 10 sec covering 15% of their body before innoculation
Incubation_Time	Up to 10 days
Incubation_Temp	-

GENE

Gene_id	1
Name	algC
Organism_id	1
Nucleotide_id	AE004945.1
Protein_id	AAG08707.1
Orf	PA5322
Classification_VGEDB	LPS Biosynthesis
Classification_Prints	-
Classification_TVFac	-
Classification_VFDB	-
Function	Phosphoglucomutase and phosphomannomutase

ORGANISM

Organism_id	1	212
Genus	Pseudomonas	Mus
Species	aeruginosa	musculus
Strain	PAO1	CF-1
Cell_type	-	-
Common_name	-	Mouse
Tax_id	208964	10090

LITERATURE

Pubmed_id	7558335
Authors	Goldberg JB, Coyne MJ Jr, Neely AN, Holder IA
Title	Avirulence of a Pseudomonas aeruginosa algC mutant in a burned mouse model of infection.
Journal	Infection and Immunity
Volume	63
Issue	10
Year	1995
Pages	4166-9
Date_Published	30-Oct-95

RESULT

Result_id	1	2	3	4	5
Experiment_id	10	10	10	10	10
Type	Killing	Killing	Killing	Killing	Killing
Dose	100 CFU	100	100	100	100
Time	1 day	2 days	3 days	4 days	5 days
Replicates	10	10	10	10	10
Tissue	-	-	-	-	-
Description	-	-	-	-	-
Units	Number killed	Number killed	Number killed	Number killed	Number killed
Wild-type	0	4	9	10	10
Mutant	0	0	0	0	0
Complement	-	-	-	-	-
Significance	-	-	-	-	-
Source	Table 1	Table 1	Table 1	Table 1	Table 1
Code	EX	EX	EX	EX	EX

6.3 Future use of the VGEDB

The contextual data present in the VGEDB can allow us to ask more sophisticated questions of virulence trends that could not easily be answered before. For example, we can ask questions such as the following: 1) Which VFs are involved in disease in different hosts (e.g. Mouse, Worm, Plant – both common and species-specific), 2) Are their functional categories of VFs that tend to be more associated with broad host-range infection?, 3) Which VFs, or VF categories, play a significant role in severe or lethal disease (e.g. mutants with highest change in LD50 across multiple species)? Studying these types of trends can take advantage of the contextual information provided by the VGEDB, and may potentially lead to genuinely new insights regarding trends in pathogen virulence.

In addition, I have performed an analysis of potentially improving the accuracy of BLAST-based identification of VFs by utilizing a larger dataset of VFs, such as the VGEDB, into the analysis. This is described further in Chapter 7.

6.4 Discussion

The VGEDB contains a set of high quality, well-annotated virulence genes that have been experimentally verified through STM and bacterial gene knockout experiments. The unique feature of this database is that rather than just providing a simple list of virulence genes, the VGEDB incorporates contextual information about the experimental conditions under which a given gene is involved in

virulence, and this information can be used a base for more sophisticated questions regarding trends in virulence.

Initially I set out to create a database that was focused on annotating all VFs, primarily based on gene knockout experiments. However, with the advent of STM data, I decided to focus more on the use of this STM data to provide a more consistent type of measure of virulence across species. This collection of STM data, in an organized fashion, has not been developed previously, let alone any development of a highly contextual database of VFs. This STM dataset collection alone will now permit powerful new analyses of these data to be performed, including identifying commonalities and differences across species and conditions, that could not be easily performed before.

In this study, I utilized the VGEDB, coupled with additional high quality datasets of known VFs, to investigate the accuracy of current VF identification methods, and propose that further expansion of the VGEDB could improve computational identification of VFs in newly sequenced genomes (discussed in Chapter 7). Additionally, I hypothesize that the rich contextual information in the VGEDB can be used to gain insights into future questions regarding trends in bacterial virulence and that only through the analysis of such resources can more complex patterns of trends in virulence be identified. There is still much to be done on this front, but the development of this more complex VF databse schema is an important start to being able to perform more sophisticated analyses of trends in virulence.

CHAPTER 7
ESTIMATING THE ACCURACY OF COMPUTATIONAL IDENTIFICATION OF VIRULENCE FACTORS

7.1 Introduction

When a bacterial pathogen genome is first sequenced researchers often have a primary interest in identifying any genes that may encode VFs. However, current approaches for computationally identifying VFs are very ad hoc – usually involving a BLAST analysis against a dataset of known VFs, such as that used to identify T3SS genes in *Chlamydia pneumonia* (Kalman et al. 1999). There are several limitations with current computational methods (discussed in section 1.8.2), including the fact that the accuracy of these methods has not been examined before. As the number of sequenced bacterial genomes continue to increase exponentially, such as from metagenomic studies for example, propagation of errors in accuracy will inevitably increase as well.

There is therefore a need to investigate the accuracy of these methods and to develop more robust methods to computationally identify potential VFs. This is not a trivial endeavour due to the very contextual nature of a VF. Some genes are VFs in one species, while not in another, due to changing genomic context. However, by investigating the accuracy of current VF prediction methods, I hypothesize that we can gain some insight into how well common methods are performing as an important first step to developing improved methods.

In this study, I examine the accuracy of a BLAST-based method using high quality VFs from the VFDB (section 1.8.1), as well as additional VFs that were identified experimentally (either through signature-tagged mutagenesis studies or additional gene knock-out experiments) from the VGEDB (Chapter 6). These datasets were used to estimate both the sensitivity (recall) and specificity (precision) of a BLAST-based approach for identifying VFs. I found that both sensitivity and specificity are quite low and discuss possibilities for improving computational identification of VFs in genomic sequence through the use of high quality VF database resources such as the VGEDB (Chapter 6).

7.2 Materials and methods

I used a dataset of 1819 VFs from the VFDB (described in section 1.8.1) to identify putative VF homologs in the deduced proteome of *P. aeruginosa* PAO1 using BLAST (Altschul et al. 1990) (e-value cut-off of 10^{-7} was used, which represents a common e-value cutoff used in such analyses and a value used for gene family identification, though other cut-offs were initially examined). This analysis is typical of what is commonly performed during genome annotation to identify possible VFs. The BLAST results were then compared against a dataset of "true" VFs, consisting of 148 reported genes identified in a *P. aeruginosa* PAO1 STM screen, where mutants with disruptions in these genes showed reduced virulence in a rat model of chronic respiratory infection (Potvin et al. 2003). The sensitivity (recall) was calculated using the formula: true positives/(true positives + false negatives), or TP/(TP+FN), and specificity (precision) using true positives/(true positives + false positives), or TP/(TP+FP).

TPs correspond to the number of "true" VFs (out of the 148 experimentally identified through STM) that were also identified by BLAST, FNs are the "true" VFs not identified by BLAST, and FPs are genes identified by BLAST that were not identified in the STM screen.

To investigate whether accuracy of a BLAST based approach was improved when using a larger VF dataset for the BLAST analysis against the test genome (*P. aeruginosa* strain PAO1), I repeated this analysis using two larger datasets of VFs for the BLAST analysis: One, I used a dataset that comprises a collection of genes from diverse species that have been experimentally identified through STM studies and are included in the VGEDB (see description of the VGEDB described in Chapter 7). Note that to avoid evaluating test data with the same data, I did exclude the148 VFs from *P. aeruginosa* described above in this dataset. The second larger dataset comprised the VGEDB as well as the VFDB datasets, combined together into one dataset.

One potential bias with this analysis is that possible false-positives in the STM-based analysis may occur due to polar effects on downstream genes. Using a survey of 10 STM papers that investigated the number of false-positives due to polar effects, I used an estimation of this error rate of 25% (see section 7.3.3).

7.3 Results

7.3.1 Overall accuracy of BLAST-based identification of virulence factors is very low

Using these well defined datasets, I found that the sensitivity of this BLAST-based approach for identifying VFs was quite low: only 25% (37/148) of

the experimentally determined genes (under this particular infection condition) were identified (Table 7.1). Additionally, the estimated specificity is only 3.4%, indicating many potentially false positive VFs are being identified.

When the VFDB and VGEDB datasets are combined, there is an increase in sensitivity to 39.2%. However, the specificity remained approximately the same at 3.6%.

Table 7.1 Accuracy of BLAST-based identification of virulence factors

VF Dataset	Sensitivity (%)	Specificity (%)
VFDB Only	25.0	3.4
VGEDB Only	26.4	3.8
VGEDB and VFDB	39.2	3.6
VFDB with correction[a]	31.0	4.3
VGEDB and VFDB with correction[a]	48.6	4.4

[a] Correction due to polar effects of VF identification through STM

7.3.2 Classification of virulence factors identified and not-identified with a BLAST-based approach

I also found that VFs identified by BLAST are disproportionately classified as 'Signal transduction mechanisms' and 'Cell motility and secretion' COG functional categories (p = 5.27E-03 and p = 3.61E-02 respectively). However, the statistical significance, while less than 0.05, was not high. Genes that are 'Unclassified' or classified as 'Function unknown' are not well identified by this approach (17.1% and 16.2% respectively) as shown in Table 7.2, however this was not statistically significant.

Table 7.2 COG Classification of VFs identified and not identified with BLAST

COG Category	VFs identified by BLAST		VFs not identified by BLAST		*p*-value[a]
	#	%	#	%	
Categories with higher percentage of VFs identified by BLAST					
Signal transduction mechanisms	7	18.9	1	0.9	5.27E-03*
Cell motility and secretion	4	10.8	0	0.0	3.61E-02*
Intracellular trafficking secretion and vesicular transport	3	8.1	0	0.0	1.03E-01
Transcription	4	10.8	3	2.7	2.76E-01
Inorganic ion transport and metabolism	3	8.1	4	3.6	8.56E-01
Posttranslational modification protein turnover chaperones	2	5.4	4	3.6	1
Carbohydrate transport and metabolism	2	5.4	5	4.5	1
Cell envelope biogenesis outer membrane	1	2.7	3	2.7	1
Coenzyme metabolism	0	0.0	0	0.0	1
Categories with higher percentage of VFs not identified by BLAST					
Unclassified	1	2.7	19	17.1	1.38E-01
Function unknown	2	5.4	18	16.2	5.67E-01
Energy production and conversion	1	2.7	10	9.0	8.78E-01
Nucleotide transport and metabolism	0	0.0	5	4.5	8.71E-01
General function prediction only	1	2.7	8	7.2	9.47E-01
Translation ribosomal structure and biogenesis	0	0.0	3	2.7	1
Lipid metabolism	1	2.7	6	5.4	1
Cell division and chromosome partitioning	0	0.0	2	1.8	1
Amino acid transport and metabolism	3	8.1	11	9.9	2
Defense mechanisms	1	2.7	4	3.6	1
DNA replication recombination and repair	1	2.7	4	3.6	1
Secondary metabolites biosynthesis transport and catabolism	0	0.0	1	0.9	1

[a] Pearson's Chi-squared test with Yates' continuity correction. Asterisks indicate statistical significance (p-value < 0.05).

7.3.3 Potential bias in virulence factors identified through signature tagged mutagenesis

One potential bias with this analysis is that possible false-positives in the STM-based analysis may occur due to polar effects on downstream genes. In a survey of 10 STM papers that investigated the number of false-positives due to polar effects, I estimated an average false positive rate of 25% (Begun et al. 2005; Collins et al. 2005; Grant et al. 2005; Ku et al. 2005; Lawlor et al. 2005; Li et al. 2005; Ojha et al. 2005; Paik et al. 2005; Shah et al. 2005; van Diemen et al. 2005). After correcting for these STM false positives, the sensitivity of identifying VFs with this approach is increased to 31%, and the specificity to 4.3% (Table 7.1).

7.4 Discussion

Overall, these results suggest that the accuracy of current computational approaches for identifying VFs is relatively low – only 31% sensitivity and 4.3% specificity for an analysis that estimated false positives due to the STM method. I show that the sensitivity can be greatly improved by incorporating a more comprehensive list of VFs by factoring in datasets of putative VFs that have been experimentally determined. However, I believe this increase is not enough to warrant the exclusive use of this method for identification of potential VFs in newly sequenced genomes. I also note that novel VFs are not well identified with this approach, or well classified with existing gene functional classification tools, since many of the genes involved were classified as "Unclassified" or "Function Unknown". This suggests that there is a need to develop classification tools more

specific for genes involved in virulence, as has been initiated with the VFDB classification scheme. Also, more novel approaches to identify both known and unknown VF genes, for example, by identifying pathogen-associated genes (discussed in Chapter 3), may prove more useful than current methods. In particular, when analyzing metagenomic data to identify possible pathogens, one must be careful with the approach of identifying VFs in sequences by BLAST and then assuming that they are either VFs, or that the microbe encoding them is necessarily a pathogen. The identification of pathogen-associated genes, versus simply genes involved in host-association, may identify more virulence- specific genes and have more utility in identifying some pathogens from metagenomic sequence data.

However, there are a number of cautionary notes regarding these data. The specificity estimate may be exceptionally low in part because the STM screen may not have comprehensively identified all VFs present in the pathogen. Since virulence is so contextual, some genes may be required for virulence under some conditions, but not others. However, based on a manual analysis of selected genes, it is clear that a significant proportion of the false positives are due to this BLAST analysis detecting similarity to genes that are not VFs but share significant similarity to such genes. As the VF dataset used in the BLAST analysis is increased in size, the specificity does not change. Accounting for some false positives in the STM data does affect specificity, but at a very minimal level. Regardless of the reason for this low specificity, it is clear that the overlap of STM data, versus common BLAST-based identification of VFs using a VF

dataset, is low. Regarding sensitivity or recall, the values were more reasonable, though still very low, but did increase when a larger VF dataset was used in the BLAST analysis. This may be in part due to the fact that STM identifies some genes that are involved in *in vivo* growth, but aren't strictly VFs. However, even when a dataset based on STM data from other species was used, the sensitivity was still below 50%. Clearly more sensitive detection methods are required – likely coupled with a more expanded dataset of VFs used for training of a computational method.

Regardless, this study suggests that the accuracy of common BLAST-based methods for the identification of VFs is likely very low and clearly has low overlap with STM-based data. While it is generally appreciated that such an analysis would not be highly accurate, to my knowledge it has not been appreciated just how poorly such a method performs. I hypothesize that as more VFs from STM studies are added to the VGEDB (see below), with contextual information about the infection conditions involved, we can create a more comprehensive, contextual, list of VFs and be able to identify potential VFs in genomic sequence with increased sensitivity. However, methods that are more precise than BLAST must be explored or coupled with a BLAST-based analysis. The use of orthology information may be useful to avoid identifying non-VF paralogs that are similar to VFs through a BLAST-based approach (Fulton et al. 2006). Likely in the end only certain classes of VFs may be accurately identified using a computational approach, but at minimum, there is a significant need to improve our current methods and use them with the understanding of the degree

of accuracy of the methods used. More contextual information regarding the genomic context and gene context (i.e. what other genes are present in the genome) for a given gene in a genome will also likely be necessary to improve computational VF identification. This will become increasingly critical as metagenomic analysis of microbes of uncultural species becomes increasingly common, necessitating more dependence on computational analyses of virulence.

CONCLUSIONS

These analyses of pathogen-associated genes and a curated dataset of VFs suggest that such genes are, on average, more associated with GIs versus non-GIs. Collectively, these results also further suggest that "offensive" and virulence-specific VFs in bacterial pathogens are more likely to be associated with GIs, versus VF homologs in non-pathogens involved in more passive host-association functions. Though there are of course certain bacteria that are exceptions, the work supports the strong role of GIs in the evolution of virulence and provides the first systematic analysis of this trend across diverse genera. I also identify pathogen-associated genes and provide evidence that certain components of T3SSs and certain types of toxins are quite selectively pathogen-associated. Additionally, I provide whole genome datasets of pathogen-associated genes in a set of completely sequenced bacterial genomes. Such pathogen-associated genes may warrant further study for their potential as anti-infective drug targets and vaccine components.

In addition, I have developed the VGEDB, a resource that incorporates detailed information about experimental conditions used to identify a given VF. The contextual information in this database will potentially enable more sophisticated analyses of virulence and VFs not easily performed before. Finally, with the continuing increase of genomic data, I propose that there is a need to

develop more robust approaches to computationally identify VFs in newly sequenced genomes.

APPENDICES

Appendix A:
Virulence Factor Database classification of "offensive" and "defensive" virulence factors

Offensive virulence factors

1) Adherence
2) Invasion
3) Toxin
    - 3.1) Toxin: membrane-acting
    - 3.2) Toxin: membrane-damaging
        - 3.2.1) Pore-forming
            - 3.2.1.1) Channel-forming involving beta-sheet-containing toxin
            - 3.2.1.2) Channel-forming involving alpha-helix-containing toxins
            - 3.2.1.3) Thiol-activated cholesterol-binding cytolysin
            - 3.2.1.4) RTX toxin
    - 3.3) Toxin: intracellular toxin
        - 3.3.1) ADP-ribosyltransferase
        - 3.3.2) Adenylate cyclase
        - 3.3.3) Deamidase
        - 3.3.4) Guanylate cyclase
        - 3.3.5) N-glycosidase
        - 3.3.6) Dnasel
4) Actin-based motility
5) Secretion system
    - 5.1) Type III secretion system
    - 5.2) Type IV secretion system
    - 5.3) Autotransporter (Type V)

Defensive virulence factors

1) Antiphagocytosis
2) Anti-proteolysis
3) Cellular metabolism
4) Phase variation
5) Serum resistance
7) Ig protease
8) Stress protein
9) Complement Protease

Nonspecific virulence factor

1) Iron uptake system
2) Magnesium uptake system
3) Exoenzyme

Regulation of virulence-associated genes

1) Regulation

Appendix B:
Pathogen-associated toxin genes

VFDB ID	Organism	Toxin Gene Name	Toxin Description	Number of Pathogen genera
VFG0011	*Bordetella pertussis* Tohama I	*ptxA*	pertussis toxin subunit 1 precursor	3
VFG0012	*Bordetella pertussis* Tohama I	*ptxB*	pertussis toxin subunit 2 precursor	1
VFG0013	*Bordetella pertussis* Tohama I	*ptxD*	pertussis toxin subunit 4 precursor	1
VFG0014	*Bordetella pertussis* Tohama I	*ptxE*	pertussis toxin subunit 5 precursor	1
VFG0015	*Bordetella pertussis* Tohama I	*ptxC*	pertussis toxin subunit 3 precursor	1
VFG0017	*Bordetella pertussis* Tohama I	*ptlB*	pertussis toxin transport protein	5
VFG0019	*Bordetella pertussis* Tohama I	*ptlD*	putative membrane protein	1
VFG0020	*Bordetella pertussis* Tohama I	*ptlI*	putative bacterial secretion system protein	1
VFG0026	*Bordetella pertussis* Tohama I	*dnt*	dermonecrotic toxin	3
VFG0074	*Listeria monocytogenes* EGD-e	*hly*	listeriolysin O precursor	5
VFG0107	*Vibrio cholerae* N16961	*ctxA*	cholera enterotoxin, A subunit	1
VFG0108	*Vibrio cholerae* N16961	*ctxB*	cholera enterotoxin, B subunit	1
VFG0115	*Pseudomonas aeruginosa* PAO1	*toxA*	exotoxin A precursor	1
VFG0147	*Pseudomonas aeruginosa* PAO1	*exoS*	exoenzyme S	3
VFG0148	*Pseudomonas aeruginosa* PAO1	*exoT*	exoenzyme T	4
VFG0150	*Pseudomonas aeruginosa* PAO1	*exoY*	adenylate cyclase ExoY	4
VFG0422	*Yersinia pestis* CO92	*ymt*	murine toxin	2
VFG0636	*Shigella flexneri* (serotype 2a) 301	*set1B*	ShET1B	1
VFG0637	*Shigella flexneri* (serotype	*set1A*	ShET1A	1

VFDB ID	Organism	Toxin Gene Name	Toxin Description	Number of Pathogen genera
	2a) 301			
VFG0676	*Bacillus anthracis* Sterne	*lef*	anthrax toxin lethal factor precursor, lef,	1
VFG0677	*Bacillus anthracis* Sterne	*pagA*	anthrax toxin moiety, protective antigen, pagA	1
VFG0678	*Bacillus anthracis* Sterne	*cya*	calmodulin sensitive adenylate cyclase, edema factor, cya,	4
VFG0835	*Escherichia coli* O157:H7 EDL933	*stx1A*	shiga-like toxin 1 subunit A encoded within prophage CP-933V	2
VFG0836	*Escherichia coli* O157:H7 EDL933	*stx1B*	shiga-like toxin 1 subunit B encoded within prophage CP-933V	2
VFG0837	*Escherichia coli* O157:H7 EDL933	*stx2A*	shiga-like toxin II A subunit encoded by bacteriophage BP-933W	2
VFG0838	*Escherichia coli* O157:H7 EDL933	*stx2B*	shiga-like toxin II B subunit encoded by bacteriophage BP-933W	2
VFG0859	*Escherichia coli* 42	*set1A*	toxin subunit Set1A	2
VFG0860	*Escherichia coli* 42	*set1B*	toxin subunit Set1B	2
VFG0863	*Escherichia coli* 42	*astA*	heat-stable enterotoxin 1	1
VFG0951	*Streptococcus pyogenes* MGAS315	*speA*	exotoxin type A precursor - phage associated	2
VFG0952	*Streptococcus pyogenes* SF370	*speI*	streptococcal exotoxin I	2
VFG0953	*Streptococcus pyogenes* MGAS315	*speK*	streptococcal pyrogenic exotoxin SpeK - phage associated	1
VFG0954	*Streptococcus pyogenes* MGAS315	*ssa*	streptococcal superantigen SSA - phage associated	2
VFG0957	*Streptococcus pyogenes* MGAS8232	*speL*	putative exotoxin precursor (SpeL)	2
VFG0958	*Streptococcus pyogenes* MGAS8232	*speM*	putative exotoxin precursor (SpeM)	2
VFG0976	*Streptococcus pyogenes* SF370	*Slo*	streptolysin O precursor	5
VFG0977	*Streptococcus pyogenes* SF370	*sagA*	streptolysin S associated protein	1
VFG0978	*Streptococcus pyogenes*	*speC*	pyrogenic exotoxin C precursor,	2

VFDB ID	Organism	Toxin Gene Name	Toxin Description	Number of Pathogen genera
	SF370		phage associated	
VFG0979	*Streptococcus pyogenes* SF370	*speG*	exotoxin G precursor	2
VFG0980	*Streptococcus pyogenes* SF370	*speJ*	putative exotoxin (superantigen)	2
VFG0981	*Streptococcus pyogenes* SF370	*smeZ*	mitogenic exotoxin Z	2
VFG0982	*Streptococcus pyogenes* SF370	*speH*	streptococcal exotoxin H precursor	2
VFG1273	*Staphylococcus aureus* MW2	*hlgA*	gamma-hemolysin chain II precursor	2
VFG1274	*Staphylococcus aureus* MW2	*hlgC*	gamma-hemolysin component C	2
VFG1275	*Staphylococcus aureus* MW2	*hlgB*	gamma-hemolysin component B	2
VFG1276	*Staphylococcus aureus* MW2	*lukF*	Panton-Valentine leukocidin chain F precursor	2
VFG1277	*Staphylococcus aureus* MW2	*lukS*	Panton-Valentine leukocidin chain S precursor	2
VFG1292	*Staphylococcus aureus* MW2	*hld*	delta-hemolysin	1
VFG1293	*Staphylococcus aureus* MW2	*hla*	Alpha-Hemolysin precursor	2
VFG1325	*Staphylococcus aureus* MW2	*sea*	staphylococcal enterotoxin A precursor	2
VFG1326	*Staphylococcus aureus* MW2	*seg2*	staphylococcal enterotoxin SeG	2
VFG1327	*Staphylococcus aureus* MW2	*sek2*	staphylococcal enterotoxin Sek	2
VFG1332	*Streptococcus agalactiae* 2603V/R	*cylE*	cylE protein	1
VFG1333	*Streptococcus agalactiae* 2603V/R	*cfb*	CAMP factor	2
VFG1363	*Streptococcus pneumoniae* TIGR4	*ply*	pneumolysin	5
VFG1800	*Staphylococcus aureus*	*Eta*	exfoliative toxin A	1
VFG1802	*Staphylococcus aureus*	*seb*	enterotoxin B	2
VFG1803	*Staphylococcus aureus*	*seh*	enterotoxin H precursor	2

VFDB ID	Organism	Toxin Gene Name	Toxin Description	Number of Pathogen genera
VFG1804	*Staphylococcus aureus*	*sec1*	staphylococcal enterotoxin C3	2
VFG1805	*Staphylococcus aureus*	*sec3*	enterotoxin C1 precursor	2
VFG1806	*Staphylococcus aureus*	*sed*	staphylococcal enterotoxin D	2
VFG1807	*Staphylococcus aureus*	*entD*	enterotoxin D precursor	2
VFG1808	*Staphylococcus aureus*	*entE*	enterotoxin E precursor	2
VFG1809	*Staphylococcus aureus* N315	*Tst*	toxic shock syndrome toxin-1	1
VFG1828	*Shigella dysenteriae* (serotype 1)	*stxA*	Shiga toxin subunit A; RNA-N-glycosidase; catalyticsubunit	2
VFG1829	*Shigella dysenteriae* (serotype 1)	*stxB*	Shiga toxin subunit B; receptor binding subunit	2

Appendix C:
Toxin genes "common" to both pathogens and non-pathogens

VFDB ID	Organism	Toxin Gene Name	Toxin Description	Number of Genera
VFG0016	*Bordetella pertussis* Tohama I	*ptlA*	pertussis toxin transport protein	4
VFG0018	*Bordetella pertussis* Tohama I	*ptlC*	putative bacterial secretion system protein	27
VFG0021	*Bordetella pertussis* Tohama I	*ptlE*	putative bacterial secretion system protein	16
VFG0022	*Bordetella pertussis* Tohama I	*ptlF*	putative bacterial secretion system protein	18
VFG0023	*Bordetella pertussis* Tohama I	*ptlG*	putative bacterial secretion system protein	28
VFG0024	*Bordetella pertussis* Tohama I	*ptlH*	putative bacterial secretion system protein	77
VFG0025	*Bordetella pertussis* Tohama I	*cyaA*	bifunctional hemolysin-adenylate cyclase precursor	38
VFG0109	*Vibrio cholerae* N16961	*zot*	zona occludens toxin	5
VFG0110	*Vibrio cholerae* N16961	*ace*	accessory cholera enterotoxin	1
VFG0157	*Pseudomonas aeruginosa* PAO1	*plcH*	hemolytic phospholipase C precursor	17
VFG0158	*Bordetella pertussis* Tohama I	*cyaC*	cyclolysin-activating lysine-acyltransferase	5
VFG0279	*Helicobacter pylori* 26695	*vacA*	vacuolating cytotoxin	2
VFG0840	*Escherichia coli* O157:H7 EDL933	*hlyA*	hemolysin toxin protein	30
VFG0841	*Escherichia coli* O157:H7 EDL933	*hlyB*	hemolysin transport protein	123
VFG0842	*Escherichia coli* O157:H7 EDL933	*hlyC*	hemolysin transport protein	10
VFG0843	*Escherichia coli* O157:H7 EDL933	*hlyD*	hemolysin transport protein	62
VFG0861	*Escherichia coli* 42	*pic*	Pic serine protease precursor	15
VFG0862	*Escherichia coli* 42	*pet*	Pet serine protease precursor	18
VFG0905	*Escherichia coli* CFT073	*hlyC*	Hemolysin C	8
VFG0906	*Escherichia coli* CFT073	*hlyA*	Hemolysin A	34
VFG0907	*Escherichia coli* CFT073	*hlyB*	Hemolysin B	123

VFDB ID	Organism	Toxin Gene Name	Toxin Description	Number of Genera
VFG0908	*Escherichia coli* CFT073	*hlyD*	Hemolysin D	61
VFG0983	*Vibrio cholerae* N16961	*rtx*	RTX toxin RtxA	55
VFG1269	*Bordetella pertussis* Tohama I	*cyaB*	cyclolysin secretion ATP-binding protein	123
VFG1270	*Bordetella pertussis* Tohama I	*cyaD*	cyclolysin secretion protein	66
VFG1271	*Bordetella pertussis* Tohama I	*cyaE*	cyclolysin secretion protein	7
VFG1394	*Mycobacterium tuberculosis* H37Rv	*plcD*	plcD	14
VFG1400	*Mycobacterium tuberculosis* H37Rv	*plcC*	plcC	15
VFG1401	*Mycobacterium tuberculosis* H37Rv	*plcB*	plcB	13
VFG1402	*Mycobacterium tuberculosis* H37Rv	*plcA*	plcA	13
VFG1447	*Escherichia coli*	*cnf1*	cytotoxic necrotizing factor 1	5
VFG1798	*Staphylococcus aureus*	*hlb*	beta-hemolysin	6
VFG1801	*Staphylococcus aureus*	*etb*	exfoliative toxin B	3
VFG1827	*Shigella flexneri* (serotype 2a) 301	*sen*	enterotoxin	3

REFERENCE LIST

Abdallah, A.M., Gey van Pittius, N.C., Champion, P.A., Cox, J., Luirink, J., Vandenbroucke-Grauls, C.M., Appelmelk, B.J., and Bitter, W. 2007. Type VII secretion--*mycobacteria* show the way. *Nat. Rev. Microbiol.* 5: 883-891.

Ahuja, N., Kumar, P., and Bhatnagar, R. 2004. The adenylate cyclase toxins. *Crit. Rev. Microbiol.* 30: 187-196.

Aktories, K. and Barbieri, J.T. 2005. Bacterial cytotoxins: targeting eukaryotic switches. *Nat. Rev. Microbiol.* 3: 397-410.

Alba, M.M., Lee, D., Pearl, F.M., Shepherd, A.J., Martin, N., Orengo, C.A., and Kellam, P. 2001. VIDA: a virus database system for the organization of animal virus genome open reading frames. *Nucleic Acids Res.* 29: 133-136.

Allaoui, A., Scheen, R., Lambert de Rouvroit, C., and Cornelis, G.R. 1995. VirG, a Yersinia enterocolitica lipoprotein involved in Ca2+ dependency, is related to exsB of *Pseudomonas aeruginosa*. *J. Bacteriol.* 177: 4230-4237.

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

Anderson, G.W.,Jr, Leary, S.E., Williamson, E.D., Titball, R.W., Welkos, S.L., Worsham, P.L., and Friedlander, A.M. 1996. Recombinant V antigen protects mice against pneumonic and bubonic plague caused by F1-capsule-positive and -negative strains of *Yersinia pestis*. *Infect. Immun.* 64: 4580-4585.

Anisimova, M., Bielawski, J., Dunn, K., and Yang, Z. 2007. Phylogenomic analysis of natural selection pressure in *Streptococcus* genomes. *BMC Evol. Biol.* 7: 154.

Ascenzi, P., Visca, P., Ippolito, G., Spallarossa, A., Bolognesi, M., and Montecucco, C. 2002. Anthrax toxin: a tripartite lethal combination. *FEBS Lett.* 531: 384-388.

Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P., et al. 2003. PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.* 31: 400-402.

Backert, S. and Meyer, T.F. 2006. Type IV secretion systems and their effectors in bacterial pathogenesis. *Curr. Opin. Microbiol.* 9: 207-217.

Baker, M.D. and Acharya, K.R. 2004. Superantigens: structure-function relationships. *Int. J. Med. Microbiol.* 293: 529-537.

Barbieri, J.T. and Sun, J. 2004. *Pseudomonas aeruginosa* ExoS and ExoT. *Rev. Physiol. Biochem. Pharmacol.* 152: 79-92.

Begun, J., Sifri, C.D., Goldman, S., Calderwood, S.B., and Ausubel, F.M. 2005. *Staphylococcus aureus* virulence factors identified by using a high-throughput Caenorhabditis elegans-killing model. *Infect. Immun.* 73: 872-877.

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

Binder, S., Levitt, A.M., Sacks, J.J., and Hughes, J.M. 1999. Emerging infectious diseases: public health issues for the 21st century. *Science* 284: 1311-1313.

Bouwer, H.G. and Hinrichs, D.J. 1996. Cytotoxic-T-lymphocyte responses to epitopes of listeriolysin O and p60 following infection with Listeria monocytogenes. *Infect. Immun.* 64: 2515-2522.

Boyce, J.D., Cullen, P.A., and Adler, B. 2004. Genomic-scale analysis of bacterial gene and protein expression in the host. *Emerg. Infect. Dis.* 10: 1357-1362.

Boyd, E.F. and Brussow, H. 2002. Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. *Trends Microbiol.* 10: 521-529.

Brown, N.F., Wickham, M.E., Coombes, B.K., and Finlay, B.B. 2006. Crossing the line: selection and evolution of virulence traits. *PLoS Pathog.* 2: e42.

Casadevall, A. and Pirofski, L.A. 1999. Host-pathogen interactions: redefining the basic concepts of virulence and pathogenicity. *Infect. Immun.* 67: 3703-3713.

Cascales, E. and Christie, P.J. 2003. The versatile bacterial type IV secretion systems. *Nat. Rev. Microbiol.* 1: 137-149.

Cassel, D. and Pfeuffer, T. 1978. Mechanism of cholera toxin action: covalent modification of the guanyl nucleotide-binding protein of the adenylate cyclase system. *Proc. Natl. Acad. Sci. U. S. A.* 75: 2669-2673.

Centers for Disease Control and Prevention. 2002. *Staphylococcus aureus* resistant to vancomycin--United States, 2002. *MMWR Morb. Mortal. Wkly. Rep.* 51: 565-567.

Champion, O.L., Gaunt, M.W., Gundogdu, O., Elmi, A., Witney, A.A., Hinds, J., Dorrell, N., and Wren, B.W. 2005. Comparative phylogenomics of the food-borne pathogen *Campylobacter jejuni* reveals genetic markers predictive of infection source. *Proc. Natl. Acad. Sci. U. S. A.* 102: 16043-16048.

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., and Jin, Q. 2005. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* 33: D325-8.

Christie, P.J. 1997. *Agrobacterium tumefaciens* T-complex transport apparatus: a paradigm for a new family of multifunctional transporters in eubacteria. *J. Bacteriol.* 179: 3085-3094.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., and Bork, P. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* 311: 1283-1287.

Collier, R.J. 2001. Understanding the mode of action of diphtheria toxin: a perspective on progress during the 20th century. *Toxicon* 39: 1793-1803.

Collins, D.M., Skou, B., White, S., Bassett, S., Collins, L., For, R., Hurr, K., Hotter, G., and de Lisle, G.W. 2005. Generation of attenuated *Mycobacterium bovis* strains by signature-tagged mutagenesis for discovery of novel vaccine candidates. *Infect. Immun.* 73: 2379-2386.

Confer, D.L. and Eaton, J.W. 1982. Phagocyte impotence caused by an invasive bacterial adenylate cyclase. *Science* 217: 948-950.

Cornelis, G.R. 2006. The type III secretion injectisome. *Nat. Rev. Microbiol.* 4: 811-825.

Cornelis, G.R. 2002a. *Yersinia* type III secretion: send in the effectors. *J. Cell Biol.* 158: 401-408.

Cornelis, G.R. 2002b. The *Yersinia* Ysc-Yop 'type III' weaponry. *Nat. Rev. Mol. Cell Biol.* 3: 742-752.

Cornelis, G.R., Boland, A., Boyd, A.P., Geuijen, C., Iriarte, M., Neyt, C., Sory, M.P., and Stainier, I. 1998. The virulence plasmid of *Yersinia*, an antihost genome. *Microbiol. Mol. Biol. Rev.* 62: 1315-1352.

Cornelis, G.R. and Van Gijsegem, F. 2000. Assembly and function of type III secretory systems. *Annu. Rev. Microbiol.* 54: 735-774.

Cotter, P.A. and DiRita, V.J. 2000. Bacterial virulence gene regulation: an evolutionary perspective. *Annu. Rev. Microbiol.* 54: 519-565.

Dale, C., Jones, T., and Pontes, M. 2005. Degenerative evolution and functional diversification of type-III secretion systems in the insect endosymbiont *Sodalis glossinidius*. *Mol. Biol. Evol.* 22: 758-766.

Dale, C., Plague, G.R., Wang, B., Ochman, H., and Moran, N.A. 2002. Type III secretion systems and the evolution of mutualistic endosymbiosis. *Proc. Natl. Acad. Sci. U. S. A.* 99: 12397-12402.

Dale, C., Young, S.A., Haydon, D.T., and Welburn, S.C. 2001. The insect endosymbiont *Sodalis glossinidius* utilizes a type III secretion system for cell invasion. *Proc. Natl. Acad. Sci. U. S. A.* 98: 1883-1888.

Davies, D.G., Parsek, M.R., Pearson, J.P., Iglewski, B.H., Costerton, J.W., and Greenberg, E.P. 1998. The involvement of cell-to-cell signals in the development of a bacterial biofilm. *Science* 280: 295-298.

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

Dozois, C.M., Daigle, F., and Curtiss, R.,3rd. 2003. Identification of pathogen-specific and conserved genes expressed in vivo by an avian pathogenic *Escherichia coli* strain. *Proc. Natl. Acad. Sci. U. S. A.* 100: 247-252.

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

Falkow, S. 1988. Molecular Koch's postulates applied to microbial pathogenicity. *Rev. Infect. Dis.* 10 Suppl 2: S274-6.

Fields, P.I., Swanson, R.V., Haidaris, C.G., and Heffron, F. 1986. Mutants of *Salmonella typhimurium* that cannot survive within the macrophage are avirulent. *Proc. Natl. Acad. Sci. U. S. A.* 83: 5189-5193.

Finlay, B.B. and Falkow, S. 1997. Common themes in microbial pathogenicity revisited. *Microbiol. Mol. Biol. Rev.* 61: 136-169.

Fulton, D.L., Li, Y.Y., Laird, M.R., Horsman, B.G., Roche, F.M., and Brinkman, F.S. 2006. Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics* 7: 270.

Galan, J.E. and Collmer, A. 1999. Type III secretion machines: bacterial devices for protein delivery into host cells. *Science* 284: 1322-1328.

Gal-Mor, O. and Finlay, B.B. 2006. Pathogenicity islands: a molecular toolbox for bacterial virulence. *Cell. Microbiol.* 8: 1707-1719.

Gandon, S., Mackinnon, M., Nee, S., and Read, A. 2003. Imperfect vaccination: some epidemiological and evolutionary consequences. *Proc. Biol. Sci.* 270: 1129-1136.

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

Geoffroy, C., Gaillard, J.L., Alouf, J.E., and Berche, P. 1987. Purification, characterization, and toxicity of the sulfhydryl-activated hemolysin listeriolysin O from *Listeria monocytogenes*. *Infect. Immun.* 55: 1641-1646.

Gerlach, R.G. and Hensel, M. 2007. Protein secretion systems and adhesins: the molecular armory of Gram-negative pathogens. *Int. J. Med. Microbiol.* 297: 401-415.

Goldberg, J.B., Coyne, M.J.,Jr, Neely, A.N., and Holder, I.A. 1995. Avirulence of a *Pseudomonas aeruginosa* algC mutant in a burned-mouse model of infection. *Infect. Immun.* 63: 4166-4169.

Gottfert, M., Rothlisberger, S., Kundig, C., Beck, C., Marty, R., and Hennecke, H. 2001. Potential symbiosis-specific genes uncovered by sequencing a 410-kilobase DNA region of the *Bradyrhizobium japonicum* chromosome. *J. Bacteriol.* 183: 1405-1412.

Grant, A.J., Coward, C., Jones, M.A., Woodall, C.A., Barrow, P.A., and Maskell, D.J. 2005. Signature-tagged transposon mutagenesis studies demonstrate the dynamic nature of cecal colonization of 2-week-old chickens by *Campylobacter jejuni*. *Appl. Environ. Microbiol.* 71: 8031-8041.

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

Hacker, J., Bender, L., Ott, M., Wingender, J., Lund, B., Marre, R., and Goebel, W. 1990. Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal *Escherichia coli* isolates. *Microb. Pathog.* 8: 213-225.

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

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

Haft, D.H., Selengut, J.D., Brinkac, L.M., Zafar, N., and White, O. 2005. Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics* 21: 293-306.

Harty, J.T. and Bevan, M.J. 1992. CD8+ T cells specific for a single nonamer epitope of *Listeria monocytogenes* are protective in vivo. *J. Exp. Med.* 175: 1531-1538.

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

He, S.Y., Nomura, K., and Whittam, T.S. 2004. Type III protein secretion mechanism in mammalian and plant pathogens. *Biochim. Biophys. Acta* 1694: 181-206.

Heesemann, J., Gross, U., Schmidt, N., and Laufs, R. 1986. Immunochemical analysis of plasmid-encoded proteins released by enteropathogenic *Yersinia* sp. grown in calcium-deficient media. *Infect. Immun.* 54: 561-567.

Heidelberg, J.F., Seshadri, R., Haveman, S.A., Hemme, C.L., Paulsen, I.T., Kolonay, J.F., Eisen, J.A., Ward, N., Methe, B., Brinkac, L.M., et al. 2004. The genome sequence of the anaerobic, sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough. *Nat. Biotechnol.* 22: 554-559.

Henderson, I.R., Navarro-Garcia, F., Desvaux, M., Fernandez, R.C., and Ala'Aldeen, D. 2004. Type V protein secretion pathway: the autotransporter story. *Microbiol. Mol. Biol. Rev.* 68: 692-744.

Hensel, M., Shea, J.E., Gleeson, C., Jones, M.D., Dalton, E., and Holden, D.W. 1995. Simultaneous identification of bacterial virulence genes by negative selection. *Science* 269: 400-403.

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

Herman, A., Kappler, J.W., Marrack, P., and Pullen, A.M. 1991. Superantigens: mechanism of T-cell stimulation and role in immune responses. *Annu. Rev. Immunol.* 9: 745-772.

Holden, M., Crossman, L., Cerdeno-Tarraga, A., and Parkhill, J. 2004. Pathogenomics of non-pathogens. *Nat. Rev. Microbiol.* 2: 91.

Holland, I.B., Schmitt, L., and Young, J. 2005. Type 1 protein secretion in bacteria, the ABC-transporter dependent pathway (review). *Mol. Membr. Biol.* 22: 29-39.

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

Hsiao, W., Wan, I., Jones, S.J., and Brinkman, F.S. 2003. IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics* 19: 418-420.

Hsiao, W.W., Ung, K., Aeschliman, D., Bryan, J., Finlay, B.B., and Brinkman, F.S. 2005. Evidence of a Large Novel Gene Pool Associated with Prokaryotic Genomic Islands. *PLoS Genet.* 1: e62.

Hueck, C.J. 1998. Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiol. Mol. Biol. Rev.* 62: 379-433.

Hung, D.T., Shakhnovich, E.A., Pierson, E., and Mekalanos, J.J. 2005. Small-molecule inhibitor of *Vibrio cholerae* virulence and intestinal colonization. *Science* 310: 670-674.

Huynen, M.A., Diaz-Lazcoz, Y., and Bork, P. 1997. Differential genome display. *Trends Genet.* 13: 389-390.

Isberg, R.R., Voorhis, D.L., and Falkow, S. 1987. Identification of invasin: a protein that allows enteric bacteria to penetrate cultured mammalian cells. *Cell* 50: 769-778.

Jenner, E. 1800. An inquiry into the causes and effects of the variolæ vaccinæ, a disease discovered in some of the western counties of England, ... and known by the name of the cow pox. By Edward Jenner, M.D.F.R.S.&c.

Jeong, H., Yim, J.H., Lee, C., Choi, S.H., Park, Y.K., Yoon, S.H., Hur, C.G., Kang, H.Y., Kim, D., Lee, H.H., et al. 2005. Genomic blueprint of *Hahella chejuensis*, a marine microbe producing an algicidal agent. *Nucleic Acids Res.* 33: 7066-7073.

Jernigan, D.B., Raghunathan, P.L., Bell, B.P., Brechner, R., Bresnitz, E.A., Butler, J.C., Cetron, M., Cohen, M., Doyle, T., Fischer, M., et al. 2002. Investigation of bioterrorism-related anthrax, United States, 2001: epidemiologic findings. *Emerg. Infect. Dis.* 8: 1019-1028.

Jerse, A.E., Gicquelais, K.G., and Kaper, J.B. 1991. Plasmid and chromosomal elements involved in the pathogenesis of attaching and effacing *Escherichia coli*. *Infect. Immun.* 59: 3869-3875.

Jerse, A.E., Yu, J., Tall, B.D., and Kaper, J.B. 1990. A genetic locus of enteropathogenic *Escherichia coli* necessary for the production of attaching and effacing lesions on tissue culture cells. *Proc. Natl. Acad. Sci. U. S. A.* 87: 7839-7843.

Johnson, T.L., Abendroth, J., Hol, W.G., and Sandkvist, M. 2006. Type II secretion: from structure to function. *FEMS Microbiol. Lett.* 255: 175-186.

Jores, J., Rumer, L., and Wieler, L.H. 2004. Impact of the locus of enterocyte effacement pathogenicity island on the evolution of pathogenic *Escherichia coli*. *Int. J. Med. Microbiol.* 294: 103-113.

Jungo, F. and Bairoch, A. 2005. Tox-Prot, the toxin protein annotation program of the Swiss-Prot protein knowledgebase. *Toxicon* 45: 293-301.

Kalman, S., Mitchell, W., Marathe, R., Lammel, C., Fan, J., Hyman, R.W., Olinger, L., Grimwood, J., Davis, R.W., and Stephens, R.S. 1999. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nat. Genet.* 21: 385-389.

Katada, T. and Ui, M. 1982. Direct modification of the membrane adenylate cyclase system by islet-activating protein due to ADP-ribosylation of a membrane protein. *Proc. Natl. Acad. Sci. U. S. A.* 79: 3129-3133.

Knapp, S., Hacker, J., Jarchau, T., and Goebel, W. 1986. Large, unstable inserts in the chromosome affect virulence properties of uropathogenic *Escherichia coli* O6 strain 536. *J. Bacteriol.* 168: 22-30.

Kotzin, B.L., Leung, D.Y., Kappler, J., and Marrack, P. 1993. Superantigens and their potential role in human disease. *Adv. Immunol.* 54: 99-166.

Ku, Y.W., McDonough, S.P., Palaniappan, R.U., Chang, C.F., and Chang, Y.F. 2005. Novel attenuated *Salmonella enterica* serovar Choleraesuis strains as live vaccine candidates generated by signature-tagged mutagenesis. *Infect. Immun.* 73: 8194-8203.

Kuhnert, P., Heyberger-Meyer, B., Burnens, A.P., Nicolet, J., and Frey, J. 1997. Detection of RTX toxin genes in gram-negative bacteria with a set of specific probes. *Appl. Environ. Microbiol.* 63: 2258-2265.

Ladant, D. and Ullmann, A. 1999. *Bordatella pertussis* adenylate cyclase: a toxin with multiple talents. *Trends Microbiol.* 7: 172-176.

Langille, M.G.I., Hsiao, W.W.L., and Brinkman, F.S.L. unpublished. Evaluation of genomic island predictors using a comparative genomics approach.

Langille, M.G.I., Zhou, F., Fedynak, A., Hsiao, W.W.L., Xu, Y., and Brinkman, F.S.L. 2008. *Mobile genetic elements and their predictions. In Press. In Xing Yu and J. Peter Gogarten (eds.), Computational Methods for Understanding Bacterial and Archaeal Genomes.* World Scientific Publishing, USA, .

Lawlor, M.S., Hsu, J., Rick, P.D., and Miller, V.L. 2005. Identification of *Klebsiella pneumoniae* virulence determinants using an intranasal infection model. *Mol. Microbiol.* 58: 1054-1073.

Leary, S.E., Williamson, E.D., Griffin, K.F., Russell, P., Eley, S.M., and Titball, R.W. 1995. Active immunization with recombinant V antigen from Yer*sinia pestis* protects mice against plague. *Infect. Immun.* 63: 2854-2858.

Li, G., Laturnus, C., Ewers, C., and Wieler, L.H. 2005. Identification of genes required for avian *Escherichia coli* septicemia by signature-tagged mutagenesis. *Infect. Immun.* 73: 2818-2827.

Macdonald-Fyall, J., Xing, D., Corbel, M., Baillie, S., Parton, R., and Coote, J. 2004. Adjuvanticity of native and detoxified adenylate cyclase toxin of Bordetella pertussis towards co-administered antigens. *Vaccine* 22: 4270-4281.

Mahan, M.J., Slauch, J.M., and Mekalanos, J.J. 1993. Selection of bacterial virulence genes that are specifically induced in host tissues. *Science* 259: 686-688.

Mantri, Y. and Williams, K.P. 2004. Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. *Nucleic Acids Res.* 32: D55-8.

Marenne, M.N., Journet, L., Mota, L.J., and Cornelis, G.R. 2003. Genetic analysis of the formation of the Ysc-Yop translocation pore in macrophages by *Yersinia enterocolitica*: role of LcrV, YscF and YopN. *Microb. Pathog.* 35: 243-258.

Marrack, P. and Kappler, J. 1990. The staphylococcal enterotoxins and their relatives. *Science* 248: 1066.

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

Mazurier, S., Lemunier, M., Hartmann, A., Siblot, S., and Lemanceau, P. 2006. Conservation of type III secretion system genes in *Bradyrhizobium* isolated from soybean. *FEMS Microbiol. Lett.* 259: 317-325.

Mazurkiewicz, P., Tang, C.M., Boone, C., and Holden, D.W. 2006. Signature-tagged mutagenesis: barcoding mutants for genome-wide screens. *Nat. Rev. Genet.* 7: 929-939.

Merkl, R. 2004. SIGI: score-based identification of genomic islands. *BMC Bioinformatics* 5: 22.

Morens, D.M., Folkers, G.K., and Fauci, A.S. 2004. The challenge of emerging and re-emerging infectious diseases. *Nature* 430: 242-249.

Mota, L.J., Sorg, I., and Cornelis, G.R. 2005. Type III secretion: the bacteria-eukaryotic cell express. *FEMS Microbiol. Lett.* 252: 1-10.

Mougous, J.D., Cuff, M.E., Raunser, S., Shen, A., Zhou, M., Gifford, C.A., Goodman, A.L., Joachimiak, G., Ordonez, C.L., Lory, S., et al. 2006. A virulence locus of *Pseudomonas aeruginosa* encodes a protein secretion apparatus. *Science* 312: 1526-1530.

Neyt, C. and Cornelis, G.R. 1999. Insertion of a Yop translocation pore into the macrophage plasma membrane by *Yersinia enterocolitica*: requirement for translocators YopB and YopD, but not LcrG. *Mol. Microbiol.* 33: 971-981.

Ochman, H. and Moran, N.A. 2001. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* 292: 1096-1099.

Ojha, S., Sirois, M., and Macinnes, J.I. 2005. Identification of *Actinobacillus suis* genes essential for the colonization of the upper respiratory tract of swine. *Infect. Immun.* 73: 7032-7039.

Orr, B., Douce, G., Baillie, S., Parton, R., and Coote, J. 2007. Adjuvant effects of adenylate cyclase toxin of *Bordetella pertussis* after intranasal immunisation of mice. *Vaccine* 25: 64-71.

Paik, S., Senty, L., Das, S., Noe, J.C., Munro, C.L., and Kitten, T. 2005. Identification of virulence determinants for endocarditis in *Streptococcus sanguinis* by signature-tagged mutagenesis. *Infect. Immun.* 73: 6064-6074.

Pallen, M.J. and Wren, B.W. 2007. Bacterial pathogenomics. *Nature* 449: 835-842.

Pasteur, L. 1880. De l'attenuation du virus du cholera des poules. *C. R. Acad. Sci. Paris* 91: 673.

Pennisi, E. 2004. The Biology of Genomes meeting. Surveys reveal vast numbers of genes. *Science* 304: 1591.

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

Pizza, M., Covacci, A., Bartoloni, A., Perugini, M., Nencioni, L., De Magistris, M.T., Villa, L., Nucci, D., Manetti, R., and Bugnoli, M. 1989. Mutants of pertussis toxin suitable for vaccine development. *Science* 246: 497-500.

Pizza, M., Scarlato, V., Masignani, V., Giuliani, M.M., Arico, B., Comanducci, M., Jennings, G.T., Baldi, L., Bartolini, E., Capecchi, B., et al. 2000. Identification of vaccine candidates against serogroup B *meningococcus* by whole-genome sequencing. *Science* 287: 1816-1820.

Plotkin, S.A. 2005. Vaccines: past, present and future. *Nat. Med.* 11: S5-11.

Pohlner, J., Halter, R., Beyreuther, K., and Meyer, T.F. 1987. Gene structure and extracellular secretion of *Neisseria gonorrhoeae* IgA protease. *Nature* 325: 458-462.

Pollack, J.R. and Neilands, J.B. 1970. Enterobactin, an iron transport compound from *Salmonella typhimurium*. *Biochem. Biophys. Res. Commun.* 38: 989-992.

Portnoy, D.A., Jacks, P.S., and Hinrichs, D.J. 1988. Role of hemolysin for the intracellular growth of *Listeria monocytogenes*. *J. Exp. Med.* 167: 1459-1471.

Potvin, E., Lehoux, D.E., Kukavica-Ibrulj, I., Richard, K.L., Sanschagrin, F., Lau, G.W., and Levesque, R.C. 2003. In vivo functional genomics of Pseudomonas aeruginosa for high-throughput screening of new virulence factors and antibacterial targets. *Environ. Microbiol.* 5: 1294-1308.

Pukatzki, S., Ma, A.T., Sturtevant, D., Krastins, B., Sarracino, D., Nelson, W.C., Heidelberg, J.F., and Mekalanos, J.J. 2006. Identification of a conserved bacterial protein secretion system in *Vibrio cholerae* using the Dictyostelium host model system. *Proc. Natl. Acad. Sci. U. S. A.* 103: 1528-1533.

Russmann, H. 2004. Inverted pathogenicity: the use of pathogen-specific molecular mechanisms for prevention or therapy of disease. *Int. J. Med. Microbiol.* 293: 565-569.

Saenz, H.L. and Dehio, C. 2005. Signature-tagged mutagenesis: technical advances in a negative selection method for virulence gene identification. *Curr. Opin. Microbiol.* 8: 612-619.

Saier, M.H.,Jr. 2004. Evolution of bacterial type III protein secretion systems. *Trends Microbiol.* 12: 113-115.

Salyers, A.A. and Whitt, D.D. 2002. *Bacterial pathogenesis : a molecular approach.* ASM Press, Washington, D.C.

Scaria, J., Chandramouli, U., and Verma, S.K. 2005. Antibiotic Resistance Genes Online (ARGO): a Database on vancomycin and beta-lactam resistance genes. *Bioinformation* 1: 5-7.

Schmidt, H. and Hensel, M. 2004. Pathogenicity islands in bacterial pathogenesis. *Clin. Microbiol. Rev.* 17: 14-56.

Shah, D.H., Lee, M.J., Park, J.H., Lee, J.H., Eo, S.K., Kwon, J.T., and Chae, J.S. 2005. Identification of *Salmonella gallinarum* virulence genes in a chicken infection model using PCR-based signature-tagged mutagenesis. *Microbiology* 151: 3957-3968.

Shankar, N., Baghdayan, A.S., and Gilmore, M.S. 2002. Modulation of virulence within a pathogenicity island in vancomycin-resistant *Enterococcus faecalis*. *Nature* 417: 746-750.

Shukla, H.D. and Sharma, S.K. 2005. *Clostridium botulinum*: a bug with beauty and weapon. *Crit. Rev. Microbiol.* 31: 11-18.

Simpson, J.A., Smith, S.E., and Dean, R.T. 1988. Alginate inhibition of the uptake of Pseudomonas aeruginosa by macrophages. *J. Gen. Microbiol.* 134: 29-36.

Smith, J. 2001. The social evolution of bacterial pathogenesis. *Proc. Biol. Sci.* 268: 61-69.

Snyder, L.A. and Saunders, N.J. 2006. The majority of genes in the pathogenic *Neisseria* species are present in non-pathogenic *Neisseria lactamica*, including those designated as virulence genes. *BMC Genomics* 7: 128.

Sokurenko, E.V., Gomulkiewicz, R., and Dykhuizen, D.E. 2006. Source-sink dynamics of virulence evolution. *Nat. Rev. Microbiol.* 4: 548-555.

Srinivasan, K.N., Gopalakrishnakone, P., Tan, P.T., Chew, K.C., Cheng, B., Kini, R.M., Koh, J.L., Seah, S.H., and Brusic, V. 2002. SCORPION, a molecular database of scorpion toxins. *Toxicon* 40: 23-31.

Stabler, R.A., Marsden, G.L., Witney, A.A., Li, Y., Bentley, S.D., Tang, C.M., and Hinds, J. 2005. Identification of pathogen-specific genes through microarray analysis of pathogenic and commensal *Neisseria* species. *Microbiology* 151: 2907-2922.

Sullivan, J.T. and Ronson, C.W. 1998. Evolution of *rhizobia* by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. *Proc. Natl. Acad. Sci. U. S. A.* 95: 5145-5149.

Tampakaki, A.P., Fadouloglou, V.E., Gazi, A.D., Panopoulos, N.J., and Kokkinidis, M. 2004. Conserved features of type III secretion. *Cell. Microbiol.* 6: 805-816.

Tang, C.M., Hood, D.W., and Moxon, E.R. 1998. Microbial genome sequencing and pathogenesis. *Curr. Opin. Microbiol.* 1: 12-16.

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

Taylor, R.K., Miller, V.L., Furlong, D.B., and Mekalanos, J.J. 1987. Use of phoA gene fusions to identify a pilus colonization factor coordinately regulated with cholera toxin. *Proc. Natl. Acad. Sci. U. S. A.* 84: 2833-2837.

Tobe, T. and Sasakawa, C. 2002. Species-specific cell adhesion of enteropathogenic *Escherichia coli* is mediated by type IV bundle-forming pili. *Cell. Microbiol.* 4: 29-42.

Toh, H., Weiss, B.L., Perkin, S.A., Yamashita, A., Oshima, K., Hattori, M., and Aksoy, S. 2006. Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Res.* 16: 149-156.

Tweten, R.K. 2005. Cholesterol-dependent cytolysins, a family of versatile pore-forming toxins. *Infect. Immun.* 73: 6199-6209.

van Diemen, P.M., Dziva, F., Stevens, M.P., and Wallis, T.S. 2005. Identification of enterohemorrhagic *Escherichia coli* O26:H- genes required for intestinal colonization in calves. *Infect. Immun.* 73: 1735-1743.

Waalwijk, C., MacLaren, D.M., and de Graaff, J. 1983. In vivo function of hemolysin in the nephropathogenicity of *Escherichia coli*. *Infect. Immun.* 42: 245-249.

Wassenaar, T.M. and Gaastra, W. 2001. Bacterial virulence: can we draw the line? *FEMS Microbiol. Lett.* 201: 1-7.

Way, S.S. and Wilson, C.B. 2005. The *Mycobacterium tuberculosis* ESAT-6 homologue in Listeria monocytogenes is dispensable for growth in vitro and in vivo. *Infect. Immun.* 73: 6151-6153.

Wheeler, D.L., Chappey, C., Lash, A.E., Leipe, D.D., Madden, T.L., Schuler, G.D., Tatusova, T.A., and Rapp, B.A. 2000. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 28: 10-14.

Whiteley, M., Lee, K.M., and Greenberg, E.P. 1999. Identification of genes controlled by quorum sensing in *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U. S. A.* 96: 13904-13909.

Whittam, T.S. and Bumbaugh, A.C. 2002. Inferences from whole-genome sequences of bacterial pathogens. *Curr. Opin. Genet. Dev.* 12: 719-725.

Wilson, B.A. and Salyers, A.A. 2003. Is the evolution of bacterial pathogens an out-of-body experience? *Trends Microbiol.* 11: 347-350.

Wizemann, T.M., Heinrichs, J.H., Adamou, J.E., Erwin, A.L., Kunsch, C., Choi, G.H., Barash, S.C., Rosen, C.A., Masure, H.R., Tuomanen, E., et al. 2001. Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. *Infect. Immun.* 69: 1593-1598.

Wooldridge, K.G. and Williams, P.H. 1993. Iron uptake mechanisms of pathogenic bacteria. *FEMS Microbiol. Rev.* 12: 325-348.

Yoon, S.H., Hur, C.G., Kang, H.Y., Kim, Y.H., Oh, T.K., and Kim, J.F. 2005. A computational approach for identifying pathogenicity islands in prokaryotic genomes. *BMC Bioinformatics* 6: 184.

Zhang, R. and Zhang, C.T. 2006. The impact of comparative genomics on infectious disease research. *Microbes Infect.* 8: 1613-1622.

Zhang, R. and Zhang, C.T. 2004. A systematic method to identify genomic islands and its applications in analyzing the genomes of Corynebacterium glutamicum and Vibrio vulnificus CMCP6 chromosome I. *Bioinformatics* 20: 612-622.

Zhang, S., Chen, Y., Potvin, E., Sanschagrin, F., Levesque, R.C., McCormack, F.X., and Lau, G.W. 2005. Comparative signature-tagged mutagenesis identifies *Pseudomonas* factors conferring resistance to the pulmonary collectin SP-A. *PLoS Pathog.* 1: 259-268.

Zhang, Y.Q., Ren, S.X., Li, H.L., Wang, Y.X., Fu, G., Yang, J., Qin, Z.Q., Miao, Y.G., Wang, W.Y., Chen, R.S., et al. 2003. Genome-based analysis of virulence genes in a non-biofilm-forming *Staphylococcus epidermidis* strain (ATCC 12228). *Mol. Microbiol.* 49: 1577-1593.

Zhou, C.E., Smith, J., Lam, M., Zemla, A., Dyer, M.D., and Slezak, T. 2007. MvirDB--a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res.* 35: D391-4.

Zhou, D. and Galan, J. 2001. *Salmonella* entry into host cells: the work in concert of type III secreted effector proteins. *Microbes Infect.* 3: 1293-1298.