VIBRIO VULNIFICUS VIRULENCE AND SURVIVAL MECHANISMS REVEALED THROUGH COMPARATIVE MICROBIAL GENOMIC ANALYSIS

by

Shatavia Sharday Morrison

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2013

Approved by:

_____
Dr. Cynthia J. Gibas

_____
Dr. Jennifer Weller

_____
Dr. Cory Brouwer

_____
Dr. ZhengChang Su

_____
Dr. James Oliver

_____
Dr. Matthew Parrow

ABSTRACT

SHATAVIA SHARDAY MORRISON. Vibrio vulnificus virulence and survival mechanisms revealed through comparative microbial genomic analysis.
(Under the direction of DR.CYNTHIA J. GIBAS)

A sound genome assembly and robust annotations are essential to the differential analysis of bacterial genomes. Using a case study data set of newly sequenced *Vibrio vulnificus* genomes, both the biology of these bacteria, and the bioinformatics processes that support identification of the similarities and differences found within the different isolates of *V. vulnificus*, were examined. The two main themes of this research are 1) identification of the virulence and survival mechanisms of clinical and environmental biotypes of *Vibrio vulnificus* and 2) quantification of the impact of different analysis choices on the overall biological conclusions of the study. Whole genome sequencing, in conjunction with comparative genomics, are current techniques used to capture the genetic and functional repertoire of organisms. It is important to consider and track analytic provenance in bacterial genomics because the impact of making alternate workflow choices can involve changing the biological interpretation of hundreds of genes, even in relatively simple bacterial genomes. Chapter 1 describes the bioinformatics analyses used to determine the draft genome sequences of three environmental genotype *Vibrio vulnificus* reference genomes and to identify genotype-specific genomic regions. Chapter 1 also highlights the functional systems including the virulence and survival genes that differentiate between clinical and environmental *Vibrio vulnificus* genotypes. Chapter 2 explores the direct impact of the parameter

and methods selected during the assembly and annotation stage of a genome project. Despite decades of advances in *ab initio* gene prediction, method and parameter choices still strongly influence the identification of genes, and therefore the biologically significant results in a comparative genomics analysis. Using a benchmarking approach based on simulation studies with a related genome, it is possible to identify an optimal assembly-to-annotation pipeline for the collection of *V. vulnificus* strains. A software framework for comparing the outcomes of different assembly-to-annotation workflows was constructed in the Taverna workflow management system and used to carry out the bioinformatics experiments described in Chapter 3. Chapter 3 expands on the analysis performed in Chapter 1 by performing an extensive comparative genomics analysis of newly sequenced *Vibrio vulnificus* genomes, each ones represents the different biological classifications found within this species. The analysis of these genomes reveals genes that are specific to each of the biotypes. Comparative analysis of representative strains from each of the established *Vibrio vulnificus* biotypes is used to identify differentiating genes, which may relate to the apparent host-specificity of the different biotypes.

ACKNOWLEDGMENTS

The Graduate Assistance Support Plan (GASP), The College of Computing and Informatics Graduate Assistance in Areas of National Need (GAANN) Fellowship Program, and the Lucille P. and Edward C. Giles Graduate School Dissertation-Year Fellowship supported the research described in this dissertation. I would like to acknowledge the many contributions the members of the laboratory of Dr. James Oliver made to Chapter 1 and 3. I would also like to thank Dr. Craig Baker-Austin at the Centre for Environment, Fisheries, and Aquaculture Science for supplying some of *Vibrio vulnificus* sequencing data used throughout this dissertation work. Dr. Raad Gharaibeh, Joshua Newton, Aurora Cain, and Myung Sik Jeon were members of the laboratory of Dr. Cynthia Gibas during my graduate career. They were invaluable colleagues that helped with the development of various methods and modules to assist in the analysis of the *V. vulnificus* sequencing data.

I would also like to acknowledge Dr. Jennifer Weller for her guidance and support during my graduate career and her participation in my doctoral committee. I would like to thank my doctoral Dr. Cory Brouwer and Dr. ZhengChang Su, Department of Bioinformatics and Genomics, Dr. Matthew Parrow, Department of Biology for their advice, and Dr. James Oliver, Department of Biology, for his research expertise with *Vibrio vulnificus* and finally, to Dr. Cynthia Gibas for her guidance and support throughout my graduate career and the inspiration to pursue my research goals.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: PYROSEQUENCING-BASED COMPARATIVE GENOME ANALYSIS OF VIBRIO VULNIFICUS ENVIRONMENTAL ISOLATES [2]

1.1 Introduction

The study of microbiology presents many opportunities and challenges around genome sequencing. Bacterial genome sequences evolve rapidly, and the gene content even closely related bacterial strains can change significantly due to processes such as horizontal gene transfer. Sequencing and comparative analysis of bacterial genomes make it possible to identify the genes and associated functional capabilities that make bacteria effective as pathogens, the role in which they are most of concern to human health. Using comparative genomics analysis techniques on collections of bacterial genome sequences can begin to elucidate the differences in function and gene content among these bacterial species. Sequencing bacterial genomes is the starting point for studies of pathogenicity, niche specialization, and evolutionary relationships among species. Given the large amount of sequencing data that can be produced using technologies, it is not uncommon to perform comparative analyses of dozens of bacterial strains simultaneously, where even 10 years ago comparison of even two strains would have been considered a wealth of information. In order to compare the multiple bacterial genomes that are now common, it is necessary to design and implement a bioinformatics infrastructure to store and track the artifacts of the analysis process.

While genome browsers might support alignment of two bacterial genomes, modern projects require a database infrastructure that supports aggregation of genomes into relevant classes (for example, strains that have been found in human infections vs. those that have only been observed in marine environments) and comparison of functional content across classes.

Described in this chapter is a large-scale approach in which multiple genomes are compared in order to discover the similarities and differences between the genomes and to stud the biology of individual strains. There are several types of analyses that can be performed with comparative genomics [1] such as the identifying differentiating genes and genomic rearrangements. In this work, the primary analysis consists of identification and comparison of protein coding sequence content, including gene content, protein content, orthologs, and paralogs. This analysis is the basis for identification of commonalities and differences between various biological classifications of *Vibrio vulnificus*, and provides a starting point for molecular investigation of previously uncharacterized differences in function.

Of all seafood-associated human pathogens, none are as critical as those of the genus *Vibrio*, and of all the food-borne pathogens, only infections caused by this genus increased (by 78%) between 1996 and 2006[3]. In the United States, a single member of this genus, *Vibrio vulnificus*, causes 95% of all deaths resulting from seafood consumption[3]. In addition to the high fatality rate there is a considerable level of productivity lost as a result of the symptoms, including nausea, hypotensive septic shock, and the formation of secondary lesions on the extremities. For a human pathogen of this importance the molecular data are surprisingly sparse: at the time of

this writing three clinical strains of *V. vulnificus* had been fully sequenced but only short read data existed for any environmental strains. Without a completely sequenced environmental strain of *V. vulnificus* it is not possible to identify the genotypic differences that lead to pathogenicity. In this chapter, we describe the sequencing, assembly, and comparative analysis of three environmental strains of *V. vulnificus*. In addition to identifying the virulence and survival mechanisms that contribute to *V. vulnificus*'s pathogenicity at a genome wide scale, we perform an assessment of the sequence assembly process to ensure that genetic content of these strains has been captured in its entirety.

Several approaches have been used to identify genotypic factors that distinguish between the virulent and avirulent isolates of *Vibrio vulnificus*. Aznar *et al*. [4] identified two groups (termed A and B) of *V.vulnificus* strains based on 16SrDNA gene polymorphism, and Nilsson *et al*. [5] showed that these two groups were associated with clinical (B, or C-type or C-genotype) or environmental (A, E-type or E-genotype) isolation. Despite employing a variety of population genetics methods, Gutacker *et al*. [6]found no association between their grouping and environmental or clinical origin. This contradiction is explained by poor resolution in the traditional molecular biology techniques used to identify this pathogen. Until recently, only local genetic differences between two genotypes have been probed, and only the genomes of clinical isolates have so far been completely sequenced [7-9]. In 2010, a comparative genomic analysis using short read data was performed on four *V. vulnificus* strains, including three environmental strains: 99-520 DP-B8, 99-738 DP-B5, and ATCC 33149[10]. However, that study employed the ABI SoLID next

generation sequencing platform, which produces very short sequence fragments. Such reads cannot be assembled *ab initio,* but must be mapped to the clinical reference genomes. This approach leaves the possibility that regions of the environmental genome, for which there are no reference in the clinical genome sequence, remain undetected. By sequencing environmental strain genomes we have developed a far more complete understanding of the differences between clinical and environmental strains than has previously been possible. This study also provided us with a better understanding of *V. vulnificus* as an agent of disease and helped to identify the molecular components that may be associated with its virulence and survival mechanisms.

1.2 Material and Methods

1.2.1 Strains, Growth Conditions, and DNA Isolation

V. vulnificus JY1305 (E-genotype and environmental isolate) was grown overnight in Bacto™ Heart Infusion (HI) broth (BD, New Jersey) at 30°C with vigorous shaking. Cells were pelleted by centrifugation and supernatants discarded. The cells were washed three times with phosphate buffered saline (PBS) before being resuspended to a final approximate concentration of 5×10^8 cell/ml. The MagMax™ Total Nucleic Acid Isolation Kit (Ambion) and All Prep DNA/RNA/Protein Mini Kit (Qiagen) were used for DNA extraction. The quality and quantity of DNA was evaluated spectrophotometrically with the NanoDrop ND1000 (Thermo Scientific, Wilmington, DE). A concentration of 50 ng/µL was used for next generation sequencing.

V. vulnificus strains E64MW (E-genotype and wound isolate) and JY1701 (E-genotype and environmental isolate) were grown overnight with shaking in 10 ml of alkaline saline peptone water (ASPW). Cells were pelleted by centrifugation and resuspended in 100 µl of ice-cold PBS. DNA was extracted using DNAzol (Invitrogen) according to manufacturer instructions, followed by incubation with RNase A. Subsequently, samples were purified using phenol/chloroform/isoamyl alcohol extraction protocol. Briefly, 40 µl of 3 M sodium acetate was added to each DNA sample, followed by 440 µl of pheno/chloroform/isoamyl alcohol. Samples were centrifuged (5 min, 13,000 rpm) and ~400 µl of supernatant was removed and mixed with an equal volume of phenol/chloroform/isoamyl alcohol. This solution was centrifuged for 5 min (13,000 rpm) and the supernatant (~200 µl) was subjected to ethanol precipitation. The DNA pellet was re-dissolved in 50 µl 1 ×TE buffer and stored at -80°C. The quality and quantity of DNA was subsequently ascertained spectrophotometrically using a NanoDrop ND 1000 (NanoDrop Technologies, Wilmington, DE).

1.2.2 Genome sequencing and assembly

V. vulnificus JY1305 was sequenced at the Virginia Commonwealth University using Roche/454 Titanium technology [11]. One complete sequencing plate was used for this genome. *V. vulnificus* E64MW and JY1701 were sequenced at the BBSRC Genome Analysis Centre (Norwich, UK) also using the Roche/454 Titanium technology [11]. Quarter plates were used for both. For all three sequencing datasets (JY1305, E64MW, and JY1701) single end reads were generated. De novo assembly with Newbler version 2.3 was initially performed at

the sequencing centers [11]. An additional assembly was performed using the MIRA 3.2.1 de novo assembler[12]. The default parameters for MIRA were used, except for the assembly quality parameter, which was changed from "normal" to "accurate", and trace information was excluded from the assembly.

1.2.3 Genome and gene characterization

Draft annotation of the sequences was performed using a pipeline of published microbial annotation tools, as follows. Feature determination for each strain was performed on the contig set from each sequence assembly. Feature identification methods included Glimmer3.02 (Glimmer) and GeneMark.hmm (GeneMark) [13,14]. Both packages are widely used feature determination applications whose output is recognized and accepted by NCBI, and both are publicly available. Glimmer was used with default parameters. An exception was that the circular chromosomes were treated as linear in the analysis. This setting was used to prevent each contig from being treated as an individual circular chromosome. GeneMark was used with the default parameters. The models used for training were the two *V. vulnificus* reference organisms (CMCP6 and YJ016). Spacer sequence was added to the ends of each contig to mimic start and stop signals. The spacer sequence was 32 nucleotides in length. We used the sequence NNNNNCACACACTTAATTAATTAAGTGTGTGNNNNN, which is used at the J. Craig Venter Institute (JCVI) to merge contigs [http://www.jcvi.org/cms/research/projects/annotation-service/submission-guide/]. Differences in interpretation may arise when it comes to combining results from the various gene identification methods into a unified annotation. Because one of our

main goals in this study was to compare the newly sequenced E-type genomes to the genome of the previously sequenced C-type strains, we maintained a consistent analytical pipeline throughout. For gene identification in each of the newly sequenced strains, one of the following criteria had to be met: (1) A gene will be included in the gene list if it can be predicted by either Glimmer or GeneMark, as long as the amino acid (aa) sequence length is equal to or greater than 150 aa. (2) A gene must be predicted by both Glimmer and GeneMark to be included in the gene list, if its amino acid sequence length less than 150 aa. (3) A gene prediction may also be included in the gene list if it occurs in a cluster of known orthologous genes found in other *Vibrio* spp., regardless of whether it meets the length criterion. A cluster is defined as a group of gene sequences that represent either orthologs or paralogs from a set of reference genomes closely related to the genome being annotated. The first two criteria were derived from Chen *et al*. 2003 [6]and were used as a consistency benchmark for different gene prediction methods across genomes. For the third criterion we defined homology as membership in a set of sequences that formed an unambiguous ortholog cluster with all genomes used in this study when analyzed using OrthoMCL[15].

tRNAScanSE was used to predict the tRNAs in the MIRA contigs for each strain[16]. RNAHMMER was used to predict the rRNAs from the MIRA contigs for each strain[17]. In all cases, default parameter settings were used.

1.2.4 Gene Clustering

OrthoMCL version 2.0 was used to cluster newly predicted genes of the three newly sequenced environmental *V. vulnificus* genomes (JY1305, E64MW, and

JY1701) with genes from other completely characterized *Vibrio* spp.[15]. OrthoMCL uses an all-against-all blastp comparison of sequences as an input step followed by application of a Markov clustering procedure. The e-value cutoff for the BlastP algorithm was 1e-5. Default parameters were used for OrthoMCL except that clusters were formed based on a shared sequence similarity of 70% rather than 50%. The increased stringency resulted in more constrained gene clusters, and reduced inappropriate clustering of partial homologs into ortholog clusters.

1.2.5 Gene Content Comparison

The OrthoMCL clustering output that was generated during the annotation step became the basis for identification of differentiating genes. Identified gene features and OrthoMCL results were stored in a locally developed OLAP data warehouse (GenoSets) that supports queries across aggregate data generated by a variety of genomic annotation and comparison methods[18], as described in Cain *et al*. Annotations for the published C-type genomes were downloaded and parsed from the EMBL-Bank public repositories. Annotations for the novel E-type genomes reported were generated as described in section 1.3.4. Feature boundaries were determined from the annotation output and stored, allowing gene presence-absence queries to be formulated in GenoSets returning gene features that differentiate the three E-types from each other, and from the C-type strains.

In order to provide a standard means of comparison for feature attributes we established relationships between features using two methods. First, we estimated orthologous relationships between genes using OrthoMCL, which uses a Markov Cluster algorithm to group putative homologs based on sequence similarity.

OrthoMCL has been shown to outperform other stand-alone methods for ortholog clustering[15]. For functional analysis, gene features identified in the newly sequenced *V. vulnificus* strains were associated with GO terms using homology determined through OrthoMCL clustering of BLASTP results. For functional comparison purposes, we used a controlled vocabulary to describe genes and other features. The Gene Ontology (GO) provides standardized terms for the description of gene products in terms of biological processes, cellular location, and molecular function [19,20]. If a GO term was associated with any gene within an ortholog cluster, all genes within that cluster were also associated with that GO term.

1.3 Results

1.3.1 Genome Sequencing and Assembly Statistics

188,710,063 bases of DNA sequence were generated for *V. vulnificus* strain JY1305. Given the known sizes and expected variability of *V. vulnificus* genomes, we estimated that this is equivalent to ~33x genome coverage depth of the *V. vulnificus* JY1305 genome, of estimated size 5.7 Mb. We obtained 671,521 reads of average length 281 bp of 454-pyrosequencing data for *V. vulnificus* JY1305. The data were assembled into 159 large contigs and 9,184 unassembled fragments using the MIRA assembler, version 3.0[12]. Table 1.1A has the complete assembly results for the three E-type strains. The coverage of each of these genomes is significantly above the recommended genome coverage (6-10x) for a whole prokaryote genome study established in a recent exhaustive simulation of outcomes of Roche 454 type sequencing in prokaryotes[21]. In Figure 1.1, we show the assembled contigs from each of the newly sequenced E genomes, aligned to the *V. vulnificus* CMCP6

genome[22]. *V. vulnificus* CMCP6 was recently re-annotated and is regarded as the most complete and accurate of the published *V. vulnificus* clinical strains genome[23]. Assembled contigs were deposited in the NCBI whole genome shotgun archive, and are available under project IDS 49015(JY1305), 67135(E64MW), and 67137(JY1701). The GenBank accessions IDs are AFSW00000000 (JY1305), AFSX00000000 (E64MW), and ASFY00000000 (JY1701) in the NCBI Whole Genome Assembly database.

1.3.2 General Properties of the Vibrio E-type Genomes

The genome of *V. vulnificus* JY1305 is composed of 2 circular chromosomes with an estimated total of approximately 5.7 MB of genomic DNA. *V. vulnificus* E64MW is estimated to be nearly identical in size to JY1305. *V. vulnificus* JY1701 slightly smaller at 5.6 Mb. Some Vibrio strains are known to have plasmids, but the *V. vulnificus* JY1305 sequence data contained no evidence of extra chromosomal DNA. PCR was assays were performed to verify this finding and no plasmid DNA (Appendix A) was found in the genomic DNA preps. It is unknown if *V. vulnificus* E64MW and *V.vulnificus* JY701 contain plasmid DNA, but no plasmid sequence with homology to the known *V. vulnificus* YJ016 plasmid sequence was identified, either in the assembled genomic sequence, or among the unassembled reads. Table 1.1B summarizes the general characteristics and predicted gene content of each sequenced draft genomes.

1.3.3 Genome Content Comparison

After annotation of the newly sequence E-genotype *Vibrio vulnificus* genomes described in section 1.3.3, we performed a comparative analysis of the

presence or absence of individual genes. We compared the E-genotype genomes to the group of previously sequenced C-genotype *V. vulnificus* genomes. Figure 1.2 summarizes the gene count differentials for the six *V. vulnificus* strains apart of this work. Genes were clustered together based on the basis of a shared sequence similarity of 70% or greater for the purpose of defining orthology, as described in section 1.3.4. The counts represent differential presence or absence of a gene ortholog in a given genome.

1.3.4 The Conserved Core of *Vibrio vulnificus*

We identified approximately 3664 orthologs common to all of the *V. vulnificus* strains analyzed in this chapter. An in-depth comparison between the two genotypes of *V. vulnificus* revealed 278 genes found only in the C-type strains, and 167 genes found only in the E-genotype strains. We also identified 43 genes common to the three C-genotype blood isolates, CMCP6, YJ016, MO6-24/O, and the E-genotype wound isolate, E64MW. The gene VV2 0404 (*vvhA*), which is commonly used in a core marker set to distinguish *V. vulnificus* from other *Vibrio* spp. in molecular assays, was found, as expected, in all six *V. vulnificus* strains, which gives us confidence in the sequencing and differential analysis. The gene encoding zinc metalloprotease, VV2-0032 (*vvpE*), another commonly-used diagnostic marker, was identified by Gulig *et al*. 2010 as being common to both E-type and C-type strains[10] , and we found this to be true in our analysis, as well. A related gene, VVA0964, the cytolysin secretion protein gene *vvhB* [24], is unique to the *V. vulnificus* genomes and may have potential as a diagnostic marker. Also, we identified Flp pilus genes common to all the *V. vulnificus* genomes. We believe this

is a novel observation, as we have not seen it discussed elsewhere. The E- and C-genotypes of *V. vulnificus* contain a nearly identical operon for the assembly of an Flp pilus, a type IV pilus that mediates adherence, including genes for Flp pilus assembly *CpaB*, *CpaC*, a conserved unknown protein, and *CpaE*. The Tad assembly protein of the Flp pilus, including TadA, TadB, TadC, and TadD, are also highly conserved and identically ordered in C7184 and YJ016. Both E- and C- type strains of *V. vulnificus* contain all the components of the Tad assembly proteins except TadD, while other *Vibrio spp*. do not. These genes may be part of a tad (tight adherence) locus, found in a wide variety of bacteria that is characteristic of horizontal gene transfer. *tad* loci are generally present as part of a mobile genetic element, specifically the "widespread colonization island" [25]. Loci such as these are known to be related to disease, both human and animal, playing a role in colonization and/or pathogenicity. In non-pathogens, *tad* loci are proposed to facilitate environmental niche colonization[26].

1.3.5 Gene and Functionally Different Regions of C- and E- type Isolates

In Table 1.2 and Table 1.3, we summarize key differences between C-type and E –type genomes, listing genes that are shared between the strains of a specific genotype, but excluded from the other genotype. A few of those differentiating genes are significance to human virulence or to survival in the estuarine/oyster environment. As in section 1.3.5, features are described using the Gene Ontology (GO) categories and terms. Functional categories having significant enrichment or depletion between genomes (at the species or genus level) were identified using the Gene Ontologizer[27]. A detailed description of how significance is estimated is

given in Cain et al[28]. Figure 1.3 summarizes differences in GO functional content between the C-genotypes and E-genotypes of *V. vulnificus*. The differential functional analysis shows that GO terms mannitol-1-phosphate 5-dehdrogenase and N-acetylneuraminidase are significantly enriched in the C-types with an adjusted p-value of 2.42 $^{E-04}$ and 1.13 $^{E-05}$, respectively. Specifically, 35% of the genes associated with mannitol-1-phosphate 5-dehydrogenase activity and nearly 100% of the genes with associated with N-acetylneuraminidase function are found to be unique to C-types. Additionally, the GO terms "chondroitin AC lyase activity" and "arylsulfatase activity" are significantly enriched with adjusted p-values of 0.0068 and 0.048, respectively and close to 100% of these genes only found in the C-genotype strain differentials. In contrast, the E-genotypes appear to be strongly enriched in genes associated with the GO functions "urea metabolic process" and "nickel ion binding". Nearly all of the genes that fall under these GO categories are only found in the E-genotypes. Both show up as statistically significant differentials with adjusted p-values of $1.52^{E-09}$and 4.37^{E-07}, respectively. E-genotypes also appear to have several unique genes that fall into GO categories associated with carbohydrate transport and transmembrane transporter activity for a variety of sugars and sugar derivatives. Understanding the overall significance of these genotypic GO functional differences will require further investigation. However, we propose that these differentiating functional categories may be relevant to the SPANC hypothesis, which describes the balance between self-preservation and nutritional competence in bacterial genomes [29,30]. Explanation and relevance of the SPANC hypothesis will be expanded in section 1.5.1.

1.3.6 Genome Sequence Assembly Comparison

The initial Newbler genome assemblies provided by the sequencing centers contained 179, 269, and 269 contigs for JY1305, E64MW, and JY1701 respectively. To ensure optimal assembly, we reassembled the sequence reads for each strain using MIRA version 3.0[11], which resulted in 159, 274, and 324 contigs for JY1305, E64MW, and JY1701 respectively. Tablet [31]was used to visualize the contigs and to investigate the apparent quality of both assemblies. Figure 1.4 illustrates the differences between the Newbler 2.3 and MIRA 3.0 assemblies, showing a side-by-side comparison of the assembled sequence covering a homologous region of a large contig found in both assemblies. The difference in coverage across this region shown in the comparison is typical of the differences in assembly results of MIRA 3.0 and Newbler 2.3. Feature prediction using the Newbler 2.3 assemblies resulted in gene undercounts, with 24 apparent genes being missed in the JY1305 Newbler assembly, and 63 and 75 genes being missed in E64MW and JY1701 respectively. Newbler left a residue of 9263, 2897, and 2706 unassembled reads for JY1305, E64MW, and JY1701, respectively, while MIRA left 9183, 3491, and 3659 reads unassembled for JY1305, E64MW, and JY1701, respectively. Based on these observations, we chose to use the MIRA version 3.0 assembly in all subsequent analyses in this work, and contigs deposited at NCBI are from those assemblies. Table 1.4 summarizes the sequence assembly statistics and differences at the initial stage of feature prediction between the two assemblies: 1.4A shows MIRA assembly statistics and 1.4B shows Newbler assembly statistics. *V. vulnificus* JY1305 had greater coverage depth, and hence the fewest number of

contigs. The MIRA contigs were of higher quality, as seen in the image below the table, which shows an example of a region where the construction quality of a MIRA contig is better than the cognate Newbler contig. This outcome justifies selection of the MIRA assembly of JY1305 genome as a reference for subsequent genome analyses.

1.3.7 Gene Retention based on Combined Length and Orthology Criteria

We manually reviewed the annotation comparison results to determine whether the stringency of our initial criteria for gene inclusion may have caused us to miss genes that are found exclusively in the accessory genomes of the E-type draft genomes. The accessory genome is defined as genes that are present in two or more strains, but not in all genomes included in the study. When we simply applied criteria similar to Chen et al. 2003 [7] to merge the Glimmer and GeneMark annotations described in section 1.3.3, numerous shorter genes were omitted. Inclusion of putative genes that were shorter than 150 amino acids in length, but were supported by their membership in an ortholog cluster spanning other completely characterized *Vibrio* spp. added over 700 genes to the gene lists for each of the newly sequenced strains. This increased the number of genes by 21.38%, 21.39%, and 20.90% for *V. vulnificus* JY1305, E64MW, and JY1701 respectively. Table 1.5 summarizes the predicted gene counts based on these criteria for each of the newly sequenced genomes.

1.4 Discussion

1.4.1 Characteristics of E-type Genomes

V. vulnificus C- and E-type have been shown to exhibit differences in pathogenicity and environmental distribution. In addition, previous examination of several housekeeping and putative virulence-associated genes has revealed a number of genetic polymorphisms suggesting that these two genotypes are in the process of diverging into distinct ecotypes [32,33]. One hypothesis of particular interest, referred to as the SPANC (self-preservation and nutritional competence) balance, could potentially offer insight into the niche adaption and differentiation seen in *V. vulnificus* C- and E-genotypes. The SPANC hypothesis has been well characterized in E. coli and demonstrates that clonal populations can experience genetic mutations and phenotypic changes as a result of physiological stress under conditions such as nutrient starvation. These changes often lead to variations in the activity of the global gene regulator, sigma factor (*rpoS*), which governs the general stress response. Decreased RpoS activity can lead to the development of specialized populations which are less resistant to stress but have broader nutritional capabilities and a higher affinity for low nutrient concentrations, whereas the original population is more stress tolerant but less nutritionally competent[29,30]. In aquatic environments, in which nutrients are often limiting and competition for resources is intense, such modifications could confer a selective advantage for these bacterial strains.

It seems plausible that this trade-off between self-preservation (stress resistance) and nutritional competence could be a factor driving the diversification of

V. vulnificus species. By completely sequencing three E-genotypes of *V. vulnificus*, we were able to identify those genes that are specific to E-genotypes. As noted in the section 1.3.5, the GO functional differences in gene content between C- and E-genotypes show that the sequenced E-type genomes are significantly enriched for metabolic functions such as urea and nitrogen cycle metabolism, suggesting that the E-genotypes may possess more versatile metabolic capabilities. Laboratory studies support this finding demonstrating that when *V. vulnificus* C- and E-genotypes are grown in co-culture, E-genotypes are favored under nutrient rich conditions (Rosche and Oliver, unpublished).

Coping with the rapid host transition, from oyster to human likely requires a variety of stress resistance genes, both protective and adaptive. Previous studies have demonstrated the need for stress regulators for adaption to conditions of starvation, osmotic stress, low pH, non-optimal temperatures, and oxidative damage[34]. Studies investigating the ability of *V. vulnificus* to survive stressful conditions have shown that C-genotypes are significantly better able to survive in complement-activated human serum than E-genotypes[35]. Rosche *et al*. demonstrated that C-genotypes exhibit better cross-protection when exposed to multiple stresses, such as osmotic shock followed by H_2O_2 exposure or elevated temperature[33]. Under conditions tested to date, C-genotypes appear to be physiologically more stress tolerant, and this suggests that the SPANC hypothesis may apply in *Vibrio vulnificus*, in that C-genotypes are more capable at self-preservation, while E-genotypes carry additional genes that suggest they may be more capable of nutritional competence. Sequence alignments of the *rpoS* gene for

all six sequence strains did not indicate any major genetic polymorphisms, with only a few amino acid substitutions. The nucleic acid sequence is ~99% identical and the coded protein 98.5% identical. Other genes that may affect the SPANC balance [34]are similarly well conserved. Future studies will need to be performed to investigate the roles of E- genotype specific genes under relevant conditions such as nutrient limitation in order to validate this hypothesis.

1.4.2 Characteristics of C-type Genomes

Mannitol transport and fermentation genes were found to be present in the C-genotype strains but not in the newly 3 sequenced E-genotype strains. Mannitol has been correlated with virulence-associated genotypes (*vcgC* and 16S rDNA type B)[25]. This lack of a mannitol operon (consisting of a dehydrogenase, a phosphotransferase system component, and an operon repressor) in the sequenced E-type strains was identified in a previous study[37,38]. This differentiating feature was also identified in a recent analysis of short-read sequence fragments from four other E-type strains[10]. It is important to note that while many E-genotype strains lack the mannitol operon, phenotypic and molecular testing by the Oliver laboratory at the University of North Carolina at Charlotte has shown that 40% of 73 total tested E-type strains in their study contained the mannitol operon and were able to ferment this sugar[37,38]. The strains sequenced in this study and in the study by Gulig et al. [10]were among those previously known, before sequencing, to be unable to ferment mannitol, and future sequencing should include E-genotype strains that are able to ferment mannitol, to provide a more extensive comparison between these two phenotypes.

Cohen *et al*. (2007) used multi-locus sequence tag (MLST) data to identify a 33-kb genomic island (region XII) on the second chromosome of *V. vulnificus* [37]. This region contained an arylsulfatase gene cluster, a sulfate reduction system, two chondroitinase genes, and an oligopeptide ABC transport system, none of which were found in their "lineage II" (our E-genotype) isolates. They suggested that this region may play a role in the pathogenic process, as both arylsulfatases and the chondroitin sulfate proteoglycan degrading chondrotinase have been speculated to be involved in the penetration of epithelial cells[40,41]. The authors thus speculated that region XII, along with others, could give members of the C-genotypes a selective advantage in their relationships with aquatic environments or human hosts, or both. Gulig *et al*. (2010), in their *V. vulnificus* sequencing study, suggested that the ability to scavenge sulfate groups could facilitate survival in the human host, where free sulfur is limited[10]. Cohen *et al*. (2007) identified region XII in 32 of the 37 lineage I genotypes included in their study are the C-genotypes, *V. vulnificus* CMCP6, *V. vulnificus* MO6-24/O and *V. vulnificus* YJ016 , but in only 3 of the 6 lineage II strains[39]. Consistent with their findings, we identified 83.3% of the XII region as being present only in the C-genotypes (YJ016, CMCP6, MO6-24/O), and not in the three newly sequenced E-genotypes.

Type IV secretion system gene *VirB4* (VV2_0638) was found to be present in C-genotype strains (*V. vulnificus* YJ016 and *V. vulnificus* CMCP6) but absent in the newly sequenced E-genotype isolates and *V. vulnificus* MO6-24/O. Type IV bacterial secretion systems (T4SS) are responsible for the translocation of molecules such as DNA, proteins, and toxins out of the cell and into the immediate

environment or the host cell[42,43]. This system is composed of the T-pilus and membrane-associated complex and is constructed from 12 VirB proteins, several other Vir proteins, and a coupling protein (*VirD4*) [43,44]. Of these proteins, *VirB4* serve as a energizing component as this gene has been associated with ATPase functionality[44,46]. Because this system is associated with the transfer of DNA (conjugation) and also toxins, it is also often implicated with pathogenicity. Our *V. vulnificus* E-genotype strain sequencing suggests that these T4SS components play a role in infections caused by C-genotypes (*V. vulnificus* YJ016 and *V. vulnificus* CMCP6). 70% of the predicted *virB* operon sequence of the T4SS has been observed to be present in the C-genotypes (*V. vulnificus* YJ016 and *V. vulnificus* CMCP6) and not in M06-24 or the E-genotypes [47]. Sequencing of more C- and E-genotypes should be performed to investigate whether the presence of this operon displays a trend towards virulent strains in *Vibrio vulnificus*.

1.4.3 Assessment of Genome Assembly and Identification

Genomic assembly and feature prediction assessment metrics are based on a numerical scale. Lower contig counts, higher gene prediction counts, and high N50 values are ideal in constructing high-quality draft genomes. N50 is defined as the size of the contig that represents 50% of the assembled genome. Smaller contig counts are used as an indicator that fewer gaps were constructed when assembling the genome, which can be interpreted as a higher probability that the genome is complete. The more complete the genomic sequence, the higher our confidence that the gene content has been completely captured for the newly sequenced organism. In this work we used comparative approaches to select the best assembly to use for

the E-type genomes to use on our comparative genomic analysis. By re-assembling the sequence read fragments for *V. vulnificus* JY1305 with an alternative de novo assembler, we constructed an improved draft genome based on decreased contig counts and increased gene prediction counts. This workflow was also applied to *V. vulnificus* E64MW and JY1701, even though the available data for these genomes led to slightly higher contig counts with MIRA than in the original Newbler assemblies. However, as noted above, gene counts were corrected by substantially improved. Baker *et al*. 2012 [48] stated that the bioinformatics community still struggles with next generation sequencing data and analysis; in part this is because benchmark data sets and algorithms are not available. An even greater problem is that the majority of the microbial comparative genomics studies do not include their bioinformatics analysis steps in sufficient detail to replicate the analysis process, so it is uncertain what precautions were taken to ensure that genetic components were captured. In this work we have taken care to produce complete computational workflow details, allowing others to identify the same genetic components in these genomes. By performing multiple assemblies and gene prediction methods we are able to validate our computational measures and identify unknown genetic characteristics with confidence. As additional sequencing data is obtained, for example to fill gaps and finish these genomes, only minor changes in differential gene list should result if the same pipeline is used.

1.5 Conclusion

Three E-genotype strains of *Vibrio vulnificus* have been sequenced to over 99% completion. The genomes have been assembled using *ab initio* methods and

contig sequences have been deposited in the NCBI Whole Genome Shotgun archive. We expect this effort to provide insights into structural rearrangements among the C-genotype and E-genotype strains, but we do not expect additional sequencing to significantly alter the membership in the list of strain-differentiating genes reported in this chapter. The presence or absence of a particular gene in a specific genotype provides an initial target for functional differentiation. This work also provides the *V. vulnificus* community with a valuable reference for functional study of determinants of virulence, survival, host-specificity and adaptation, and facilitates the use of high-throughput approaches to assess the functional differences via the study of the *V. vulnificus* transcriptome and the possibility to investigate the evolutionary event or series of events that led to the environmental niche specification seen among the *V.vulnificus* genotypes.

Also in this chapter, we began to investigate the types of metrics used to evaluate the quality of a draft genome and its annotations and the steps that can be taken to determine that they are sufficiently accurate and complete to capture the true genetic make-up of an organism. We showed that combining *ab initio* gene predictions and comparative information we can identify and interpret gene content in a comparative genome analysis. In chapter 2, this analysis is expanded to include approaches to benchmarking when a reference genome is available, and to systemically test the outcomes of different workflow choices in microbial genome assembly and annotation.

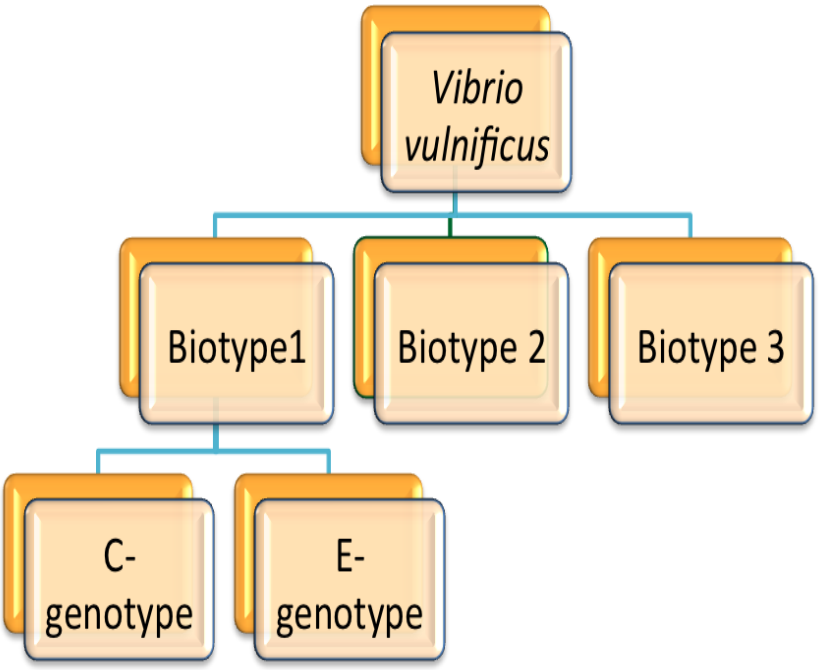
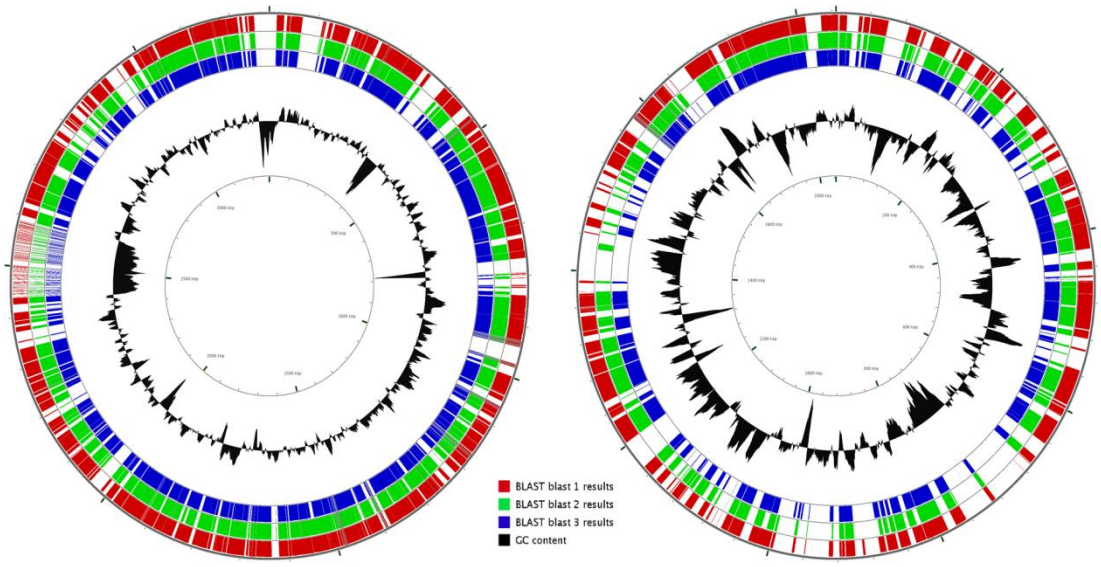


FIGURE 1.1: *Vibrio vulnificus* biological classifications. Biotype 1- primarily associated with human death, Biotype 2 – primarily pathogen of marine organisms, and Biotype 3- to date only reported in wound infections. C-genotype – strains isolated from clinical sources, most commonly found in human infections and E-genotype – strains isolated from environment, rarely cause human disease.



Chr .1
Size: 3.8 Mb

Chr .2
Size: 1.9 Mb

FIGURE 1.2 Circular maps of the sequencing contigs of *V. vulnificus* JY1305, E64MW, and JY1701.From the outside in, the first circle (red) represents *V. vulnificus* JY1305 genomic contigs, the second circle (green) represents *V. vulnificus* JY1701 genomic contigs, and third circle (blue) represents *V. vulnificus* E64MW genomic contigs. The circles represent BLAST alignment of contigs against the *V. vulnificus* CMCP6 reference genome. Circle 4 shows GC content. Figure generated using CGView.

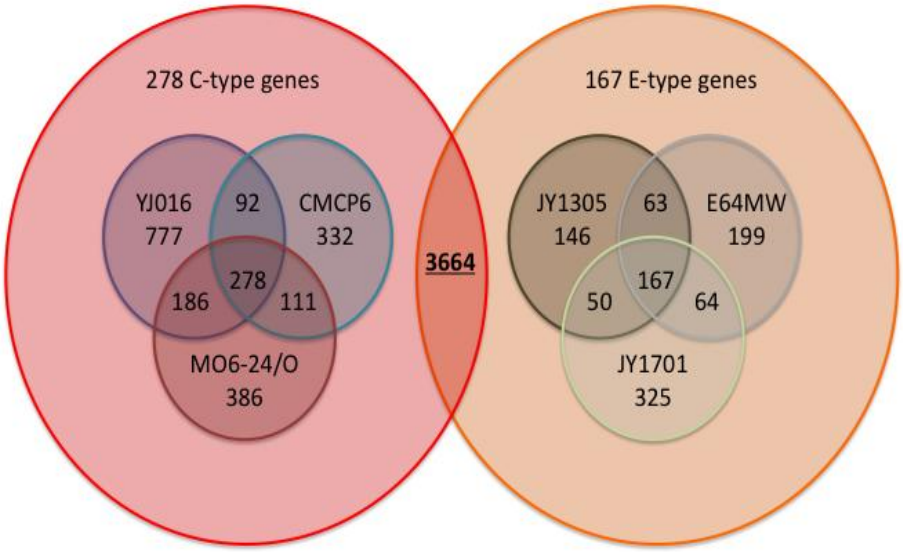


FIGURE 1.3: *Vibrio vulnificus* genomic content differential Venn diagram.

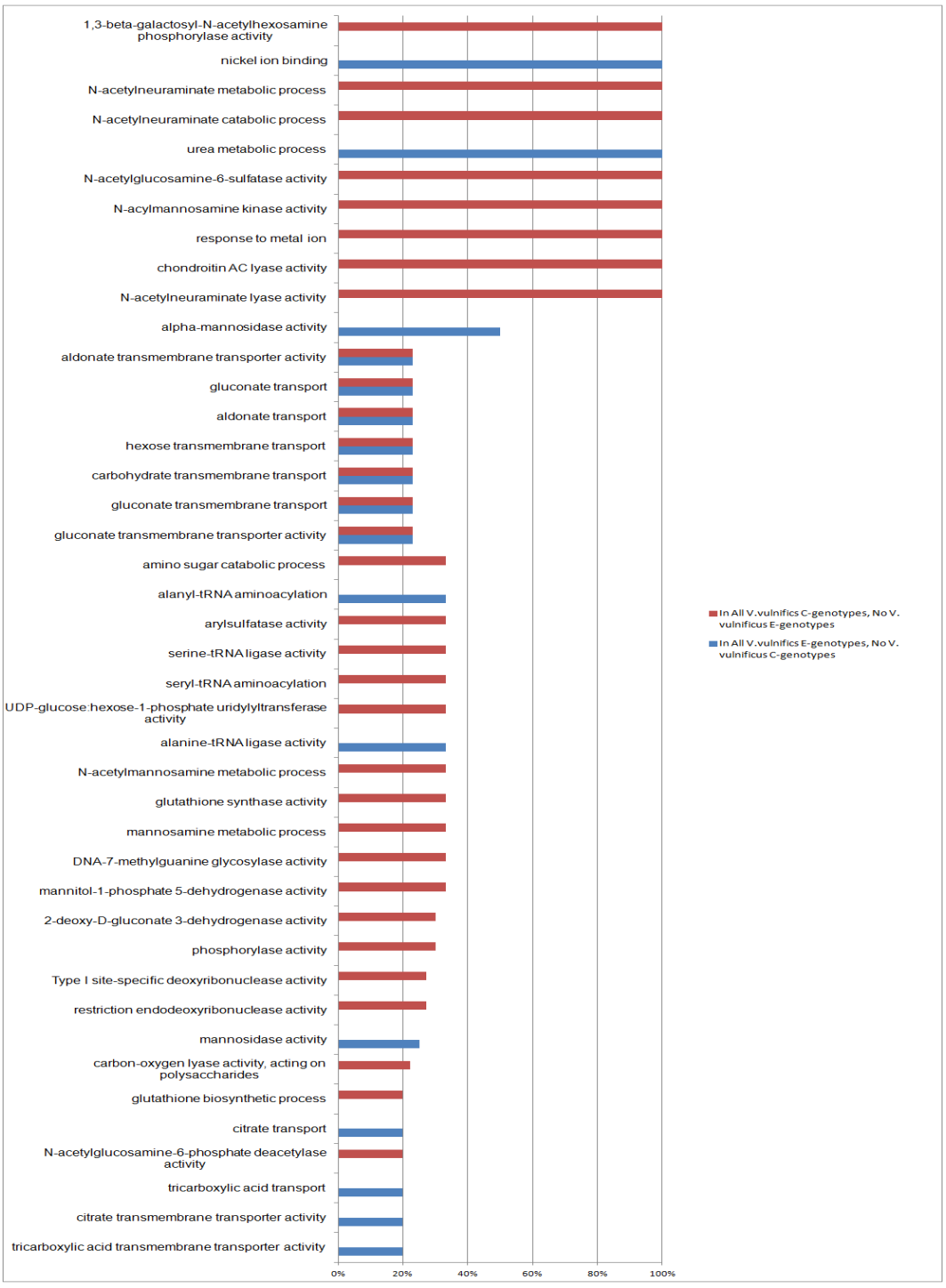


FIGURE 1.4: Gene Ontology (GO) functional differences between C- and E- genotypes.Figure shows GO functional categories which are enriched in C-genotypes of *V. vulnificus* relative to E-genotypes (blue) or E-genotypes relative to C-genotypes (red). Percentages represent percent of genes under each category that are differential between the genotypes. Percentages of less than 20% are not depicted.

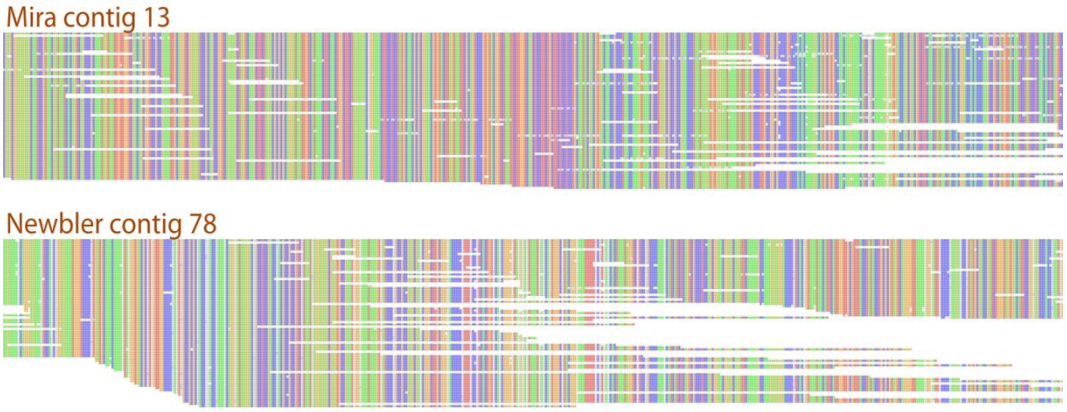


FIGURE 1.5: Homologous sequence contig comparison between MIRA 3.0 and Newbler 2.3.

TABLE 1.1: Summary of A.) Assembly and B.) Genomic characteristics for *V. vulnificus* JY1305, E64MW, and JY1701.

A.)

Sequencing Statistics	*V. vulnificus* JY1305	*V. vulnificus* E64MW	*V. vulnificus* JY1701
# of reads	671,521	376,287	321,091
# of nucleotides sequenced	188,710,063 bp	96,530,017 bp	73,115,338 bp
Average read length	281 bp	257 bp	228 bp
# of contigs	159	271	329
N50	237659 bp	69696 bp	36756 bp
N90	54287 bp	14424 bp	9249 bp
Largest Contig	489256 bp	163962 bp	112761 bp
Depth Coverage	~33x	~17x	~13x
Estimate Genome Size	5.7 Mb	5.7Mb	5.6Mb
Genome Coverage	~99.9%	~99%	~99%

B.)

Genome Features	*V. vulnificus* JY1305	*V. vulnificus* E64MW	*V. vulnificus* JY1701
Chromosome Number	2	2	2
Plasmid	None	N/A	N/A
G+C content %	46.7%	46.7%	46.5%
Predicted Genes	4235	4301	4425
# of predicted tRNAs	115	109	96
# of predicted rRNAs	23	17	15

TABLE 1.2: Key differential genes found in *V. vulnificus* C-genotypes that are NOT present in the E-genotypes.

Strain	Chr	Locus Tag	Product Description	GO ID	GO Term
CMCP6	2	VV2_0726	Sialic acid-induced transmembrane protein YjhT	GO:0005975	carbohydrate metabolic process*
	2	VV2_0729	sialic acid utilization regulator RpiR family	GO:0005975	carbohydrate metabolic process*
	2	VV2_0730	N-acetylneuraminate lyase	GO:0008747	N-acetylneuraminate lyase activity*
	2	VV2_0731	TRAP-type transport system large permease component	GO:0016021	Integral to membrane+
	2	VV2_0732	TRAP-type transport system small permease component	N/A	N/A
	2	VV2_0733	TRAP-type system periplasmic component	GO:006810	transport+*
	2	VV2_1509	Putative two –component response regulator & GGDEF f protein YeaJ	GO:0009190	cyclic nucleotide biosynthetic process+*
	2	VV2_1510	Response regulator	GO:0000156	two-component response regulator activity+*
	2	VV2_1106	Arysulfastase A	GO:0008484	sulfuric ester hydrolase activity*
	2	VV2_1107	Arylsulfatase regulator	GO:0008152	metabolic process+*
	2	VV2_1108	Arylsulfatase A	GO:0008449	N-acetylglucosamine-6-sulfatase activity*
	2	VV2_1109	Arysulfatase	GO:0008484	Sulfuric ester hydrolase activity*
	2	VV2_0074	RsbS, negative regulator of sigma-B	N/A	N/A
	2	VV2_0075	anti-sigma B factor RsbT	GO:0005524	ATP binding+
	2	VV2_0076	Serine phosphatase RsbU, regulator of sigma subunit	GO:0008152	metabolic process+*
	2	VV2_0077	Two-component system sensor protein	GO:0004673	protein histidine kinase activity+*
	2	VV2_0735	N-acylmannosamine kinase	GO:0009384	N-acylmannosamine kinase activity*

TABLE 1.2: (Continued).

MO6-24/O	**2**	**VVMO6_03282**	**Putative two-component response regulator & GGDEF family protein YeaJ**	**GO:0009190**	**cyclic nucleotide biosynthetic process+***
	2	VVMO6_03283	Putative two-component response regulator	GO:0003677	DNA binding*
	2	VVMO6_04101	Sialic acid-induced transmembrane protein YjhT	GO:0005975	Carbohydrate metabolic process*
	2	VVMO6_04102	Sialic acid utilization regulator RpiR family	GO:0005975	Carbohydrate metabolic process*
	2	VVMO6_04103	N-acetylneuraminate lyase	GO:0008747	N-acetylneuraminate lyase activity*
	2	VVMO6_04104	TRAP-type transport system large permease component	GO:0016021	integral to membrane+
	2	VVMO6_04105	TRAP-type transport system small permease component	N/A	N/A
	2	VVMO6_04106	TRAP-type system periplasmic component	GO:0006810	transport+*
	2	VVMO6_04498	Arysulfastase A	GO:0008484	sulfuric ester hydrolase activity*
	2	VVMO6_04499	GALNS arysulfatase regulator (Fe-S oxidoreductase)	GO:0008152	metabolic process+*
	2	VVMO6_04500	Choline-sulfatase	GO:0008449	N-acetylglucosamine-6-sulfatase activity*
	2	VVMO6_04501	Arysulfastase	GO:0008484	Sulfuric ester hydrolase activity*
	2	VVMO6_03523	rsbS, negative regulator of sigma-B	N/A	N/A
	2	VVMO6_03524	anti-sigma B factor RsbT	GO:0005524	ATP binding+
	2	VVMO6_03525	serine phosphatase RsbU, regulator of sigma subunit	GO:0003824	Catalytic activity+*
	2	VVMO6_03526	two-component system sensor protein	GO:0004673	protein histidine kinase activity+*
	1	VVMO6_02633	PTS system, mannitol-specific IIC component	GO:0016301	kinase activity+*
	1	VVMO6_02634	Mannitol-1-phosphate 5-dehydrogenase	GO:0008926	mannitol-1-phosphate 5-dehydrogenase activity*
	1	VVMO6_02635	Mannitol operon repressor	N/A	N/A

TABLE 1.2: (Continued).

YJ016	**2**	**VVA0202**	**Transcriptional regulator**	**GO:0003677**	**DNA binding***
	2	VVA0325	Putative fimbrial protein Z, transcriptional regulator	GO:0003677	DNA binding*
	2	VVA0326	GGDEF family protein	GO:0009190	cyclic nucleotide biosynthetic process+*
	2	VVA0327	Putative fimbrial protein Z, transcriptional regulator	GO:0003677	DNA binding*
	2	VVA1199	Putative N-acetylneuraminate lyase	GO:0008747	N-acetylneuraminate lyase activity*
	2	VVA1200	TRAP-type C4-dicarboxylate transport system, large permease component	GO:0016021	integral to membrane+*
	2	VVA1201	TRAP-type C4-dicarboxylate transport system, small permease component	N/A	N/A
	2	VVA1202	TRAP-type C4- dicarboxylate transport system, periplasmic component	GO:0006810	Transport+*
	2	VVA1632	Arysulfastase A	GO:0008484	sulfuric ester hydrolase activity*
	2	VVA1633	Arylsulfatase regulator	GO:0055114	oxidation-reduction process+*
	2	VVA1634	Arylsulfatase A	GO:0008449	N-acetylglucosamine-6-sulfatase activity*
	2	VVA1635	Arysulfatase A	GO:0008484	sulfuric ester hydrolase activity*
	2	VVA0581	anti-anti-sigma regulatory factor	N/A	N/A
	2	VVA0582	anti-sigma regulatory factor	GO:000552	ATP binding+
	2	VVA0583	indirect negative regulator of sigma-B activity	GO:0003824	Catalytic activity+*
	2	VVA0584	conserved hypothetical protein	GO:0016310	phosphorylation+*

TABLE 1.3: Key differential genes found in *V. vulnificus* E-genotypes that are NOT present in the C-genotypes.

Strain	Chr. Alignment	Locus tag	Product Description	GO ID	GO Term
JY1305	No LCB alignment	VvJY1305_2152	Hypothetical protein	GO:0019627	urea metabolic process*
	LCB in Vv. CMCP6 chr 1	VvJY1305_1632	Permease	GO:0016020	membrane+*
	LCB in Vv. CMCP6 chr 2	VvJY1305_2975	PTS system, glucose-specific IIBBC component	GO:0006810	transport+*
	LCB in Vv. CMCP6 chr 2	VvJY1305_3160	PKD domain containing protein	N/A	N/A
E64MW	No LCB alignment	VvE64MW_4158	Hypothetical protein	GO:0016151	nickel ion binding*
	LCB in Vv. CMCP6 chr 1	VvE64MW_1434	Permease	GO:0015128	gluconate transmembrane transporter
	LCB in Vv. CMCP6 chr 2	VvE64MW_3479	PTS system, glucose-specific IIBBC component	GO:0005351	hydrogen symporter activity+*
	No LCB alignment	VvE64MW_3886	PKD domain containing protein	N/A	N/A
JY1701	No LCB alignment	VvJY1701_4279	Hypothetical protein	GO:0019627	urea metabolic process*
	LCB in Vv. CMCP6 chr 1	VvJY1701_1508	Permease	GO:0016020	membrane+*
	LCB in Vv. CMCP6 chr 2	VvJY1701_3646	PTS system, glucose-specific IIBBC component	GO:0019627	transport+*
	LCB in Vv. CMCP6 chr 2	VvJY1701_4020	PKD domain containing protein	NA	N/A

TABLE 1.4: Sequence assembly statistics and the preliminary feature predictions between Newbler and MIRA for E-genotype genomes.

A.)

Genome	Estimated genome size	Coverage Depth	% of genome covered	# of contigs	Largest contig	N50	Feature Identification
JY1305	5.7 Mb	~33x	99.9%	159	489256 bp	237659 bp	2974
E64MW	5.7 Mb	~17x	99%	271	163962 bp	69696 bp	2977
JY1701	5.6 Mb	~13x	99%	329	112761 bp	36756 bp	3040

B.)

Genome	Estimated genome size	Coverage Depth	% of genome covered	# of contigs	Largest contig	N50	Feature Identification
JY1305	5.7 Mb	33x	99.9%	179	396819 bp	184539 bp	2950
E64MW	5.7 Mb	17x	99%	269	464851bp	131953 bp	2914
JY1701	5.6 Mb	13x	99%	269	177862 bp	64400 bp	2965

TABLE 1.5: Gene prediction criteria counts for *V. vulnificus* JY1305, E64MW, and JY1701.

Criteria for Prediction Inclusion	***V.vulnificus* JY1305**	***V. vulnificus* E64MW**	***V.vulnificus* JY1701**
Total Predicted	4889	5173	5403
Amino acid length > 150	3482	3535	3652
Amino acid length < 150, but predicted by both Glimmer and GeneMark	7	8	8
Total predicted genes following criteria from Chen *et al.*	3489	3543	3660
Amino acid length < 150, with orthologs in other *Vibrio spp.*	746	758	765
Total genes included in final count	4235	4301	4425
Predicted but not included	654	872	978
Percentage gene gain using orthology criterion	21.38%	21.39%	20.90%

CHAPTER 2: IMPACT OFANALYTIC PROVENANCE IN GENOME ANALYSIS

2.1 Introduction

Comparative genomics studies are executed on the premise that completely characterized and closed reference genomes are used to represent the organisms in the analysis. However, since the development of the next generation sequencing (NGS) technologies, it is more common in microbial comparative genomic studies for incomplete or draft genomes to be used. The majority of newly sequenced bacterial species do not have a closely related species completely sequenced and characterized, making reference-based methods inappropriate to construct their genome. The weakness of reference-based assembly is that it cannot accurately represent regions of which there is no equivalent sequence in the reference. These differentiating regions are often the regions of greatest interest in a comparative genomics study. In the case of *Vibrio vulnificus* two different reference genomes exist, and currently there is not a quantitative way to select which reference genome would be the most appropriate to use in assembly of the newly sequenced genomes. When a newly sequenced genome cannot be assembled based on a reference sequence, *de novo* assembly and *ab initio* gene finders are used.

There are many computational tools for *de novo* assembly and *ab initio* gene-finding on next generation sequencing data. Both of these stages in a genome project are vital for accurate interpretation of genomic data in a comparative genomics study.

Ab initio gene-finders use signal or pattern recognition techniques based in known prokaryotic gene features to identify probable genes. However, gene-finders are very sensitive, both to parameters and training sets used within the application itself, and to the methods chosen to construct the underlying genome that is being annotated. This chapter focuses on the *de novo* assembly and *ab initio* annotation of 5 newly sequenced *V. vulnificus* strains. An automated pipeline approach was used to exhaustively test combinations of methods and parameters at the assembly and annotation stages. The outcome of the comparative genomics study, including identification of enriched gene function categories that often point the way to gene candidates for further molecular investigations, are heavily dependent on the analysis workflow, and a benchmarking approach is recommended in order to establish the optimal approach.

2.2 Background

Many computational methods are available for assembly and annotation of newly sequenced microbial genomes. However, when new genomes are reported in the literature, there is frequently very little critical analysis of choices made during the sequence assembly and gene annotation stages. These choices have a direct impact on the biologically relevant products of a genomic analysis – for instance identification of common and differentiating regions among genomes in a comparison, or identification of enriched gene functional categories in a specific strain. That is, there are consequences both for biological and clinical relevance of the results in terms of accuracy and completeness (or sensitivity and specificity). Inconsistencies arise from the algorithms selected, the parameters used in those algorithms, and the order in

which operations are carried out. The impact of such inconsistencies is multiplied genomes to be compared are analyzed with different workflows. Tracking the analysis history of the data – its analytic provenance – is critical for reproducible analysis of genome data. Here, we examine the outcomes of different assembly and analysis steps in typical workflows, using as a data set the comparison of assembly and features across strains of *Vibrio vulnificus*.

Next generation sequencing (NGS) has revolutionized the study of microbial genomics, by making the data required to complete a genome available within days. The bottleneck has thus moved to the analysis stage of the experiment. To handle the millions of sequence read fragments produced by the NGS platforms, a variety of assembly approaches have been developed [49-51]. In most instances the assembler produces a set of contigs or scaffolds, which still leaves the genome in dozens to hundreds of pieces. Until a group adopts the organism for full analysis it is no longer common to completely finish or close a newly sequenced genome. Usually, we evaluate the "success" of the draft assembly with two metrics: the number of contigs produced and the N50 value. Lower contig counts and higher N50 values are considered optimal. N50 is defined as the size of the contig that represents 50% of the assembled genome. A contig is a consensus of overlapping DNA sequencing reads that represent a region of the newly sequenced organism's DNA. However, Parra *et al*. [52] and others[48] reported that choosing assemblies with higher N50 values frequently results in conserved genes going undetected in benchmark studies. If a gene sequence is omitted due to errors at the assembly stage it will not be annotated, leading to inconsistencies in downstream analyses.

There have been several efforts to assess the quality of assemblies produced by *de novo* methods. *De novo* assembly is defined as assembling DNA sequencing reads without the aid of a reference genome. The GAGE [53] and the Assemblathon [54] projects provided gold-standard data sets and a consistent environment for peer evaluation of assembly methods. Recently, NGS read assemblers were evaluated using bacterial datasets in the GAGE-B study. Magoc *et al*. [55] showed that a single library prep and deep (100x -250x) sequencing coverage is sufficient to capture the genomic content of most bacterial species, but from the same base data demonstrated that there is wide variation in the final assemblies produced by different methods.

Analysis of genomes does not stop at assembly, however. There exists a wide range of methods for finding features, or annotation of the assembled data. Genome annotation includes identification of the gene sequences within a contig, and assignment of function based on similarity to known genes or sequence patterns. *Ab initio* gene finders and methods for functional assignment each have their own associated assumptions and errors, and results from one method are unlikely to agree completely with those from another [48]. Assembly and annotation are the two major components of the bacterial genomics workflow, and there are an astonishing number of combinations of methods that can be used to carry out just these two steps.

When we survey the literature in microbial genomics, we find that investigators depositing microbial sequences have not come to a consensus on the best pipeline for genome analysis. Several different assemblers are in common use. Annotation methods may include anything from simply comparing the genome to a reference by using BLAST, to using *ab initio* genefinders, to using integrated

annotation pipelines provided by sequencing centers. Despite over a decade of literature on the performance of *ab initio* genefinders and annotation pipelines, [56-59] nearly any reasonable workflow seems able to pass peer review (Figure 2.1), and so the genome annotations found in the public databases vary widely in analytic provenance. Especially in the absence of ground truth, the proliferation of analysis options can lead to inconsistencies (comparing apples to oranges) and ultimately to errors in biological interpretation. It is not possible to distinguish a true target, such as a gene that differentiates one genome from its near relatives, from an artifact introduced at the assembly or annotation steps. Yet investigators often seem to remain unaware of the impact of their choices, and how the selection of Glimmer[13] rather than GeneMark[14] (for example) may result in a greatly altered story when they begin to analyze the apparent content of a newly sequenced genome. Figure 2.1 is a summary of the major elements of current genomic workflows based on a census of 2013 bacterial genome announcements in recent issues of the journal GenomeA (American Society of Microbiology)[60].

In this study, we assess the scope of the data interpretation problem caused by variation in pipeline choices. Starting with five *V. vulnificus* strains for which paired-end Illumina sequence was collected by the laboratory of Dr. Craig Baker-Austin (personal communication), and one *V. vulnificus* genome with a high quality finished sequence that has been continually revised and updated[23], we apply well-regarded assembly and annotation methods, in different combinations, to the data. We have chosen to focus on the most popular methods in each category, because workflow

construction from multiple options is a combinatorial problem, and not all combinations make sense.

The case study data demonstrate the influence of choices made during the assembly and annotation stages on biological interpretation of newly sequenced genomes. *Vibrio vulnificus* is a bacterium commonly found in estuarine waters and mollusks. It is responsible for 95% of all deaths resulting from seafood consumption in the United States[3]. There are both clinical isolates and environmental genotypes associated with this bacterium, making it a prime candidate for a clinically relevant comparative genomics study. In the present study, we demonstrate the direct impact of parameter and method choices on the output of a comparative genomics analyses among newly sequenced strains of *Vibrio vulnificus*. The results highlight the need for contributors of genomic data to provide complete information about the workflow (analytic provenance) of their assembled and annotated genomes as they do for library preparation steps, and for application of consistent workflows, justified by benchmark testing where possible, to be used throughout a project.

2.3 Material and Methods

2.3.1 Genome Sequencing and Sequencing Simulation

V. vulnificus strains were sequenced at The Genome Analysis Centre (TGAC) using the Illumina HiSeq2000 platform. Sequencing was carried out on pooled libraries, using pools of 12 strains in one lane of the Illumina HiSeq 2000, and producing on average 100 base pair (bp) paired-end (PE) reads. *V. vulnificus* CMCP6 chromosome 1 and 2 genome sequences were used to construct a simulated data set of 100 bp PE reads. The simulated read (SR) set was constructed with ART version 1.5.0 using the program art_illumina[61]. The simulation parameters used were as

follows: data type “paired end”, read length “100”, fold coverage “100”, and quality score “20” (forward and reverse sequence reads). This dataset was used as a benchmark to evaluate the performance of the *de novo* assemblers, gene prediction algorithms, and annotation methods to reproduce the published sequence and annotations of the CMCP6 genome. *V. vulnificus* CMCP6 was recently re-annotated and is regarded as the most complete and accurate of the published *V. vulnificus* genomes at the time of this writing.

2.3.2 Data Cleansing

FastQC was used to evaluate the quality of the sequence reads for each strain[62]. Any repetitive sequence identified by FastQC was removed from the dataset using an in-house perl script. If a sequence read contained 1 or more ‘N’, both the read and its pair were removed. After the data-cleansing steps were completed we sampled a subset of reads for each strain that was equivalent to 100x coverage based on the Lander and Waterman statistic[63]. After the data-cleansing steps were completed each newly sequenced isolate read set contained 11,400,000 paired reads. In the case of *V. vulnificus* CMCP6, the ART sequencing simulation program art-illumina generated 6,620,286 paired reads for CMCP6 using an identical threshold. This difference may be due to use of an alternative mathematical formula for calculating genome coverage in ART.

2.3.3 Sequence Assembly

Initially, each read set was assembled with VelvetOptimiser version 2.2.0 and Velvet 1.0.17 in order to identify an optimal kmer value for assembly and construct an initial contig set. A kmer value is used in a *de novo* assembly to set a minimum length

on the number of contiguous nucleotides that should overlap to construct a contig sequence. The optimal kmer values were 79 for *V. vulnificus* CIP8190 and CECT5763, 83 for *V. vulnificus* CMCP6 and 87 for *V. vulnificus* CECT5198, CECT4606, and CECT4886. The VelvetOptimiser parameters were then used to initiate the Velvet assembler. The VelvetOptimiser hash value (kmer) was set to a range of 73 to 93. The read description parameter was set to "-shortPaired". The VelvetOptimiser optimal kmer value was also used as the input kmer value for ABySS version 1.2.6 (abyss-pe) and SOAPdenovo version SOAPdenovo127mer. The default paired-end parameters were used for both assemblers.

2.3.4 Contig Comparison

MUMmer 2.3 was used to create sequence alignments between assembled contigs, within collections of assemblies for the same genome and among genomes.

2.3.5 Genome Annotation

Ab initio gene-finding and functional annotation for each contig set was performed using the in-house workflow microbial assembly and annotation pipeline constructed in the Taverna workflow management system [64]. This workflow executes parallel assembly-to-analysis pipelines on a genomic data set. The *ab initio* annotation methods implemented include Glimmer3.02, GeneMark.hmm and the Rapid Annotation using Subsystem Technology (RAST) [65] web service. The training model used for *ab initio* gene finding with Glimmer and GeneMark was constructed based on published *Vibrio vulnificus* annotations available in the NCBI database. The RAST web service parameters used were as follows: the genetic code

was set to 11 for bacteria, taxonomy id was set to 672 for genus Vibrio, and the corresponding sequencing statistics for each strain were provided to the web service.

2.3.6 Ortholog Identification and Functional Annotation

OrthoMCL[15] was used to cluster gene predictions with reference genes in the *Vibrio vulnificus* CMCP6 genome. For this application a cluster threshold of 95% identity was used. OrthoMCL[15] was also used to make connections between orthologs among sequenced *Vibrio vulnificus* strains, with a clustering threshold of 70% identity. Gene ontology (GO) terms were assigned using the BLAST2GO software [66]. BLAST2GO was used to perform a BLASTP against the nr (non-redundant) protein database, with e-value cut-off set to 1^{E-6}. GO annotations were assigned based on the BLAST2GO database version b2g_mar13. BLAST2GO assigns GO terms based on a weighted system of evidence codes.

2.3.7 Content and Functional Comparison

For comparison of assembly-to-annotation workflow outcomes and for comparisons of genomic content, we used the GenoSets software application[18]. The annotations produced by each workflow were loaded into the GenoSets application, which enables comparisons among multiple genomes. Each alternate annotation was treated as a separate "genome" in the GenoSets system. We followed the same gene clustering procedure used in Morrison *et al*. 2012 [2] to define sets of genes that differentiate between genomes. To differentiate between the assembly-to-analysis pipeline outcomes, the approach was modified to reflect the expectations that gene sequences arising from different analysis workflows would be highly similar. OrthoMCL clustering was performed against the *Vibrio vulnificus* reference genome

CMCP6 and clusters were formed based on a shared sequence similarity of 90%, instead of the OrthoMCL default parameter of 50%. The increase in stringency to 90% shared sequence similarity results in tightly constrained gene clusters, and allows for the possible of identified genes on the ends of contig that may have not been predicted in their entirety.

2.4 Results

2.4.1 Workflow Dependent Outcomes in Simulated Assembly Case

As a basis for choosing an appropriate analysis pipeline for newly sequenced *V. vulnificus* genomes, we first generated simulated read data from the genome of *V. vulnificus* CMCP6. This genome was initially sequenced using Sanger sequencing and a traditional genome finishing approach in 2003, [8] and was updated with revised annotation in 2011[23]. The published sequence and annotations served as ground truth for evaluation of pipeline options.

We performed *de novo* sequence assemblies of the simulated data with Velvet (V), ABySS (A), and SoapDenovo (S). GeneMark.hmm (GeneMark)[14] and RAST[65] were then used to identify gene sequences for each contig set. We used OrthoMCL[15] with a stringent similarity cutoff to cluster predicted genes with their counterparts in the 2011 *V. vulnificus* CMCP6 annotation.

The contig counts observed were 205, 144, and 269 for the V, A, and S assemblies, respectively. Table 2.1 summarizes gene counts obtained for each assembly followed by each gene annotation method, for the simulated *V. vulnificus* CMCP6 genomes. To avoid ambiguity, the percentage of genes recovered refers only to predicted genes, which clustered uniquely with one gene in the reference

annotation. Less than 1% of predicted genes cluster with apparent paralogs in the reference genome when clustered at a 95% threshold.

The results presented in Table 2.1 suggest that, while the Velvet assembler[49] does not assemble the simulated data into the smallest number of contigs, it produces the most accurate assembly of the simulated *V. vulnificus* CMCP6 data. Velvet, in combination with the GeneMark[14] *ab initio* gene-finder, may produce the best results on novel *V. vulnificus* sequence data. This type of simple two-step workflow is representative of genome analysis workflows found in the genome announcements surveyed in Figure 2.1. However, it should be noted that the best-performing workflow still resulted in a loss of over 200 previously annotated genes, when reanalyzing simulated *V. vulnificus* CMCP6 data.

2.4.2 Workflow Dependent Outcomes on Novel Genome Data

The published *Vibrio vulnificus* genomes are mainly composed of 2 circular chromosomes, and some are known to have plasmids. The size of the *V. vulnificus* genome is estimated at 5.6Mb-5.8Mb of DNA, and this size is consistent among known strains. The newly sequenced isolates *V. vulnificus* CIP8190, CECT5198, CECT4606, CECT5763, and CECT4886 are all known to have 2 chromosomes and 2, 3, 1, 2, and 2 plasmids, respectively. Table 2.2 describes each genome used in this study and its genomic characteristics, as well as the number of sequence reads available for each genome.

Our analysis here is primarily focused on the performance of the assembly and annotation steps typically used during the construction of a draft genome. Biological findings for these genomes will be the focus of another manuscript, currently in

preparation. Using the workflow framework shown in Figure 2.2, we assembled contig sets and annotation sets for each *V. vulnificus* strain. After the removal of sequence reads containing 'N' characters, and random sampling of read pairs to obtain 100x genome coverage based on the Lander Waterman statistic[63], there were 11400000 paired end reads in the final read sets for each of the newly sequenced strains. The same coverage depth was simulated for *V. vulnificus* CMCP6.

Using the same *de novo* assemblers we applied to the simulated data set, we constructed contig sets ranging in size from 180-630 contigs for each of the input genomes. Table 2.3 summarizes the output of Velvet, Soap, and ABySS assemblies for each *V. vulnificus* strain. We then used MUMmer 2.3[67] to align the contig sets for each strain, using an all-against-all alignment to identify contigs that were similarly constructed between the assemblers. Contig pairs that exceeded coverage and sequence identity cut-offs of 95% were identified as similarly constructed. Figure 2.3 summarizes the conservation of contigs across assemblies. Although counts varied from genome to genome, we observed on average 43 contigs constructed by all three assemblers, 133 found by any combination of two of the three assemblers, and 445 contigs that were uniquely constructed by a specific assembler.

In our analysis of the novel *Vibrio vulnificus* genomes, we included the Glimmer3.0[13] *ab initio* gene-finding method in addition to GeneMark[14] and RAST[65]. Glimmer3.0 is demonstrated to be approximately 96% accurate in gene identification, [13] which is similar to the accuracy that we observed for GeneMark in the CMCP6 case study above. In Table 2.4, we summarize the gene predictions by each of the three prediction methods for each of the three assemblies constructed for

each *V. vulnificus* strain. We find that RAST and GeneMark tend to identify more regions as putative genes sequences than Glimmer for these strains. However, this is not a case of simple over-prediction, since the Glimmer gene sequences are not strictly a subset of the predictions by other methods. As an example, in Figure 2.4 we detail the number of gene overlaps between all possible assembly-to-annotation permutations for *V. vulnificus* CECT4606.

Figure 2.5 summarizes the gene overlaps for *Vibrio vulnificus* CMCP6 and CECT4606 datasets for different genefinders applied to assemblies generated by the Velvet assembler. Gene overlaps are defined as two genes that have the same start and stop signals and strand orientation on the same contig sequence. This is a stringent definition of similarity among predictions. Glimmer tends to predict fewer genes that are outside the common "core" of predictions produced by all three genefinders. It is possible that this reflects greater accuracy, or it may be that Glimmer alone is more conservative in its gene-identification model. RAST (which uses Glimmer in an initial annotation pass) and GeneMark both make, and agree upon, predictions that are excluded from the Glimmer prediction set. It is possible that these two methods are potentially capturing more species-specific genes.

2.4.3 Workflow Dependent Outcomes in Functional Analysis

An archetypal result presented in genomic analyses is the categorization of genes into functional categories. This type of analysis is frequently used to draw conclusions about the energy sources an organism can use for survival, or about the genome's capacity to code for systems related to pathogenicity. To illustrate the impact of workflow choice on interpretation of functional content, we performed a

comparative analysis among the results of six assembly-to-annotation workflows applied to the genome of *V. vulnificus* CECT4866. We used the GenoSets analysis system to perform the comparison of analysis outcomes, treating the annotation set produced by each workflow as if it were an independent "genome".

Each workflow's gene set was assigned Gene Ontology (GO) terms[19,20] as described in Cain *et al*., 2012[28]. GO categories and individual genes having functionality significant enrichment or depletion between the various annotation versions were identified using the Gene Ontologizer[27]. Table 2.6 summarizes the complete GO enrichment set for each of the workflow combinations examined. We first compared annotations produced by a workflow that used the Velvet assembler, followed by either Glimmer or GeneMark. 134 genes appeared in the Glimmer predictions, but not in the GeneMark predictions, resulting in the appearance of statistically significant enrichment or depletion in two GO functional categories. Deoxyribose phosphate metabolic process and deoxyribose phosphate catabolic process p-values were 0.0066 and 0.0072, respectively. 120 genes were identified solely with GeneMark annotations. Use of GeneMark resulted in the appearance of enrichment in GO terms associated with response to stress and iron ion binding, with p-values at 5.99^{E-12} and 0.0017, respectively. The GO terms associated with iron utilization are especially of interest in the context of *Vibrio vulnificus* genomics, because as a pathogen it is especially dangerous to hosts in a condition of iron overload[68]. Iron-protein binding and stress response are potentially regarded as factors contributing to *V. vulnificus*'s pathogenicity. Several studies have reported on the correlation between *V. vulnificus* infections and increased levels of iron in animal

models and infected individuals [3,68,69]. Wright *et al.*[68] showed the injecting mice with iron prior to *V. vulnificus* infection significantly lowered the LD_{50}. Amaro *et al*. [69] showed that after the injection of *V. vulnificus* to an iron-overload mice, they always died within a 48 hour period of inoculation. In this case, changing the assembly-to-annotation analysis pipelines result in a significant change in detected gene content, in a category that is directly relevant to the biology of the pathogen.

We next examined pipelines using the ABySS assembler followed by RAST or Glimmer. 1880 genes were unique to the RAST annotation. Of these, 132 significant GO enrichment terms were identified. In this set we find both iron-binding protein and terms associated with response to stress, again suggesting that the choice of assembly-to-annotation pipeline has the potential to significantly alter biological interpretation. Only 148 gene clusters were unique to the Glimmer set, and only 5 functional categories showed apparent statistically significant enrichment. Comparison of RAST and GeneMark annotations on a SOAPdenovo assembly resulted in approximately 10 statistically significant differences in functional content in either direction, although none of these categories were identified as significant to the biology of *V. vulnificus* in a previous study[2].

While these results are not conclusive, they indicate that at least in the case of *V. vulnificus*, RAST or GeneMark predictions may best reflect the presence of genes in key functional categories, known to be significant in the biology of these organisms.

2.4.4 Workflow Dependent Outcomes in Genome Content Comparison

Another archetypal figure found in nearly every comparative genomics analysis paper is the Venn diagram or its conceptual equivalent. The Venn diagram

provides a convenient method to summarize what the microbiologist really wants to know: what is in strain (or species) A that makes it function differently from strain B? In Figure 2.6, we show the effect on this commonly generated analysis product when different assembly-to-annotation pipelines are used to generate the input data. As an illustrative example, we performed gene content comparisons between *V. vulnificus* strain CMCP6 (clinical genotype) and strain CECT5198 (environmental genotype). In each comparison, the same assembly-to-annotation pipeline was used on each of the genomes being compared. We tested four combinations of assembler and gene-finder. In Figure 2.6, we show that the majority of differences are seen when different annotation methods are used. In contrast, when different assemblers are used with the same annotation method, the number of differential genes is highly conserved. Given the large number of non-identical genes found when different pipelines are used on the same genome, as we saw in the previous examples, the result is as expected – the valuable biological "end product", the set of differentiating genes around which the biologist will build their scientific conclusions, can vary by dozens if not hundreds of members.

2.5 Discussion

Many factors can have an impact on the assembly of next generation sequence data. Typical information captured about the provenance of sequence data focuses on laboratory procedures and conditions, as we see in the MIGS standard for genomic data[70], or in the experiment information preserved in, for example, the NCBI's Gene Expression Omnibus[71]. However, assuming that samples were properly handled and prepared in the laboratory and those procedures and conditions are consistent, there is

still an entire layer of provenance information to be considered. Here, we have considered the analytic provenance of genome sequence data, that is, the computational steps that are executed to process the data and to attach features and functional information that allows for interpretation.

Despite an attitude on the part of researchers and publishers that microbial genome analysis is a solved problem, application of multiple assembly-to-annotation pipelines to the same data demonstrates that analysis outcomes are heavily dependent on pipeline choice. These choices carry forward into comparative content analysis and functional analysis of genomes, and have the potential to significantly impact scientific conclusions.

It is now typical to report on novel microbial genomes in terse genome announcements, abstract-style papers that give little information about parameterization and execution of bioinformatics processes. A survey of these typical papers shows that a wide variety of genome analysis pipelines using combinations of bioinformatics tools, from simple to sophisticated, will pass peer review. However, on closer examination typical pipelines do not produce identical or even similar results. And while in the hands of trained bioinformaticians, the pipelines we tested in this paper may be fine-tuned to produce somewhat more accurate results, the literature surveyed suggests that this is not what is happening “on the ground” in analysis of bacterial genomes. If the protocols outlined in recent genome reports are accurate, in many cases these protocols are no more complex than the simple one assembler, one gene-finder workflows we have analyzed here.

While in many cases, ground truth for novel genome assemblies and annotations is not available, we recommend that creators of microbial genome datasets consider the following strategies to ensure high quality, reproducible analysis. First, if possible, benchmark proposed analysis pipelines using simulated data derived from a high-quality genome sequence that is as closely related to the novel sequences as possible. Second, maintain an awareness of the variability of assembly-to-annotation results. Perform parallel analyses and assess downstream results for pipeline dependence. Finally, maintain a detailed record of the analytic provenance of the secondary data generated from your raw sequence reads, including pipeline steps and parameters.

2.6 Summary

Inconsistencies in genomic analysis can arise depending on the choices that are made during the assembly and annotation stages. These inconsistencies can have a significant impact on the interpretation of an individual genome's content. The impact is multiplied when comparison of content and function among multiple genomes is the goal. Tracking the analysis history of the data – its analytic provenance – is critical for reproducible analysis of genome data.

The work described in this chapter makes clear the importance of keeping consistent annotation methods when constructing draft genomes. In chapter 3, the benchmarking and analysis optimization techniques described here are applied to a population sized sequencing dataset of 25 newly sequenced *Vibrio vulnificus* strains. This dataset contains representatives of each of the three known biotypes of *V. vulnificus*. An in-depth comparative genomics analysis of this magnitude can begin to

investigate the genomic difference of the *V. vulnificus* biotypes and see if they support the literature in the known composite differences between them or perhaps begin to facilitate discussion within the *Vibrio vulnificus* community if potential re-classification of the biotypes and potentially the genotypes is necessary to coincide this new differential genomic data with traditional molecular diagnostic techniques.

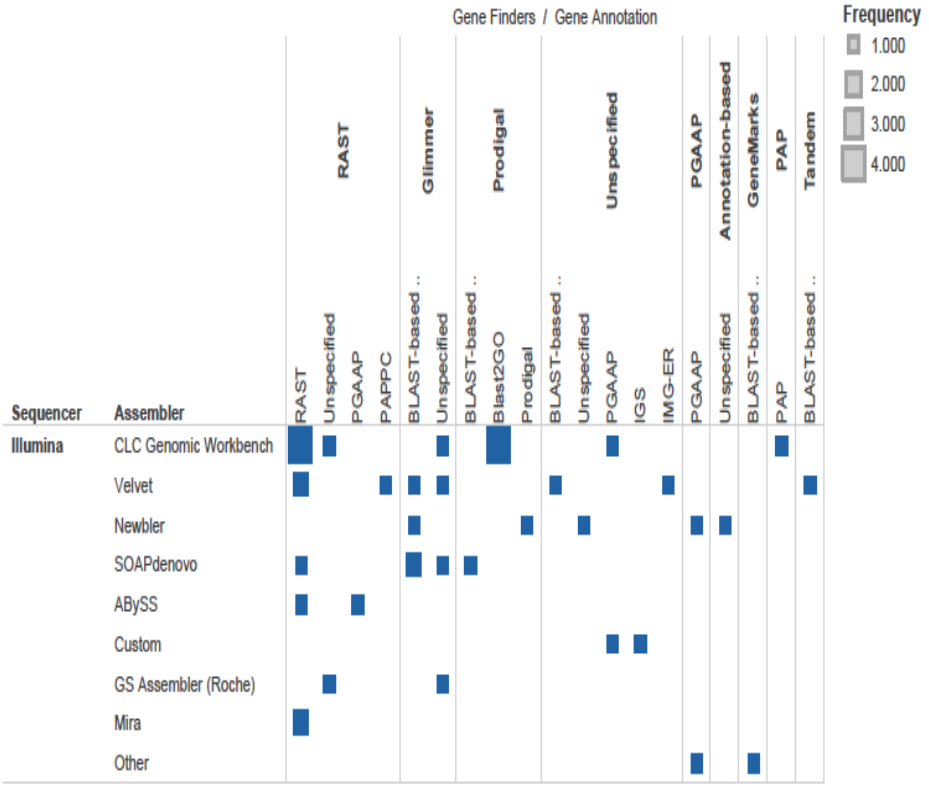


FIGURE 2.1: Crosstab map of frequency levels of assembler and annotation method applied to Illumina data.Summary of the major elements of current genomic workflows based on a census of 2013 bacterial genome announcements in recent issues of the journal GenomeA and Journal of Bacteriology. Frequency represents the number of times that particular combination of sequencer, assembler, and annotation was encountered in survey of 40 papers.

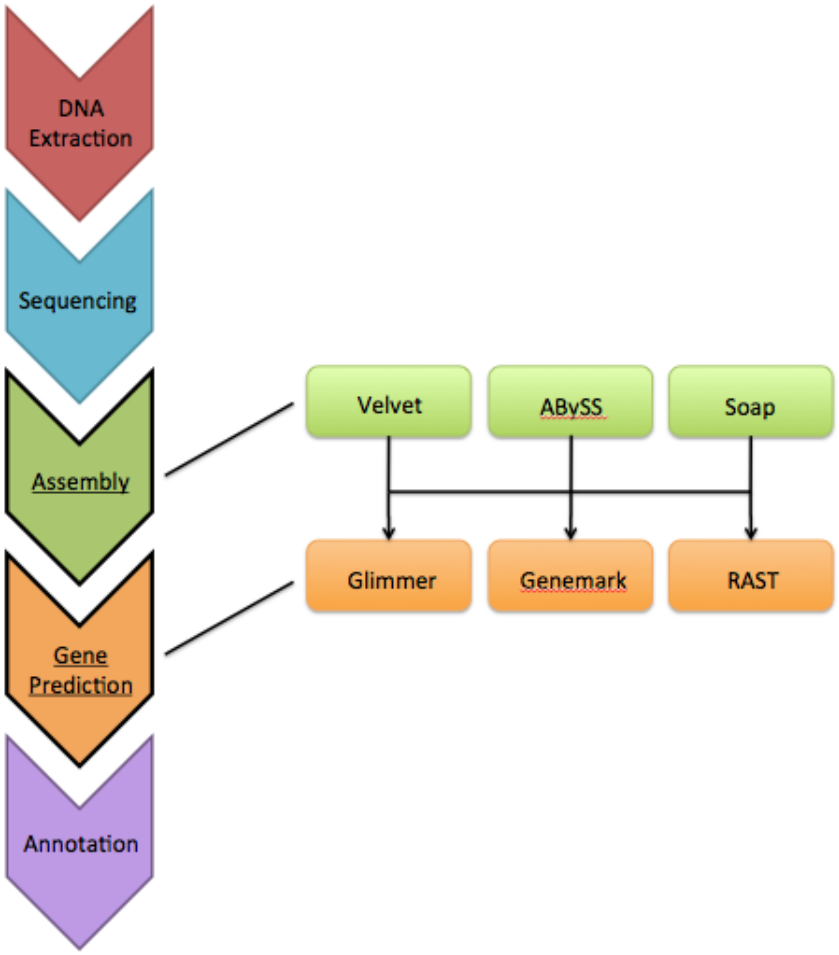


FIGURE 2.2: Workflow framework of assembler and annotation methods.

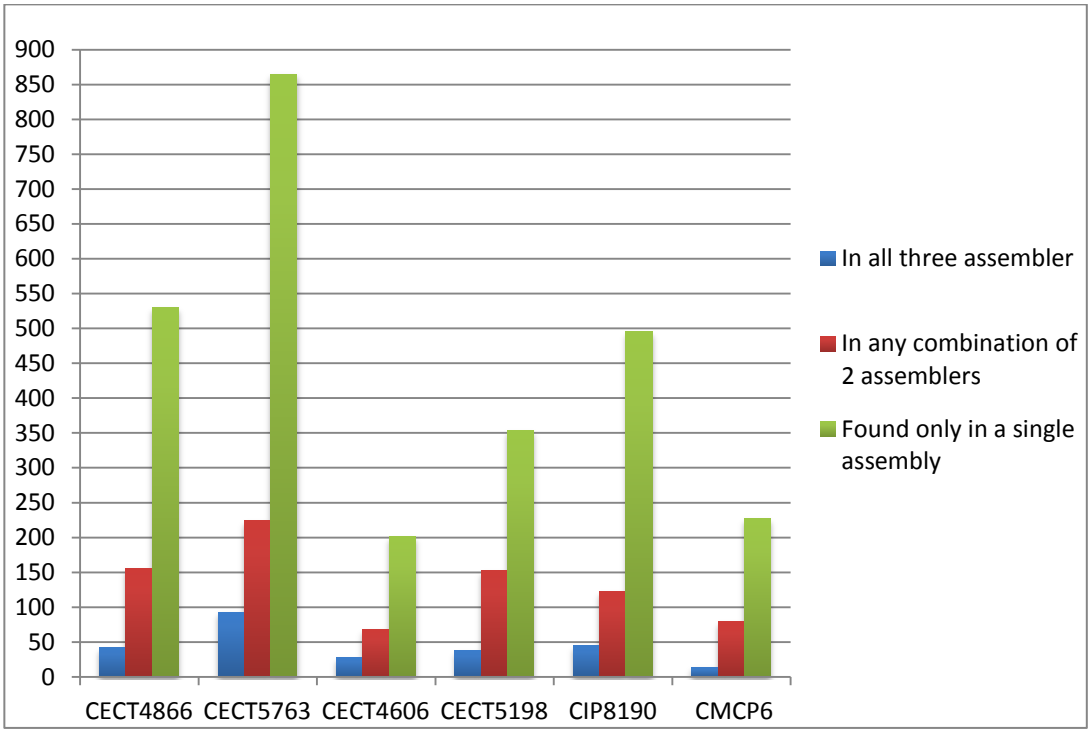


FIGURE 2.3: Comparison count of highly conserved contigs for all *V. vulnificus* strains included in this study.

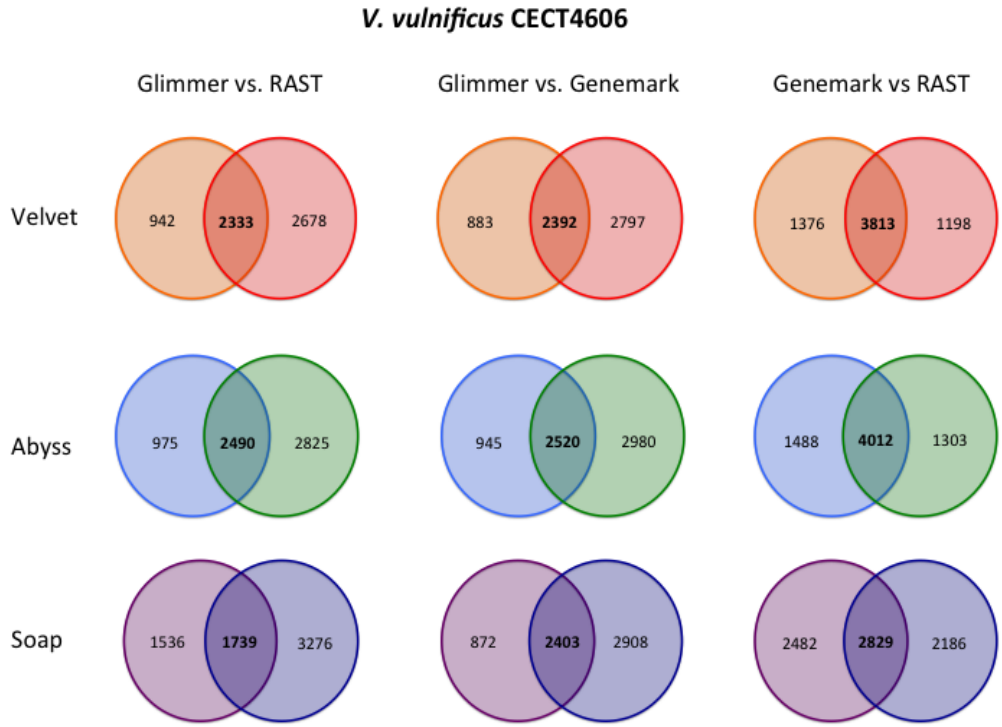


FIGURE 2.4: *Vibrio vulnificus* CECT4606 gene overlap counts.Figure shows the number of gene overlaps between all possible assembly-to annotation permutations for *V. vulnificus* CECT4606. Gene overlaps are defined as two genes that have the same start and stop signals and strand orientation on the same contig sequence.

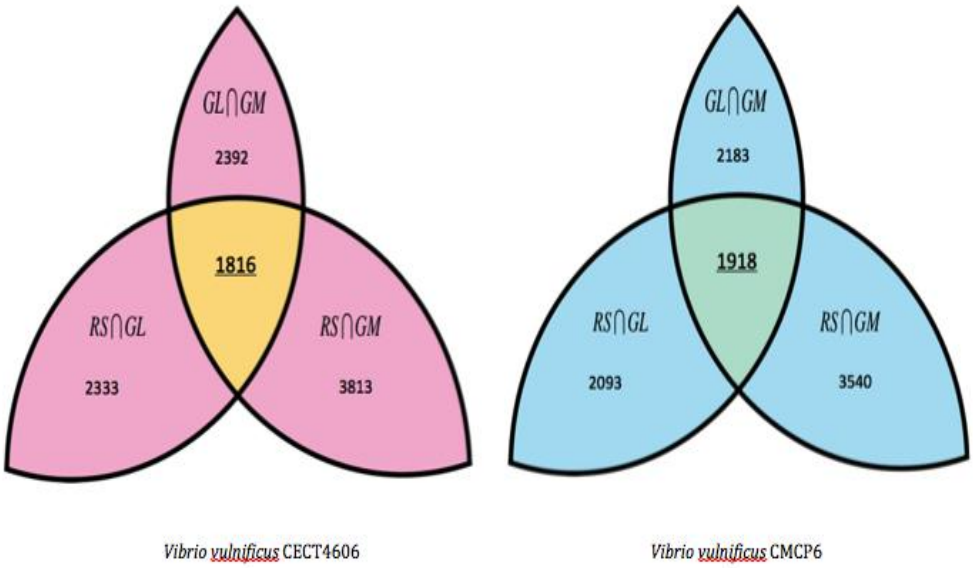


FIGURE 2.5: *Vibrio vulnificus* CECT4606 and CMCP6 gene overlap counts.Each segment in the Venn diagram represents the intersection of the number of genes that were identified in any combination of 2 gene prediction methods. GL ∩ GM = the number of genes that were identified in Glimmer and GeneMark. RS ∩ GM = the number of genes that were identified in RAST and GeneMark. GL ∩ RS = the number of genes that were identified in Glimmer and RAST. The Velvet assembly was used for this comparison.

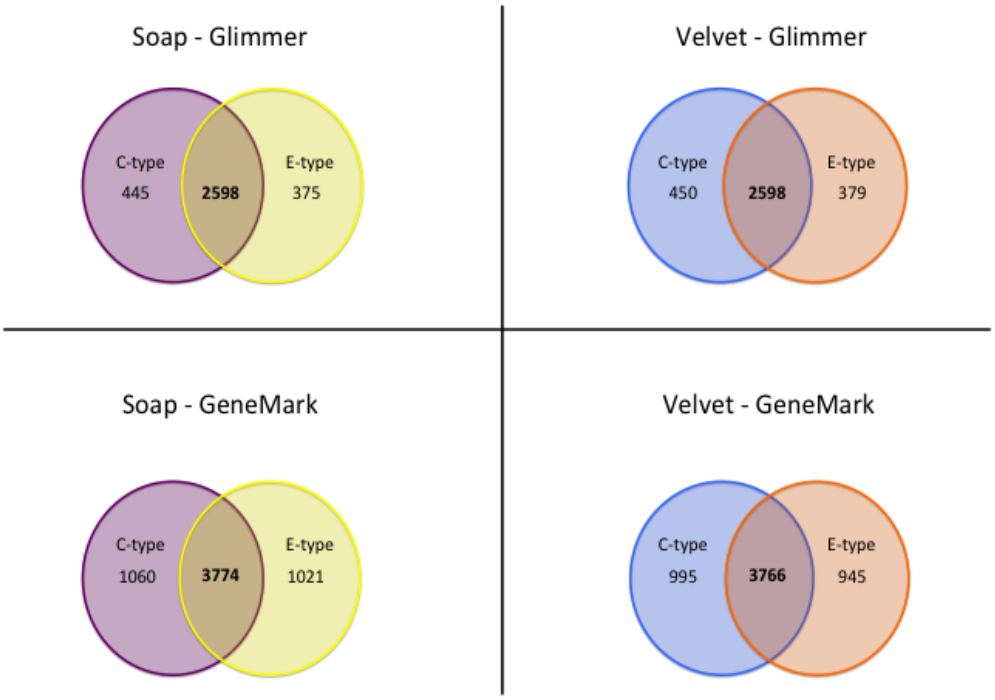


FIGURE 2.6: Genome content comparison for *Vibrio vulnificus* CMCP6 and CECT5198. The Venn diagrams represent the differences in differential gene counts identified when using the Velvet and SoapDenovo (Soap) assembly, each with the Glimmer and GeneMark annotation. *V. vulnificus* CMCP6 is classified as a C-genotype and *V. vulnificus* CECT5198 is classified as an E-genotype.

TABLE 2.1: Assembly and annotation statistics for *V. vulnificus* CMCP6.

Assembly method	**Velvet**	**ABySS**	**Soap**
# of contigs	205	144	269
Assembly + RAST performance			
# of genes predicted	4684	5095	4720
# of genes with match in CMCP6	3890	3777	3863
% of known genes recovered	91.8%	89.2%	91.2%
Assembly + GeneMark performance			
# of genes predicted	4761	5051	4833
# of genes with match in CMCP6	4019	3754	3844
% of known genes recovered	94.9%	88.6%	90.7%

TABLE 2.2: Genomic characteristics of *V. vulnificus* CMCP6, CIP8190, CECT5198, CECT4606, CECT5763, and CECT4866.

Genomic Characteristic	**CMCP6**	**CIP8190**	**CECT5198**	**CECT4606**	**CECT5763**	**CECT4866**
Biotype	1	2	2	2	2	2
Genotype	C	C	E	E	E	C
Chr Number	2	2	2	2	2	2
Plasmid Number	None	2	3	2	2	2
Average G+C content	46.6 %	46.5%	46.5%	46.2%	46.3%	46.5%
# of reads generated	6620286*	26869740	14366914	23523786	18852452	33792718
N50 for Velvet	196375	71778	60906	316446	51991	65142
N50 for ABySS	187671	57867	66098	154882	54273	64876
N50 for Soap	196396	71391	62139	165040	52087	65144

TABLE 2.3: Total number of contigs for *V. vulnificus* CMCP6, CIP8190, CECT5198, CECT4606, CECT5763, and CECT4866 using Velvet, SoapDenovo, and ABySS assemblers

Strain	Velvet	Abyss	Soap
***V. vulnificus* CMCP6**	205	144	269
***V. vulnificus* CIP8190**	284	364	507
***V. vulnificus* CECT5198**	302	289	448
***V. vulnificus* CECT4606**	129	148	267
***V. vulnificus* CECT5763**	492	743	845
***V. vulnificus* CECT4866**	404	366	519

TABLE 2.4: A.) Glimmer, B.) RAST and C.) GeneMark gene prediction counts for *V. vulnificus* strains included in this study.

A.)

Glimmer	Abyss	Soap	Velvet
***V. vulnificus* CMCP6**	3226	3042	3047
***V. vulnificus* CIP8190**	3233	3030	3032
***V. vulnificus* CECT5198**	3289	2973	2977
***V. vulnificus* CECT4606**	3465	3275	3275
***V. vulnificus* CECT5763**	3253	3079	3083
***V. vulnificus* CECT4866**	3301	3024	3031

B.)

RAST	Abyss	Soap	Velvet
***V. vulnificus* CMCP6**	5095	4720	4684
***V. vulnificus* CIP8190**	4963	4600	4623
***V. vulnificus* CECT5198**	5021	4554	4563
***V. vulnificus* CECT4606**	5315	5015	5011
***V. vulnificus* CECT5763**	5038	4732	4752
***V. vulnificus* CECT4866**	5035	4605	4631

C.)

GeneMark	Abyss	Soap	Velvet
***V. vulnificus* CMCP6**	5051	4833	4761
***V. vulnificus* CIP8190**	5084	4912	4787
***V. vulnificus* CECT5198**	5187	4795	4710
***V. vulnificus* CECT4606**	5500	5311	5189
***V. vulnificus* CECT5763**	5489	5346	5062
***V. vulnificus* CECT4866**	5243	4931	4839

TABLE 2.5: Workflow descriptions applied to *V. vulnificus* CECT4866 for differential functional analysis.

Workflow Assignment	Assembly Type	Annotation Method	Number of genes
A	Velvet	Glimmer	3031
B	Velvet	GeneMark	4839
C	Abyss	RAST	5035
D	Abyss	Glimmer	3301
E	Soap	GeneMark	4931
F	Soap	RAST	4605

TABLE 2.6: Summarizes the differential GO enrichment terms for the workflow descriptions listed in Table 5. If there were more than 20 genes that were above the significant p-value cut-off of .005, only the top 20 genes were shown for each differential category.

Category	GO Identifier	GO Name	P-value
Velvet - GeneMark Not Glimmer	GO:0005198	structural molecule activity	5.72E-29
Velvet - GeneMark Not Glimmer	GO:0043226	organelle	1.42E-26
Velvet - GeneMark Not Glimmer	GO:0030529	ribonucleoprotein complex	1.33E-23
Velvet - GeneMark Not Glimmer	GO:0003676	nucleic acid binding	2.79E-22
Velvet - GeneMark Not Glimmer	GO:0032991	macromolecular complex	1.83E-20
Velvet - GeneMark Not Glimmer	GO:0043229	intracellular organelle	1.77E-17
Velvet - GeneMark Not Glimmer	GO:0043170	macromolecule metabolic process	3.12E-17
Velvet - GeneMark Not Glimmer	GO:0044260	cellular macromolecule metabolic process	1.88E-16
Velvet - GeneMark Not Glimmer	GO:0044267	cellular protein metabolic process	5.32E-16
Velvet - GeneMark Not Glimmer	GO:0044444	cytoplasmic part	9.10E-16
Velvet - GeneMark Not Glimmer	GO:0006412	translation	1.93E-15
Velvet - GeneMark Not Glimmer	GO:0016410	N-acyltransferase activity	9.63E-15
Velvet - GeneMark Not Glimmer	GO:0019538	protein metabolic process	7.22E-14
Velvet - GeneMark Not Glimmer	GO:0005623	cell	7.45E-13
Velvet - GeneMark Not Glimmer	GO:0044464	cell part	7.45E-13
Velvet - GeneMark Not Glimmer	GO:0005840	ribosome	3.76E-12
Velvet - GeneMark Not Glimmer	GO:0006950	response to stress	5.99E-12
Velvet - GeneMark Not Glimmer	GO:0019843	rRNA binding	1.03E-11
Velvet - GeneMark Not Glimmer	GO:0016407	acetyltransferase activity	2.98E-11
Velvet - GeneMark Not Glimmer	GO:0048519	negative regulation of biological process	8.64E-09
Velvet - Glimmer Not GeneMark	GO:0019692	deoxyribose phosphate metabolic process	0.006635071
Velvet - Glimmer Not GeneMark	GO:0046386	deoxyribose phosphate catabolic process	0.00729927
Soap - RAST Not GeneMark	GO:0005198	structural molecule activity	2.58E-06
Soap - RAST Not GeneMark	GO:0032991	macromolecular complex	1.22E-05
Soap - RAST Not GeneMark	GO:0044267	cellular protein metabolic process	1.77E-05
Soap - RAST Not GeneMark	GO:0019538	protein metabolic process	5.65E-05
Soap - RAST Not GeneMark	GO:0030529	ribonucleoprotein complex	6.35E-05
Soap - RAST Not GeneMark	GO:0043226	organelle	1.01E-04
Soap - RAST Not GeneMark	GO:0006412	translation	1.35E-04
Soap - RAST Not GeneMark	GO:0044444	cytoplasmic part	3.70E-04
Soap - RAST Not GeneMark	GO:0043229	intracellular organelle	0.001603198
Soap - GeneMark Not RAST	GO:0004803	transposase activity	1.21E-07
Soap - GeneMark Not RAST	GO:0032196	transposition	5.88E-07
Soap - GeneMark Not RAST	GO:0003676	nucleic acid binding	4.62E-06
Soap - GeneMark Not RAST	GO:0032991	macromolecular complex	7.65E-05
Soap - GeneMark Not RAST	GO:0005198	structural molecule activity	9.91E-04
Soap - GeneMark Not RAST	GO:0006313	transposition, DNA-mediated	0.001075903
Soap - GeneMark Not RAST	GO:0006259	DNA metabolic process	0.001745285
Soap - GeneMark Not RAST	GO:0009987	cellular process	0.002874757
Soap - GeneMark Not RAST	GO:0019213	deacetylase activity	0.003067666
Soap - GeneMark Not RAST	GO:0051704	multi-organism process	0.004927739
ABySS - RAST Not Glimmer	GO:0005198	structural molecule activity	2.62E-27

TABLE 2.6: (Continued).

ABySS - RAST Not Glimmer	GO:0030529	ribonucleoprotein complex	5.00E-26
ABySS - RAST Not Glimmer	GO:0043226	organelle	8.03E-26
ABySS - RAST Not Glimmer	GO:0032991	macromolecular complex	1.67E-18
ABySS - RAST Not Glimmer	GO:0003676	nucleic acid binding	3.97E-17
ABySS - RAST Not Glimmer	GO:0043170	macromolecule metabolic process	5.26E-17
ABySS - RAST Not Glimmer	GO:0044444	cytoplasmic part	1.38E-16
ABySS - RAST Not Glimmer	GO:0044267	cellular protein metabolic process	1.48E-16
ABySS - RAST Not Glimmer	GO:0044260	cellular macromolecule metabolic process	1.01E-15
ABySS - RAST Not Glimmer	GO:0005840	ribosome	1.10E-14
ABySS - RAST Not Glimmer	GO:0006412	translation	1.25E-14
ABySS - RAST Not Glimmer	GO:0016410	N-acyltransferase activity	2.80E-14
ABySS - RAST Not Glimmer	GO:0019538	protein metabolic process	5.71E-13
ABySS - RAST Not Glimmer	GO:0005623	cell	1.62E-12
ABySS - RAST Not Glimmer	GO:0044464	cell part	1.62E-12
ABySS - RAST Not Glimmer	GO:0019843	rRNA binding	2.57E-12
ABySS - RAST Not Glimmer	GO:0006950	response to stress	5.03E-12
ABySS - RAST Not Glimmer	GO:0016407	acetyltransferase activity	1.22E-11
ABySS - RAST Not Glimmer	GO:0003735	structural constituent of ribosome	1.98E-09
ABySS - Glimmer Not RAST	GO:0019692	deoxyribose phosphate metabolic process	5.80E-05
ABySS - Glimmer Not RAST	GO:0046386	deoxyribose phosphate catabolic process	1.40E-04
ABySS - Glimmer Not RAST	GO:0009262	deoxyribonucleotide metabolic process	2.65E-04
ABySS - Glimmer Not RAST	GO:0009264	deoxyribonucleotide catabolic process	7.12E-04
ABySS - Glimmer Not RAST	GO:0005515	protein binding	0.002764846

CHAPTER 3: COMPARATIVE GENOMIC ANALYSIS OF VIBRIO VULNIFICUS BIOTYPES 1, 2 AND 3

3.1 Introduction

The species *Vibrio vulnificus* comprises three known biotypes each of which is capable of causing life-threatening infections in humans and in aquatic species. Biotype 1 is commonly associated with human infection, primarily through the consumption of raw or undercooked mollusks containing *V. vulnificus,* or by entry through an open wound[3]. In Chapter 1, we examined the differences between clinical (C) and environmental (E) genotypes of Biotype 1. Biotype 2 is a pathogen of eels and other marine species, and rarely causes human infection[3]. To date, biotype 3 has only been observed as the causative agent in an outbreak among fish market workers in Israel[72]. The *V. vulnificus* biotypes are currently distinguished based on their biochemical, serological, and molecular characteristics [73-75]. While these characteristics can give some insight into phenotypic variation between the biotypes, they do not provide detailed information about the genetic and functional differences. To date, there has not been an in-depth comparative genomics study incorporating sequences from all three biotypes of *Vibrio vulnificus*. Identifying differences in genomic content among the biotypes will lead to greater insight into their mechanisms of pathogenesis and survival in the environment.

In this chapter, we describe the sequencing, assembly, and comparative analysis of 10 *V. vulnificus* strains, with the intent of establishing the core virulence and survival mechanisms shared between the biotypes, as well as identifying the virulence and survival mechanisms that are specific to each biotype. These strains are representative of a larger data set that includes 25 biotypes, including additional Biotype 1 and Biotype 2 strains. The analysis approach, and results, described here preview key elements of the approach and findings from the comprehensive study of 25 strains.

3.2 Background

In previous comparative studies of *V. vulnificus*, biochemical markers along with sequences of selected genomic regions have commonly been used to distinguish between the biotypes. There are 13 biochemical characteristics used to differentiate between the biotypes. Of these tests, the indole production reaction is commonly reported in the identification of *V. vulnificus* biotypes in bench-work settings. The indole test is a biochemical test performed on bacterial species to determine the ability of the organism to convert tryptophan into indole. It is commonly reported that Biotype 1 isolates have positive indole reactions, while Biotype 2 have negative reactions [74-76]. The Biotype 3 indole test is positive, but in conjunction with other biochemical properties it can generally distinguish Biotype 3 from the other two biotypes [73]. However, the classification of the biotypes by these methods is not entirely clear cut. Biosca *et al*. [77] reported on a *V. vulnificus* isolate that is classified as Biotype 2 and virulent to eels, but has a positive indole reaction. This suggests a

diversity of biochemistry that may not be correctly represented by the small biomarker set currently used for *V. vulnificus* classification.

When DNA sequence is used as a marker for biotype in *V. vulnificus*, Biotype 1 isolates are associated with the presence of the genomic XII region. This region was identified by Cohen *et al*. [39], and some genes within this region are suggested to play a role in Biotype 1 pathogenicity. In comparison, the majority of Biotype 2 isolates contain at least two plasmids, a virulence plasmid and a putative conjugative plasmid[78]. To date, this extra-chromosomal content has not been reported in any Biotype 1 isolates and its presence can be used as a diagnostic for Biotype 2.

Until recently, only the molecular characteristics of the Biotypes could be probed, and only Biotype 1 genomes had been completely sequenced. In 2010, short read data was made available for a single Biotype 2 genome, ATCC 33149[10] and in 2013, the sequencing of one Biotype 3 environmental genome *Vibrio vulnificus* VVyb1, was reported[79]. The Danin-Poleg *et al.* [79] study identified 217 unique protein-coding sequences that were not in any of the known *V. vulnificus* genomes. The earlier study that produced Biotype 2 data employed the ABI SoLID next generation sequencing platform, which produces very short read fragments, thus making it impossible to assemble *ab initio*. Since the Biotype 2 genome in that study was assembled with a Biotype 1 genome as its reference, it was not possible to identify Biotype 2 specific regions. In the more recent Danin- Poleg *et al*. [79] study, the Biotype 3 genome was sequenced using the Illumina next generation sequencing platform. The draft genome assembly comprises 140 contigs. While the Illumina sequencing platform produces sufficient read lengths to assemble the reads *ab initio* and identify Biotype 3 specific

regions there still remains a level of uncertainty around genome content, due to its lack of closure. As we observed in Ch. 2, sequenced data for bacterial genomes obtained using the Illumina platform with a single insert size fails to completely capture gene content, potentially missing about 1% of genes even when paired reads in excess of 100x coverage are produced. However, it is likely that the identification of over 200 Biotype 3-unique genes in the strain examined in that study is reasonably accurate.

By sequencing and comparing a wider variety of Biotype 1, 2, and 3 strains that are differ in their isolation method and genotype can we begin to understand more about the differences between the Biotypes than has previously been possible. Using methods similar to those described in Chapter 1 we will be able to observe the differences in gene content in strain to strain comparisons within the same biotype, biotype to biotype comparisons, and isolation to isolation comparisons. This study will provide us with a better understanding of the virulence and survival mechanisms that are shared between the biotypes as well as those that are biotype-specific. Comparison of large number of strains may also call into question the relevance of the current biotype designations. The key to understanding these classifications will be observing differences between the core genomes of the biotypes. If we observe consistent and specific genome content for each biotype, regardless of strain-to-strain differences within biotype, then it is more likely that the traditional biotype designations are representing the underlying evolutionary history of *V. vulnificus*.

3.3 Materials and Methods

3.3.1 Growth Conditions and DNA Isolation

V. vulnificus strains were grown in the laboratory of Dr. Craig Baker-Austin at the Center for Environment, Fisheries, and Aquatic Science in Weymouth, UK. Cells were grown at 28°C for 24 hours in tryptone soy broth or on solid agar media, supplemented with 5 g/liter NaCl. Strains were cryogenically stored at -80°C prior to use, supplemented with 20% (vol/vol) glycerol. Late-logarithmically grown bacterial suspensions were pelleted and DNA extracted using a Mini-prep protocol. The quality and quantity of DNA was subsequently ascertained spectrophotometrically using NanoDrop ND1000 (NanoDrop Technologies, Wilmington, DE). Extracted DNA was run on 2% agarose gels to further check the quality and quantity of extracted DNA. Table 3.1 summarizes the provenance information for each of the 25 newly sequenced *V. vulnificus* strains discussed in this chapter, as well as the *V.vulnificus* reference genomes.

3.3.2 Genome Sequencing and Assembly

V. vulnificus strains were sequenced at The Genome Analysis Centre (TGAC, Norwich, UK) using the Illumina HiSeq2000 platform. Sequencing was carried out on libraries in pools of 12 strains in one lane of an Illumina plate. Sequence reads are 100 bp (PE) reads. All analysis subsequent to sequencing was then performed at UNC Charlotte.

FastQC was used to evaluate the quality of the sequence reads for each library[62]. We removed sequenced reads that contained any ‘N’ calls as a data-cleansing step. If a sequence read contained 1 or more ‘N’, both the read and its pair would be removed. After the data-cleansing step we sampled a subset of reads that was equivalent to 100x coverage based on the Lander and Waterman statistic[63]; for

a few data sets, lower coverage was used due to the amount of sequence remaining after 'N' removal. *Vibrio vulnificus* ORL 1506 did not go through the random read sampling protocol see Appendix B.

We used Velvet version 1.0.12 [49] and VelvetOptimiser version 2.2.0 to assemble reads into contigs. This decision was guided by the simulation results obtained in Chapter 2. The VelvetOptimiser parameters were used to initiate the Velvet assembler. The VelvetOptimiser hash value ranged from 73 to 93. The read description parameter was set to "-shortPaired". Table 3.2 summarizes the sequence assembly statistics for each of the newly sequenced *V. vulnificus* strains included in this study.

3.3.3 Genome and Gene Characterization

Feature identification for each strain was performed on the contig set for each isolate using an in-house pipeline of published microbial annotation tools constructed using the Taverna workflow management system[64] . The feature identification methods that were used were Glimmer3.02[13] (Glimmer) and GeneMark.hmm[14] (GeneMark). Both packages are widely used *ab initio* gene finding applications recognized and accepted by NCBI, and both are publicly available. Glimmer was used with default parameters. An exception was that the circular chromosomes were treated as linear in the analysis. This setting was used to prevent each contig from being treated as an individual circular chromosome. Both Glimmer and GeneMark.hmm were trained using sequence from known *Vibrio* genomes. The training sets chosen for each method are not identical, and training set choices were constrained by the capabilities and needs of the software. The Glimmer training set

was constructed with all completely sequenced *Vibrio spp*. genomes available as of June 2013. Table 3.3 contains a list of all the completely sequenced genomes used as training data in this study. GeneMark was used with its default parameters, and the model was trained on chromosome 1 of the *V. vulnificus* CMCP6 genome. To ensure that we could compare these newly sequenced genomes to the previous ones constructed in Chapter 1, and to the completely closed *Vibrio vulnificus* Biotype 1 genomes we followed the same annotation merging procedure described in Chapter 1. After the gene sequences were identified, we used Blast2Go to annotate the gene sequences [80]. Blast2Go performs a BlastP protocol against the non-redundant database at NCBI. Based on the results of the BlastP search, Gene Ontology (GO) terms are associated to each gene were associated on the results. The detailed description of the Blast2Go annotation methodology is in Conesa *et al*. [80].

3.3.4 Gene Clustering

OrthoMCL version 2.0[15] was used to cluster the genes predicted in the newly sequenced genomes with genes from other completely characterized *Vibrio* spp. The purpose of the ortholog clustering procedure is to establish relationships among genes from genome to genome. For the purpose of content comparison, a cluster of genes that contains one ortholog from each genome in a study is considered to represent the same gene when content is being compared. The implied relationship between the genes is that they have an evolutionary common ancestor. OrthoMCL has been shown to outperform other stand-alone methods for ortholog clustering [15]. OrthoMCL uses an all-against-all BlastP comparison of sequences as an input step, followed by application of a Markov clustering procedure. The e-value cutoff for the

BlastP algorithm was 1e-5. Default parameters were used for OrthoMCL, except that clusters were formed based on a shared sequence similarity of 70%. Details for the parameter cut-off are explained in 1.2.4.

3.3.5 Gene Content Comparison

The OrthoMCL [15]clustering generated during the annotation step was used as the basis for identification of differentiating genes. Identified gene features and OrthoMCL results were stored in a locally developed OLAP data warehouse (GenoSets) that supports queries across aggregate data generated by a variety of genomic annotation and comparison methods as described in Cain *et al*.[28]. Annotations for the novel genomes reported were generated as described in section 1.3.4. Feature boundaries were determined from the annotation and stored in a GenoSets database. The software facilitates gene presence-absence queries to be formulated in GenoSets at different levels, and was used to investigate differences among Biotype groups as well as individual strains.

In order to provide a standard means of comparison for feature attributes we established relationships between features using two methods. First, we estimated orthologous relationships between genes using OrthoMCL as described in the previous section. For functional analysis, gene features identified in the newly sequenced *V. vulnificus* strains used the GO terms assigned by Blast2Go. For functional comparison purposes, we used a controlled vocabulary to identify genes and other features. The Gene Ontology (GO) [19,20] provides standardized terms for the description of gene products in terms of biological processes, cellular location, and molecular function.

3.3.6 Phylogenetic Analysis

There were 2651 single-copy ortholog clusters identified within the 25 newly sequenced *V. vulnificus* strains. Ortholog clusters were constructed based on the criteria defined in section 3.3.4. Table 3.1 contains a list of the 25 newly sequenced *V. vulnificus* strains included in the phylogenetic analysis. The strains included in this analysis represent all the known biotypes and subtypes found within *V. vulnificus*. We performed a phylogenetic analysis following the methods used in [81, 82, and 2]. We randomly selected protein sequences of approximately 10% of the single-copy ortholog clusters identified (266 genes) and used the sample as a basis for construction of a maximum likelihood tree, following the approach used in Hasan *et al* 2010. MUSCLE version 3.8.31 was used to align sequence members of each ortholog cluster independently [83]. Individual alignments were used to minimize rearrangements within the multiple sequence alignment [2]. Once each individual protein alignment was built, the independent alignments were concatenated. phyML 3.0, a maximum likelihood method, was used to generate a phylogenetic species tree with 100 replicates for bootstrapping [84]. The tree was visualized with Figtree [85]. Three independent samplings were tested and all three produced trees with highly similar topologies.

3.4 Results and Discussion

3.4.1 Genome Sequencing and Assembly Statistics

An average of 785,799,011 paired end sequencing reads were generated for the newly sequenced *V. vulnificus* strains included in this chapter. Table 3.2 details the number of paired end reads generated for each of the strains. The read length was 100

bp. The number of contigs constructed ranged from 78 to 4658. The average N50 value was 19156.32. The raw genome data for each of the newly sequenced genomes were randomly sampled to 100x genome coverage, with the exception of those genomes that did not have enough sequence reads after the removal of 'N' characters in the sequence and *V. vulnificus* CDC 9030-95 (ORL 1506), see Appendix B. Magoc *et al.* [55] reported that 100x is sufficient genome coverage to construct a microbial genome from paired end Illumina sequence reads. In Chapter 2, we tested multiple assembly methods on read sets at different coverage depth and found that using additional reads beyond 100x coverage did not appreciably improve assemblies, while adding significantly to the computational time required to assemble the genomes.

3.4.2 General Properties of the Biotype 1, 2 and 3 Genomes

All of the Biotype 1, 2 and 3 genomes are composed of 2 circular chromosomes containing an estimated total of approximately 5.6 MB – 5.8 MB of genomic DNA. The Biotypes however do differ in the extra-chromosomal DNA. Of the newly sequenced Biotype 1 genomes only 1 has plasmid DNA present, *Vibrio vulnificus* CECT 4606. All of the newly sequenced Biotype 2 and 3 genomes have 1 or 2 plasmids. Table 3.1 summarizes the predicted gene content of each sequenced draft genomes.

3.4.3 Gene Content that Characterizes *V. vulnificus* Biotypes

We identified approximately 3690 common genes between the *V. vulnificus* biotypes. The differential gene counts are based on using by *V. vulnificus* strains CMCP6 (BT1), CECT4606 (BT2), 11028 (BT3), and 2(BT3) to represent biotypes 1, 2 and 3, respectively. *V. vulnificus* CMCP6 was used to represent the Biotype 1

genomes due to its recent re-annotation [23]. We identified 384 genes that are in *V. vulnificus* CMCP6 isolates that are not in any of the Biotype 2 or Biotype 3 isolates. There are 628 genes in the representative Biotype 2 isolate that are not present in any of the Biotype 1 or Biotype 3 isolates. Biotype 3 has 420 genes that are present in both representative Biotype 3 isolates, but not present in the Biotype 1or Biotype 2 genomes. Figure 3.1 summarizes the gene count differentials among the biotypes. There are many possible queries that can be constructed within this data set, each of which will produce a lengthy list of differentiating genes. Here, we focus on several of the most significant comparisons. 3.4.4 General characteristics of Biotype 1 strains isolated from clinical sources

All of the known *V. vulnificus* biotypes have members that have been isolated from human infections (clinical isolates). This distinction is independent of the "C-type" and "E-type" classifications discussed in Chapter 1; Biotype 2 and 3 clinical isolates identified thus far are nonetheless characterized as E genotypes by molecular methods. In this differential analysis, we examine the differences among *clinical isolate* strains of each of the three biotypes. The result set represents those genes, which are common among those strains that cause human infection, and identifies different functional capabilities in clinically significant strains of each of the three biotypes. Here, we discuss several typical examples of genes that only appear in the Biotype 1 clinical strains. *V. vulnificus* strains CMCP6, MO6-24/O, YJ016, ATL-9824, NSV 5830, ORL 1506, and C718AV represented the biotype 1 strains, *V. vulnificus* strains 11028 and 12 represented the biotype 3 strains, and *V. vulnificus* strains CIP8190, CECT4866, and 94-8-112 represented biotype 2 strains. The results

in Table 3.4 represent the differences in gene content between *V. vulnificus* CMCP6, YJ016, and MO6-24/O. These strains are completely characterized and their functional annotations were transferred to the annotations of the newly sequenced biotype 1 clinically isolated strains. Of the genes that were identified as only appearing in biotype 1 strains relative to clinical isolates of biotypes 2 and 3, we confirm six functional systems that were also identified as characteristic of Biotype 1 clinical isolates in our previous differential analysis, Morrison *et al*.[2].

3.4.4 General Characteristics of Biotype 1 Strains Isolated From Clinical Sources

In Table 3.4, we highlight Biotype 1 genes from clinically isolated C-type strains. The mannitol-1-phosphate 5-dehydrogenase and mannitol operon repressor genes were identified in our previous study[25] as a common feature of clinical Biotype 1 strains, and they are suggested as being associated with clinical genotype virulence[25]. We also previously identified Biotype 1 clinical isolates possessing unique GGDEF family proteins (GGDEF family protein YeaJ) located in an operon with a putative two-component response regulator and a fimbrial protein Z transcriptional regulator. These genes are unique to the Biotype 1 clinical strains relative to the other two Biotypes as well [2].

In addition to the previously identified mannitol-associated genes and the fimbrial protein Z operon [2], we find other groups of genes, which characterize Biotype 1 relative to Biotypes 2 and 3. One such group is methyl-accepting chemotaxis proteins. Chemotaxis is the process by which the movement of cells is directed by chemicals in the environment. Since *V. vulnificus* is commonly found in estuarine environments, the assumption would be that all *V. vulnificus* have some sort

of drive to direct movement toward nutrient rich environments. The additional chemotaxis genes characterizing Biotype 1 may result in increased responsiveness to environmental stimuli relative to the other two biotypes. Biotype 1 strains are the most frequently reported sources of human infection. Experiments targeting these genes would need to conducted, to determine whether enhanced chemotactic capabilities are perhaps a phenotypic advantage for Biotype 1 clinical isolates, or play a role in their recognition of human tissue as a nutrient rich environment.

Another set of genes, which seem to be present in Biotype 1 and absent in other biotypes is a portion of an arsenic resistance operon. Arsenic is a contaminant of water supplies around the world and one of the most toxic inorganic ions [86]. It is likely that *V. vulnificus* encounters this toxin in some environments and that arsenic resistance may provide a survival advantage for some strains. It has been reported in the literature that arsenic levels increase in fresh and marine water when crude oil is present [87]. The oil interrupts the natural filtration process of sediments bonding with arsenic, which results in increased levels of arsenic in the body of water [87,88]. Tao *et al*. [89] reported on the prevalence of *Vibrio vulnificus* cells surviving in tar balls collected as a result of the 2010 BP Deepwater Horizon oil spill. Their results showed total aerobic bacterial counts were ($> 10^6$ CFU/g) in tar balls collected from Alabama and Mississippi (USA)[89]. While they did report that *V. vulnificus* did not grow when exposed to tar-ball enriched seawater agar [89], it is plausible that this operon is contributing to the survival of *V. vulnificus* within the tar balls. Again, experiments targeting the arsenic resistance operon would need to be conducted to clarify the role of arsenic-resistance in survival mechanisms of *V. vulnificus*

3.4.5 Biotype Differentials Among E-genotype Strains

Vibrio vulnificus JY1305 is traditionally classified as a biotype 1 strain, isolated from an environmental source and therefore designated as an E-genotype (B1E). As stated in the previous section, all Biotype 2 and 3 strains are classified as E-genotypes, regardless of their source of isolation. Our goal in this comparison was to contrast E-genotype strains from each Biotype classification. The genes identified in this comparison represent a common core of E-genotype genes across the biotypes, and the differential genes should be characteristic of biotype. Strain JY1305 was used to represent biotype 1, E-genotype strains in a comparative genome query. The strains included in this comparison are *V. vulnificus* JY1305, representing biotype 1 E-type strains, *V. vulnificus* CECT4606 for biotype 2 E-type strains, and *V. vulnificus* 11028 for biotype 3 E-type strains.

Of the 328 specific genes identified as specific to Biotype 1 in this analysis, we highlight a gene associated with iron utilization. Iron utilization is especially of interest in the context of *Vibrio vulnificus* pathobiology, since this pathogen is dangerous to hosts in a condition of iron overload [68, 69]. Several studies have reported on the correlation between *V. vulnificus* infections and increased levels of iron in animal models and infected individuals [3, 68, and 69]. However there have been no clinical cases of infection by JY1305. Differences in expression of iron utilization genes were observed in an RNA-Seq study of Biotype 1 C and E strains, in which expression under human serum and artificial seawater conditions was compared (E. Blackman and T. Williams, personal communication). Differences in iron utilization might also potentially support the prevalence of biotype 1 strains in human

infection over biotype 2 and 3[4]. Iron utilization gene differentials were not identified in the comparison of the clinical isolates described in the previous section and demonstrates the power of changing the data aggregation strategy when carrying out comparisons of a large number of bacterial strains of varied origin and type.

Vibrio vulnificus CECT4606 is traditionally classified as a biotype 2 strain, isolated from an environmental source and therefore designated as an E-genotype (B2E). This strain was used to represent biotype 2, E-genotype strains in the comparison among E-genotypes of different biotype. Of the 648 B2E specific genes identified, several were characterized as being associated with bacterial pathogenesis. Table 3.5 lists a selection of potentially significant differential genes for *V.vulnificus* CECT4606. There is no evidence that supports the ability of *V. vulnificus* CECT4606 to cause infection within either human or animal hosts. Nevertheless, several secretion proteins of the type II (T2SS), type IV (T4SS), and type VI (T6SS) secretion systems were identified. These systems (T4SS) have traditional been associated with *V. vulnificus* Biotype 1 strains (CMCP6 and YJ016) [2], which are known to cause infections within humans. If their presence is shown to be consistent across the remaining Biotype 2 strains, it could provide an explanation of the occasional isolation of Biotype 2, E-genotype strains from clinical samples. The presence or absence of secretion system genes alone does not fully explain the capability of the strain to act as a pathogen, however. There is evidence that not all Biotype 1 strain C-type strains (MO6-24/O) that are capable of causing infection indeed have all components of the T4SS system [2]. Uptake of pathogen-associated genes of this type is thought to be the result of horizontal gene transfer.

Another biotype 2 specific gene suggests horizontal gene transfer activity is the zonula occludens toxin (*ZOT*). *ZOT* has been documented in *Vibrio cholerae* infections as being associated with the symptom of diarrhea [90]. In 1970, in a clinical case of *V. vulnificus*, it was reported that the patient had symptoms of diarrhea, vomiting, and hemorrhagic rash; all of which are common symptoms of *V. cholerae* infections [91-93]. These symptoms suggest the presence of the cholera-associated ZOT, although molecular confirmation of the presence of ZOT in similar cases of infection would be necessary to support that hypothesis.

A third interesting finding was that several of the mannitol-associated genes identified in Chapter 1 as characteristic of C-type strains were identified as gene differentials for this B2E strain. That the two groups share this common functionality may suggest a closer or more ambiguous relationship between the Biotype 2 strains and Biotype 1 C-genotype strains than previously suspected. The presence of mannitol-associated genes in the C-genotype strains is commonly associated with virulence capabilities [25] and sees section 1.4.2. In the Oliver laboratory, 40% of the E-type strains tested have been found to contain the mannitol operon and are able to ferment this sugar [37, 38]. CECT4606 is one of those E-type strains included in this category.

While further comparative analysis will be necessary to sort out the complex relationships within and between the groups, these findings suggest that a single gene or even a group cannot be used to differentiate between the biotypes with certainty and that a more detailed classification system may be needed.

Vibrio vulnificus 11028 is classified as a biotype 3 strain, isolated from a clinical source with an E-genotype classification (B3E). This strain was used to represent biotype 3, E-genotype strains in the comparative analysis of E-genotypes. *V. vulnificus* 11028 was isolated from a human sample though classified as an E-genotype strain using molecular criteria. In chapter 1 it was established that E-type strains rarely cause human infections. The genes characteristic of this infectious Biotype 3 strain may therefore give some insight into the varying means by which *V. vulnificus* genomes acquire the ability to act as human pathogens.

Of the 476 B3E specific genes identified, genes associated with toxins RelE and RelB are highlighted. The *relBE* operon inhibits translation during nutritional stresses [94-96. In the Yamamoto *et al.* (2002) study [97], they presented a case when the *Escherichia coli relE* gene was expressed inducibly in a human osteosarcoma cell line and it caused growth inhibition and cell death by apoptosis. The functional implications of its presence in *V. vulnificus remain unclear*. It may play a role in this organism's ability to cause symptoms related to cell death in the infected host, causing conditions such as blistering dermatitis. It also possible that *relE* may play a role in BE3 cell deaths in nutrient limiting environments, which may possibly explain the limited number of cases of biotype 3 infection reported. Further investigation of the role of RelE and RelB in the biotype 3s will need to be done in order to elucidate their role. These toxin genes may serve as a differential characteristic of *V. vulnificus* biotype 3 strains from either a clinical or environmental source, and may be used in combination with other genes as a diagnostic marker for biotype 3 strains.

3.4.6 Phylogeny of V. vulnificus Biotypes and Genotypes

Figure 3.2 is a phylogeny of the *Vibrio vulnificus* strains listed in Table 3.1. The phylogeny includes all of the 25 recently sequenced strains of *V. vulnificus* along with previously sequenced strains. The overall consensus of the three trees is similar to the evolutionary relationships previously observed between the biotypes; with the biotype 3 strains placed between the biotype 1 and 2 strains [98-102] and also follows the pattern of divergence between C and E genotypes seen in Morrison *et al*. [2]. However, there are quite a few strains that do not follow previously published *V. vulnificus* biotype phylogenies. As previously noted, over 25% of *V. vulnificus* strains are atypical in their response to one or more molecular assays, which suggests that there may be more diversity among strains and biotypes than is adequately represented by the traditional molecular assays. The phylogenetic tree presented here is based on a significant number of conserved genes. It was constructed using an approach, which we have previously used to produce a phylogeny of the genus Vibrio; that phylogeny was congruent with accepted ideas of the phylogeny of that genus. It is likely that the phylogeny presented here represents the basic relationships among the strains accurately, although addition of an out-group species from within the genus would clarify the tree topology. The two biotype 3 strains are consistently placed between biotype 1 and 2 strains within the three sampling trees. There are three biotype 2 strains (CECT4606, CECT5769, and 95-8-162) that are closer in evolutionary relationship to the other two biotypes than to the remainder of the biotype 2 strains, which consistently form a cluster elsewhere in the tree. It is possible that the meta-data collected for these strains is incorrect, but it may also be that the division between Biotype 1 and Biotype 2 is not as unambiguous as previously thought. Molecular

assays need to be performed on the strains (CECT4606, CECT5769, and 95-8-162) to confirm their classification as Biotype 2.

3.5 Summary

Ten strains of various *Vibrio vulnificus* biotypes and genotypes have been sequenced, assembled and annotated into draft genomes. These draft genomes have provided the basis of a novel in-depth comparative genomics study of *V. vulnificus* biotypes and biotype-to-genotype combination. As a result of the differential analysis we have identified Biotype 1, 2, and 3 specific genomic regions. These insights can be used to establish an improved classification system for *V. vulnificus*. Approximately 25% percent of *V. vulnificus* strains are known to have some sort of molecular, functional, or biochemical discrepancy from the 'norm' associated with their currently assigned biological classification. We anticipate that the regions we have identified in this comparison may provide insights into the infection and survival mechanisms specific to each of the biotypes and genotypes. This work will facilitates further molecular investigation of gene and biochemical pathway targets that can be used to assess the relationship of genomic differences to function in a bench work setting.

3.6 Conclusion of Work

Overall, this work demonstrates the benefits of large-scale sequencing to develop differential datasets for microbes. It represents an important scientific step in a significant collaborative effort, in which methods and knowledge from multiple disciplines are used to solve a complex problem. The bioinformatics work presented here provides a foundation of tools and analysis techniques for future studies in

Vibrios and other bacteria. The biological outcome – analysis and comparison of *V. vulnificus* Biotypes – provides a wealth of targets for future investigation by collaborators and future students in the Gibas group.

With the capabilities of NGS technologies, sequencing of complete microbial genomes can be accomplished within a matter of days. It is foreseeable that in-depth comparative genomics studies of many closely related strains will soon be standard practice for researchers investigating the biology of microbes. The contributions of this work are two-fold. First of all, it contributes to specific understanding of the biology of *Vibrio vulnificus*, establishing the differences in gene content that define clinical and environmental genotypes of Biotype 1 strains, as shown in Chapter 1 and in Morrison *et al*. 2012.That study was the first to report on several E-type specific genomic regions. An important result was the observation that the SPANC theory may potentially be one of the driving forces between the diversification of the genotypes in *V. vulnificus*. Chapter 3 lays the groundwork for a similar manuscript defining the differences in genetic content between Biotypes 1, 2 and 3. Again, many of the differential results in this chapter will be novel findings for *V. vulnificus*, which have not been observed previously, or have been observed only in part. Identification of these gene differentials provides microbiologists with the molecular tools to investigate new aspects of the different survival mechanisms of the biotypes in different isolation sources. Also, these findings may bring the attention of microbiologists to focus on genes that previously were deemed insignificant in distinguishing characteristics between the biological classifications of *V. vulnificus*.

Secondly, this work contributes to an understanding of bioinformatics best practices for microbial genome assembly and annotation. The results in chapter 2 can be a point of reference in future sequencing projects, and provide a guide for other researchers on the importance of maintaining consistent analysis practices when identifying the similarities and differences in NGS datasets.

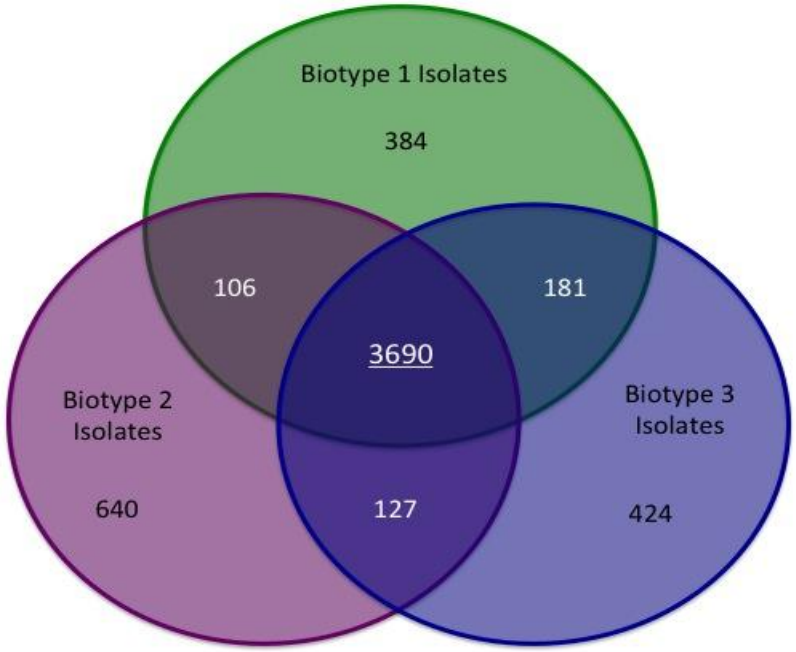


FIGURE 3.1: *Vibrio vulnificus* biotypes gene content differential Venn diagram.

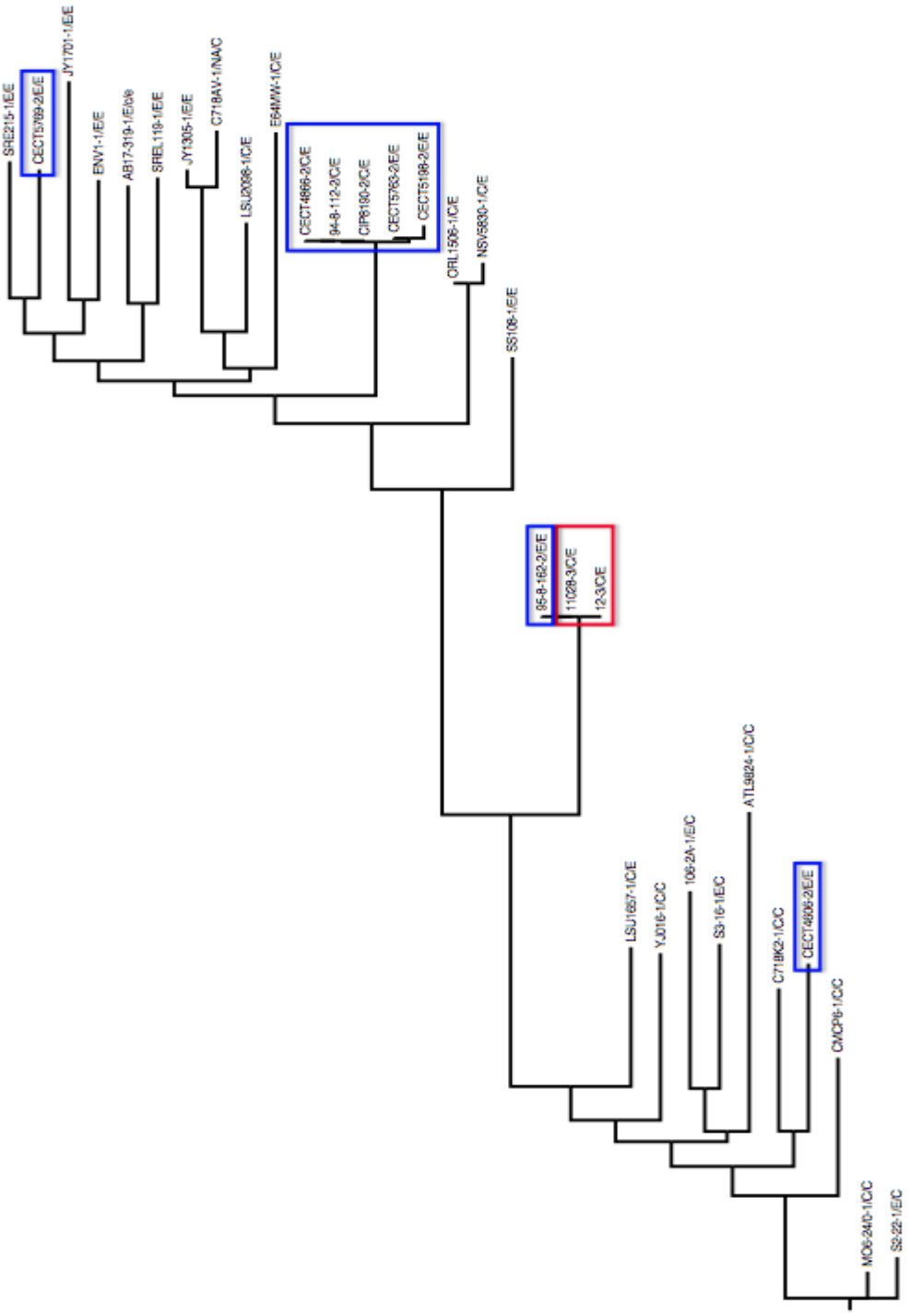


FIGURE 3.2: Phylogenetic relationships among newly sequenced *Vibrio vulnificus* genomes, completely characterized *V. vulnificus* genomes, and genomes described in Morrison *et al.* Blue boxes = biotype 2 strains and Red box = biotype 3 strains

TABLE 3.1: Genomic characteristics for 25 newly sequenced *V. vulnificus* strains included in this study, strains from Morrison *et al.* (2012), completely sequenced *V. vulnificus* genomes.

Strain	Biotype	Source (Clinical or Environmental)	# of Chromosomes	# of Plasmids	# of Genes
CECT 5198	2	E – Diseased eel	2	3	4306
CECT 4606	2	E – Diseased eel	2	1	3814
CECT 5769	2	E – Diseased eel	2	2	4445
CECT 5763	2	E – Eel tank water	2	2	4503
CIP 8190	2	C – Human blood	2	2	4379
CECT 4866	2	C – Human blood	2	2	4401
94-8-112	2	C – Human wound	2	1	4381
95-8-162	2	E – Diseased eel	2	3	4497
11028	3	C – Human sample	2	2	4491
2	3	C – Human disease	2	2	7022
SS108-A3A	1	E – Oyster	2	NA	4243
LSU 1657	1	E – Wound	2	NA	4319
SS109-3B2	1	E – Water	2	NA	4180
AB17-319	1	C/E – Oyster	2	NA	4403
SREL 214	1	E – Water	2	NA	4329
LSU 2098	1	E – Wound	2	NA	4221
C7184/AV	1	C – avirulent form of C7184	2	No information	6052
SREL 119	1	E – Sediment	2	NA	4331
C7184/K2	1	C – Blood	2	NA	4415
ATL-9824	1	C – NA	2	NA	4738
CDC 9348-95 (NSV 5830)	1	C –NA	2	No information	4255
CDC 9030-95 (ORL 1506)	1	C – NA	2	NA	4222
S2-22	1	E – NA	2	NA	4315
S3-16	1	E – NA	2	NA	4392
106-2A	1	E – NA	2	NA	4378
YJ016	1	C – Human Blood Chen et al 2003	2	1	5028
CMCP6	1	C - Kim et al 2011	2	NA	4433
MO6-24/O	1	C- Human sample Park et al 2011	2	NA	4562
JY1305	1	E – environmental Morrison et al 2012	2	NA	4235
E64MW	1	E – clinical Morrison et al 2012	2	NA	4301
JY1701	1	E – environmental Morrison et al 2012	2	NA	4425

TABLE 3.2: Summary of sequencing and assembly characteristics for 25 newly sequenced *V. vulnificus* strains.

Strain	# of reads	# of reads after sampling	# of contig	N50 Values
CECT 5198	14366914	11400000	302	60906
CECT 4606	23523786	11400000	129	316446
CECT 5769	27070214	11400000	211	211418
CECT 5763	18852452	11400000	492	51991
CIP 8190	26869740	11400000	284	71778
CECT 4866	33792718	11400000	404	65142
94-8-112	20119070	11400000	317	71397
95-8-162	22823524	11046690	252	52010
11028	18695952	9816264	263	51953
12	13897430	9319912	573	39774
SS108-A3A	27257942	11400000	288	146794
LSU 1657	45564622	11400000	153	212238
SS109-3B2	38172306	11400000	214	284089
AB17-319	78300154	11400000	184	250605
SREL 214	17603919	9857920	167	160449
LSU 2098	25350168	11400000	216	192046
C7184/AV	16171946	11400000	4658	2343
SREL 119	32445288	11400000	267	270843
C7184/K2	35552340	11400000	232	194696
ATL-9824	36824944	1473633	188	234976
CDC 9348-95 (NSV 5830)	39030306	11400000	111	504803
CDC 9030-95 (ORL 1506)	28284148	22519278	167	415694
S2-22	61463250	11400000	78	397969
S3-16	31474754	11400000	148	192080
106-2A	52291124	9951161	144	336218

Table 3.3 List of completely characterized *Vibrio* spp. reference genomes used in this study.

Vibrio anguillarum 775	*Vibrio cholerae* LMA 3984-4
Vibrio cholerae M66-2	*Vibrio cholerae* MJ-1236
Vibrio cholerae O1 biovar El Tor str. N16961	*Vibrio cholerae* O1 str. 2010EL-1786
Vibrio cholerae O395	*Vibrio furnissi* NCTC 11218
Vibrio harveyi ATCC BAA-1116	*Vibrio parahaemolyticus* RIMD 2210633
Vibrio sp. EJY3	*Vibrio sp.* Ex25
Vibrio splendidus LGP32	*Vibrio vulnificus* CMCP6
Vibrio vulnificus MO6-24/O	*Vibrio vulnificus* YJ016
Vibrio fischeri ES114	*Vibrio fischeri* MJ11
Vibrio parahaemolyticus BB220P	*Vibrio cholerae* IEC2244

TABLE 3.4: Key differential genes found in Biotype 1 clinically isolated genomes *V. vulnificus* CMCP6, Y016, and MO6-24/O that are NOT present in ANY Biotype 2 or Biotype 3 clinically isolated strains.

Strain	Chr	Locus Tag	Product Description	GO ID	GO Term
CMCP6	2	VV2_0893	arsenical-resistance protein	GO:0008508 GO:00162021	Bile acid sodium symporter activity Integral to membrane
	2	VV2_1508	two-component response regulator	GO:0000156 GO:0003700 GO:0043565	Phosphorelay response regulator activity Sequence-specific DNA binding transcription factor Sequence-specific DNA binding
	2	VV2_1509	two-component response regulator and GGDEF family protein YeaJ	GO:0016849 GO:0009190 GO:0035556	Phosphorus-oxygen lyase activity Cyclic nucleotide biosynthetic process Intrcellular signal transduction
	2	VV2_1510	response regulator	GO:0003700 GO:0043565 GO:0005622 GO:0006351	Sequence-specific DNA binding transcription factor activity Sequence-specific DNA binding Intracellular Transcription, DNA-dependent
	1	VV1_0640	mannitol repressor protein	NA	NA
	1	VV1_0639	mannitol-1-phosphate 5-dehydrogenase	GO:0008926 GO:0050662 GO:0019594	Mannitol-1-phosphate 5-dehydrogenase activity Coenzyme binding Mannitol metabolic process

TABLE 3.4: (Continued).

YJ016	**2**	**VVA1362**	**arsenite efflux pump ACR3**	**GO:0008508 GO:0016021**	**Bile acid: Sodium symporter activity Integral to membrane**
	1	VV0504	mannitol-1-phosphate 5-dehydrogenase	GO:0008926 GO:0050662 GO:0019594	Mannitol-1-phosphate 5- dehydrogenase activity Coenzyme binding Mannitol metabolic process
	1	VV0503	mannitol repressor protein	NA	NA
	2	VVA0325	fimbrial protein Z, transcriptional regulator	GO:0000156 GO:0003700 GO:0043565 GO:0005622 GO:0006351 GO:0035556	phosphorelay response regulator activity sequence-specific DNA binding transcription factor activity sequence-specific DNA binding intracellular transcription, DNA-dependent intracellular signal transduction
	2	VVA0326	GGDEF family protein	GO:0016849 GO:0009190 GO:0035556	Phosphorus-oxygen lyase activity Cyclic nucleotide biosynthetic process Intracellular signal transduction
	2	VVA0327	fimbrial protein Z, transcriptional regulator	GO:0000156 GO:0003700 GO:0043565 GO:0005622 GO:0006351 GO:0035556	phosphorelay response regulator activity sequence-specific DNA binding transcription factor activity sequence-specific DNA binding intracellular transcription, DNA-dependent intracellular signal transduction

TABLE 3.4: (Continued).

MO6-24/O	**2**	**VVMO6_04275**	**arsenical-resistance protein ACR3**	**GO:0008508 GO:0016021**	**Bile acid: Sodium symporter activity Integral to membrane**
	1	VVMO6_02634	mannitol-1-phosphate 5-dehydrogenase	GO:0008926 GO:0050662 GO:0019594	Mannitol-1-phosphate 5- dehydrogenase activity Coenzyme binding Mannitol metabolic process
	1	VVMO6_02635	mannitol operon repressor	NA	NA
	2	VVMO6_03281	two-component response regulator	GO:0000156 GO:0003700 GO:0043565	Phosphorelay response regulator activity Sequence-specific DNA binding transcription factor Sequence-specific DNA binding
	2	VVMO6_03282	two-component response regulator and GGDEF family protein YeaJ	GO:0016849 GO:0009190 GO:0035556	Phosphorus-oxygen lyase activity Cyclic nucleotide biosynthetic process Intrcellular signal transduction
	2	VVMO6_03283	two-component response regulator	GO:0000156 GO:0003700 GO:0043565	Phosphorelay response regulator activity Sequence-specific DNA binding transcription factor Sequence-specific DNA binding

TABLE 3.5: Key differential gene blast homolog results for *V. vulnificus* CECT 4606- biotype 2 and *V. vulnificus* 11028-biotype 3.

Strain	Biotype	Gene length	Blast homology result
***V. vulnificus* CECT4606**	2	1145	YP_002311894.1 Zonular occludens toxin (ZOT) *Shewanella piezotolerans* WP3
***V. vulnificus* CECT4606**	2	2399	ACV96422.1 Type-IV secretion protein TraC *Vibrio cholerae* Mex1
***V. vulnificus* CECT4606**	2	1403	WP_006071610.1 Type-IV secretion protein VirD2 *Vibrio shilonii*
***V. vulnificus* CECT4606**	2	1181	WP_000808629.1 Type II secretion system protein F *Vibrio mimicus*
***V. vulnificus* CECT4606**	2	890	WP_017790953.1 Type IV secretion protein Rhs *Vibrio vulnificus*
***V. vulnificus* CECT4606**	2	1700	WP_001929930.1 Type IV secretion protein Rhs *Vibrio cholerae*
***V. vulnificus* CECT4606**	2	461	WP_008153828.1 Type IV secretion protein Rhs *Pseudomonas sp.* GM41(2012)

TABLE 3.5: (Continued).

***V. vulnificus* CECT4606**	**2**	**1580**	**WP_000392189.1** **Type IV secretion protein Rhs** ***Vibrio mimicus***
***V.vulnificus* CECT4606**	2	1403	WP_000426088.1 Type VI secretion protein *Vibrio albensis*
***V. vulnificus* CECT4606**	2	680	WP_000647406.1 Type VI secretion protein *Vibrio mimicus*
***V. vulnificus* CECT4606**	2	482	WP_001045313.1 Type VI secretion protein *Vibrio cholerae*
***V. vulnificus* CECT4606**	2	506	WP_000031394.1 Type VI secretion protein *Vibrio mimicus*
***V. vulnificus* 11028**	3	329	WP_017788802.1 Addition module toxin RelE *Vibrio vulnificus*
***V. vulnificus* CECT4606**	2	530	YP_004189860.1 Mannitol operon repressor *Vibrio vulnificus* MO6-24/O
***V. vulnificus* CECT4606**	2	1148	NP_933297.1 Mannitol-1-phosphate 5-dehydrogenase *Vibrio vulnificus* YJ016
***V. vulnificus* CECT4606**	2	1952	NP_759624.2 PTS system mannitol-specific transport subunit IIC *Vibrio vulnificus* CMCP6

REFERENCES

1. Wei, L Liu Y, Dubchak I, Shon J, Park J (2002) Comparative genomics approaches to study organism similarities and differences. J Biomed Inform 35:142-150.

2. Morrison SS, Williams T, Cain A, Froelich B, Taylor C, et al. (2012) Pyrosequencing-based comparative genome analysis of Vibrio vulnificus environmental isolates. PLoS One 7: e37553.

3. Jones MK, Oliver JD (2009) Vibrio vulnificus: disease and pathogenesis. Infect Immun 77: 1723-1733.

4. Aznar R, Ludwig W, Amann RI, Schleifer KH (1994) Sequence determination of rRNA genes of pathogenic Vibrio species and whole-cell identification of Vibrio vulnificus with rRNA-targeted oligonucleotide probes. Int J Syst Bacteriol 44: 330-337.

5. Nilsson WB, Paranjype RN, DePaola A, Strom MS (2003) Sequence polymorphism of the 16S rRNA gene of Vibrio vulnificus is a possible indicator of strain virulence. J Clin Microbiol 41: 442-446.

6. Gutacker M, Conza N, Benagli C, Pedroli A, Bernasconi MV, et al. (2003) Population genetics of Vibrio vulnificus: identification of two divisions and a distinct eel-pathogenic clone. Appl Environ Microbiol 69: 3203-3212.

7. Chen CY, Wu KM, Chang YC, Chang CH, Tsai HC, et al. (2003) Comparative genome analysis of Vibrio vulnificus, a marine pathogen. Genome Res 13: 2577-2587.

8. Kim YR, Lee SE, Kim CM, Kim SY, Shin EK, et al. (2003) Characterization and pathogenic significance of Vibrio vulnificus antigens preferentially expressed in septicemic patients. Infect Immun 71: 5461-5471.

9. Park JH, Cho YJ, Chun J, Seok YJ, Lee JK, et al. (2011) Complete genome sequence of Vibrio vulnificus MO6-24/O. J Bacteriol 193: 2062-2063.

10. Gulig PA, de Crecy-Lagard V, Wright AC, Walts B, Telonis-Scott M, et al. SOLiD sequencing of four Vibrio vulnificus genomes enables comparative genomic analysis and identification of candidate clade-specific virulence genes. BMC Genomics 11: 512.

11. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380.

12. Chevreux B, Wetter T, Suhai S (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. In Proceedings of German Conference on Bioinformatics. pp. 45-56.

13. Salzberg SL, Delcher AL, Kasif S, White O (1998) Microbial gene identification using interpolated Markov models. Nucleic Acids Res 26: 544-548.

14. Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 26: 1107-1115.

15. Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178-2189.

16. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25: 955-964.

17. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, et al. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35: 3100-3108.

18. Cain AA, Kosara R, Gibas CJ (2012) GenoSets: visual analytic methods for comparative genomics. PLoS One 7: e46401.

19. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.

20. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, et al. (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32: D258-261.

21. Li J, Jiang J, Leung FC (2012) 6-10x pyrosequencing is a practical approach for whole prokaryote genome studies. Gene 494: 57-64.

22. Stothard P, Wishart DS (2005) Circular genome visualization and exploration using CGView. Bioinformatics 21: 537-539.

23. Kim HU, Kim SY, Jeong H, Kim TY, Kim JJ, et al. (2011) Integrative genome-scale metabolic analysis of Vibrio vulnificus for drug targeting and discovery. Mol Syst Biol 7: 460.

24. Senoh M, Miyoshi S, Okamoto K, Fouz B, Amaro C, et al. (2005) The cytotoxin-hemolysin genes of human and eel pathogenic Vibrio vulnificus strains: comparison of nucleotide sequences and application to the genetic grouping. Microbiol Immunol 49: 513-519.

25. Drake SL, Whitney B, Levine JF, DePaola A, Jaykus LA Correlation of mannitol fermentation with virulence-associated genotypic characteristics in Vibrio vulnificus isolates from oysters and water samples in the Gulf of Mexico. Foodborne Pathog Dis 7: 97-101.

26. Kachlany SC, Planet PJ, DeSalle R, Fine DH, Figurski DH (2001) Genes for tight adherence of Actinobacillus actinomycetemcomitans: from plaque to plague to pond scum. Trends Microbiol 9: 429-437.

27. Bauer S, Gagneur J, Robinson PN GOing Bayesian: model-based gene set analysis of genome-scale data. Nucleic Acids Res 38: 3523-3532.

28. Cain AA, Kosara R, Gibas CJ (2012) GenoSets: visual analytic methods for comparative genomics. PLoS One 7: e46401.

29. Ferenci T (2005) Maintaining a healthy SPANC balance through regulatory and mutational adaptation. Mol Microbiol 57: 1-8.

30. King T, Ishihama A, Kori A, Ferenci T (2004) A regulatory trade-off as a source of strain variation in the species Escherichia coli. J Bacteriol 186: 5614-5620.

31. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, et al. (2010) Tablet--next generation sequence assembly visualization. Bioinformatics 26: 401-402.

32. Rosche TM, Yano Y, Oliver JD (2005) A rapid and simple PCR analysis indicates there are two subgroups of Vibrio vulnificus which correlate with clinical or environmental isolation. Microbiol Immunol 49: 381-389.

33. Rosche TM, EA B, Oliver JD (2010) Vibrio vulnificus genome suggest two distinct ecotypes. Environmental Microbiology Reports 2: 128-132.

34. Hulsmann A, Rosche TM, Kong IS, Hassan HM, Beam DM, et al. (2003) RpoS-dependent stress response and exoenzyme production in Vibrio vulnificus. Appl Environ Microbiol 69: 6114-6120.

35. Bogard RW, Oliver JD (2007) Role of iron in human serum resistance of the clinical and environmental Vibrio vulnificus genotypes. Appl Environ Microbiol 73: 7501-7505.

36. Finkel SE (2006) Long-term survival during stationary phase: evolution and the GASP phenotype. Nat Rev Microbiol 4: 113-120.

37. Froelich BA, Oliver JD (2011) Orientation of mannitol related genes can further differentiate strains of Vibrio vulnificus possessing the vcgC allele. Adv Stud Biol: 151-160.

38. Froelich BA, Oliver JD (2008) Arrangement of Mannitol Genes as an Indicator of Virulence in C-genotype Strains of Vibrio vulnificus. 108th Gen Meet Amer Soc Microbiol; Boston, MA

39. Cohen AL, Oliver JD, DePaola A, Feil EJ, Boyd EF (2007) Emergence of a virulent clade of Vibrio vulnificus and correlation with the presence of a 33-kilobase genomic island. Appl Environ Microbiol 73: 5553-5565.

40. Hoffman JA, Badger JL, Zhang Y, Huang SH, Kim KS (2000) Escherichia coli K1 aslA contributes to invasion of brain microvascular endothelial cells in vitro and in vivo. Infect Immun 68: 5062-5067.

41. Smith AJ, Greenman J, Embery G (1997) Detection and possible biological role of chondroitinase and heparitinase enzymes produced by Porphyromonas gingivalis W50. J Periodontal Res 32: 1-8.

42. Cascales E, Christie PJ (2003) The versatile bacterial type IV secretion systems. Nat Rev Microbiol 1: 137-149.

43. Grohmann E, Muth G, Espinosa M (2003) Conjugative plasmid transfer in gram-positive bacteria. Microbiol Mol Biol Rev 67: 277-301, table of contents.

44. Christie PJ, Atmakuri K, Krishnamoorthy V, Jakubowski S, Cascales E (2005) Biogenesis, architecture, and function of bacterial type IV secretion systems. Annu Rev Microbiol 59: 451-485.

45. Ward DV, Draper O, Zupan JR, Zambryski PC (2002) Peptide linkage mapping of the Agrobacterium tumefaciens vir-encoded type IV secretion system reveals protein subassemblies. Proc Natl Acad Sci U S A 99: 11493-11500.

46. Christie PJ, Vogel JP (2000) Bacterial type IV secretion: conjugation systems adapted to deliver effector molecules to host cells. Trends Microbiol 8: 354-360.

47. Mao F, Dam P, Chou J, Olman V, Xu Y (2009) DOOR: a database for prokaryotic operons. Nucleic Acids Res 37: D459-463.

48. Baker M (2012) Denovo genome assembly: what every biologist should know. Nature Methods 9: 333-337.

49. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821-829.

50. Luo R, Liu B, Xie Y, Li Z, Huang W, et al. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1: 18.

51. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, et al. (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res 18: 810-820.

52. Parra G, Bradnam K, Ning Z, Keane T, Korf I (2009) Assessing the gene space in draft genomes. Nucleic Acids Res 37: 289-297.

53. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, et al. (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res 22: 557-567.

54. Earl D, Bradnam K, St John J, Darling A, Lin D, et al. (2011) Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Res 21: 2224-2241.

55. Magoc T, Pabinger S, Canzar S, Liu X, Su Q, et al. (2013) GAGE-B: An Evaluation of Genome Assemblers for Bacterial Organisms. Bioinformatics.

56. Rawat A, Elasri MO, Gust KA, George G, Pham D, et al. (2012) CAPRG: sequence assembling pipeline for next generation sequencing of non-model organisms. PLoS One 7: e30370.

57. Barriuso J, Valverde JR, Mellado RP (2011) Estimation of bacterial diversity using next generation sequencing of 16S rDNA: a comparison of different workflows. BMC Bioinformatics 12: 473.

58. Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95: 315-327.

59. Zhang W, Chen J, Yang Y, Tang Y, Shang J, et al. (2011) A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. PLoS One 6: e17915.

60. (2013) Genome Annoucements. Genome Annoucements 1.

61. Huang W, Li L, Myers JR, Marth GT (2012) ART: a next-generation sequencing read simulator. Bioinformatics 28: 593-594.

62. Andrews S (2010) FastQC.

63. Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2: 231-239.

64. Oinn T, Addis M, Ferris J, Marvin D, Senger M, et al. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 20: 3045-3054.

65. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, et al. (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9: 75.

66. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676.

67. Delcher AL, Salzberg SL, Phillippy AM (2003) Using MUMmer to identify similar regions in large sequence sets. Curr Protoc Bioinformatics Chapter 10: Unit 10 13.

68. Wright AC, Simpson LM, Oliver JD (1981) Role of iron in the pathogenesis of Vibrio vulnificus infections. Infect Immun 34: 503-507.

69. Amaro C, Biosca EG, Fouz B, Toranzo AE, Garay E (1994) Role of iron, capsule, and toxins in the pathogenicity of Vibrio vulnificus biotype 2 for mice. Infect Immun 62: 759-763.

70. Field D, Garrity G, Gray T, Morrison N, Selengut J, et al. (2008) The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol 26: 541-547.

71. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30: 207-210.

72. Bisharat N, Raz R (1996) Vibrio infection in Israel due to changes in fish marketing. Lancet 348: 1585-1586.

73. Bisharat N, Agmon V, Finkelstein R, Raz R, Ben-Dror G, et al. (1999) Clinical, epidemiological, and microbiological features of Vibrio vulnificus biogroup 3 causing outbreaks of wound infection and bacteraemia in Israel. Israel Vibrio Study Group. Lancet 354: 1421-1424.

74. Biosca EG, Oliver JD, Amaro C (1996) Phenotypic characterization of Vibrio vulnificus biotype 2, a lipopolysaccharide-based homogeneous O serogroup within Vibrio vulnificus. Appl Environ Microbiol 62: 918-927.

75. C. Amaro EGB, C. Esteve, B. Fouz, A.E. Toranzo (1992) Comparative study of phenotypic and virulence properties in Vibrio vulnificus biotypes 1 and 2 obtained from a European eel farm experiencing mortalities. Disease of Aquatic Organisms 13: 29-35.

76. Amaro C, Biosca EG (1996) Vibrio vulnificus biotype 2, pathogenic for eels, is also an opportunistic pathogen for humans. Appl Environ Microbiol 62: 1454-1457.

77. Biosca EG, Amaro C, Larsen JL, Pedersen K (1997) Phenotypic and genotypic characterization of Vibrio vulnificus: proposal for the substitution of the subspecific taxon biotype for serovar. Appl Environ Microbiol 63: 1460-1466.

78. Roig FJ, Amaro C (2009) Plasmid diversity in Vibrio vulnificus biotypes. Microbiology 155: 489-497.

79. Danin-Poleg Y, Elgavish S, Raz N, Efimov V, Kashi Y (2013) Genome Sequence of the Pathogenic Bacterium Vibrio vulnificus Biotype 3. Genome Announc 1: e0013613.

80. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676.

81. Suzuki H, Lefebure T, Hubisz MJ, Pavinski Bitar P, Lang P, et al. Comparative genomic analysis of the Streptococcus dysgalactiae species group: gene content, molecular adaptation, and promoter evolution. Genome Biol Evol 3: 168-185.

82. Hasan NA, Grim CJ, Haley BJ, Chun J, Alam M, et al. (2010) Comparative genomics of clinical and environmental *Vibrio mimicus*. Proc Natl Acad Sci USA 107(49): 21134–21139. doi: 10.1073/pnas.1013825107.

83. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.

84. Guindon S, Gascuel (2003) A Simple, Fast, and Accurate Algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52(5): 696–704. doi:

85. Rambaut A (2010) Figtree. http://tree.bio.ed.ac.uk/software/figtree.

86. Touw DS, Nordman CE, Stuckey JA, Pecoraro VL (2007) Identifying important structural characteristics of arsenic resistance proteins by using designed three-stranded coiled coils. Proc Natl Acad Sci U S A 104: 11969-11974.

87. Wainipee W, Weiss DJ, Sephton MA, Coles BJ, Unsworth C, et al. (2010) The effect of crude oil on arsenate adsorption on goethite. Water Res 44: 5673-5683.

88. Smith C (2 July 2010) Oil spills raise arsenic levels in the ocean,says new research. Imperial College London.

89. Tao Z, Bullard S, Arias C (2011) High numbers of Vibrio vulnificus in tar balls collected from oiled areas of the north-central Gulf of Mexico following the 2010 BP Deepwater Horizon oil spill. Ecohealth 8: 507-511.

90. Fasano A, Baudry B, Pumplin DW, Wasserman SS, Tall BD, et al. (1991) Vibrio cholerae produces a second enterotoxin, which affects intestinal tight junctions. Proc Natl Acad Sci U S A 88: 5242-5246.

91. Chiang SR, Chuang YC (2003) Vibrio vulnificus infection: clinical manifestations, pathogenesis, and antimicrobial therapy. J Microbiol Immunol Infect 36: 81-88.

92. Roland FP (1970) Leg gangrene and endotoxin shock due to vibrio parahaemolyticus--an infection acquired in New England coastal waters. N Engl J Med 282: 1306.

93. Blake PA MM, Weaver RE, Hollis DG, Heublien PC (1979) Disease caused by a marine vibrio. Clinical characteristics and epidemiology. N Engl J Med 300: 1-5.

94. Pandey DP, Gerdes K (2005) Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. Nucleic Acids Res 33: 966-976.

95. Christensen SK, Mikkelsen M, Pedersen K, Gerdes K (2001) RelE, a global inhibitor of translation, is activated during nutritional stress. Proc Natl Acad Sci U S A 98: 14328-14333.

96. Christensen SK, Pedersen K, Hansen FG, Gerdes K (2003) Toxin-antitoxin loci as stress-response-elements: ChpAK/MazF and ChpBK cleave translated RNAs and are counteracted by tmRNA. J Mol Biol 332: 809-819.

97. Yamamoto TA, Gerdes K, Tunnacliffe A (2002) Bacterial toxin RelE induces apoptosis in human cells. FEBS Lett 519: 191-194.

98. Bisharat N, Cohen DI, Harding RM, Falush D, Crook DW, et al. (2005) Hybrid Vibrio vulnificus. Emerg Infect Dis 11: 30-35.

99. Bisharat N (2010) Population genetics of vibrios. Hoboken, NJ: Wiley. 378-401 p.

100. Bisharat N, Bialik A, Paz E, Amaro C, Cohen DI (2011) Serum antibodies to Vibrio vulnificus biotype 3 lipopolysaccharide and susceptibility to disease caused by the homologous V. vulnificus biotype. Epidemiol Infect 139: 472-481.

101. Bisharat N, Cohen DI, Maiden MC, Crook DW, Peto T, et al. (2007) The evolution of genetic structure in the marine pathogen, Vibrio vulnificus. Infect Genet Evol 7: 685-693.

102. Bisharat N, Bronstein M, Korner M, Schnitzer T, Koton Y (2013) Transcriptome profiling analysis of Vibrio vulnificus during human infection. Microbiology 159: 1878-1887.

APPENDIX A: POLYMERASE CHAIN REACTION PRIMERS FOR VIBRIO VULNIFICUS DNA CONTENT VALIDATION

Primer Name	Primer Sequence	Identification Purpose
csr AupF1 csr AupF2	5'-CGACCTTATTGCTTCCCGAT 5'-GTCAGCCTCTATCATTCAGAG	*V. vulnificus* Chromosome 1
Rpod UP Rpod DOWN	5'-GACCAAGCACGTACGATTC 5'-GCATTTGCATACGCTCTG	*V. vulnificus* Chromosome 1
vvhA F vvhA R	5'-AGCGGTGATTTCAACG 5'-GGCCGTCTTTGTTCACT	*V. vulnificus* Chromosome 2
pepRF F2 pepR3	5'-AGTTGTCCATATGCCTGCCTC 5'-ACGAGAGTTTCCGCTGATGA	*V. vulnificus* Chromosome 2
vvSSF1 vvSSR1	Seq 5' GGCAAAGCCTCTTGTAGACAC Seq 3' TGATAGAGTGGCAAGGGTGCC	Plasmid content
vvF2 vvR2	Seq 5' ACACACCGCATCAACGGATTGAAC (plus) Seq 5' GCAAGGGTGCATAAAAGGAGTGCC (minus)	Plasmid content

Two sets of Primers were generated (the Primer 3 software) using the conserved regions of Plasmid YJ016 and PC4602-1 with expected product length of 244 and 209bps. The conserved sequenced used for primer generation were blasted against the genomic sequence of *Vibrio vulnificus* CMCP6 and YJ016 stains to ensure that they were exclusively for two plasmid sequences.

APPENDIX B: VIBRIO VULNIFICUS CDC 9030-95 (ORL 1506)

Inaccurate provenance information of the *V. vulnificus* CDC 9030-95 (ORL 1506) genome sequence data caused it to be overlooked during the removal of the 'N' character step. Each genome was run independently, human error is at fault.