UiT

THE ARCTIC
UNIVERSITY
OF NORWAY

FACULTY OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF CHEMISTRY
MOLECULAR BIOSYSTEMS RESEARCH GROUP

An Insight into the *Aliivibrio* genus

A comparative study on relationships and traits of species within the genus Aliivibrio

Terje Klemetsen

KJE-3900 Master's Thesis in Chemistry
May 2016

Acknowledgements

I am grateful for the time I have been granted to work with bioinformatics. I would especially thank my supervisor Nils Peder Willassen for making this masters degree possible, and for all the talks and good suggestions. My work has been both challenging and highly rewarding. Experience on the field has grown in these passing years, and so too, the interest and desire for more knowledge.

I also thank my very inspiring office mates Cecilie Bækkedal and Espen Åberg who as well contribute to the field of bioinformatics. Moreover, my family has supported me all the way since I started as a student seven years ago and I am very grateful for all they have done.

Terje Klemetsen
Tromsø, May 2016

Abstract

Background

Few studies have emphasized on the genus *Aliivibrio* as a whole and lags behind the better known *Vibrio*. Nevertheless, the *Aliivibrio* has for several decades been associated with species expressing bioluminescence like the *Aliivibrio fischeri*, but has also been linked to costly diseases in the fish farming industry such as *Aliivibrio salmonicida*. In an attempt to gain insight in the genus on a broad level, *Aliivibrio* genomes were sequenced, assembled and annotated prior to phylogenetic and pan-genome analysis. Additionally, mapping of genes related to quorum sensing and the CRISPR defense system was performed in a comparative manner. Works like this have never been carried out before on this scale for the *Aliivibrio* and is needed to better understand the complexity of this genus.

Results

This project found the *Aliivibrio* to be distinct in the *Vibrionaceae* family and harbors a diverse range of species with a relatively large set of dispensable genes. Among these appeared five new proposed species, and comparison showed deviations in quorum sensing genes with novel additions to previously described systems. This comes as well as proving the presence of Type I-F CRISPR system in most strains of *A. wodanis*.

Conclusion

Aliivibrio was discovered to be a distinct and diverse genus on more than one level, and is likely to harbor additional species and traits that still remain to discover.

Abbreviations

AAI	-	Average Amino acid Identity
AHL	-	Acyl Homoserine Lactone
ANI	-	Average Nucleotide Identity
BIC	-	Bayesian Information Criterion
BLAST	-	Basic Local Alignment Search
cas genes	-	CRISPR Associated genes
CDS	-	Coding DNA Sequences
CLC	-	CLC genomic workbench
COG	-	Clusters of Orthologous Groups
CPU	-	Central Processing Unit
CRISPR	-	Clustered Regularly Interspaced Short Palindromic Repeat
crRNA	-	CRISPR RNAs, spacer and repeat products
DDBJ	-	DNA Data Bank of Japan
DNA	-	Deoxyribonucleic Acid
ENA	-	European Nucleotide Archive
GC	-	Guanine-Cytosine content
GTR	-	General Time Reversible model
G	-	Gamma distributed
I	-	Invariant positions
HGT	-	Horizontal Gene Transfer
IUPAC	-	International Union of Pure and Applied Chemistry
KEGG	-	Kyoto Encyclopedia of Genes and Genomes
MLSA	-	MultiLocus Sequence Analysis
NCBI	-	National Center for Biotechnology Information
NGS	-	Next Genetation Sequencing
PAM	-	Protospacer Adjacent Motif
PCR	-	Polymerase Chain Reaction
QS	-	Quorum Sensing
RNA	-	Ribonucleic Acid
ML	-	Maximum likelihood

Table of contents

Background

The *Vibrionaceae* and *Aliivibrio*

The exploration of the prokaryotic universe is an ongoing process in the scientific community. The share amount of bacterial in the earth's soil and water is astounding. It is roughly estimated that the number of bacteria in the world's ocean outnumbers the stars in our known cosmos by a 100 million times (Copley, 2002). By such numbers we are still starting to understand the diversity of these single cell organisms.

As with all organisms bacteria are classified in ranks from the top domain down to the species level. The family of *Vibrionaceae* is one amidst the large proteobacteria phyla and its name likely stems from one of the oldest and most publicly known genera, the *Vibrio* who was proposed back in 1854 (Gomez-Gil et al., 2003). Holding the well known *Vibrio cholera*, the *Vibrionaceae* also comprises a large number of species shared by its eleven genera, according to the NCBI taxonomy listings (Federhen, 2012). Its species are known to inhabit a wide selection of environments and niches and is known to be both abundant and ubiquitous in aquatic environment as well as in symbiotic relationships with marine and freshwater organisms (Thompson et al., 2004). Within the *Vibrionaceae* family is the *Aliivibrio fischeri*, the first organism to shed knowledge about the specific genes responsible for bioluminescence (Miyashiro and Ruby, 2012). This *Aliivibrio* genus also harbors the *A. salmonicida* and *A. wodanis* who are related to the costly illnesses of coldwater vibriosis (Hjerde et al., 2008) and winter-ulcers (Hjerde et al., 2015) in farmed fish. But the *Aliivibrio* genus as a whole has not received full attention in the literature and information is still lacking as concerns its diversity and species traits.

The relationship of *Aliivibrio* species has evolved in the recent decades and has long been associated with *Photobacteria* due to similar phenotypic traits and molecular characters. Until 2007 it was defined as part of the *Vibrio* genus, but became reclassified as *Aliivibrio* due to its ability to exhibit monophyletic relationship with its species (Urbanczyk et al., 2007). Identification and classification of the *Aliivibrio* species started with the use of microscopes and standard laboratory tests where identification was strictly phenotypic. Small scale sequencing of 16S ribosomal sequences by applying PCR technology made molecular phylogenetic analysis possible, but this particular biomarker has shown to be ambiguous due to low resolution (Urbanczyk et al., 2007). In a recent study by Sawabe and collaborators multi locus sequence analysis (MLSA) of eight housekeeping genes was carried out on

Vibrionaceae species. This study included 96 taxa in a split network tree (Figure 1) to represent most of the known species in *Vibrio* as well as a selection of the remaining genera represented in *Vibrionaceae* (Sawabe et al., 2014). The majority of these formed distinct and separable clades, including the "*Fischeri*" which constitutes five of the known *Aliivibrio* taxa; *A. fischeri*, *A. sifiae*, *A. wodanis*, *A. salmonicida* and *A. logei*. Their study gave a rough overview of the interconnections between the included genera and species in this family, but showed little detail and emphasis on the *Aliivibrio*. Similar multi locus analysis from 2009 by Ast and collaborators demonstrated the *Aliivibrio* being distinct from other genera, forming well supported clades with its species (Ast et al., 2009). They also introduced two new species at the time, the *A. thorii* and *A. sifiae* from previously ambiguous taxa.

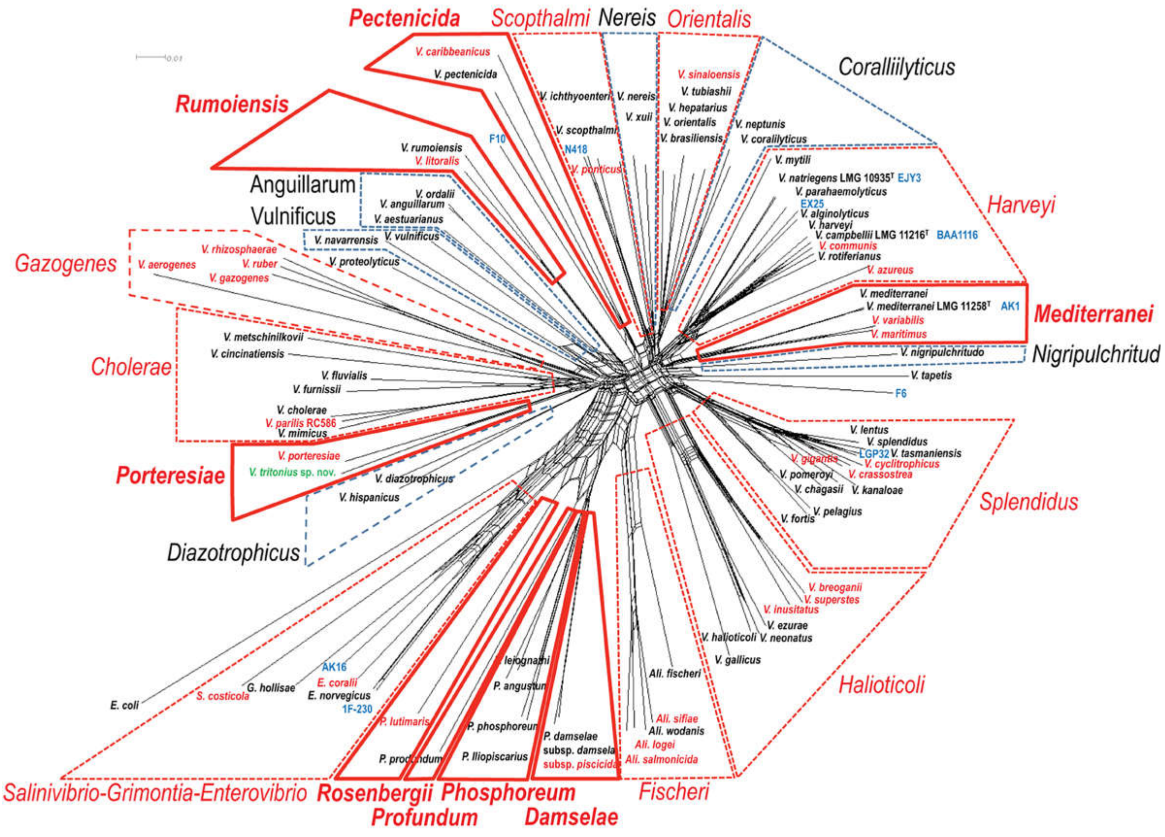


Figure 1. Split network tree of 96 taxa based on the 8 housekeeping genes *ftsZ*, *gapA*, *gyrB*, *mreB*, *pyrH*, *recA*, and *topA*. The network emphasizes cladistic relationship among *Vibrio* species of the *Vibrionacea*, which represents the genus *Aliivibrio* as "*Fischeri*". Figure adapted from the work of Sawabe and collaborators (Sawabe et al., 2014).

Sequencing technology and post processing

High-throughput – next generation sequencing (NGS) act as an important tool in life sciences these days and are capable of processing astonishing amounts of genetic data (Pareek et al., 2011). Leaving behind the Sanger sequencing technology, NGS has went through numerous improvements since it was publicly introduced to the science community a decade ago. Its processing speed, read length and throughput has steadily increased during these years with corresponding cost reduction reaching the $1000 per human genome target. This goal was claimed to be achieved back in 2014 when Illumina introduced their HiSeq X machine (van Dijk et al., 2014, Hayden, 2014). Several methods and platforms have arisen during these competing decades of genetic sequencing and have classified three generations of DNA sequencers. The first generation mainly refers to the dideoxy method developed in the 70s and 80s, while the second and third generation of high-throughput NGS sequencing emerged in the 21st century (Pareek et al., 2011). The second generation is most frequently applied in recent projects and involves technologies developed by Roche, Illumina or SOLiD which are evolve around massive parallel sequencing of fragmented DNA. The third generation is still in the development stage, but aims for the single-molecule DNA sequencing. Several platforms are under testing and include the Heliscope, Nanopore, Single molecule real time (SMRT) sequencer, Single molecule real time (RNAP) sequencer, Real time single molecule DNA sequencer by VisiGen Biotechnologies and Multiplex polony (Pareek et al., 2011).

In massive parallel sequencing, like Illumina Miseq, DNA or RNA samples undergo fragmentation in order to obtain short sequences of ~50 to 500bp. From here the sequences becomes ligated to adapters and barcodes in order to build libraries prior to sequencing (van Dijk et al., 2014). Libraries is loaded onto the Illumina flow cells where the DNA fragments binds and the bridge amplification takes place and results in clusters. Ilulumina sequencing, with similarity to Sanger sequencing, applies a system of terminators to restrict polymerization, adding a single base at a time. These terminators also work as single-color fluorescent labels which are detectable in clusters by a camera. By only using a single-color requires cycling of the four bases during synthesis and imaging. At the end of each cycle the reversible terminators are removed to prepare for a new cycle (Stuart, 2012). This continues until the clustered fragments are completely synthesized. The products generated are amassed reads, often in the 30-40 millions (MiSec), as fasta or fastq format. These formats are only used in an interchange step where the fastq carries information about each single read and its base quality score (Cock et al., 2010). This enables downstream processes to examine and compile statistics on the whole dataset of sequenced reads.

With the fastq files accessible the coming step becomes importing the raw data while applying filters adjusted to remove failed reads and reads below a given quality threshold.

Manual approach in performing quality control (QC) with applications like Prinseq or FastQC (Trivedi et al., 2014) can also be performed prior to import for a greater perspective of the obtained data and the possibility to remove obscure read parts.

The term *de novo* assembly in bioinformatics refers to the computational demanding process of building long contigs or scaffolds from overlapping reads. These will ultimately represent the fragmentary draft genome of the sequenced sample, which contains gaps in the product code often due to repeat sequences. Utilization of larger libraries have been suggested to give better assembly results (Chen et al., 2015) and if read lengths hit 7Kb or more the tricky repeat sequences will almost be extinct, achieving high quality draft genomes (Koren and Phillippy, 2015). Until such technology emerges or significant improvements are made for existing technology, actions can be taken to obtain better accuracy with the contigs or scaffolds assembled. This involves the crude mapping against a reference sequence where contigs will be reoriented and ordered which facilitates better comparison between genomes and clarify variations (Assefa et al., 2009).

Draft genomes can act as good starting point for identification of genes. This step is often automated and termed gene prediction where coding DNA sequences (CDS) becomes determined on the basis of several factors. The simplest applications will only look for start and stop codons and are found to result in numerous false positive predictions, which also are the main issue in this field. Advanced algorithms like Glimmer3 and Prodigal search the genome for CDS by applying interpolated Markov models or base its prediction on advanced GC computations (Aggarwal and Ramaswamy, 2002, Hyatt et al., 2010). These algorithms are relatively fast and have been implemented in piped processes where the product becomes transferred to the next process for functional assignment of predicted genes.

The functional assignment of a single gene is a straightforward task for a bioinformaticians and involves the appending of homology, motif and pattern based information on global and local sequence similarities. Performing such task on thousands of genes for dozens of genomes quickly becomes tedious and unenforceable. Automation of this work has become available with pipelines like RAST (Aziz et al., 2008) and other web-services as well as local installations which completes a task, at best, within hours for whole bacterial genomes (Richardson and Watson, 2013).

The Pan-genome era

In recent years it has been cheaper and easier to obtain draft genomes of one's own samples with sequencing technology, and the number of high quality closed genomes is steadily increasing in public databases. This has led to possibilities where whole genome sequences and/or its content of genes can be compared in a large scale, where the goal is to unravel how species evolve, their gene functions as well as understand the noncoding regions of genomes (Sivashankari and Shanmughavel, 2007). Manny of the techniques involve recognition of homolog sequences and how they are evolving on over time.

Comparative analysis of bacteria, archaea and eukaryote genomes have shown that a significant fraction of gene content to be xenolog, which is likely the result of horizontal gene transfer (HGT) (Koonin et al., 2002). Further studies supporting these reports have additionally shown great variations among genomes, even between the same species, and are as well believed to be the result of HGT events (Riley and Lizotte-Waniewski, 2009). This plasticity in genome content raise questions about what really explains a species when there are such fluctuations. The core genome hypothesis (Lan and Reeves, 1996) has been suggested to represent a species best. The hypothesis build upon that conserved genes contribute greatly to species phenotypic traits as well as functions of maintenance, replication, translation and cellular homeostasis (Tettelin et al., 2005, Medini et al., 2005).

The core of a species genome is only part of the much larger pan-genome, which represents the whole gene repertoire of all genomes considered. The pan-genome can be recognized as open or closed depending on whether additionally added genomes will increase the pan-genome size or remain unchanged (Guimaraes et al., 2015). It is further divided in an accessory (dispensable) component and a unique component, the latter often regarded as species-specific or strain-specific part depending on the study design.

The dispensable accessory genome represents to a greater extent the diversity of a species and may include additive functions expressed in biological pathways that serves as selective advantages (Medini et al., 2005). These functions might be highly beneficial for the given bacterial cell but remain dispensable as the genes are not a essential for proliferation. Strain-specific genes have also been studied where they have shown to occupy as much as 5 to 35% of a single genome. Many of these represent paralogous genes duplicated in a tandem pattern (Jordan et al., 2001)and are thought to be related with pathogenic behavior (Guimaraes et al., 2015).

A pan-genome analysis performed by Kahlke and co-workers from 2012 included 64 *Vibrionacea* genomes where the focus was on functions expressed by the unique core genes, a sub selection of the accessory genome (Kahlke et al., 2012). Here it was concluded that the unique core genes had conserved metabolic functions and can be applied to classification of bacteria on the genomic level. Nevertheless, their analysis only included four *Aliivibrio* genomes and no pan-genome analysis has ever been performed solely on the species of the *Aliivibrio* genus.

The pan-genomic era has required sophisticated methods to perform these multi genome analyses, but has given programmers the opportunity to bring new tools to the field. These are often standalone installations where minimal programming knowledge is required, but are capable of performing a range of tasks with only a few lines of Linux code. CMG-biotools (Vesth et al., 2013), Get_homologues (Contreras-Moreira and Vinuesa, 2013), Bacterial Pan Genome Analysis tool (BPGA) (Chaudhari et al., 2016) are only a small selection of available applications. These usually require known file formats like the GenBank or fasta files containing information about the whole strain and/or its protein or nucleotide sequences, as input. The initial process usually involves blasting all sequences against each other to achieve an accessible foundation for further computation. Specified algorithms, like OrthoMCL (Li et al., 2003) or other less complex rules, are then utilized for the purpose of defining homologous sequence clusters based on blast results. The end product of homologous gene clustering can be further analyzed to gain genome wide insight about; core genes, accessory (dispensable) and unique (strain-specific) genes, gene synteny, phylogeny, GC content, KEGG or COG mapping, codon usage or to generate blastp atlases based on a reference genome. These are only a limited selection of analysis which can be performed on available genomes and will likely be more accessible with the development of even more user-friendly software.

Quorum sensing in bacteria

Cell-density regulated gene expression, better known as quorum sensing, is by definition the smallest number required for initiation of a cellular response. Constituently active quorum sensing is known in bacteria that possess this trait and results in the continuous synthesis and diffusion of autoinducers. Bacteria are found to regulate behaviors like virulence, biofilm formation, bioluminescence, motility and sporulation with quorum sensing as cell numbers reach a prerequisite density. This has been adapted as an essential strategy to save energy as expression by small numbers or a single cell would be futile in achieving the desired effect (Atkinson et al., 2006, Williams et al., 2007).

Quorum sensing was first discovered occurring in high density of *Aliivibrio fischeri* originating from the light-producing organ of the Hawaiian bobtail squid (Nealson et al., 1970). Since the emergence of quorum sensing in the 60s *Escherichia coli*, *Salmonella enterica*, *Pseudomonas aeruginosa*, *Acinetobacter* sp., *Aeromonas* sp. and *Yersinia* are species being referred to in articles containing such systems (Williams et al., 2007). Both gram-positive and gram-negative bacteria may apply modules of quorum sensing and has been proven able to communicate across species. In gram-negative bacteria there has been discovered several types of N-Acyl Homoserine Lactones (AHL) as well as the autoinducer-2 (AI-2) being synthesized (Reading and Sperandio, 2006). These are the main signaling molecules managed and involve the employment of the LuxR/AHL and LuxS/AI-2 systems, where the former involves the bioluminescence activity of *A. fischeri*.

The mechanism of luminescence in *A. fischeri* is governed by the enzymatic action of LuxI that synthesize the N-3-oxohexanoyl-homoserine lactone. When the concentration of this particular AHL has surpassed a given threshold level it will bind to the transcription factor LuxR (Verma and Miyashiro, 2013). This LuxR/AHL complex then readily binds upstream of the lux operon and recruits RNA polymerase for transcription of its genes (Stevens et al., 1994). Subsequent translation of the transcript forms the dimer luciferase (luxAB), LuxG and the reductase complex (luxCDE). Both LuxG and the reductase complex work in concert to supply specific long-chain fatty acids and the reduced oxidizer flavin mononucleotide ($FMNH_2$) as fuel for the luciferase enzyme, thus enforcing luminescence in *A. fischeri* (Verma and Miyashiro, 2013).

Quorum sensing and bioluminescence are well known in the *Vibrionacea* and, in particular, the *Aliivibrio* due to *A. fischeri*. The genus was introduced to a new bioluminescent species in 1978 when Bang and co-workers identified the *Photobacterium logei* sp. nov., classified as *A. logei* today (Bang et al., 1978). In their study the *A. logei* strains showed high phenotypic similarity to *A. fischeri*, but with the inability to grow at 30 degrees Celsius. Later on *A. salmonicida* also became announced but required additional treatment with an aliphatic aldehyde or a specific AHL in order to exhibit bioluminescent characteristics (Fidopiastis et al., 1999). Luminescence in *A. salmonicida* proved in a succeeding study to be partially faulty in comparison with *A. logei*. It was discovered to harbor a deformity in its LuxD coding gene

(Manukhov et al., 2011). Further studies of *A. salmonicida* have been undertaken in order to investigate its pathogenic potential. The importance of LitR, a homologue of the QS master regulator in *A. fischeri*, has proven to be a crucial and temperature sensitive regulator of quorum sensing and biofilm formation (Bjelland et al., 2012, Hansen et al., 2014).

Works like these have mainly focused on well known samples related to distinct phenotypic characteristics or complications related to health and disease. Gaps still has to be filled concerning the diversity of *Aliivibrio* and identify carriers of quorum sensing on a general level.

CRISPR – the antiviral defense system in bacteria

The stand-alone system employed by various prokaryotes that confer an individual cell protective role is named after its arrangement in the genome. Clustered regularly interspaced short palindromic repeat, abbreviated CRISPR, serves as an adaptive and inheritable immune system. It targets invading DNA or RNA from either phages and/or plasmids where it has the ability to learn and recognize their sequences. What would be known as CRISPR was first described in *E. coli* in 1987, but would not be fully realized before *in silico* studies in the early 21st century when the CRISPR RNAs (crRNA) and CRISPR associated (cas) genes were discovered. A couple of years passed until, in 2005, it became clear that it was a link between the observed spacer sequences and phages (Marraffini, 2015).

The CRISPR system works by capturing fragments of the invading DNA, integrate it in the cell's own DNA, express it to achieve hybridization with the target and degrade the invading sequence with cas genes. Due to this integration of novel DNA, known as spacer acquisition, the system becomes inheritable and passes on to the offspring during proliferation. Phages and viruses in general are known to be lacking repair system for its DNA/RNA and results in unrestricted mutation. Thus, mutations can competently circumvent the previously acquired spacers and lead to events where the viruses escape the barrier. In these cases additional new spacers must be obtained to keep up with the rapid phage evolution (Marraffini, 2015, Rath et al., 2015).

The CRISPR system is proposed to be divided in two classes based on if there is a multi subunit (Class 1) or a single subunit (Class 2) target binding protein, the crRNA-effector module. These classes are further split in five types, three for Class 1 and two for Class 2, depending on particular signature genes responsible for the actual target cleavage (Makarova et al., 2015). The scientific community has mainly embraced the Type 2 system of Class 2, carrying the signature *cas9* gene. This particular system has the simplest design and do not processively degrade its target as cas3. Utilization of the Type 2 system has proven to make gene-targeted modifications tasks, like correction of genetic diseases, straightforward and more economically achievable (Rath et al., 2015, Kim and Kim, 2014, Ma et al., 2014).

The complete mechanism of action performed by the CRISPR system can be split in the tree, adaptation, expression and interference (Rath et al., 2015). The adaptation process is the learning technique of the system and has the goal of obtaining spacer information from the invading target. The type I-E system has two proposed ways of acquiring new spacers with the help of the universal cas1-cas2 protein complex (Yosef et al., 2012). One is the naive action when completely unknown, novel sequence information is integrated. The second occurs if existing spacers of the CRISPR system recognize the invading DNA/RNA. If this happens the primed spacer acquisition becomes stimulated which accelerate the mechanism, increasing the number of spacers from the same target and making it less likely to escape the system (Rath et al., 2015). The belligerent role against invading genetic material starts with the expression of attained spacers and repeats as a continuous transcript called pre-crRNA. One of the repeat sequences roles is forming secondary structure hairpins to recruit cas proteins. These will bind and cleave the repeat sequence, discharging the individual spacers as mature crRNA. Interference with the incoming target nucleic acid is achieved when the crRNA, bound to cas proteins, positively identifies a complementary perfect match (protospacer) along with a protospacer adjacent motif (PAM). When the crRNA and viral protospacer hybridize, cleavage (cas9) or degradation (cas3) of the target will be executed depending on system (Makarova et al., 2015, Rath et al., 2015).

Few studies have focused on species of *Vibrionacea* in efforts to map CRISPR systems. Only *Vibrio parahaemolyticus* and *Vibrio cholerae* has been screened for CRISPR systems. In a study by Sun and co-workers 154 strains of *V. parahaemolyticus* resulted in six different CRISPR sequence types. Comparative analysis showed association with known virulence factors and was hypothesized to indicate the virulence potential of *V. parahaemolyticus* strains (Sun et al., 2015). *V. cholerae* has also been proved to benefit from the CRISPR system in a later study. Box and collaborators identified the Type I-E system in the classical biotype of *V. cholerae* where it was described to be prevalent. However, there was no system proven for the El Tor biotype, but under laboratory conditions they showed that the CRISPR system is transferable from the classical biotype to the El Tor biotype (Box et al., 2015). No further dedicated studies have mapped the presence of neither *Vibrionacea* as a whole nor the *Aliivibrio* genus.

Aims of this project

The *Aliivibrio* genus has been largely overshadowed by the focus and attention given to the genera *Vibrio* and *Photobacteri*a with their many species. Thus, little is known about the diversity of the *Aliivibrio* species, their phylogenetic relationship and genome composition.

Reduced cost of sequencing marine bacteria may lead to the discovery of novel species, and some of these might affiliate with the uncharted *Aliivibrio*. This project will centralize on this unfamiliar genus with both its known and ambiguous species with the goal of unraveling its diversity. Phylogenetic and pan-genome analysis was of great interest in achieving heritable relationships and differences in genome composition. As these form the foundation of this project, in-depth mapping and analysis of quorum sensing and the CRISPR viral defense systems will additionally be discussed.

Material

Genomes and house-keeping genes

A total of 81 bacterial strains within the class Gammaproteobacteria and the family *Vibrionaceae*, with the exception of the selected outgroup, were selected for phylogenetic and comparative genome analysis in this project. Of these, 45 were sequenced locally at the UiT The Arctic University of Norway (UiT) using the in-house Illumina MiSeq sequencer. Overview of in-house sequenced genomes is shown in Table 1.

Table 1. In-house sequenced strains included in this project.

Species	Strain	Taxonomy ID	Isolation source	Time	Comment	Location	Reference	PubMed (PMID)
A. finisterrensis	DSM 23419 (CMJ 11.1)	511998	Ruditapes philippinarum	2004/2005	Cultured Manila clam	North-western coast of Spain	Beaz-Hidalgo et al., 2010	19648323
A. friggae sp. nov.	SA12	511678	Sepiola affinis	01.07.1995	Light organ	France: Banyuls-sur-Mer	Fidopiastis et al., 1998	9422593
A. friggae sp. nov.	SR6	511678	Sepiola robusta	01.07.1995	Light organ	France: Banyuls-sur-Mer	Fidopiastis et al., 1998	9422593
A. logei	A11-3	688	Salmo salar	-	Challenge	Norway: Solbergstrand, Frogn	This study	-
A. logei	ATCC 29985	688	Mytilus edulis (arctic mussel)	-	Gut	-	Bang et al., 1978	-
A. logei	MR17-66	688	Styela rustica (Ascidiacea)	01.05.2009	Body surface	Norway: Barent Sea	Purohit et al., 2013	23725044
A. logei	MR17-77	688	Porifera indet	01.05.2009	Body	Norway: Barent Sea	Purohit et al., 2013	23725044
A. logei	MR17-80	688	Porifera indet	01.05.2009	Body	Norway: Barent Sea	Purohit et al., 2013	23725044
A. logei	SES03-1	688	Gadus morhua	01.05.2009	Intestine	Norway: Barent Sea	Purohit et al., 2013	23725044
A. logei	SES03-5	688	Gadus morhua	01.05.2009	Intestine	Norway: Barent Sea	Purohit et al., 2013	23725044
A. magni sp. nov.	R8-63	511678	Eurythenes gryllus (Amphipoda)	01.05.2009	Body	Norway: Barent Sea	Purohit et al., 2013	23725044
A. magni sp. nov.	R8-67	511678	Eurythenes gryllus (Amphipoda)	01.05.2009	Body	Norway: Barent Sea	Purohit et al., 2013	23725044
A. modi sp. nov.	A25	511678	Salmo salar	-	Challenge experiment	Norway: Solbergstrand, Frogn	This study	-
A. modi sp. nov.	A9-1	511678	Salmo salar	-	Challenge experiment	Norway: Solbergstrand, Frogn	This study	-
A. modi sp. nov.	A9-2	511678	Salmo salar	-	Challenge experiment	Norway: Solbergstrand, Frogn	This study	-
A. raniae sp. nov.	A11-1	511678	Salmo salar	-	Challenge experiment	Norway: Solbergstrand, Frogn	This study	-
A. raniae sp. nov.	A15	511678	Salmo salar	-	Challenge experiment	Norway: Solbergstrand, Frogn	This study	-
A. raniae sp. nov.	A22	511678	Salmo salar	-	Challenge experiment	Norway: Solbergstrand, Frogn	This study	-
A. thrudae sp. nov.	2208-14	511678	Cyclopterus lumpus	01.08.2014	Challenge experiment	Norway: Kårvika, Troms	This study	-
A. salmonicida	12	40269	Salmo salar	Fall-82	Diseased fish	Norway: Fiskebøl, 8317 Strønstad	This study	-
A. salmonicida	250	40269	Salmo salar	Spring-87	Diseased fish	Norway: Sætrelaks, 5950 Brekke	This study	-
A. salmonicida	289	40269	Salmo salar	Spring-87	Diseased fish	Norway: Alta laks, 9530 Kviby	This study	-
A. salmonicida	378	40269	Salmo salar	Summer-87	Diseased fish	Norway: Tromsølaks, 9022 Krokelvdalen	This study	-
A. salmonicida	554	40269	Salmo salar	01.04.1994	Diseased fish	Norway: Frøya Edelfisk A/S, 7270 Dylvik	This study	-
A. salmonicida	561	40269	Salmo salar		Diseased fish	Norway:	This study	-
A. salmonicida	574	40269	Oncorhynchus mykiss	01.01.2002	Diseased fish	Norway:	This study	-
A. salmonicida	B9-15	40269	Dendrodoa aggregata (Ascidiacea)	01.05.2009	Body	Norway: Barent Sea	Purohit et al., 2013	23725044
A. salmonicida	LFI-180	40269	Salmo salar		Diseased fish	Norway:	This study	-
A. salmonicida	N5541	40269	Salmo salar		Diseased fish	Norway:	This study	-

A. salmonicida	R5-43	40269	Gersemia rubiformis (soft coral)	01.05.2009	Body	Norway: Barent Sea	Purohit et al., 2013	23725044
A. salmonicida	R8-68	40269	Eurythenes gryllus (Amphipoda)	01.05.2009	Body	Norway: Barent Sea	Purohit et al., 2013	23725044
A. salmonicida	R8-70	40269	Eurythenes gryllus (Amphipoda)	01.05.2009	Body	Norway: Barent Sea	Purohit et al., 2013	23725044
A. salmonicida	TEO	40269	Salmo salar		Diseased fish	Norway:	This study	-
A. sifae	26449	566293	Seawater	29.06.1905	Surface seawater	Japan: Harumi Pier in Tokyo Bay	Yoshizawa et al., 2010	21282907
A. wodanis	Vw1	80852	Salmo salar	01.12.1989	Outbreak	Norway: Saltfjellvik, Frei	This study	-
A. wodanis	Vw11	80852	Salmo salar	01.04.2001	Outbreak	Norway: Svanøybukt	This study	-
A. wodanis	Vw12	80852	Salmo salar	01.04.1988	Vaccination experiments	Norway: Svanøybukt	This study	-
A. wodanis	Vw130426	80852	Salmo salar	01.04.2013	Outbreak	Norway: Hammerfest, Husfjord	This study	-
A. wodanis	Vw27	80852	Salmo salar	01.04.2006	Outbreak	Norway: Buksevika, Flekkefjord	This study	-
A. wodanis	Vw29	80852	Salmo salar	28.06.1905	Outbreak	Norway: Knivskjeneset, Gjemnes	This study	-
A. wodanis	Vw35	80852	Salmo salar	01.09.2006	Experiment with outbreak	Norway: VESO Vikan	This study	-
A. wodanis	Vw5	80852	Salmo salar	01.01.1990	Outbreak	Norway: Straumen, Gratangsbotn	This study	-
A. wodanis	Vw7	80852	Salmo salar	01.01.2002	Outbreak	Norway: Halsanaustan	This study	-
A. wodanis	Vw8	80852	Salmo salar	01.04.2002	Outbreak	Norway: Bogen, Frei	This study	-
A. wodanis	VwK7F1	80852	Salmo salar	01.04.2013	Experiment with outbreak	Norway: Solbergstrand, Frogn	This study	-

Of the 81 genomes, 34 were downloaded from the NCBI GenBank® database (http://www.ncbi.nlm.nih.gov/genbank/) to supplement the locally sequenced genomes. These are summarized in Table 2 and were accessible from the database either as drafts or closed genomes.

Table 2. Genomes obtained from the NCBI GenBank® database.

Species	Strain	GenBank assembly accession	Taxonomy ID	Isolation source	Time	Comments	Location	Reference	PubMed (PMID)
A. fischeri	ES114	GCF_000011805. 1	312309	Euprymna scolopes	01.03.1988	Light organ	USA: Hawaii, Kaneohe bay	Boettcher et al., 1990	2163384
A. fischeri	MJ11	GCF_000020845. 1	388396	Metanephrops Japonicus	1991	Light organ	USA: California	Mandel et al., 2009	19182778
A. fischeri	SR5	GCF_000241785. 1	1088719	Sepiola robusta (bobtail squid)		Light organ	Mediterranean Sea	Gyllborg et al, 2012	22374964
A. fischeri	ZF-211	GCF_000287175. 1	617135	Filtered seawater (64um filter)	38961	Surface seawater	USA: Massachusetts	Cordero et al., 2012	22955834
A. logei	5S-186	GCF_000286935. 1	626086	Filtered seawater (5um filter)	01.04.2006	Surface seawater	USA: Massachusetts	Cordero et al., 2012	22955834
A. salmonicida	LFI1238	GCA_000196495.1	316275	Gadus morhua	39965	Diseased fish	Norway: Hammerfest	Hjerde et al. 2008	19099551
A. wodanis	06/09/139	GCA_000953695.1	80852	Salmo salar	01.03.2006	Outbreak	Norway: Kvangardsnes, Volda	Hjerde et al., 2015	26059548
P. angustum	ATCC 25915	GCF_000950005.1	661	Seawater	-	Seawater at depth of 750 m	-	Reichelt et al., 1976	1015934
P. damselae	CIP 102761	GCF_000176795.1	675817	Chromis punctipinnis	-	Ulcer of a damsel fish	USA: California	Smith et al., 1991	1742198
P. laumondii	TTO1	GCF_000196155.1	243265	Heterorhabditis bacteriophora	-	Symbiontwith H. bacteriophora	Trinidad and Tobago	Duchaud et al., 2003	14528314
P. leiognathi	ATCC 25521	GCF_000950415.1	553611	Leiognathidae	-	Light organ	-	Boisvert et al., 1967	5624740
V. anguillarum	775	GCF_000217675.1	882102	Oncorhynchus kisutch (Coho salmon)	-	Clinical isolate	USA: Pacific Ocean cost	Crosa et al., 1977	924679
V. anguillarum	M3	GCF_000462975.1	882944	Paralichthys olivaceus (flounder)	-	Skin ulcer	China: Shandong	Li et al., 2013	24072867
V. anguillarum	NB10	GCF_000786425.1	55601	Oncorhynchus mykiss	1986	Clinical isolate	Sweden: Boden, Gulf of Bothnia	Rehnstam et al., 1989	2782871
V. furnissii	NCTC11218	GCF_000184325.1	903510	Sediment	-	Estuary, Intertidal zone	England: Hull, River Humber	Lux et al., 2011	21217006
V. harveyi	AOD131	GCF_000347555.1	1287887	Epinephelus lanceolatus (Giant grouper)	-	Outbreak	Taiwan: Kaohsiung	Unpublished	-
V. harveyi	ATCC 14126	GCA_000400305.1	1219071	Talorchestria sp	-	Dead luminescing amphipod	USA: Massachusetts	Urbanczyk et al., 2013	23710045
V. harveyi	E385	GCA_000493315.1	1352943	Epinephelus coioides	40087	Diseased cage-cultured grouper	China: Daya Bay of Guangdong Province	Yu et al., 2013	24336361

V. harveyi	ZJ0603	GCF_000275705.1	1191522	Epinephelus coioides (Orange-spotted grouper)	-	Diseased grouper	China: Guangdong	Huang et al., 2012	23144396
V. ordalii	12B09	GCF_000287135.1	314865	Seawater	37712	Seawater	USA: Massachusetts	Cordero et al., 2012	22955834
V. ordalii	FS-144	GCF_000287115.1	617134	Seawater	01.04.2006	Filtered seawater	USA: Massachusetts	Cordero et al., 2012	22955834
V. splendidus	12E03	GCF_000272105.1	1191305	Seawater	37712	Seawater	USA: Massachusetts	Cordero et al., 2012	22955834
V. splendidus	12F01	GCA_000256485.1	530557	Seawater	01.04.2006	Seawater	USA: Massachusetts	Shapiro et al., 2012	22491847
V. splendidus	1S-124	GCF_000272305.1	1191313	Seawater	38808	Filtered seawater	USA: Massachusetts	Cordero et al., 2012	22955834
V. splendidus	ZF-90	GCF_000272125.1	617147	Seawater	01.04.2006	Filtered seawater	USA: Massachusetts	Cordero et al., 2012	22955834
V. vulnificus	99-578 DP-B1	GCA_000788325.1	672	Oyster	1998	Environmental strain	USA: Loisiana	Phillips et al., 2015	25593245
V. vulnificus	99-796 DP-E7	GCA_000788315.1	672	Oyster	20.06.1905	Environmental strain	USA: Florida	Phillips et al., 2015	25593245
V. vulnificus	ATCC 33147	GCA_000764895.1	672	Eel	29098	Diseased eel	Japan:	Tison et al., 1982	7138004
V. vulnificus	CMCP6	GCA_000039765.1	216895	Homo sapiens	25.06.1905	Clinical isolate	South Korean:Gwangju	Kim et al., 2003	14500463

For phylogenetic analysis eight additional strains without any genome sequences were included to get a broader map of the *Aliivibrio* clade. The housekeeping genes *gapA*, *gyrB*, *pyrH*, *recA*, *rpoA* and *16S rRNA* for these strains were obtained from NCBI GenBank® (Table 3). The strain N16961 of *Vibrio cholera* (O1 El tor) was used as the reference sequence in the phylogenetic part and thus only housekeeping loci were required.

Table 3. House-keeping genes obtained from the GenBank® database.

Species	Strain	Accession ID						Taxonomy ID	Isolation source	Time	Comments	Location	Reference	PubMed (PMID)
		16S	gapA	gyrB	pyrH	recA	rpoA							
A. sifiae	H1-1	AB464964.1	AB464977.1	AB464968.1	AB464972.1	AB464996.1	AB465000.1	566293	Seawater	2007	Surface seawater	Japan: Harumi Pier, Tokyo Bay	Yoshizawa et al., 2010	21282907
A. sifiae	H1-2	AB464965.1	AB464978.1	AB464969.1	AB464973.1	AB464997.1	AB465001.1	566293	Seawater	2007	Surface seawater	Japan: Harumi Pier, Tokyo Bay	Yoshizawa et al., 2010	21282907
A. thorii	MdR7	EU185839.1	EU185868.1	EU185897.1	EU185925.1	EU185948.1	EU185971.1	1087367	Seawater	-	Costal seawater	USA: Marina del Rey, California	Fidopiastis et al., 1998	9422593
A. thorii	SA5	EU185840.1	EU185869.1	EU185898.1	EU185926.1	EU185949.1	EU185972.1	326926	Sepiola affinis	34881	Light organ	France: Banyuls-sur-Mer	Urbanczyk et al. 2009	19481895
A. thorii	SA6	EU185841.1	EU185870.1	EU185899.1	EU185927.1	EU185950.1	EU185973.1	491936	Sepiola affinis	34881	Light organ	France: Banyuls-sur-Mer	Urbanczyk et al. 2009	19481895
A. logei	Kch1	FJ858206.1	JF342802.1	JF342803.1	JF342806.1	JF342804.1	JF342805.1	688	Myoxocephalus polyacanthocephalus (goby fish)	-	Intestine	Sea of Okhotsk (Kamchatka peninsula)	Khrulnova et al., 2009	-
V. cholerae	N16961	NR_074810	VC2000	VC0015	VC2258	VC0543	VC2571	243277	Homo sapiens	1971	Fecal	Bangladesh:	Heidelberg et al., 2000	10952301

Software

CLC genomics workbench (v8.0.3)

CLCgenomics workbench (CLC Bio-Qiagen, Aarhus, Denmark) is a multipurpose software solutions featuring workflow design, read mapping, de novo assembly, variant detection, RNA-Seq and tools for epigenomics among others. During this project the assembly processes, creation of local BLAST databases and local BLAST searches were performed with this software.

Web page: https://www.qiagenbioinformatics.com/

ABACAS: algorithm-based automatic contiguation of assembled sequences (v1.3.1)

ABACAS (Assefa et al., 2009) is a Perl based script for automatic orienting, aligning and ordering of assembled contigs to a reference sequence. Providing an easier comparison of syntenic elements. This script was applied to all assembled genomes with a closed reference in this project.

Web page: http://abacas.sourceforge.net/index.html

Unipro UGENE (v1.19.0)

Unipro UGENE (Okonechnikov et al., 2012) is an open-source software package for analyzing various bioinformatic data. The software core contains many tools for basic sequence analysis but also integrates with external modules to perform aligning of sequences, RNA-seq analysis and read assembly with more. UGENE was important in the managing and preparation of multiple sequence alignments in this project.

Web page: http://ugene.net/

MEGA: Molecular Evolutionary Genetics Analysis (v6.0)

MEGA (Tamura et al., 2013) performs sequence alignment and phylogenetic analysis on datasets. It is integrated with both ClustalW and MUSCLE for the alignment module while the analysis part of the program infers phylogeny, performs model estimates and analysis based on molecular evolution. The contribution of MEGA in this project was to perform the model testing of alignments and infer the corresponding phylogenetic relationship using the maximum likelihood method.

Web page: http://megasoftware.net/

SplitsTree4 (version)

SplitsTree4 (Huson and Bryant, 2006) is a software for calculating unrooted phylogenetic networks based on aligned sequence data. By using split decomposition and neighbor-net among others, the program represents taxa by nodes and evolutionary relationships by edges. SplitsTree4 was applied in this project to amplify the visualization of diverging species.

Web page: http://www.splitstree.org/

Artemis (V16.0.0)

Artemis (Carver et al., 2008) is a genome browser and an editing tool for annotation of DNA and protein sequences. It was applied for visually inspection of genome content, frame shift mutations in particular as well as obtaining basic statistical values for this project.

Web page: http://www.sanger.ac.uk/science/tools/artemis

BioEdit (V7.2.5)

BioEdit (Hall, 1999) is a sequence and annotation editing program. It also provides built inn algorithms for alignment and statistical computations. BioEdit came to use in this project when investigating pair wise sequence identities between homologous genes.

Web page: http://www.mbio.ncsu.edu/bioedit/page2.html

Get_homologues (v1.4)

Get homologues (Contreras-Moreira and Vinuesa, 2013) is a standalone package for performing homolog sequence clustering and analysis. It is integrated with BLASTall and OrthoMCL as well as two additional clustering algorithms to perform pan-genome analysis. Pan and core-genome analysis with the acquisition of cluster files in addition to statistical analysis were obtained for this project with this software.

Web page: https://github.com/eead-csic-compbio/get_homologues/releases

R (v3.2.3)

R (https://www.r-project.org/) is the base program and scripting language for a wide range of modules provided by the public. R was used in this project to visualize the contents of distance matrices as violin plots.

Web-services and databases

GenBank(®) database

The GenBank(®) (Clark et al., 2016) is a publicly open nucleotide database which is continually updated and synchronized with both DDBJ and ENA. This database was used to acquire some of the draft and closed genomes as well as individual gene sequences for this project.

Web page: http://www.ncbi.nlm.nih.gov/genbank/

Galaxy web server (V16.01)

The Galaxy free web-based service (Goecks et al., 2010) is a collection of tools to perform various bioinformatics tasks. The user is provided with 250Gb storage for projects and dozens of connectable tools for the workflow manager to create pipelines. In this project Galaxy was utilized for managing the concatenation and alignment process of sequence elements in a piped process.

Web page: https://usegalaxy.org/

ClustalW: Clustal weigth (V2.01)

ClustalW (Thompson et al., 2002) is a alignment algorithm for datasets of multiple sequences. The implementation of ClustalW in Galaxy was used to perform the major alignment procedure of the phylogenetic analysis during this project.

Web page: http://www.clustal.org/clustal2/

RAST: Rapid Annotation using Subsystem Technology (V2.0) – FIGfam (release v.70)

RAST (Aziz et al., 2008) is a gene prediction and annotation service available for bacteria and archaea. The RAST server was used in this project to annotate 45 Aliivibrio draft or closed genomes in order to obtain similarly annotation.

Web page: http://rast.nmpdr.org/rast.cgi

CRISPRfinder (update 2014-08-05)

CRISPRfinder (Grissa et al., 2007b) is a web-service for identification of direct repeat sequences. It provides a rich output of spacers and repeats and was used in this project as a first step to locate repeat positions, acquire spacer sequences and determine system type.

Web page: http://crispr.u-psud.fr/Server/CRISPRfinder.php

CRISPRmap (v1.3.0-2013)

The CRISPRmap (Lange et al., 2013) is a web-service for identifying sequence family and secondary structure of direct repeats. On the basis of this information it predicts the affiliation to the most likely CRISPR system. During this project the web service was used as the second step in determining the presence of CRISPR systems based on repeats found by CRISPRfinder.

Web page: http://rna.informatik.uni-freiburg.de/CRISPRmap/Input.jsp

BlastKOALA (v2.1, update of March 4, 2016)

BlastKOALA (Kanehisa et al., 2016) is an annotation tool associated with KEGG who assigns K-numbers to submitted protein sequences. In this project BlastKOALA was applied on the pan-genome of *Aliivibrio* to deduce the protein family contents of the genus.

Web page: http://www.kegg.jp/blastkoala/

Equipment and hardware

ICE2 computer cluster

Work performed with CLC genomics workbench was performed on the local server cluster ice2 (ice2.cs.uit.no). It operates under Linux with 10 nodes utilizing a total of 40CPUs, 320GB DRAM and 40TB hard disc capacity.

Web page:
https://uit.no/forskning/forskningsgrupper/sub?p_document_id=347053&sub_id=356799

Method

Assembly and Genome Annotation

Sequence assembly

Locally sequenced strains, shown in Table 1, belonged all to the *Aliivibrio* genus and were assembled with the CLC genomics workbench package (CLC) (http://www.clcbio.com). All raw data from sequenced strains (fastq.gz files) were initially stored on our local computer cluster ice2, which also act as the platform for running CLC. The sequence read files were then imported as paired-end for each strain. Options for the import were set to remove failed reads while the quality scores scheme of *NCBI/Sanger or Illumina Pipeline 1.8 and later* was applied. De novo assemblies of the imported paired reads were then executed with parameters adjusted to auto-detection of paired distances and to perform scaffolding with a minimum contig length of 500bp. We also applied the mapping options where reads were mapped back to contigs with the *update contigs* feature checked, other parameters remained as default. The completed assemblies were exported as multi-fasta files before either first being mapped to a gold standard reference or annotated.

Contig mapping

Mapping of contigs was performed on *A. salmonicida* or *A. wodanis* strains as well as one strain of *A. fischeri* due to the availability of closed genomes for these species. Full length nucleotide sequences of concatenated and ordered chromosomes with appending plasmids from *A. salmonicida* LFI1238 and *A. wodanis* 0609139 (see Table 2) were used with ABACAS (Assefa et al., 2009). A total of 14 *A. salmonicida* drafts and 11 *A. wodanis* drafts were mapped against their respective reference. Parameters applied with ABACAS were set to execute mapping with nucmer and to print the ordered contigs in a multi-fasta file. Unmapped contigs were written to a separate multi-fasta file which was appended to the mapped contigs. The same procedure was executed for the strain ZF-211 of *A. fischeri* mapped against the reference strain ES114.

Annotation

Annotation of the 45 *Aliivibrio* strains used in the pan-genome analysis was performed using RAST (Aziz et al., 2008). Seven strains were omitted due to low assembly quality and to reduce CPU work load in the upcoming pan-genome analysis. These were *A. salmonicida* strain 12, 378 and 574 while strain Vw5, Vw7, Vw27 and Vw29 were omitted from the *A.*

wodanis genus. Remaining genomes constituted all available *Aliivibrio* strains both locally sequenced and published. Individual genomes were uploaded to the RAST server for annotation. The RAST annotation parameters were set to use GLIMMER-3 for gene prediction, applying the FIGfam version 70. The RAST server was also set to build metabolic models.

Inferring the Phylogenetic relationship of *Aliivibrio*, *Vibrio* and *Photobacteria*

Phylogenetic Design

The multilocus sequence analysis (MLSA) were performed according to the work of Sawabe and collaborators, applying 6 of 9 housekeepong genes from this publication (Sawabe et al., 2007). These included *16S rRNA*, *gapA* (glyceraldehyde-3-phosphate dehydrogenase, subunit α), *gyrB* (DNA gyrase, subunit β), *pyrH* (Uridylate kinase), *recA* (recombinase, subunit α) and *rpoA* (DNA-directed RNA polymerase, subunit α) which were concatenated using the same sub selection of each locus as described by Sawabe and collaborators (Sawabe et al., 2007).

Obtaining the multilocus sequences

The strain *Vibrio cholerae* O1 biovar El Tor str. N16961 was used as reference for the six nucleotide sequences; *16S rRNA*, *gapA*, *gyrB*, *pyrH*, *recA* and *rpoA*. These were obtained from the GenBank database with accession or locus tags; NR_074810, VC2000, VC0015, VC2258, VC0543 and VC2571 respectively. These sequences were stored locally on our computer cluster ice2 (ice.cs.uit.no) and imported to CLC where they became integrated in a local BLAST database. The *V. cholera* reference sequences were queried against sequenced and downloaded genomes, shown in Table 1 and Table 2, using the BLASTN implemented in CLC. Hits against these six sequences were stored in multi-fasta format representing each individual locus where the complete aggregates contained 81 nucleotide sequences for each gene. In addition, the six sequences from strains not available as drafts or as closed genomes were also downloaded from the NCBI database (see Table 3). The headers of each sequence in the multi-fasta files were then organized to represent the four letter gene symbol, genus or specie name and finally the strain ID, all separated by the pipe symbol "|".

Alignment construction

In preparation for the alignment process all IUPAC nucleotide ambiguity codes where replaced by N's by performing searches with the regular expression [^-AGTCN]. Missing data in any of the collected sequences caused by gap or short sequence products were defined as "?" signs. Each individual gene, represented by its multi-fasta file, was then loaded into UGENE (Okonechnikov et al., 2012). Each sequence was manually adjusted to match the following conserved motifs; tTGACGTT, AAgTGGg, GGtGtgCC, TaAAaGAacT, TtTAcGC and GAGCC in *16S rRNA*, *gapA*, *gyrB*, *pyrH*, *recA* and *rpoA* respectively. Here, majuscule characters represent conserved sites and lowercase represents non-conserved sites. These motifs were close to the starting position of each applied loci regions as described by Sawabe (Sawabe et al., 2007). The applied regions (see Table 6) of each gene locus were then cut and exported in multi-fasta format using the function *Save subalignment* in UGENE.

The full size MLSA alignment was concatenated based on all six sub-selections and aligned with the galaxy web-service (Goecks et al., 2010) in a piped process applying ClustalW2 (Thompson et al., 2002). The pipeline constructed for this specific task was published under the name "6L - MLSA merger & aligner".

Model testing of the concatenated MLSA design was performed with MEGA v6 (Tamura et al., 2013). MEGA reported the lowest Bayesian Information Criterion (BIC) for the General Time Reversible model with site rates being Gamma distributed (G) and having Invariant positions (I), indicating this as the best fitting model. A phylogenetic Maximum likelihood (ML) tree was then constructed on the basis of this report. Parameters were set to apply the GTR+G+I substitution model with the default number of 5 discrete gamma categories and *Complete deletion* as *Gaps/Missing Data Treatment*. Test of phylogeny was additionally set to perform 100 bootstrap replications. The same procedure with model testing and tree construction was also carried out applying all sites of the concatenated dataset.

A splitstree network was generated from the MLSA dataset using SplitsTree4 (Huson and Bryant, 2006). Parameters were set to apply the Jukes-Cantor model with the NeighborNet method applying position filtering to exclude gapped sites and allow 0% missing data per site. This was performed to simulate complete deletion by MEGA to obtain comparable ML trees and split-networks. *Vibrio*, *Photobacteria* and the outgroup, *Photorabdus*, were filtered out in the final network analysis to focus on the *Aliivibrio* genus.

Finding the discriminating power of individual gene loci

Each locus sub-selection saved from UGENE was individually aligned applying ClustalW2 implemented in the Galaxy web-service. Finished alignments were then analyzed by MEGA to obtain general statistics as well as obtaining the best substitution models for the ML method (see Table 7). Use of all sites and complete deletion as gap/missing data treatment were applied to construct the trees. Test of phylogeny was set to perform 100 bootstrap replications in constructed trees. Additionally, BioEdit (Hall, 1999) was applied to obtain identity matrixes of the final sequence alignments including the concatenated MLSA. These were corrected by removing half of the matrices including the diagonal values. The matrices values were then converted to individual vectors in R (https://www.r-project.org/) and joined as single data frames representing each locus and the MLSA. A violin-box plot were then created on the these distances applying the ggplot2 (Wickham, 2009) package for R. See R code for violin-box plots in the appendix for console lines used.

Phylogenetic reconstruction of quorum sensing genes

The *ainS* and *ainR* gene sequences were retrieved from their respective clusters generated by Get_homologues (described below). The two gene clusters were then individually aligned with ClutalW. Model testing was carried out with MEGA to find the best substitution model for reconstruction based the ML method using all sites (see Table 7). In the phylogenetic tree construction robustness was tested with 100 bootstrap replications.

Pan-genome analysis of the genus *Aliivibrio*

Homologous gene Clustering

GenBank files based on RAST annotations of 45 *Aliivibrio* genomes (Table 4) were included in the clustering process with Get_homologues (Contreras-Moreira and Vinuesa, 2013). Parameters for the clustering process were set to apply the OrthoMCL (Li et al., 2003) algorithm (-M) on amino acid sequences. The algorithm was also configured to request all generated clusters (-t 0), perform genome composition analysis on 10 replications (-c), generate identity matrix along with the output (-A) when engaging four processor cores (-n 4). The estimated cluster size of both the pan-genome and core-genome were calculated from the generated core- and pan- tab-files created. The analysis was executed using the *plot_pancore_matrix.pl* script applying *core_Tettelin* (Tettelin et al., 2005) and *pan* parameters separately. A pangenome matrix was additionally produced from the data by applying the *compare_clusters.pl* script, addressing the -m flag. This matrix was transposed and analyzed manually to determine the number of genes representing the core, accessory and unique genes of the pan genome. The compare_clusters.pl script was run again to additionally focus on clusters being syntenic by adding the "-s" flag to its parameters.

Multi-fasta files representing the amino acid sequences of the core, accessory and unique parts of the pan-genome were created applying the first sequence of each cluster. Each of the three pan-genomes were uploaded and separately annotated with the BlastKOALA service (Kanehisa et al., 2016). The taxonomy group was set to Bacteria and the KEGG GENES database set to apply genus_prokaryotes for K-number assignment. The resulting K-numbers (Kegg Ontology identifiers) were finally downloaded and applied to the KEGG Reconstruct BRITE mapping tool (Kanehisa, 2016).

An additional clustering run was executed on the same samples, this time clustering the nucleotide sequences by addressing the "-a" flag in *get_homologues.pl*. The resulting nucleotide identity matrix from this run and the previously obtained amino acid matrix were analyzed in R to generate violin plots (see Finding the discriminating power of individual gene loci above).

Mapping of Quorum Sensing Genes

The lux operon and related genes

Genes related to quorum sensing and the lux operon was searched for by applying the CLC genomics workbench on the RAST annotated genomes (see Table 4). The five systems comprising *varS/varA-Csr*, *luxI/luxR*, *luxM/luxN*, *luxS/luxPQ* and *ainS/ainR* were mapped by manually searching for the syntenic positions of their related gene loci. Any frame shift discrepancies were examined by Artemis (Carver et al., 2008) if the same gene appeared annotated more than once in a row. Additional genes coding *Fis*, *Hfq* and *LitR* with their relative neighborhood were also accounted for. Both direction and placement relative to neighboring and conserved flanking genes were noted and managed. Contigs was also evaluated and marked if ends were appearing in close vicinity or inside of the systems or genes. Finally, a consensus array of the gene synteny of each species was drawn for comparison.

Three phylogenetic trees were constructed for the genes coding *ainR* and *ainS* as described in Phylogenetic reconstruction of quorum sensing genes.

Sequences identities between of the *luxR* and *luxR2* loci were obtained by selecting the *luxR* cluster that had neighboring genes were *luxG*, *luxI*, *mscS* or *luxC*. These were all located in a single cluster comprising 50 sequences generated in the pan-genome analysis. Sequences of strains not having both *luxR* loci were removed before alignment with ClustalW2. The final values were then achieved by creating an identity matrix in BioEdit.. The average pair-wise sequence identities within and between *luxR1* and *luxR2* were measured on the basis of all strains of *A. salmonicida, A. logei and A. finisterrensis* individually.

Identification of CRISPR repeats and cas genes

Identification of active CRISPR systems

Discovery of confirmed CRISPR arrays was performed by providing CRISPRfinder (Grissa et al., 2007b) with each of the 45 *Aliivibrio* genomes as contigs. Each confirmed CRISPR hit by the web service was saved and the consensus repeat sequences were collected in a multi-fasta file. CLC was then utilized to manually investigate the regions containing CRISPR arrays and their neighboring genes. This was performed on the RAST-annotated GenBank files as well as searching for other CRISPR related genes. Annotations related to CRISPER systems were graphically mapped focusing on the syntenic relationship of neighboring loci both upstream and downstream. Finally, the consensus direct repeats previously collected were uploaded to CRISPRmap (Lange et al., 2013) utilizing the 1.3.0-2013 version to determine sequence and secondary structure affiliation.

Results

Being able to perform pan-genome analysis and further component analysis requires a minimum of annotated draft genomes as input. To achieve this we performed assembly of sequenced reads for each strain to obtain contigs. By using ABACAS (Assefa et al., 2009) these contigs were mapped to gold standard references obtained from the GenBank database to improve synteny and sequence correlation between the strains. In order to attain similar annotations, all genomes were processed with the RAST pipeline (Aziz et al., 2008), which also provided GenBank files for further pan-genome analysis.

Assembly and annotation overview

Sequence assembly

On average we provided the assembly process with 3M reads, but obtained 450K reads with less successful sequencing runs. The best and the most productive sequencing run reached a peak of 8.4M reads. As seen in Table 4, these extremes were both observed in *A. salmonicida* for strain 378 and R8-70 respectively. The percentage of reads able to map into contigs did not fall below 95.88% nor exceed 99.5%. No trends in the mapping percentages were found in relation to species or read numbers for the assembled genomes.

Table 4. General statistics of assembled and annotated genomes with CLC genomics workbench and RAST. Missing data in Genes and RNA represents genomes not annotated by the RAST service.

Strain	Filtered reads	Matched reads		Bases in matched reads	Contigs obtained	Average contig length	Bases in contigs	Computed coverage	Genes	RNA
DSM 23419	2 618 850	2 510 921	95.88 %	680 096 591	223	16 792	3 744 825	182	3 424	84
ES114	-	-	-	-	3	-	4 273 718	-	3 883	156
MJ11	-	-	-	-	3	-	4 503 336	-	4 058	152
SR5	-	-	-	-	2	-	4 273 614	-	3 707	106
ZF-211	-	-	-	-	197	-	4 037 340	-	3 665	22
SA12	2 920 764	2 803 772	95.99 %	754 277 903	244	17 369	4 238 275	178	3 862	91
SR6	2 828 680	2 755 424	97.41 %	702 849 717	413	10 507	4 339 648	162	3 887	85
5S-186	-	-	-	-	162	-	4 451 230	-	4 037	20
A11-3	3 969 498	3 925 005	98.88 %	909 683 056	48	97 851	4 696 856	194	4 260	92
ATCC 29985	3 209 922	3 085 946	96.14 %	806 137 385	333	13 469	4 485 327	180	4 101	85
MR17-66	2 876 440	2 793 607	97.12 %	746 678 809	338	13 553	4 581 156	163	4 136	83
MR17-77	2 960 184	2 928 622	98.93 %	670 333 406	90	51 119	4 600 717	146	4 205	85
MR17-80	2 834 806	2 741 328	96.70 %	723 327 359	452	10 208	4 614 459	157	4 199	75
SES03-1	3 589 448	3 495 760	97.39 %	941 017 168	341	13 943	4 754 736	198	4 328	86
SES03-5	2 873 828	2 758 100	95.97 %	754 922 564	326	14 573	4 750 909	159	4 343	87
R8-63	2 124 774	2 107 414	99.18 %	474 478 948	107	39 012	4 174 311	114	3 883	89
R8-67	3 738 482	3 611 306	96.60 %	956 470 527	285	14 396	4 103 025	233	3 885	98
A25	1 971 994	1 951 426	98.96 %	455 896 316	40	108 080	4 323 200	105	4 004	82
A9-1	1 625 408	1 604 789	98.73 %	363 569 779	75	57 494	4 312 105	84	3 976	81
A9-2	1 511 080	1 495 669	98.98 %	335 960 977	47	91 684	4 309 175	78	3 972	78
A11-1	1 650 696	1 623 974	98.38 %	376 294 065	42	103 609	4 351 615	86	4 021	86
A15	1 666 286	1 640 204	98.43 %	328 067 098	83	55 226	4 583 781	72	4 205	91
A22	1 262 154	1 251 241	99.14 %	210 124 393	144	31 751	4 572 249	46	4 168	85

2208-14	3 865 466	3 816 172	98.72 %	964 487 865	132	34 323	4 530 751	213	4 185	98
LFI1238	-	-	-	-	6	-	4 655 660	-	4 455	141
12	627 924	620 138	98.76 %	130 238 620	761	5 344	4 067 399	32	-	-
250	1 662 992	1 652 251	99.35 %	274 108 788	353	11 987	4 231 421	65	4 178	82
289	2 191 742	2 179 930	99.46 %	349 141 440	327	12 413	4 059 163	86	3 988	82
378	448 642	442 828	98.70 %	92 526 570	773	5 251	4 059 434	23	-	-
554	3 432 062	3 363 595	98.01 %	581 345 158	265	15 126	4 008 463	145	4 009	88
561	3 742 046	3 689 951	98.61 %	737 682 734	271	15 210	4 122 112	179	3 989	81
574	491 288	481 779	98.06 %	94 468 878	1 407	2 747	3 865 059	24	-	-
B9-15	3 491 598	3 393 554	97.19 %	922 378 370	390	12 477	4 866 404	190	4 505	91
LFI-180	2 989 000	2 973 948	99.50 %	786 357 401	274	15 075	4 130 753	190	4 144	80
N5541	3 432 062	3 363 595	98.01 %	581 345 158	265	15 126	4 008 463	145	3 897	80
R5-43	3 033 730	2 934 273	96.72 %	817 773 456	475	10 552	5 012 226	163	4 650	78
R8-68	3 570 524	3 542 567	99.22 %	794 230 899	258	15 750	4 063 612	195	3 967	79
R8-70	8 428 252	8 359 012	99.18 %	1 648 032 419	255	16 058	4 094 855	402	4 086	88
TEO	2 122 276	2 107 634	99.31 %	350 434 044	332	12 199	4 050 321	87	4 017	81
26449	2 520 196	2 421 807	96.10 %	659 547 230	410	11 370	4 661 838	141	4 264	88
609139	-	-	-	-	6	-	4 635 126	-	4 257	110
Vw1	2 500 528	2 481 250	99.23 %	656 839 453	159	28 804	4 579 782	143	4 150	100
Vw11	2 978 336	2 883 129	96.80 %	671 278 186	240	18 688	4 485 162	150	4 148	104
Vw12	5 210 276	5 158 842	99.01 %	1 395 442 144	212	23 241	4 926 987	283	4 600	94
Vw130426	4 498 266	4 454 192	99.02 %	1 159 007 726	381	12 009	4 575 394	253	4 186	80
Vw27	4 438 290	4 391 246	98.94 %	1 181 687 817	870	5 308	4 617 778	256	-	-
Vw29	3 359 090	3 292 929	98.03 %	874 616 401	590	7 688	4 535 631	193	-	-
Vw35	4 583 280	4 549 444	99.26 %	1 200 665 454	103	45 609	4 697 770	256	4 385	104
Vw5	4 754 230	4 702 084	98.90 %	1 262 642 841	602	7 461	4 491 639	281	-	-
Vw7	4 377 868	4 333 031	98.98 %	1 142 870 279	569	7 935	4 514 860	253	-	-
Vw8	4 719 808	4 647 981	98.48 %	1 169 537 373	251	17 922	4 498 377	260	4 144	96
VwK7F1	4 728 680	4 663 586	98.62 %	1 183 717 283	415	10 551	4 378 839	270	3 916	96

The assembled draft genomes varied in size from 3.7Mb to 5.0Mb where *A. finisterrensis* ranged as the physically smallest having the smallest chromosome. The environmental strain R5-43 of *A. salmonicida* was found to be the physically largest with all other *Allivibrio* genomes having smaller chromosomes. With *A. modi* strain A25 we obtained the most complete draft genome defined by only 40 contigs. The median regarding the achieved number of contigs describing the chromosome size was 273 in overall. *A. salmonicida* contributed greatly to this number by reaching no less than 255 contigs observed for the R8-70 strain. Strains of most other species were not following any trend and were seen with both high and low contig numbers. Considering the coverage of completed assemblies we observed a correlation between the initial number of reads, the percentage of mapped reads and the average contig length. Again we discover strain 378 and R8-70 of *A. salmonicida* sharing the extremes of having the lowest (23x) and highest (402x) coverage, putting other strains in between at an average of 165x. Still, these two trains only vary in size by 35Kb.

Contig mapping of *A. salmonicida*, *A. wodanis* and *A. fischeri*

The mapping result of draft genome contigs against a reference sequence varied greatly based on species and between strains of the same species. Most contigs of *A. salmonicida* strains were easily mapped on LFI1238 by ABACAS, as shown in Table 5, and only a few did not match the criterion by the nucmer algorithm. Exceptions were observed for the environmental strains of B9-15 and R5-43 where only 45% and 49% of all contigs became mapped. Excluding the environmental strains of *A. salmonicida*, ABACAS did on average map 98% of contigs while on *A. wodanis* the same average number of contigs mapped were only 67%. But strains within *A. wodanis* also varied as seen for Vw35 and Vw5 which had the lowest (38%) and the highest percentage (96%) of mapped contigs respectively.

Table 5. The number of contigs mapped with ABACAS for each given strain against its listed reference. Strains marked "*" were mapped in a later stage and were not in a mapped condition during the pan-genome analysis.

	Contigs	
Strain	**Mapped**	**Unmapped**
***A. salmonicida* LFI1238**		
12	752 (98.8 %)	9
250	334 (94.6 %)	19
289	327 (100.0 %)	0
387	767 (99.2 %)	6
554	250 (97.7 %)	6
N5541	261 (98.5 %)	4
561	263 (97.0 %)	8
574	1 401 (99.6 %)	6
LFI-180	262 (95.6 %)	12
R8-68	258 (100.0 %)	0
R8-70	251 (98.4 %)	4
TEO	331 (99.7 %)	1
B9-15*	175 (44.9 %)	215
R5-43*	235 (49.5 %)	240
***A. fischeri* ES114**		
ZF-211	85 (43.1 %)	112
***A. wodanis* 0609139**		
Vw1	138 (86.8 %)	21
Vw11	102 (42.5 %)	138
Vw12	121 (57.1 %)	91
Vw130426	266 (69.8 %)	115
Vw27	641 (73.7 %)	229
Vw29	465 (78.8 %)	125
Vw35	39 (37.9 %)	64
Vw37	70 (45.8 %)	83
Vw5	580 (96.3 %)	22
Vw7	400 (70.3 %)	169
Vw8	137 (54.6 %)	114
VwK7F1	364 (87.7 %)	51

Annotation

The RAST annotation service provided equal prediction and annotation to genomes as summarized in Table 4. Here we observed the physically largest and smallest genomes of *A. salmonicida* R5-43 and *A. finisterrensis* DSM 23419 respectively, also had the fewest and most numerous predicted genes by 3424 and 4650. These extremes make the genus *Aliivibrio* differed with 1226 genes. Focusing on strains within the same species showed similar tendencies with highly varied numbers of genes predicted for both *A. salmonicida* and *A. wodanis*. In the former's two environmental strains we observed approximately 500 additionally predicted genes in comparison to the remaining disease strains. On the contrary *A. friggae*, *A. magni* and *A. modi* all shared a common tendency of having quite equal amounts of genes between their individual stains.

Compared to genes, the amount of RNA's discovered did not correlate with the genomes physical size but were rather centered on species. A few exceptions were observed for the strain ZF-211 of *A. fischeri* and 5S-186 of *A. logei* only possessing 22 and 20 RNA's compared to the average of 89. On the upper scale we found *A. salmonicida* LFI1238, *A. fischeri* ES114 and MJ11, all closed genomes, having roughly 50 additional RNA's on top of the average for same species.

Phylogenetic relationships based on MLSA

In an attempt to establish the phylogenetic relationship of *Aliivibrio*, *Vibrio* and *Photobacteria* we included a wide selection of strains. These mainly represented the *Aliivibrio* clade, but also in selections of *Vibrio* and *Photobacteria* to compare with. To achieve this we applied multi locus sequence analysis (MLSA) with familiar housekeeping genes and the *16S rRNA* locus. Analysis of each component loci in the MLSA; *16S rRNA*, *gapA*, *gyrB*, *pyrH*, *recA* and *rpoA* was also undertaken to understand their capability in distinguishing between our genera, species and strains.

Multilocus sequence analysis

The concatenated MLSA tree represented by Figure 2 was well supported by bootstrap values and successfully discriminate all three genera into their respective monophyletic clades. The *Vibrio* clade obtained a good support for its relative placement with a bootstrap value of 94 and separates from the *Aliivibrio* clade by a total branch length of 0.108. Still, the sub-clades of *V. harveyi* and *V. cholerae*/*V. furnissii* show questionable supportive values. Comparison with the MLSA tree applying all sites (see Appendix figure 1) reveal a slightly different placement of this clade. Superior support is obtained for all nodes representing the *A. fischeri*, *A. thorii*, *A. salmonicida* and *A. logei* clades for the *Aliivibrio*. These share the maximum possible support on higher level nodes with no less than a bootstrap value of 99. There are several less well supported nodes as well. These were within the range of 77 to 81 in the *Aliivibrio* clade and involved *A. finisterrensis*, the whole sub-clade comprising *A. wodanis*, *A. raniae*, *A. friggae* and *A. sifiae*. In addition, the same lower range wasseen in the sub-clade involving *A. magni*, *A. thrudae* and *A. modi* as well. Even less supported nodes are found below in the range of 44 to 67 and involves the individual species groups of *A. magni*, *A. thrudae*, *A. modi* and *A. sifiae*. Nodes at the lower levels (not shown in Figure 2) separating the individual strains of *A. salmonicida* share supportive values of zero or close to zero. The same uncertainty was also observed at seemingly random nodes within *A. logei* and *A. wodanis* where bootstrap values fall to 18 at lowest, but with most strains remain decently supported.

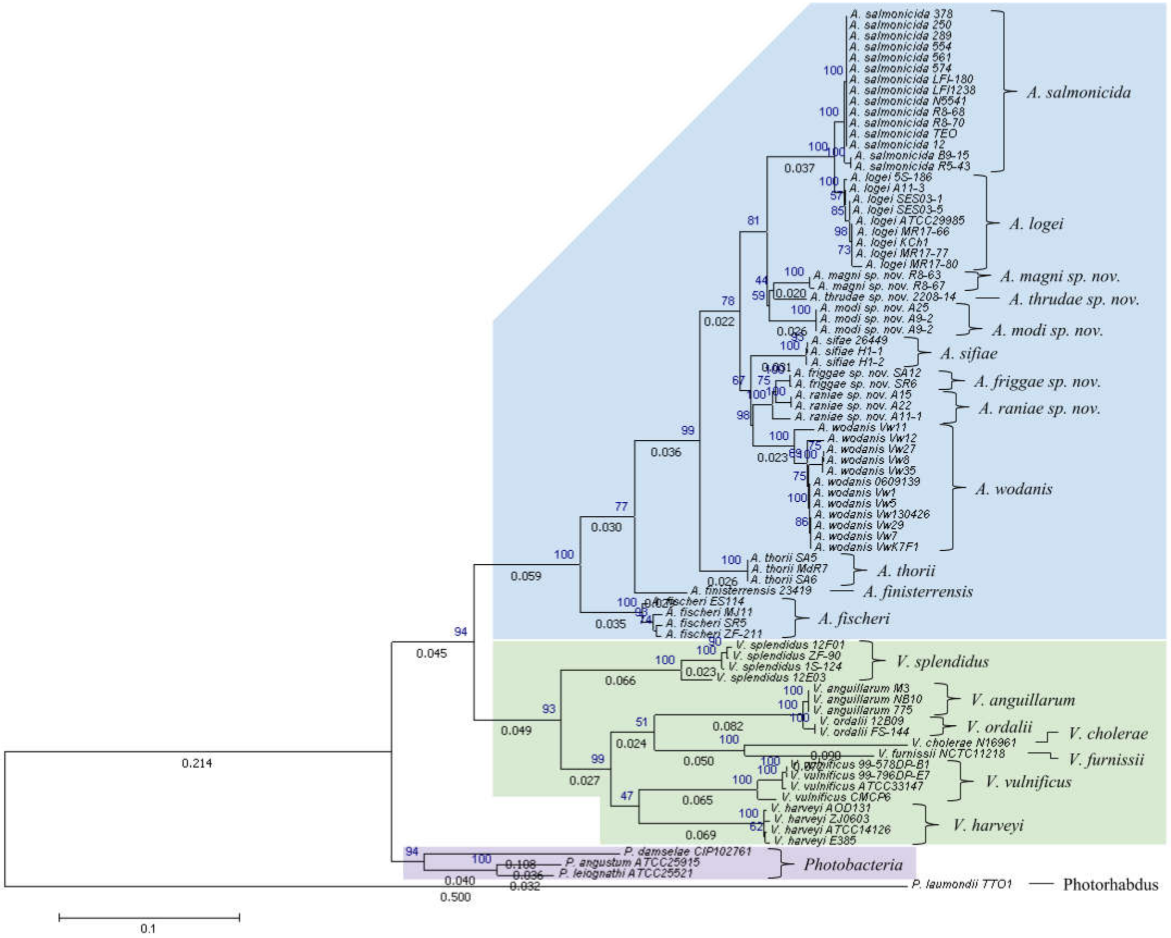


Figure 2. Concatenated maximum likelihood tree based on the six loci *16S rRNA*, *gapA*, *gyrB*, *pyrH*, *recA* and *rpoA*. The *Aliivibrio* genus is highlighted in light blue while the genera *Vibrio* and *Photobacteria* are shown in green and purple respectively.

Species groups containing two or more strains considered in the MLSA alignment differed in their intra-species identities ranging from 90.8 % to 100 %. Peak identities were among the *A. modi*, *V. anguillarum* and *V. ordalii* species showing no variation. Species within the 99 % layer were *A. salmonicida*, *A. thorii*, *A. magni* and *V. harveyi* which shared a minimal degree of variation with less than 0.5 % identity deviation between strains. Most other species kept high identities with little deviation, but in some cases the identities fell below 96 %. This was observed between strains of *A. sifiae* and *V. splendidus* but was only marginally below this value. A much clear breach was caused within the *A. logei* group by the strain Kch1, which failed to show more than 93.2 % identity towards any other strain within this species. Remaining strains of *A. logei* ranged between 96.9 % to being fully identical with an average at 97.7 %.

Network of *Aliivibrio*

The MLSA network diagram shown in Figure 3-A manages to separate all three genera and abides the same cladistic relationships as seen in the ML based tree. This happens despite being distance based and applying the Jukes-Cantor model. The *A. fischeri*, *A. finisterrensis* and *A. thorii* group's in Figure 3-B also matches with the ML tree as noticeably diverging from the remaining tree.

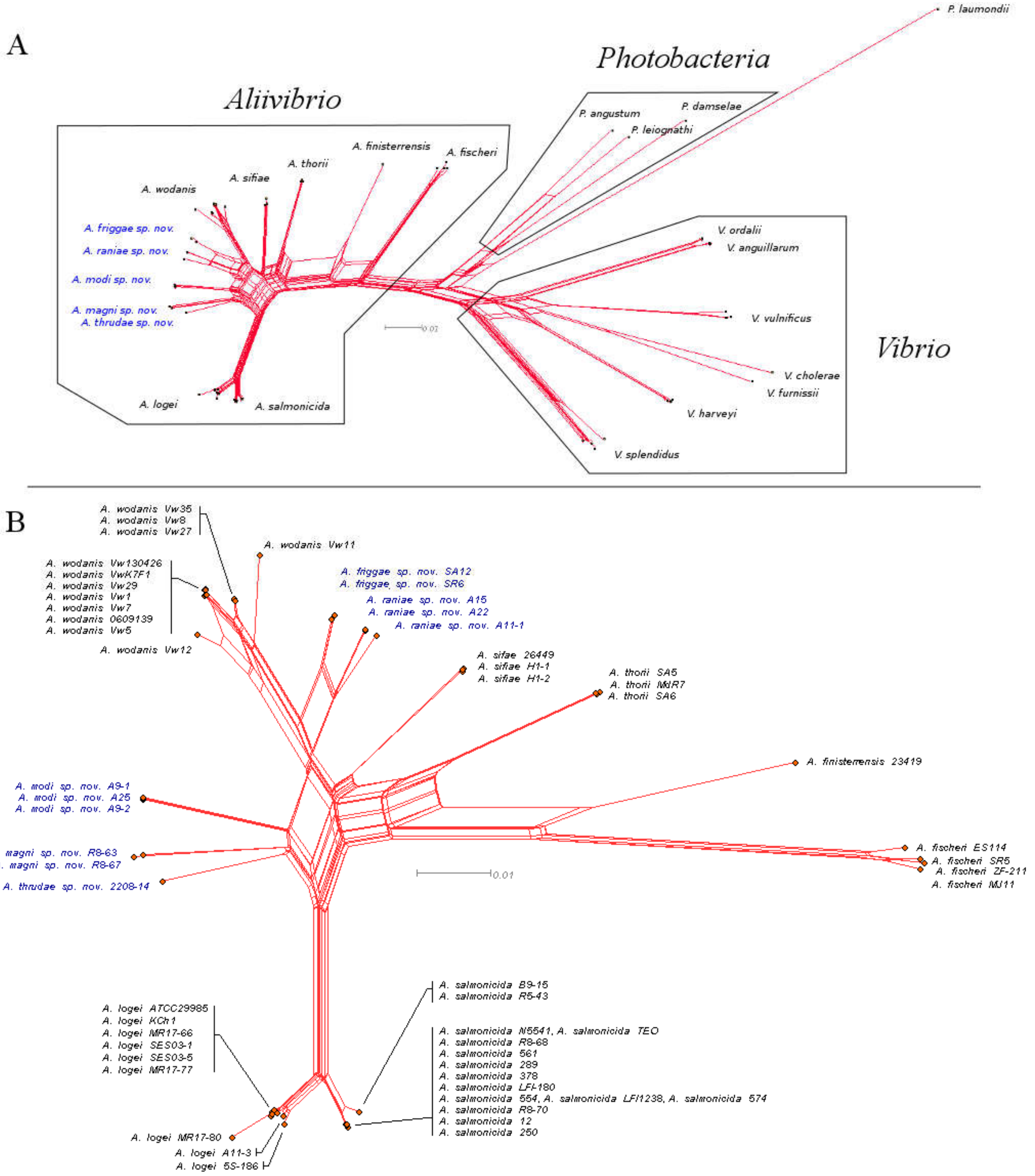


Figure 3. Network of all 81 included strains (A) and an individual network of the 58 *Aliivibrio* strains concerning 12 species (B). Lengths of edges and the scale bar are comparable to the ML tree. Species proposed as new are highlighted in blue.

The clades comprising *A. logei* and *A. salmonicida* in the network gives an explicit distinction of the strains known to be environmental samples (MR17-80, B9-15 and R5-43 in particular). All five species proposed as new are also clearly separated in the network as in the ML tree, but both methods were unable to group A11-1 of *A. raniae* with the remaining two strains. Focusing on the *A. wodanis* clade, which has the largest spread of leaves in the NeighborNet-based network, preserves the same grouping as in the ML tree. More prominent was the divergence of the two strains Vw11 and Vw12 with their edges poles apart in separate directions. The protruding edge of Vw11, and to a lesser degree the Vw35, Vw8 and Vw27 strains, shares some of the parallel trend which is observed in the right wing part of the network from *A. friggae* to *A. thorii* due an earlier splitting event.

General statistics of housekeeping genes

The sub-selections of genes applied in the MLSA corresponded to approximately 49% on the full sequence length of all six combined loci in *Vibrio cholerae* strain N16961. Each of the loci had the observed features as given in Table 6 with a total concatenated length of 3763bp. The alignment process by ClustalW2 (Thompson et al., 2002) further expanded the datasets total length to 3778 by introducing 15 gapped positions.

Table 6. Overview of general statistics before and after the alignment process of each applied locus and the concatenated MLSA design.

Gene	Full length (bp)	Sub-length (bp)	Average % GC, (μ ± σ)	Alignment clustalW (positions)			
				Alignment length	Gaps introduced	Variable	Missing data/gapped
16S rRNA	1535	626	52.33 ± 0.43	626	0	112 (17.9 %)	111
gapA	996	668	45.19 ± 0.85	671	3	269 (40.1 %)	172
gyrB	2418	590	42.20 ± 2.71	599	9	300 (50.1 %)	18
pyrH	732	445	46.85 ± 2.12	445	0	180 (40.4 %)	31
recA	1065	605	41.69 ± 2.24	605	0	264 (43.6 %)	84
rpoA	993	829	45.05 ± 0.87	832	3	283 (34.0 %)	27
MLSA	7739	3763	45.12 ± 1.38	3778	15	1408 (37.3 %)	442

The GC values appear averaged at ~45% in the concatenated MLSA dataset compared to the individual loci. Here, the *recA* and *gyrB* loci were observed with the lowest GC percentage, but had the widest standard deviations (σ) in this overview. Oppositely, the *16S rRNA* has GC content more than 10% higher with a much tighter standard deviation than any of the other loci considered.

The proportion of variable positions in the datasets also differs greatly among the loci. The *16S rRNA* clearly had the least variable positions, but also suffers from 111 missing data or gapped sites. Variable positions in the remaining loci ranged from 34% to 50.1% where *gapA* and *gyrB* were observed with the largest and lowest portion of missing data or gapped sites respectively.

Pair-wise identity values based on individual loci, illustrated in Figure 4, shows *16S rRNA* was most conserved, averaging at 96.1% with ~75% of all sample values above 95%. The three loci *gyrB*, *pyrH* and *recA* showed distribution patterns which appeared to have more balanced central tendency compared to *16S rRNA*, *gapA* and *rpoA*. Sample identities in both *gapA* and *rpoA* appear skewed, but in opposite directions: The *gapA* gene clusters below ~87% identity for half of its population while the opposite was observed in *rpoA* where the same population size was above 95%. The MLSA identities illustrate a more even distribution of samples but accumulate slightly more in the lower half. The main tendencies of the kernel plots, clearly seen in *pyrH*, are to split the data in three main sections. The lower section of the plots represents the between-genera identities while the mid section resembles between-species identities. Finally, the top section shows the inter-species identities. Based on these sections and the transition areas between them; *gyrB*, *pyrH*, *recA* and the MLSA datasets manages to clearly separate between the three levels. The *rpoA* locus easily separates the genera but has problems giving clear distinction between species and within species identities. The same trend is also evident in *gapA* and *16S rRNA*, but in the latter there is a visible degree

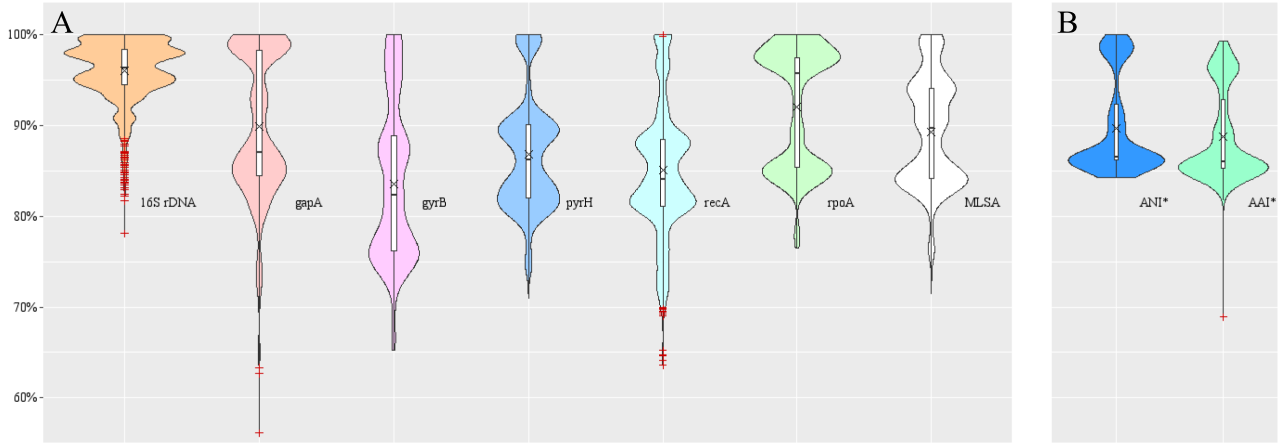


Figure 4. Violin plots (A) with overlaid box plots representing all paired identity values from each locus identity matrix including the MLSA. Medians are given by a black bar in the box plots, mean value by "X" and outliers by "+". Shape of the plots shows the estimated kernel probability density of the data. In (B) the average amino acid (AAI) and average nucleotide identity (ANI) matrixes are included in the same plot type which originates from the core genome (see Pan-genome analysis). *AAI and ANI values are based on 45 genomes only originating from the *Aliivibrio*.

of separation even if the identities are compressed in the upper layer. Outliers in *16S rRNA* in Figure 7 are mainly due to strain 5S-186 of *A. logei,* which contained missing data in position 451-472 and 809-893 of the applied loci section.

Overview of gene loci

In preparation of single loci phylogenetic analysis, model tests were performed with MEGA (Table 7) on the datasets and ranged from the simple K2 model to the complex GTR model. Here the T92 model (Tamura 92) was established to qualify for both the *gyrB* and *recA* loci independent of all sites or complete deletion were utilized. These loci also represent the lowest %GC values with the most variations in all of the aligned datasets examined. The K2 model (Kimura 2-parameter, also abbreviated K80) was only assigned the *16S rRNA* data in the case when missing data and gaps were removed, all other datasets qualified for the GTR model by the testing procedure.

Table 7. Model tests performed with MEGA v6 for each locus and the concatenated MLSA dataset as well as the *ainR* and *ainS* gene. Abbreviations: General Time Reversible (GTR), Kimura 2-parameter (K2), Tamura 3-parameter (T92), Bayesian information criterion (BIC).

MLSA	Missing data	Model	BIC
16SrDNA	Complete deletion	K2+G	4327.166
16SrDNA	Use all sites	GTR+G+I	6436.226
gapA	Complete deletion	GTR+G	6853.735
gapA	Use all sites	GTR+G	9698.239
gyrB	Complete deletion	T92+G+I	13787.557
gyrB	Use all sites	T92+G+I	13952.772
pyrH	Complete deletion	GTR+G+I	8891.958
pyrH	Use all sites	GTR+G+I	9471.702
recA	Complete deletion	T92+G+I	11426.727
recA	Use all sites	T92+G+I	12897.831
rpoA	Complete deletion	GTR+I	9689.732
rpoA	Use all sites	GTR+G	9904.456
MLSA	Complete deletion	GTR+G+I	49872.214
MLSA	Use all sites	GTR+G+I	57544.978

Lux genes	Missing data	Model	BIC
ainR	Use all sites	GTR+G	48738.586
ainS	Use all sites	GTR+G+I	22850.497

Both complete deletion and use of all sites manage to separate between the *Aliivibrio* and *Vibrio* genera when considering the inferred topology. This goes for most loci with the exception for *gapA* (see Appendix figure 2 and Appendix figure 3) and *16S rRNA,* the latter shown in Figure 5. Applying all sites with the *16S rRNA* locus managed to gain additional resolution and separate the out-group as well from the remaining tree. Using all sites, the *Aliivibrio* and *Vibrio* genera are separated with a combined branch length of 0.037 substitutions per site. But the inferred tree fails to give sufficient division between the *Vibrio* group and *Photobacteria* group. The tree is not well supported in general and is further reduced in the variant treated with complete deletion where the branch lengths also are sharply reduced when comparing the scales provided in Figure 5.

The resulting topology based on *gapA* and *recA* loci (see Appendix figure 2 and Appendix figure 3) manages to mix the *V. splendidus* group with the *Aliivibri* clade. This disturbs the *Aliivibrio* clade, separating *A. finisterrensis* and the *A. fischeri* group. These loci are also unable to distinguish the *Photobacteria* from the *Vibrio* clade properly. The *recA* gene was additionally incapable of clustering *P. damselae* with the remaining *Photobacteria*, making it appear as a second out-group in the tree. Bootstrap values are generally discovered to be low in the *Vibrio* group for these loci. When *gapA* is treated with complete deletion it improves supportiveness for some nodes including the *V. splendidus* branch. This node increases its bootstrap value from 28 to 84, but the main tendency within the tree was slightly reduced branch lengths. Branch lengths of *recA* increases with the same treatment but supportive values remain mostly unchanged.

The inferred tree based on *gyrB* easily discriminates the *Aliivibrio* from the *Vibrio*, but fails to separate the *Photobacteria* from the *Vibrio* clade. Bootstrap support was mainly kept high for nodes in the *Aliivibrio* clade, but several nodes amid the *Vibrio* and *Photobacteria* genera share questionable values, making their relative placement uncertain. Removal of missing data while applying complete deletion did slightly extend the overall branch lengths of the inferred tree based on *gyrB.* Supportiveness was unchanged.

Successful separation of all genera and the out-group was achieved when inferring the phylogenetic relationship of *pyrH* and *rpoA*. A total branch length of 0.106 and 0.164 substitutions per site separates the *Aliivibrio* from the *Vibrio* applying these genes. The *Photobacteria* was parted from these by a branch length of 0.149 and 0.103 respectively in the inferred trees. Only marginally shorter branch lengths resulted when complete deletion was applied, but there were slightly lower supportive values for the *pyrH* locus. Oppositely, an apparent increase in bootstrap values was observed for *rpoA* with the same treatment, but the tree performed poorly when attempting to define the *Vibrio* clade placement which swapped position with the *Photobacteria*.

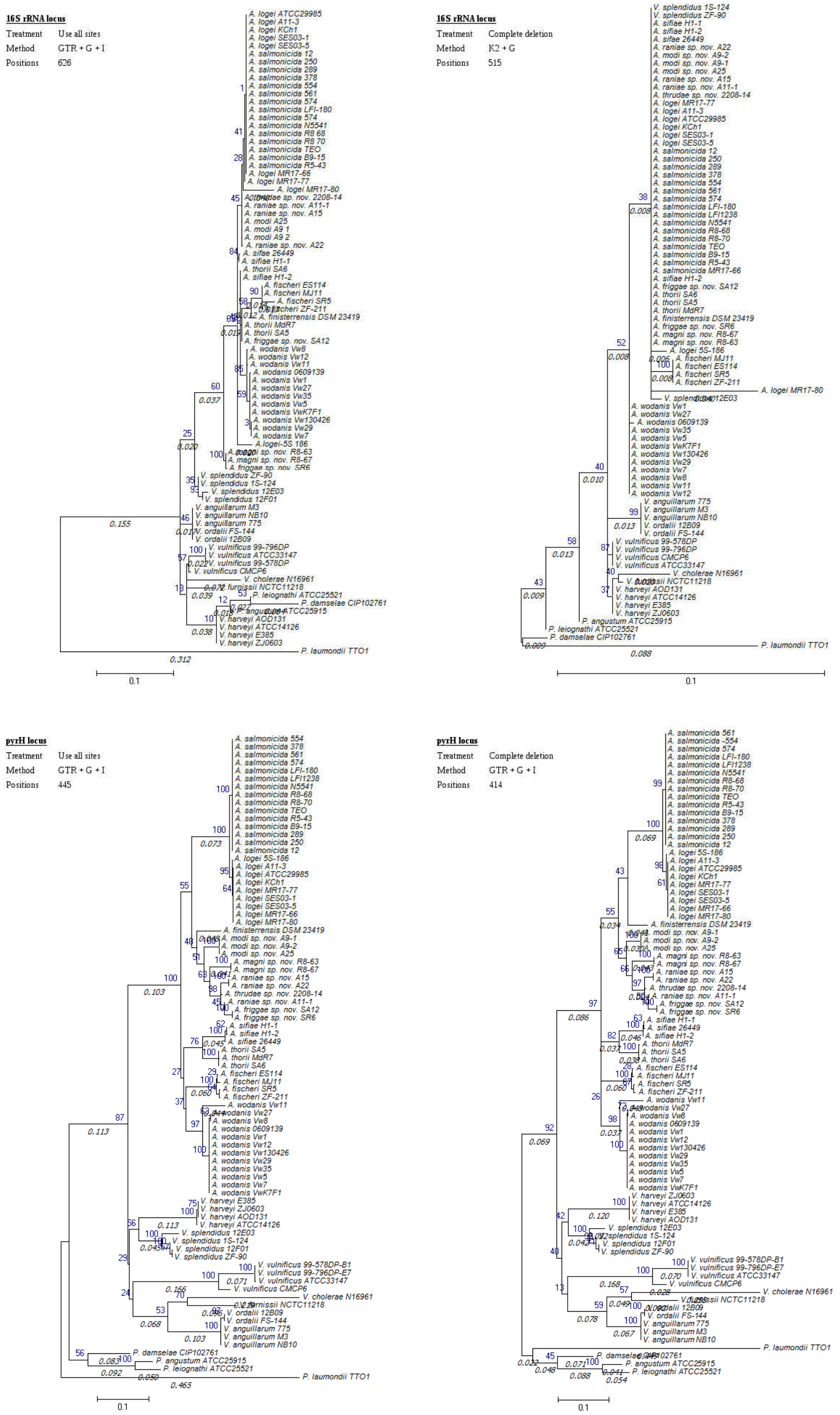


Figure 5. Phylogenetic inference of *16S rRNA* and *pyrH* based on 81 nucleotide sequences emphasizing the effect of applying all sites compared to complete deletion of missing data. Robustness shown in marine blue is based on 100 bootstrap replications.

Pan-genome analysis of 45 *Aliivibrio* genomes

Homologue gene clustering was performed on 45 strains within the *Aliivibrio* genera to help determine the diversity within this genus. Identifying the core, accessory and unique genes were performed with the Get_homologues script package (Contreras-Moreira and Vinuesa, 2013). This was performed to gain insight about preserved genes and pathways as well as understanding the flow of new genes. Further analysis by BlastKOALA (Kanehisa et al., 2016) attempted to discover biological inequalities between the core- accessory- and unique genome in terms of mapping to protein families in the BRITE database. We additionally obtained the average nucleotide identity (ANI) and average amino acid identities (AAI) of the core genome to identify the discriminatory power of these.

General statistics

Time consumption

Homologue genes clustering of 45 genomes ranging in size from 3424 to 4650 took 14.32 hours on an Intel Core i7-4800MQ laptop, engaging four of eight 2.70GHz cores. The second run performed on nucleotide sequences, also included BLASTall and the OrthoMCL algorithm but without resampling, amounted to 1.21 hours.

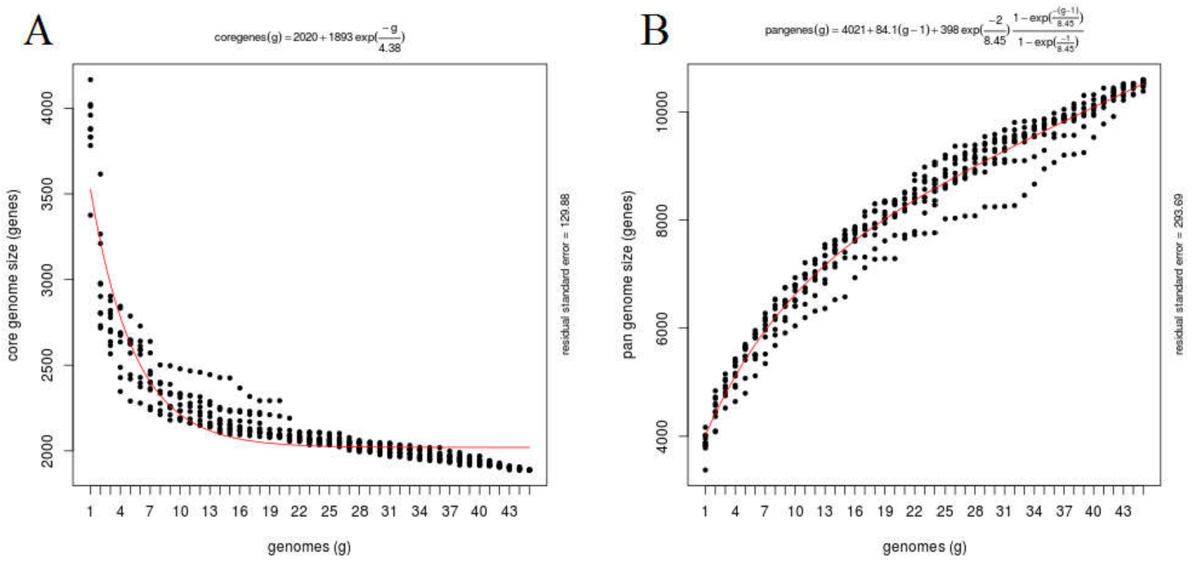


Figure 6. Estimations of the pan- and core-genome size based on 10x resampling. The convergence of the core-genome is illustrated in (A) while the extrapolation of the pan-genome shown in (B). Residual standard errors explain the goodness of fit for the functions against the resampled values.

Homologous clusters

A total of 183969 protein sequences were included in the pan-genome analysis and resulted in the formation of 15922 clusters when applying OrthoMCL. This was achieved with a minimum coverage requirement of 75% for pair wise BLAST alignments and an inflation value set to 1.5. The run performing homologous clustering of nucleotide sequences resulted in 20842 clusters being formed applying the same stringency on parameters.

Pan and core genome sizes

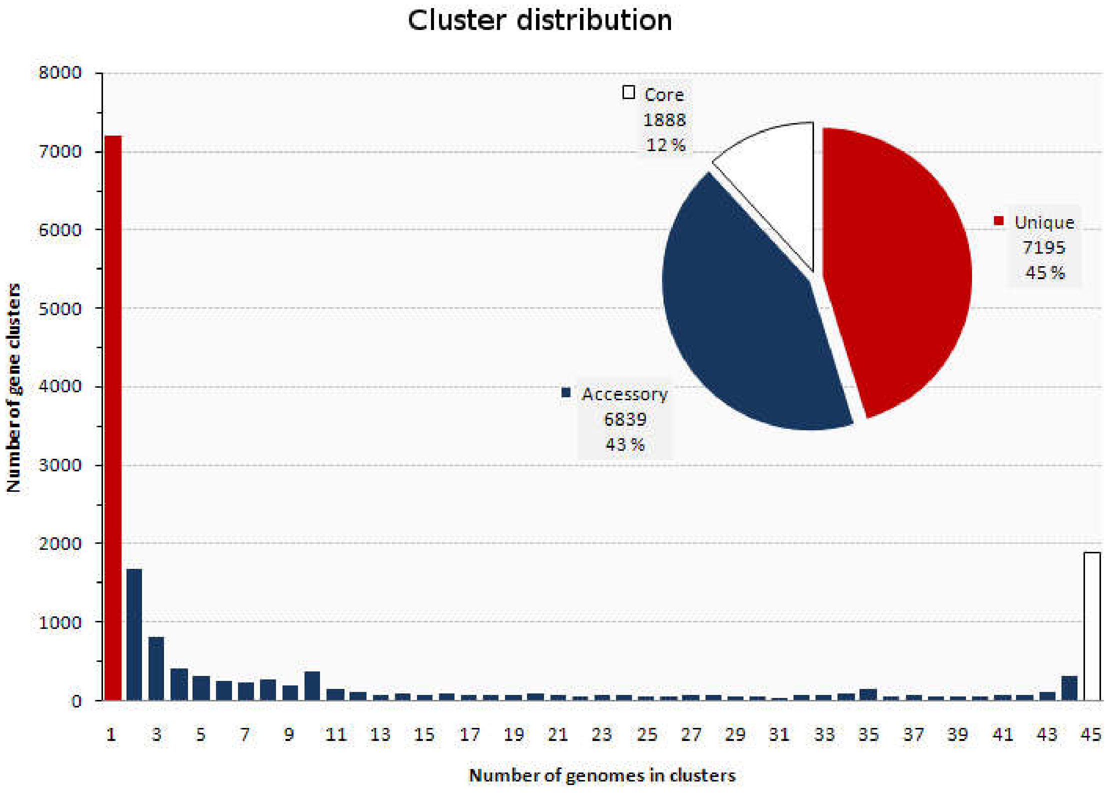


Figure 7. Illustration shows the percentages of genes being classified as core, accessory and unique as a pie chart. Bar-plot shows the distribution of genomes (strains) being present in cluster.

Genome composition analysis

The computed theoretic sizes of the pan- and core-genomes are illustrated as functions in Figure 6. The fitted and extrapolated function in Figure 6-A indicated that approximately 2020 genes will fall in to pre-existing clusters as new *Aliivibrio* genomes are added. The second plot in Figure 6-B imply that we can expect approximately 84 new and novel genes being discovered for each *Aliivibrio* genome added to the pool. The convergence of the curve explaining the core has leveled out at a rather stable level, but the function extrapolating the pan-genome do not show any sign of leveling out based on our applied strains. The goodness of fit is more certain for the core with a residual standard error of 129.88 in comparison to the pan-genome curve with 293.69.

Cluster distribution

The pan-genome matrix created by *compare_clusters.pl* revealed the core to occupy 1888 gene clusters based on the 45 genomes as shown in Figure 7. The accessory size of the pan-genome, which involves the presence of 2 to 44 strains in a single cluster, totaled 6839 gene clusters. The pan-genome counted 7195 unique gene clusters represented by a single genome. The cluster distribution in the accessory was discovered to be uneven with decreasing values from a cluster size of 2 to 12 before stabilizing in the mid zone. Slightly increased numbers of clusters were seen in 34 to 35 and 43 to 44 range.

Core contents

We discovered three mobile element proteins and two hypothetical protein clusters among the ten largest clusters. Sequences within these clusters were usually represented by a few genomes or by single genomes. These were present in excessive amounts for *A. salmonicida* LFI1238 and *A. wodanis* 0609139. Also draft genomes contained huge numbers of mobile elements. One of these clusters totaled 162 sequences and was shared among seven of the *A. salmonicida* strains including the LFI-180 and 554, having 63 and 48 sequences respectively.

Syntenic and hypothetic genes clusters

Filtering syntenic genes with the *compare_clusters.pl* script resulted in a total of 4477 (28.1%) clusters, having a common neighboring gene in other clusters. Further insight revealed that most syntenic genes are found in the accessory and unique section while none were detected in the core. In the accessory part of the pan-geneome there were 1129 clusters with syntenic genes identified while the same amount for the unique part was 3348.

The amount of clusters containing hypothetical proteins was highest among the unique genes by 4397 (61.1%) followed by 3104 (45.4%) in the accessory and 188 (10.0%) in the core.

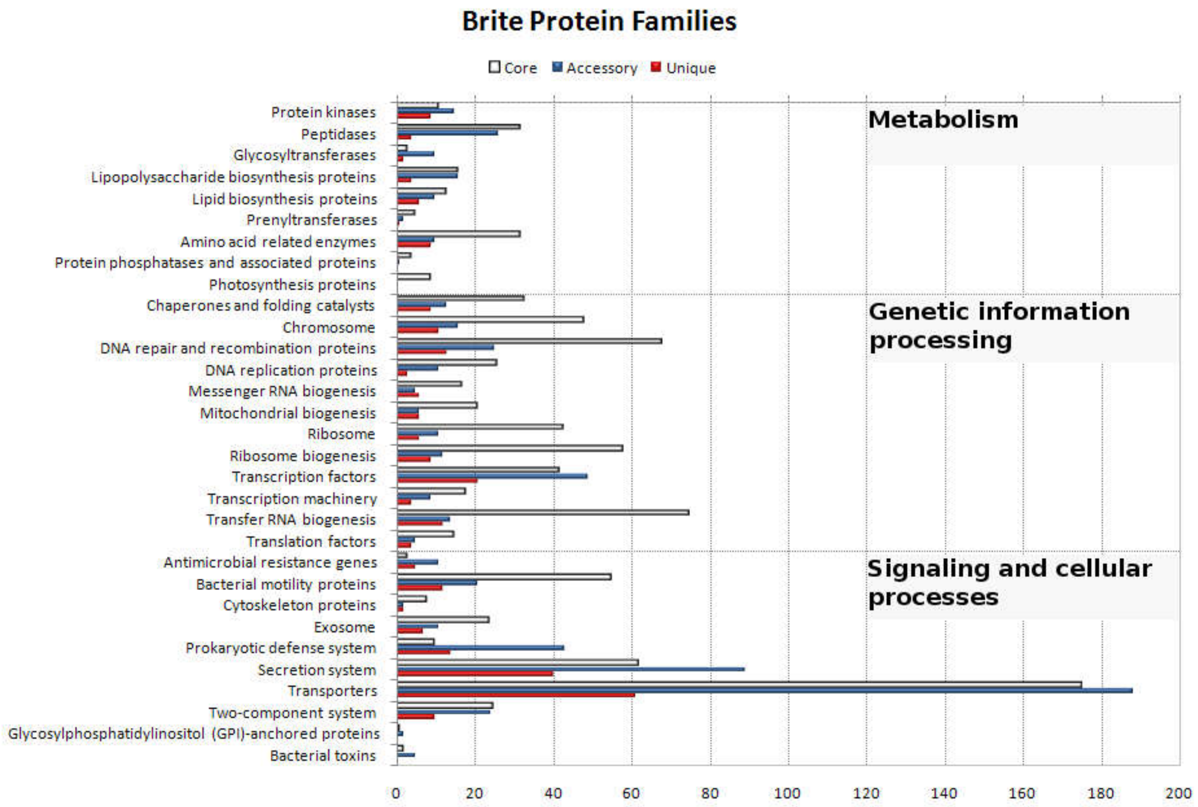


Figure 8. Direct counts of K-numbers (Kegg Ontology identifiers) from core, accessory and unique gene clusters mapped to the Brite protein families.

Protein families of the core, accessory and unique genes

Only partial mapping to the Brite protein families was achieved to obtain K-numbers (Kegg Ontology identifiers) for the pan-genome. The core, accessory and unique part had 1455, 1664 and 747 K-numbers mapped to the protein families respectively. These represented 77%, 24% and 10% of all gene clusters in the given pan-genome. The core was most prominent in the metabolism and genetic information processing categories as illustrated in Figure 8. Certain families, like transcription factors and glycosyltransferases, contained more mappings from the accessory genome than the core. In the signaling and cellular processes we found the accessory genome to dominate several categories. In particular the family representing secretion systems, prokaryotic defense systems and antimicrobial resistance genes, all shows a higher presence in the accessory than in the core. In two of the mentioned protein families, anti microbialresistance and prokaryotic defense systems, we also found the unique gene clusters being higher represented than the core, but in general the unique pan-genome were less represented compared to core and accessory.

Core identities

Identity matrices from the two homologue clustering runs were found to be directly comparable with the loci and MLSA analysis (Figure 4-A). These were based on average amino acid identity (AAI) and average nucleotide identity (ANI) respectively. Half of the data for both ANI and AAI are firmly positioned below 86% identity with the latter possessing an outlier at 68%, as shown in Figure 4-B. This outlier originated from the reference strains LFI1238 and 0609139 of *A. salmonicida* and *A. wodanis* respectively. The shapes representing the kernel probability densities are rather similar but have a slight down shift in AAI compared to ANI. They clearly separate in two distinct ranges at 85 to 86% identity and 96 to 98% identity with a thin transition area between them. This contrasts when comparing the shapes of *gyrB*, *pyrH* and *recA* which were observed with three ranges separated by transition areas.

Quorum sensing genes in *Aliivibrio* genomes

To understand the variations in the lux operon and its auxiliary systems we studied their presence and synteny within the species of *Aliivibrio*. By further identifying the *ainS* and *ainR* genes we attempted to gain an overview of this system's spread and phylogenetic relationships. The luxI/luxR complex with its operon was also thoroughly mapped in this project to obtain a comprehensive context of the system. The *luxR* and *luxR2* gene loci were also studied to reveal their identities for those strains possessing both in the lux operon. Conservation of other known quorum sensing systems were additionally mapped for the *Aliivibrio* species.

The ainS/ainR and luxM/luxN systems

The manual investigation of quorum sensing genes revealed three conditions for the *ainS/ainR* system. The system was found either present, absent or duplicated as shown in Figure 9. The duplication was only found in three of the newly proposed species of *A. modi*, *A. friggae* and *A. raniae* and in the *A. wodanis* strain Vw11. It was not observed in *A. finisterrensis* and *A. magni*. Strains of the latter had, in contrast to others, contig breaks that suggested the *ainS/ainR* system has been lost.

The duplicated *ainS/ainR* system was recorded with unequal gene lengths for the strains possessing this feature. The length of *ainS1*, *ainR1* and *ainR2* were approximately 1200bp,

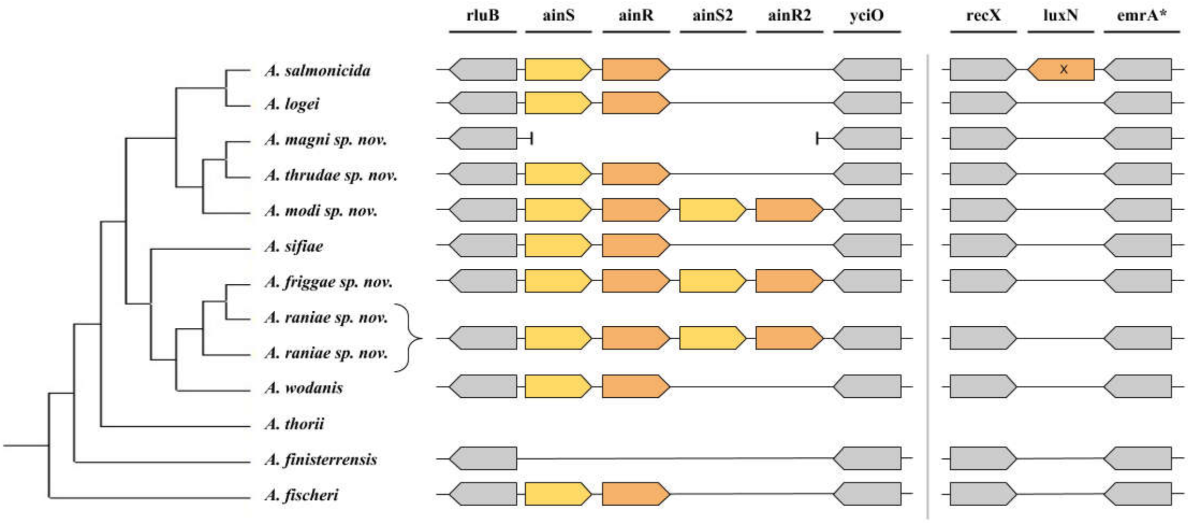


Figure 9. General trend among species observed with the ainR/ainS system as well as the luxN gene. Phylogenetic tree is an extract from the ML method applied in the MLSA analysis (see Figure 2). * The emrA gene was not conserved among all strains in species (see Appendix figure 4 and Appendix figure 5 for details). Frameshift mutations are indicated by “x”.

2460bp and 2460bp respectively for all strains. A shorter version of *ainS1*, having a length of 900bp was observed in *A. raniae* strain A11-1 and was the only deviation not resulting from contig breaks. The *ainS2* gene was discovered with three different lengths of approximately 650, 900 and 990 for *A. modi*, *A. friggae* and *A. raniae* respectively, thus shorter than the *ainS1*. The *A. wodanis* strain Vw11 also had the same length as the latter species based on the gene predictions.

In all *A. salmonicida* strains except for the environmental strains of R5-43 and B9-15 we found a possible frameshift mutation in what is believed to be the *luxN* gene. Here two slightly overlapping *luxN* annotated genes were located amid the flanking genes of *recX* and *emrA*. The same *luxN* annotation in this area was observed for strain ATCC 29985 of *A. logei* (not shown in Figure 9) as a single occurrence. Comparison of the *A. logei luxN* with the *A. salmonicida luxN* revealed them to be identical except for the fractured area in the latter species.

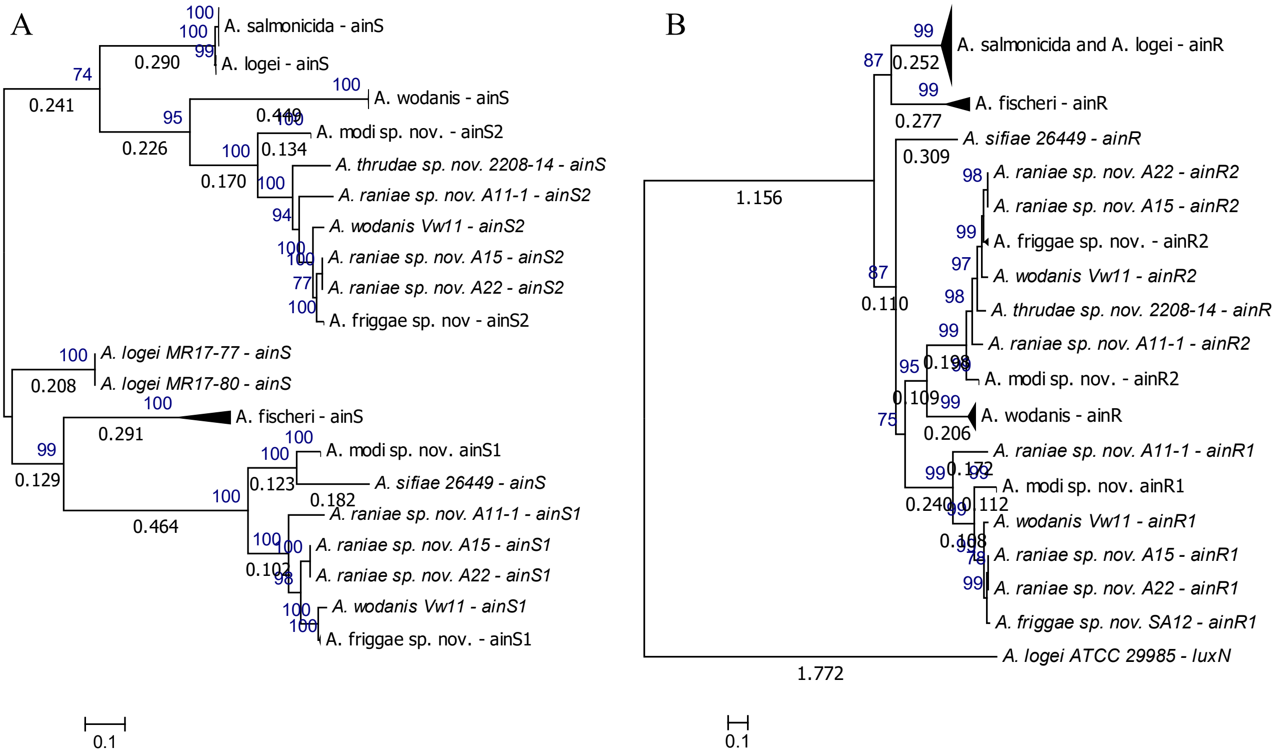

A
A. salmonicida - ainS
A. logei - ainS
A. wodanis - ainS
A. modi sp. nov. - ainS2
A. thrudae sp. nov. 2208-14 - ainS
A. raniae sp. nov. A11-1 - ainS2
A. wodanis Vw11 - ainS2
A. raniae sp. nov. A15 - ainS2
A. raniae sp. nov. A22 - ainS2
A. friggae sp. nov - ainS2
A. logei MR17-77 - ainS
A. logei MR17-80 - ainS
A. fischeri - ainS
A. modi sp. nov. ainS1
A. sifiae 26449 - ainS
A. raniae sp. nov. A11-1 - ainS1
A. raniae sp. nov. A15 - ainS1
A. raniae sp. nov. A22 - ainS1
A. wodanis Vw11 - ainS1
A. friggae sp. nov. - ainS1
0.241
0.290
0.226
0.449
0.134
0.170
0.208
0.291
0.129
0.464
0.123
0.182
0.102
0.1
B
A. salmonicida and A. logei - ainR
A. fischeri - ainR
A. sifiae 26449 - ainR
A. raniae sp. nov. A22 - ainR2
A. raniae sp. nov. A15 - ainR2
A. friggae sp. nov. - ainR2
A. wodanis Vw11 - ainR2
A. thrudae sp. nov. 2208-14 - ainR
A. raniae sp. nov. A11-1 - ainR2
A. modi sp. nov. - ainR2
A. wodanis - ainR
A. raniae sp. nov. A11-1 - ainR1
A. modi sp. nov. ainR1
A. wodanis Vw11 - ainR1
A. raniae sp. nov. A15 - ainR1
A. raniae sp. nov. A22 - ainR1
A. friggae sp. nov. SA12 - ainR1
A. logei ATCC 29985 - luxN
0.252
0.277
0.309
1.156
0.110
0.198
0.109
0.206
0.172
0.240
0.112
0.108
1.772
0.1

Figure 10. Phylogenetic relationship of the *ainS* (A) and the *ainR/luxN* (B) based on nucleotide sequences of their loci. Trees were constructed applying the ML method utilizing all available positions with supportive values represented on 100 bootstrap replications.

The phylogenetic reconstruction of all *ainS* genes in Figure 10-A showed a clear distinction between the *ainS1* loci, defined by having *rluB* and *ainR1* as neighbors, and the *ainS2.* The *ainS2* was defined with the neighboring *ainR1* and *ainR2/yciO* genes. The inferred tree showed the *ainS1* version of this locus share relatedness with both *A. fischeri* and *A. sifiae* while the *ainS2* type is clearly more similar to *A. wodanis*, *A. logei* and *A. salmonicida.* All represented as single systems carriers.

The phylogenetic tree constructed from the *ainR* loci in Figure 10-B shows the *ainR1* and *ainR2* versions to be clearly separated as well, but here only *A. wodanis* was integrated amid the duplicated systems. All of *A. salmonicida*, *A. logei*, *A. fischeri* shared similar *ainR* genes closely related in the inferred tree. Together with *A. sifiae,* these relate less with the duplicated systems than *A. wodanis*. The *luxN* gene of *A. logei* strain ATCC 29985 was highly divergent from the *ainR* and acted as an outgroup.

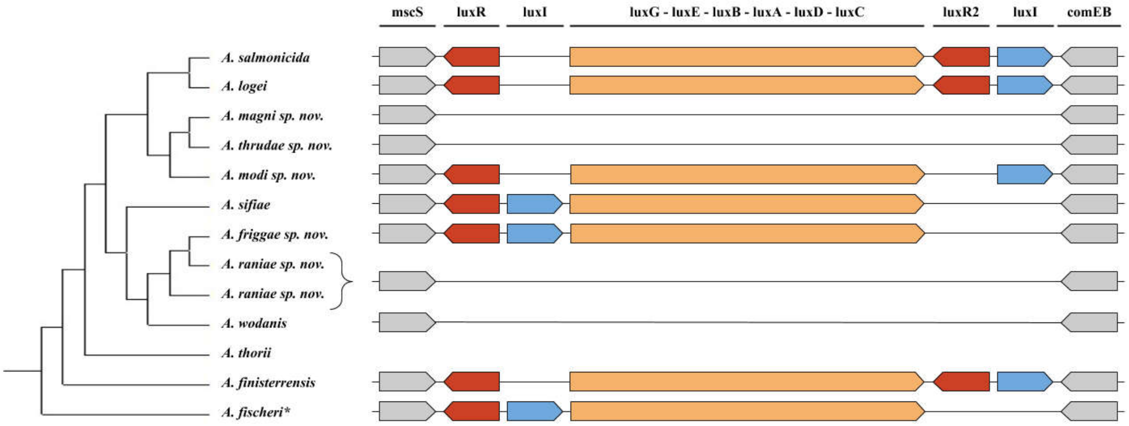


Figure 11. The genes comprising the lux operon and its two conserved flanking genes *mscS* within the *Aliivibrio* genera. The conserved gene position of *luxG, luxE, luxB, luxA, luxD* and *luxC* is shown as a single entity.

The lux operon

The lux operon, illustrated in Figure 11, was discovered in either of four settings in the *Aliivibrio* clade. The operon was not present in *A. wodanis*, *A. magni*, *A. thrudae* and *A. raniae*. Remaining species had variations where *luxI* and *luxR* shifted in position or number. *A. salmonicida* along with *A. logei* and *A. finisterrensis* possessed the duplicated *luxR2* gene. In the dual *luxR* systems, one was located upstream of the operon and one alongside the *luxI* gene downstream of the operon. The operon structure known from *A. fischeri* was also found in *A. sifiae* and the newly proposed species *A. friggae*. Unlike the layout in *A. salmonicida* this system contained no duplicate *luxR* gene and had the *luxI* gene located just ahead of the core lux genes instead of in the end of the operon. A novel design was discovered for the tree *A. modi* strains sharing the luxI placement known in *A. salmonicida*, but without having the luxR2 gene.

The sequence identities between the *luxR* and its duplicate *luxR2* were measured for the three species *A. salmonicida*, *A. logei* and *A. finisterrensis*. Computation showed 99.2 and 99.8 percent identities internally between the *luxR2* sequences on average for *A. salmonicida* and *A. logei*. For *luxR* we measured 86.6 and 93.8 percent identity on average between strains in *A. logei* and *A. salmonicida* respectively. Comparison of the *luxR* with *luxR2* resulted in an average deviation by of 61.5 and 61.9% identity for strains within the two given species respectively. In the only strain of *A. finisterrensis* the *luxR* and *luxR2* showed 53.2% identity. No comparison was made between the species.

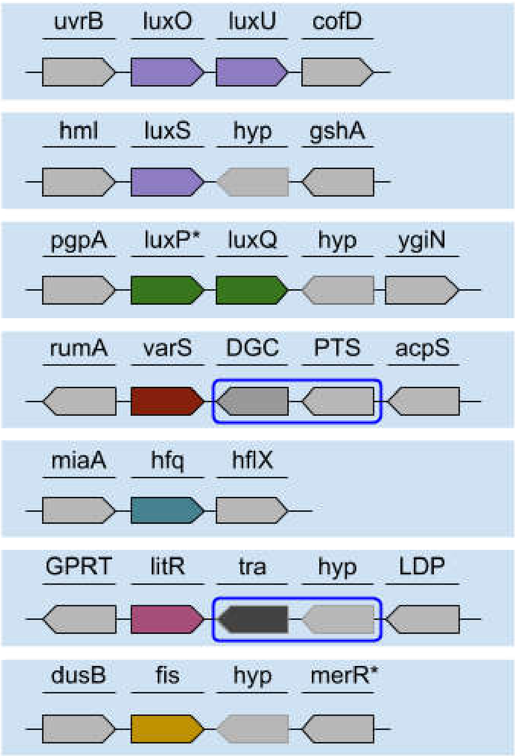


Figure 12. Systems related to the quorum sensing function. Marked in blue boxes are inserted elements only observed in few of the species considered. Complete overview of names and abbreviations can be found in Appendix figure 4 and Appendix figure 5.

Other quorum sensing systems

Genes in relation, with similar functionality or involved in the management of the lux operon or *ainS/ainR* system were discovered to be inconsistent between species. Some had variations in their synteny between strains as shown in Figure 12. The central workings of *luxU/luxO* and *luxQ* with the auxiliary *luxS/luxP* along with *fis*, *hfq* and *litR* were highly conserved in regards to their loci and synteny. The three genes *luxO*, *luxU* and *luxS* with their neighboring genes were identical for all 45 strains while the only deviation for *luxP* was observed in *A. salmonicida*. Here we observed one or two frameshift mutations splitting the gene. This was not seen in the environmental strains B9-15 and R5-43. Also relevant for *A. salmonicida* was the insertion of transposases in front of the *litR* gene. There were observed three transposases and a hypothetical protein for the LFI1238 strain, but in other strains there were contig breaks resulting in either none or a single transposase annotation. Also in this case the two environmental strains of *A. salmonicida* were the exceptions, having none of these transposases inserted. The receptor coding gene of *varS* was highly conserved among the species considered, but the additional two genes *DGC* and *PTS* were common for most *A. wodanis* strains as well as for *A. raniae* strain A15 and A22. These two genes were not observed in strain Vw11 and A11 of the respective species.

Detection of the CRISPR defense system

Species and strains carrying the CRISPR system in *Vibrionacea* are largely unknown. Here we attempted to gain insight in the presence of CRISPR systems in the included *Aliivibrio* species as well as identifying which system types occurred among these. Details on repeat sequences and cas genes were mapped to gain knowledge about familiar trends in this genus.

Identified carriers of the CRISPR system

There were found CRISPR arrays in 11 of the 45 genomes when applying the CRISPRfinder (Grissa et al., 2007b) default search algorithm. For these the overall numbers of positively confirmed arrays by the web-application were 21 and were distributed as either one or two arrays per genome. All *A. wodanis* strains had these arrays which all contained 28bp long direct repeats (DR) with up to forty spacer sequences (see Table 8). Others were *A. salmonicida* strain B9-15 and *A. magni* strain R8-63 which also possessed positively confirmed arrays.

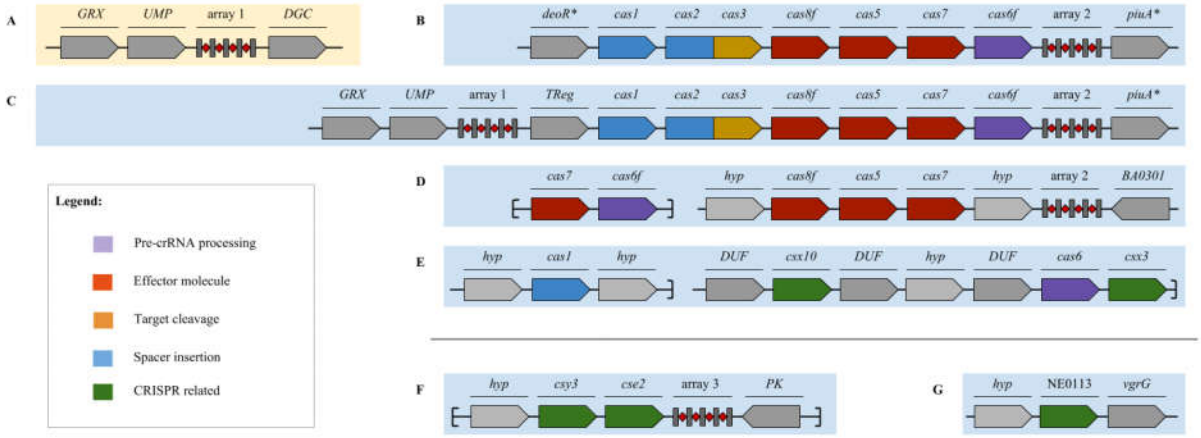


Figure 13. CRISPR arrays and their neighboring cas-genes and flanking genes. Abbreviations: Hypotetical protein (*hyp*), Transcriptional regulator (*TReg*), Integral membrane protein (*IMP*), COG0398: uncharacterized membrane protein (*UMP*), Glutaredoxin (*GRX*), diguanylate cyclase (*DGC*), Membrane protease family protein BA0301 (*BA0301*), Pyruvate kinase (*PK*), DUF324 domain-containing protein (*DUF*), valine-glycine repeat G (*vgrG*), deoxyribose operon repressor (*deoR*), hydroxamate-type ferrisiderophore receptor (*piuA*). "*" indicate genes not fully conserved among strains.

The arrays themselves were found to be in four relatively distinct positions flanked by somewhat conserved genes as shown in Figure 13. The position of array 1 (Figure 13-A) was exclusive to *A. wodanis* and had flanking genes coding Glutaredoxin, a membrane protein and a GDDF domain containing protein. Array 2 was found in all CRISPR containing strains except R8-63 of *A. magni* and, in contrast to array 1, always found directly downstream of the cas-genes (Figure 13-B). A deviation was seen in *A. wodanis* strain Vw11 where array 1 appeared upstream of the Cas-genes instead of the usual far off position known from reference strain 0609139 of *A. wodanis*. The array was observed in this position together with two of its known flanking genes (Figure 13-C). Array 3 (Figure 13-F) was found downstream of a solo effector complex (Makarova et al., 2015) possessing a hitherto unknown synteny of the *cse2* and *csy3* (*cas7*) annotated genes. This complex was found in six strains including *A. wodanis* Vw 35, *A. raniae* A15 and A22 , *A. fischeri* MJ11 and ZF-211 and R8-63 of *A. magni*. Only *A. magni* strain R8-63 had a confirmed CRISPR hit within this effector complex. This binary system was notably plastic but seemed to be conserved with a pyrevat kinase downstream and a hypothetical protein upstream. Another cas related gene annotated as NE0113, also known as *csx1* (Makarova et al., 2011) (Figure 13-G), was found in 8 of the strains; *A. wodanis* Vw12 and Vw130426, *A. salmonicida* R5-43, *A. raniae* A15 and A22, *A. logei* SES03-1 and SES03-5 as well as *A. finisterrensis* DSM 23419. Strains having NE0113 did not have any confirmed CRISPR arrays with the exception of *A. wodanis*.

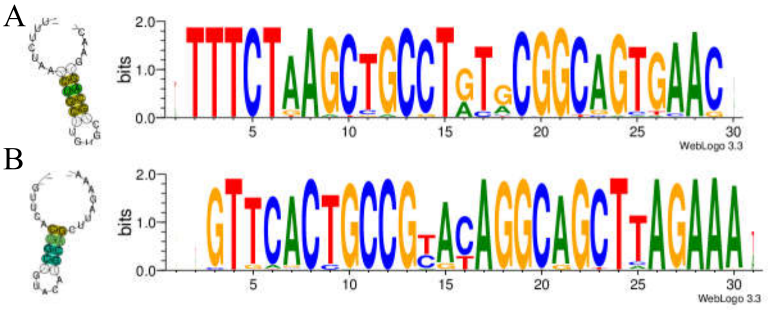


Figure 14. Secondary structure and logo of consensus repeat sequences. A represents structure family 6 with sequence family 8 and B represents structure family 4 with sequence family 5 based on CRISPRmap v.1.3.0-2013 classification.

Investigation the gene order with CLC revealed the systems to consist of 6 cas-genes in an operon annotated as; *cas1*, *cas3*, *csy1*, *csy2*, *csy3* and *csy4*. Here, the second gene, *cas3*, was significantly larger than the other with a length of 3411bp (*A. wodanis* strain 0609139). This system was seen in most *A. wodanis* and the *A. salmonicida* B9-15, but the three strains Vw130426, Vw35 and Vw8 of *A. wodanis* did not possess complete CRISPR systems. Vw130426 and Vw8 were missing the genes coding effector molecules while genes coding spacer insertion and target cleavage functions were missing in Vw35. The latter strain was also found to have a short contig carrying *cas7* and *cas6f* by itself, as shown in Figure 13-D. The strain *A. magni* R8-63 had the *cas2*, *cas6*, *csx3* and *csx10* (Figure 13-E) in a region surrounded by a vast number of hypothetical proteins, both downstream and upstream of its relative position on the contig. R8-63 did not have any repeat array nearby these cas genes, but was identified with a confirmed array in its solo effector complex.

Sequence specific identification of system type

The consensus sequences of direct repeats were largely identified as either of two secondary structure families, directly linked to their corresponding sequence families by CRISPRmap (Lange et al., 2013). These were structure motif 4 and 6 which relates to sequence family 5 and 8 respectively as illustrated in Figure 14. Both structure motifs model a similar stem-loop by complementary binding CUGC, forming a 7bp head. In 5 of the 7 strains containing array 1 the motif structure and sequence family were 4 and 6 respectively (Table 8). In array 2 on the other hand, half of the strains had 6 and 8 as structure motif and sequence family respectively. Both structure motifs and sequence families indicated CRISPR subtype I-F and were indicated by CRISPRmap to harbor the cas-genes; *cas1, cas2, cas3', csy1, csy2, csy3, cas6f* as well as *cas3"*, *cas7* and *cas10*. The sequence family could not be determined in 7 of 21 submitted repeats. Two of these could not be mapped to any secondary structure motif while three cases had unfamiliar motifs which were not found in the CRISPRmap database. Strains having these unknown sequence families were *A. wodanis* strain Vw1 and Vw35, *A. raniae* strain A11-1 and *A. magni* strain R8-63.

Table 8. Hits with CRISPRmap based on the submitted consensus direct repeat sequences for all confirmed arrays. Motif structure and sequence family refers to the systematic classification described by Lange et al., 2013.

	Array 1				**Array 2**			
Strain	Repeat length	Repeat number	Motif structure	Sequence family	Repeat length	Repeat number	Motif structure	Sequence family
A. Wodanis								
0609139	28	25	6	8	28	40	6	8
Vw1	28	24	4	5	28	39	6	8
Vw12	28	15	4	5	28	25	6	8
Vw130426	-	-	-	-	28	7	6	8
Vw35	28	6	4	5	29	7	4	?
Vw8	28	5	6	8	28	21	4	5
VwK7F1	28	13	4	5	-	-	-	-
Vw11	28	35	4	5	28	6	4	5
A. salmonicida								
B9-15	-	-	-	-	28	22	4	5
A. magni	**Array 3**							
R8-63	28	3	4	?	-	-	-	-

Discussion

Assembly and Annotation

Draft genome assembly proves the presence of repeat sequences

Draft genomes were obtained from Illumina paired-end reads by applying CLC Genomics Workbench for filtering and assembly. Reads can often be of low quality and requires filtering and quality control in early steps before the assembly process is started (Zhou and Rokas, 2014, Guo et al., 2014). The assembly merged overlapping and paired 150bp-reads in an attempt to form as large contigs as possible from the data of sequenced strains. We had highly varied amount of reads which ranged from 450k to 8.4m paired-end reads and resulting in contig numbers between 40 and 1407. Draft genomes with high number of contigs can be linked to interference and inaccuracy in downstream analysis and is often found as a result of technical (low coverage) or biological (repetitive sequences) nature (Koren and Phillippy, 2015). We found strains of *A. salmonicida*, in particular, to generally result in elevated amounts of contigs compared to other species. From homologues gene clustering we know strain LFI1238 has substantial number of repeating sequences, like transposases and hypothetical proteins, which are equally problematic in draft genomes of this species. The Illumina paired-end feature is indeed designed to reduce repeated sequences like these, but there are other biological effects like variable tandem repeats which is not properly correctable even with this technology (Koren and Phillippy, 2015). On the technical side we discovered a clear correlation in the number of reads and the obtained contig numbers after assembly. The fewer reads the higher the number of contigs, followed by lower contig coverage. Thus, *A. salmonicida* is on this basis a difficult species to perform assembly on and very difficult to obtain high quality draft genomes of. It is likely that coming sequencing technologies, where longer reads are supported, overcome the repeat lengths of *A. salmonicida* and will achieve higher quality draft genomes (Koren and Phillippy, 2015).

Contig mapping explain greater DNA variation among *A. wodanis*

Contiguation (ordering, aligning and orientation) of contigs against a reference were carried out with ABACAS. We obtained contigs from strains of *A. salmonicida* and *A. wodanis* as well as one strain on *A. fischeri* which was contiguated against their corresponding references LFI1238, 0609139 and ES114 respectively. ABACAS is intended as a tool for helping researchers plan an approach to close genomes (Assefa et al., 2009), but its functionality can be exploited to reduce the random factor when viewing genomes in browsers, and to some extent, connect contigs and improve gene synteny. Contiguating performed with *A. salmonicida* strains revealed almost all contigs became mapped (~98%) with the exception of strain B9-15 and R5-43 (45% and 49%). The latter two strains are known to be environmental strains from body samples of Ascidiacea and a soft coral respectively. These possessed approximately 0.8Mb larger genomes than those originating from mainly diseased fish samples. The low numbers of contigs mapped in these are most likely a combination of larger genomes and dissimilarity between the contigs and the reference being mapped against. On the contrary, we found contigs belonging to strains of *A. wodanis* and *A. fischeri* ranging between ~40 to ~90% being mapped against their respective references even as their chromosome sizes were highly similar (~4.5Mb for *A. wodanis*). By default ABACAS map nucleotide sequences sharing a minimum of 40% identity and 40% overlap. This shows that strains of *A. wodanis* share more variation in comparison to the reference 0609139 than *A. salmonicida* strains against the LFI1238 reference.

Number of annotated genes varies greatly within the *Aliivibrio* genus

Annotated genomes were achieved using the RAST web-service (Aziz et al., 2008). After the initial identification of open reading frames (ORF), these were automatically annotated with the closest relative homologue to obtain information about the translated product. We supplied the service with strains as contigs (draft genome) or in closed condition and achieved between 3424 and 4650 predicted and annotated genes. This shows the *Aliivibrio* to be highly diverse when the genus can differ with more than 1200 genes. Normally small genomes are associated with an intracellular, host dependent behavior while larger genomes represent free living cells less dependent of other species (Martinez-Cano et al., 2014). When comparing genome size of *A. logei*, representing samples of intestinal origin, body surface and filtered seawater, we discovered an opposite trend as stated in the literature. Here the intestinal and body samples, likely living as symbionts with the host, possessed significantly (up to 300) more genes in comparison to the free living cell from filtered seawater. B9-15 and R5-43, both environmental strains of *A. salmonicida*, were found to better represent the theory by having larger chromosomes and larger genomes compared to strains from diseased fish. Genes in the hundreds, present or not, also represent uncertainty since there only were a few environmental strains to compare the remaining symbionts with. This as well as the quality of the obtained draft genomes adds to inaccurate predictions and annotations of genomes.

Phylogeny

High resolution achieved on the basis on six housekeeping loci

The *Vibrio* genus is widely studied and outlined in many publications, but its family member *Aliivibrio* has lagged behind in knowledge about its phylogenetic relations. We have attempted to fill this gap with a Multi Locus Sequence Analysis (MLSA) to better understand the residing species of *Aliivibrio* and how they relate.

We achieved a complete MLSA design by concatenating the six housekeeping genes *16S rRNA, gapA*, *gyrB*, *pyrH*, *recA* and *rpoA*. These mainly included strains from the *Aliivibrio* genera but also *Vibrio* and *Photobacteria* for comparison. The MLSA was aligned with ClustalW and further processed with MEGA to obtain model test and infer the phylogenetic relationship. This was performed with the Maximum Likelihood method. The inferred MLSA tree showed a clear distinction between the three genera and the outgroup, indicating the discriminatory power at the genus level is satisfactory. We thereby confirm the findings of Ast et al. that the *Aliivibrio* clade with its species is distinct from the *Vibrio* and *Photobateria* of *Vibrionacea* (Ast et al., 2009).

At the species level we obtained good classification in general for both the *Vibrio* and *Aliivibrio* clades. With the included strains there appears to be twelve species within the *Aliivibrio* genus based on our MLSA. Seven of these are previously described and were forming stable distinct phylogenetic clade. This made us able to classify and propose five new species which appeared in relation to the inner group of *A. salmonicida*, *A. logei*, *A. sifiae* and *A. wodanis*.

The work of Ast et al. demonstrated a similar tree structure of *Aliivibrio* where *A. sifiae* and *A. wodanis* were slightly separated from *A. logei* and *A. salmonicida* (Ast et al., 2009). Here they also found *A. fischeri* forming a distinct group aside from the other mentioned species, but were also described to split in to two clades. They additionally introduced two new species; the *A. sifiae* and the *A. thorii*, which both were included in this study. The former was localized in agreement to the study of Ast et. al., but the latter were more distinct and positioned closer to *A. fischeri* than *A. sifiae*.

Network analysis of *Aliivibrio* emphasized the existence of novel species

It has become increasingly popular to represent taxa in phylogenetic network diagrams or as an addition to regular phylogenetic trees. These networks have the ability to illustrate the evolutionary relationship of taxa in an explicit manner as nodes connected by edges (Huson and Bryant, 2006).

In an early study from 2007 Sawabe and collaborators constructed a split network tree based on MLSA applying nine concatenated gene loci of 58 *Vibrio* related species (Sawabe et al., 2007). These were found to form 14 monophyletic clades, one of these included the *A. wodanis* and *A. fischeri* species, in a Jukes-Cantor corrected network diagram. A second by the same team in 2013 included 96 taxa in a similar MLSA with eight housekeeping genes (Sawabe et al., 2014). Here they expanded the split network diagram and discovered 26 clades within the *Vibrionacea*. The clades they defined as *Fischeri*, *Phosphoreum*, *Anguillarum*, *Cholerae*, *Vulnificus*, *Splendidus* and *Harveyi* all included all species used in this study's MLSA and spatially agreed very well with our reconstructed network tree. We included 59 *Aliivibrio* strains in the network analysis compared to the five species samples in the work of Sawabe and collaborators (2013), thus achieved a higher resolution for this genus in our splits network diagram. By eliminating non-*Aliivibrio* strains we focused on the local taxa arrangement which showed to strengthen the likelihood that there are up to five new species not previously described.

In light of the finalized MLSA, both reconstructing the phylogenetic relationship applying the ML method and the split network method, we propose the existence of five new species. These are the *A. modi*, *A. thrudae*, *A. magni*, *A. friggae*, and *A. raniae* which uphold the *Allivibrio* nomenclature relating its species to the Norse gods. Móði (Modi) and Magni, the brave and the strong son of Odin as well as his daughter Thrud (Þrúðr), meaning strength (Rudolf, 2008, Margaret, 1994) has contributed with their names. Additionally, the names originating from Frigga and Rán, which according to the old Norse mythology are Odin's wife and the goddess of the sea (Rudolf, 2008, Carolyne, 1999). These were selected for the remaining two species discovered.

Individual gene loci share unique features

Individual data as sequence statistics, pair wise identities (distances) and maximum likelihood based tree representations were achieved for five housekeeping genes (*gapA*, *gyrB*, *pyrH*, *recA* and *rpoA*) and the *16S rRNA* locus. These data were based on the same strains as included in the MLSA, mostly comprising the *Aliivibrio* genera with additional strains from the *Vibrio* genera and *Photobacteria* for comparison. We found consistent patterns among the data for each locus that complemented a coherent picture of their features. GC content in these was generally consistent with little variation between strains and was found ranging from ~42% to ~52% for the six loci. The *16S rRNA* locus had the highest value and thereby probably represents the most stable locus followed by *pyrH* while *recA* had the lowest GC content (Yakovchuk et al., 2006). Our values showed similarity with the study of Thompson and collaborators which discovered the GC content of *rpoA* (46%±0.9%), *recA* (46.6%±1.9%) and *pyrH* (48.7%±0.9%) in *Vibrio* species (Thompson et al., 2005).

The loci comprising *16S rRNA*, *gapA* and *rpoA* were found to be less valuable in characterizing strains in the genera and species levels and had the least variable positions compared to their lengths. The aggregate of their pair-ways distances were skewed or highly focused, which became evident as individual problematic factors in the inferred phylogenetic trees.

The *16S rRNA* and *gapA* loci were found to be unable to adequately separate the *Aliivibrio* and *Vibrio* genera. The trend for the *gapA* gene locus was described as "fuzzy" and complex by Thompson and collaborators when trying to distinguish *V. harveyi* with other *Vibrio* species (Thompson et al., 2007). We thereby propose *gapA* to be applied with care since it tends to perform poorly in some cases, but may still be included to provide support in concatenated datasets. The least informative and concluding results were obtained from the *16S rRNA* locus. This locus has been studied extensively by Sawabe and collaborators where they used 58 *Vibrio* taxa to compare the contributing effect of *16S rRNA* with 8 housekeeping genes (Sawabe et al., 2014). It was concluded from this research that the sequences from *16S rRNA* were neither critical nor necessary for reconstructing the phylogeny of *Vibrio* on the basis of MLSA.

Removal of gapped or missing data in *16S rRNA* proved to negatively impact the inferred ML tree more than for other loci. Most of the 111 positions removed were in the start and end of our loci selection and corresponded to the areas 451-472 and 809-893 in the reference sequence of *V. cholera*. These sections corresponds to a part of variable region 3 (V3) and the entire V5 of *E. coli 16S rRNA* (Yarza et al., 2014) and is most likely the reason for the heavy loss of resolution. Further application of *16S rRNA* should at least ensure that variable regions become included to gain any discriminatory effect.

When comparing our work with the work of Thompson and collaborators (2005) we found agreement with the *rpoA* gene being highly capable of distinguishing on the genera level (Thompson et al., 2005). This became clarified in the violin plot (Figure 4) and inferred maximum likelihood trees (Appendix figure 3), but also showed lacking abilities to separate at the strain and species level. We believe the *rpoA* locus is a crucial component in concatenated MLSA designs due to its ability to separate well on the genus level. This remains true within the *Vibrionaceae* family and it is thereby suggested to always include *rpoA*.

In the second part we discuss the *gyrB*, *pyrH* and *recA* loci which show more consistency in separating species and genera. In the violin plots of all these loci we observed shapes that had three focused regions with slimmer transition ranges between them. This indicates visible boundaries between genera, species and strains of *Vibrionacea* and was best explained by these three loci. We experienced a few issues with *gyrB* and *recA* being unable to manage stable grouping of the *Photobacteria* by interfering with the *Vibrio* clade or the outgroup in the inferred ML trees. Otherwise, these performed well in separating species and genera with *pyrH* and *gyrB* having the best proposed resolutions. We did not experience the same problems distinguishing between species as described by Thompson and collaborators (2007) where *gyrB* proved to have the lowest discriminatory power within the *Vibrio* genera.

When examining the pair ways similarity data of our MLSA design (Figure 4) we found a slightly more complex model compared to the individual genes. It showed narrow transition states between genera, species and strains which was confirmed by the ML tree. This was managed in the tree generating process despite of losing 442 of its 3778 original positions .

Pan-genome analysis

Pan-genome analysis proves *Aliivibrio* to be a diverse genus with a moderately conserved core

The pan-genome and its clustered gene content were obtained by utilizing Get_homologues with RAST annotated genomes. This was achieved with 45 selected *Aliivibrio* genomes (Table 4 of annotated genomes) representing available species in draft or closed condition.

We experienced the clustering process of 45 genomes, in the size order of 3.7 to 5.0Mb, with 3500 to 4500 protein sequences each taking slightly more than half a day applying a high end laptop computer. This result emphasizes the possibility of performing pan-genome analysis on moderately large datasets without the need of a computer cluster (Contreras-Moreira and Vinuesa, 2013), but due to the nature of blasting all against all there is a practical user limit for adding additional genomes until the necessity of a computer cluster becomes obvious.

We were able to compute from regression analysis, developed by Tettelin and collaborators (Tettelin et al., 2005), the most likely number of novel and conserved genes when a new genome is added to the pool based on the provided genomes. These were found to be ~84 novel and ~2020 conserved genes respectively from the genome composition analysis and represent a vague estimate of what can be expected in general for the *Aliivibrio* genera with its current species. Observing the trend lines provided in Figure 6 for the core genome extrapolations indicate that the *Aliivibrio* appears to be represented well by a core of approximately 2000 genes as the curve has flattened out. The pan-genome on the contrary showed a steep continuing curve, still climbing, which can be interpreted as it will take considerably extra genomes to obtain a collection that encompass all genes in the *Aliivibrio* genera. It also tells us that the genus is highly diverse, but it is likely that the low quality in some assembled genomes and few closed genomes contributed to incorrect genome sizes, thus leading to inaccurate calculations.

The initial clustering analysis of the 45 genomes reviled 15922 gene clusters being distributed in the core, accessory or unique pan-genome (Tettelin et al., 2005). In this particular study we found 1888 gene clusters representing the conserved core which represent 12% of the pan genome, but approximately 46% of the average genome size (4100 genes) of a given *Aliivibrio* strain. This genus represents a considerably less conserved core compared to the species core of eight *Streptococcus agalactia* strains that were found to retain approximately

80 percent of its core in a pan-genome study (Tettelin et al., 2005). Despite this, our study indicates that one half of the average genome is, to some degree, free to be reinstated by other functions or depleted, in particular the unique part. The accessory and unique pan-genome are considerably larger by 6839 and 7195 gene clusters respectively, illustrating the great variation observed in the *Aliivibrio* genera. In the accessory there were noteworthy more gene clusters with up to 10 genomes present, sharing homologues sequences, and also gene clusters with 44 genomes present were prominent. The first case is likely the cause of *A. salmonicida*, *A. wodanis* and *A. logei* which all consist of eight to twelve strains while the remaining species are represented with fewer strains, often tree or four. These naturally share a higher proportion of homologous genes since they represent the same species. In the other end we have the clusters with 44 genomes present where the main portion of these are likely belonging to the core part, but are expelled to this cluster as a result of erroneous sequencing, contig assembly or gene prediction. More likely is also the high number of uniquely clustered sequences, only found in a single genome, to be either in the accessory part or in the core part as these might be fragmented genes originating from erroneous assembly. Careful parameter adjustment might be able to optimize these results, as for this project the OrthoMCL (Li et al., 2003) algorithm remained at a 75% coverage cutoff value for genes. Based on this it is likely that a number of interrupted genes became classified as unique instead of their real cluster affiliation leaving a high false negative rate. Lowering the coverage may solve this but will likely need to be counterbalanced with a higher identity threshold (currently 1%) or adjusted inflation value (1.5).

It has been stated that it is likely more biologically relevant to define the pan-genome core in a less strict manner by defining a ‘soft core’ which include genes in 95% of all considered genomes (Kaas et al., 2012). Implementing this will likely result in less false negatives excluded from the core than false positives included. Applying the soft core rule in this study would have included additionally 471 gene clusters forming a core of 2359 genes.

The content of the pan-genome was syntenic only at the species level and comprised a wealth of inconsistent and unknown gene products

We investigated the rough content of unknown gene products, hypothetical genes, and transposases as well as mobile genetic elements in the obtained gene clusters. Hypothetical genes were highly present in the dispensable part of the pan-genome consisting of the accessory (45.4% unknown gene clusters) and the unique (61.1% unknown gene clusters) while the core only had 10% gene clusters with unknown functions. This means that the *Aliivibrio* genus is still in an early stage of exploration with copious gene products still to discover and classify. It agrees with a previous study on *S. agalactia* where these undefined genes as well as mobile elements also were found in the same dispensable parts of the pan-genome (Tettelin et al., 2005). The authors also link mobile elements, which were highly present in both *A. wodanis* and *A. salmonicida* strains, as well as extrachromosomal elements to a theory that the majority of specific traits are the result of lateral gene transfer. This might be the case in our mentioned species as these redundant elements are expressed in draft genomes as high contig numbers and occurs due to assembly errors. Many of these sequences are likely lost in such draft genomes and would be better represented if the genomes were closed.

Syntenic genes are described by Tang H. and collaborators as “inferred homology relationship between genes which are supported by sharing a common genomic neighborhood” and were additionally investigated in the *Aliivibrio* genus with Get_homologues (Moreno-Hagelsieb et al., 2001, Tang et al., 2015). Surprisingly there was no synteny discovered in the core of the pan-genome whereas the accessory and unique parts had 1129 and 3348 syntenic gene clusters. It means that the 45 strains considered do not share any common neighborhood among the core and thus the species they represent seems to differ in their genomic architecture. The accessory part of the pan-genome has obvious reasons for sharing genomic neighborhoods since it deals with all possible smaller combinations also including separate species, and involves the unique species core (Kahlke et al., 2012). This then implies the existence of syntenic homology at the species level and probably between some of the species, but might be present in the *Aliivibrio* core genome if all genomes have been in a closed condition.

We additionally performed mapping of K-numbers (Kegg Ontology identifiers) with BlastKOALA (Kanehisa et al., 2016) in order to achieve the proportion of protein families being present in each of the pan-genome parts. The mappings achieved with the web service correlated with the percentage of hypothetical gene clusters observed and again proves there are many genes left to classify in the *Aliivibrio* genera. In previous studies there were proposed to be a relationship between the core content and housekeeping genes with some of their functions involving management of cell envelope, regulation, transport and binding (Tettelin et al., 2005, Sarkar and Guttman, 2004). In this study we found the pan-genome core heavily represented in the metabolism and genetic information processing categories of the

KEGG protein families. These harbor many of the housekeeping functions mentioned and thus show similarity with the literature. In the signaling and cellular processes category several families, including antimicrobial defense systems, secretion systems, and prokaryotic defense systems were more present in the accessory and unique parts of the pan-genome. This indicate species specific carriers of resistance to antibiotics, type I-VI secretion systems among others and several defensive systems like CRISPR-Cas, restriction and modification systems are present within the *Aliivibrio* genus, but distributed only among a selection of its species.

The genome core distances showed favorable resolution in separating species

Average identity matrices based on the core genome, representing amino acid identity (AAI) and average nucleotide identity (ANI), were obtained in two separate runs with Get_homologues and processed thru R with the ggplot2 package (Team, 2008, Wickham, 2009) to compute violin plots. The first observation when comparing these plots against the gene loci plots was the presence of two, instead of tree, distinct shapes explaining the amassment of values. These were centered around 84 to 87% and 96 to 98% identity, slightly shifted between the two identity types, and refer to the between-species and within-species identity expected when comparing the core distances of *Aliivibrio* genomes. These are comparable with the loci and closely resemble the lower value centering of *rpoA*. But when the core identities illustrate the between-species identities, the *rpoA* locus resembles the between-genera identities and thus indicating a significantly greater resolution by applying the core which is unmatched by any of the loci.

In a recent study on the species definition and concept Rossello-Mora and collaborators measured the ANI of all present genomes in the NCBI database applying the fast MUMmer algorithm (ANIm). Here they suggest distance ranges which encompass intra-species (96 to 100% identity) and intra-genus (77 to 93% identity) values as well as a genomovar/fuzzy zone (93 to 96% identity) where phenotypic similar species proves phylogenetic different (Rossello-Mora and Amann, 2015). Our ANI values overlap well both in the intra-species and intra-genus ranges and have the least values in the genomovar range. Based on their work this would mean in theory that the *Aliivibrio* species resemble phenotypic distinct species which also easily should be distinguished phylogenetic.

Quorum sensing

The quorum sensing ability is known among the *Aliivibrio* species *A. fischeri*, *A. logei*, and *A. salmonicida* and constitutes several sensors and synthases along with regulators to manage responses to certain auto inducers. What concerns other species in this genus is still not known.

The AinR/AinS system is not consistent in the *Aliivibrio* genera

The gene coding AinS synthesize auto inducers in *A. salmonicida* and *A. fischeri* under the right conditions (Hansen et al., 2015, Kimbrough and Stabb, 2013), (Kimbrough and Stabb, 2013) and are connected with the receptor AinR. Collectively they form one of the available quorum sensing systems which is the prerequisite for population density-dependent luminescence (Lupp and Ruby, 2004).

We charted the syntenic neighborhood of the AinR and AinS coding genes with the help of CLC genomics workbench for all 45 considered genomes. This was based on RAST annotations where these genes initially were defined as LuxM and LuxN, implying a problem related to accuracy in the process.

We noted species with and without this system as well as discovering duplicate systems in *A. modi*, *A. friggae* and *A. raniae*, all equally conserved between the two flanking genes *rluB* and *yciO* independent of the system condition. The duplicated *ainS/ainR* systems, denoted as *ainS/ainR 1* and *2* in this project, were not identical copies and have probably evolved over a relatively short time period in comparison to the homologous luxN. It is also likely the duplication has occurred twice due to its existence in *A. modi*, which is not a close relative and rules out a monophyletic relationship with the *A. friggae* and *A. raniae*. *A. modi* also diverges slightly in the phylogenetic reconstruction of the *ainS* and *ainR* loci strengthening the likelihood of two separate duplication events, while *A. friggae* and *A. raniae* might have obtained this duplicate system prior to the speciation event, but at a later stage circumventing *A. wodanis* (Via, 2009). These events seems to have involved both genes at the same time and determining the functional pair is uncertain, but the stability seems to have favored the *ainR/S 1* pair due to its well conserved lengths between species.

An unconfirmed existence of the system was reported for *A. magni* due to contig breaks and missing genes, but may shed light on the possible mechanism behind the duplication event. Paired-end reads have been utilized when sequencing all strains, including the *A. magni* strains, and has been proven to perform poorly on variable number tandem repeats (VNTRs) (Koren and Phillippy, 2015), thus resulting in contig endpoints. The same repeat sequences can further be linked to tandem duplication/deletion by homologous recombination or single-strand annealing (Reams and Roth, 2015, Reams et al., 2014). Further study on the subject might reveal the presence of tandem repeats in this region and its involvement in the *ainR/ainS* system.

The LuxN coding gene is known in *Vibrio harveyi* as a sensor kinase, along with LuxQ, working in parallel sensing-systems responding to auto inducer-I and II (AI-1 and AI-2) (Freeman et al., 2000). This gene was found among all the non-environmental strains of *A. salmonicida* and the *A. logei* strain ATCC 29985 and had *recX* as downstream neighboring gene. In all *A. salmonicida* these had frameshift mutations which likely made it a psedogene as previously described by Hjerde and collaborators (19099551), but a seemingly intact version of the gene was found in the *A. logei* strain ATCC 29985. It is then possible that this particular *A. logei* strain is capable of responding to the AI-1.

Novel form of lux operon identified in *A. modi*

The lux operon is known to contribute with the bioluminescent in bacteria by synthesizing the luciferase enzyme and is well known among species in the *Vibrionaceae* (Miyashiro and Ruby, 2012, Meighen, 1993). There is a minimum of five genes needed, the *luxCDABE*, to perform biosynthesis but additional nonessentials like luxG and luxH are also found (Meighen, 1991, O'Grady and Wimpee, 2008) alongside the auto inducer synthase luxI and the acyl-homoserine lactone (AHL) receptor luxR (Fuqua et al., 1994). We revealed the presence of this operon in all species except *A. magni*, *A. thrudae*, *A. raniae* and *A. wodanis*. The operon synteny was discovered to be dissimilar between some species and was represented as tree different types regarding the *luxR*-*luxI* arrangement. The well known *A. fischeri* type where *luxR*-*luxI* are upstream of the *luxGEBADC* genes (Dunlap, 1999), the *A. salmonicida* type carrying the additional *luxR2* along *luxI* downstream of the operon (Nelson et al., 2007), and a novel hybrid type discovered in *A. modi* carrying the *luxR* upstream and *luxI* downstream of the operon. It can be hypothesized that the *A. modi* type is a divergent representative of the operon architecture in *A. salmonicida* and has remained similar while the operon became lost within *A. thrudae* and *A. magni*. Despite this, it has the required components for responding to AHL in theory, but further studies might test the system integrity *in vitro*, and perhaps gain insight to possible mutations by sequence comparison.

The *A. fischeri* is known as capable of bioluminescence by synthesizing luciferase from the *luxA-luxB* genes (Miyashiro and Ruby, 2012). Work has been done to uncover defect systems (Manukhov et al., 2011, O'Grady and Wimpee, 2008), but there is a lack of knowledge with respect to the behavior of operon deficiency in *A. fischeri*, as we revealed in the strain ZF-211. This strain was otherwise similar to the remaining representative's considered, but originated from a filtered seawater sample rather than the light organs of the host (Cordero et al., 2012). It proves that there is no necessity of the lux operon in *A. fischeri*. This might be explained by natural selection, out ruling the operon, since this sample likely represents a free living cell who finds no involvement in symbiotic dependent bioluminescence.

The duality of *luxR* and *luxR2* is known in both *A. salmonicida* and *A. logei* (Nelson et al., 2007, Manukhov et al., 2011), but were also found for *A. finisterrensis* during this project. The average conservation of the *luxR2* locus was greatest (~99%) for each individual species considered. It was followed by *luxR* having moderately lower identity between strains, but was not identical for *A. logei* (86.6%) and *A. salmonicida* (93.8%). Sequence identities between *luxR* and *luxR2* were equally low for *A. logei* and *A. salmonicida*, numbering ~62%, while *A. finisterrensis* showed ~53% identity comparing these genes. The similar identities between the *luxR-luxR2* genes for *A. salmonicida* and *A. logei* may indicate the dual system originated prior to speciation, but since then the individual luxR genes have evolved at different rates.

Recent work on sixteen *A. logei* strains by Khrulnova and collaborators has concluded that *luxR2* is the main activator of the lux operon due to its higher sensitivity towards auto inducer concentrations (Khrulnova et al., 2016). This correlates well with our comparison of the sequence conservation in *A. logei*, and is probably following the same trend in *A. salmonicida*. Here the functional *luxR2* gene is maintained stable while the *luxR* gene is likely to further evolve. In an additional study by Konopleva and collaborators they proposed a more specific role for LuxR. In this study a cooperative relation between LuxR and LuxR2 is introduced for regulation of the lux operon in *A. logei* by heterodimerization of LuxR1/LuxR2 (Konopleva et al., 2016).

Additional quorum sensing genes were highly conserved

Prior work has uncovered additional two-component systems of AI synthases and sensor kinases like VarS/A and LuxS/PQ (Milton, 2006) as well as the phosphorelay protein luxU and the response regulator luxO (Freeman and Bassler, 1999). Also associated with quorum sensing is the genes coding hfq, fis and litR acting as mediator of regulatory RNAs, regulator and known homolog to the master regulator of quorum sensing in *V. fischeri* (Lenz et al., 2004, Lenz and Bassler, 2007, Bjelland et al., 2012).

These gene neighborhoods were in general highly conserved for all considered species. Few discrepancies were uncovered and only found in *A. salmonicida*, *A. wodanis* and *A. raniae* as additional transposase regions, frameshift mutations or additional genes.

The neighborhood of the *litR* coding gene had an upstream transposase region in all *A. salmonicida* strains except the environmental linked. In this region of the reference strain LFI1238, known for possessing high numbers of transposases (Hjerde et al., 2008), there is tree transposases which also is likely o be present in the remaining host associated *A. salmonicida* strains. The same strains contained frameshift mutation in luxP, but were not present in the environmental samples. Work by Hjerde and collaborators have also described the existence of a psaudogene variant of *luxP* in the LFI1238 strain of *A. salmonicida*, a sample from diseased atlantic cod (Hjerde et al., 2008). Here we suggest environmental strains of *A. salmonicida* might commonly be carriers of an intact *luxP* gene and have less frequent transposase activity.

Additional genes with the annotation referring to Diguanylate cyclase (DGC) and the wide group of ABC-type phosphate transport system (PTS) were found in all *A. wodanis* and *A. raniae* except strains Vw11 and A11-1 of the respective species. These genes were added upstream of VarS and might only represent redundancy, but due to DGC's involvment in bacterial biofilm formation (Sisti et al., 2013) it can probably be linked to the quorum sensing apparatus of VarS/VarA system.

CRISPR

The CRISPR system is believed to be an antiviral defense system applying a RNA-interference-like mechanism and is harbored in some bacterial species while almost all archaea has it (Sorek et al., 2008). Since there is little knowledge of this system among the *Vibronacea* family in general as well as its *Aliivibrio* genera, we charted the positive hits of within 45 genomes with CRISPRfinder and classified their system type based on the repeat sequences.

The CRISPR system was present mainly in *A. wodanis* strains

According to the February 2016 update of CRISPRdb (Grissa et al., 2007a) the *Vibrionacea* has seven species carrying CRISPR where most are found within the *Vibrio* and with a single case in the *Photobacterium* genera. In this project we discovered CRISPR systems in at least one species of the *Aliivibrio* genera and in one environmental sample of a species with no previously identified systems.

In a review from 2015 by Makarova and collaborators are five main CRISPR systems with additional sub-systems described. These can easily be identified based on their signature genes and gene synteny which comprise adaptation, expression and interference genes (Makarova et al., 2015). Since the palindromic repeats also are coupled to the proteins working on them, these sequences are able to backtrack on the system type (Lange et al., 2013). We identified the direct repeats of 45 *Aliivibrio* genomes by CRISPRfinder and applied these to CRISPRmap as well as investigating the synteny of annotated cas-genes. Both repeat sequences and synteny with annotation indicated the I-F sub-system as being the only one present. This is likely correct due to factors like the large *cas3* gene, which is found to be *cas2* and *cas3* fused together (Makarova et al., 2015), and the knowledge of system I-C/E/F and II being strongly coupled to their direct repeats (Lange et al., 2013).

CRISPR subtype I-F was identified in *A. wodanis* and *A. salmonicida*

Strains possessing the I-F system belonged mainly to *A. wodanis* with the exception of one environmental strain of *A. salmonicida*. Despite of all of *A. wodanis* having direct repeats and spacers with significant sizes, only half of these seemed to have functional systems due to missing cas-genes. The *cas1* and *cas2* genes transcribe a protein complex which is the most universal among the systems and performs the acquisition of new spacer sequences (Nunez et al., 2014). One or both of these was not observed in strain Vw130426, Vw35 or VwK7F1 while *cas5*, *cas7* and *cas8f* were missing in Vw8. The necessity of these genes is essential for the adaptation or interference steps, thus making these systems rather unlikely to be fully functional (Cass et al., 2015). The hypothesis based on these observations suggests a possible ongoing termination of the CRISPR systems in these strains.

As mentioned, the direct repeats have a regulatory role in the adaptation process, coupling specific sets of cas genes to arrays. These arrays contain direct repeats with lengths varying between 19-48bp for bacteria and can currently be assigned to 40 individual sequence families in CRISPR map v1.3.0-2013 (Lange et al., 2013). The complete genome reference strain

0609139 of *A. wodanis* was found to contain two CRISPR arrays in separate positions in its genome. The other strains contained either one or two arrays where one of these was located downstream of the cas genes. Still, both of these arrays may be active even if one is remote to the acting cas genes. Both arrays contained an identical repeat length of 28bp and constituted similar secondary structures associating with the same I-F system. The number of spacers in the two arrays was found to be rather similar proving both array positions have been or is actively maintained by the system. The main hypothesis concerning the distant array regards the possibility that it has been horizontally transferred or moved with transposases in an earlier event. This hypothesis builds on the arrangement seen in strain Vw11 of *A. wodanis* where array 1, with two of its syntenic neighboring genes, and Array 2 flank the cas genes. This greatly contrasts the arrangement seen in reference strain 0609139.

The average strain of *A. salmonicida* was not observed with any proper CRISPR system even if a few were in possession of solo effector complexes (described below). The exception applied to environmental strain B9-15, which carried the same system discovered in *A. wodanis*. Its sequence family of the CRISPR array and the gene synteny matched *A. wodanis*, in particular strain Vw11. These findings supports the fact that *A. salmonicida* can also obtain the CRISPR system under circumstances that allow uptake and use.

Irregular systems were observed in several strains

Solo effector complexes are thought to have evolved from the main systems, often I-B or I-F, and are commonly associated with transposon activity or plasmids (Makarova et al., 2015). Such complexes were discovered in 6 of the 45 strains in this project and featured the *cse2* and *cas7* genes. This arrangement is not previously described but both genes, especially *cse2*, are specific for the I-E system. These were usually identified with questionable CRISPR arrays by CRISPRfinder and contained few repeats if any at all. It is unlikely that these constitute any function for the cell and probably exist as a reminder of how plastic and unrestrained the horizontal flow of genetic elements is. Nevertheless, these were found to exist within four separate species and should have a common origin.

Strain R8-63 of *A. magni* was one six strains with a solo effector complex, but the only one with a confirmed CRISPR array possessing three spacers. Still, it had a highly fragmented and incomplete CRISPR system not seen in any other strain in the project nor described. It was lacking genes for spacer insertion, target cleavage and effector molecules as well as having cas genes related to system III-D, all located amidst dozens of hypothetical genes. This unordered arrangement is likely not in working condition and might be hypothesized as remnants of system I or III which has been evolving in unstable vectors before being inserted in this strain.

Conclusion

Knowledge about the *Aliivibrio* has been limited, hence this study has focused on comparative analysis of this genus to gain insight about its species variation. From the initial steps of genome assembly and annotation we showed significant deviations in both chromosome and genome size, emphasizing a wide diversity.

Applying multi locus sequence analysis of six concatenated housekeeping genes we illustrated sufficient discriminatory power to unambiguously separate *Aliivibrio*, *Vibrio* and *Photobacteria* as well as their species. The magnitude of resolution paved way for five new proposed species; the *A. Magni* sp. nov., *A. Thrudae* sp. nov., *A. Modi* sp. nov., *A. Friggae* sp. nov and *A. Raniae* sp. nov.

Further pan-genome analysis demonstrated a conserved core of genes representing 46% of the average *Aliivibrio* genome. Abundant number of dispensable and strain-specific genes indicated a great variability outside of the core genome, where computed models signal a still open pan-genome. We also showed quorum sensing to be partially conserved in *Aliivibrio* species, identified duplicated *ainR/ainS* systems and proved *A. fischeri* is not self-evident to express bioluminescence. In addition, repeat sequences have been linked to the CRISPR defense system in *A. wodanis*. The system, identified as Type I-F, is likely familiar in the *Aliivibrio* genus due to its additional presence in the environmental strain B9-15 of *A. salmonicida*.

References

AGGARWAL, G. & RAMASWAMY, R. 2002. Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER. *J Biosci,* 27**,** 7-14.

ASSEFA, S., KEANE, T. M., OTTO, T. D., NEWBOLD, C. & BERRIMAN, M. 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics,* 25**,** 1968-9.

AST, J. C., URBANCZYK, H. & DUNLAP, P. V. 2009. Multi-gene analysis reveals previously unrecognized phylogenetic diversity in *Aliivibrio*. *Syst Appl Microbiol,* 32**,** 379-86.

ATKINSON, S., CHANG, C. Y., SOCKETT, R. E., CAMARA, M. & WILLIAMS, P. 2006. Quorum sensing in *Yersinia enterocolitica* controls swimming and swarming motility. *J Bacteriol,* 188**,** 1451-61.

AZIZ, R. K., BARTELS, D., BEST, A. A., DEJONGH, M., DISZ, T., EDWARDS, R. A., FORMSMA, K., GERDES, S., GLASS, E. M., KUBAL, M., MEYER, F., OLSEN, G. J., OLSON, R., OSTERMAN, A. L., OVERBEEK, R. A., MCNEIL, L. K., PAARMANN, D., PACZIAN, T., PARRELLO, B., PUSCH, G. D., REICH, C., STEVENS, R., VASSIEVA, O., VONSTEIN, V., WILKE, A. & ZAGNITKO, O. 2008. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics,* 9**,** 75.

BANG, S. S., BAUMANN, P. & NEALSON, K. H. 1978. Phenotypic characterization of *Photobacterium logei* (sp. nov.), a species related to *P. fischeri. Curr. Microbiol. 1:285–288.*

BJELLAND, A. M., SORUM, H., TEGEGNE, D. A., WINTHER-LARSEN, H. C., WILLASSEN, N. P. & HANSEN, H. 2012. LitR of *Vibrio salmonicida* is a salinity-sensitive quorum-sensing regulator of phenotypes involved in host interactions and virulence. *Infect Immun,* 80**,** 1681-9.

BOX, A. M., MCGUFFIE, M. J., O'HARA, B. J. & SEED, K. D. 2015. Functional Analysis of Bacteriophage Immunity through a Type I-E CRISPR-Cas System in *Vibrio cholerae* and Its Application in Bacteriophage Genome Engineering. *J Bacteriol,* 198**,** 578-90.

CAROLYNE, L. 1999. *The Poetic Edda* Oxford University Press.

CARVER, T., BERRIMAN, M., TIVEY, A., PATEL, C., BOHME, U., BARRELL, B. G., PARKHILL, J. & RAJANDREAM, M. A. 2008. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics,* 24**,** 2672-6.

CASS, S. D., HAAS, K. A., STOLL, B., ALKHNBASHI, O. S., SHARMA, K., URLAUB, H., BACKOFEN, R., MARCHFELDER, A. & BOLT, E. L. 2015. The role of Cas8 in type I CRISPR interference. *Biosci Rep,* 35.

CHAUDHARI, N. M., GUPTA, V. K. & DUTTA, C. 2016. BPGA- an ultra-fast pan-genome analysis pipeline. *Sci Rep,* 6**,** 24373.

CHEN, T. W., GAN, R. C., CHANG, Y. F., LIAO, W. C., WU, T. H., LEE, C. C., HUANG, P. J., LEE, C. Y., CHEN, Y. Y., CHIU, C. H. & TANG, P. 2015. Is the whole greater than the sum of its parts? De novo assembly strategies for bacterial genomes based on paired-end sequencing. *BMC Genomics,* 16**,** 648.

CLARK, K., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J. & SAYERS, E. W. 2016. GenBank. *Nucleic Acids Res,* 44**,** D67-72.

COCK, P. J., FIELDS, C. J., GOTO, N., HEUER, M. L. & RICE, P. M. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res,* 38**,** 1767-71.

CONTRERAS-MOREIRA, B. & VINUESA, P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl Environ Microbiol,* 79**,** 7696-701.

COPLEY, J. 2002. All at sea. *Nature,* 415**,** 572-4.

CORDERO, O. X., WILDSCHUTTE, H., KIRKUP, B., PROEHL, S., NGO, L., HUSSAIN, F., LE ROUX, F., MINCER, T. & POLZ, M. F. 2012. Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance. *Science,* 337**,** 1228-31.

DUNLAP, P. V. 1999. Quorum regulation of luminescence in *Vibrio fischeri*. *J Mol Microbiol Biotechnol,* 1**,** 5-12.

FEDERHEN, S. 2012. The NCBI Taxonomy database. *Nucleic Acids Res,* 40**,** D136-43.

FIDOPIASTIS, P. M., SORUM, H. & RUBY, E. G. 1999. Cryptic luminescence in the cold-water fish pathogen *Vibrio salmonicida*. *Arch Microbiol,* 171**,** 205-9.

FREEMAN, J. A. & BASSLER, B. L. 1999. Sequence and function of LuxU: a two-component phosphorelay protein that regulates quorum sensing in *Vibrio harveyi*. *J Bacteriol,* 181**,** 899-906.

FREEMAN, J. A., LILLEY, B. N. & BASSLER, B. L. 2000. A genetic analysis of the functions of LuxN: a two-component hybrid sensor kinase that regulates quorum sensing in *Vibrio harveyi*. *Mol Microbiol,* 35**,** 139-49.

FUQUA, W. C., WINANS, S. C. & GREENBERG, E. P. 1994. Quorum sensing in bacteria: the LuxR-LuxI family of cell density-responsive transcriptional regulators. *J Bacteriol,* 176**,** 269-75.

GOECKS, J., NEKRUTENKO, A., TAYLOR, J. & GALAXY, T. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol,* 11**,** R86.

GOMEZ-GIL, B., THOMPSON, F. L., THOMPSON, C. C. & SWINGS, J. 2003. *Vibrio pacinii* sp. nov., from cultured aquatic organisms. *Int J Syst Evol Microbiol,* 53**,** 1569-73.

GRISSA, I., VERGNAUD, G. & POURCEL, C. 2007a. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics,* 8**,** 172.

GRISSA, I., VERGNAUD, G. & POURCEL, C. 2007b. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res,* 35**,** W52-7.

GUIMARAES, L. C., FLORCZAK-WYSPIANSKA, J., DE JESUS, L. B., VIANA, M. V., SILVA, A., RAMOS, R. T., SOARES SDE, C. & SOARES SDE, C. 2015. Inside the Pan-genome - Methods and Software Overview. *Curr Genomics,* 16**,** 245-52.

GUO, Y., YE, F., SHENG, Q., CLARK, T. & SAMUELS, D. C. 2014. Three-stage quality control strategies for DNA re-sequencing data. *Brief Bioinform,* 15**,** 879-89.

HALL, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series,* 41**,** 4.

HANSEN, H., BJELLAND, A. M., RONESSEN, M., ROBERTSEN, E. & WILLASSEN, N. P. 2014. LitR is a repressor of syp genes and has a temperature-sensitive regulatory effect on biofilm formation and colony morphology in *Vibrio* (*Aliivibrio*) *salmonicida*. *Appl Environ Microbiol,* 80**,** 5530-41.

HANSEN, H., PUROHIT, A. A., LEIROS, H. K., JOHANSEN, J. A., KELLERMANN, S. J., BJELLAND, A. M. & WILLASSEN, N. P. 2015. The autoinducer synthases LuxI and AinS are responsible for temperature-dependent AHL production in the fish pathogen *Aliivibrio salmonicida*. *BMC Microbiol,* 15**,** 69.

HAYDEN, E. C. 2014. Technology: The $1,000 genome. *Nature,* 507**,** 294-5.

HJERDE, E., KARLSEN, C., SORUM, H., PARKHILL, J., WILLASSEN, N. P. & THOMSON, N. R. 2015. Co-cultivation and transcriptome sequencing of two co-existing fish pathogens *Moritella viscosa* and *Aliivibrio wodanis*. *BMC Genomics,* 16**,** 447.

HJERDE, E., LORENTZEN, M. S., HOLDEN, M. T., SEEGER, K., PAULSEN, S., BASON, N., CHURCHER, C., HARRIS, D., NORBERTCZAK, H., QUAIL, M. A., SANDERS, S., THURSTON, S., PARKHILL, J., WILLASSEN, N. P. & THOMSON, N. R. 2008. The genome sequence of the fish pathogen Aliivibrio salmonicida strain LFI1238 shows extensive evidence of gene decay. *BMC Genomics,* 9**,** 616.

HUSON, D. H. & BRYANT, D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol,* 23**,** 254-67.

HYATT, D., CHEN, G. L., LOCASCIO, P. F., LAND, M. L., LARIMER, F. W. & HAUSER, L. J. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics,* 11**,** 119.

JORDAN, I. K., MAKAROVA, K. S., SPOUGE, J. L., WOLF, Y. I. & KOONIN, E. V. 2001. Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res,* 11**,** 555-65.

KAAS, R. S., FRIIS, C., USSERY, D. W. & AARESTRUP, F. M. 2012. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC Genomics,* 13**,** 577.

KAHLKE, T., GOESMANN, A., HJERDE, E., WILLASSEN, N. P. & HAUGEN, P. 2012. Unique core genomes of the bacterial family *vibrionaceae*: insights into niche adaptation and speciation. *BMC Genomics,* 13**,** 179.

KANEHISA, M. 2016. KEGG Bioinformatics Resource for Plant Genomics and Metabolomics. *Methods Mol Biol,* 1374**,** 55-70.

KANEHISA, M., SATO, Y. & MORISHIMA, K. 2016. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol,* 428**,** 726-31.

KHRULNOVA, S. A., BARANOVA, A., BAZHENOV, S. V., GORYANIN, II, KONOPLEVA, M. N., MARYSHEV, I. V., SALYKHOVA, A. I., VASILYEVA, A. V., MANUKHOV, I. V. & ZAVILGELSKY, G. B. 2016. Lux-operon of the Marine Psychrophilic Bacteria *Aliivibrio logei*: a Comparative Analysis of the LuxR1/LuxR2 Regulatory Activity in *Escherichia coli* cells. *Microbiology*.

KIM, H. & KIM, J. S. 2014. A guide to genome engineering with programmable nucleases. *Nat Rev Genet,* 15**,** 321-34.

KIMBROUGH, J. H. & STABB, E. V. 2013. Substrate specificity and function of the pheromone receptor AinR in *Vibrio fischeri* ES114. *J Bacteriol,* 195**,** 5223-32.

KONOPLEVA, M. N., KHRULNOVA, S. A., BARANOVA, A., EKIMOV, L. V., BAZHENOV, S. V., GORYANIN, II & MANUKHOV, I. V. 2016. A combination of luxR1 and luxR2 genes activates Pr-promoters of psychrophilic *Aliivibrio logei* lux-operon independently of chaperonin GroEL/ES and protease Lon at high concentrations of autoinducer. *Biochem Biophys Res Commun*.

KOONIN, E. V., MAKAROVA, K. S. & ARAVIND, L. 2002. Horizontal Gene Transfer in Prokaryotes: Quantification and Classification. . *Annual Reviews Collection, National Center for Biotechnology Information (US).*

KOREN, S. & PHILLIPPY, A. M. 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr Opin Microbiol,* 23**,** 110-20.

LAN, R. & REEVES, P. R. 1996. Gene transfer is a major factor in bacterial evolution. *Mol Biol Evol,* 13**,** 47-55.

LANGE, S. J., ALKHNBASHI, O. S., ROSE, D., WILL, S. & BACKOFEN, R. 2013. CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Res,* 41**,** 8034-44.

LENZ, D. H. & BASSLER, B. L. 2007. The small nucleoid protein Fis is involved in Vibrio cholerae quorum sensing. *Mol Microbiol,* 63**,** 859-71.

LENZ, D. H., MOK, K. C., LILLEY, B. N., KULKARNI, R. V., WINGREEN, N. S. & BASSLER, B. L. 2004. The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in *Vibrio harveyi* and *Vibrio cholerae*. *Cell,* 118**,** 69-82.

LI, L., STOECKERT, C. J., JR. & ROOS, D. S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res,* 13**,** 2178-89.

LUPP, C. & RUBY, E. G. 2004. *Vibrio fischeri* LuxS and AinS: comparative study of two signal synthases. *J Bacteriol,* 186**,** 3873-81.

MA, Y., ZHANG, L. & HUANG, X. 2014. Genome modification by CRISPR/Cas9. *FEBS J,* 281**,** 5186-93.

MAKAROVA, K. S., HAFT, D. H., BARRANGOU, R., BROUNS, S. J., CHARPENTIER, E., HORVATH, P., MOINEAU, S., MOJICA, F. J., WOLF, Y. I., YAKUNIN, A. F., VAN DER OOST, J. & KOONIN, E. V. 2011. Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol,* 9**,** 467-77.

MAKAROVA, K. S., WOLF, Y. I., ALKHNBASHI, O. S., COSTA, F., SHAH, S. A., SAUNDERS, S. J., BARRANGOU, R., BROUNS, S. J., CHARPENTIER, E., HAFT, D. H., HORVATH, P., MOINEAU, S., MOJICA, F. J., TERNS, R. M., TERNS, M. P., WHITE, M. F., YAKUNIN, A. F., GARRETT, R. A., VAN DER OOST, J., BACKOFEN, R. & KOONIN, E. V. 2015. An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol,* 13**,** 722-36.

MANUKHOV, I. V., KHRUL'NOVA, S. A., BARANOVA, A. & ZAVILGELSKY, G. B. 2011. Comparative analysis of the lux operons in *Aliivibrio logei* KCh1 (a Kamchatka Isolate) and *Aliivibrio salmonicida*. *J Bacteriol,* 193**,** 3998-4001.

MARGARET, C., ROSS. 1994. *Prolonged echoes I, Old Norse myths in medieval northern society*

MARRAFFINI, L. A. 2015. CRISPR-Cas immunity in prokaryotes. *Nature,* 526**,** 55-61.

MARTINEZ-CANO, D. J., REYES-PRIETO, M., MARTINEZ-ROMERO, E., PARTIDA-MARTINEZ, L. P., LATORRE, A., MOYA, A. & DELAYE, L. 2014. Evolution of small prokaryotic genomes. *Front Microbiol,* 5**,** 742.

MEDINI, D., DONATI, C., TETTELIN, H., MASIGNANI, V. & RAPPUOLI, R. 2005. The microbial pan-genome. *Curr Opin Genet Dev,* 15**,** 589-94.

MEIGHEN, E. A. 1991. Molecular biology of bacterial bioluminescence. *Microbiol Rev,* 55**,** 123-42.

MEIGHEN, E. A. 1993. Bacterial bioluminescence: organization, regulation, and application of the lux genes. *FASEB J,* 7**,** 1016-22.

MILTON, D. L. 2006. Quorum sensing in vibrios: complexity for diversification. *Int J Med Microbiol,* 296**,** 61-71.

MIYASHIRO, T. & RUBY, E. G. 2012. Shedding light on bioluminescence regulation in *Vibrio fischeri*. *Mol Microbiol,* 84**,** 795-806.

MORENO-HAGELSIEB, G., TREVINO, V., PEREZ-RUEDA, E., SMITH, T. F. & COLLADO-VIDES, J. 2001. Transcription unit conservation in the three domains of life: a perspective from *Escherichia coli*. *Trends Genet,* 17**,** 175-7.

NEALSON, K. H., PLATT, T. & HASTINGS, J. W. 1970. Cellular control of the synthesis and activity of the bacterial luminescent system. *J Bacteriol,* 104**,** 313-22.

NELSON, E. J., TUNSJO, H. S., FIDOPIASTIS, P. M., SORUM, H. & RUBY, E. G. 2007. A novel lux operon in the cryptically bioluminescent fish pathogen *Vibrio salmonicida* is associated with virulence. *Appl Environ Microbiol,* 73**,** 1825-33.

NUNEZ, J. K., KRANZUSCH, P. J., NOESKE, J., WRIGHT, A. V., DAVIES, C. W. & DOUDNA, J. A. 2014. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat Struct Mol Biol,* 21**,** 528-34.

O'GRADY, E. A. & WIMPEE, C. F. 2008. Mutations in the lux operon of natural dark mutants in the genus *Vibrio*. *Appl Environ Microbiol,* 74**,** 61-6.

OKONECHNIKOV, K., GOLOSOVA, O., FURSOV, M. & TEAM, U. 2012. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics,* 28**,** 1166-7.

PAREEK, C. S., SMOCZYNSKI, R. & TRETYN, A. 2011. Sequencing technologies and genome sequencing. *J Appl Genet,* 52**,** 413-35.

RATH, D., AMLINGER, L., RATH, A. & LUNDGREN, M. 2015. The CRISPR-Cas immune system: biology, mechanisms and applications. *Biochimie,* 117**,** 119-28.

READING, N. C. & SPERANDIO, V. 2006. Quorum sensing: the many languages of bacteria. *FEMS Microbiol Lett,* 254**,** 1-11.

REAMS, A. B., KOFOID, E., DULEBA, N. & ROTH, J. R. 2014. Recombination and annealing pathways compete for substrates in making rrn duplications in *Salmonella enterica*. *Genetics,* 196**,** 119-35.

REAMS, A. B. & ROTH, J. R. 2015. Mechanisms of gene duplication and amplification. *Cold Spring Harb Perspect Biol,* 7**,** a016592.

RICHARDSON, E. J. & WATSON, M. 2013. The automatic annotation of bacterial genomes. *Brief Bioinform,* 14**,** 1-12.

RILEY, M. A. & LIZOTTE-WANIEWSKI, M. 2009. Population genomics and the bacterial species concept. *Methods Mol Biol,* 532**,** 367-77.

ROSSELLO-MORA, R. & AMANN, R. 2015. Past and future species definitions for Bacteria and Archaea. *Syst Appl Microbiol,* 38**,** 209-16.

RUDOLF, S. 2008. *A Dictionary of Northern Mythology*.

SARKAR, S. F. & GUTTMAN, D. S. 2004. Evolution of the core genome of *Pseudomonas syringae*, a highly clonal, endemic plant pathogen. *Appl Environ Microbiol,* 70**,** 1999-2012.

SAWABE, T., KITA-TSUKAMOTO, K. & THOMPSON, F. L. 2007. Inferring the evolutionary history of vibrios by means of multilocus sequence analysis. *J Bacteriol,* 189**,** 7932-6.

SAWABE, T., OGURA, Y., MATSUMURA, Y., FENG, G., AMIN, A. K., MINO, S., NAKAGAWA, S., SAWABE, T., KUMAR, R., FUKUI, Y., SATOMI, M., MATSUSHIMA, R., THOMPSON, F. L., GOMEZ GIL, B., CHRISTEN, R., MARUYAMA, F., KUROKAWA, K. & HAYASHI, T. 2014. Corrigendum: Updating the *Vibrio* clades defined by multilocus sequence phylogeny: proposal of eight new clades, and the description of *Vibrio tritonius* sp. nov. *Front Microbiol,* 5**,** 583.

SISTI, F., HA, D. G., O'TOOLE, G. A., HOZBOR, D. & FERNANDEZ, J. 2013. Cyclic-di-GMP signalling regulates motility and biofilm formation in *Bordetella bronchiseptica*. *Microbiology,* 159**,** 869-79.

SIVASHANKARI, S. & SHANMUGHAVEL, P. 2007. Comparative genomics - a perspective. *Bioinformation,* 1**,** 376-8.

SOREK, R., KUNIN, V. & HUGENHOLTZ, P. 2008. CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol,* 6**,** 181-6.

STEVENS, A. M., DOLAN, K. M. & GREENBERG, E. P. 1994. Synergistic binding of the *Vibrio fischeri* LuxR transcriptional activator domain and RNA polymerase to the lux promoter region. *Proc Natl Acad Sci U S A,* 91**,** 12619-23.

STUART, M. B. 2012. Sequencing-by-Synthesis: Explaining the Illumina Sequencing Technology. *Bitesize Bio*.

SUN, H., LI, Y., SHI, X., LIN, Y., QIU, Y., ZHANG, J., LIU, Y., JIANG, M., ZHANG, Z., CHEN, Q., SUN, Q. & HU, Q. 2015. Association of CRISPR/Cas evolution with *Vibrio parahaemolyticus* virulence factors and genotypes. *Foodborne Pathog Dis,* 12**,** 68-73.

TAMURA, K., STECHER, G., PETERSON, D., FILIPSKI, A. & KUMAR, S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol,* 30**,** 2725-9.

TANG, H., BOMHOFF, M. D., BRIONES, E., ZHANG, L., SCHNABLE, J. C. & LYONS, E. 2015. SynFind: Compiling Syntenic Regions across Any Set of Genomes on Demand. *Genome Biol Evol,* 7**,** 3286-98.

TEAM, R. D. C. 2008. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.*

TETTELIN, H., MASIGNANI, V., CIESLEWICZ, M. J., DONATI, C., MEDINI, D., WARD, N. L., ANGIUOLI, S. V., CRABTREE, J., JONES, A. L., DURKIN, A. S., DEBOY, R. T., DAVIDSEN, T. M., MORA, M., SCARSELLI, M., MARGARIT Y ROS, I., PETERSON, J. D., HAUSER, C. R., SUNDARAM, J. P., NELSON, W. C., MADUPU, R., BRINKAC, L. M., DODSON, R. J., ROSOVITZ, M. J., SULLIVAN, S. A., DAUGHERTY, S. C., HAFT, D. H., SELENGUT, J., GWINN, M. L., ZHOU, L., ZAFAR, N., KHOURI, H., RADUNE, D., DIMITROV, G., WATKINS, K., O'CONNOR, K. J., SMITH, S., UTTERBACK, T. R., WHITE, O., RUBENS, C. E., GRANDI, G., MADOFF, L. C., KASPER, D. L., TELFORD, J. L., WESSELS, M. R., RAPPUOLI, R. & FRASER, C. M. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A,* 102**,** 13950-5.

THOMPSON, F. L., GEVERS, D., THOMPSON, C. C., DAWYNDT, P., NASER, S., HOSTE, B., MUNN, C. B. & SWINGS, J. 2005. Phylogeny and molecular identification of vibrios on the basis of multilocus sequence analysis. *Appl Environ Microbiol,* 71**,** 5107-15.

THOMPSON, F. L., GOMEZ-GIL, B., VASCONCELOS, A. T. & SAWABE, T. 2007. Multilocus sequence analysis reveals that *Vibrio harveyi* and *V. campbellii* are distinct species. *Appl Environ Microbiol,* 73**,** 4279-85.

THOMPSON, F. L., IIDA, T. & SWINGS, J. 2004. Biodiversity of vibrios. *Microbiol Mol Biol Rev,* 68**,** 403-31, table of contents.

THOMPSON, J. D., GIBSON, T. J. & HIGGINS, D. G. 2002. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics,* Chapter 2**,** Unit 2 3.

TRIVEDI, U. H., CEZARD, T., BRIDGETT, S., MONTAZAM, A., NICHOLS, J., BLAXTER, M. & GHARBI, K. 2014. Quality control of next-generation sequencing data without a reference. *Front Genet,* 5**,** 111.

URBANCZYK, H., AST, J. C., HIGGINS, M. J., CARSON, J. & DUNLAP, P. V. 2007. Reclassification of *Vibrio fischeri*, *Vibrio logei*, *Vibrio salmonicida* and *Vibrio wodanis* as *Aliivibrio fischeri* gen. nov., comb. nov., *Aliivibrio logei* comb. nov., *Aliivibrio salmonicida* comb. nov. and *Aliivibrio wodanis* comb. nov. *Int J Syst Evol Microbiol,* 57**,** 2823-9.

VAN DIJK, E. L., AUGER, H., JASZCZYSZYN, Y. & THERMES, C. 2014. Ten years of next-generation sequencing technology. *Trends Genet,* 30**,** 418-26.

VERMA, S. C. & MIYASHIRO, T. 2013. Quorum sensing in the squid-*Vibrio* symbiosis. *Int J Mol Sci,* 14**,** 16386-401.

VESTH, T., LAGESEN, K., ACAR, O. & USSERY, D. 2013. CMG-biotools, a free workbench for basic comparative microbial genomics. *PLoS One,* 8**,** e60120.

VIA, S. 2009. Natural selection in action during speciation. *Proc Natl Acad Sci U S A,* 106 Suppl 1**,** 9939-46.

WICKHAM, H. 2009. ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York.*

WILLIAMS, P., WINZER, K., CHAN, W. C. & CAMARA, M. 2007. Look who's talking: communication and quorum sensing in the bacterial world. *Philos Trans R Soc Lond B Biol Sci,* 362**,** 1119-34.

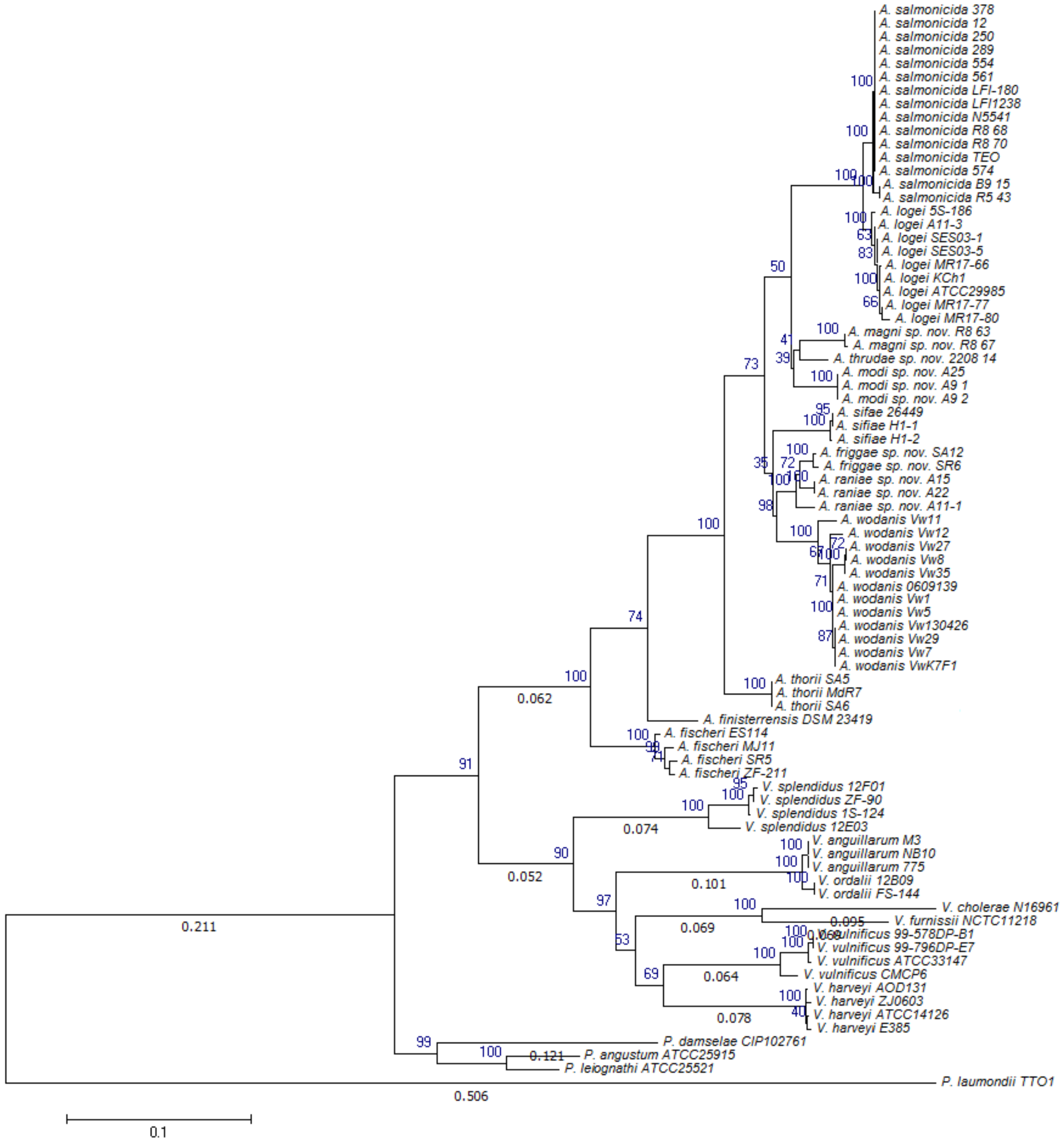
YAKOVCHUK, P., PROTOZANOVA, E. & FRANK-KAMENETSKII, M. D. 2006. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res,* 34**,** 564-74.

YARZA, P., YILMAZ, P., PRUESSE, E., GLOCKNER, F. O., LUDWIG, W., SCHLEIFER, K. H., WHITMAN, W. B., EUZEBY, J., AMANN, R. & ROSSELLO-MORA, R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol,* 12**,** 635-45.

YOSEF, I., GOREN, M. G. & QIMRON, U. 2012. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res,* 40**,** 5569-76.

ZHOU, X. & ROKAS, A. 2014. Prevention, diagnosis and treatment of high-throughput sequencing data pathologies. *Mol Ecol,* 23**,** 1679-700.

Appendix



Appendix figure 1. Phylogenetic tree based on the concatenated genes of *16S rRNA*, *gapA*, *gyrB*, *pyrH*, *recA* and *rpoA*. Parameters were set to apply all sites and perform 100 bootstrap replications. Robustness is shown in marine blue and branch lengths above 0.05 substitutions per site are shown.

R code for violin-box plots

The following code lines were executed in R, applying the ggplot module, after importing the matrix values and arranging them in data frames (df).

Formula representing the individual loci and MLSA (Figure 4-A):

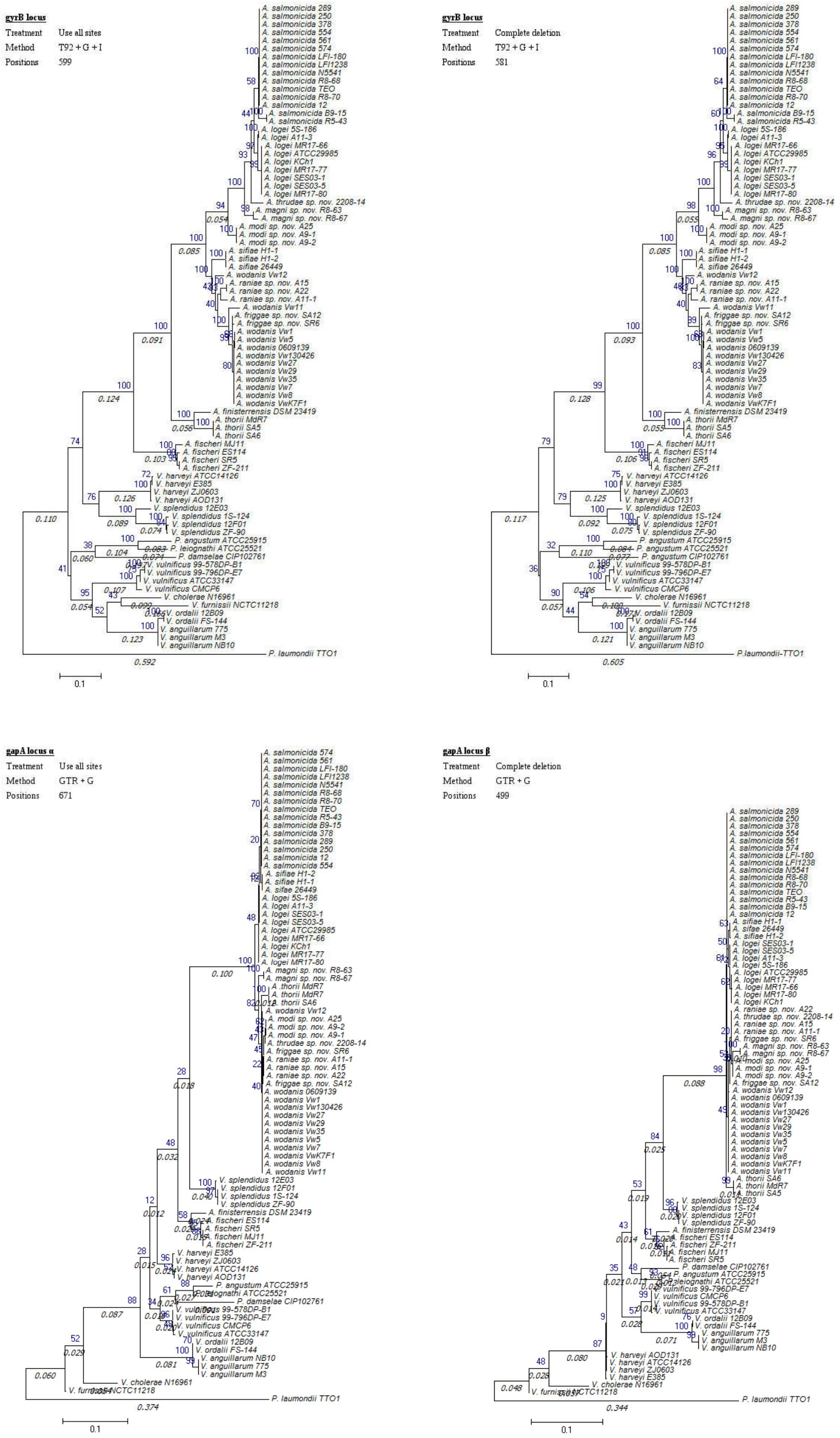
```
ggplot(stack(df), aes(x = ind, y = values, fill=ind))
       + geom_violin()
       + geom_boxplot(width=0.05,fill="white",outlier.size = 1.5,
         outlier.shape=3,outlier.colour="#CC0000")
       + stat_summary(fun.y=mean, geom="point",shape=4, size=2.5, color="black")
       + scale_fill_manual(values=c("#FFCC99", "#FFCCCC", "#FFCCFF",
        "#99CCFF","#CCFFFF", "#CCFFCC", "#FFFFFF"))
       + ylim(0.5, 1.01)
```

df = data frame containing all matrix values of each individual locus (*16S rRNA*, *gapA*, *gyrB*, *pyrH*, *recA* and *rpoA*).

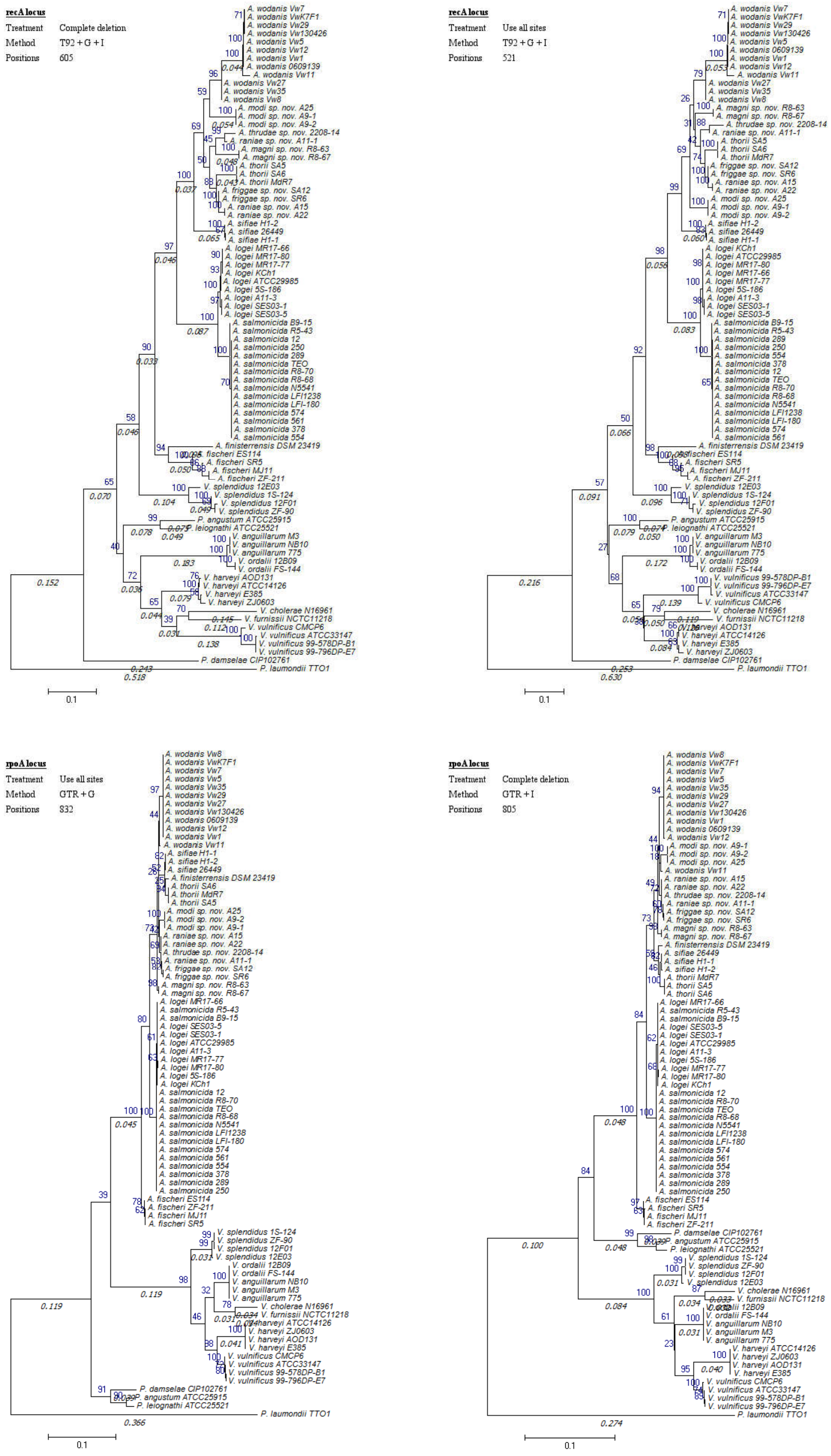
Formula representing the ANI and AAI of the pan-genome analysis (Figure 4-B):

```
ggplot(stack(dfo), aes(x = ind, y = values, fill=ind))
       + geom_violin()
       + geom_boxplot(width=0.05,fill="white",outlier.size = 1.5,
         outlier.shape=3,outlier.colour="#CC0000")
       + stat_summary(fun.y=mean, geom="point",shape=4, size=2.5, color="black")
       + scale_fill_manual(values=c("#3399FF", "#0066FF")) + ylim(50, 101)
```

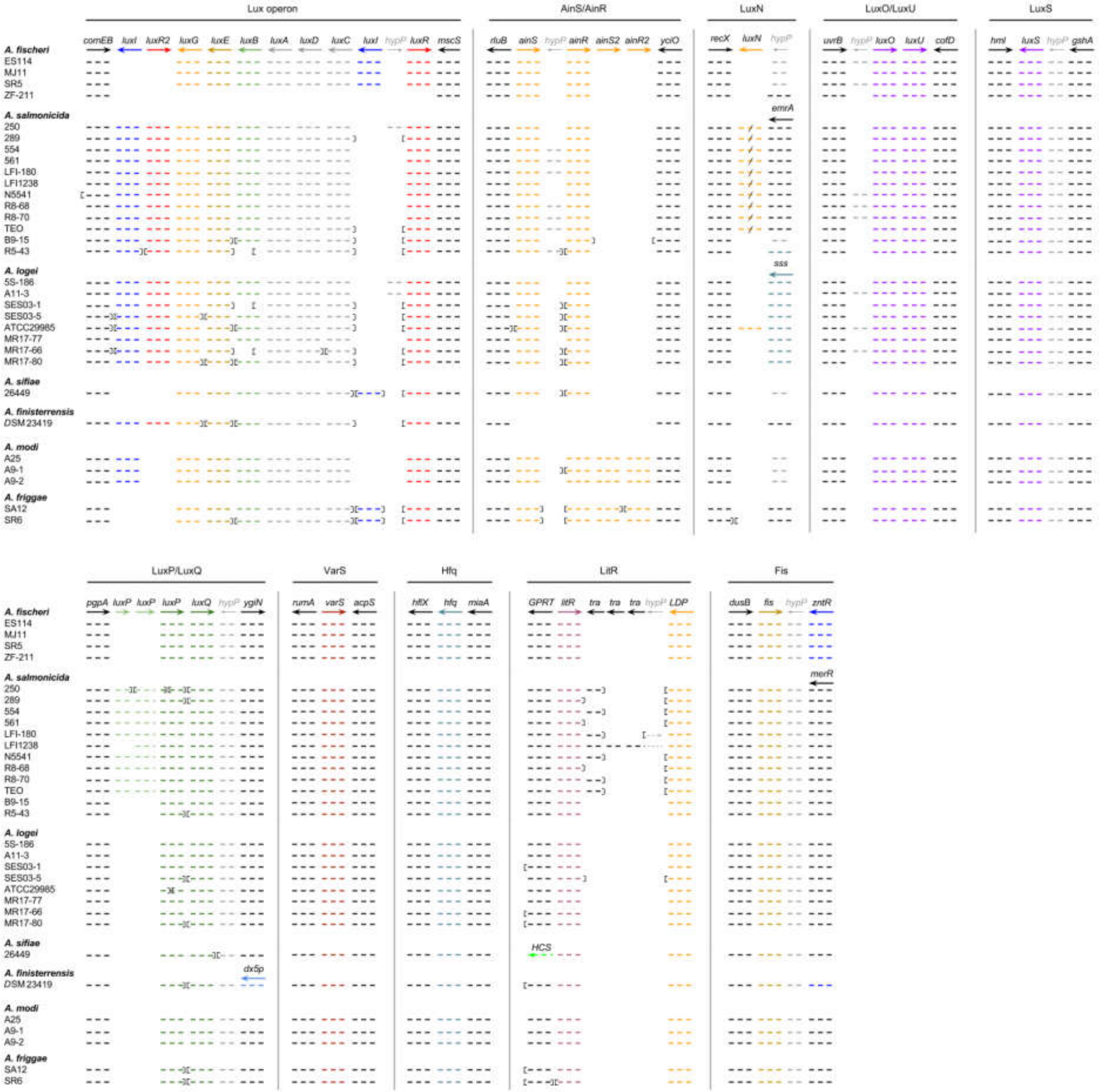
dfo = data frame containing all matrix values of ANI and AAI.



Appendix figure 2. Phylogenetic trees constructed on the basis of the *gyrB* and *gapA* loci where all sites (left) and complete deletion (right) were applied. Shown in marine blue is bootstrap supports based on 100 replications for the 81 included taxa.



Appendix figure 3. Phylogenetic trees based on the individual *recA* and *rpoA* gene loci for 81 taxa of *Aliivibrio*, *Vibrio* and *Photobacteria*. Bootstrap supports of 100 replications are shown in marine blue.

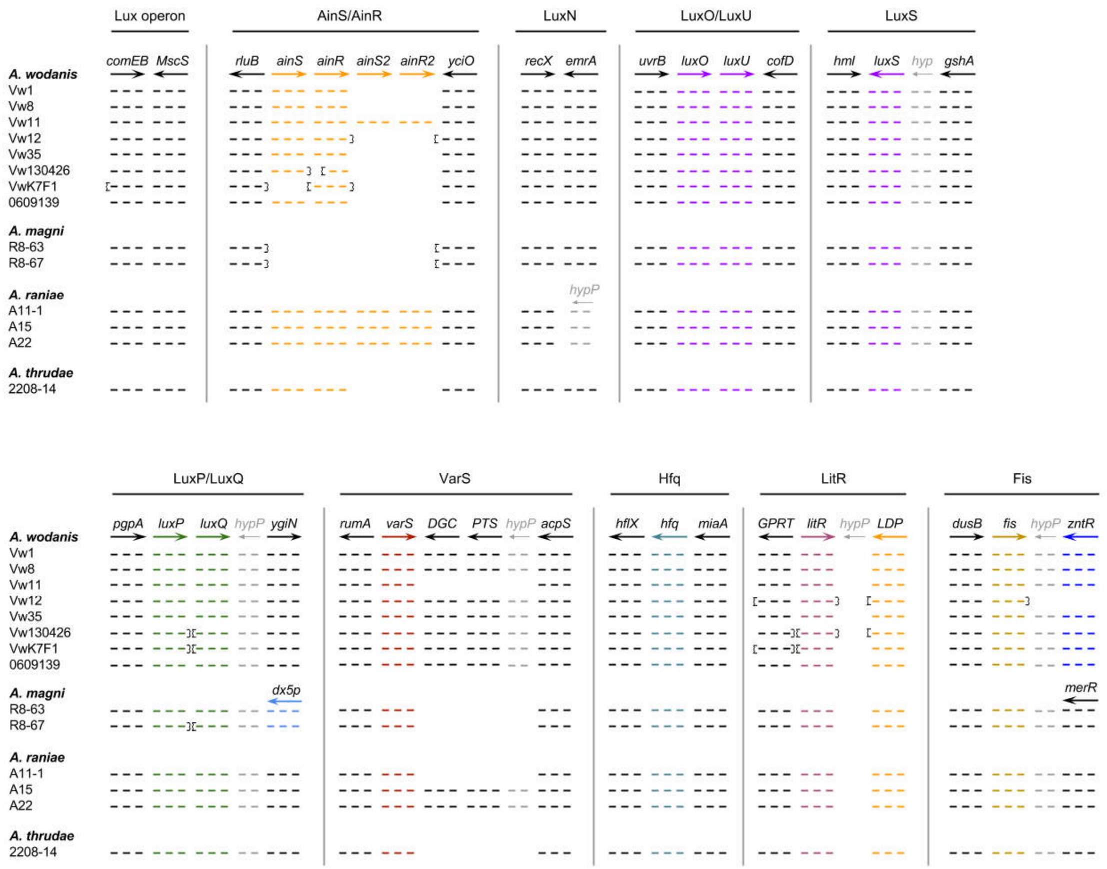


Legend

→	Transcriptional direction of gene
)(	Start/endpoint indicator of contigs
- - -	Gene exists for the given strain and follows direction given in the color code
- ⁄ -	Gene is observed with frameshift mutation
[- - -]	A single gene is the only item on the given contig

Abbreviation	Product	EC number
acpS	Holo-[acyl-carrier protein] synthase	EC 2.7.8.7
arcD	Arginine/ornithine antiporter ArcD	
cofD	FIG002813: LPPG:FO 2-phospho-L-lactate transferase like, CofD-like	
comEB	dCMP deaminase	EC 3.5.4.12
dusB (E. coli)	tRNA dihydrouridine synthase B	EC 1.-.-.-
dx5p	1-deoxy-D-xylulose-5-phosphate synthase activity	
emrA	Probable Co/Zn/Cd efflux system membrane fusion protein	
GPRT	Hypoxanthine-guanine phosphoribosyltransferase	EC 2.4.2.8
gshA	Glutamate--cysteine ligase	EC 6.3.2.2
HCS	Biotin operon repressor / Biotin-protein ligase	EC 6.3.4.15
hflX	GTP-binding protein HflX	
hml	Hemolysins and related proteins containing CBS domains	
hypP	Hypothetical protein	
LDP	Dihydrolipoamide dehydrogenase of pyruvate dehydrogenase complex	EC 1.8.1.4
mepA	Multi antimicrobial extrusion protein (Na(+)/drug antiporter), MATE family of MDR efflux pumps	
merR	Transcriptional regulator, MerR family	
miaA	tRNA dimethylallyltransferase	EC 2.5.1.75
mscS	Small-conductance mechanosensitive channel	
pgpA	Phosphatidylglycerophosphatase A	EC 3.1.3.27
recX	Regulatory protein recX	
rluB	Ribosomal large subunit pseudouridine synthase B	EC 4.2.1.70
rumA	23S rRNA (Uracil-5-) -methyltransferase RumA	EC 2.1.1.-
sss	Sodium-solute symporter, putative	
tra	Mobile element protein / Transposase	
uvrB	Excinuclease ABC subunit B	
yciO	Hypothetical YciO protein, TsaC/YrdC paralog	
ygiN	Antibiotic biosynthesis monooxygenase	
zntR	HTH-type transcriptional regulator zntR	

Appendix figure 4. Gene synteny in relation to the lux operon, AinR/S, LuxN, LuxO/U/, LuxP/Q/S and VarS systems as well as some of the auxiliary products Hfq, LitR and Fis. Species commonly carrying the lux operon in *Aliivibrio* are shown as they appeared in the CLC genome browser, focusing on the presence of genes on assembled contigs.



Legend

→	Transcriptional direction of gene
][	Start/endpoint indicator of contigs
- - -	Gene exists for the given strain and follows direction given in the color code
-/-	Gene is observed with frameshift mutation
[- - -]	A single gene is the only item on the given contig

Abbreviation	Product	EC number
arcD	Arginine/ornithine antiporter ArcD	
acpS	Holo-[acyl-carrier protein] synthase	EC 2.7.8.7
cofD	FIG002813: LPPG:FO 2-phospho-L-lactate transferase like, CofD-like	
comEB	dCMP deaminase	EC 3.5.4.12
DGC	Diguanylate cyclase/phosphodiesterase (GGDEF & EAL domains) with PAS/PAC sensor(s)	
dusB (E.coli)	tRNA dihydrouridine synthase B	EC 1.-.-.-
dxp5	1-deoxy-D-xylulose-5-phosphate synthase	
GPRT	Hypoxanthine-guanine phosphoribosyltransferase	EC 2.4.2.8
gshA	Glutamate--cysteine ligase	EC 6.3.2.2
hflX	GTP-binding protein HflX	
hml	Hemolysins and related proteins containing CBS domains	
hyp	hypothetical protein	
LDP	Dihydrolipoamide dehydrogenase of pyruvate dehydrogenase complex	EC 1.8.1.4
mepA	Multi antimicrobial extrusion protein (Na(+)/drug antiporter), MATE family of MDR efflux pumps	
miaA	tRNA dimethylallyltransferase	EC 2.5.1.75
mluB	Ribosomal large subunit pseudouridine synthase B	EC 4.2.1.70
mMerR	Transcriptional regulator, MerR family	
mscS	Small-conductance mechanosensitive channel	
pgpA	Phosphatidylglycerophosphatase A	EC 3.1.3.27
PTS	ABC-type phosphate transport system, periplasmic component	
rumA	23S rRNA (Uracil-5-) -methyltransferase RumA	EC 2.1.1.-
uvrB	Excinuclease ABC subunit B	
yciO	Hypothetical YciO protein, TsaC/YrdC paralog	
ygiN	Antibiotic biosynthesis monooxygenase	
zntR	HTH-type transcriptional regulator zntR	

Appendix figure 5. Supplementary to Appendix figure 4 showing the species with strains known to not carry the lux operon.