Studies of the genome and regulatory processes of *Vibrio parahaemolyticus*

Saylem M. Ingalls

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Masters of Science
in
Biological Sciences

Ann M. Stevens, Chair
Timothy J. Larson
David L. Popham
Richard A. Walker

December 8th, 2010

Blacksburg, VA

Keywords: *Vibrio parahaemolyticus*, genome sequencing, annotation, OpaR, CsrA

Studies of the genome and regulatory processes of *Vibrio parahaemolyticus*

Saylem M. Ingalls

Abstract

Vibrio parahaemolyticus is considered to be an emerging, yet understudied, human pathogen. The *V. parahaemolyticus* BB22OP genome was sequenced to allow for a comparative analysis between the genome of BB22OP and another previously sequenced, pathogenic strain of *V. parahaemolyticus*, RIMD2210633. *V. parahaemolyticus* BB22OP is interesting because it exhibits a spontaneous phenotypic switch in colony morphology due to the loss of a functional OpaR; this also influences virulence. OpaR is the major quorum-sensing regulator in *V. parahaemolyticus* homologous to LuxR from *V. harveyi*. When *opaR* is removed from the RIMD2210633 genome, the same phenotypic switch is not seen indicating a difference between the quorum-sensing systems in these two strains. Understanding the regulatory variation in these two strains has the potential to provide key insights into the control of pathogenesis in this organism.

Initially, the BB22OP genome sequencing results aligned into 125 contigs. The genome has now been assembled into two distinct chromosomes with only two gaps remaining to be filled. These gaps are located in the integron region, which is difficult to assemble due to its structure. The integron is a series of gene cassettes separated by inverted repeats that facilitate recombination events that build the integron. The integron region is further evidence of genetic differences between the two strains. The integron in the RIMD2210633 strain is comprised of 69 gene cassettes, while the BB22OP integron contains at least 86 gene cassettes. There are 313 genes novel to the BB22OP genome, which could result in the phenotypic differences seen in these two strains. Additionally five of the 313 genes are predicted to be transcriptional regulators indicating the potential for differential gene regulation. Further comparative analysis will likely reveal more phenotypic divergence between the physiology of RIMD2210633 and BB22OP.

Additionally, the CsrA regulatory network was explored in RIMD2210633. CsrA was first characterized in *E. coli* as a global regulator of carbon storage and metabolism. RIMD2210633 contains a CsrA homolog and was predicted to contain four CsrA-regulating sRNAs (CsrB1-3 and CsrC), and this work confirmed that these sRNAs regulate CsrA in the same manner as in *E. coli*. CsrA and the same CsrA-regulating sRNAs were found in the BB22OP genome as well. Since CsrA is known to regulate glycogen production, a qualitative iodine-staining plate assay and a quantitative glycogen assay were used to indirectly measure CsrA activity in the presence and absence of individual regulatory sRNAs. The RIMD2210633 CsrA, CsrB1, CsrB2, CsrB3 and CsrC were shown to have the predicted physiological role in recombinant *E. coli*, with higher glycogen levels observed when CsrA was active and lower levels when each of the sRNAs was overexpressed. CsrA is also known to regulate biofilm production and virulence factors. In an attempt to develop a screening method for potential CsrA targets, a transcriptional/translational fusion system was developed. Transcriptional and translational fusions to β-galactosidase were created to P_{dksA}, P_{glgC1} and P_{toxR} from RIMD2210633. CsrA or CsrB2 was overexpressed in recombinant *E. coli* containing each of the fusion constructs in order to see what happens to the gene expression from these promoters at low and high CsrA activity levels. Surprisingly, changing the activity levels of CsrA impacted both transcriptional and translational levels making the results of the assay difficult to interpret.

Collectively these efforts have enhanced our understanding of *V. parahaemolyticus*. In particular, the sequencing of BB22OP has allowed for a comparative analysis between the BB22OP and RIMD2210633 strains. These strains have remarkably conserved genomes despite the phenotypic differences they exhibit. It appears there is variation in the quorum-sensing systems of these two strains. Further analysis will reveal how the quorum-sensing regulons differ and how this impacts the virulence of these two pathogenic *V. parahaemolyticus* strains.

Acknowledgements

First I would like to thank my advisor, Dr. Ann Stevens, for her constant patience, understanding and guidance throughout this process. I would also like to thank the other members of my committee, Drs. David Popham, Tim Larson and Richard Walker who have been an immense source of advice with my research project.

I'd like to thank our collaborators Dr. Roderick Jensen, who provided a tremendous amount of help and support in regards to the genome-sequencing project, and Dr. Linda McCarter who's knowledge has been immensely helpful. Without their expertise, the completion of this project would not have been possible.

I would like to thank my colleagues in the Stevens' lab for providing a constantly entertaining and cooperative work environment. Specifically I'd like to thank Dr. Joshua Williams for helping get this project off the ground. I would like to especially thank Revathy Ramachandran for her constant patience when listening to my questions and doubts and for being a good friend.

Finally, I would like to thank my friends and family for constantly encouraging me, believing in me and acting as a sounding board when I get frustrated – which they will gladly remind you is more often than not.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The pathogen, *Vibrio parahaemolyticus*

Vibrio parahaemolyticus is a Gram-negative marine bacterium capable of causing foodborne gastroenteritis in humans. This organism is an emerging pathogen and a leading cause of seafood-associated gastroenteritis in humans in the United States (Daniels et al, 2000). Illness caused by this organism is normally self-limiting, but there is epidemiological evidence that immunocompromised individuals are more susceptible to developing potentially lethal septicemia (FDA, 2001). *V. parahaemolyticus* is also capable of causing wound infections. After hurricane Katrina, 22 individuals developed wound-associated *Vibrio* infections. Three of these were attributed to *V. parahaemolyticus*, and two were fatal (CDC, 2005). In 1996, a *V. parahaemolyticus* strain, serotype 03:K6, emerged in Calcutta, India. Following this outbreak, strains of the identical serotype emerged pandemically in Southeast Asia, Japan and the United States (Okuda, et al. 1997; CDC, 1999)

V. parahaemolyticus strain RIMD2210633 is a pathogenic clinical isolate of the pandemic serotype 03:K6, capable of causing gastroenteritis and travellers' diarrhea. All strains of the serotype 03:K6 contain a common plasmid, pO3K6, with similar gene arrangement to a filamentous phage known to be associated with *V. parahaemolyticus*. It was determined that pO3K6 is a replicative form of the phage genome, containing an open reading frame, ORF8, unique to O3:K6 strains isolated after 1996 (Nasu et al, 2000). The genome of RIMD2210633 has been sequenced, facilitating a systems level analysis of the organism's virulence and colonization

capabilities. The genome of RIMD2210633 is 5.16 Mb arranged in two chromosomes containing approximately 4800 genes (Makino, et al 2003).

The disease mechanisms of *V. parahaemolyticus* are not fully understood, but are known to be distinct from those of other pathogenic *Vibrio* species. To date, several key virulence factors have been identified. (1) The *V. parahaemolyticus* RIMD genome contains two sets of a type III secretion system (TTSS) gene clusters (Honda et al, 2008; Makino et al, 2003). (2) *V. parahaemolyticus* also utilizes a thermostable direct hemolysin (TDH) to form pores in the host cell membrane (Honda et al, 1992). (3) Some strains, including RIMD2210633, contain an additional hemolysin, TDH-related hemolysin (Sochard and Colwell, 1977). (4) ToxR is a key regulator involved in transcriptional control of virulence genes among *Vibrio* species. For example, ToxR regulates the cholera toxin operon in *Vibrio cholerae* (Osorio and Klose, 2000; Miller et al, 1987). It is a member of the AraC family of transcriptional regulators and is well distributed among *Vibrio* species, including *V. parahaemolyticus*. Frequently, *toxR* is the target of PCR methods to identify *Vibrio* species (Bauer and Rørvik, 2007; Kim et al, 1999).

V. parahaemolyticus strain BB22OP is the other strain under investigation in this work. BB22OP is a pre-1983 environmental isolate from Bangaldesh and is the best genetically characterized strain. It is capable of a switch in colony morphology from opaque to translucent in response to the loss of functional OpaR, resulting in altered biofilm production (McCarter, 1998; Figure 1.1). The opaque strain colonies are smaller in size and sticky. When touched with a toothpick, the entire colony will

come off the agar in one piece. The translucent colonies are larger and mucoid. The translucent strain becomes pathogenic, whereas the opaque strain is avirulent. The loss of functional OpaR can be due to a spontaneous loss of function mutation in *opaR* or a spontaneous mutation in *luxO* resulting in a hyperstimulated LuxO (McCarter, 1998). OpaR and LuxO are two constituents of the *V. parahaemolyticus* quorum-sensing system (Figure 1.2). Interestingly, when *opaR* is deleted from RIMD2210633, the same phenotype is not seen (Linda McCarter, personal communication). Chapter Two describes the genome sequencing, assembly and annotation process for BB22OP. The genome sequence will allow for a comparative analysis between BB22OP and RIMD2210633 so that the mutually conserved and novel attributes of each strain may be defined. In particular, there appears to be differences in the quorum-sensing networks of the two strains.

Quorum Sensing in Vibrios

Quorum sensing is a method of cell-cell communication in bacteria. Bacteria capable of quorum sensing release chemical signal molecules as they grow and divide. Therefore, the extracellular concentration of the signals, called autoinducers, increases with the cell density of the population. At some threshold concentration, the autoinducer signal induces differential gene expression in the population. This allows the bacterial population to coordinate activities beneficial to the population as a whole that may be impossible for a single organism to accomplish alone (reviewed in Waters and Bassler, 2005). Quorum sensing has been implicated in the

expression of genes involved in virulence and biofilm formation (Fuqua et al, 2001, Reading and Sperandio, 2006).

In Gram-negative proteobacteria, the autoinducer signal molecule is most commonly an acylated homoserine lactone (Visick and Fuqua, 2005). The paradigm for quorum sensing in Gram-negative bacteria is the *Vibrio fischeri* system, which was first described in 1970 by Nealson, Platt and Hastings (Nealson et al, 1970). However, the quorum-sensing system in *V. parahaemolyticus* is more similar to the *Vibrio harveyi* model (Figure 1.2). The *V. harveyi* quorum-sensing system produces three autoinducer molecules via synthase enzymes, which are then recognized by cell surface receptors. The signals from these three autoinducer molecules culminate in a single regulatory pathway consisting of a two-component phospho-relay system that includes LuxU and LuxO. LuxU is phosphorylated by the autoinducer receptors and, in turn, phosphorylates LuxO. Phosphorylated LuxO activates the expression of five sRNAs, Qrr1-5. These sRNAs work in conjunction with Hfq to inhibit the expression of $LuxR_{Vh}$, the major quorum sensing regulator in *V. harveyi*, by binding $luxR_{Vh}$ mRNA. *V. parahaemolyticus* contains homologs (Figure 1.2) to all of the major regulatory proteins found in the *V. harveyi* system (reviewed in Waters and Bassler, 2005).

OpaR System

OpaR is the major quorum-sensing regulator in *V. parahaemolyticus* and is homologous to *V. harveyi* $LuxR_{Vh}$. $LuxR_{Vh}$ homologs have been found in a variety of other *Vibrio* species, including *V. cholerae*, *V. vulnificus*, *V. angustum*, *V. anguillarium*,

and *V. alginolyticus* (McDougald et al, 2000). $LuxR_{Vh}$ homologs differ from the *Vibrio fischeri* LuxR homologs in that they do not directly bind autoinducer. $LuxR_{Vh}$ homologs respond to the phosphorelay cascade in the quorum-sensing system. The structure of SmcR, the $LuxR_{Vh}$ homolog and major quorum-sensing regulator in *V. vulnificus*, was solved at 2.1 Å resolution. The structure reveals SmcR shares structural similarity to the TetR superfamily, containing a DNA binding domain at the N-terminus and a dimerization domain at the C-terminus. There is evidence that SmcR may have the ability to recognize different target promoters with differing affinities, which suggests an additional level of sophistication to this protein family's ability to regulate gene expression (Kim et al, 2010). While the TetR family of proteins are primarily considered repressors, it has been demonstrated that $LuxR_{Vh}$ and SmcR function as both an activator and a repressor (Pompeani et al, 2008; Kim et al, 2010).

OpaR was first discovered in a screen of a cosmid gene bank of the *V. parahaemolyticus* BB22OP strain genome looking for genes with the ability to induce light expression from the *V. harveyi* luminescence operon *luxCDABE*. The gene responsible for inducing light expression was identified and found to encode a protein with 96% amino acid identity to the *V. harveyi* quorum-sensing regulator, LuxR (McCarter, 1998).

As previously mentioned, *V. parahaemolyticus* BB22OP cells are capable of exhibiting a differential colony morphology, designated opaque or translucent. OpaR has been implicated in the switch from a translucent to an opaque phenotype.

A translucent *V. parahaemolyticus* strain containing an *opaR*::Tn5 insertion complemented with an IPTG-inducible copy of *opaR* on a plasmid will switch to the opaque phenotype upon induction of *opaR* expression with IPTG. The colony morphology of the IPTG-induced opaque strain appears identical to the wild-type opaque strain (McCarter, 1998). When stained with ruthenium red and viewed under an electron microscope, opaque cells are surrounded by a thick, electron-dense layer that is not found in translucent cells. The opaque strain was found to produce abundant levels of capsular polysaccharide in comparison to the translucent cell type due to activation by OpaR (Enos-Berlage and McCarter, 2000). Extracellular polysaccharide production has been demonstrated to influence biofilm development (reviewed in Sutherland, 2001). This suggests a linkage between quorum sensing and biofilm production in *V. parahaemolyticus* that may influence colonization and pathogenesis.

Another important phenotype that distinguishes the opaque and translucent cell types is their ability to swarm. *V. parahaemolyticus* BB22OP can assume two distinct flagellar patterns. In liquid, cells produce a single polar flagellum for swimming motility. The lateral flagellar (*laf*) system is expressed when the organism is growing on a solid and more viscous surface. Production of the *laf* genes results in a swarmer cell with multiple lateral flagella (Stewart and McCarter, 2003). Opaque strains produce copious amounts of capsular polysaccharide but are incapable of swarming motility, whereas translucent strains produce little capsular polysaccharide and swarm proficiently. Initially, it was hypothesized that either the excess capsular polysaccharide inhibited swarming or OpaR repressed swarming.

Opaque strains produce significantly less lateral flagellin than translucent strains, supporting the hypothesis that OpaR genetically represses swarming. OpaR was found to inhibit swarming by repressing *laf* gene expression, extending its role as an important regulator in *V. parahaemolyticus* (Jaques and McCarter, 2006).

CsrA

CsrA is predicted to be another important regulator of virulence and colonization in *V. parahaemolyticus*. This hypothesis was examined through the work described in Chapter Three of this thesis. CsrA, or carbon storage regulator, is a global regulator that was first identified in *Escherichia coli* where it is involved in the transition from exponential to stationary growth, controlling the switch between gluconeogenesis and glycolysis, which can be measured through glycogen synthesis and catabolism (Romeo, 1998). It can indirectly influence quorum sensing in *V. cholerae* by influencing LuxO levels (Lenz et al, 2005). CsrA is a RNA binding protein that post-transcriptionally regulates gene expression through binding specific mRNA near the ribosome-binding site (Liu et al, 1997). CsrA homologs remain highly conserved among different bacterial species, and are known to play roles in biofilm formation and virulence in both plant and animal pathogens (Altier et al, 2000; Jackson et al, 2002; Ma et al, 2001).

CsrA activity is controlled by small RNAs, which are capable of binding multiple copies of CsrA to titrate it away from its target mRNAs. In *E. coli* these sRNAs are called CsrB and CsrC (Figure 1.3; Weilbacher et al, 2003). Expression of the genes encoding these sRNAs is activated by the BarA-UvrY two-component

system (Suzuki et al, 2002). γ-proteobacteria that contain CsrA homologs also contain systems homologous to the BarA-UvrY system, suggesting that CsrA-regulating sRNAs are a conserved method of CsrA control.

A program called CSRNA_FIND was developed to identify putative CsrA-regulating sRNAs (Kulkarni et al, 2006). These predictions included two CsrA-regulating sRNAs in *Vibrio fischeri*, CsrB1 and CsrB2, which have been experimentally verified (Kulkarni et al, 2006). This program also predicted four CsrA-regulating sRNAs in *V. parahaemolyticus* RIMD2210633, CsrB1-3 and CsrC, which are being analyzed as part of this research.

sRNA Regulation

There are two major methods of sRNA regulation in prokaryotes, via homologous binding to target mRNA or via modulation of the activity of RNA-binding proteins. Both involve the expression of short RNA transcripts that are capable of post-transcriptional regulation of gene expression. The most common form of sRNA regulation involves sRNAs that can base pair to mRNA targets; this regulation can occur in *cis* or in *trans*. *Cis*-encoded base-pairing sRNAs are encoded on the opposite strand of DNA from their gene targets. This creates a region of highly specific base pairing between the sRNA and the mRNA. *Cis*-encoded sRNAs are capable of binding the 5' untranslated region (UTR) of the mRNA transcript, inhibiting translation and promoting degradation of the mRNA. They are also able to bind between two genes encoded in a polycistronic mRNA. This interaction can cause cleavage of the mRNA or transcription termination after the first cistron

(reviewed in Waters and Storz, 2009). In *E. coli*, a *cis*-encoded sRNA, GadY, binds *gadXW* mRNA leading to cleavage of the mRNA between the *gadX* and *gadW* coding sequences (Opdyke et al, 2004; Tramonti et al, 2008). This mechanism is a common way to maintain the proper copy number of a plasmid, such as ColE1 RNA I, which acts by inhibiting replication (Tomizawa et al, 1981).

Trans-encoding base-pairing sRNAs have limited complementarity to their mRNA targets because they are encoded elsewhere on the genome from their target genes. Kawamoto et al. determined that the SgrS sRNA has the potential to form a 23 base-pair complex with *ptsG* mRNA; however, only four base-pairs between the two RNA molecules are essential for destabilization of *ptsG* mRNA (Kawamoto et al, 2006). This limited complementarity results in the ability of sRNAs to bind several different mRNA targets (reviewed in Gottesman, 2005; Massé et al, 2005). *Trans*-encoding sRNAs bind the 5' UTR of their target mRNAs; this interaction can result in translation inhibition by blocking of the ribosome-binding site, which is often coupled with mRNA degradation. This method of action is utilized by the Qrr sRNAs to repress the expression of *hapR* mRNA in *V. cholerae* (Svenningsen et al, 2008). *Trans*-encoding sRNAs can also relieve secondary structures in the mRNA that block the ribosome-binding site, and by doing so actually promote translation of the mRNA target. The Qrr sRNAs in *V. cholerae* are also capable of positive regulation (Hammer and Bassler, 2007; reviewed in Waters and Storz, 2009).

In the second major mode of action, small RNAs are also capable of modulating protein activity. The paradigm of this kind of interaction is regulation of

CsrA by CsrB and CsrC in *E. coli* (Figure 1.3). CsrB and CsrC are referred to as CsrA-regulating sRNAs. As mentioned previously, CsrA is an RNA binding protein that binds GGA motifs in the 5' UTR of its target mRNAs affecting the stability or translation of the mRNA. CsrB and CsrC contain several GGA-binding sites to mimic the 5' UTR of CsrA-regulated mRNAs; therefore, CsrB and CsrC are capable of binding 22 and 13 CsrA proteins, respectively (Liu et al, 1997; reviewed in Waters and Storz, 2009).

CsrA-Regulating sRNAs in *Vibrio cholerae*

In *E. coli*, transcription of the genes for *csrB* and *csrC* are induced by the BarA-UvrB two-component system. There is a homologous system in *Pseudomonas aeruginosa*, GacA-GacS, that induces the expression of CsrA-regulating sRNAs (Kay et al, 2006; reviewed in Waters and Storz, 2009). Among the Vibrios, this regulatory system is best understood in *V. cholerae* where it is referred to as VarS/VarA (Lenz et al, 2005).

Due to the presence of the VarS/VarA as well as the presence of a CsrA homolog, it was determined that *V. cholerae* likely encodes one or more CsrA-regulating sRNAs. A BLAST search using the sequence for *E. coli csrB* as a query, revealed a CsrB homolog in *V. cholerae*. A subsequent search using this sequence revealed two more putative CsrA-regulating sRNAs, deemed CsrC and CsrD. Secondary structure predictions for these sequences suggest that they share the same secondary structure as other known CsrA-regulating sRNAs. The looped regions contain AGGA and AGGGA CsrA binding motifs, suggesting that CsrBCD of *V.*

cholerae likely act in a manner similar to their homologues in *E. coli* and control CsrA by binding multiple copies and titrating it from the environment and away from its target mRNAs (Lenz et al, 2005).

Sequence alignment of the upstream regions of the genes encoding CsrBCD revealed a putative VarA binding site based on the UvrY binding site upstream of *csrB* in *E. coli*. Transcriptional reporter fusions confirmed that the VarS/VarA system regulate *csrBCD* expression. Northern blot analysis confirmed this hypothesis. Each sRNA is expressed at a lower level than the wild-type in both a *varS* and *varA* mutant (Lenz et al, 2005).

V. parahaemolyticus RIMD2210633 is hypothesized to have four CsrA-regulating sRNAs, CsrB1, CsrB2, CsrB3 and CsrC. It is likely that these CsrA-regulating sRNAs regulate CsrA activity similarly to the CsrA-regulating sRNAs in both *E. coli* and *V. cholerae*; however, since there are different CsrA-regulating sRNAs in *V. parahaemolyticus,* this system likely has unique attributes.

Research Plan

The research in Chapter Two aims to provide a comparative analysis between the genomes of *V. parahaemolyticus* strains RIMD2210633 and BB22OP, focusing on what is unique to each of the two strains (Figure 1.4). The first step in evaluating this goal required the sequencing, assembly and annotation of the BB22OP genome. The two strains exhibit phenotypic differences that may be attributed to unique genes found in each. So after the assembly and annotation was completed, the genomes were compared to determine regions that were different in

or novel to the BB22OP strain. The genes novel to BB22OP were evaluated for their potential effect on the physiology of the organism.

The third chapter of this research aims to confirm the presence and functionality of CsrA-regulating sRNAs, CsrB1, CsrB2, CsrB3 and CsrC in *V. parahaemolyticus* RIMD2210633. Their functions were evaluated both qualitatively via iodine glycogen staining plate assays and quantitatively via glycogen production assays in recombinant *E. coli*. To measure post-transcriptional regulation by CsrA, a transcriptional and translational fusion system of putative CsrA target promoters was developed to enable the screening of potential target gene promoters in the future.

Figure 1.1: Comparison of the opaque and translucent colony phenotypes. *V. parahaemolyticus* BB22 opaque (right) appears darker and the colonies are smaller in size. BB22 translucent spontaneous mutant (left) appears lighter and the colonies are wider. The strains were grown overnight at 30°C on HI medium with 2% agar to prevent swarming.

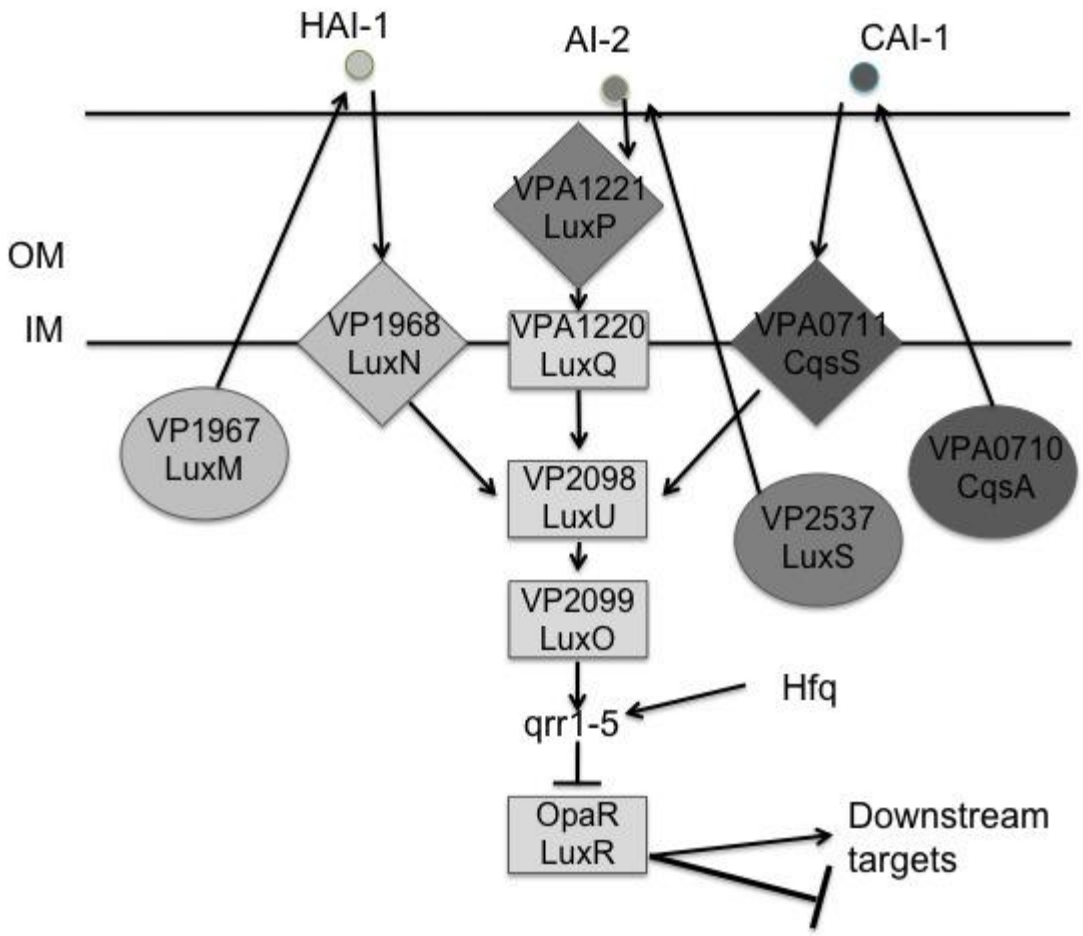

Figure 1.2: Proposed mechanism of quorum sensing in *V. parahaemolyticus*. *V. parahaemolyticus* genes are on top with their *V. harveyi* homologs below. The identities of *V. harveyi* autoinducers are indicated. HAI-1 is *N*-(3-hydroxybutanoyl) homoserine lactone (Cao et al, 1989). AI-2 is 3A-methyl-5,6-dihydro-furo(2,3-D)(1,3,2)dioxaborорle-2,2,6,6A-tetraol (Chen et al, 2002). CAI-1 is (*S*)-3-hydroxytridecan-4-one (Higgins et al, 2007).

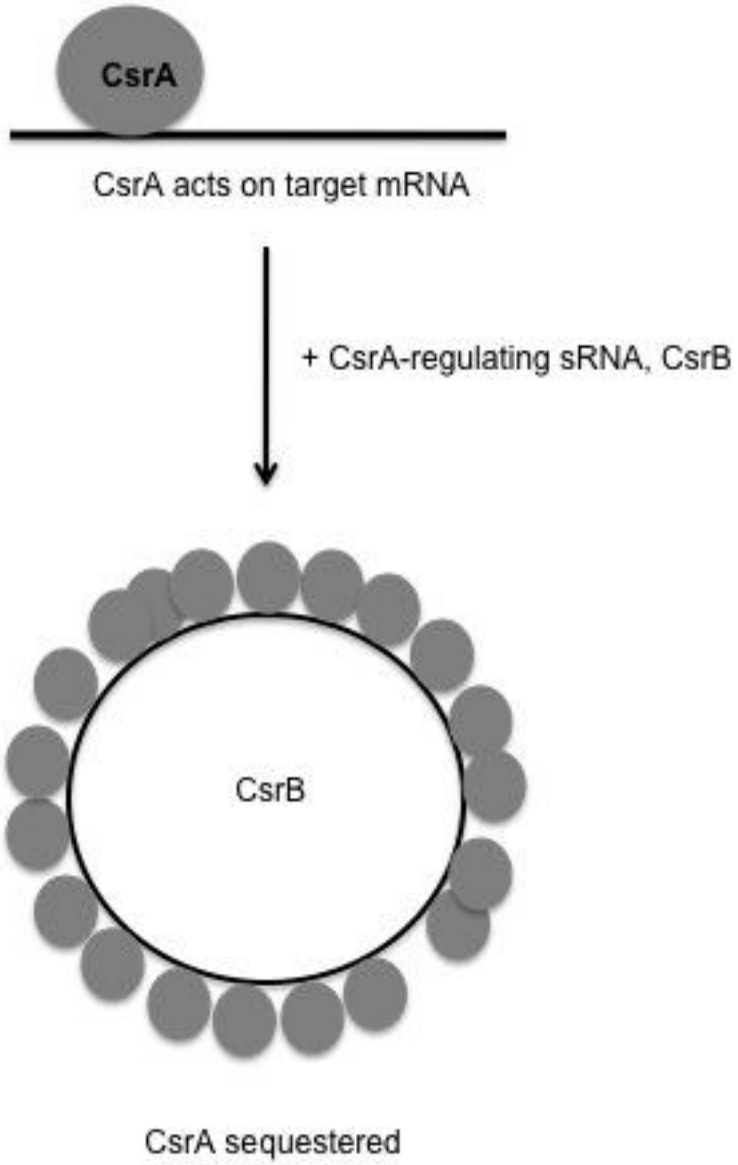

Figure 1.3: Regulation of CsrA by protein-binding sRNA CsrB. CsrA binds GGA motifs on the mRNA altering expression of the transcripts. When CsrB levels increase, CsrA binds GGA motifs on CsrB, sequestering it from the environment.

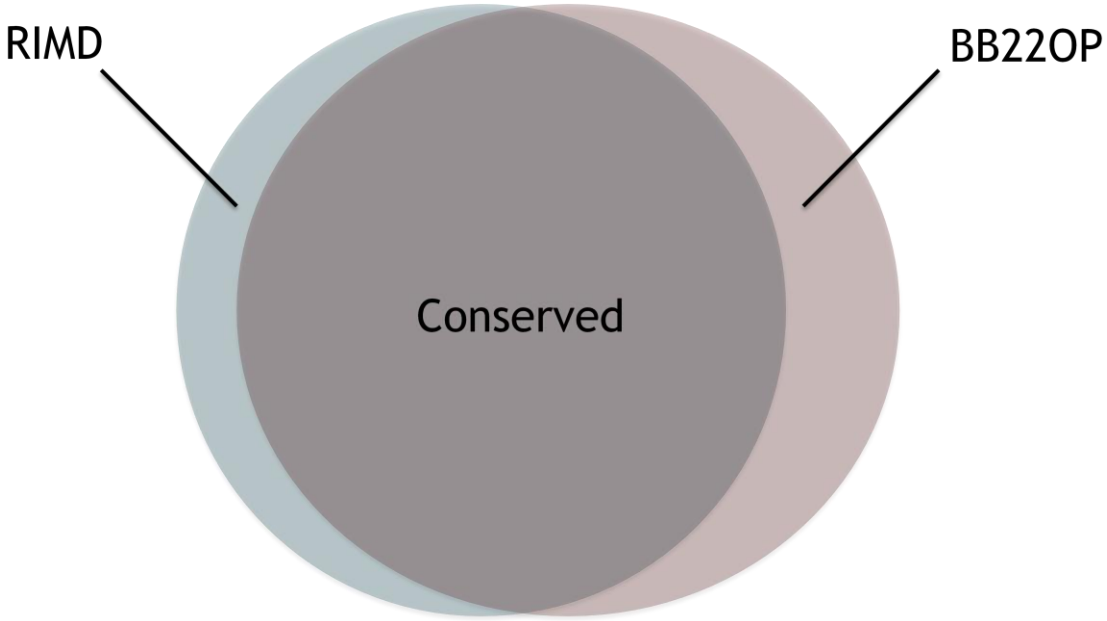

Figure 1.4: Venn diagram demonstrating the degree of conservation between the two *V. parahaemolyticus* strains, RIMD2210633 and BB22 opaque (BB22OP). There are approximately 100 genes unique to RIMD2210633 and 300 genes unique to BB22 (see Chapter 2).

Chapter 2

Genome-scale comparative analysis of two *Vibrio parahaemolyticus* strains

Abstract

The marine bacterium, *Vibrio parahaemolyticus*, is a leading cause of seafood-associated gastroenteritis in the United States and an emergent food-borne pathogen worldwide. The genome of *V. parahaemolyticus* strain, BB22OP, has now been sequenced. The Roche/454 GS FLX Titanium sequencing system provided approximately 200x coverage of the BB22OP genome, allowing for a comparative analysis with the previously sequenced, pathogenic RIMD2210633 strain. Strain BB22 is one of the best genetically characterized *V. parahaemolyticus* strains. It has the ability to undergo a phase variable switch from an opaque (BB22OP) to a translucent colony phenotype (BB22TR). This switch alters many cell surface characteristics of the cell, including capsular polysaccharide production. The opaque strain is not pathogenic; whereas, genetic alterations in either *opaR* or *luxO* cause the translucent strain to be pathogenic. *opaR* encodes the major quorum-sensing output regulator, and *luxO* encodes a transcriptional regulator upstream of *opaR* in the quorum-sensing network. It is hypothesized that a comparative genome-level sequence analysis of BB22OP and RIMD will reveal novel genetic components in the BB22OP strain. A *de novo* sequence assembly using the Roche/454 Newbler software resulted in the alignment of 98% of the reads into 125 contigs covering 5.05 Mbp. These contigs have been assembled into two chromosomes. Alignment to the RIMD sequence has revealed the presence of approximately 320,000 bp unique to BB22OP containing more than 300 novel genes, including 5 transcriptional regulators, a hemolysin, and several MSHA- and type IV pili-associated genes. Comparative analysis provides insight into both the

commonly shared and the distinctive genetic determinants in the two strains under investigation.

Methods and Materials

DNA Preparation

The wild-type opaque *V. parahaemolyticus* BB22OP strain LM5312 (McCarter, 1998) was streaked for individual colonies on HI medium (25 g/L Heart Infusion, 20 g/L NaCl) with 2% agar to prevent swarming. The strain was streaked out a total of three times to ensure an opaque colony was selected, as the strain is capable of spontaneously switching to the translucent phenotype. A single opaque colony was chosen to inoculate 5 mL of HI broth that was grown at 30°C with shaking at 250 rpm overnight. The QIAGEN DNeasy Blood & Tissue Kit (QIAGEN, Valencia, CA) requires a maximum of 2×10^9 cells for optimal purification, so 1.3 mL of the culture was harvested by centrifugation. The genomic DNA was prepared using the protocol optimized for Gram-negative bacteria. The optional RNase treatment was used and the DNA was eluted off the column using 100 µL of AE buffer (provided with the kit). A second elution was performed using 100 µL of AE buffer. Serial dilutions of the genomic DNA were run on a 1% agarose gel to ensure the quality and 10 µg of DNA were sent to the Virginia Bioinformatics Institute for sequencing.

DNA Sequencing

DNA sequencing was performed by the Virginia Bioinformatics Institute Core Laboratory Facility using the Roche/454 GS FLX Titanium Series system (Roche Diagnostics Corporation, Branford, CT). The system provides over 1 million reads

averaging approximately 400 base pairs in length in one 10 hour sequencing run (http://www.454.com/products-solutions/system-features.asp). Two full sequencing runs generated 2,977,326 reads and 970,168,049 bp of sequence for the BB22OP genome.

Genome Assembly

A *de novo* sequence assembly of the reads was performed using the Roche/454 Newbler software. In this case, the reads were aligned to each other with no additional input resulting in a set of 125 "*de novo*" contigs ranging in size from 101 bp to 883,073 bp. For each of the large *de novo* contigs (> 100 bp), the Roche/454 Newbler assembly software provides files, called .ace files, that contain information for how each individual read aligns to the contigs and the different ways the contig may connect to other contigs. Utilizing the information in the .ace files, a computer program was designed using Cytoscape (www.cytoscape.org) to map all of the connections between the contigs (Tian Hong, personal communication). Using this map (Figure 2.1), in combination with manually analyzing the information in the .ace files, the connections between contigs were determined. PCR primers were designed to confirm the junctions (Table 2.1). An effort was made to avoid making primers to contigs that were thought to be repeated throughout the genome. In these instances, the PCR primers were designed to form a product that included the repeated region. If necessary, longer PCR products (> 500 base pairs) were sequenced from each end, with both the forward and reverse primers.

Primers (Table 2.1) were ordered from Integrated DNA Technologies (Coralville, IA). PCR reactions were performed according to the *Taq* PCR Master Mix protocol (QIAGEN). Each reaction contained 2.5 units *Taq* DNA polymerase, 1x QIAGEN PCR Buffer (containing 1.5 mM $MgCl_2$), 0.5 µM of each primer, 1 µg BB22OP genomic DNA and RNase-free water to 100 µL. The products were separated on a 1% agarose gel, and then they were excised and eluted from the gel using the QIAGEN Gel Extraction Kit. The products were then sent to the Virginia Bioinformatics Core Laboratory Facility for sequencing using the appropriate primer.

In addition to the Newbler *de novo* assembly, a MIRA (http://www.chevreux.org/projects_mira.html) *de novo* assembly was performed by Thero Modise, a graduate student working for Dr. Roderick Jensen, in order to provide another set of 324 contigs to assist in filling gaps in the genome assembly. An assembly to the RIMD2210633 genome as a reference (described in detail below) using the Lasergene 8 software package (DNASTAR, Madison, WI) was particularly useful for visualization of the alignments of the reads and contigs to the RIMD2210633 reference genome.

Assembly of "the boneyard"

The sequencing reads were also assembled to chromosome 1 and chromosome 2 of the reference sequence (RIMD2210633) using the Newbler software package. All of the reads that did not align to the reference sequence, the "boneyard" were then assembled into the "novel" contigs. These novel contigs are

potentially all of the regions of DNA that are unique to strain BB22OP and not found in RIMD2210633.

Genome Annotation

The genome annotation was approached by several methods (Figure 2.2). When the *de novo* contigs could be successfully mapped to the RIMD2210633 reference genome using the Lasergene 8 software package, the GenBank annotation of the RIMD2210633 (NC_004603 and NC_004605) was used to identify and annotate genes. Annotation of the novel contigs was completed using both GeneMark-P* and GeneMark.hmm-P (http://exon.biology.gatech.edu, Georgia Tech), using *V. parahaemolyticus* chromosome 1 as the species input, and ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html, NCBI), using the bacterial code. The open reading frames predicted by GeneMark were BLAST searched against the non-redundant database using the blastx algorithm (default settings) on the NCBI website. Open reading frames predicted by ORF Finder were BLAST searched against the non-redundant database using the blastp algorithm (default settings). The predicted function of the open reading frame was determined by analyzing the BLAST results for conservation of function, conservation of sequence and the significance of the e-value. The e-value was used as a guide for determining the significance of the BLAST results. However, a specific cut-off value cannot be stated because shorter regions of DNA can have very good sequence homology with a higher e-value and still be considered significant if there is conservation of function among the BLAST hits.

Stand-alone BLAST

Stand-alone BLAST is a process that allows one to utilize the BLAST algorithm to search specific databases that are created by the user. Stand-alone BLAST was used to search the novel contigs against a database of the *de novo* contigs to determine where within the *de novo* contigs the novel contigs are located. This was done by using a program called "formatdb" to first create the database of the *de novo* contigs, and a program called "megablast" to complete the alignment, using the novel contigs as the query input and an e-value cut-off of e^{-100}. All of these programs are freely available for download from the NCBI website.

Results and Discussion

Genome Assembly

The initial *de novo* assembly using the Roche/454 Newbler software resulted in the alignment of 98% of the reads into 125 contigs greater than 100 bp, and numerous smaller contigs, covering 5.05 Mbp. These "*de novo*" contigs are the result of the assembly program forcing a break wherever there are repeated regions or ambiguity of a connection. The .ace files provided as output from the 454/Newbler assembly software provided information used to predict connections between the contigs. This information was used to create a map that shows each predicted connection (Figure 2.1). Using the information from the .ace files and the contig map, predicted connections were confirmed with PCR (Table 2.1).

Table 2.1 shows the junctions that were confirmed by PCR and subsequent sequencing, including any additional bases predicted in between contigs. Additional base pairs located between contigs can be predicted by analysis of the .ace files and then confirmed by PCR and sequencing. For example, the predicted additional bases in between *de novo* contigs 80 and R89 were amended from "GTGGAA" to "TGGA" after analysis of the sequencing results. These results indicate that it is important not to solely rely on automated methods and computer algorithms in order to assemble a genome. Subtle discrepancies can be missed by automated methods that are easily detectable by PCR and sequencing.

Assemblies of the BB22OP sequencing reads to the RIMD2210633 reference strain using the Newbler and Lasergene software confirmed gaps in the BB22OP genome, originally identified by DNA microarray data from the McCarter lab (Linda McCarter, personal communication) corresponding to RIMD2210633 DNA sequence and associated genes that are missing from the BB22OP genome. Moreover, assembling all of the BB22OP reads that failed to align to the RIMD2210633 genome resulted in 102 contigs ranging from 510 bp to 32996 bp that did not align to the reference sequence. These novel contigs contain genes that are unique to BB22OP.

A third assembly using MIRA was performed by Thero Modise. Collectively, using the PCR and sequencing results, the contigs resulting from the MIRA assembly and the Lasergene alignment to the RIMD2210633 genome as reference, the genome of *V. parahaemolyticus* strain BB22OP has been assembled into two distinct chromosomes. Final proofreading of the genome sequence is being performed by Elizabeth Harbolick, an undergraduate researcher working for Dr. Roderick Jensen.

There are two significant issues left to resolve in the genome assembly. The first problem areas are the tRNA/rRNA clusters, consisting of one or two copies of the ribosomal RNA subunit genes followed by several tRNA genes. There is one copy of this cluster on chromosome two and 10 copies on chromosome one. The tRNA/rRNA cluster is highly repetitive and therefore difficult to assemble using bioinformatic tools. It is also too long (>5000 bp) to easily sequence across and because it is repetitive, internal sequencing primers cannot be designed. Because tRNAs and rRNAs are highly conserved and there are no additional open reading

frames in this region, the tRNA/rRNA clusters from RIMD2210633 will be used in place of the BB22OP tRNA/rRNA cluster sequence. The second problem area is an integron that is found in both genomes, but is not highly conserved. The integron region will be discussed in more detail in the annotation section of this chapter.

Genome Annotation

When the Lasergene assembly to the reference sequence (RIMD2210633 genome) was performed, the regions of DNA that did align to the reference sequence were annotated by Lasergene 8 based on the similarity of the BB22OP genome to the RIMD2210633 genome. For those sequences that did not align to the RIMD2210633 genome, the novel contigs, the annotation was complicated by the discrepancy between the outputs of GeneMark and ORF Finder. GeneMark predicts a total of 304 open reading frames, while ORF Finder predicts 1081 open reading frames greater than 100 base pairs. GeneMark is generally considered a more accurate predictor of genes because it takes into account potential ribosome binding sites in addition to start and stop codons using Hidden Markov Models (exon.biology.gatech.edu, GeneMark: Background, Georgia Tech). Hidden Markov Models assign a probability to each base to be a ribosome-binding site or a start or stop codon. ORF Finder generally predicts the same open reading frames as GeneMark, though there can be differences between start sites.

The ORF Finder output was mined for any potentially legitimate open reading frames that may have been missed by GeneMark (Table 2.2). Legitimacy was determined by homology with conserved genes in other organisms. Any ORFs

with no significant homology were not considered further because ORF Finder does not search for promoter elements. There is substantial subjectivity in determining if a BLAST result is significant. Generally, a result was not considered significant if there were only a few hits reported that do not share a conserved function and limited sequence homology. The majority of the ORF Finder predicted ORFs in BB22OP with no significant homology do not produce any hits when BLAST searched against the non-redundant protein database. Hypothetical proteins were included because they have been implicated as ORFs in other organisms. There are nine potentially legitimate ORFs predicated by ORF Finder and not GeneMark. Seven of these ORFs are hypothetical proteins and have no predicted function. One is a putative epimerase/dehydratase, which are usually involved in sugar metabolism. The last is a phage transcriptional activator on novel contig 32. Novel contig 32 is a 33 kb insert in the BB22OP genome that is not found in the RIMD2210633 genome. Analysis of the annotation for the contig (Appendix I) reveals that it is comprised solely of bacteriophage genes and hypothethical proteins indicating that it may be a bacteriophage genome insertion. The putative phage transcriptional activator predited by ORF Finder fits into this region, suggesting that ORF Finder may predict legitimate ORFs that GeneMark can miss.

Most of the ORF Finder predicted open reading frames that are not also predicted by GeneMark are short (less than 100 bp) and rarely have any significant or conserved BLAST hits to the NCBI non-redundant database. This could be due to the fact that these predicted open reading frames are either too short to encode functional proteins or that coding regions this short are underrepresented in

research. Therefore, it is unclear as to which algorithm is a better predictor of open reading frames until more research is performed on small proteins. GeneMark appears to be a good predictor of larger genes and those that have a predicted function, but it may be important to incorporate ORF Finder results to supplement important open reading frames GeneMark may miss or to help determine the correct start site.

Many of the open reading frames predicted by both GeneMark and ORF Finder are located at the ends of contigs. This causes an issue when annotating these regions because it is likely that they are actually longer than predicted, and their true length will not be revealed until the entire genome is assembled. A fully complete and accurate annotation of the BB22OP genome can only be completed after the genome is fully proofread and assembled; however, annotation of the individual contigs provides considerable insight into the differences between strain BB22OP and RIMD2210633.

Predicted physiological roles of genes novel to BB22OP

A list of interesting annotated novel genes to BB22OP can be found in Table 2.3. This list includes several genes that are associated with colonization and virulence, including MSHA pilin-associated proteins, which are important for adherence. MSHA pilins are mannose sensitive, and there are numerous mannose recognition and metabolism genes novel to BB22OP. Additional type IV pili system proteins can be found in the boneyard annotation. Type IV pili are also important to adherence. Several genes found in the O-antigen cluster in *E. coli* are found in novel

contig 9, including dTDP-D-glucose 4,6 dehydratase, glucose-1-phosphate-thymidylyltransferase, a WxcM-like protein, WblQ and a glycosyl transferase family protein. The arrangement of these genes suggests that they are arranged in an operon together. The O-antigen in polysaccharide found on the outside of Gram-negative bacteria that elicits a strong immune response in host cells. The Na^+/glucose symporter may contribute the Na^+ membrane potential, which is known to power the polar flagellar motor in *V. parahaemolyticus*. The polar flagellum in *V. parahaemolyticus* is sheathed and there is a polar flagellar sheath protein novel to BB22, which could indicate a difference between the flagellar sheaths in these two organisms. Notably, there is an additional hemolysin found on novel contig 84, which is not found in the RIMD2210633 genome. Hemolysins are an endotoxin produced by bacteria to lyse red blood cells in the host. There are five transcriptional regulators novel to BB22OP indicating the potential for differential regulation between the two strains. Due to the tiered effect of gene regulation, these five additional transcriptional regulators could have a profound effect on gene expression in BB22OP. Further comparative analysis of the two strains may reveal more phenotypic diversity as a result of differential regulation.

Interestingly, the BB22OP genome encodes bleomycin and ampicillin antibiotic resistance genes, as well as a penicillin-binding protein likely involved in cell wall synthesis. Penicillin is an antibiotic that disrupts cell wall synthesis. As this strain's differential phenotype displays altered outer membrane characteristics, this antibiotic resistance may play a role in determining the differential cell surface characteristics. The penicillin-binding protein is found in an integron on

chromosome 1. Integrons are notorious for carrying acquired antibiotic resistance genes. An integron is a series of gene cassettes separated by repeat sequences that facilitate recombination events allowing an organism to acquire genetic material via horizontal gene transfer, commonly antibiotic resistance and virulence genes (reviewed in Cambray et al, 2010). The integron in *V. parahaemolyticus* RIMD2210633 is comprised of 69 gene cassettes (reviewed in Cambray et al, 2010). The integron in *V. parahaemolyticus* BB22OP is not completely assembled due to the repetitive regions, but is already known to contain at least 86 gene cassettes some of which are conserved with RIMD2210633 and some of which are completely novel and share no sequence homology to any gene with known function. The annotation of the integron is being performed by Elizabeth Harbolick.

An interesting feature of integrons is that the gene cassettes are often promoterless, which would indicate they are prone to being lost due a lack of selective pressure. However, integrons in *Vibrio* species can contain as many as the 217 gene cassettes found in *Vibrio vulnificus* CMCP6. It is thought that gene cassettes are maintained by addiction modules, genetic elements that result in toxicity and cell death when their gene expression is disrupted (reviewed in Cambray et al, 2010). One type of addiction module is a toxin-antitoxin (TA) cassette, which are commonly found on plasmids or within prophages to maintain these structures by preventing proliferation of progeny cells lacking the TA cassette. They likely perform the same function in the integron. The TA cassette is organized as an operon encoding a stable toxin molecule, which builds up in the cell, and an unstable antitoxin molecule that inactivates the toxin. If a progeny cell lacks the TA

cassette, the toxin from the mother cell will be lethal because the unstable antitoxin is not being expressed to inactivate it (reviewed in Cambray et al, 2010). In the "boneyard" annotation of *V. parahaemolyticus* BB22OP, there are four additional toxin-antitoxin cassettes, of which at least one is located in the integron (Table 2.2). There are expected to be more TA cassettes located in the integron region, however this cannot be confirmed until the integron assembly and final annotation are completed.

Stand-alone BLAST using a draft of the integron as a query and the novel contigs as the subject and an e-value cut-off of e^{-50} reveals that novel contigs 17, 20, 21, 23, 37, 38, 43, 44, 45, 50, 57, 61, 64, 80, 88, 125, 305 and 308 are part of the integron. This is not a comprehensive list, as much of the integron is conserved between both the RIMD2210633 and BB22OP strains and many of the *de novo* contigs are also part of the integron. The annotation for these novel contigs can be found in Appendix I. Many of the ORFs located in the novel regions of the integron are annotated as hypothetical proteins or do not have any significant hits to the database implying they are potentially completely novel genes that have not been seen before.

Analysis of the sequencing results indicated that *de novo* contig 70 contains a 25 kb insert of DNA that is not found in the RIMD2210633 strain genome. This segment of DNA is comprised of novel contigs 7 and 8 (Appendix I). These predictions are based on stand-alone BLAST results where the novel contigs were used as a query against the *de novo* contigs as a database. This process revealed the

location of the novel contigs in relation to the *de novo* contigs. The majority of the genes have no assigned function, so it will be interesting to see what impact these novel genes have on the physiology of BB22OP.

Current genome annotation tools available today do not predict the presence of other important genome features, such as small RNAs that play a significant role in gene regulation. Stand-alone BLAST was used to confirm the presence of four CsrA-regulating sRNAs in the BB22OP genome based on their homology to those found in RIMD2210633 genome (Table 2.4). The subject used was the CsrA-regulating sRNA sequences from the RIMD2210633 genome (Kulkarni, et al, 2006; see Chapter 3) and the query was the BB22OP genome draft dated 9/16/10 with an e-value cut off of e^{-100}. BLAST results predicted the presence of the same four CsrA-regulating sRNAs in both strains with very high sequence conservation. All of CsrA-regulating sRNAs found in BB22OP were found on the same chromosome as the corresponding sRNA in the RIMD2210633 genome. The presence of five predicted Qrr sRNAs in BB22OP was also confirmed (Table 2.4). The subject used was the Qrr sequences from BB22OP itself (Linda McCarter, personal communication), and the query was the BB22OP genome draft date 9/16/10 with an e-value cut-off of e^{-20}. This e-value is relatively high because the *qrr* sequences are very short (approximately 100 base pairs each). While the presence of these Qrr sRNAs has already been established, stand-alone BLAST was used to ensure that the list was comprehensive now that the genome has been sequenced and assembled. No additional Qrr sRNAs were found; however, the predicted sequences and chromosomal locations were confirmed.

In total, there are 313 (304 predicted by GeneMark and 9 predicted by ORF Finder) predicted open reading frames novel to the BB22OP genome, pending the final assembly and annotation of the genome. To date, 58 ORFs are not annotated to their full length because they are located at the ends of contigs. Of the 313 predicted ORFs, only 130 are assigned a putative function based on homology to other annotated genes in the Genbank database. While it is interesting to investigate what role these genes have in the differential phenotype of BB22, there is potentially a lot to be learned from the 183 genes that have no assigned function because they are either hypothetical proteins (125) or have no significant homology (58) to any genes in the Genbank database. It is possible that the 58 ORFs with no significant homology are not genes at all, but some could potentially be genes that are completely novel to *V. parahaemolyticus* BB22OP and play a significant role in its differential phenotype.

Acknowledgements

I would like to emphasize the fact that the sequencing, annotation, and assembly of the *V. parahaemolyticus* BB22OP genome has been a collaborative effort by a team of individuals, and would not be complete without each one of them. I would like to thank Ann Stevens, Roderick Jensen and Linda McCarter for writing the grant proposal (VBI #555175) that provided us with the funding for this project, as well as for their help and guidance throughout this process. Secondly, I would like to thank Linda McCarter for providing not only the strain, but her knowledge of it as well. Thank you to Clive Evans at the Virginia Bioinformatics Institute Core Laboratory for providing guidance on the genomic DNA preparation and submission. Roderick Jensen performed the bulk of the bioinformatics, with the help of his graduate student, Thero Modise, and his Spring 2010 Bioinformatics and Computation Genomics class, specifically Tian Hong who predicted a number of the gaps that were filled by PCR and sequencing. Cimarron Smith, a summer 2010 NSF REU student I supervised performed a number of the PCR reactions and designed primers to fill gaps predicted by Tian. I would also like to thank Elizabeth Harbolick, an undergraduate student of Roderick Jensen, for completing the integron annotation and helping with the final proofreading of the genome assembly.

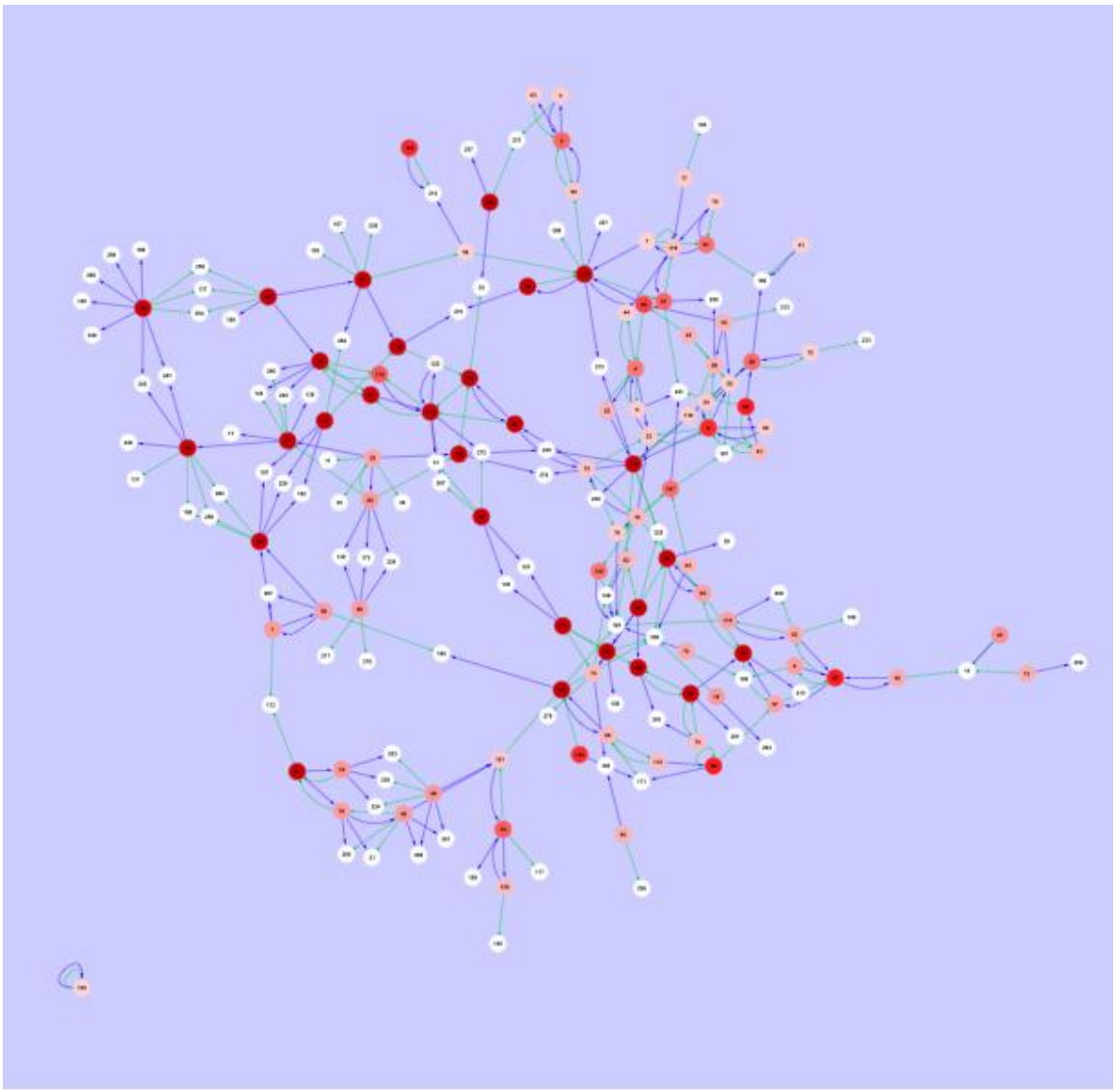

Figure 2.1: Preliminary map of *de novo* contigs. Each contig is shown with its connections to other contigs based on the individual read information found in the .ace files. Contigs numbered higher than 125 do not have any .ace file information and are likely very small (< 100 base pairs). The color indicates the coverage of that particular contig based on the number of reads that comprise the contig and the average length of the reads (approximately 300 base pairs). The dark red contigs have very high coverage and are likely repeated regions. This map was created by Tian Hong and is reproduced with his permission.

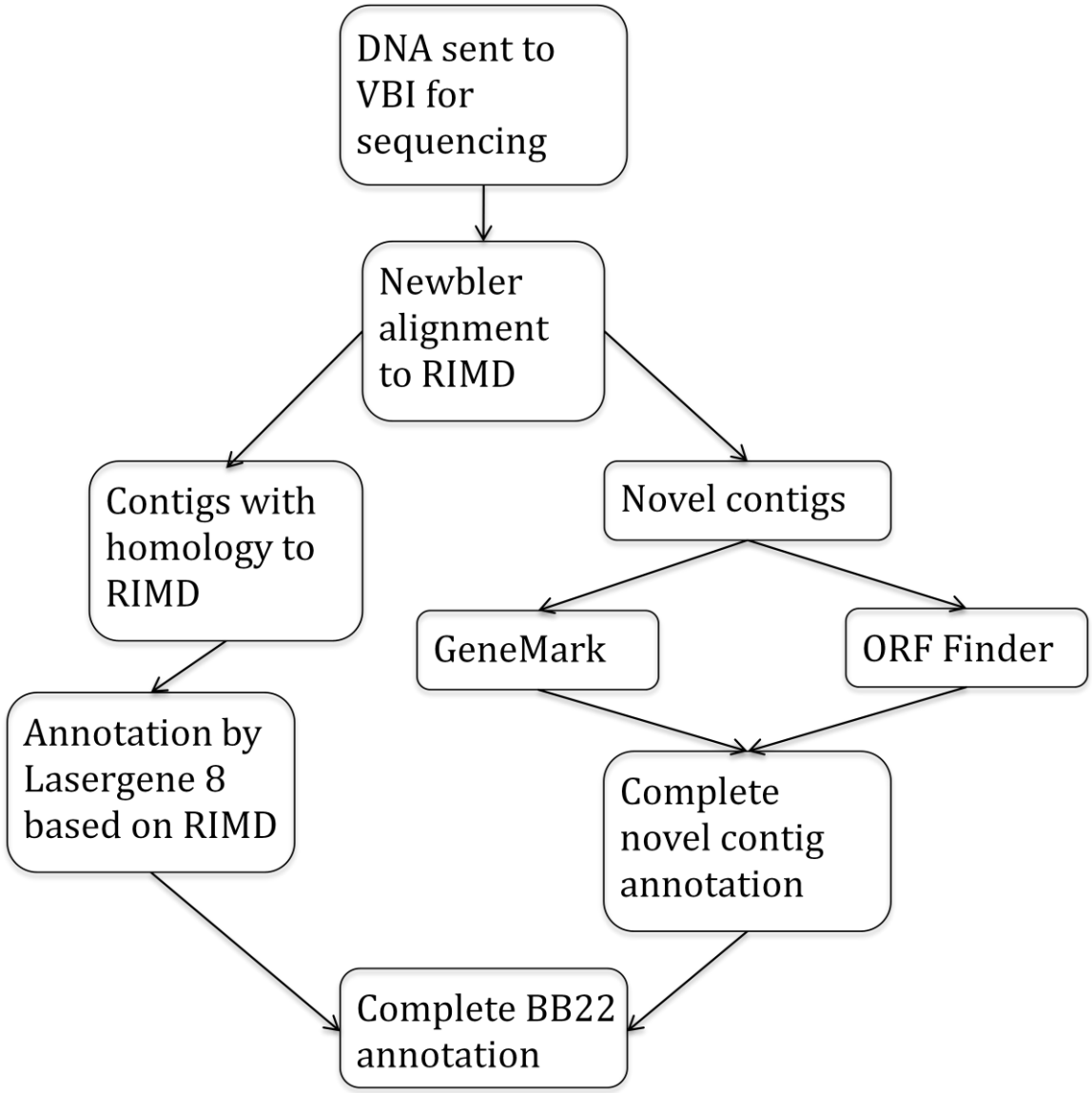

Figure 2.2: Schematic of the annotation process.

Table 2.1: Primers used in this study grouped in pairs[a].

Connection[b]	Primers	Sequence (5' --> 3')
109 to AA to R101	109F	CCTCGCACTTACCTTGCTTGGTTAGGCG
	R101R	GCACGGTGGCTGTCTACGTATTTCAGTGCCGC
R101 to R176	R101F	GCGATGCTTACTTCACAAATGAGGCGC
	R76R	GGGCACACAAATATTCTTTGTGATGCC
11 to 25	11F	CGCGTTACGAAAGTTTCACCACGACGTGG
	25R	CGCTTCGCCTTCTTGAGGCAATTCGC
25 to A to 80	25F	GGCATCGCCATATATTTCTTCTGGCCCC
	80R	GCCTTGGCGATCGTTCGGAAAATGC
80 to TGGA to R89	80F	CCTCTTCCACTTCTTCACCTGGCGCTACC
	R89R	CGGGTTCACTGTTTGAAGACATGTTTGCCG
108 to R117	108F	GCCCTCTCGGCTCAATTTGTAGAAACTTGAGC
	R117R	CGTATTTCAGTGCCGCATCAATGATCGCG
dn70 to n7	dn70F	GGAAAACGCGTTTTTCCTTGTGCGCGCG
	n7R	GCATCTGAAGCGATGCGCTCCACTAATTCGG
Rn8 to dn49	Rn8F	GGGGAGACTGAGTATTCAGCAGCAAGTGTTAGG
	dn49R	GGAAACGGCTTCAGTCGCATCACTGACGC
n7 to GGG to Rn8	n7F	CCAGTTCTTGTACCAAGTCCGTACTTAGACAGGG
	Rn8R	CCAAGATGGGCTCCCATAGGAGG
15 to 124	15F	CGGTGGTATTCCTGATACAGACGAGC
	124R	GCCGCCACTTCTGGCCATTTCTTCACG
124 to 3	124F	CCGAGTCTGCAAACAGAAGGTGAGTCAATCTCG
	3R	CCGGTTCACTTTTATTCGCCGCAACAGTCGACGC
109 to 84	109Fc	CCGATGATGAATTATTAGAGATGATTGACTAGATACGC
	84R	CCCCTGCTAAGGGAGTATACGGTTTATCCCG
R56 to R7	R56Fc	CGTGGCGGTCACACCGTCGGCACCGATAATCGC
	R7Rc	CGGTGGACGATGTGTATGCCGAAGGCG
39 to R63	39F	CGAAGAGAAATTAACACCAATCTACCC
	R63Rc	GGCTATGCAGCGGATTGCAAATCCG
R11 to R45	R11Fc	CGGTCTACCACCAACAGCAGGTCAAGG
	R45Rc	CGCGCCACAAACCCATCTTTGCGAACGGG
47 to 46 (Insert)	47F92010	CAGCCATATCTGCGGTATGCACACC
	46R92010	CGTGCCGGAATCGTCCCTTGTTCAC
13 to 11 (Insert)	13F91510	GAAGGGCTAACGGATGCTCCTCTTCG
	11R91510	CCCTTTGAGGCGTACCGAGTGAGTG
I60 to I76 (Integron)	I60F092110	CTCCACTTACTTCTAGAGTTATCTAATGG
	I76R092110	CGTTTAGTTAATGAAGTTCGACG
Integron to M70	I1F111510	GTCACCGCACCAAAGAATAAATCAACCG
	I1R111510	GCCAAGGTTATCGTAAAAGAACTGG
M70 to Integron	I2F111510	CCCTGAACTTCGATATACATCTGGC
	I2R111510	CGAGTCTGTCCATCATCATTTCGTACG

[a] number refers to contig number, R indicates the contig is in the reverse orientation, n indicates a novel contig, dn indicates a *de novo* contig, M indicate a MIRA contig, I indicates the contig is found in the integron

[b] additional base pairs between contigs are shown in the "Connection" column.

Table 2.2: Potential legitimate open reading frames predicted by ORF Finder.

Contig	ORF	Length (bp)	Annotation
4	8	117	hypothetical protein
32	24	252	phage transcriptional activator, Ogr/delta
34	5	105	hypothetical protein
49	2	105	hypothetical protein
55	5	189	hypothetical protein
56	1	201	hypothetical protein
76	2	114	hypothetical protein
146	2	288	putative epimerase/dehydratase
306	6	102	hypothetical protein

Table 2.3: Novel BB22OP genes of interest. The information in this table corresponds to some of the data found in Appendix I.

Contig	ORF	Length (bp)	Annotation
4	15	201	transcriptional regulator
9	2	1065	dTDP-D-glucose 4,6 dehydratase
9	3	867	glucose-1-phosphate-thymidylyltransferase
9	5	471	WxcM-like protein
9	6	1104	WblQ
9	18	723	Glycosyl transferase family protein
9	19	132[a]	lipid A biosynthesis
10	1	4596[a]	MshA biogenesis protein MshQ
13	3	1266	UDP-N-acetyl-D-mannosamine dehydrogenase
13	10	1107	GDP-mannose 4,6-dehydratase
13	12	471	GDP-mannose mannosylhydrolase
13	13	1443	mannose-1-phosphate guanylyltransferase
13	14	1428	phosphomannomutase
13	15	1182	mannose-6-phosphate isomerase
14	3	1632	Na^+/glucose symporter
21	1	162[a]	glyoxalase/bleomycin resistance protein
26	2	915	LysR-family transcriptional regulator
26	4	1020	beta-lactamase domain-containing protein
27	2	4749	OmpA family protein
27	3	1080	putative lipoprotein
35	2	618	putative Tfp pilus assembly protein PilW
35	3	495[a]	pili retraction protein PilT
38	2[b]	291	RelE protein
38	3[b]	249	stability protein StbD
41	1[b]	303	RelE1
41	2[b]	258	RelB1
42	1[b]	288	plasmid stabilization system protein
42	2[b]	282	antitoxin of toxin-antitoxin stability system
50	1	489	penicillin-binding protein
66	1	1416[a]	54K polar flagellar sheath protein A
71	1[b]	243	ParD
71	2[b]	300	ParE
84	1	885[a]	hemolysin
91	1	477[a]	MSHA pilin protein MshC
97	1	699[a]	nitric oxide reductase transcriptional regulator
98	1	157[a]	MSHA pilin protein MshA
98	2	408[a]	MSHA pilin protein MshA
100	1	405	type IV pilin PilA
306	2	990	transcriptional regulator, LysR family
306	5	882	transcriptional regulator

[a] indicates the genes is located at the end of a contig and is not annotated to its full length

[b] indicates the gene is part of a toxin-antitoxin cassette

Table 2.4: Presence of sRNAs confirmed by stand-alone BLAST. The start and end values indicate the nucleotide position in the BB22 draft genome dated 9/16/10.

Query	Subject	% ID	Length	Subject start	Subject end
CsrB1 (RIMD)	Chromosome 1	99.52	419	3189510	3189928
CsrB2 (RIMD)	Chromosome 1	99.51	406	2366231	2365826
CsrB3 (RIMD)	Chromosome 2	100	293	171667	171375
CsrC (RIMD)	Chromosome 1	100	303	118152	118454
Qrr1 (BB22)	Chromosome 1	100	95	2151070	2151164
Qrr2 (BB22)	Chromosome 2	100	102	1667757	166765
Qrr3 (BB22)	Chromosome 2	100	102	1284489	1284590
Qrr4 (BB22)	Chromosome 2	100	109	200469	200577
Qrr5 (BB22)	Chromosome 2	100	108	1667757	1667650

Chapter 3

Exploring CsrA regulation of select targets in *Vibrio parahaemolyticus* RIMD2210633

Abstract

CsrA, or carbon storage regulator, is a global regulator that is involved in the transition from exponential to stationary growth in *E. coli*. It is involved in the switch between gluconeogenesis and glycolysis, controlling glycogen synthesis and catabolism. CsrA is a RNA-binding protein that regulates gene expression post-transcriptionally by binding target mRNA near the ribosome-binding site. CsrA and its homologs have been implicated in biofilm formation and regulation of virulence factors in both plant and animal pathogens. In *V. parahaemolyticus* RIMD2210633 CsrA is hypothesized to be regulated by four sRNAs. These CsrA-regulating sRNAs bind multiple copies of CsrA away from its target mRNAs. The regulation of CsrA by CsrA-regulating sRNAs was tested by a qualitative iodine-staining plate assays and quantitative glycogen production assays. After this relationship was confirmed, a transcription and translational fusion system was developed to screen potential CsrA targets from RIMD2210633 in recombinant *E. coli*.

Methods and Materials

DNA manipulation

Standard DNA manipulation techniques (Sambrook et al, 1989) were used for cloning. PCR purification, gel extraction and plasmid purification kits were obtained from Qiagen (Valencia, CA). High-fidelity Deep Vent DNA Polymerase (New England Biolabs; Ipswich, MA) and Phusion Polymerase (Finnzymes via Thermo Scientific; Lafayette, CO) were used to generate PCR products for cloning steps.

Cloning of *csrA*, *csrB1*, *csrB2*, *csrB3* and *csrC* from *Vibrio parahaemolyticus*

The gene coding for CsrA was PCR amplified from *V. parahaemolyticus* RIMD chromosomal DNA with primers CsrAF and CsrAHis2 (Table 3.1). The resulting PCR product encoded CsrA with a C-terminal His_6 tag, flanked by EcoRI and HindIII restriction sites. The coding regions for each of the four sRNAs, as predicted by Kulkarni et al, were also PCR amplified from genomic DNA (Table 3.1). All PCR products were individually ligated into the pGEM-T vector (Promega, Madison, WI) and sequenced (Virginia Bioinformatics Institute Core Laboratories). The EcoRI-HindIII fragment from pGEM-T encoding the genes of interest were was ligated into pKK223-3 (Amann et al 1983). This vector contains the P_{tac} promoter, which is inducible by isopropyl-β-D-thiogalactopyranoside (IPTG).

Qualitative glycogen production assays

Recombinant *E. coli* MG1655 strains were grown at 37°C in Luria-Bertani (LB) medium supplemented with ampicillin (100 μg/ml). The *E. coli* strains expressing *V. parahaemolyticus* RIMD CsrB1, CsrB2, CsrB3, CsrC, CsrA or containing the pKK223-3 empty vector (Table 3.2) were streaked on Kornberg agar plates (1.1% K_2HPO_4, 0.85% KH_2PO_4, 0.6% yeast extract, 1% glucose and 1.5% agar) plates supplemented with 100 μg/ml ampicillin and 0.2 mM IPTG from overnight cultures. The plates were incubated at 30°C for approximately 8 hours (until noticeable growth was present) then inverted over iodine crystals until a noticeable change in color could be detected.

Quantitative glycogen production assays

Recombinant *E. coli* MG1655 strains containing pKK223-3 or pVP1-5 (Table 3.2, encoding *csrA*, *csrB1*, *csrB2*, *csrB3* and *csrC* respectively) were grown in 5 mL LB containing 100 μg/mL ampicillin to late log phase and then subcultured to an OD_{600} of 0.05 into 100 mL of LB containing 100 ug/mL ampicillin and 0.2 mM IPTG. Fifty mL of cells at an OD_{600} of 1 were harvested by centrifugation (5000 rpm for 10 minutes) . The cells were stored at -20°C prior to being utilized in an assay. The cells were resuspended in 1.5 mL of H_20 and lysed via sonication (six 30 second bursts at 25% followed by 30 seconds rest; Fisher Scientific Sonic Dismembrator Model 500). The cell lysate was prepared by boiling the sample for 5 minutes to inactivate enzymes and then centrifuged at 13000 rpm for 5 minutes to separate insoluble material. The supernatant was assayed in triplicate from two independent

experimental sets according to the BioVision Glycogen Assay Kit (Mountain View, CA) instructions for fluorescence. A glycogen standard was provided with the kit and was diluted according to the kit instructions (from 0 to 0.2 ug glycogen per sample). Serial dilutions of the cell lysate were assayed to check that the fluorescence output was within the linear range of the assay (Table 3.3).

Cloning of transcriptional and translational fusion constructs

Transcriptional and translational promoter fusions to *lacZ* were created by PCR amplifying the promoter regions of interest (P_{dksA}, P_{glgC1}, P_{toxR}; Table 3.1) from *V. parahaemolyticus* RIMD chromosomal DNA. P_{dksA} and P_{glgC1} were chosen because *dksA* and *glgC* have been shown to be regulated by CsrA in *E. coli* (Adrianne Edwards, personal communication; reviewed in Timmermans and Van Melderen, 2010). As mentioned in Chapter 1, ToxR is a key virulence determinant in *V. parahaemolyticus*, so P_{toxR} was chosen in an attempt to connect CsrA directly to virulence in RIMD2210633. The promoters for the transcriptional fusions were amplified from the end of the upstream gene to, but not including, the ATG start codon of the gene of interest. The promoters for the translational fusions were amplified approximately 15 amino acids into the gene of interest to include the ribosome-binding site and were amplified to insert in-frame to the fusion vector. The amplified regions were ligated into the pGEM-T vector (Promega) and sequenced (Virginia Bioinformatics Institute). The BamHI/EcoRI fragment from pGEM-T was cloned into the respective vector, pSP417 (Podkovyrov and Larson,

1995) for the transcriptional fusion and pMC1403 (Casadaban et al., 1980) for the translational fusion (Table 3.2).

β-galactosidase assay strain construction

In order to create the necessary *E. coli* strains for assays, *csrA* (pVP1) and *csrB2* (pVP3) were cut from pKK223-3 with the P_{tac} promoter using BamHI and HindIII and ligated into pBBR1MCS2 (Kovach et al, 1995), which confers kanamycin resistance (Table 3.2). This was done so that the CsrA and CsrB2 overexpression constructs and the pBBR1MCS2 vector control could be co-transformed into competent *E. coli* MG1655 along with the fusion constructs (ampicillin resistant). The co-transformation was unsuccessful for the translational fusion for P_{toxR}, so pVP8 was transformed into competent *E. coli* MG1655 and the CsrA and CsrB2 overexpression constructs and pBBR1MCS2 vector control were then transformed into these strains. Strains were constructed to contain pBBR1MCS2, pBBR1MCS2 P_{tac} *csrA*, or pBBR1MCS2 P_{tac} *csrB2* and the respective transcriptional or translational fusions (Table 3.4). Qualitative iodine-staining plate assays were performed with and without IPTG induction to ensure CsrA levels were being regulated as expected in these strains.

β-galactosidase assays

β-Galactosidase assays were performed in order to quantify expression from transcriptional and translational fusions of promoters to *lacZ* in the presence of high and low levels of active CsrA protein. Cells were grown in 5 mL of LB with 100 μg/mL ampicillin and 50 μg/mL kanamycin to an OD_{600} of 0.5. 5 μL aliquots were

stored at -70°C prior to analysis of LacZ expression via chemiluminescent β-galactosidase assays (Applied Biosystems, Bedford, MA). The 5 µL aliquots were diluted 1:200 in Z Buffer (60 mM $Na_2PO_4.7H_20$, 40 mM $NaH_2PO_4.H_20$, 10 mM KCl, 1 mM $MgSO_4.7H_20$, 400 nM DTT) and permeabilized using 50 µL chloroform. LacZ expression was measured by using 10 µL of cell lysate, and the experiment was done in triplicate from two independent sets. Light output was measured at a wavelength of 492 nm with an integration time of 1.0 second using a single-point luminescence method on a LD-400S luminescence detector (Beckman Coulter; von Bodman et al, 2003).

One mL of cells was harvested by centrifugation for one minute at 15,000 rpm to be used to determine the protein concentration of the crude cell extract. This cell pellet was resuspended in 1 mL of Z buffer and 50 µL of chloroform to permeabilize the cells. The cell lysate was assayed for total protein concentration by the Bradford assay using Bio-Rad Protein Assay Dye Reagent Concentrate (Bio-Rad, Hercules, CA). Cell lysate (500 µL) was mixed with 200 µL of the protein dye and 300 µL of water and mixed by vortex for several seconds. The samples were incubated at room temperature for 20 minutes, and then the absorbance at 595 nm was taken. A BSA standard curve was used to determine the protein concentration in the crude cell extract.

The samples overexpressing CsrB2 clumped when grown due to increased glycogen production, so it was difficult to obtain an accurate OD_{600} value. In order to normalize the β-galactosidase levels, the total protein concentration for each

sample was divided by the total protein concentration for sample dksA-1. The dksA-1 sample was chosen because the cells did not clump and the culture was grown exactly to an OD_{600} of 0.5. This ratio was then multiplied by the individual RLU values to obtain a value relative to dksA-1. The relative RLU values were then averaged and the standard deviation calculated.

Results and Discussion

Qualitative glycogen production assays

The iodine-staining plate assay (Weilbacher et al, 2003) was used as a qualitative measure of glycogen production in the heterologous host *E. coli* MG1655 overexpressing *V. parahaemolyticus* RIMD2210633 CsrA, CsrB1, CsrB2, CsrB3, and CsrC to verify the presence and functionality of these sRNAs (Figure 3.1). The *E. coli* strain overexpressing CsrA stained visibly lighter than the pKK223-3 vector control indicating overexpression of CsrA leads to a decrease in glycogen accumulation in *E. coli*. The strains overexpressing the *V. parahaemolyticus* RIMD CsrA-regulating sRNAs showed noticeable darker staining than both the strain overexpressing *V. parahaemolyticus* CsrA and the pKK223-3 vector control. The darker staining indicates an increase in glycogen production caused by sRNA inactivation of CsrA. The *E. coli* strain overexpressing CsrC stains lighter than the other sRNAs and similarly to the pKK223-3 vector control, indicating there is some difference between the CsrA-regulation of CsrC compared to the CsrBs. When grown on LBS agar plates, the recombinant *E. coli* strains overexpressing CsrC produce noticeably larger colonies. In summary, the *csrA*, *csrB1*, *csrB2*, *csrB3*, and *csrC* genes from *V. parahaemolyticus* have been successfully cloned and have demonstrated biological activity in recombinant *E. coli*, indicating the *V. parahaemolyticus* RIMD CsrA-regulating sRNAs are capable of interacting with *E. coli* MG1655 CsrA. Initially, these assays were performed using *V. fischeri* CsrA while the *V. parahaemolyticus*

CsrA overexpression construct was being made, and the assay results were the same.

Quantitative glycogen production assays

Quantitative glycogen production assays were performed on recombinant *E. coli* MG1655 overexpressing *V. parahaemolyticus* RIMD CsrB1, CsrB2, CsrB3, CsrC and CsrA (Figure 3.2). The results were consistent with those found by the iodine-staining plate assays performed on the same strains. The strain overexpressing CsrA produced an average of 0.009 μg of glycogen per μL of cell lysate assayed, which is about an eight-fold decrease in the amount of glycogen produced by the pKK223-3 vector-only control (0.07 μg per μL of cell lysate assayed). The strains over expressing the CsrA-regulating sRNAs produced significantly higher amounts of glycogen than both the vector-only control and the strain overexpressing CsrA; however, these results had a high degree of error as is evident from the standard deviation. While this inconsistency could be a result of experimental or human error, given the consistency of the controls (pKK223-3 and overexpressing CsrA), the results may also be an indication of instability in the CsrA-regulating sRNAs themselves. Based on the averages, a trend emerges indicating that the CsrB sRNAs repress CsrA activity to a greater degree than CsrC, which is consistent with the differential staining of these three strains in the iodine-staining plate assay. However, the observed differences are not statistically significant.

β-galactosidase assays

For the β-galactosidase assays, CsrB2 was chosen to represent the CsrA-regulating sRNAs. Though CsrB3 shows potentially the highest level of repression of CsrA, it also proved to be the most inconsistent. CsrB2 showed the highest average glycogen production and was more consistent than CsrB3. Regardless of which CsrA-regulating sRNA was chosen to represent the others, all produced significantly more glycogen than the strain overexpressing CsrA indicating their ability to repress CsrA activity in *E. coli* MG1655.

β-Galactosidase assays were performed on the strains found in Table 3.3 in order to determine the effect of CsrA levels on the transcription and translation from P_{dksA}, P_{glgC1}, and P_{toxR} (Figure 3.3). CsrA is a RNA-binding protein that regulates gene expression post-transcriptionally, so the level of transcription was not expected to change in response to CsrA levels. However, transcriptional levels of β-galactosidase changed for all three samples. CsrA is a global regulator and it would appear that changing its levels in the cell may have indirectly altered transcription at these promoters. The ratios of transcription to translation was calculated by dividing the average β-galactosidase level of the transcriptional fusions by the translational fusions in order to determine differences in translation in response to CsrA levels (Table 3.5). There does not appear to be any correlation between CsrA level and translation of mRNAs directed from these promoters. While it is possible that these targets are not regulated by CsrA in this system, it appears that this system is not an effective method for testing *V. parahaemolyticus* CsrA targets.

Overall conclusions

The presence and functionality of RIMD2210633 CsrA and the CsrA-regulation sRNAs has been demonstrated in recombinant *E. coli*. RIMD2210633 CsrA-regulating sRNAs are capable of interacting with and down-regulating *E.coli* MG1655 CsrA, as is evident by the increase of glycogen production when the CsrA-regulating sRNAs are overexpressed in *E. coli* demonstrated both qualitatively and quantitatively. This relationship is maintained in the transcriptional/translational fusions system designed to test *V. parahaemolyticus* RIMD2210633 CsrA targets in *E. coli*. However, the results from the β-galactosidase assays are inconclusive in regards to CsrA regulation.

Table 3.1: Primers used in this study

Primer	Sequence (5' → 3')	Use
CsrAF	GAATTCATGCTAATTTTGACTCGCCGCG	CsrA-His_6 overexpression
CsrAHis2	AAGCTTGTGGTGGTGGTGGTGGTGTTAGTA GTTACCTGAAGCAACGTTGCC	CsrA-His_6 overexpression
CsrB1F	GAATTCGTCAGCAGGAAGCAGACACGGAAC AGG	CsrB1 overexpression
CsrB1R	AGCTTGAAAGACCCCGACACGAATGTATCG	CsrB1 overexpression
CsrB2F	GAATTCGTCGGAAGGATGCTGACACGGAACAGG	CsrB2 overexpression
CsrB2R	AAGCTTGCAGGTCGGGCGGGAAATGGCAGAAAAAG	CsrB2 overexpression
CsrB3F	GAATTCGTCGACAGGATGTTGGCGGGAACAGG	CsrB3 overexpression
CsrB3R	AAGCTTAAATAAAAACCCCGCCTGGTTTCCC	CsrB3 overexpression
CsrCF	GAATTCGCGTTCACAGGGGGTGAATTGCAGGATTTC	CsrC overexpression
CsrCR	AAGCTTAAACTGTGTTGCACGCATTTAGTGC	CsrC overexpression
DksAFEcoRI	GAATTCGAGAAAAAATGATGTCGTCTTCACATTGACG	P_{dksA} fusions
DksAtlfR	GGATCCGCTAGGATGCCGATCGTTTTCTTCTTTGC	P_{dksA} translational fusion
DksAtcfR	GGATCCACAGCATCTCCTACTTACACCTAGTCAACTGC	P_{dksA} transcriptional fusion
GlgC1FEcoRI	GAATTCACTCAACTTGTAAAATAAAGGCAGCAAATACG	P_{glgC1} fusions
GlgC1tlfR	GGATCCGCAAGAATCATTCCCAAAACACCAGCCAT	P_{glgC1} translational fusion
GlgC1tcfR	GGATCCTCGTTGTTCTCCATATCGTTATAGC	P_{glgC1} transcriptional fusion
ToxRFEcoRI	GAATTCTGTGTGCTCCATAACATCTAAACTAACAGG	P_{toxR} fusions
ToxRtlfR	GGATCCGTAAACCTTTGAGCAAGTAGAAATTTGGTGCC	P_{toxR} translational fusion
ToxRtcfR	GAATTCTGTGTGCTCCATAACATCTAAACTAACAGG	P_{toxR} transcriptional fusion

Table 3.2: Plasmids created for this study

Plasmid	Description	Antibiotic Resistance
pVP1	*csrA* in pKK223-3 for overexpression in *E. coli* flanked by HindIII and EcoRI sites	Ap^R
pVP2	*csrB1* in pKK223-3 for overexpression in *E. coli* flanked by HindIII and EcoRI sites	Ap^R
pVP3	*csrB2* in pKK223-3 for overexpression in *E. coli* flanked by HindIII and EcoRI sites	Ap^R
pVP4	*csrB3* in pKK223-3 for overexpression in *E. coli* flanked by HindIII and EcoRI sites	Ap^R
pVP5	*csrC* in pKK223-3 for overexpression in *E. coli* flanked by HindIII and EcoRI sites	Ap^R
pVP6	*dksA* promoter region *lacZ* translational fusion in pMC1403 flanked by BamHI and EcoRI sites	Ap^R
pVP7	*glgC1* promoter region *lacZ* translational fusion in pMC1403 flanked by BamHI and EcoRI sites	Ap^R
pVP8	*toxR* promoter region *lacZ* translational fusion in pMC1403 flanked by BamHI and EcoRI sites	Ap^R
pVP9	*dksA* promoter region *lacZ* transcriptional fusion in pSP417 flanked by BamHI and EcoRI sites	Ap^R
pVP10	*glgC1* promoter region *lacZ* transcriptional fusion in pSP417 flanked by BamHI and EcoRI sites	Ap^R
pVP11	*toxR* promoter region *lacZ* transcriptional fusion in pSP417 flanked by BamHI and EcoRI sites	Ap^R
pVP12	*csrB2* with P_{tac} promoter from pKK223-3 in pBBR1MCS2 flanked by BamHI and EcoRI sites	Km^R
pVP13	*csrA* with P_{tac} promoter from pKK223-3 in pBBR1MCS2 flanked by BamHI and EcoRI sites	Km^R

Table 3.3: Dilution and volume assayed for each strain for quantitative glycogen assay

	Sample set #1		**Sample set #2**	
Strain[a]	**Dilution of cell extract**	**Volume assayed**	**Dilution of cell extract**	**Volume assayed**
pKK223-3	1:20	10	1:20	10
pVP1 (CsrA)	1	10	1	5
pVP2 (CsrB1)	1:50	10	1:50	5
pVP3 (CsrB2)	1:50	10	1:50	5
pVP4 (CsrB3)	1:50	10	1:50	5
pVP5 (CsrC)	1:20	20	1:20	10

[a] *E. coli* MG1655 contained the indicated plasmid

Table 3.4: β-galactosidase assay strains.

	P_{dksA}		**P_{glgC1}**		**P_{toxR}**	
	pVP9[a]	**pVP6[b]**	**pVP10[a]**	**pVP7[b]**	**pVP11[a]**	**pVP8[b]**
pBBR1MCS2	dksA-1	dksA-4	glgC1-1	glgC1-4	toxR-1	toxR-4
pVP13 (CsrA)	dksA-2	dksA-5	glgC1-2	glgC1-5	toxR-2	toxR-5
pVP12 (CsrB2)	dksA-3	dksA-6	glgC1-3	glgC1-6	toxR-3	toxR-6

[a] transcriptional fusion
[b] translational fusion

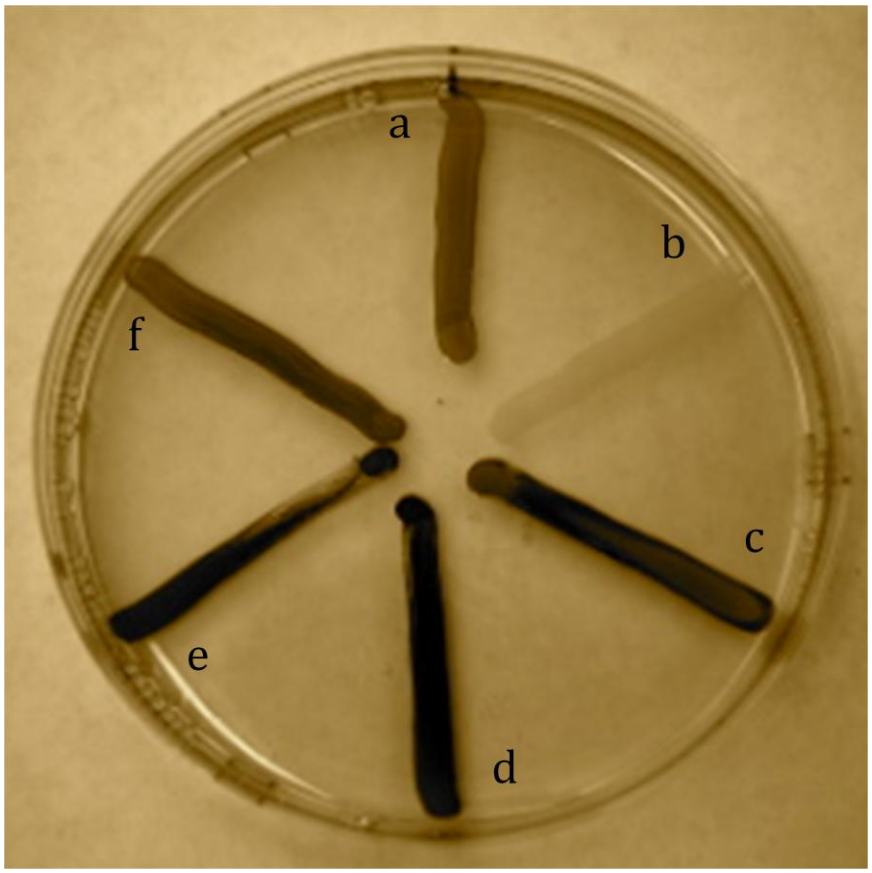

Figure 3.1: Effect of sRNA overexpression on glycogen production. Recombinant *E. coli* MG1655 overexpressing *V. parahaemolyticus* RIMD (c) CsrB1, (d)CsrB2, (e) CsrB3, and (f) CsrC. Controls were pKK223-3 (a) and *V. parahaemolyticus* RIMD CsrA (b).

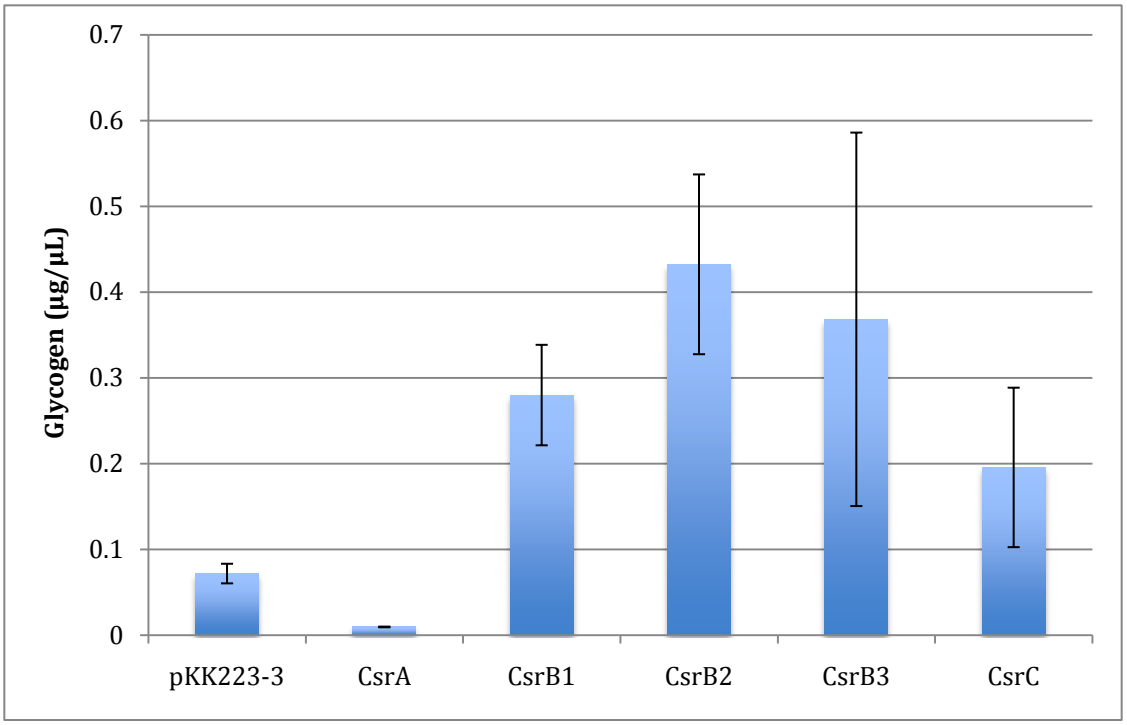

Figure 3.2: Effect of CsrA and sRNA overexpression on glycogen production. Recombinant *E. coli* MG1655 overexpressing *V. parahaemolyticus* RIMD CsrA, CsrB1, CsrB2, CsrB3 and CsrC, with pKK223-3 as a vector only control. The x-axis shows the protein or sRNA being overexpressed in *E. coli* MG1655. The y-axis gives the amount of glycogen in µg per µL of cell lysate assayed. The error bars represent +/- one standard deviation from the average.

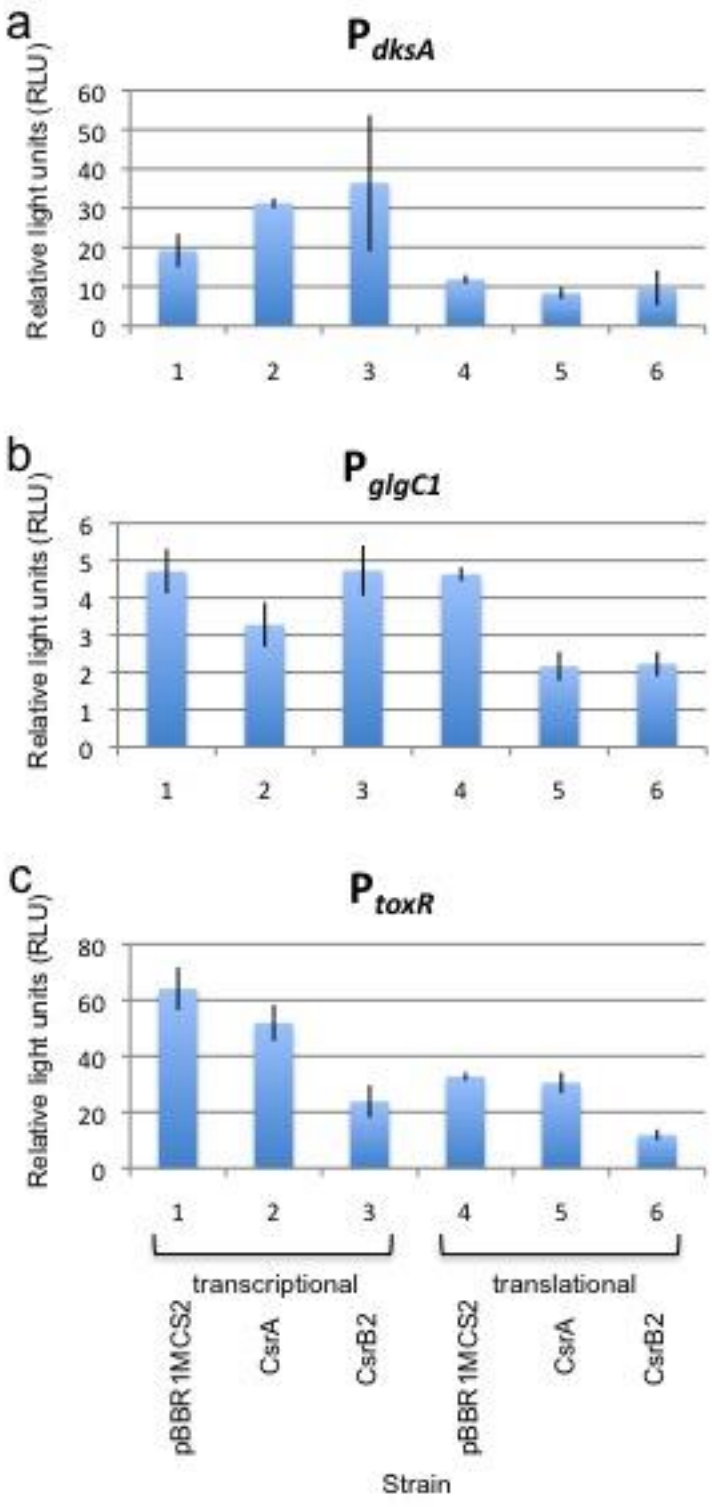

Figure 3.3: β-galactosidase activity assay results: a) P_{dksA}, b) P_{glgC1}, c) P_{toxR}. The strain numbers on the x-axis correspond to those found in Table 3.4. The y-axis gives the average relative light units. The data in this figure has been normalized to the protein concentration. The error bars represent +/- one standard deviation from the average.

Table 3.5: Ratio of transcription and translation of β-galactosidase from selected promoters.

	$\mathbf{P}_{dksA}$	$\mathbf{P}_{glgC1}$	$\mathbf{P}_{toxR}$
pBBR1MCS2	1.6243	1.017	1.9557
CsrA	3.7445	1.514	1.7013
CsrB2	3.7500	2.120	2.0056

Chapter 4

Overall conclusions

The *Vibrio parahaemolyticus* BB22OP Genome

The comparison of the *V. parahaemolyticus* BB22OP and RIMD2210633 genomes is providing important insights into both the conserved colonization and virulence factors essential to the infectious process utilized by *V. parahaemolyticus*, as well as the distinctive genetic determinants in the two strains under investigation. Conserved genes are likely important to general metabolism and colonization. However, the two strains under investigation also exhibit differential phenotypes. Strain RIMD2210633 is constitutively virulent. It is hypothesized that this is due to a constitutively active quorum-sensing system. On the other hand, only the translucent strain of *V. parahaemolyticus* BB22 is virulent, due to inactivation of OpaR. Unique genes to each strain may ultimately be found to have an influence on virulence. Genomic comparisons have already provided new information about differential regulation between these two strains. For example, there are five transcriptional regulators unique to the BB22OP genome. Due to the tiered effect of transcriptional regulation, is it is to be expected these transcriptional regulators will be found to have a profound effect on gene regulation in the BB22OP strain.

To begin defining these regulatory networks, in the future, the direct targets of OpaR could be established by utilizing a modified ChIP sequencing procedure. Additionally, it would be interesting to understand the entire breadth of the quorum-sensing regulon via total transcriptome-level analysis using next generation sequencing technologies. Once the quorum-sensing regulons are defined

for each strain, it will need to be determined where the systems diverge. It has been hypothesized that *V. parahaemolyticus* RIMD2210633 has maintained a mutation in *luxO* that results in a constitutively active LuxO protein, meaning the quorum-sensing system would be perpetually turned on and OpaR is never expressed. This mutation would result in the constitutively pathogenic phenotype. Quorum sensing likely controls the gene expression of a significant number of genes, so if this is the case, how does the RIMD2210633 strain compensate?

There are numerous hypothetical genes in the "boneyard" of *V. parahaemolyticus* BB22OP. More importantly, there are at least 58 ORFs, potentially many more if the ORF Finder results are considered, which show no significant homology to anything found in the non-redundant protein database. These genes may turn out to be truly novel to *V. parahaemolyticus* BB22OP. It would be interesting to see if any of these ORFs encode a functional protein that contributes to the unique phenotypes and regulatory patterns of *V. parahaemolyticus* BB22OP. The fully assembled and annotated *V. parahaemolyticus* BB22OP genome will provide the platform for a complete comparative analysis between these two strains, which are of particular interest due to the variations in their phenotypic profiles.

CsrA

It has been established that *V. parahaemolyticus* RIMD2210633 CsrA is regulated by four CsrA-regulating sRNAs in a similar fashion to *E. coli* CsrA. The presence of the same four CsrA-regulating sRNAs has been established in *V.*

parahaemolyticus BB22OP. It is expected that CsrA is regulated in the same manner in BB22OP.

Unfortunately, this project was not successful in identifying any RIMD2210633 CsrA targets using a transcriptional/translation fusion reporter system. It may be necessary to take a more direct approach in determining CsrA targets in *V. parahaemolyticus*. CsrA is an RNA-binding protein that regulates gene expression post-transcriptionally by binding mRNA targets and influencing their translation. CsrA could be co-purified with its mRNA targets, and then the mRNA could be reverse transcribed into cDNA and sequenced. The results would provide the direct RIMD2210633 CsrA targets.

Final Remarks

By sequencing the *V. parahaemolyticus* BB22OP genome, the platform has been provided for a comparative analysis between the BB22OP and RIMD2210633 genomes. The two strains are distinctly different, but share significant sequence homology. Due to the highly conserved genomes, it is plausible that the differential expression of phenotypes is largely a result of regulatory differences as opposed to genetic differences. That being said, it will be interesting to see how the unique genes contribute to the individuality of these two strains. Further comparative analysis and genetic manipulations will reveal the function and physiological role of the unique genes to *V. parahaemolyticus* BB22OP and this information can be extrapolated to RIMD2210633. Collectively, studies of these two pathogenic strains

will contribute to a more complete understanding of the virulence of *V. parahaemolyticus*, an emerging human pathogen.

References

Altier, C., M. Suyemoto, and S. D. Lawhon. 2000. Regulation of *Salmonella enterica* serovar typhimurium invasion genes by csrA. Infect Immun **68:**6790-7.

Amann, E., Brosius, J., and Ptashne, M. 1983. Vectors bearing a hybrid *trp-lac* promoter useful for regulated expression of cloned genes in *Escherichia coli*. Gene **25:** 167-178.

Bauer, A. and Rørvik L. M. 2007. A novel multiplex PCR for the identification of *Vibrio parahaemolyticus*, *Vibrio cholerae* and *Vibrio vulnificus*. Lett Appl Microbiol **45:** 371-375.

Cambray, G., Guerout, A. M., and Mazel, D. 2010. Integrons. Annu Rev Genet **44:** 141-66.

Cao, J. G., and Meighen, E. A. 1989. Purification and structural identification of an autoinducer for the luminescence system of *Vibrio harveyi*. J Bacteriol **264:**21670-21676.

Casadaban, M. J., Chou, J., and Cohen, S. N. 1980. In vitro gene fusions that join an enzymatically active beta-galactosidase segment to amino-terminal fragments of exogenous proteins: *Escherichia coli* plasmid vectors for the detection and cloning of translational initiation signals. J Bacteriol **143:** 971-980.

Chen, X., Schauder, S., Potier, N., Van Dorsselaer, A., Pelczer, I., Bassler, B. L., and Hughson, F. M. 2002. Structural identification of a bacterial quorum-sensing signal containing boron. Nature **415:**545-549.

CDC. 2005. *Vibrio* Illnesses After Hurricane Katrina – Multiple States, August – September 2005. MMWR **54(37):** 928-931.

CDC. 1999. Outbreak of *Vibrio parahaemolyticus* infection associated with eating raw oysters and clams harvested from Long Island Sound- Connecticut, New Jersey, and New York, 1998. MMWR **48**.

Daniels, N. A., Ray, B., Easton, A., Marano, N., Kahn, E., McShan, A. L. Del Rosario, L., Baldwin, T., Kingsley, M. A., Puhr, N. D., Wells, J. G., and Angulo, F. J. 2000. Emergence of a new *Vibrio parahaemolyticus* serotype in raw oysters: A prevention quandary. JAMA **284:** 1541-1545.

Enos-Berlage, J. L., and McCarter, L. L. 2000. Relation of capsular polysaccharide production and colonial cell organization to colony morphology in *Vibrio parahaemolyticus*. J Bacteriol **182:** 5513-5520.

FDA. 2001. Draft risk assessment on the public health impact of *Vibrio parahaemolyticus* in raw molluscan shellfish online. <www.cfsan.fda.gov>.

Fuqua, C., M. R. Parsek, and E. P. Greenberg. 2001. Regulation of gene expression by cell-to-cell communication: acyl-homoserine lactone quorum sensing. Annu Rev Genet **35:**439-68.

Georgia Tech. GeneMark™- Free Gene Prediction Software. Web. 23 Nov. 2010. <http://exon.biology.gatech.edu>.

Gottesman, S. 2005. Micros for microbes: non-coding regulatory RNAs in bacteria. Trends Genet **21:** 399-404.

Hammer, B. K, and Bassler, B. L. 2007. Regulatory small RNAs circumvent the conventional quorum sensing pathway in pandemic *Vibrio cholerae*. Proc Natl Acad Sci U S A **104:** 11145-11149.

Higgins, D. A., Pomianek, M. E., Kraml, C. M., Taylor, R. K., Semmelhack, M. F., and Bassler, B. L. 2007. The major *Vibrio cholerae* autoinducer and its role in virulence factor production. Nature **450:**883-886.

Honda, T., Iida, T., Akeda, Y., and T. Kodama. 2008. Sixty years of *V. parahaemolyticus* research. Microbe **3:** 462-466.

Honda, T., Y. Ni, T. Miwatani, T. Adachi, and J. Kim. 1992. The thermostable direct hemolysin of *Vibrio parahaemolyticus* is a pore-forming toxin. Can J Microbiol **38:**1175-80.

Jackson, D. W., K. Suzuki, L. Oakford, J. W. Simecka, M. E. Hart, and T. Romeo. 2002. Biofilm formation and dispersal under the influence of the global regulator CsrA of *Escherichia coli*. J Bacteriol **184:**290-301.

Jaques, S., and McCarter, L. L. 2006. Three new regulators of swarming in *Vibrio parahaemolyticus*. J Bacteriol **188:** 2625-2635.

Kawamoto, H., Koide, Y., Morita, T., and Aiba H. 2006. Base-pairing requirement for RNA silencing by a bacterial small RNA and acceleration of duplex formation by Hfq. Mol Microbiol **61:** 1013-1022.

Kay, E., Humair, B., Dénervaud, V., Riedel, K., Spahr, S., Eberl, L, Valverde, C., and Haas, D. 2006. Two GacA-dependent small RNAs modulate the quorum-sensing response in *Pseudomonas aeruginosa*. J Bacteriol **188:** 6026-6033.

Kim, Y., Kim, B. S., Park, Y. J., Choi, W. C., Hwang, J., Kang, B. S., Oh, T. K., Choi, S. H., and Kim, M. H. 2010. Crystal structure of SmcR, a quorum-sensing master regulator of *Vibrio vulnificus*, provides insight into its regulation of transcription. J Biol Chem **285:** 14020-30.

Kim, Y. B., Okuda, J., Matsumoto, C., Takahashi, N., Hasimoto, S., and Nishibuchi, M. 1999. Identification of *Vibrio parahaemolyticus* strains at the species level by PCR targeted to the *toxR* gene. J Clin Microbiol **37:** 1173-1177.

Kovach, M. E., Elzer, P. H., Hill, D. S., Robertson, G. T., Farris, M. A, Roop, R. M. 2nd, and Peterson, K. M. 1995. Four new derivatives of the broad-host-range cloning vector pBBR1MCS, carrying different antibiotic-resistance cassettes. Gene **166:**175-176.

Kulkarni, P. R., X. Cui, J. W. Williams, A. M. Stevens, and R. V. Kulkarni. 2006. Prediction of CsrA-regulating small RNAs in bacteria and their experimental verification in *Vibrio fischeri*. Nucleic Acids Res **34:**3361-3369.

Lenz, D. H., Miller, M. B., Zhu, J., Kulkarni, R. V., and Bassler, B. L. 2005. CsrA and three redundant small RNAs regulate quorum sensing in *Vibrio cholerae*. Mol Microbiol **58:** 1186-1202.

Liu, M. Y., Gui, G., Wei, B., Preston J. F., Oakford, L., Yüksel, U., Giedroc, D. P., and Romeo T. 1997. The RNA molecule CsrB binds to the global regulatory protein CsrA and antagonizes its activity in *Escherichia coli*. J Biol Chem **272:** 17502-17510.

Ma, W., Y. Cui, Y. Liu, C. K. Dumenyo, A. Mukherjee, and A. K. Chatterjee. 2001. Molecular characterization of global regulatory RNA species that control pathogenicity factors in *Erwinia amylovora* and *Erwinia herbicola* pv. gypsophilae. J Bacteriol **183:**1870-80.

Makino, K., K. Oshima, K. Kurokawa, K. Yokoyama, T. Uda, K. Tagomori, Y. Iijima, M. Najima, M. Nakano, A. Yamashita, Y. Kubota, S. Kimura, T. Yasunaga, T. Honda, H. Shinagawa, M. Hattori, and T. Iida. 2003. Genome sequence of *Vibrio parahaemolyticus*: a pathogenic mechanism distinct from that of *V cholerae*. Lancet **361:**743-9.

Massé, E., Vanderpool, C. K., and Gottesman, S. 2005. Effect of RyhB small RNA on global iron use in *Escherichia coli*. J Bacteriol **187:** 6962-6971.

McCarter, L. L. 1998. OpaR, a homolog of *Vibrio harveyi* LuxR, controls opacity of *Vibrio parahaemolyticus*. J Bacteriol **180:** 3166-3173.

McDougald, D., Rice, S. A., and Kjelleberg, S. 2000. The marine pathogen *Vibrio vulnificus* encode a putative homologue of the *Vibrio harveyi* regulatory gene, *luxR*: a genetic and phylogenetic comparison. Gene **248:** 213-21.

Miller, V. L., Taylor, R. K., and Mekalanos, J. J. 1987. Cholera toxin transcriptional activator ToxR is a transmembrane DNA binding protein. Cell **48:** 271-279.

Nasu, H., Iida, T., Sugahara, T., Yamaichi, Y., Park, K. S., Yokoyama, K., Makino, K., Shinagawa, H., and Honda, T. 2000. A filamentous phage associated with recent pandemic *Vibrio parahaemolyticus* O3:K6 strains. J Clin Microbiol **38:**2156-61.

NCBI. ORF Finder. Web. 23 Nov. 2010. <http://www.ncbi.nlm.nih.gov/gorf/gorf.html>

Nealson, K. H., Platt, T., and Hastings, J. W. 1970. Cellular control of the synthesis and activity of the bacterial luminescent system. J Bacteriol **104:**313-322.

Okuda, J., Ishibashi, M., Hayakawa, E., Nishino, T., Takeda, Y., Mukhopadhyay, A. K., Garg, S., Bhatacharya, S. K., Nair, G. B., Mishibuchi, M. 1997. Emergence of a unique O3:K6 clone of *Vibrio parahaemolyticus* in Calcutta, India, and isolation of strains from the same clonal group from Southeast Asian travelers arriving in Japan. J Clin Microbiol **35:** 3150-3155.

Opdyke, J. A., Kang, J. G., and Storz, G. 2004. GadY, a small-RNA regulator of acid response genes in *Escherichia coli*. J Bacteriol **186:** 6698-6705.

Osorio, C. R., and Klose, K. E. 2000. A region of the transmembrane regulatory protein ToxR that tethers the transcriptional activation domain to the cytoplasmic membrane displays wide divergence among *Vibrio* species. J Bacteriol **182:** 526-528.

Podkovyrov, S. M. and Larson, T. J. 1995. A new vector-host system for construction of *lacZ* transcription fusions where only low-level gene expression is desirable. Gene **156:** 151-152.

Pompeani, A. J., Irgon, J. J., Berger, M. F., Bulyk, M. L., Wingreen, N. S., and Bassler, B. L. 2008. The *Vibrio harveyi* master quorum-sensing regulator, LuxR, a TetR-type protein is both an activator and a repressor: DNA recognition and binding specificity at target promoters. Mol Microbiol **70:** 76-88.

Products & Solutions - System Features: 454 Life Sciences, a Roche Company. Web. 23 Nov. 2010. <http://www.454.com/products-solutions/system-features.asp>.

Reading, N. C., and V. Sperandio. 2006. Quorum sensing: the many languages of bacteria. FEMS Microbiol Lett **254:**1-11.

Romeo, T. 1998. Global regulation by the small RNA-binding protein CsrA and the non-coding RNA molecule CsrB. Mol Microbiol **29:**1321-30.

Sambrook, J., Fritsch, E. F., Maniatis, T. 1989. Molecular Cloning: A Laboratory Manual, 2nd edition. Cold Spring Harbor, NY Cold Spring Harbor Laboratory Press.

Sochard, M. R., and R. R. Colwell. 1977. Toxin isolation from a Kanagawa-phenomenon negative strain of *Vibrio parahaemolyticus*. Microbiol Immunol **21:**243-54.

Stewart, B. J., and McCarter, L. L. 2003. Lateral flagellar gene system of *Vibrio parahaemolyticus*. J Bacteriol **185:** 4508-4518.

Sutherland, I. 2001. Biofilm expolysaccharides: a strong and sticky framework. Microbiology **147:**3-9.

Suzuki, K., X. Wang, T. Weilbacher, A. K. Pernestig, O. Melefors, D. Georgellis, P. Babitzke, and T. Romeo. 2002. Regulatory circuitry of the CsrA/CsrB and BarA/UvrY systems of *Escherichia coli*. J Bacteriol **184:**5130-40.

Svenningsen, S. L., Waters, C. M., and Bassler, B. L. 2008. A negative feedback loop involving small RNAs accelerates *Vibrio cholerae's* transition out of quorum-sensing mode. Genes Dev **22:** 226-238.

Timmbermans, J. and Van Melderen, L. 2010. Post-transcriptional global regulation by CsrA in bacteria. Cell Mol Life Sci **67:**2897-2908.

Tomizawa, J., and Itoh T. 1981. Plasmid ColE1 incompatibility determined by interaction of RNA I with primer transcript. Proc Natl Acad Sci U S A **78:** 6096-6100.

Tramonti, A., De Canio, M., and De Biase, D. 2008. GadX/GadY-dependent regulation of the *Escherichia coli* acid fitness island: transcriptional control at the *gadY-gadW* divergent promoters and identification of four novel 42 bp GadX/GadW-specific binding sites. Mol Microbiol **70:** 965-982.

Visick, K. L., and Fuqua, C. 2005. Decoding microbial chatter: cell-cell communication in bacteria. J Bacteriol **187:**5507-5519.

von Bodman, S. B., Ball, J. K., Faini, M. A., Herrera, C. M., Minogue, T. D., Urbanowski, M. L., and Stevens, A. M. 2003. The quorum sensing negative regulators EsaR and ExpR(Ecc), homologues with the LuxR family, retain the ability to function as activators of transcription. J Bacteriol **185:** 7001-7.

Waters, C. M., and B. L. Bassler. 2005. Quorum sensing: cell-to-cell communication in bacteria. Annu Rev Cell Dev Biol **21:**319-346.

Waters, L. S., and Storz, G. 2009. Regulatory RNAs in bacteria. Cell **136:** 615-628.

Weilbacher, T., Suzuki, K., Dubey, A. K., Wang, X., Gudapaty, S., Morozov, I., Baker, C. S., Georgellis, D., Babitzke, P. and Romeo, T. 2003. A novel sRNA component of the carbon storage regulatory system of *Eschericia coli*. Mol. Microbiol **48:**657-670.

Appendix 1: Novel BB22OP gene annotation

Contig #	Number	Strand	from	to	length	Annotation[b]	e-value
4	1	-	<3	104	102[a]	No significant similarity found	-
4	2	+	271	1497	1227	site-specific recombinase, phage integrase family (*Vibrio mimicus* VM223)	0
4	3	-	1601	3214	1614	hypothetical protein Bpse14_40543 (*Burkholderia pseudomallei* 14), Hypoth_Ymh superfamily domain, HATPase_c superfamily domain	0
4	4	-	3310	4488	1179	cytosine specific DNA methyltransferase (*Escherichia coli* SE15), AdoMet_Mtases superfamily domain	2.00E-145
4	5	+	4680	5153	474	DNA repair protein RadC (*Vibrio cholerae* CT 5369-93), MPN superfamily domain	7.00E-87
4	6	+	5150	5587	438	hypothetical protein VV0524 (*Vibrio vulnificus* YJ016), DUF2787 superfamily domain	4.00E-77
4	7	+	5639	6082	444	hypothetical protein (*Vibrio cholerae* O1 biovar El tor), DUF2787 superfamily domain	8.00E-80
4	8	+	6086	6262	177	hypothetical protein VV0522 (*Vibrio vulnificus* YJ016)	5.00E-22
4	9	+	6401	7135	735	hypothetical protein VFA_000653 (*Vibrio furnissii* CIP 102972)	2.00E-137
4	10	+	7166	7534	369	hypothetical protein VCB_000202 (*Vibrio cholerae* TMA 21)	2.00E-61
4	11	+	7582	7809	228	hypothetical protein VMA_000184 (*Vibrio mumicus* VM223)	3.00E-30
4	12	+	8103	8213	111	hypothetical protein VCB_000200 (*Vibrio cholerae* TMA 21)	3.00E-10
4	13	-	8242	9342	1101	hypothetical protein VMB_20200 (*Vibrio mimicus* VM603)	0
4	14	-	9467	9904	438	ribonuclease HI (*Vibrio cholerae* TMA 21)	9.00E-79
4	15	-	9974	10174	201	transcriptional regulator (*Vibrio* sp. RC341), Phage_AlpA superfamily domain	5.00E-31

4	16	-	10243	10842	600	hypothetical protein VC0496 (*Vibrio cholerae* O1 biovar El tor str. N16961), DUF3296 domain	6.00E-112
4	17	-	10895	11569	675	hypothetical protein VCJ_000312 (*Vibrio* sp. RC341), DUF3296 domain	1.00E-128
4	18	-	11631	12311	681	CP4-6 prophage conserved protein (*Vibrio mimicus* VM603), DUF3296 domain	2.00E-131
4	19	+	12864	13748	885	hypothetical protein VMA_00179 (*Vibrio mimicus* VM223)	2.00E-165
7	1	+	552	803	252	conserved hypothetical protein (*Vibrio parahaemolyticus* 16)	4.00E-07
7	2	+	1051	1983	933	hypothetical protein V12G01_21268 (*Vibrio alginolyticus* 12G01) contains dsrB domain	0
7	3	+	1973	2401	429	hypothetical protein V12G01_21273 (*Vibrio alginolyticus* 12G01)	3.00E-76
7	4	+	2894	5035	2142	KAP family P-loop domain protein (*Teredinibacter turnerae* T7901)	0
7	5	+	5151	5618	468	No significant hits	-
7	6	+	6321	6572	252	hypothetical protein BpOF4_03635 (*Bacillus pseudofirmus* OF4)	2.00E-07
7	7	-	6629	6805	177	No significant hits	-
7	8	+	7224	8201	978	hypothetical protein (*Pseudomonas mendocina*)	8.00E-129
7	9	-	8217	9797	1581	hypothetical protein PE36_07332 (*Moritella* sp. PE36)	0
8	1	+	<3	110	108[a]	N-terminal end contains P-loop NTPase superfamily domain no hits with sig. e-value	-
8	2	+	198	2537	2340	hypothetical protein A79_5175 (*Vibrio parahaemolyticus* AQ3810)	0
8	3	+	2531	4942	2412	site-specific recombinase, phage integrase family domain protein, putative (*Vibrio parahaemolyticus* AQ3810)	0
8	4	+	4954	5484	531	conserved hypothetical protein (*Vibrio parahaemolyticus* AQ3810)	5.00E-92

8	5	+	5484	6848	1365	site-specific recombinase, phage integrase family protein (*Vibrio alginolyticus* 12G01) add. Conserved domains on NCBI	0
8	6	-	7069	13215	6147	hypothetical protein MED222_16571 (*Vibrio* sp. MED222)	0
8	7	-	13302	14177	876	hypothetical protein MED222_16576 (*Vibrio* sp. MED222)	1.00E-168
8	8	-	14787	15578	792	amidase family protein (*Proteus mirabilis*)	5.00E-136
8	9	+	15637	15765	129	No significant similarity found	-
9	1	+	<2	103	102[a]	No significant similarity found	-
9	2	+	196	1260	1065	dTDP-D-glucose 4;6dehydratase (*Vibrio parahaemolyticus* AN-5034), NADB_Rossman superfamily domain	0
9	3	+	1260	2126	867	glucose-1-phosphate-thymidylyltransferase (*Vibrio alginolyticus* 12G01), Glyco_transf-GTA-type superfamily domain	1.00E-162
9	4	+	2127	2546	420	WblP protein (*Vibrio alginolyticus* 12G01), Cupin_2 superfamily domain	9.00E-75
9	5	+	2524	2994	471	WxcM-like protein (*Vibrio alginolyticus* 12G01), LbetaH superfamily domain, PRK12461 domain	2.00E-82
9	6	+	2987	4090	1104	WblQ protein (*Vibrio parahaemolyticus* AN-5034), numerous domains predicted - too many to list	0
9	7	+	4096	5025	930	hypothetical protein VparAN_08535 (*Vibrio parahaemolyticus* AN-5034), Glyco_tranf_GTA_type superfamily domain, WcaA domain	2.00E-177
9	8	-	5083	6012	930	hypothetical protien VparAN_08540 (*Vibrio parahaemolyticus* AN-5034), UPF0104 superfamily domain	9.00E-160
9	9	-	6022	6435	414	acid phosphatase/vanadium-dependent haloperoxidase related protein (*Vibrio parahaemolyticus* AN-5034), DUF212 superfamily domain	1.00E-73
9	10	-	6457	7308	852	putative prenyltransferase (*Vibrio parahaemolyticus*), UbiA superfamily domain	9.00E-162
9	11	+	7350	7745	396	hypothetical protein (*Vibrio parahaemolyticus*), GtrA superfamily domain	3.00E-70

9	12	+	7747	9030	1284	oxidoreductase, FAD-binding, putative (*Vibrio parahaemolyticus* AN-5034), FAD_binding_4 superfamily domain, ALD superfamily domain, GlcD domain	0
9	13	+	9033	9770	738	short chain dehydrogenase (*Vibrio parahaemolyticus* AN-5034), NADB_Rossman superfamily domain, DltE domain	6.00E-140
9	14	+	9767	11389	1623	hypothetical protein VparAN_80570 (*Vibrio parahaemolyticus* AN-5034)	0
9	15	-	11417	12562	1146	3-deoxy-D-manno-octulosonic-acid transferase (*Vibrio parahaemolyticus* AN-5034), Glycos_transf_N superfamily domain, Glycos_transf_1 superfamily	2.00E-166
9	16	-	12826	13695	870	putative α-1,2-fucosyltransferase (*Vibrio parahaemolyticus* AN-5034)	3.00E-140
9	17	-	13716	14771	1056	ADP-heptose-LPS hyptosyltransferase II (*Vibrio parahaemolyticus* AN-5034), GT1_LPS_heptosyltransferase domain, Glycosyltransferase_GTB_type superfamily domain, RfaF domain	0
9	18	-	14768	15490	723	glycosoly transferase family protein (*Vibrio parahaemolyticus* AN-5034)	1.00E-57
9	19	-	15487	>15618	132[a]	lipid A biosynthesis (KDO)2-(lauroyl)-lipid IVA acyltransferase (*Vibrio parahaemolyticus* AN-5034)	2.00E-05
10	1	-	<1	4596	4596[a]	MshA biogenesis protein MshQ (*Vibrio parahaemolyticus* 16	3.00E-79
12	1	-	46	951	906	No significant hits	-
13	1	+	102	1094	993	OtnB protein (*Vibrio parahaemolyticus* RIMD2210633), Wzz superfamily domain	6.00E-170
13	2	+	1281	2405	1125	UDP-N-acetylglucosamine 2-epimerase (*Vibrio vulnificus* YJ016), GT1_UDP-GlcNAc_2-Epimerase domain, Glycosyltransferase_GTB_type superfamily domain, wecB domain	0
13	3	+	2423	3688	1266	UDP-N-acetyl-D-mannosamine dehydrogenase (*Vibrio vulnificus* CMCP6), UDPG_MGDP_dh superfamily domain, UDPTG_MGDP_dh_C superfamily domain, wecC domain	0
13	4	+	3791	5239	1449	polysaccharide biosynthesis protein (*Geobacter metallireducens* GS-15), Polysacc_synt superfamily domain	1.00E-16

13	5	+	5287	6357	1071	WcgA (*Bacteroides fragilis*)	2.00E-52
13	6	+	6354	7541	1188	glycosyl transferase, group 1(*Methanoculleus marisnigri* JR1), GT1_wlbH_like domain, Glycosyltransferase_GTB_type superfamily domain	2.00E-32
13	7	+	7538	8641	1104	putative glycosyltransferase (*Bacteroides fragilis* 3_1_12), GT1_YqgM_like domain, Glycosyltransferase_GTB_type superfamily domain, RfaG domain	1.00E-28
13	8	+	8662	9813	1152	No significant hits	-
13	9	+	9806	10891	1086	glycosyl transferase, group 1 family protein (*Shigella dysenteriae* 1012), GT1_YqgM_like domain, Glycosyltransferase_GTB_type superfamily domain	1.00E-74
13	10	+	10892	11998	1107	GDP-mannose 4,6-dehyratase (*Vibrio parahaemolyticus* AN-5034), NADB_Rossman superfamily domain, Gmd domain	0
13	11	+	12082	13041	960	putative nucleotide di-P-sugar epimerase or dehydratase (*Vibrio parahaemolyticus* AN-5034), Epimerase domain	0
13	12	+	13069	13539	471	putative GDP-mannose mannosylhydrolase (*Vibrio parahaemolyticus* AN-5034), GCPMH domain, Nudix_Hydrolase superfamily domain,	1.00E-83
13	13	+	13577	15019	1443	mannose-1-phosphate guanylyltransferase/mannose-6-phosphate isomerase (*Vibrio parahaemolyticus* AN-5034), GDP-M1P_Guanylyltransferase domain, Glyco_tranf_GTA_type superfamily domain, Cupin_2 superfamily domain, GMP_PMI domain	0
13	14	+	15042	16469	1428	phosphomannomutase (*Vibrio parahaemolyticus* AN-5034), ManB domain, phosphohexomutase superfamily domain, ManB domain	0
13	15	+	16544	17725	1182	mannose-6-phosphate isomerase (*Vibrio* sp. RC586), manA superfamily domain (2x), PMI_typeI domain	7.00E-107
13	16	+	17733	18473	741	glycosyl transferase (*Helicobacter pylori* P12), GT_2_WfgS_like domain, Glyco_transf_GTA_type superfamily domain	6.00E-61

13	17	+	18470	19414	945	probably UDP-galactose 4-epimerase (*Vibrio vulnificus*), lots of putative domains	1.00E-127
13	18	+	19420	19968	549	undecaprenyl-phosphate β-N-acetyl-D-fucosaminephosphotransferase (*Vibrio fischeri* ES114), Bac_transf superfamily domain	3.00E-92
13	19	+	20150	>20548	399[a]	putative epimerase/dehydratase (*Vibrio parahaemolyticus* RIMD 221633)	3.00E-07
14	1	-	73	708	636	unnamed protein product (*Vibrio parahaemolyticus*), numerous hits to DNA polymerase III epsilon subunit with lower e-values	2.00E-39
14	2	-	779	2620	1842	Signal transduction protein (*Grimontia hollisae* CIP 101886), CAP_ED superfamily daomin, CBS_pair superfamily domain, NT_Pol-beta-like superfamily domain, DUF294_C superfamily domain, COG2905 domain	0
14	3	-	2684	4315	1632	Na+/glucose symporter (*Vibrio parahaemolyticus*), SSF superfamily domain, PRK10484 domain	0
15	1	-	73	507	435	hypothetical protein VIBHAR_01804 (*Vibrio harveyi* ATCC BAA-1116)	2.00E-29
15	2	-	507	1073	567	Archaeal/vacuolar-type H+-ATPase subunit B (*Vibrio* sp. Ex25)	5.00E-52
16	1	+	293	895	603	hypothetical protein V12G01_04806 (*Vibrio alginolyticus* 12G01)	2.00E-87
16	2	+	895	2088	1194	predicted lipase (*Vibrio alginolyticus* 12G01), Lipase_3 domain, Esterase_lipase superfamily domain	0
17	1	-	21	707	687	HAD-superfamily hydrolase, subfamily 1A, varient 1 (*Paenibacillus* sp. JDR-2), HAD_like domain, HAD_like superfamily domain, COG1011 domain	4.00E-31
18	1	+	372	965	594	No significant hits	-
18	2	+	965	2056	1092	No significant hits	-
18	3	+	2056	2394	339	No significant hits	-
18	4	-	2465	2890	426	hypothetical protein VIC_004034 (*Vibrio coralliilyticus* ATCC BAA-450), NTF2_like superfamily domain	3.00E-47

18	5	+	3078	3416	339	hypothetical protein P3TCK_01320 (*Photobacterium profundum* 3TCK), WHTH_GntR superfamily domain	7.00E-41
19	1	-	35	847	813	No significant hits	-
19	2	+	870	1004	135	No significant hits	-
19	3	-	1189	>1851	663[a]	No significant hits	-
21	1	+	<3	164	162[a]	hypothetical protein VMC_26540 (*Vibrio alginolyticus* 40B), hits to glyoxalase/blemycin resistance protein with higher e-values	5.00E-20
21	2	+	327	746	420	Acetyltransferases, including N-acetylases of ribosomal proteins (*Vibrio* sp. Ex25), GNAT superfamily domain	8.00E-59
21	3	+	900	1526	627	putative threonine efflux protein (*Vibrio vulnificus* CMCP6), LysE superfamily domain	3.00E-101
21	4	+	2559	2798	240	No significant similarity found	-
22	1	+	<2	1195	1194[a]	conserved hypothetical protein (*Vibrio harveyi* 1DA3)	2.00E-152
22	2	+	1195	2076	882	No significant hits	-
22	3	+	2081	2686	606	No significant hits	-
23	1	+	460	828	369	No significant hits	-
24	1	+	178	774	597	hypothetical protein VEA_002669 (*Vibrio* sp. Ex25)	3.00E-101
25	1	+	382	2481	2100	hypothetical protein VCG_002159 (*Vibrio cholerae* 12129(1)), P-loop NTPase superfamily domain	0
25	2	-	2634	3272	639	No significant hits	-
25	3	-	3269	4069	801	No significant hits	-
25	4	-	4296	4472	177	No significant hits	-

25	5	-	5417	6202	786	resolvase domain-containing protein (*Shewanella halifaxensis* HAW-EB4), SR_ResInv domain, Ser_Recombinase superfamily domain	6.00E-48
25	6	-	6577	6867	291	integrase family protein (*Shewanella baltica* OS223), DNA_BRE_C superfamily domain	6.00E-17
25	7	-	6934	7452	519	integrase family protein (*Shewanella baltica* OS223)	3.00E-28
26	1	+	<1	51	51[a]	No significant hits	-
26	2	-	250	1164	915	transcriptional regulator of LysR family protein (*Psychromonas ingrhamii* 37)	1.00E-11
26	3	+	1307	2266	960	hypothetical protein M23134_05654 (*Microscilla marina* ATCC 23134), Esterase_lipase superfamily domain, Aes domain	1.00E-09
26	4	+	2682	3701	1020	beta-lactamase domain-containing protein (*Psychromonas ingrahamii* 37), Lactamase_B superfamily	5.00E-88
26	5	+	3685	5190	1506	hypothetical protein Ping_3113 (*Psychromonas ingrhamii* 37)	3.00E-86
26	6	+	5190	7691	2502	ATP-dependent RNA helicase, DEAD box family protein (*Psychromonas ingrhamiii* 37), DEXDc superfamily domain (2x), COG4581 domain	0
27	1	+	<3	164	162[a]	sigma 70 anti-sigma factor (*Vibrio* sp. AND4)	3.00E-07
27	2	-	205	4953	4749	OmpA family protein (*Vibrio* sp. Ex25), OmpA_C like domain, OmpA_C-like superfamily domain	7.00E-173
27	3	-	4953	6032	1080	putative lipoprotein (*Vibrio fischeri* ES114), tolB domain	4.00E-54
27	4	-	6063	>7070	1035[a]	hypothetical protein 1103602000597_AND4_09062 (*Vibrio* sp. AND4)	0
28	1	+	125	1345	1221	hypothetical protein SKA34_01662 (*Photobacterium* sp. SKA34)	4.00E-151
28	2	+	1338	3602	2265	hypothetical protein SKA34_01667 (*Photobacterium* sp. SKA34)	0
28	3	+	3615	6398	2784	hypothetical protein SKA34_01672 (*Photobacterium* sp. SKA34)	0
28	4	+	6385	6933	549	hypothetical protein SKA34_01677 (*Photobacterium* sp. SKA34)	1.00E-53

28	5	+	7014	7472	459	hypothetical protein Shewana3_2918 (*Shewanella* sp. ANA-3)	9.00E-66
28	6	+	7469	>7984	516[a]	DNA helicase/exodeoxyribonuclease V, α subunit (*Shewanella* sp. ANA-3)	8.00E-66
29	1	-	185	1246	1062	DNA polymerase IV (*Photobacterium* sp. SKA34), Pol_IV_kappa domain, Poly_Y superfamily domain, PRK02406 domain	3.00E-154
29	2	+	1687	2034	348	No significant similarity found	-
29	3	-	2046	3158	1113	hypothetical protein V12B01_04848 (*Vibrio splendidus* 12B01)	2.00E-153
29	4	+	3384	4613	1230	hypothetical protein VOA_001942 (*Vibrio* sp. RC586)	4.00E-159
29	5	+	4902	6419	1518	Type I restriction enzyme M protein (*Vibrio* splendidus 12B01), HsdM_N superfamily domain, N6_Mtase domain	0
29	6	+	6409	7644	1236	Type I site-specific deooxyribonuclease (*Vibrio splendidus* 12B01), Methylase_S superfamily domain (2x), HsdS domain	1.00E-57
29	7	+	7644	8519	876	hypothetical protein V12B01)04686 (*Vibrio splendidus* 12B01), GIY-YIG superfamily domain	9.00E-148
29	8	+	8544	9881	1338	KAP P-loop domain-containing protein (*Shewanella putrefaciens* CN-32), KAP_NTPase superfamily	5.00E-167
29	9	+	9895	13161	3267	HsdR family type I site-specific deoxyribonuclease (*Shewanella putrefaciens* CN-32), DEXDc domain, DEXDC superfamily domain, HSDR_N superfamily domain, hsdR domain	0
29	10	+	13161	13937	777	putative predicted metal-dependent hydrolase (*Vibrio splendidus* 12B01), DUF45 superfamily domain	4.00E-136
29	11	+	14331	16547	2217	hypothetical protein APECO1_4465 (*Escherichia coli* APEC O1)	7.00E-137
29	12	-	17027	>18052	1026[a]	DNA helicase/exodeoxyribonuclease V, alpha subunit (*Shewanella* sp. ANA-3)	2.00E-126
31	1	-	51	578	528	conserved hypothetical protein (*Vibrio harveyi* HY01)	6.00E-70
31	2	-	575	1429	855	hypothetical protien (*Photobacterium profundum* SS9)	4.00E-51

31	3	-	1426	>1512	87[a]	No significant similarity found	-
32	1	+	<2	496	495[a]	A/G-specific DNA glycosylaase (*Vibrio splendidus* 12B01)	3.00E-58
32	2	+	506	1195	690	hypothetical protein VIBHAR_05048 (*Vibrio harveyi* ATCC BAA-116)	2.00E-108
32	3	+	1186	1950	765	hypothetical protein 1103602000597_AND4_09652 (*Vibrio* sp. AND4)	4.00E-55
32	4	+	1953	2150	198	gp12 protein (Vibrio phage VP58.5)	2.00E-08
32	5	-	2869	3279	411	hypothetical protein (*Pelodictyon phaeoclathratiforme* BU-1)	4.00E-32
32	6	-	3258	3689	432	hypothetical protein Ppha_2523 (*Pelodictyon phaeoclathratiforme* BU-1)	5.00E-45
32	7	+	3829	4167	339	hypothetical protein VSAK1_13831 (*Vibrio shilonii* AK1)	2.00E-15
32	8	+	4448	5521	1074	phage integrase family protein (*Vibrio shilonii* AK1), DNA_BRE_C superfamily domain, XerD domain	0
32	9	-	5704	6333	630	hypothetical membrane protein (*Desulfovibrio magneticus* RS-1)	1.00E-04
32	10	-	6380	6943	564	No significant hits	-
32	11	-	7005	7832	828	hypothetical protein VSAK1_13631 (*Vibrio shilonii* AK1)	4.00E-70
32	12	+	8003	8284	282	hypothetical protein VSAK1_13636 (*Vibrio shilonni* AK1), PyocinActivator superfamily domain	8.00E-17
32	13	+	8341	8880	540	phage regulatory protein like CII (*Vibrio furnissii* CIP 102972), Phage_CP76 superfamily domain	7.00E-46
32	14	+	8889	9206	318	hypothetical protein VSAK1_13666 (*Vibrio shilonii* AK1)	8.00E-20
32	15	+	9306	9626	321	No significant hits	-
32	16	+	9623	10024	402	hypothetical protein V12B01_20922 (*Vibrio splendidus* 12B01)	3.00E-41
32	17	+	10021	10614	594	ATP-dependent 26S proteasome regulatory subunit (*Vibrio shilonii* AK1), DnaQ_like_exo superfamily domain	3.00E-77

32	18	+	10611	11063	453	hypothetical protein V12B01_20927 (*Vibrio splendidus* 12B01), ASCH superfamily domain	3.00E-47
32	19	+	11060	11287	228	hypothetical protein VFA_0002222 (*Vibrio furnissii* CIP 102972)	4.00E-24
32	20	+	11284	11508	225	No significant hits	-
32	21	+	11505	14201	2697	phage replication protein (*Vibrio furnissii* CIP 102972)	6.00E-146
32	22	+	14295	14957	663	phage DNA methylase (*Vibrio shilonii* AK1)	6.00E-82
32	23	+	14971	15348	378	No significant hits	-
32	24	-	15349	15600	252	phage transcriptional activator, Ogr/delta (*Vibrio* sp. AND4), Ogr_Delta superfamily domain	2.00E-26
32	25	-	15667	16701	1035	phage portal protein (*Aliivibrio salmonicida* LFI1238), Phage_portal superfamily domain	5.00E-141
32	26	-	16698	18470	1773	terminase (*Vibrio cholerae* O395), Terminase_5 superfamily domain, COG4374 superfamily domain, Terminase_6 domain	0
32	27	+	18734	19519	786	phage capsid scaffolding protein (*Aliivibrio salmonicida* LFI1238)	9.00E-61
32	28	+	19563	20576	1014	phage major capsid protein, P2 family (*Vibrio cholerae* NCTC 8457), Phage_cap_P2 superfamily	8.00E-176
32	29	+	20615	21328	714	phage terminase, endonuclease subunit (*Aliivibrio salmonicida* LFI1238), Phage_term_smal superfamily	1.00E-92
32	30	+	21439	21858	420	prophage PSPPH06, putative head completion/stabilization protein (*Vibrio splendidus* 12B01), Phage_GPL superfamily domain	3.00E-49
32	31	+	21855	22346	492	putative phage gene (*Vibrio splendidus* 12B01), P2_Phage_GpR superfamily domain	4.00E-77
32	32	+	22330	22986	657	prophage PSPPH06, virion morphogenesis protein (*Vibrio splendidus* 12B01)	1.00E-98

32	33	+	22990	24120	1131	prophage PSPPH06, putative tail sheath protein (*Vibrio splendidus* 12B01), DUF2586	9.00E-180
32	34	+	24124	24579	456	prophage PSPPH06, putative tail tube protein (*Vibrio splendidus* 12B01), DUF2597 superfamily domain	1.00E-75
32	35	+	24592	24810	219	DnaK suppressor protein (*Vibrio harveyi* ATCC BAA-1116), zf-dskA_traR superfamily domain	1.00E-12
32	36	+	24807	25232	426	peptidase M15A (*Vibrio parahaemolyticus* AN-5034) additional peptidase M15A hits from other species with lower e-values	4.00E-05
32	37	+	25235	25465	231	hypothetical protein VIBHAR_05036 (*Vibrio harveyi* ATCC BAA-1116)	2.00E-27
32	38	+	25468	25704	237	hypothetical protein VIBHAR_05037 (*Vibrio harveyi* ATCC BAA-1116)	1.00E-28
32	39	+	25701	25973	273	hypothetical protein VIBHAR_05038 (*Vibrio harveyi* ATCC BAA-1116), DUF2765 superfamily domain	2.00E-18
32	40	+	26015	26158	144	hypothetical protein VHA_000299 (*Grimontia hollisae* CIP 101886)	7.00E-11
32	41	+	26167	28056	1890	prophage PSPPH06, tail tape measure protein, TP901 family (*Vibrio splendidus* 12B01), PhageMin_Tail domain	0
32	42	+	28056	28385	330	putative phage gene (*Vibrio splendidus* 12B01), DUF2590 superfamily domain	3.00E-47
32	43	+	28378	29565	1188	putative bacteriophage protein (*Vibrio* sp. AND4), Baseplate_J superfamily domain	1.00E-142
32	44	+	29552	30154	603	putative bacteriophage protein (*Vibrio splendidus* 12B01)	5.00E-88
32	45	+	30167	32944	2778	phage-related tail fiber protein (Vibrio *splendidus* 12B01), DUF3751 superfamily domain only covers the beginning of the predicted gene	3.00E-136
33	1	+	142	1650	1509	putative atp binding protein (*Burkholderia thailandensis* MSMB43), ABC_ATPase domain, P-loop NTPase superfamily domain	7.00E-84
34	1	+	<1	171	171[a]	ABC-type Fe3+ transport system periplasmic component (*Vibrio harveyi* ATCC BAA-1116)	3.00E-16

34	2	+	433	717	285	No significant similarity found	-
34	3	-	824	931	108	No significant similarity found	-
34	4	+	1170	1382	213	No significant similarity found	-
34	5	+	1453	1556	105	hypothetical protein VIBHAR_04879 (*Vibrio harveyi* ATCC BAA-1116)	8.00E-08
35	1	-	24	1334	1311	conserved hypothetical protein (*Vibrio alginolyticus* 40B)	0
35	2	-	1331	1948	618	putative Tfp pilus assembly protein PilW (*Vibrio splendidus* LGP32) better hits to hypothetical proteins	5.00E-44
35	3	-	1929	>2423	495[a]	pili retraction protein PilT (*Vibrio alginolyticus* 12G01)	3.00E-55
36	1	+	54	950	897	No significant similarity found	-
37	1	+	114	521	408	No significant similarity found	-
37	2	+	659	1063	405	conserved hypothetical protein (*Vibrio harveyi* HY01)	1.00E-49
38	1	+	88	408	321	hypothetical protein (*Vibrio* sp. DAT722), DUF2834 superfamily domain	8.00E-43
38	2	-	535	825	291	RelE protein (*Vibrio coralliilyticus* ATCC BAA-450), Plasmid_stabil superfamily domain	3.00E-47
38	3	-	815	1063	249	stability protein StbD (*Vibrio cholerae 12129(1)), PhdYeFM superfamily*	3.00E-37
39	1	-	9	524	516	hypothetical protein MED92_14048 (*Oceanospirillum* sp. MED92)	2.00E-19
40	1	+	50	598	549	hypothetical protein VCB_002440 (*Vibrio cholerae* TMA 21)	4.00E-14
41	1	-	19	321	303	RelE1 (*Vibrio vulnificus*), Plasmid_stabil superfamily domain	3.00E-50
41	2	-	309	566	258	RelB1 (*Vibrio vulnificus*), PhdYeFM superfamily	4.00E-41
42	1	-	11	298	288	plasmid stabilization system protein (*Vibrio cholerae* 623-39), Plasmid_stabil superfamily domain	3.00E-48
42	2	-	295	576	282	antitoxin of toxin-antitoxin stability system (*Vibrio vulnificus* CMCP6), PhdYEFM	4.00E-45

43	1	+	56	544	489	No significant hits	-
44	1	-	31	777	747	conserved hypothetical protein (*Campylobacter rectus* RM3267)	1.00E-20
45	1	+	65	460	396	No significant hits	-
45	2	+	511	1116	606	transposase (*Vibrio cholerae* non-O1/non-O139), Transposase_11 domain	1.00E-110
45	3	+	1291	1431	141	transposase (*Vibrio* sp. RC586)	5.00E-16
45	4	+	1535	2029	495	No significant hits	-
46	1	-	22	786	765	hypothetical protein Shewana3_3449 (*Shewanella* sp. ANA-3)	5.00E-110
47	1	+	211	540	330	putative Rhs-family protein (*Vibrio alginolyticus* 12G01)	2.00E-29
47	2	+	542	>622	81[a]	No significant similarity found	-
48	1	+	<3	1163	1161[a]	maltoporin (*Vibrio harveyi* HY01)	6.00E-51
48	2	+	1321	2151	831	maltose operon periplasmic protein MalM (*Vibrio coralliilyticus* ATCC BAA-450), MalM superfamily	6.00E-114
48	3	+	2313	2831	519	glycosidase (*Vibrio harveyi* HY01)	7.00E-41
49	1	-	35	481	447	histone acetyltransferase HPA2 (*Vibrio vulnificus* CMCP6)	8.00E-69
49	2	+	525	629	105	hypothetical protein A79_0873 (*Vibrio parahaemolyticus* AQ3810)	1.00E-04
49	3	-	606	1088	483	hypothetical protein VMC_00030 (*Vibrio alginolyticus* 40B)	2.00E-85
49	4	-	1251	2315	1065	No significant hits	-
50	1	+	44	532	489	N-carbamoyl-L-amino acid amidohydrolase (*Alteromonadales bacterium* TW-7) hits to penicillin binding protein with higher e-values	2.00E-59
50	2	+	799	>978	180[a]	hypothetical protein (*Vibrio vulnificus*)	3.00E-14
52	1	+	54	638	585	hypothetical protein VV1821 (*Vibrio vulnificus* YJ016)	7.00E-84

53	1	+	441	1472	1032	hypothetical protein, ADP-ribose binding protein (*Aliivibrio salmonicida* LFI1238), Macro_Poa1p_like domain, Macro superfamily domain numerous hits to Appr-1-p processing protein with higher e-values	5.00E-121
53	2	-	1503	2084	582	hypothetical protein ETAE_3071 (*Edwardsiella tarda* EIB202), PTS_2-RNA superfamily domain	9.00E-49
54	1	+	<2	64	63[a]	No significant similarity found	-
54	2	-	123	1859	1737	bifunctional UDP-sugar hydrolase/5'-nucleotidase periplasmic precursor (*Vibrio harveyi* ATCC BAA-1116), MPP_UshA_N domain, MPP_superfamily superfamily domain, 5_nucleotid_C superfamily domain, ushA domain	0
55	1	-	204	395	192	No significant similarity found	-
55	2	+	732	1082	351	4-amino-4-deoxy-L-arabinose transferase (*Vibrio* sp. Ex25), PKR09272 superfamily domain	1.00E-57
55	3	+	1593	2336	744	conserved hypothetical protein (*Vibrio parahaemolyticus* AQ3810)	4.00E-137
55	4	+	2358	3020	663	conserved hypothetical protein (*Vibrio parahaemolyticus* AQ3810)	3.00E-124
55	5	+	3126	3314	189	hypothetical protein VIC_004279 (*Vibrio coralliilyticus* ATCC BAA-450), DUF1289 domain, DUF1289 superfamily domain	8.00E-20
55	6	-	3345	>3452	108[a]	No significant similarity found	-
56	1	-	2	202	201	Vco29 (*Vibrio parahaemolyticus* AQ3810), numerous hits to glyoxalase family protein	8.00E-08
56	2	-	49	336	288	glyoxalase family protein (*Vibrio* sp. Ex25), Glyoxalase superfamily domain	2.00E-37
56	3	-	520	867	348	hypothetical protein V12G01_21068 (*Vibrio alginolyticus* 12G01)	4.00E-46
57	1	-	37	489	453	hypothetical protien A5A_A0397 (*Vibrio cholerae* MZO-2)	9.00E-63
58	1	+	278	361	84	No significant similarity found	-
59	1	+	254	1024	771	No significant hits	-

59	2	+	1026	1565	540	No significant hits	-
60	1	+	<2	835	834[a]	No significant hits	-
61	1	+	<2	97	96[a]	No significant similarity found	-
61	2	-	106	957	852	hypothetical protein VCA_000159 (*Vibrio cholerae* bv. Albensis VL426), TIR-like superfamily domain	1.00E-161
63	1	+	<3	2129	2127[a]	gametolysin peptidase M11 family (*Vibrio parahaemolyticus* AQ3810)	0
63	2	-	2234	4483	2250	conserved hypothetical protein (*Vibrio alginolyticus* 40B)	0
64	1	-	27	650	624	putative orphan protein (*Shewanella denitrificans* OS217)	3.00E-54
65	1	+	72	704	633	hypothetical protein P3TCK_11078 (*Photobacterium profundum* 3TCK)	4.00E-105
66	1	+	<3	1418	1416[a]	54K polar flagellar sheath protein A (*Vibrio parahaemolyticus* AQ3810)	0
67	1	+	<3	680	678[a]	No significant hits	-
68	1	+	109	417	309	transposase and inactivated derivative (*Vibrio cholerae* MZO-2), HTH_Hin_like superfamily domain	1.00E-51
68	2	+	414	1265	852	transposase InsF for insertion sequence IS3A/B/C/D/E/fA (*Vibrio harveyi* HY01)	4.00E-99
68	3	+	1428	2258	831	Phosphoglycerate dehydrogenase (*Vibrioanles bacterium* SWAT-3)	5.00E-102
69	1	+	45	863	819	hypothetical protein VMC_04160 (*Vibrio alginolyticus* 40B), PKc_like superfamily domain several hits to homoserine kinase and aminoglycoside phosphotransferase with higher e-values	3.00E-158
71	1	+	130	372	243	ParD protein (antitoxin to ParE) (*Vibrio* sp. RC586), PhdYeFM superfamily domain	1.00E-31
71	2	+	380	679	300	plasmid stabilization system protein protein ParE (*Vibrio vulnificus* YJ016), Plasmid_stabil superfamily	9.00E-46

72	1	+	222	1082	861	conserved hypothetical protein (*Vibrio* sp. Ex25)	1.00E-159
73	1	+	<1	147	147[a]	No significant similarity found	-
73	2	+	555	4073	3519	putative superfamily I DNA helicase (*Vibrio furnissii* CIP 102972)	0
74	1	+	97	687	591	hypothetical protein VVA0317 (*Vibrio vulnificus* YJ016)	5.00E-57
75	1	-	120	>1919	1800[a]	outer membrane vitamin B12 receptor BtuB (*Vibrio* sp. Ex25)	0
76	1	-	<1	426	426[a]	OtnA protein (*Vibrio parahaemolyticus* AN-5034)	2.00E-58
76	2	-	493	606	114	hypothetical protein VparAN_08625 (*Vibrio parahaemolyticus* AN-5034)	1.00E-12
76	3	-	609	>941	333	hypothetical protein VparAN_08625 (*Vibrio parahaemolyticus* AN-5034)	5.00E-44
77	1	-	<1	>1830	1830[a]	TonB-dependent receptor (*Aeromonas salmonicida* subsp. Salmoncida A449)	4.00E-152
78	1	-	<2	>1246	1245[a]	large exoproteins involved in heme utilization or adhesion (*Vibrio* sp. Ex25)	5.00E-88
79	1	+	<1	237	237[a]	putative epimerase/dehydratase (*Vibrio parahaemolyticus* K5030)	2.00E-15
79	2	+	394	>693	300[a]	nucleotide sugar dehydrogenase (*Vibrio harveyi* ATCC BAA-1116)	1.00E-37
80	1	-	57	581	525	hypothetical protein VV1_2471 (*Vibrio vulnificus* CMCP6) several hits to acetyltransferase with higher e-value	5.00E-97
81	1	-	6	503	498	No significant hits	-
82	1	+	<3	>572	570[a]	putative Rhs-family protein (*Vibrio alginolyticus* 12G01)	8.00E-69
84	1	+	<1	885	885	hemolysin (*Vibrio parahaemolyticus* AQ3810)	3.00E-108
85	1	-	48	569	522	acetyltransferase (*Vibrio vulnificus* CMCP6), GNAT superfamily domain	4.00E-65
86	1	+	<3	644	642[a]	SAM-dependent methyltransferase (*Vibrio orientalis* CIP 102891)	2.00E-79
87	1	+	<3	>545	543[a]	DNA helicase/exodeoxyribonuclease V, alpha subunit (*Shewanella* sp. ANA-3)	4.00E-73
88	1	+	159	500	342	No significant hits	-

88	2	+	671	1108	438	No significant hits	-
89	1	+	292	735	444	No significant hits	-
90	1	-	<3	>587	585[a]	oxaloacetate decarboxylase beta chain (*Vibrio parahaemolyticus* AQ3810)	4.00E-87
91	1	+	<3	479	477[a]	msha pilin protein MshC (*Vibrio harveyi* HY01)	5.00E-37
91	2	+	476	>559	84[a]	No significant similarity found	-
92	1	-	95	385	291	TM2 domain containing protein (*Pectobacterium wasabiae* WPP163), TM2 domain, XynA domain, TM2 superfamily domain few hits to tfp pilus assembly protein	1.00E-36
93	1	+	69	944	876	No significant hits	-
94	1	+	44	502	459	hypothetical protein A79_0874 (*Vibrio parahaemolyticus* AQ3810)	2.00E-76
95	1	-	5	>1054	1050[a]	integrase (Vibrio mimicus VM223)	2.00E-174
97	1	+	<3	>701	699[a]	nitric oxide reductase regulator (*Vibrio harveyi* ATCC BAA-1116) numerous hits to transcriptional regulator	2.00E-76
98	1	-	<2	157	156[a]	MSHA pilin protein MshA (*Vibrio* sp. Ex25)	5.00E-11
98	2	-	227	>634	408[a]	MSHA pilin protein MshA (*Photobacterium* sp. SKA34)	2.00E-36
99	1	-	32	469	438	hypothetical protein Shew_0612 (*Shewanella loihica* PV-4)	2.00E-59
100	1	-	89	493	405	type IV pilin PilA (*Vibrio* sp. RC341)	4.00E-41
101	1	-	22	420	399	acetyltransferase, gnat family (*Vibrio cholerae* AM-19226)	2.00E-32
102	1	+	<2	>1042	1041[a]	diaminobutyrate-pyruvate transaminase/L-2,4-diaminobutyrate decarboxylase (*Vibrio* sp. Ex25), AAT_I superfamily	0
106	1	-	<3	>518	516[a]	No significant hits	-
109	1	+	<3	68	66[a]	No significant similarity found	-

109	2	+	94	531	438	No significant hits, GIY-YIG superfamily domain	-
114	1	-	<3	167	165[a]	glutamate racemase (*Vibrio alginolyticus* 40B)	1.00E-14
114	2	-	199	>615	417[a]	ATPase of the PP-loop superfamily (*Vibrio* sp. Ex25)	6.00E-56
117	1	-	52	354	303	hypothetical protein VPA1343 (*Vibrio parahaemolyticus* RIMD 2210633)	7.00E-34
125	1	+	157	528	372	hypothetical protein DP1716 (*Desulfotales psychrophila* LSv54)	6.00E-49
128	1	+	<1	384	384[a]	excinuclease ABC subunit C (*Vibrio* sp. AND4)	6.00E-39
128	2	+	433	915	483	putative carboxynorspermidine dehydrogenase (*Vibrio parahaemolyticus* RIMD2210633), LYS9 domain also hits to saccharopine dehydrogenase	7.00E-89
135	1	-	<3	461	459[a]	ADP-L-glycero-D-mannoheptose-6-epimerase (*Vibrio parahaemolyticus* AN-5034)	2.00E-30
146	1	+	<1	363	363[a]	No significant similarity found	-
146	2	+	290	576	288	putative epimerase/dehydratase (*Vibrio parahaemolyticus* 2210633), Polysacc_synt_2 domain	4.00E-42
305	1	-	<1	714	714[a]	hypothetical protein GalfDRAFT_1274 (*Gallionella ferruginea* ES-2)	8.00E-23
306	1	+	<1	96	96[a]	No significant hits	-
306	2	-	170	1159	990	transcriptional regulator, LysR family (*Vibrio parahaemolyticus* AQ3810)	2.00E-160
306	3	+	1278	2306	1029	soluble lytic murein transglycosylase (*Vibrio parahaemolyticus* AQ3810), DUF1254 superfamily domain, DUF1214 superfamily domain, COG5361 domain	0
306	4	+	2338	3390	1053	conserved hypothetical protein (*Vibrio parahaemolyticus* AQ3810) hits to soluble lytic murein transglycosylase with higher e-value, DUF1254 superfamily domain, DUF1214 superfamily domain	0
306	5	+	3513	4394	882	transcriptional regulator (*Vibrio parahaemolyticus* AQ3810) hits to LysR family transcriptional regulator with higher e-value	7.00E-170

306	6	-	4383	4484	102	hypothetical protein A79_5290 (*Vibrio parahaemolyticus* AQ3810)	5.00E-10
308	1	+	<2	652	651[a]	hypothetical protein PputW619_3918 (*Pseudomonas putida* W619)	3.00E-14

[a] ORF is not annotate to its full length because it is located at the end of a contig

[b] annotation column includes the annotation for each ORF, the organism the best BLAST hit came from, as well as any predicted domains within the ORF