Complete Genome Sequences of Seven *Vibrio anguillarum* Strains as Derived from PacBio Sequencing

Kåre Olav Holm*, Cecilie Bækkedal, Jenny Johansson Söderberg, and Peik Haugen*

Department of Chemistry and Center for Bioinformatics (SfB), Faculty of Science and Technology, UiT – The Arctic University of Norway, Tromsø, Norway

***Corresponding authors:** E-mails: kare.olav.holm@uit.no; peik.haugen@uit.no.

Accepted: April 6, 2018

Data deposition: New GenBank assembly numbers (from table 1): GCA_002211505.1, GCA_002211985.1, GCA_002212005.1, GCA_002212025.1, GCA_002287545.1, GCA_002291265.1, and GCA_002310335.1.

Abstract

We report here the complete genome sequences of seven *Vibrio anguillarum* strains isolated from multiple geographic locations, thus increasing the total number of genomes of finished quality to 11. The genomes were de novo assembled from long-sequence PacBio reads. Including draft genomes, a total of 44 *V. anguillarum* genomes are currently available in the genome databases. They represent an important resource in the study of, for example, genetic variations and for identifying virulence determinants. In this article, we present the genomes and basic genome comparisons of the 11 complete genomes, including a BRIG analysis, and pan genome calculation. We also describe some structural features of superintegrons on chromosome 2 s, and associated insertion sequence (IS) elements, including 18 new ISs (ISVa3 – ISVa20), both of importance in the complement of *V. anguillarum* genomes.

Key words: *Vibrio anguillarum*, chromosomal integrons, integrases, insertion sequences, IS-elements, PacBio sequencing.

Introduction

Vibrio (Listonella) anguillarum is a marine bacterium and the causative agent of hemorrhagic septicemia (or vibriosis), in fish, molluscs, and crustaceans (Frans et al. 2011). The pathogenic nature of *V. anguillarum* and its global impact on the aquaculture industry continues to keep this bacterium in the spotlight. In efforts to elucidate virulence determinants and/or to analyze genetic variations among strains, 44 genome sequences have been determined (Agarwala et al. 2018).

Recently, Holm et al. (2015) reported the complete genome of the virulent strain NB10, originally isolated from diseased rainbow trout (*Oncorhynchus mykiss*) on the Swedish coast of the Gulf of Bothnia. Its genome, which is typical in size (average 4.31 Mb), is 4,373,835 bp in total, and consists of two circular chromosomes and a pJM1-like plasmid named p67 (66.8 kb). This is 255 kb larger than the genomes of strains 775 and M3, which were published in 2011 and 2013, respectively (Naka et al. 2011; Li et al. 2013). The majority of the 255-kb DNA represent prophages, genomic islands, and genes of unknown function/hypothetical protein genes (Holm et al. 2015). Strain 775 was isolated from Coho salmon (*Oncorhynchus kisutch*) on the United States Pacific coast, and strain M3 was isolated from Japanese flounder (*Paralichthys olivaceus*) off the coast of China. Another previously available complete genome includes that of strain 90-11-286 (Castillo et al. 2017). The initially complete strains NB10, 775, and M3 all harbor a pJM1-like plasmid, strain 90-11-286 has no plasmid.

Materials and Methods

Bacterial Isolates

Vibrio anguillarum strains 87-9-116, JLL237, S3 4/9, CNEVA NB11008, VIB43, and VIB12 were kindly provided by Prof. Hans Rediers (KU Leuven Association, Sint-Katelijne-Waver, Belgium). Strain ATCC-68544 (synonym 775) was acquired from the ATCC Bacteriology Collection. Bacteria were routinely grown at 22°C on BD Difco Marine Agar 2216 (Fisher Scientific) and in liquid cultures in BD Bacto Tryptic Soy Broth (Fisher Scientific).

Table 1
Complete *Vibrio anguillarum* Genomes

Strain	Size (bp) Chr1/Chr2	Plasmid[a]	Assembly	Serovar	Technology[b]	Reference
NB10	3,119,695/1,187,342	66,798	GCA_000786425.1	01	454 and PacBio	(Holm et al. 2015)
775	3,063,912/988,135	65,009	GCA_000217675.1	01	454	(Naka et al. 2011)
M3	3,063,587/988,134	66,164	GCA_000462975.1	01	454	(Li et al. 2013)
90-11-286	3,048,854/1,293,370	No	GCA_001660505.1	01	Illumina PacBio	(Rasmussen et al. 2016)
87-9-116	3,130,467/1,207,658	No	GCA_002211505.1	01	PacBio Illumina*	This study
JLL237	3,122,822/1,164,167	No	GCA_002211985.1	01	PacBio Illumina*	This study
S3 4/9	2,955,425/1,227,548	No	GCA_002212005.1	01	PacBio Illumina*	This study
CNEVA NB11008	3,132,527/1,123,902	No	GCA_002212025.1	03	PacBio Illumina*	This study
VIB43	3,239,943/1,152,744	15,178	GCA_002287545.1	01	PacBio Illumina*	This study
ATCC-68554	3,078,846/998,051	65,009	GCA_002291265.1	01	PacBio	This study
VIB12	3,323,092/1,282,503	292,095	GCA_002310335.1	02	PacBio Illumina*	This study

[a]Total sizes are as listed in the NCBI genomes resource. The 775 assembly does not include the pJM1 plasmid (AY312585.1/65,009 bp), but has been added to this table for clarity.
[b]Illumina sequences (scaffold level) associated with an asterisk are publically available (Busschaert et al. 2015).

DNA Isolation and DNA Sequencing

Total DNA was isolated from 6 ml overnight cultures at stationary phase using Genomic-tip 100/g (Qiagen) according to the manufacturer protocol. The final DNA concentration and quality were measured using a Nanodrop 2000c (Thermo Scientific) instrument. Integrity of high-molecular weight DNA was examined on a 1% agarose gel. DNA samples were sequenced at the Norwegian Sequencing Centre (NSC: a national sequencing core facility located in Oslo).

Genome Analysis

SMRT sequencing was performed at NSC. Libraries were constructed using Pacific Biosciences 20-kb library preparation protocol. Size selection of the final library was performed using BluePippin with a 7-kb cut-off. Libraries were sequenced on Pacific Biosciences RS II instrument using P6-C4 chemistry with 360-min movie time. Reads were assembled using HGAP v3 (Pacific Biosciences, SMRT Analysis Software v2.3.0). Contigs were circularized using Minimus2 software of Amos package (Schatz et al. 2013).

For CNEVA NB11008 and JLL237, the number of circular contigs were bioinformatically corrected. The BRIG software (Alikhan et al. 2011) was used to compare the 11 complete genomes (with the NB10 strain chromosomes as a reference). The genomes of ATCC-68544 and 775 were globally compared using the Artemis Comparison Tool (ACT), and the comparison file was produced with the DOUBLE ACT v.2 server (Carver et al. 2005).

Results and Discussion

As of March 2018, 11 *V. anguillarum* genomes of finished quality are available in the genome databases (table 1). A rough overview of the location at which these strains originate is shown in supplementary figure S1, Supplementary Material online (although exact geographical positions for most of them are unavailable). All strains originate from Europe, except 775/ATCC-68544 (from the United States Pacific coast) and M3 (from China).

The sequencing statistics for strains sequenced in this study are shown in supplementary table S1, Supplementary Material online. The complete genomes were assembled from Pacific Biosciences (PacBio) sequence reads (32,981–139,094) produced at the Norwegian Sequencing Centre (NSC). These sequences produced 84.4–207.8× genome coverage, and assembled into circular contigs (i.e., chromosomes and plasmids). The total genome sizes ranged from 4.14 to 4.89 Mb, with an average GC content of 44.4%.

Figure 1 shows a global BLAST comparison of the 11 complete genomes generated using the BRIG software (Alikhan et al. 2011). BLAST matches of sequences from each strain were mapped onto Chromosome 1 and Chromosome 2 of NB10 (i.e., the reference). Overall, the figure shows that the majority of sequences present in NB10 are also present in all others strains. However, the BRIG analysis does not display sequences not found in NB10. To add more information we therefore calculated the pan genome using the GET_HOMOLOGUES tool (Contreras-Moreira and Vinuesa 2013). The orthoMCL algorithm was used with default

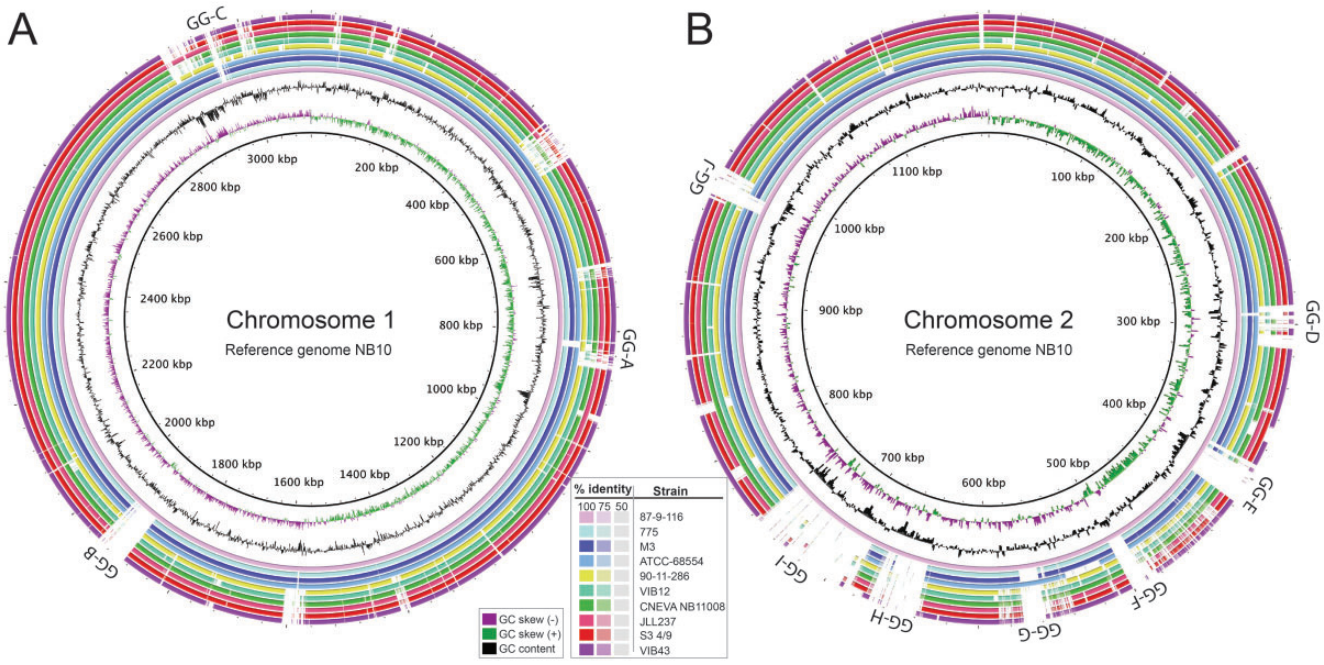

Fig. 1.—Comparison of 11 complete *Vibrio anguillarum* genomes. The figure was generated using the BLAST Ring Image Generator (BRIG) tool, and Chromosome 1 (*A*) and Chromosome 2 (*B*) of strain NB10 as central references (black ring in center). Genomic gaps (GG-A–GG-J) are the same as previously described (Holm et al. 2015). BLAST matches between NB10 and other strains are shown as concentric colored rings on a sliding scale according to percentage identity (100%, 75%, or 50%). GC content and skew are also shown.

parameters and the 11 complete genomes as input. In brief, the numbers from GET_HOMOLOGUES suggest that the pan genome include 7,667 gene clusters in total, 2,574 core clusters, 2,183 accessory clusters, and finally 2,910 unique clusters. This clearly demonstrates that the total number of genes greatly exceeds the number of genes in each genome (complete genomes contain between 3,426 and 4,127 genes). Moreover, figure 1 shows that the genome gaps (GGs) B, C, and E–J, that have previously been described as sequences present in NB10, but not in 775 and M3 (Holm et al. 2015), are also missing from the majority of strains in the current analysis. However, NB10 sequences in GGs A and D are present in most strains. In general, sequences located in the GGs represent hypothetical CDSs, genomic islands, or prophages. Other GGs are also present, but will not be described in further detail in this work. Finally, based on the BRIG analysis, and as the only complete strain, 87-9-116 appears to contain all (or close to all) CDSs that are present in NB10. A likely explanation relies on the fact that both strains have been sampled from relatively close geographical locations in the Gulf of Bothnia, well adapted to the environment and local biotic factors, even though from different salmonid species.

Notably, according to the ATCC Bacteriology Collection, the strain names ATCC-68554 and 775 are synonymous. To verify this relationship ATCC-68554 was acquired directly from the ATCC Bacteriology Collection, cultured under standard conditions, and finally sequenced. A global pairwise genome comparison of the two genomes was done using the Artemis Comparison Tool (ACT) and the DOUBLE ACT v.2 server (Carver et al. 2005; see supplementary fig. S2, Supplementary Material online). This shows that both Chromosome 1 and Chromosome 2 sequences are highly similar, except for a region located approximately between positions 470,000–597,000 on Chromosome 2 (according to the ATCC-68554 sequence). This region represents a so-called "superintegron" (SI), which normally begins with an integron integrase, but in ATCC-68554 it is truncated and thus nonfunctional (locus tag CLI14_17245). *AttC*-containing regions (i.e., the SIs) are marked in yellow in supplementary figure S2*C*, Supplementary Material online. In ATCC-68554 this sequence is 54 kb longer than that of 775. In 775, the SI is followed by a 26-kb region, which is not present in ATCC-68554. The latter 775-specific sequence contains CDSs of various functions, and one tRNA-Gly gene. These discrepancies may be explained by technical artefacts during sequencing and assembly, or by real differences in the genomes, perhaps as a result of subculturing of the bacterium in different laboratories. The SIs and associated insertion sequences (ISs) are described in more detail below.

SIs are subsets of chromosomal integrons (CI) found in vibrios and a wide range of other gram negative bacterial species (for review, see, Cambray et al. 2010). Integrons contain a functional platform (i.e., the integrase encoding gene, *intI*, a primary integration site, *attI*, and a primary promoter,

P_C) administering integrated gene cassettes [i.e., ORF(s) followed by a recombination site *attC*]. Superintegron *attC* sites are species specific (Mazel 2006; Cambray et al. 2010), with a high degree of identity and a common set of characteristics that enable them to be identified, which is the reason why we focused on these cassette components as SI-markers. The supplementary figure S3, Supplementary Material online, shows an alignment of *attC*-sites from the serovar O1 NB10 strain, with a consensus 31-nt "cassette-identifier": 5′-TAACAAACGnnTCAAGAGGGAnnGnCAACGC-3′. This cassette-identifier constitutes part of the quality assurance system in the final assembly of the genomes, enabling calculations of the number of cassettes within the SI-part of chromosome 2 s. The SI gene content (*attC*-span) of finished genomes varies relative to chromosome 2-sizes, between 6.9% in strains 775 and M3 (harboring equally sized and the smallest chromosome 2 s, both with 64 attC-sites) and 28.9% in strain S3 4/9 (containing 147 *attC*-sites/cassettes). Worth mentioning in this context is the low number of *attC*-sites in the published *partially* complete genomes (span: 1–46; average: 22 cassette identifiers; see supplementary table S2, Supplementary Material online), most likely due to their missing genes (cassettes).

Vibrio anguillarum CIs harbor clusters of highly diverse gene cassettes (VAR; *Vibrio anguillarum* repeats), mostly of unknown function, but among others toxin/antitoxin cassettes and genes involved in substrate modification or interactions with virulence factors and DNA modification, similar to in *Vibrio cholera* (Rowe-Magnus et al. 2003).

Also embedded in *V. anguillarum* genomes, and especially within SIs, are numerous insertion sequences (IS: i.e., transposases, and sometimes one or two accessory genes). The *V. anguillarum* SIs encode a specific integrase denoted VangIntIA (based on the naming of VchIntIA, a specific integrase in *V. cholera* [O1] El Tor strain N16961; Mazel et al. 1998).

A striking observation is that the VangIntIA gene is truncated in many strains, nearly always due to the insertion of an ISVa5-element (see supplementary fig. S4, Supplementary Material online). It is also worth mentioning that this truncation apparently co-occurs with the presence of a pJM1-like plasmid (carrying two ISVa5 elements, see supplementary table S2, Supplementary Material online; bungled only by strain 87-9-116). Whether there is a functional link between these two genetic coincidences is unknown. A further exhaustive scrutiny of *V. anguillarum* CI/SI genes is not within the scope of this study. However, the completion of seven additional genomes means that we are nevertheless able to present the scientific community with significant new knowledge.

The presence of repetitive IS-elements may present major technical challenges during sequencing and assembly of microbial genomes, especially when using short read methods, and the majority of genomes in the archives are therefore frequently found in a large number of contigs (Busschaert et al. 2015). Our work revealed 18 new IS-elements (ISVa3-ISVa20), which are available in the "ISfinder" database (Siguier et al. 2006) (see supplementary table S3 and data file S1, Supplementary Material online). Resolving the order of a high number of contigs by using, for example, long-range PCRs is very time-consuming and costly. As an alternative, we used PacBio sequencing, which offers long-sequence reads, and is therefore excellent for resolving regions with repetitive DNA. The resulting sequences were therefore de novo assembled into circular, gap free contigs without ambiguous bases. For strains CNEVA NB11008 and JLL237, discrepancies after PacBio sequencing and assembly were bioinformatically resolved. Two of the three PacBio circular contigs in strain CNEVA NB11008 were found to contain subsets of a superintegron located on Chromosome 2 (based on the presence and distribution of *attC* sites). Regarding the three PacBio circular contigs from strain JLL237, their size distribution clearly suggested a merger of the two smallest into a complete circularized Chromosome 1; a reassembly of the two was also supported by their lack of *attC*-sites. The Artemis Comparison Tool (ACT) was used to make comparisons between the respective PacBio contigs and the NB10 genome (LK021130/LK021129), forming the basis of the bioinformatic correction of their final chromosome sequences.

In summary, we have in this work sequenced seven strains of *V. anguillarum* to completion using the PacBio method, thus bringing the total number of finished genomes to 11 (as of March 2018). A pan genome based on the 11 genomes was calculated, and includes 7,667 gene clusters in total; 2,574 core clusters, 2,183 accessory clusters, and 2,910 unique clusters. These numbers show that the total number of genes among the strains is much greater than those found in each individual strain, which suggests considerable variation among strains, and that more genomes should be sequenced to completion in order to perform detailed genome comparisons, thus significantly further increasing the supply of resources for future studies of this important fish pathogen.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Prof. Hans Rediers (KU Leuven Association, Sint-Katelijne-Waver, Belgium) for providing strains for this study. The publication charges for this article have been funded by a grant from the publication fund of UiT The Arctic University of Norway.

Literature Cited

Agarwala R, et al. 2018. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 46(D1):D8–D13.

Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 12:402.

Busschaert P, et al. 2015. Comparative genome sequencing to assess the genetic diversity and virulence attributes of 15 *Vibrio anguillarum* isolates. J Fish Dis. 38(9):795–807.

Cambray G, Guerout AM, Mazel D. 2010. Integrons. Annu Rev Genet. 44:141–166.

Carver TJ, et al. 2005. ACT: the Artemis Comparison Tool. Bioinformatics 21(16):3422–3423.

Castillo D, et al. 2017. Comparative genome analyses of *Vibrio anguillarum* strains reveal a link with pathogenicity traits. mSystems 2:1–14.

Contreras-Moreira B, Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol. 79(24):7696–7701.

Frans I, et al. 2011. *Vibrio anguillarum* as a fish pathogen: virulence factors, diagnosis and prevention. J Fish Dis. 34(9):643–661.

Holm KO, Nilsson K, Hjerde E, Willassen NP, Milton DL. 2015. Complete genome sequence of *Vibrio anguillarum* strain NB10, a virulent isolate from the Gulf of Bothnia. Stand Genomic Sci. 10:60.

Li G, Mo Z, Li J, Xiao P, Hao B. 2013. Complete genome sequence of *Vibrio anguillarum* M3, a serotype O1 strain isolated from Japanese flounder in China. Genome Announc. 1(5):e00769-13.

Mazel D. 2006. Integrons: agents of bacterial evolution. Nat Rev Microbiol. 4(8):608–620.

Mazel D, Dychinco B, Webb VA, Davies J. 1998. A distinctive class of integron in the *Vibrio cholerae* genome. Science 280(5363):605–608.

Naka H, et al. 2011. Complete genome sequence of the marine fish pathogen *Vibrio anguillarum* harboring the pJM1 virulence plasmid and genomic comparison with other virulent strains of *V. anguillarum* and *V. ordalii*. Infect Immun. 79(7):2889–2900.

Rasmussen BB, et al. 2016. *Vibrio anguillarum* is genetically and phenotypically unaffected by long-term continuous exposure to the antibacterial compound tropodithietic acid. Appl Environ Microbiol. 82(15):4802–4810.

Rowe-Magnus DA, Guerout AM, Biskri L, Bouige P, Mazel D. 2003. Comparative analysis of superintegrons: engineering extensive genetic diversity in the Vibrionaceae. Genome Res. 13(3):428–442.

Schatz MC, et al. 2013. Hawkeye and AMOS: visualizing and assessing the quality of genome assemblies. Brief Bioinformatics 14(2):213–224.

Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 34(90001):D32–D36.

Associate editor: Howard Ochman