UNIVERSITY OF TROMSØ UIT

FACULTY OF HEALTH SCIENCES
DEPARTMENT OF MEDICAL BIOLOGY

Analysis of the *Vibrionaceae* Pan-Genome

Tim Kahlke

A dissertation for the degree of Philosophiae Doctor

April 2013

Analysis of the *Vibrionaceae* Pan-Genome

Tim Kahlke

Thesis for the degree of Philosophiae Doctor

Faculty of Health Sciences

Department of Medical Biology

University of Tromsø

9037 Tromsø

Norway

April 2013

Evaluating Committee:

Prof. Brian Austin
Institute of Aquaculture,
University of Stirling,
Stirling, Stirlingshire, FK9 4LA,
Scotland, United Kingdom
E-mail: brian.austin@stir.ac.uk

Prof. Finn Drabløs
Department of Cancer Research and
Molecular Medicine,
Norwegian University of Science and
Technology,
N-7006, Trondheim, Norway
E-mail: finn.drablos@ntnu.no

Prof. Johanna Ericson Sollid
Department of Medical Biology,
Faculty of Health Sciences,
University of Tromsø,
9037, Tromsø, Norway
E-mail: johanna.e.sollid@uit.no

Academic dissertation for the degree of Philosophiae Doctor in Natural Sciences to be presented for public criticism at Faculty of Health Sciences, University of Tromsø, Norway, on April 2013

This work was typeset using LaTeX 2ε

CONTENTS

ACKNOWLEDGEMENTS

The presented study was carried out at the Faculty of Health Sciences, Department of Medical Biology, University of Tromsø, Norway, from September 2007 to April 2013. Financial support for this study was provided by the University of Tromsø.

First and foremost I would like to thank my supervisors Professor Nils-Peder Willassen, Professor Peik Haugen, Professor Ingebrit Sylte and Jacob Koehler for giving me the opportunity to realize this thesis. I would also like to thank my group leader Professor Nils-Peder Willassen for the financial support that I received in the extension period of my thesis.

Special thanks go to Peik for guiding me through all phases of my PhD, from the moment you picked me up at the airport to the proof reading of this work. Thank you for always having an open door, for your endless patience with my constant resistance to the suggested changes in my manuscripts, for the tolerance regarding my "computer-guy" terminology, for the enlightening discussions about science, all the world and his brother, for your support, your knowledge and all the chocolate I got in your office. Thank you.

I would also like to thank Erik Hjerde for the help, discussions and for introducing me to norwegian culture and to rock climbing. It was highly appreciated.
Also, thanks to the rest of our focus group, Rafi, Espen, Chris and our special friend Peter. You guys made lunch an interesting experience.

Thanks to my fellow sufferers, especially to Aili, Annfrid, Alex, Jörn, Makoto (ima made iroiro arigatou), Man Kumari, Marc, Miriam and last but not least *Skirt Girl* a.k.a. Taiana for support, parties, fun, fights and all the weird stories that made the

stay in norway memorable. The same also counts for those who were not part of the PhD crew, specifically Adele and Rhys, Yvonne and all the colleagues and friends at the Department of Chemistry and the Department of Medical Biology.

Finally, I want to say thanks to Jasmin, Sven & Heike, Bo, Pit and my Mum just for being there. Thank you.

Tromsø, April 2013

Tim Kahlke

ABSTRACT

The vast advances in molecular genetics in the last two decades opened new and fascinating ways to study bacterial evolution on a genome level. Starting with the genome sequences of single isolates of particular clinical or economical importance, it is now possible to compare multiple genomes of closely related bacteria at once. The investigation of strains of even the same bacterial species allows the determination of specific genetic features and sheds light on the molecular processes of niche adaptation and bacterial speciation. In recent years the pan-genome concept became widely used to describe the diversity of groups of bacterial genomes. The determination of sets of conserved and unique genes enables the investigation of bacterial evolution on various levels, such as the determination of genes specific to any group of genomes or the identification of changes in the gene sequence of conserved genes.

In the presented work the bacterial family *Vibrionaceae* was used as a model to investigate bacterial diversity on a gene level and to analyze the underlying concepts of bacterial niche adaptation and evolution. First the pan-genome of a diverse dataset of *Vibrionaceae* genomes from various environments and temperature zones was determined and subsequently analyzed using existing as well as newly developed bioinformatic tools. In **Paper I** differences in the gene sets of groups of *Vibrionaceae* genomes were investigated to determine genes specific to particular taxa, i.e., species and genera. These genes contribute to specific metabolic and phenotypical traits and are not only important for clinical diagnostics but might also aid the demarcation of bacteria on a gene level. In **Paper II** the distribution of pan-genes on the two chromosomes of *Vibrionaceae* isolates was investigated. The results reveal the impact of the specific chromosomal location of a gene on its expression levels. Furthermore, the results of this study imply that interchromosomal translocations might be important for the evolution of *Vibrionaceae* species. Finally, a study presented in **Paper III** investigated adaptation strategies on gene sequence level by comparison of conserved membrane

proteins of *Vibrionaceae* isolates from three different temperature zones.
In summary, this study highlights the variety of different evolutionary processes that contribute to the adaptation and speciation of bacteria in general and *Vibrionaceae* in particular. Additionally, the results presented here can help in the development of a genome based concept of bacterial species.

LIST OF PAPERS

Paper I

Tim Kahlke, Alexander Goesmann, Erik Hjerde, Nils-Peder Willassen and Peik Haugen (2012), **Unique core genomes of the bacterial family *Vibrionaceae*: insights into niche adaptation and speciation**. *BMC Genomics*, 13:179

Paper II

Tim Kahlke, Alexander Goesmann and Peik Haugen, **The *Vibrionaceae* pan-genome hints at gene expression as the major driving force for unequal gene distributions on *Vibrionaceae* chromosomes**. Manuscript in preparation.

Paper III

Tim Kahlke and Steinar Thorvaldsen (2012), **Molecular characterization of cold adaptation of membrane proteins in the *Vibrionaceae* core-genome.** *PLoS One*, 7:e51761

ABBREVIATIONS

A	adenine
C	cytosine
CDS	coding DNA sequence
Chr I	Chromosome I
Chr II	Chromosome II
contig	short contiguous sequence
DDH	DNA-DNA hybridization
DNA	desoxyribolucleic acid
G	guanine
GC	guanine-cytosine
GO	Gene Ontology
HGT	horizontal gene transfer
IP	Internet Protocol
IT	Information Technology
Mbp	million base pairs
MLSA	multi locus sequence analysis
$oriC_{I}$	origin of replication of Chr I
$oriC_{II}$	origin of replication of Chr II
PTS	phosphotransferase system

Rfam	RNA Families Database
rRNA	ribosomal RNA
T	thymine
TCP	Transmission Control Protocol
terC	replication terminator
UniProt	Universal Protein Resource
WHO	World Health Organization
XML	Extensible Markup Language

Part I

INTRODUCTION

1

BACKGROUND

Since the early years of sedentism approximately 12.000 years ago selective breeding was used to enhance specific traits of plants and animals that would improve human life (Wieczorek and Wright, 2012). However, it was not until the mid-19th century that the underlying concepts of inheritance were studied scientifically. In 1859, the English naturalist Charles Darwin published his revolutionary work *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (Darwin, 1859). Based on observations from his expeditions around the world he introduced the idea that specific traits of animals, including humans, are changed over time due to natural selection. Darwin proposed that new species originate by elimination of the weak individuals in a population and the "survival of the fittest", i.e., the preservation of those characteristics that are best suited for the survival in specific environments. A short time after Darwin's publication the Augustinian friar Gregor Mendel discovered that certain traits in pea plants can be explained by the combination of traits from the parent generation (Mendel, 1866). He showed that at least some, if not all characteristics of an organisms are represented as distinct units which can be combined independently. This ushered in the era of genetics and, together with the theory of evolution, forms the basis for various scientific disciplines, such as population genetics, comparative genomics and phylogenetics. Since then, vast advances have been made in molecular genetics, including the discovery of DNA as the carrier of the genetic information (Avery et al., 1944) and the publication of the first complete genome sequence (Fleischmann et al., 1995). Today, the combination of computer science, Information Technology (IT) and biology opens ways for the analysis of genomic information in astonishing detail.

1.1 BACTERIA

Bacteria are single-celled microorganisms most of which are only a few micrometers (μm) in length. With an estimated number of $4-6 \times 10^{30}$ individuals (Whitman et al., 1998) bacteria are without question the predominant life form on earth. The biomass of all bacteria on our planet is estimated to be $350-550$ billion tons of carbon which is $60-100\%$ of the biomass of all plants. Bacteria are found in almost all habitats, in fresh water as well as marine and terrestrial ecosystems. They populate even the most hostile environments, such as the Antarctic ice (Price, 2000), hot springs (Yim et al., 2006) and the deep sea (Nogi et al., 1998).

In the general public bacteria are mostly associated with low hygiene or diseases. In fact, many of the most devastating diseases in the human history are caused by pathogenic bacteria, such as cholera and bubonic plaque. However, the majority of all bacteria is not only not pathogenic but important for humans and essential for life in general. In nature bacteria transform chemical elements, such as nitrogen and carbon, into molecular forms that are otherwise not usable for plants or animals (Gould, 1996; Canfield et al., 2010). Additionally, bacterial photosynthesis contributes significantly to the amount of oxygen in our atmosphere. In fact, even chloroplasts, the organelles that are responsible for photosynthesis in plants, evolved from symbiotic cyanobacteria (McFadden, 2001). In addition to the impact that bacteria have on the ecosystem of earth, they are also important for human health and our daily life. About 10^{14} $(100,000,000,000,000)$ bacteria are found on the skin, inside the guts and even in the blood of an average human (Berg, 1996). Although the interactions of many of these bacteria with the human body are yet poorly understood, studies show that they play an important role for our health (Grice et al., 2009; Qin et al., 2012).

In addition to the beneficial contributions of bacterial life on human health and the environment, bacteria are also of major economic importance due to their biocatalytic abilities. For example, lactic acid bacteria are widely used in the food industry for the fermentation of vegetables, meat or milk products (Asmahan, 2010). Furthermore, they are involved in the production of fine chemical such as alcohols, peptides or amino acids and are widely used in the chemical industry (Schimdt et al., 2001).

Altogether, the study of bacteria, their evolution and interaction with other bacteria, organisms or the environment is of high interest for many different scientific disciplines and branches of economy. Also, the fact that bacteria represent the most simple form of life makes them an excellent study object for any geneticist and molecular biologist.

1.1.1 *Bacterial genomes*

The genome of bacteria, just as the genome of any living organism, consists of desoxyribolucleic acid (DNA). DNA is a macromolecule build of sub-units, so called nucleotides, which carry one of four nucleobases: adenine (A), guanine (G), cytosine (C) and thymine (T). The genetic material of a cell is organised in one or more chromosomes which are composed of DNA molecules that form a double-helix structure (Figure 1) and associated proteins. The nucleobases in the two DNA strands are coupled

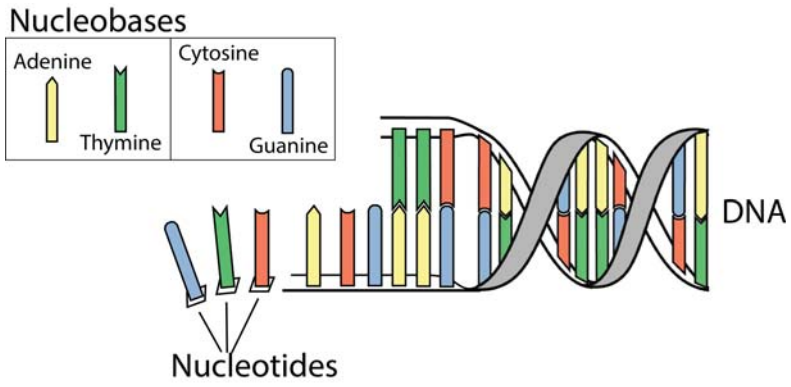


Figure 1: **Schematic view of the DNA double helix.** (Figure modified from http://www.nasa.gov)

in base-pairs: always one guanine and one cytosine or one adenine and one thymine are paired at a specific position in the DNA. Hence, the two nucleotide strands of a chromosome are *complementary*, i.e., the information on each strand is sufficient to replicate the second strand.

In the early years of bacterial genomics it was assumed that all bacteria possess one circular chromosome, i.e., the DNA forms one closed ring. This was considered a main

Table 1: Number, structure and size of bacterial chromosomes. Data taken from `http://www.ncbi.nlm.nih.gov`, January 2012

Bacterial species	# circular chromosomes	Size in Mbp	# linear chromosomes	Size in Mbp
Agrobacterium tumefaciens	1	12.8	1	2
Escherichia coli	1	5.2	-	-
Borrelia burgdorferi	-	-	1	0.9
Photobacterium profundum	2	4	-	-
		2.2		
Streptomyces griseus	-	-	1	8.5
Paracoccus denitrificans	2	2.8	-	-
		1.7		
Ureaplasma urealyticum	1	0.9	-	-
Vibrio splendidus	2	3.3	-	-
		1.7		

characteristic of bacterial genomes and used to distinguish them from other organisms. However, in 1989 Suwanto and Kaplan presented the genome sequence of the bacterium *Rhodobacter sphaeroides* which possesses multiple chromosomes (Suwanto and Kaplan, 1989). In the same year Saint-Girons and co-workers showed that the DNA in the chromosome of *Borrelia burgdorferi* is not circular but rather linear (Baril et al., 1989). Today many bacteria are known that possess multiple circular as well as linear chromosomes in their genome (Table 1).

Not only the number of chromosomes but also the size of bacterial chromosomes varies significantly. With 160,000 base-pairs the bacterium *Carsonella ruddii* possesses one of the smallest genomes known today (Nakabachi et al., 2006). In contrast, one of the largest known bacterial chromosomes, that of *Sorangium cellulosum*, is approximately 80 times larger and contains 13 million base pairs (Mbp) (Schneiker et al., 2007). In addition to chromosomes, many bacterial isolates also carry dynamic DNA molecules, so called *plasmids*. Plasmids are extra-chromosomal DNA molecules that share many characteristics with chromosomes. The main difference between chromosomes and plasmids is that plasmids are not essential for the survival of a particular bacterial species. Where chromosome loss inevitably leads to the death of a bacterial cell, the loss of a plasmid may or may not be disadvantageous for the host cell (Egan et al.,

2005). Although plasmids are commonly smaller in size than chromosomes, some of them, so called *megaplasmids*, have the size of regular bacterial chromosomes and can include more than 1 Mbp (Barnett et al., 2001). Together, chromosomes and plasmids form the genome of a bacterium.

1.1.2 *Genes and coding DNA sequences*

The genetic information of the DNA is stored in specific regions often referred to as *genes*. However, the meaning of the term *gene* is highly controversial and lacks a universal definition (Pearson, 2006). It originated in a pre-genomics era denoting a fundamental unit of heredity regardless of its physical representation on the DNA (Johannsen, 1905). One of the first molecular definitions of a gene was the *one-gene-one-enzyme* hypothesis (Beadle, 1941). It was based on the idea that cells can be seen as interconnected systems of chemical reactions. These reaction were proposed to be performed and regulated by specific biocatalytic proteins, i.e., enzymes. Thus, a gene was thought of as a region or feature on the DNA that encodes for a particular protein. Today, an ever-growing number of genes has been identified that either encode functional molecules other than proteins or that represent regions on the DNA that are involved in the regulation, inhibition or activation of gene expression. In fact, recent studies report that as little as 2% of the human genome encode proteins although approximately 80% of the DNA is functionally important (Shabalina et al., 2001; Sana et al., 2012; Dunham et al., 2012). However, in bacterial genomes >85% of the genetic material is composed of coding DNA sequences (CDSs), i.e., genetic regions which are further translated into proteins. Therefore, if not denoted otherwise, in bacterial genomics the terms gene and CDS are often used interchangeably. Also, the computer-guided prediction of CDSs in genome sequences is mostly referred to as *gene prediction*.

1.2 VIBRIONACEAE

Vibrionaceae denotes a family of curved rod-shaped gram-negative γ-Proteobacteria. Representatives of this bacterial family are motile due to a polar flagellum suited for motility in liquid medium (Atsumi et al., 1992) (Figure 2). Additionally, some species possess lateral flagella used for locomotion on viscous surfaces. The motility is a major morphological trait of *Vibrionaceae*, hence their name (l. *vibrare* - riste, vibrere: to vibrate).

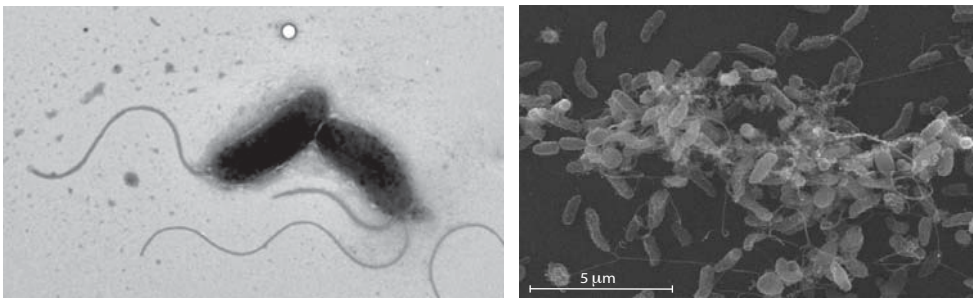


Figure 2: **Electron microscopy images of *Aliivibrio salmonicida*:** (A) single cells and (B) early stage of biofilm formation. Pictures courtesy of Hilde Hansen, University of Tromsø

Currently the *Vibrionaceae* family is divided into the genera *Aliivibrio*, *Catenococcus*, *Enterovibrio*, *Grimontia*, *Photobacterium*, *Salinivibrio* and *Vibrio*. Together they enclose 138 different species with new species being discovered frequently (`www.vibriobiology.net`). Vibrio species can be found in almost all aqueous environments and represent the majority of all culturable marine and estuarine bacteria (Okada et al., 2005). Members of this family populate marine environments as well as fresh or brackish waters and show an astonishing adaptiveness to different, often hazardous environments. For example, they are found in the Arctic ocean with temperatures close to the freezing point of water as well as in the deep sea with hydrostatic pressure many times greater than that of shallow waters. In general, *Vibrionaceae* bacteria are highly versatile. They exist as free swimming cells, posses the ability to form bacterial biofilms and are additionally associated with various plant and animal hosts, including corals, fish and even humans. Some Vibrio species are advantageous for

their hosts, e.g. the bioluminescent bacterium *Aliivibrio fischeri* which populates the light organ of the hawaiian squid *Euprymna scolopes* (Figure 3) where it produces its luminescence (Ruby and Lee, 1998).

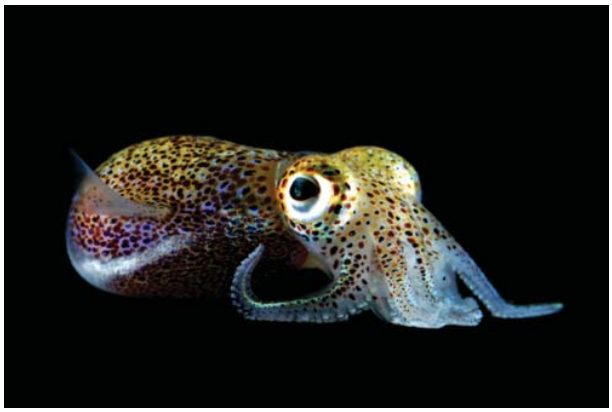


Figure 3: **The hawaiian bobtail squid (*Euprymna scolopes*)** is the host for the bioluminescent bacteria *A. fischeri*. Picture courtesy of Eric Roettinger (http://www.kahikaiimages.com/)

Other Vibrio species gained notoriety due to their pathogenicity, among which *Vibrio cholerae* is the best known. According to the World Health Organization (WHO) an estimate of 3-5 million humans are infected with *V. cholerae* every year through contaminated drinking water. The cholera disease causes more than 100.000 deaths every year due to severe diarrhea and vomiting of infected patients (WHO, 2012). Other *Vibrionaceae*, such as *Vibrio parahaemolyticus* and *Vibrio vulnificus*, are also severe human pathogens, although less devastating compared to *V. cholerae*. Given the effects of pathogenic vibrios it is not surprising that the majority of studies published today focuses on pathogenic *Vibrionaceae* species. However, pathogens represent only a small fraction of the complete diversity of this bacterial family. Even the majority of environmental isolates from species that have been reported pathogenic for humans do not carry pathogenicity genes (Yamaichi et al., 1999). In fact, many *Vibrionaceae* play an important role in the nutrient cycle of their habitat, such as *V. natrigens* which provides its environment with fixed nitrogen (Coyer et al., 1996).

1.2.1 *Genome structure of* Vibrionaceae *species*

All *Vibrionaceae* genomes sequenced today possess a bipartite genome, i.e., the genomic material is divided into two circular chromosomes and it is therefore assumed that it is a general genomic feature of this bacterial family. Interestingly, closely related bacterial families such as *Aeromonadaceae* and *Plesiomonaceae* exclusively contain single chromosomes (Okada et al., 2005) which suggests that the origin of the bipartite genome of *Vibrionaceae* may go back to the diversification of this bacterial family. Although the guanine-cytosine (GC) content of both chromosomes is roughly the same, they differ significantly in size. Additionally, the chromosomes show distinct patterns in gene conservation as well as distribution of functional genes. The larger of the two chromosomes, Chromosome I (Chr I), ranges from roughly 3 Mbp in *A. fischeri* to >4 Mbp in *Photobacterium profundum* whereas the smaller Chromosome II (Chr II) ranges from approximately 1 Mbp in certain *V. cholerae* strains to 2.2 Mbp in *P. profundum* and *V. vulnificus* isolates (Table 2). The majority of essential genes involved in replication and basic metabolic functions, as well as most genes conserved among all *Vibrionaceae* is located on Chr I (Heidelberg et al., 2000; Ruby et al., 2005; Thompson et al., 2009). Additionally, genes on Chr I in general tend to be expressed on a higher level than genes on Chr II. This can in parts be explained by the gene dosage effect, which reflects that genes on Chr I are found, in average, in higher copy numbers due to a delayed replication start of Chr II (Dryselius et al., 2008).

Another interesting feature of the bipartite genome of *Vibrionacae* species is the difference in gene sequence conservation between the two chromosomes. For example, genes conserved on Chr II of *V. cholerae* show a higher substitution rate and less codon usage bias in their amino acid sequence than those on Chr I. This indicates that genes on Chr II evolve faster in comparison to genes on Chr II (Cooper et al., 2010). Taking into account that Chr II in general carries only few conserved genes, this led to the hypothesis that Chr II might act as an "evolutionary test bed" for new genetic features that may play a role in the evolution of bacteria that possess multipartite genomes.

Table 2: Characteristics of Chr I and Chr II of different *Vibrionaceae* species. Data taken from `http://www.ncbi.nlm.nih.gov`, January 2012

Isolate	Mbp Chr I	GC content	No. of CDS	Mbp Chr II	GC content	No. of CDS
Aliivibro fischeri MJ11	2.91	38.9%	2,590	1.42	37.2%	1,254
Aliivibro salmonicida LFI1238	3.33	39.2%	2,820	1.21	38.2%	984
Photobacterium profundum SS9	4.09	42	3,416	2.24	41.2	2,006
Vibrio cholerae O1 N16961	2.96	47.7	2,741	1.07	46.9	1,093
Vibrio furnissii NCTC 11218	3.29	50.7	3,006	1.62	50.5	1,449
Vibrio harveyi ATCC BAA-1116	3.77	45.5	3,548	2.2	45.3	2,373
Vibrio parahaemolyticus RIMD 2210633	3.29	45.4	3,080	1.88	45.4	1,752
Vibrio splendidus LGP32	3.3	44	2,947	1.68	43.6	1,485
Vibrio vulnificus CMCP6	3.28	46.5	2,896	1.84	47.1	1,537

1.2.2 *Origin of a bipartite genome*

The first bacterium with a bipartite genome (*Rhodobacter sphaeroides*) was discovered during the late 1980's (Suwanto and Kaplan, 1989). Since then, the number of analyzed bacteria that posses multiple chromosomes has been growing. The origin of such a multipartite genome is however unclear and under debate. Despite the reported differences in gene distribution and size of the two *Vibrionaceae* chromosomes, the resemblance in GC content indicates a long evolutionary co-existence of the chromosomes (Dryselius et al., 2007). Theoretically, multipartite genomes in bacteria can originate in three different ways: (i) by duplication of a single chromosome, (ii) by split of a chromosome into two or more parts or (iii) by the acquisition of a plasmid that becomes persistent (Cooper et al., 2010). Heidelberg *et al.* proposed that Chr II was originally a megaplasmid that was acquired by an ancient ancestor of all *Vibrionaceae* species. This hypothesis is supported by the fact that the two chromosomes show a distinct replication machinery (Makino et al., 2003; Egan and Waldor, 2003). The origin of replication of Chr I ($oriC_I$) shows sequence similarity to the replication origin found in *Escherichia coli*. Additionally, the replication of Chr I is initiated by the ATPase DnaA, which is known to initiate the replication in *E. coli* and other bacteria (Fuller et al., 1984). On the other hand, the origin of replication of Chr II ($oriC_{II}$) includes repeat regions sim-

ilar to regions found in the replication origin of plasmids (Chattoraj, 2000). Also, the replication initiator of Chr II, RtcB, differs from DnaA (Egan and Waldor, 2003; Pal et al., 2005).

1.2.3 *Persistence of Chr II*

Regardless of its origination Chr II is an inherent part of the genome of all *Vibrionaceae* isolates sequenced today. It was hypothesized that certain Vibrio species may either loose Chr II or increase its copy numbers under specific environmental conditions but this remains to be shown. Instead, recent studies seem to refute this hypothesis. Specifically, Rasmussen *et. al* (2007) report a delayed initiation of the replication of Chr II compared to Chr I, and a linked replication termination of both chromosomes (Rasmussen et al., 2007) which assures equal chromosome copy numbers. Moreover, the fact that Chr II carries essential genes contradicts the possibility of the loss of this chromosome. Therefore, the loss of Chr II as well as the increase in copy numbers of either of the two chromosomes is disputable and would also violate the definition of a chromosome and imply a plasmid-like nature of Chr II (Egan et al., 2005).
Another question yet to be answered is the reason for the obligatory persistence of Chr II. Assuming that it was acquired as a megaplasmid it had to provide a biological advantage to its host in order to be retained. The most likely explanation is an inter-chromosomal rearrangement event that led to the translocation of essential genes from Chr I onto Chr II. Thus, the loss of Chr II would be lethal for descendants of this lineage. Another possible scenario is that the plasmid *ab initio* carried genes beneficial for the host genome, e.g. genes involved in host interaction or adaptation to certain environmental conditions. Interestingly, most bacteria with multipartite genomes interact in some way with hosts from other phyla (Egan et al., 2005). However, it is challenging to subsequently determine what led to the persistent incorporation of Chr II into the *Vibrionaceae* genome and even a combination of multiple scenarios is conceivable.

1.2.4 *Advantages of multiple chromosomes*

The fact that the two chromosomes are conserved as separate replicons in the genome of *Vibrionaceae* species raises the question which selective advantages this genome architecture offers to the host bacterium. One possible explanation is based on the benefits the replication of two relatively smaller chromosomes may offer compared to the replication of one large chromosome. *Vibrionaceae* representatives are among the fastest replicating bacteria known today with doubling times of less than half an hour reported for *V. parahaemolyticus* (12-14 min), *V. cholerae* (16-20 min) and *V. vulnificus* (18-22 min) (Dryselius et al., 2008). Therefore, it was proposed that the partition of the genome into multiple replicons may play a role for the fast replication as it enables simultaneous replication of genetic material and additionally reduces the number of overlapping replication cycles (Rasmussen et al., 2007). Also, the delayed replication start of Chr II may result in an energetically more efficient replication process in fast growing bacteria. In addition to the possible benefits for the replication process, it was proposed that the difference in distribution of gene functions between the two chromosomes provides an evolutionary advantage for *Vibrionaceae*. For example, when grown *in vitro* under aerobic conditions significantly less genes located on Chr II are expressed in *V. cholerae* in comparison to *in vivo* conditions of a rabbit's ileal loop (Xu et al., 2003). Therefore, Chr II may be important for the adaptation to changes of environmental conditions and might play a role in the adaptation to certain ecologic niches.

1.3 PSYCHROPHILIC BACTERIA

Bacteria populate almost all habitats on our planet which are, in fact, mostly cold environments with average temperatures <5 °C (Russel, 2009). This includes polar regions as well as mountainous areas and the deep sea which, by volume, represents 90% of all sea water on earth. Bacteria that maintain their metabolic functions and even proliferate in naturally cold environments are called *psychrophilic* bacteria or *psychrophiles*. Although the definition of psychrophiles is not always clear (Helmke and Weyland, 2004; D'Amico et al., 2006) one common definition is that psychrophilic bacteria are able to grow at temperatures of $\leqslant 4$ °C and show an optimal and maximum growth temperature of 15 °C and <30 °C, respectively (Morita, 1975; Moyer and Morita, 2007; Siddiqui and Cavicchioli, 2006). This distinguishes them from mesophilic bacteria with an optimal growth temperature >20 °C and thermophiles, which can proliferate at temperatures as high as 120 °C (Takai et al., 2008).

Although microorganisms are not the only life forms that populate environments with extremely low temperatures, they are least protected to the cold. Higher organisms that live in polar regions, high altitudes or the deep sea, e.g. sea mammals or birds, are commonly insulated with fur, skin and fat tissue. Microorganisms, on the other hand, lack these layers of protection and thus their internal temperature is almost identical with that of the surrounding medium. Therefore, psychrophiles need various strategies to adapt to the the low temperature in order to overcome the deleterious effects that stresses of the cold have on their metabolism and cellular machinery.

1.3.1 *Cold adapted enzymes*

One crucial factor in the adaptation to low temperature is to sustain the biocatalytic properties of enzymatic reactions. In general, enzymes catalyse chemical reactions by

forming a complex with a particular substrate S and convert it into one or several products P by

$$E + S \longleftrightarrow ES \longleftrightarrow E + P \tag{1}$$

where E is the enzyme and ES represents the enzyme-substrate complex. In this process the turnover rate k_{cat} of the enzyme-substrate complex is denoted by

$$k_{cat} = \frac{k_B T}{h} k \exp(-\Delta G^{\#}/RT) \tag{2}$$

where k is the transmission coefficient, k_B is the Boltzmann constant, h is the Planck constant, R is the universal gas constant, $\Delta G^{\#}$ is the activation energy and T is the absolute temperature in kelvin (Siddiqui et al., 2004). The transmission coefficient k in (2) is dependent on the viscosity of the medium, i.e., water, which increases significantly with decreasing temperature (Kestin et al., 1978). Hence, the reaction rate of enzymes is decreasing in cold environments due to the low temperature T and also decreasing k. However, psychophilic enzymes show high reaction rates specifically at low temperature in comparison to their mesophilic counterparts (Feller et al., 1996). On the other hand, it has been shown that psychrophilic enzymes are more heat-labile, i.e., they denaturate at significantly lower temperatures. This led to the hypothesis of the activity-flexibility-stability relationship: by increasing the flexibility of their protein sequence cold-adapted enzymes increase their catalytic activity at low temperatures but at the expense of structural stability (Alvarez et al., 1998; Feller et al., 1999; Russel, 2000). For example, comparison of psychrophilic, mesophilic and thermophilic homologs of the enzyme xylanase revealed activation optima of 35 °C, 62 °C and 80 °C, respectively (Collins et al., 2003). At the same time, the melting point of the psychrophilic xylanase (52.6 °C) is significantly lower than its mesophilic (63.1 °C) and thermophilic homologs (80.7 °C).

To adapt to low temperature the protein sequences of psychrophilic enzymes show differences in their amino acid composition in comparison to mesophilic homologs. However, identification of general adaptation strategies in terms of amino-acid substitutions remains challenging because the effects of certain amino acid substitutions vary with the position in the protein as well as with its function. For example, whereas

psychrophilic enzymes show a decrease of hydrophobic residues in their hydrophobic core (Russel, 2000; D'Auria et al., 2009) an increase of hydrophobicity is reported for loop regions of psychrophilic proteins (Metpally and Reddy, 2009). In general the study of cold adapted enzymes reveals that more than one strategy of psychrophilic bacteria exists to increase flexibility and thereby activity of their enzymes in order to countervail the deleterious effects of the cold.

1.3.2 *Membranes of psychrophilic bacteria*

The plasma membrane of bacteria represents a biological barrier that separates the inside of a cell, its cytoplasma, from the surrounding medium, the extracellular fluid. The fundamental structural units of a cell membrane are lipid molecules, mostly phospholipids, which consist of a polar head group and a hydrophobic tail of fatty acyl chains. These lipid molecules form a bilayer with an hydrophobic interior which serves as a matrix for a variety of membrane proteins (Figure 4).

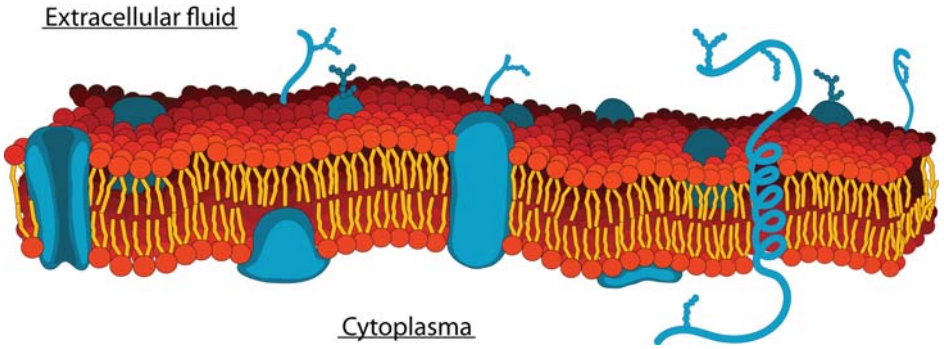


Figure 4: **Schematic view of a cell membrane.** Lipids shown in orange/yellow with round polar head groups (orange) and hydrophobic acyl chains (yellow). Membrane proteins are shown in blue. (Figure modified from http://www.wikipedia.org)

The membrane is impermeable for most solvents and molecules and thereby plays a crucial role in controlling the internal concentrations of molecules and optimal conditions for the metabolism (Konings et al., 2002). The exchange of ions and molecules is provided by specific membrane proteins, e.g. ABC transporters and phosphotransferase

systems (PTSs) (Kotrba et al., 2001; Speelmans et al., 1993). In addition to the transport systems, membrane proteins are also involved in the reception of environmental conditions, signal transduction and energy metabolism, which makes them essential for the vital functions of a cell (Speelmans et al., 1993; Goudreau and Stock, 1998; Konings et al., 2002).

Plasma membranes are no static structures. In fact, they show the behavior of liquid cristals: on one hand they represent a physical barrier that is also involved in shaping the cell but at the same time show properties of a liquid, enabling the lateral movement of the embedded proteins and lipids. This *liquid mosaic model* (Singer and Garth, 1972) is widely accepted as the basic structure of all cell membranes. The fluidity of the lipid bilayer is of crucial importance for the membrane proteins to optimally perform their functions (Lenaz, 1987; Andersen and Koeppe, 2007). However, low temperature severely affects the viscosity and thereby the fluidity of the cell membrane. With decreasing temperature the fluidity of the membrane decreases, until the membrane transitions into a gel-phase which eventually leads to the loss of all functions of the membrane.

The main strategy of psychrophilic bacteria to maintain membrane fluidity at low temperatures is the alteration of the lipid composition of the membrane. By synthesizing lipids with a lower melting temperature the gel-liquid transition point of the membrane decreases. This *homeoviscous adaptation* (Sinensky, 1974) can be achieved in various ways, e.g., by increasing the amount of shorter, branched or unsaturated acyl chains of the membrane lipids (Weber et al., 2001; Russel, 1997, 1984). For example, the average acyl chain length of the phospholipids in the membrane of *Micrococcus cryophilus* decreases with decreasing temperature and vice versa (McGibbon and Russel, 1983) which also alters the gel-liquid transition point accordingly. Additionally, *M. cryophilus* shows an isomeric preference at low temperature for the lipid isomer with the lower melting point.

1.4 THE PAN-GENOME CONCEPT

A crucial difference in the life cycle of prokaryotes and most multicellular eukaryotes is the way in which they reproduce. Sexual reproduction of eukaryotes requires the development of haploid gametes or spores through meiosis, which will carry the genetic material of each of the parent individuals. Prokaryotes, on the other hand, reproduce asexually by simple duplication of the genetic material of the parent cell and subsequent fission into two individual cells. These differences in reproduction are of major importance for the genetic variability of prokaryotes in comparison to eukaryotic genomes.

In eukaryotes new genetic traits have to be acquired in the germline of eukaryotes to be persistent and passed on to individuals of the next generation. Acquisition of new genetic material in somatic cells will not be inherited by the offsprings and therefore will not contribute to the evolution of a eukaryotic species or taxon. Additionally the pairing of homologous chromosomes in the prophase I of the meiosis of most eukaryotic taxa is based on DNA homology (Bozza and Pawlowski, 2008). This process prohibits the insertion or deletion of genetic material in only one of the homologous chromosomes (Mira et al., 2010). Therefore, sexual reproduction is a major cause for the low diversity in the gene repertoire of closely related eukaryotes. In fact, even the chromosomal location of homologous genes is conserved among members of the same eukaryotic species which allows the creation of chromosomal maps showing the exact location of particular genes (Figure 5). The high level of conservation in structure and numbers of chromosomes of related eukaryotes enables the demarcation of eukaryotic taxa into groups with distinct geno- and phenotypes. Furthermore, conclusions can be drawn regarding the gene set of a species or genus only by knowing the gene sets of few individuals.

Prokaryotes on the other hand show a high degree of genetic variability, even on a species level. In the absence of the preserving molecular mechanisms of sexual reproduction bacterial genomes are prone to constant genomic rearrangements, such as lineage-specific gene loss (Ehrlich et al., 2008; Georgiades and Raoult, 2010), the duplication of genes (Gevers et al., 2004; Jordan et al., 2001) and the horizontal acquisition

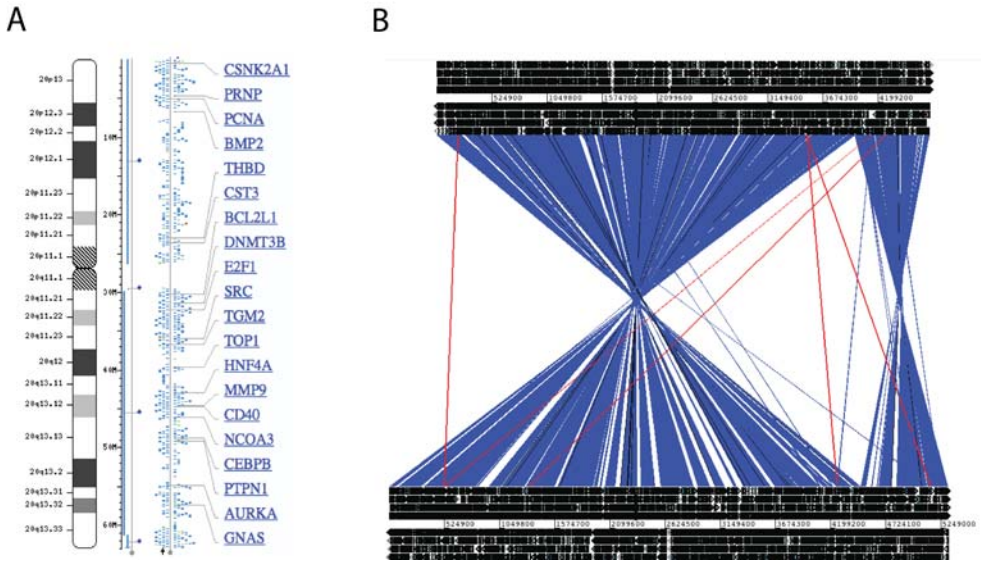


Figure 5: **(A)** Chromosomal map of the human chromosome 20 (Modified from http://www.ncbi.nlm.nih.gov/mapview/)
(B) ACT comparison of ChrII of *E. coli* K12 substr. H10B and *E. coli* O104-h4. Diagonal blue lines show inversions.

of foreign DNA (Doolittle, 1999; Koonin et al., 2001; Ragan, 2001). Additionally, inversions and translocation events alter the organization of the existing genetic material of a bacterial cell (Figure 5B). Therefore, even representatives of the same bacterial species can vary significantly in their geno-, sero- and phenotype (Jordan et al., 2001; Lerat et al., 2005; Lefébure and Stanhope, 2007; Laing et al., 2011) due to the fact that all non-fatal genome re-arrangements are passed on to the next generation.

1.4.1 *Pan-genomes and the distributed genome hypothesis*

The fact that the genomes of representatives of the same species can vary significantly in size was already discovered by pulse-field gel electrophoresis experiments in the 90ies of the last century (Bergthorsson and Ochman, 1995; Thong et al., 1997). However, whole genome sequence comparison of isolates of the same species revealed a much higher degree of intra-species variability than expected (Mira et al., 2010; Laing et al., 2011). This led to the development of the *distributed genome hypothesis* (Ehrlich

et al., 2005; Baumdicker et al., 2012). It states that bacterial taxa, e.g. species or genera, possess a *distributed genome* in which the entirety of variable or unique genes of a taxon, its so called *pan-genome*, will exceed the gene set of any of its representatives by several magnitudes (Tettelin et al., 2008; Hogg et al., 2007). Thus, no single bacterial isolate contains the complete genetic repertoire of its phylogenetic lineage but a subset of *pan-genes* that is unique to this isolate.

In general the pan-genome of a group of bacteria is defined as the union of three distinct sets of genes: *core* genes, *accessory* genes and *unique* genes. Each of the three gene sets shows certain characteristics, e.g., the number of genes that are included in it and the distribution of functional classes (Tettelin et al., 2005; Hiller et al., 2007; Huynen et al., 1998; Lapierre and Gogarten, 2009; Callister et al., 2008). Therefore, core, accessory and unique genes can be used for different applications and provide different information about the group of investigated genomes.

1.4.2 *Core genes*

Core genes denote genes that are present in all genomes of an investigated group of bacteria. The entirety of all core genes form the *core genome*, which builds the genetic backbone of the bacterial group of interest, e.g., a bacterial species. The core genome mainly consists of housekeeping genes, i.e., genes involved in maintaining basic metabolic functions, replication of DNA, the constitution of the cell envelope or binding proteins (Charlebois and Doolittle, 2004; Tettelin et al., 2005; Hiller et al., 2007). Furthermore, genes that encode extrachromosomal functions or are horizontally acquired are commonly underrepresented in the core genome.

Although the core genome is conserved for a group of bacteria it is not invariable: over time it will be shaped by horizontal gene transfer (HGT) and natural selection (Glasner et al., 2008; Lefébure and Stanhope, 2007). Thus, bacterial taxa show significant differences in the number and functional distribution of their pan-genes. For example, an investigation of 17 *E. coli* representatives determined a core genome that includes ~47% of the gene set of each of the included isolates (Rasko et al., 2008). In contrast, a similar study carried out with 17 *Streptococcus pneumoniae* genomes revealed a much

higher level of conservation, including ~73% of the genes in each strain (Hiller et al., 2007). Hence, the size of the core genome, i.e., the number of genes per genome that are shared among all investigated isolates, provides a measure of conservation and, at the same time, diversity of the investigated genomes.

Due to the fact that core genes are, per definition, present in all genomes of an investigated group, they are important for inferring phylogenetic relationships, e.g. through multi locus sequence analysis (MLSA) (Daubin et al., 2003; Zeigler, 2003; Thompson et al., 2009). Furthermore, the size of the core genome itself can be used for phylogenetic inference (Snel et al., 1999; Wolf et al., 2002).

Another potential application for core genes is the determination of a minimal gene set that is needed to maintain bacterial life (Koonin, 2003). Especially core genes of bacteria which naturally enclose a small genome such as *Mycoplasma genitalium* may be suitable for the determination of a minimal gene set. In addition, in case a minimal genome can be determined, it may also shed light on the metabolic machinery of a universal common ancestor of all bacteria and thus will provide an insight into the beginning of life itself.

1.4.3 *Unique genes*

The set of unique genes of a pan-genome is defined as those genes present in only one isolate of a group of bacteria. Thus, these genes show no or only weak homology to genes of any other investigated isolate, i.e., they are specific to one genome. Where the functional annotation of unique genes is possible it reveals a high percentage of genes related to phage genes, HGT and mobile genetic elements (Hiller et al., 2007; Rasko et al., 2008). A pan-genome analysis of 13 *Haemophilus influenzae* strains showed that ~25% of the determined unique genes are homologous to phage associated genes (Hogg et al., 2007). Additionally, unique genes tend to show an unusual codon usage (Hogg et al., 2007) and therefore it became widely accepted that a significant fraction of unique genes originate from HGT. Given the importance of HGT for bacterial evolution (Ochman et al., 2000; Koonin et al., 1996, 2001) it seems legitimate to assume that unique genes contribute significantly to the evolution and speciation of bacterial

taxa. However, the main characteristic of unique genes, their uniqueness, makes their functional classification challenging. The determination of the function of newly discovered genes is based on the comparison to already known genes. Due to the fact that unique genes are per definition not conserved among related bacteria, a high percentage of unique genes is often annotated as *protein of unknown function*. For example, the function of >50% of the unique genes of the pan-genome of *E.coli* and *S. pneumoniae* pan-genomes is unknown and they are annotated as *hypothetical genes* (Rasko et al., 2008; Hiller et al., 2007).

1.4.4 *Accessory genes*

The third group of genes a pan-genome is composed of is the set of accessory genes, also called distributed (Hiller et al., 2007) or dispensable (Tettelin et al., 2005) genes. The accessory genome encloses all genes that are neither core nor unique genes, i.e., genes found in at least two but not all investigated genomes. They are presumably not involved in essential metabolic functions but provide an important pool for genetic variability. Accessory genes of a species' pan-genome are often involved in adaptation to a specific niche (Legault et al., 2006; Laing et al., 2011; Sim et al., 2008) or manifestation of a specific phenotype, such as host adaptation or pathogenicity. For example, pathogenicity islands that discriminate pathogenic strains from their environmental counterparts, are part of their species' accessory genome (Schmidt and Hensel, 2004). Thus, accessory genes are used to identify pathogenic strains of certain bacterial species. Furthermore, in studies that include higher phyla, such as genera or families, the accessory genome also includes genes that are specific to any sub-taxon, e.g., species specific genes. Therefore, analysis of the accessory genome of higher phylogenetic lineages is important for the identification of specific biological markers and provides insights into bacterial evolution and separation of bacterial species.
Regarding functional annotation, accessory genes in general show a lower fraction of genes of unknown function than unique genes but higher than core genes. Furthermore, the higher the number of genomes that share a particular homolog, the lower is

the probability of it being horizontally transferred, and vice versa (Hogg et al., 2007). Thus, accessory genes are more prone to be horizontally transferred than core genes but less than unique genes. Likewise is the diversity of functional classes encoded by accessory genes higher than those of core genes but lower than determined for unique genes (Lapierre and Gogarten, 2009).

1.4.5 *Determination of the pan-genome*

Regardless of the goal of a pan-genome study, whether it is the investigation of a specific molecular function or the genetic diversity of a group of bacteria, the determination of a pan-genome is always based on clusters of homologous sequences. Therefore, the clustering process is of major importance and has significant effects on the outcome of any pan-genome analysis (Bentley, 2009). The first step of most clustering algorithms is the determination of pairwise sequence alignments of either the protein or nucleotide sequences of the genes in a dataset. This is commonly done by an initial *all-versus-all* comparison with bioinformatics tools, such as *blast* (Altschul et al., 1999), *fasta* (Lipman and Pearson, 1985) or *ssearch* (Pearson and Lipman, 1988). For the subsequent clustering of homologous sequences a variety of different algorithms and programs is available, e.g., *orthoMCL* (Li et al., 2003) or *Inparanoid* (Remm et al., 2001). However, no gold-standard for homology clustering exists and the results of any sequence alignment and clustering process are highly dependent on the algorithm and parameters chosen (Chen et al., 2007). For example, a too stringent sequence similarity cut-off will result in a decreased number of core genes and an increased number of accessory or unique genes of a dataset simply because homologous sequences are separated into different homology clusters. On the other hand, a similarity cut-off that is set too low will lead to an increase in core genes and a decrease of unique and accessory genes. This might falsely be interpreted as a higher level of conservation and lower genetic diversity among the investigated genomes.

1.4.6 *The pan-genome size: open or closed?*

The size of a pan-genome of a bacterial lineage is, theoretically, mostly dependent on two sets of genes: (i) the number of genes found in all members of a phylogenetic lineage, i.e., core genes, and (ii) the number of unique genes that will be added to the pan-genome with each newly sequenced isolate. Regarding bacterial systematics, this raises two questions: do all members of a bacterial taxon share a set of common genes even if an increasing number of isolates is added to the investigated group of genomes? And, additionally, does each newly sequenced strain add new, yet undiscovered genes to the pan-genome? To approach this question Fraser and co-workers used the gene sets of eight completely sequenced *Streptococcus agalactiae* genomes to extrapolate the average number of core and unique genes of all strains of this bacterial species (Tettelin et al., 2005). By calculating the number of core genes for all possible permutations of genome combinations in their dataset they estimated that all *S. agalactiae* isolates share in average roughly 80% (~1800) of their gene sets (Figure 6). The number of ~30 unique genes per strain was determined similarly (Figure 7). These results indicate that, despite their genetic variability, isolates of the same bacterial species share a set of core genes. This was also confirmed for additional bacterial species and even higher taxa, such as genera and families (Hogg et al., 2007; Tettelin et al., 2008; Hiller et al., 2007). In fact, a recent study even estimated the amount of core genes shared by all bacteria to be ~250 (Lapierre and Gogarten, 2009). Another implication of the results presented by Tettelin *et al.* (2005) is that the pan-genome of a bacterial species is *open*, i.e., its size is infinite due to the fact that each newly sequenced isolate will add new genes to the species' pan-genome. Interestingly, the same study proposed that the pan-genome of *Bacillus anthracis* is closed, i.e., the complete genetic repertoire of this species can be described by the gene set of as few as 4 isolates. This was explained by the fact that the genomes of *B. anthracis* and *Bacillus cereus* isolates differ only in the acquisition of a plasmid that carries the anthrax toxin. It was therefore proposed that *B. anthracis* is not a true bacterial species and that all true species and higher taxa possess an open pan-genome.

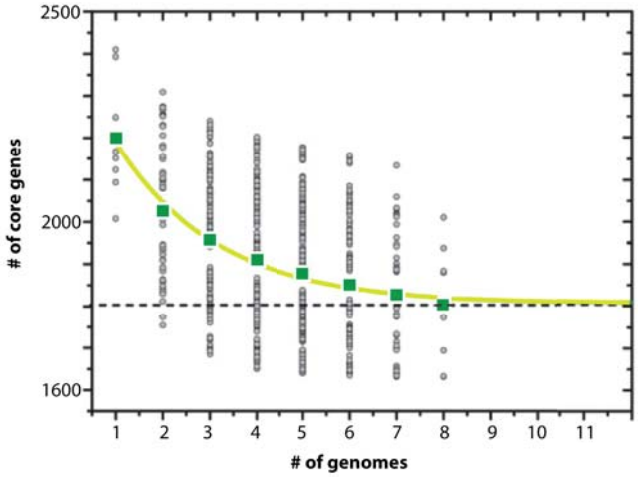


Figure 6: **Number of core genes of eight *S. agalactiae* strains**. Gray circles represent all possible permutations at each x-value. The curve (light green) represents the curve progression estimated for increasing number of genomes. Green squares show the mean value of the core genome size for x. (Figure modified from Tettelin *et al.*, 2005)

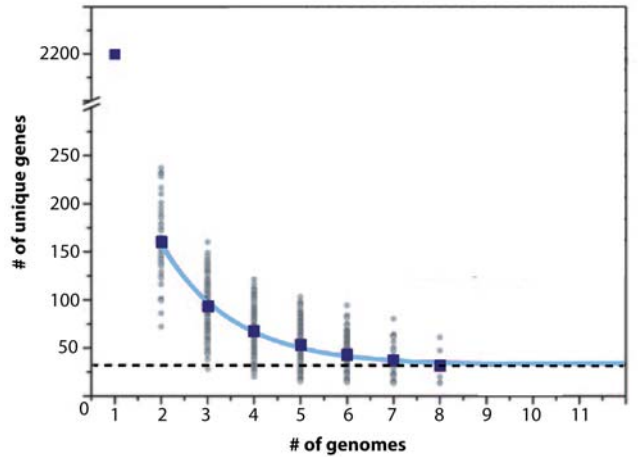


Figure 7: **Number of unique genes of eight *S. agalactiae* strains**. Gray circles represent all possible permutations at each x-value. The curve(light blue) represents the curve progression estimated for increasing number of genomes. Blue squares show the mean value of unique genes for x. (Figure modified from Tettelin *et al.*, 2005)

1.5 PHYLOGENETICS AND THE DEMARCATION OF BACTERIAL TAXA

> Species are groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups.
> *- E. Mayr, 1942 -*

In biological systematics a theory-based concept of the taxon *species* is of particular importance as it defines the basis for the demarcation of organisms into distinct groups. In his book *Systematics and the Origin of Species, from the Viewpoint of a Zoologist* (1942), Ernst Mayr introduced a species concept that combined taxonomical classification of organisms with evolutionary, i.e., phylogenetic relationships (Mayr, 1942). This concept still forms the basis for today's classification of higher organisms, such as animals and plants, into related taxa. Unfortunately, this concept can not be applied to bacteria due to their asexual reproduction. In fact, the question whether or not a theory-based concept for the classification of bacteria exists is and has always been controversially discussed in the scientific community. Until the second half of the last century most scientists denied even the possibility of classifying bacteria in terms of evolutionary relationships (McInerney et al., 2008). Bacteria were instead mainly separated on the basis of morphological or general phenotypical features. Often the ability of causing certain diseases defined representatives of bacterial species, e.g., *V. cholerae* or *Neisseria meningitides* (Gevers et al., 2005). Additionally, physio-chemical properties, such as fatty acid composition and GC content were used to separate bacteria into distinct clusters (Cohan, 2002; Staley, 2006). Nevertheless, this operational-based concept did not reflect the phylogenetic relationships in the bacterial kingdom.

1.5.1 *DNA-DNA hybridization*

With the discovery of DNA molecules as phylogenetic markers the view on bacterial systematics changed (Zuckerkandl and Pauling, 1965). In 1973, J. L. Johnson investigated the DNA sequence similarity of bacteria that were grouped into taxa according to their phenotypical traits (Johnson, 1973). By hybridizing bacterial DNA with DNA

of a given species' type strain he could show that representatives of most bacterial species show intra-species DNA similarity of >70%. His observations introduced phylogenetics into bacterial systematics and made DNA-DNA hybridization (DDH) its gold-standard. Despite its reliability DDH is not free of criticism: mainly the selection of a species' type strain as well as the determination of the threshold of 70% sequence identity are not theory based but rather artificial (Cohan, 2002). However, despite all criticism, DDH is still considered the gold standard in bacterial systematics (Konstantinidis et al., 2006; Auch et al., 2010; Chan et al., 2012).

1.5.2 *16S ribosomal RNA*

Another milestone in bacterial systematics was the discovery of small sub-unit or 16S ribosomal RNA (rRNA). For almost one decade they were proposed to be the "ultimate molecular chronometer" (Fox et al., 1977; Woese and Fox, 1977; Woese, 1987). In contrast to DDH the analysis of 16S rRNA is relatively easy and enabled scientists to accurately classify bacteria into families and genera. However, the major advantage of 16S rRNA, the presence and conservation among all known organisms, is also their biggest drawback. Few mutation sites and low mutation rates in general limit the resolution of phylogenetic classification of closely related bacteria based on 16s rRNA (Fox et al., 1992; Henz et al., 2004). For example, isolates that share <97% sequence similarity in their 16S rRNA almost always show <70% similarity in DDH experiments. This indicates a 16S rRNA similarity cut-off of <97% to distinguish bacteria that belong to different bacterial species. However, when the similarity exceeds the cut-off of 97% the phylogenetic signal of 16S rRNA molecules does not provide sufficient information to separate isolates on a species level: here the determined DDH similarity values may or may not be >70% sequence similarity (Stackebrandt and Goebel, 1994). Additionally, mutation rates of single genes can differ between bacterial species and even between genes of the same genome. Thus, the history of a single gene might not reflect the evolutionary history of the complete organism (Henz et al., 2004). In fact, as more and more genomes became available increasing numbers of inconsistencies of phylogenies of other marker genes, such as ATPase, and 16S rRNA were reported (Hilario and

Gogarten, 1993; Phillipe and Forterre, 1999). This revealed another problem regarding single gene phylogenies: the relevance of HGT on the bacterial evolution and its impact on phylogenetic analysis.

1.5.3 *Multi-Locus Sequence Analysis*

One approach to overcome the limitations of single gene phylogenies is the concatenation of multiple gene sequences. MLSA has been widely adapted in the separation of closely related bacteria (Daubin et al., 2003; Thompson et al., 2005, 2009) and has proven its reliability even for species where biochemical properties and 16s rRNA analysis alone did not result in congruent phylogenies (McTaggart et al., 2010). Genes used for MLSA have to be conserved among all investigated genomes, i.e., they are core genes, which reduces the amount of HGT (Lerat et al., 2003) and increases the consistency of such phylogenies. Furthermore, to avoid conflicts based on variable evolution of paralogs, only single copy genes should be chosen for phylogenetic inference (Zeigler, 2003; Lerat et al., 2003; Thompson et al., 2009). Additionally, to ensure that all genes included in the analysis carry enough variable sites, a minimal gene length of 900 nucleotides was proposed (Zeigler, 2003).
However, even with its advantages over other approaches MLSA, just as 16s rRNA analysis, disregards certain important aspects of bacterial evolution, such as HGT, lineage-specific gene expansion or lineage-specific gene loss. Therefore, even if conserved genes reflect the phylogeny of the *genetic backbone* of a species it does not reflect its complete evolutionary history.

1.5.4 *Phylogenies based on gene content*

The amount of complete bacterial genomes that are available to scientists nowadays opens additional ways for classifying bacteria into distinct taxa: comparison of the complete gene content of bacteria. Given that closely related bacteria share a large portion of their gene sets, and furthermore, that the amount of genes shared decreases

with increasing evolutionary distance, the comparison of the gene sets of bacterial isolates may be used for phylogenetic analysis. This presence/absence model of homologous genes is comparable to the presence or absence of certain morphological features (Fitz-Gibbons and House, 1999). Phylogenies based on gene content have been shown to resemble those of other phylogenetic methods on the kingdom and domain level as well as on the species level (Snel et al., 1999; Fitz-Gibbons and House, 1999; Wolf et al., 2001; Huson, 2004). However, phylogenies based on gene content show weaknesses in the demarcation of bacterial families, genera or species' that experienced major gene loss (Wolf et al., 2002). Additionally, without a universal definition on how to weight HGT, lineage-specific gene expansion or gene loss, the model applied for the calculation of the phylogeny significantly affects the result of such an analysis (Wolf et al., 2002).

Another major weakness of this approach lies within its basic assumption: to provide a phylogenetic signal based on gene content all members of a bacterial taxon must share genes not found in closely related species even if more bacterial sequences become available. But in case of a continuous genetic spectrum this phylogenetic signal will vanish with the availability of new bacterial genome sequences (Gevers et al., 2005).

1.6 BIOINFORMATICS

With the introduction of whole genome sequencing techniques in the last decade of the past century the amount of data produced in life sciences increased dramatically. For example, in 1990 the human genome was initiated to accomplish the sequencing of the first complete human genome. Until 2001, the human genome project had produced roughly 5.5 billion base pairs of sequence data, including short sequence parts as well as complete chromosomes, although only one third of the final sequence being completed (Chial, 2008). In book form this would result in a book of more than 1.8 million pages (supposing a common word-document with an average 3000 letters per page). Furthermore, new advances in biotechnology and the development of high-throughput sequencing methods additionally increased the speed with which new

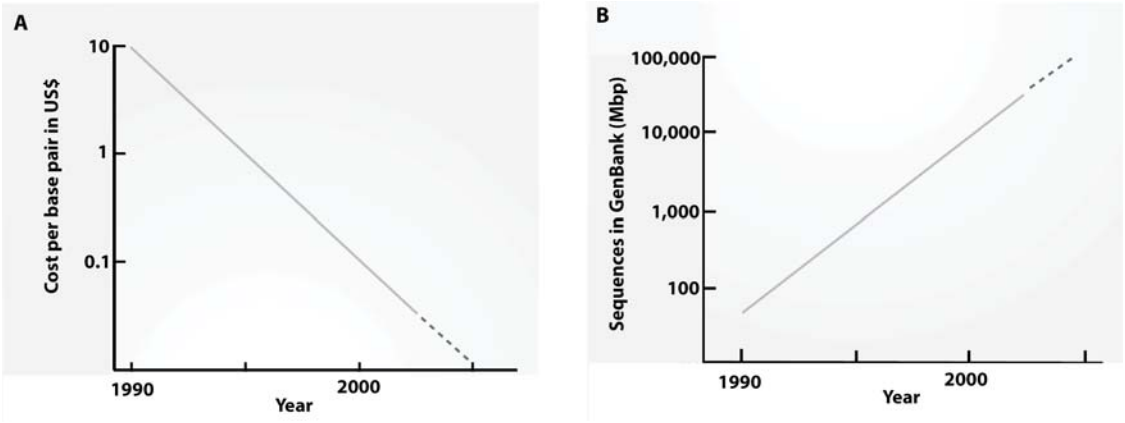


Figure 8: **Development of genome sequencing**. (A) Development of sequencing costs per base pair 1999 – 2005. (B) Increase in DNA sequences in GenBank 1999 – 2005 (Figure modified from Collins *et al.*, 2003)

sequences were published (Figure 8). This created the need for new approaches in storage, handling and analysis of biological data. Also, new methods for the integration of large dataset, such as whole genome sequences, protein interaction networks or microarray data had to be developed.

Approximately at the same time as whole genome sequencing became available for biologist, new advances were made in IT that would enhance biological science of coming years. With the introduction of the Transmission Control Protocol (TCP) and the Internet Protocol (IP) in 1982 the fundamentals were created for one of the most influential developments of the past two decades: the internet (Ruthfield, 1995; Baxevanis and Ouellette, 2001). Together with the installation of high-speed data links and WIFI connections, even the largest biological data sets are now available online and data can be shared among scientists all around the world basically without time loss. Additionally, prices for computer hardware, such as processors and hard drives, decreased significantly and the efficiency and performance roughly doubled each second year, as already proposed by Moore's Law (Moore, 1965, 1975) (Figure 9). This made IT an inconceivable component of life sciences as we know it today and resulted in a new and fast growing discipline: *Bioinformatics*.

The term *bioinformatic* is used for a variety of different fields which have in common that they address biological problems using computational approaches. This includes

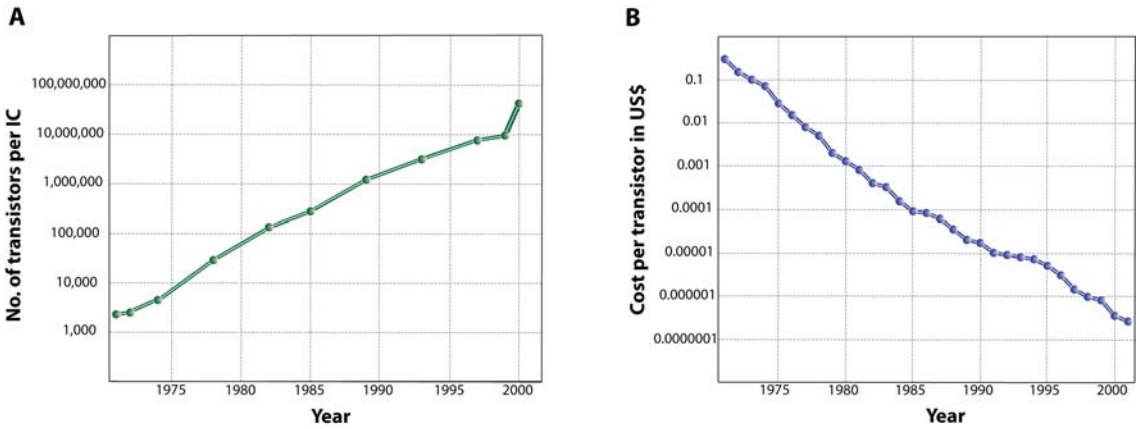


Figure 9: **History of transistor density and prices**. **A**: # of transistors per integrated circuit produced by Intel Corporation. Data taken from `http://download.intel.com/museum/Moores_Law/Printed_Materials/`. **A** Development of prices per single transistor. Data taken from `http://www.singularity.com/charts/page59.html`.
Y-axis' shown in logarithmic scale.

fields such as phylogenetics, protein modelling and the modelling and simulation of biological systems. Additionally, bioinformatics plays a key role in genomics and is involved in all steps of a sequencing projects, from the assembly of reads and short contiguous sequences (contigs) to genome annotation and the subsequent comparison of newly predicted features to those of already annotated genomes. Without the use of computers and computational methods none of these tasks would be conceivable.

1.6.1 *Genome annotation*

The annotation of a genome denotes the process of adding information to the nucleotide sequence of an organism's DNA. Without annotation the DNA is nothing else for a scientist than a very long sequence of the four nucleotides A, C, G and T. It is comparable to a book written in a foreign language of which we only know the alphabet but neither the syntax nor the semantics of the language are known. To understand the information encoded in the book text one first has to identify the words and sentences, then determine the grammar of the language and finally assign meaning to the sentences. Similarly, to understand the information encoded in the DNA

scientists first have to identify genetic features, such as protein coding sequences or genes that encode for functional RNAs, as well as binding sites and other sequence features. This *structural annotation* is mostly based on the identification of motifs or patterns that are characteristic for individual features. Subsequent to the structural annotation of a sequence, a *functional annotation* is performed to try to assign a function to each feature. Functional annotations are mostly based on the comparison of new features to databases of those with already known function, such as the Universal Protein Resource (UniProt) database for proteins (The UniProt Consortium, 2012) or the RNA Families Database (Rfam) (Gardner et al., 2011).

It goes without saying that genome annotation is a crucial step in each sequencing project. However, the annotation of a genomes is always only a snapshot of current knowledge. With each new sequence or sequence motif submitted to a public databases, with each newly identified function or novel algorithm the previous annotation becomes incomplete and eventually outdated (Salzberg, 2007; van den Berg et al., 2010). Also, genome sequences that were initially published as draft genomes, i.e., nucleotide sequences truncated into contigs, may be completed subsequent publication and features that were not included in the draft genome will obviously not be included in the annotation of the complete genome. Furthermore, the quality of genome annotation, in general, varies significantly, depending on whether the sequence was exclusively automatically annotated or manually curated by a skilled scientist. Manually curated annotations generally show a higher level of accuracy as automatic annotation pipelines often are based on simplified rules to determine the accuracy of an annotated CDS. Therefore, re-annotation of genomes becomes more and more important for any genomics project that includes sequences from various sources and dates.

1.6.2 *Genome annotation systems*

To facilitate genome annotation a wide range of software solutions and annotation systems are available, which differ greatly, e.g. in the number of incorporated tools, platform dependency or available features. Because genome annotation often requires vast computational resources, online services such as the Rast-Server (Aziz et al., 2008)

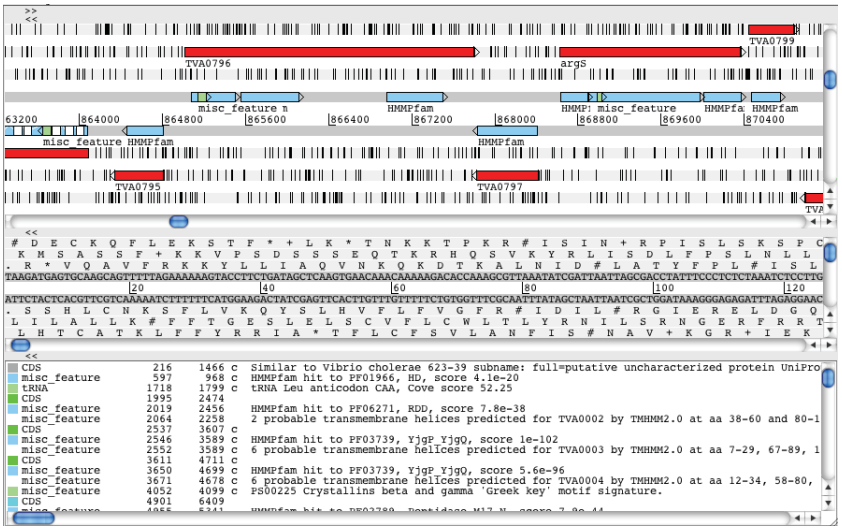


Figure 10: Screenshot of the genome view of the artemis software.

or the genome annotation system GenDB (Meyer et al., 2003) provide a cost effective solution especially for smaller research groups and projects with limited funding. The downside, however, is that these systems are only available as long as an internet connection is established. Additionally, the upload of unpublished data to servers of another university or company may raise the question of data privacy and security, especially if commercial partners are involved in the project. In these cases the installation of a local genome annotation system may be preferred over online services. A bottleneck of local annotation systems is that platform dependence, use of specific software versions as well as poor documentation and support, especially of open-source software, can result in an unintuitive and difficult installation process. Hybrid systems, such as the sequence visualization and annotation tool *Artemis* (Rutherford et al., 2000), make the installation obsolete and enable the use of local bioinformatics tools as well as the use of remote systems, i.e., computer grids or clusters (Figure 10). However, even the configuration of these out of the box solutions can be demanding for a normal user.

One problem that is found throughout all available annotation systems is that the incorporation of newly available bioinformatic tools is at best challenging if not prohibited. Thus, the user is left with those tools that are already included in the existing system and in various cases does not allow a proper configuration of the pipeline. Fur-

thermore, as annotation of metagenomes, prokaryotic genomes or eukaryotic genomes require different sets of tools, e.g. for gene prediction, different annotation pipelines have to be used for different projects. Therefore, prior to the annotation process the project leader or responsible annotator has to decide for the annotation pipeline that is most suitable regarding the current project requirements.

2

AIM OF THE STUDY

Main objectives

The main objective of the presented work was the determination and analysis of the pan-genome of the bacterial family *Vibrionaceae* using bioinformatics tools and approaches. The pan-genome of a group of organisms provides a wide range of genomic and phylogenetic information on different levels, from general implications on a group of organisms to specific genetic features of individual isolates. This study focused on the following objectives:

1. The determination of the genetic diversity among representatives of the bacterial family *Vibrionaceae*.
2. Inferring general implications of bacterial evolution and speciation based on the *Vibrionaceae* pan-genome.
3. The analysis of adaptation strategies of *Vibrionaceae* isolates using psychrophilic membrane proteins as a case model.

Secondary objectives

The development of bioinformatic tools for the annotation and analyze of prokaryotic genomes.

3

SUMMARY OF PAPERS

PAPER I

Unique core genomes of the bacterial family *Vibrionaceae*: insights into niche adaptation and speciation.

Tim Kahlke, Alexander Goesmann, Erik Hjerde, Nils-Peder Willassen and Peik Haugen (2012), *BMC Genomics*, 13:179

Bacterial genomes show a high level of genetic variability due to constant genome re-arrangement events, gene loss and the acquisition of genetic features through HGT. Especially the horizontal exchange of genetic material over taxon borders raises the question whether bacterial taxa can, in general, be distinguished by specific genetic features. The presented work addressed this question by identifying all genes that are shared exclusively by sub-groups of *Vibrionaceae* genomes. These genes were termed *unique core genes* as they are conserved genes, i.e., core genes, that are unique to a group of isolates. We investigated whether unique core genes are found in phylogenetically non-coherent (genophyletic) groups of genomes or in groups that included all isolates of a specific taxon (monophyletic groups). Furthermore, the importance of unique core genes on niche adaptation of bacterial species was determined for unique core genes of *V. cholerae*.

In summary, the results presented in this work reveal that all investigated taxa of the bacterial family *Vibrionaceae* possess sets of unique core genes which may aid the demarcation of bacterial taxa on a genome level. Furthermore, analysis of the unique core genes of *V. cholerae* indicate that these genes contribute to specific phenotypic features and play an important role in the adaptation of their host bacteria to its ecological niche. Although, we also determined unique core genes in genophyletic groups

of genomes these were shown to either include only few genomes of genophyletic groups or include only few genes.

PAPER II

The *Vibrionaceae* pan-genome hints to gene expression as the major driving force for unequal gene distributions on *Vibrionaceae* chromosomes

Tim Kahlke, Alexander Goesmann and Peik Haugen (Manuscript)

Representatives of the bacterial family *Vibrionaceae* possess two chromosomes which differ in gene composition and conservation. The larger chromosome, Chr I, carries the majority of conserved genes and shows high relative gene order conservation among *Vibrionaceae* species. The smaller chromosome, Chr II, is less conserved regarding gene order and is proposed to carry the majority of unique and taxon specific features. This led to the hypothesis that Chr II acts as an "evolutionary test bed" and plays an important role in the evolution of *Vibrionaceae* isolates. The aim of the presented study is the analysis of the distribution of core, unique and taxon specific genes on Chr I and Chr II to investigate the proposed role of Chr II in *Vibrionaceae* genomes.

The results presented in this paper do not reveal a prevalence for unique genetic features on Chr II. In fact, the main difference in the gene composition of Chr I and Chr II is an imbalance in the distribution of core genes which are mostly located on Chr I. In general, the results indicate that gene expression is the major driving force for the observed gene distribution on Chr I and Chr II. Also, no indications were found that support the hypothesis that Chr II plays a prevalent role in speciation and evolution of *Vibrionaceae* isolates.

PAPER III

Molecular characterization of cold adaptation of membrane proteins in the *Vibrionaceae* core-genome

Tim Kahlke and Steinar Thorvaldsen (2012), *PLoS One*, 7:e51761

Cold adapted (psychrophilic) bacteria have to overcome certain deleterious effects low temperature has on their metabolic machinery as well as on their morphology. For example, to maintain catalytic activity when subjected to cold conditions psychrophilic bacteria developed certain adaptation strategies that increase protein flexibility at the cost of protein stability. Also, low temperature significantly affects the fluidity of the lipid bilayer of the bacterial plasma membrane. Reduced membrane fluidity decreases the function of proteins embedded into or associated with the cell membrane. As the proper function of membrane proteins is vital for the cell, e.g., for molecule transport and signal transduction, psychrophilic bacteria alter the fatty acid composition of their lipid bilayer to maintain membrane fluidity. Yet, little is known about cold adaptation of membrane proteins on a molecular level. In this work the amino acid sequence of 66 membrane proteins present in the *Vibrionaceae* core genome is compared between groups of mesophilic, psychrotolerant and psychrophilic isolates to determine general adaptation strategies of membrane proteins to low temperature. The performed bioinformatical and statistical analysis reveals that those parts of the proteins that are located outside of the membrane and are in contact with the aqueous environment show structural changes similar to those of the loop regions of cold-adapted enzymes. However, sequence residues that are embedded inside the membrane do not show statistically significant changes which indicates that psychrophilic bacteria countervail the deleterious effects of low temperature on their cell membrane mainly by changes in the fatty acid composition of the lipid bilayer.

4

RESULTS AND DISCUSSION

The presented work is based on the comparison of all completely sequenced genomes of the bacterial family *Vibrionaceae* that were available in 2009, either in public databases or through in-house sequencing projects. The initial step of the project was the determination of the *Vibrionaceae* pan-genome which included the annotation of all 64 genomes and the clustering of homologous protein sequences. Subsequently, an in-depth analysis was performed to shed light on certain genetic characteristics.
This chapter is divided into two parts. Part one describes the main characteristics of the *Vibrionaceae* pan-genome and additionally introduces the bioinformatics framework *GePan* that was developed during this project and used for the annotation of the data set. The second part will discuss the results presented in **Paper I**, **Paper II** and **Paper III** in the context of bacterial adaptation and taxonomy.

4.1 DETERMINATION AND ANNOTATION OF THE *vibrionaceae* PAN-GENOME

The main objective of the presented thesis is the determination and analysis of the *Vibrionaceae* pan-genome. For this purpose a set of 64 completely sequenced *Vibrionaceae* genomes was compiled including clinical and environmental isolates, pathogenic and non-pathogenic strains as well as genome sequences from symbiotic *Vibrionaceae* (see Appendix for a detailed list). The CDSs of all genomes were re-predicted and annotated using the gene prediction and annotation framework *GePan* which was developed during this project. As described in **Paper I** a subsequent clustering step using the software orthoMCL was performed to determine clusters of homologous proteins. Unique genes were determined based on blast homology searches and a conservative percent identity cut-off of 70% over the complete sequence (see **Paper II** for details).

4.1.1 *The* Vibrionaceae *pan-genome*

As reported in **Paper II** the core genome of all analyzed strains is comprised of 758 clusters of homologous genes, which corresponds to 18% of the predicted 270,403 CDSs in our dataset. Additionally, the accessory genome and unique genes represent 78% and 3% of the CDSs, respectively. The average gene set of the analyzed genomes is therefore composed of 768 core genes, 3,314 accessory and 143 unique genes. Taking into account the variety of different ecological niches that are populated by representatives of this bacterial family it is not surprising that the majority of genes is part of the accessory genome. However, as discussed in **Paper II**, previous studies (Gu et al., 2009; Lilburn et al., 2010) reported a significantly higher amount of core genes among *Vibrionaceae* isolates. One possible explanation for the low number of conserved clusters in the presented study is that we excluded those homology clusters that were not consistent over multiple clustering processes with varying parameters (see **Paper I** for details). However, even the highest number of core clusters determined in any single orthoMCL run (859) is significantly smaller than reported in previous studies. Therefore, we favor the explanation that the diversity and size of the dataset analyzed led to the lower amount of core genes as it is by far the largest number of complete *Vibrionaceae* genomes ever included in a pan-genome analysis.

Comparison of the distribution of functional classes among core, unique and accessory genes was performed by assigning Gene Ontology (GO) terms to the complete gene set based on best blast hits to the Uniprot database (Figure 11). A conservative threshold of 70% sequence identity over the complete protein sequence was chosen to avoid false positive matches. GO terms were assigned to 25% (12351) of the core genes and 12% (26,272) of the accessory genes but no similarity above the chosen identity cut-off was found for any of the unique genes. As expected, the results show that certain functional classes associated with basic metabolic functions and housekeeping genes are over-represented in the core genome in comparison to the accessory genome. On the other hand, genes involved in adaptation to diverse environmental conditions, such as response to various stimuli, regulation and transport, are significantly over-represented among accessory genes. This supports the hypothesis that

accessory genes serve as a gene pool for the adaptation of bacterial isolates to different ecological niches.

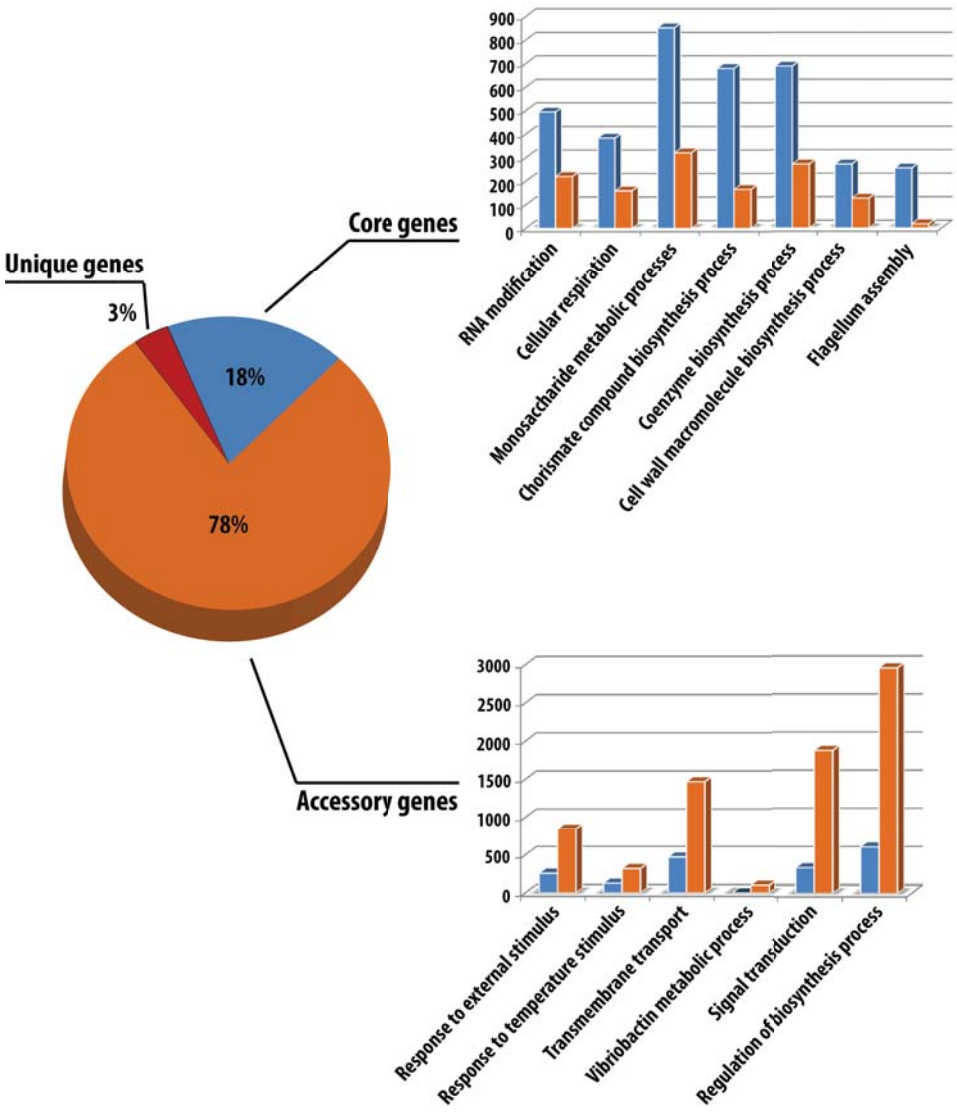


Figure 11: **The Vibrionaceae pan-genome.** Shown in the pie chart is the distribution of core genes (blue), accessory genes (orange) and unique genes (dark red) of all 270,403 CDSs in the dataset. The bar charts show the number of GO terms in the core genome (blue bars) and accessory genome (orange bars) for selected GO classes. No GO terms were assigned to unique genes due to low sequence similarities to known sequences.

4.1.2 *GePan - A bioinformatic framework for gene prediction and annotation*

A crucial step in the determination of the *Vibrionaceae* pan-genome was the initial re-annotation of all 64 complete genome sequences to provide equal annotation quality. Unfortunately, available solutions for the annotation of prokaryotic genomes are either costly or bound to a specific set of bioinformatics tools. Therefore, a software framework was developed that enabled the sequential execution of user defined tools to annotate all 64 investigated genomes. A first prototype consisted of ~65 Perl modules and classes containing ~12,000 lines of programming code. The modular implementation and a set of defined XML tags enabled the inclusion of new databases and bioinformatics tools into the pipeline with only few changes.

The annotation of a genome using the GePan framework includes three steps: (1) an initial gene prediction step, (2) the subsequent determination of potential gene annotations and (3) a final annotation of the predicted genes (Figure 12). The annotation of each gene is divided into functional, structural and transferred annotation, based on the different tools and databases used for the annotation. For example, by performing homology searches, e.g. using blast on databases of proteins of known function, it is possible to transfer the annotation of a high scoring hit sequence to a newly predicted gene. Similarly, the determination of functional domains of a particular CDS is used for its functional annotation. A final annotation step incorporates the results of the different annotation tools for each gene based on a set of predefined rules. Additionally, confidence levels are assigned to the annotation of each gene to indicate the quality of the determined annotation. The finished annotation of the complete genome is then exported into various file formats, including EMBL and Extensible Markup Language (XML).

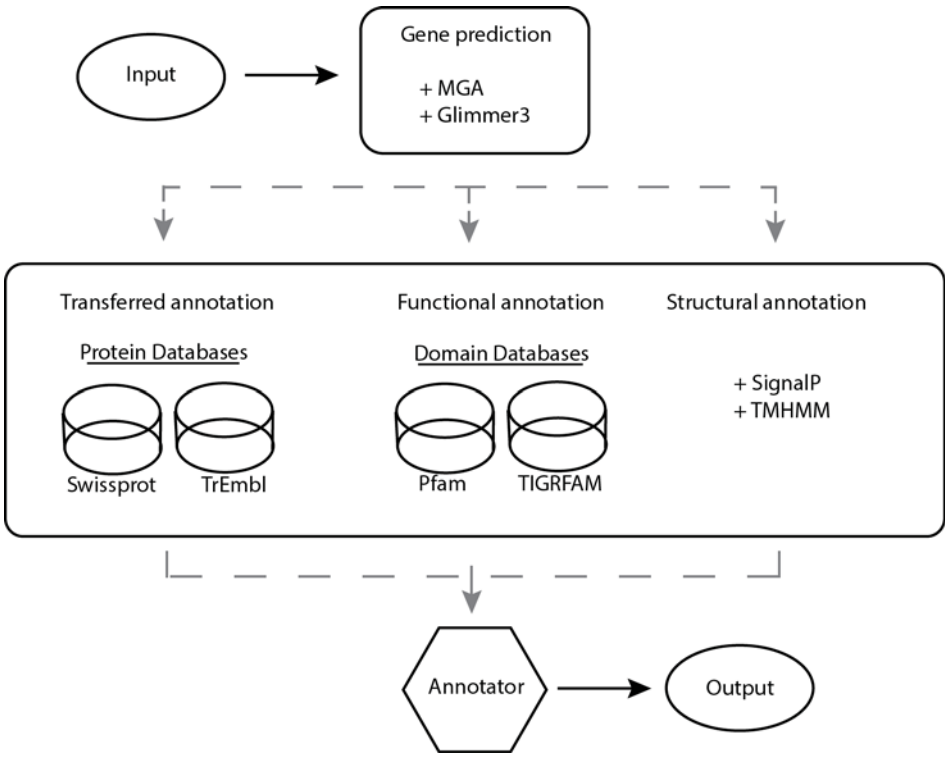


Figure 12: **Workflow of the bioinformatic GePan framework**. Shown are the tools and databases used in GePan for this thesis. Bacterial genome sequences in nucleotide fasta file are given as input files. Initial gene prediction is performed with user defined gene prediction software, e.g. Glimmer3. Subsequently, all predicted CDSs are compared to known protein sequences and functional domains. Additionally, user defined structural prediction tools can be applied, such as SignalP. The annotator incorporates the determined information of each CDS and exports a final genome annotation.

4.2 BACTERIAL SYSTEMATICS AND EVOLUTION

The plasticity of bacterial genomes is the cause for an ongoing controversy on how to define bacterial taxa. In contrast to bacteria, eukaryotic taxa are demarcated based on their genomic similarity which reflects their evolutionary relationship. On a species level they are defined by the ability of two individuals to sexually reproduce. Further classification into higher taxa is mostly based on morphological, metabolic or behavioral traits which also reflect genetic similarity. However, it is obvious that the eukaryotic species concept is doomed to fail when applied to bacteria due to their asexual reproduction. Thus, the classification of bacterial taxa, including bacterial species, must be based on the identification of other features such as similarity of gene or protein sequences as well as metabolic and morphological traits. However, the various adaptation strategies that lead to specific phenotypic traits and eventually to the separation into distinct taxa complicate the definition of what accounts for a bacterial taxon.

4.2.1 *Implications of unique core genes on bacterial taxonomy*

One of the most evident approaches to identify taxon specific features is the differential comparison of the gene sets of related organisms (Huynen and Bork, 1998). The identification of genes that are (a) found in all representatives of a bacterial taxon and (b) are not present in closely related organisms can aid the demarcation of bacteria based on distinct genetic and therefore metabolic features. However, as these *unique core genes* are likely to encode evolutionary advantageous functions they also tend to be acquired horizontally by isolates of different taxa that populate the same ecological niche. Thus, although it is theoretically possible that bacterial taxa possess unique core genes, their general existence was put into question (Gevers et al., 2005).
As reported in **Paper I** we investigated all available *Vibrionaceae* genomes for the presence of unique core genes and analyzed their phylogenetic relationship and functional annotation. Surprisingly, we were able to identify unique core genes for all *Vib-*

rionaceae species and genera in our dataset. In fact, even for species as closely related as *V. cholerae* and *Vibrio mimicus* we were able to identify 12 and 67 unique core genes, respectively. These results were also supported by a recent publication that identified unique core genes (called signature genes) for all major branches of the bacterial domain (Dutilh et al., 2008).

Despite the obvious implications of taxon specific unique core genes, e.g. for the development of vaccines and clinical diagnostics, the question of how unique core genes contribute to the challenges of bacterial systematics still remains. As presented in **Paper I**, our results indicate that taxon specific unique core genes contribute to specific characteristics of their phylogenetic lineage, such as the ability of *V. cholerae* to sense oxygen as well as the specific iron uptake and utilization through the iron chelator Vibriobactin. As these functions are not present in representatives of closely related species they may be used to build up bacterial systematics based on sets of unique metabolic characteristics. Unfortunately, not all unique core genes contribute to the specific phenotype of the host taxon (Figure 13). For example, gene artifacts such as disrupted metabolic functions and horizontally acquired genes, e.g., phage assembly genes, may also be found exclusively in a sub-group of investigated organisms without any contribution to a specific phenotype. Thus, in order to identify "true" unique core genes, i.e., genes that contribute to specific characteristics on a gene level, accurate annotation of the genome sequences in question is crucial. However, many genes still lack functional annotations which makes the identification of true unique core genes challenging. This is especially true for genes of isolates that are not closely related to model organisms, such as *E. coli*. Furthermore, as reported in **Paper I** some unique core genes are found in groups of genomes that do not belong to the same or a closely related taxon. These genes are most likely horizontally acquired and their contribution to the ecological fitness of the individual strains remains unknown.

In summary, our data suggests that unique core genes can be used to aid the demarcation of bacterial taxa by identifying specific phenotypic traits that are not found in related lineages. Furthermore, unique core genes are of major importance for clinical diagnostics and can help to develop vaccines that are specific to a phylogenetic lineage.

However, unique core genes can not be used for the demarcation of bacterial taxa by themselves.

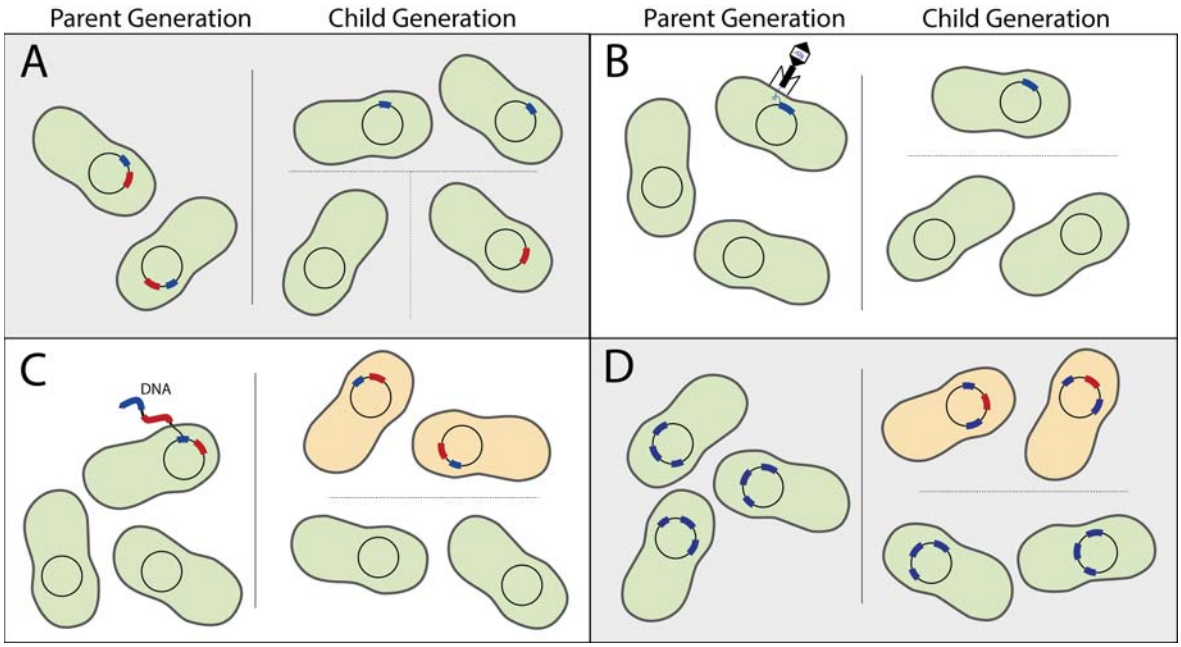


Figure 13: **Differences in unique core genes**. Bacterial cells (green and orange) with one circular chromosome are shown. Top row: possible events that lead to child generations with identical phenotypic characteristics but differences in the unique core genomes. (A) Incomplete deletion of a genetic feature from the parent generation that was encoded by two genes (blue and red). (B) Insertion of phage assembly genes (blue) into the genome of one representative of the parent generation.
Bottom row: possible events that lead to child generations that differ in unique core genomes as well as in phenotype. (C) HGT of a two gene feature (blue and red) into the genome of one representative of the parent generation. (D) Mutation/adaptation of an existing phenotypic feature by mutation of one (red) of three genes (blue) involved in its expression.

4.2.2 *Core genes and niche adaptation*

A major driving force in bacterial evolution is the mutation of existing genes and the "survival of the fittest" based on natural selection. Hence, the exclusive determination

of the absence or presence of genes will inevitably overlook those specific adaptations that are based on differences in the sequence of conserved, i.e., core genes. As reported in **Paper III**, we investigated the amino acid sequence of membrane proteins that are part of the *Vibrionaceae* core genome to identify general adaptation strategies of psychrophilic bacteria to stresses of the cold. Low temperature has deleterious effects on the lipid bilayer of the cell membrane and it is therefore legitimate to assume that proteins of psychrophilic bacteria that are embedded into the membrane or associated to it will undergo certain conformational changes. However, the results reveal that only sequence parts that are located outside of the lipid bilayer show adaptational changes in their amino acid sequence. On the other hand, sequence parts that are embedded into the membrane revealed no or only weak differences in their amino acid sequence. The results presented in **Paper III** have various implications on bacterial systematics. First they indicate how relatively subtle the genomic changes can be that lead to specific characteristics, i.e., the adaptation to a certain ecological niche. For example, as shown in Table 2 of **Paper III** the sequence similarity of membrane proteins among organisms of the same temperature group can be as low as among bacteria of different temperature groups. Hence, differences in the protein sequence can be caused by neutral mutations due to phylogenetic distance as well as by adaptation to certain environmental conditions. This not only complicates the identification of bacterial adaptation strategies but also shows how adaptational changes can bias phylogenetic analysis such as MLSA as it is based on concatenated core genes. In fact, to our knowledge no bacterial species includes representatives that populate different temperature zones, e.g. psychrophilic and mesophilic organisms.

In addition to the above, the results presented in **Paper III** also hint to another possible adaptation strategy: changes in the expression of core genes. As reported in **Paper III** those parts of the membrane proteins that are embedded in the lipid bilayer show no significant differences between psychrophilic and mesophilic *Vibrionaceae*. Thus, psychrophilic bacteria mainly countervail the deleterious effects of low temperature on membranes by alteration of its fatty acid composition (Shivaji and Prakash, 2010). However, this *homeoviscous adaptation* does not require the synthesis of fatty acids that are exclusively found in membranes of psychrophilic bacteria. It is rather based on

a change in the expression levels of common housekeeping genes, i.e., core genes. Similarly, it has been hypothesized that bacteria adapt to changing environmental conditions by gene duplication in order to increase expression levels of certain genes (Gevers et al., 2004; Kondrashov, 2012). This adds another layer of complexity to the mechanisms that lead to the development of specific phenotypes of bacteria and eventually to demarcation into separate taxa.

4.2.3 *Does interchromosomal translocation play a role in niche adaptation?*

Despite the constant genome re-arrangements, bacterial chromosomes show certain patterns of gene organization that are mostly associated with gene expression (Rocha, 2004, 2008). For example, the relative distance of a gene to oriC can have significant effects on its expression level. Genes located closer to oriC than to the replication terminator terC tend to be highly expressed due to higher copy numbers in the early phase of the replication (gene dosage effects). In **Paper II** we determined the distribution of pan-genes on Chr I and Chr II for those 12 genomes in our dataset that are completely assembled. Our results reveal that the majority of core genes is located on Chr I with a prevalence of genes located closer to oriC than to replication terminator (terC). This is in accordance with the hypothesis that core genes tend to be highly expressed. Nevertheless, a small fraction of core genes is also located on Chr II which was proposed to be the reason for its preservation in the genomes of *Vibrionaceae* species (Heidelberg et al., 2000; Cooper et al., 2010). However, our results show that the set of core genes that is located on Chr II is not consistent among the different isolates. In fact, only an average of 50% of the core genes located on Chr II is located on this chromosome in all 12 investigated genomes (Table 3) which indicates constant interchromosomal translocations of even core genes. This is surprising due to the significant differences in the expression levels of genes located on Chr I in comparison to genes on Chr II (Dryselius et al., 2008; Toffano-Nioche et al., 2012). Therefore, the translocation of genes from one chromosome to the other will most likely affect their expression in some way. A study recently published by Morrow *et al.* (2012) on *Burkholderia* genomes indicates that the expression level of genes change when translocated from one chromosome to another.

Table 3: **Core genes with conserved location on Chr I and Chr II**. Shown is the amount of core genes with conserved location on the same chromosome in all isolates. Percentage is given in relation to the total of core genes found on the particular chromosome.

Strain name	**Number of core genes conserved on Chr I**	**Number of core genes conserved on Chr II**
V. cholerae str. M66-2	649 (93.7%)	34 (49.3%)
V. cholerae str. MJ1236	649 (93.7%)	34 (49.3%)
V. cholerae str. O1 N16961	650 (93.7%)	34 (49.3%)
V. cholerae str. O395	649 (93.7%)	34 (49.3%)
V. harveyi str. ATCC-BAA 1116	652 (93%)	34 (53.1%)
V. parahaemolyticus str. RIMD 2210633	649 (93%)	35 (55.6%)
V. splendidus str. LGP32	650 (93.5%)	34 (49.3%)
V. vulnificus str. CMCP6	651 (93%)	34 (52.3%)
V. vulnificus str. YJ016	650 (93%)	34 (54%)
A. fischeri str. ES114	654 (97%)	34 (36.2%)
A. salmonicida str. LFI1238	652 (96.7%)	34 (38.2%)
P. profundum str. SS9	653 (91.7%)	35 (58.3%)

But does the change in expression levels affect the bacterium? Given the importance of core genes as conserved housekeeping genes, it is unlikely that a change in their expression values will not have any impact on the host organism. This is also shown by the fact that genome rearrangements which significantly disrupt the relative gene order, e.g., the distance to oriC are often fatal for the bacterium (Eisen et al., 2000; Mackiewicz et al., 2001). Therefore, it is legitimate to assume that the translocation of core genes from Chr I to Chr II and vice versa affects the corresponding metabolic functions and thus might contribute to specific behavior of the individual organism. Future in-depth analysis will show whether it is possible to link certain adaptations to the expression levels of core genes that are located on different chromosomes in different *Vibrionaceae* species.

5

CONCLUDING REMARKS

Representatives of the bacterial family *Vibrionaceae* populate almost all aquatic habitats whether it is fresh, sea or brackish water. They are found even under the most hostile environmental conditions, e.g., in the deep sea with pressure of several hundred atmosphere or the Arctic ocean with temperature close to zero. Additionally, many *Vibrionaceae* species include bacteria that interact with eukaryotic hosts, either as pathogens or symbionts. Due to this diversity, the comparison of this bacterial family can shed light on the different strategies and evolutionary mechanisms that lead to the adaptation of such a wide range of ecological niches and eventually to the demarcation into distinct taxa. The investigation of the *Vibrionaceae* pan-genome presented here revealed the abundance of mechanisms that lead to the development of specific phenotypic traits and distinct bacterial taxa. As reported in **Paper I**, some specific characteristics are based on genes that are found exclusively in isolates of certain taxa. Here, they either represent complete genetic features, e.g., aerotaxis in *V. cholerae*, or contribute to conserved features in a new way, such as the iron utilization of *V. cholerae* through Vibriobactin. However, adaptation does not necessarily relate to the introduction or deletion of complete genes or sets of genes. More often, the adaptation to specific environmental conditions is based on relatively subtle modifications of the protein sequences of bacteria. As shown in **Paper III**, even the deleterious effects of low temperature can be countervailed with relatively moderate adaptations of the amino acid sequence of core genes. Furthermore, the results of **Paper II** and **Paper III** indicate the impact changes in the expression values of existing genes can have on bacterial evolution.

The findings presented in **Paper I** of this study may enable the future development a bacterial taxonomy that is based on gene content. The determination of core and unique core genes of a broader range of isolates might lead to the identification of

genetic fingerprints for all known bacterial taxa which can be used in the same way as genetic marker genes are used to identify certain bacterial pathogens today. Furthermore, the identification of bacterial adaptation strategies as presented in **Paper II** and **Paper III**, increases the general knowledge about bacterial evolution which is of major importance, e.g., to prevent bacterial antibiotic resistance and to increase productivity of bacterial strains that are used in commercial biotechnology.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., 1999. Basic local alignment search tool. Journal of Molecular Biology 215, 403–410.

Alvarez, M., Zeelen, J. P., Mainfroid, V., Rentier-Delrue, F., Martial, J. A., Wyns, L., Wierenga, R. K., Maes, D., 1998. Triose-phosphate isomerase (tim) of the psychrophilic bacteriumvibrio marinus. kinetic and structural properties. The Journal of Biological Chemistry 273, 2199–2206.

Andersen, O. S., Koeppe, R. E., 2007. Bilayer thickness and membrane protein function: An energetic perspective. Annu. Rev. Biophys. Biomol. Struct. 36, 107–130.

Asmahan, A. A., 2010. Beneficial role of lactic acid bacteria in food preservation and human health: A review. Research Journal of Microbiology 5, 1213 – 1221.

Atsumi, T., McCartert, L., Imae, Y., 1992. Polar and lateral flagellar motors of marine vibrio are driven by different ion-motive forces. Nature 355, 182 – 184.

Auch, A. F., von Jan, M., Klenk, H.-P., Göker, M., 2010. Digital dna-dna hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci 2 (1), 117–134.
URL http://dx.doi.org/10.4056/sigs.531120

Avery, O. T., MacLeod, C. M., McCarty, M., 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. The Journal of Experimental Medicine 79, 137–158.

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.
URL http://dx.doi.org/10.1186/1471-2164-9-75

Baril, C., Richaud, C., Baranton, G., Saint-Girons, I., 1989. Linear chromosome of borrelia burgdorferi. Res Microbiol 140, 507 – 516.

Barnett, M., Fisher, R., Jones, T., Komp, C., Abola, A. P., Barloy-Hubler, F., Bowser, L., Capela, D., Galibert, F., Gouzy, J., Gurjal, M., Hong, A., Huizar, L., Hyman, R. W., Kahn, D., Kahn, M. L., Kalman, S., Keating, D. H., Palm, C., Peck, M. C., Surzycki, R., Wells, D. H., Yeh, K. C., Davis, R. W., Federspiel, N. A., Longm, S. R., 2001. Nucleotide sequence and predicted functions of the entire sinorhizobium meliloti psyma megaplasmid. Proceedings of the National Academy of Sciences of the United States of America 98, 9883–8.

Baumdicker, F., Hess, W. R., Pfaffelhuber, P., 2012. The infinitely many genes model for the distributed genome of bacteria. Genome Biol Evol 4 (4), 443–456.
URL http://dx.doi.org/10.1093/gbe/evs016

Baxevanis, A. D., Ouellette, B. F. F. (Eds.), 2001. Bioinformatics - A practical guide to the analysis of genes and proteins, 2nd Edition. Wiley-Interscience.

Beadle, G. W., Nov 1941. Genetic control of biochemical reactions in neurospora. Proceedings of the National Academy of Sciences 27 (11), 499–506.
URL http://dx.doi.org/10.1073/pnas.27.11.499

Bentley, S., 2009. Sequencing the species pan-genome. Nature 7, 258 – 259.

Berg, R. D., 1996. The indigenous gastrointestinal microflora. Trends in Microbiology 4, 430 – 435.

Bergthorsson, U., Ochman, H., Oct 1995. Heterogeneity of genome sizes among natural isolates of escherichia coli. J Bacteriol 177 (20), 5784–5789.

Bozza, C. G., Pawlowski, W. P., 2008. The cytogenetics of homologous chromosome pairing in meiosis in plants. Cytogenet Genome Res 120 (3-4), 313–319.
URL http://dx.doi.org/10.1159/000121080

Callister, S. J., McCue, L. A., Turse, J. E., Monroe, M. E., Auberry, K. J., Smith, R. D., Adkins, J. N., Lipton, M. S., 2008. Comparative bacterial proteomics: analysis of the core genome concept. PLoS One 3 (2), e1542.
URL http://dx.doi.org/10.1371/journal.pone.0001542

Canfield, D. E., Glazer, A. N., Falkowski, P. G., 2010. The evolution and future of earth?s nitrogen cycle. Science 330, 192–196.

Chan, J. Z.-M., Halachev, M. R., Loman, N. J., Constantinidou, C., Pallen, M. J., 2012. Defining bacterial species in the genomic era: insights from the genus acinetobacter. BMC Microbiol 12, 302.
URL http://dx.doi.org/10.1186/1471-2180-12-302

Charlebois, R. L., Doolittle, W. F., Dec 2004. Computing prokaryotic gene ubiquity: rescuing the core from extinction. Genome Res 14 (12), 2469–2477.
URL http://dx.doi.org/10.1101/gr.3024704

Chattoraj, D. K., 2000. Control of plasmid dna replication by iterons: no longer paradoxical. Molecular Microbiology 37, 467 – 476.

Chen, F., Mackey, A. J., Vermunt, J. K., Roos, D. S., 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS One 2 (4), e383.
URL http://dx.doi.org/10.1371/journal.pone.0000383

Chial, H., 2008. Dna sequencing technologies key to the human genome project. Nature Education 1, 1.

Cohan, F. M., 2002. What are bacterial species? Annual Review of Microbiology 56, 457–487.

Collins, T., Meuwis, M.-A., Gerday, C., Feller, G., 2003. Activity, stability and flexibility in glycosidases adapted to extreme thermal environments. J. Mol. Biol. 328, 419–428.

Cooper, V. S., Vohr, S. H., Wrocklage, S. C., Hatcher, P. J., Apr 2010. Why genes evolve faster on secondary chromosomes in bacteria. PLoS Comput Biol 6 (4), e1000732.
URL http://dx.doi.org/10.1371/journal.pcbi.1000732

Coyer, J. A., Cabello-Pasini, A., Swift, H., Alberte, R. S., Apr 1996. N2 fixation in marine heterotrophic bacteria: dynamics of environmental and molecular regulation. Proc Natl Acad Sci U S A 93 (8), 3575–3580.

D'Amico, S., Collins, T., Marx, J.-C., Feller, G., Gerday, C., 2006. Psychrophilic microorganisms: challenges for life. EMBO reports 7, 385–389.

Darwin, C. R., 1859. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray.

Daubin, V., Moran, N. A., Ochman, H., Aug 2003. Phylogenetics and the cohesion of bacterial genomes. Science 301 (5634), 829–832.
URL http://dx.doi.org/10.1126/science.1086568

D'Auria, S., Aurilia, V., Marabotti, A., Gonnelli, M., Strambini, G., 2009. Structure and dynamics of cold-adapted enzymes as investigated by phosphorescence spectroscopy and molecular dynamics studies. 2. the case of an esterase from pseudoalteromonas haloplanktis. The Journal of Physical Chemistry B 113, 13171–13178.

Doolittle, W. F., 1999. Lateral genomics. Trend in Cell Biology 9, M5–8.

Dryselius, R., Izutsu, K., Honda, T., Iida, T., 2008. Differential replication dynamics for large and small vibrio chromosomes affect gene dosage, expression and location. BMC Genomics 9, 559.
URL http://dx.doi.org/10.1186/1471-2164-9-559

Dryselius, R., Kurokawa, K., Iida, T., 2007. Vibrionaceae, a versatile bacterial family with evolutionarily conserved variability. Res Microbiol 158 (6), 479–486.
URL http://dx.doi.org/10.1016/j.resmic.2007.04.007

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., et al., Sep 2012. An integrated encyclopedia of

DNA elements in the human genome. Nature 489 (7414), 57–74.
URL http://dx.doi.org/10.1038/nature11247

Dutilh, B. E., Snel, B., Ettema, T. J. G., Huynen, M. A., Aug 2008. Signature genes as a phylogenomic tool. Mol Biol Evol 25 (8), 1659–1667.
URL http://dx.doi.org/10.1093/molbev/msn115

Egan, E. S., Fogel, M. A., Waldor, M. K., Jun 2005. Divided genomes: negotiating the cell cycle in prokaryotes with multiple chromosomes. Mol Microbiol 56 (5), 1129–1138.
URL http://dx.doi.org/10.1111/j.1365-2958.2005.04622.x

Egan, E. S., Waldor, M. K., Aug 2003. Distinct replication requirements for the two vibrio cholerae chromosomes. Cell 114 (4), 521–530.

Ehrlich, G. D., Hiller, N. L., Hu, F. Z., 2008. What makes pathogens pathogenic. Genome Biology 9, 225.

Ehrlich, G. D., Hu, F. Z., Shen, K., Stoodley, P., Post, J. C., Aug 2005. Bacterial plurality as a general mechanism driving persistence in chronic infections. Clin Orthop Relat Res (437), 20–24.

Eisen, J. A., Heidelberg, J. F., White, O., Salzberg, S. L., 2000. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol 1 (6), RESEARCH0011.
URL http://dx.doi.org/10.1186/gb-2000-1-6-research0011

Feller, G., D'Amico, D., Gerday, C., 1999. Thermodynamic stability of a cold-active r-amylase from the antarctic bacterium alteromonas haloplanctis. Biochemistry 38, 4613–4619.

Feller, G., Narinx, E., Arpigny, J. L., Aittaleb, M., Baise, E., Genicot, S., Gerday, C., 1996. Enzymes from psychrophilic organisms. FEMS Microbiology Reviews 18, 189–202.

Fitz-Gibbons, S. T., House, C. H., 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Ressearch 27, 4218–4222.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., Jul 1995. Whole-genome random sequencing and assembly of haemophilus influenzae rd. Science 269 (5223), 496–512.

Fox, G. E., Pechman, K. R., Woese, C. R., 1977. Comparative cataloging of 16s ribosomal ribonucleic acid: Molecular approach to procaryotic systematics. International Journal of Systematic and Evolutionary Microbiology 27, 44–57.

Fox, G. E., Wisotzkey, J. D., Jurtshuk, JR, P., 1992. How close is close: 16s rrna sequence identity may not be sufficient to guarantee species identity. International Journal of Systematic Bacteriology 42, 166–170.

Fuller, R. S., Funnell, B. E., Kronberg, A., 1984. The dnaa protein complex with the e. coli chromosomal replication origin (oric) and other dna sites. Cel 38, 889–900.

Gardner, P. P., Daub, J., Tate, J., Moore, B. L., Osuch, I. H., Griffiths-Jones, S., Finn, R. D., Nawrocki, E. P., Kolbe, D. L., Eddy, S. R., Bateman, A., Jan 2011. Rfam: Wikipedia, clans and the "decimal" release. Nucleic Acids Res 39 (Database issue), D141–D145. URL http://dx.doi.org/10.1093/nar/gkq1129

Georgiades, K., Raoult, d., 2010. Defining pathogenic bacterial species in the genomic era. Frontiers in Microbiology 1, 151.

Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., FeFeil, E., Stackebrandt, E., Van de Peer, Y., Vandamme, P., Thompson, F. L., Swing, J., 2005. Re-evaluating prokaryotic species. Nature Reviews 3, 733–739.

Gevers, D., Vandepoele, K., Cedric, S., Van de Peer, Y., 2004. Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends in Microbiology 12, 148–154.

Glasner, J. D., Marquez-Villavicencio, M., Kim, H.-S., Jahn, C. E., Ma, B., Biehl, B. S., Rissman, A. I., Mole, B., Yi, X., Yang, C.-H., Dangl, J. L., Grant, S. R., Perna, N. T., Charkowski, A. O., Dec 2008. Niche-specificity and the variable fraction of the pec-

tobacterium pan-genome. Mol Plant Microbe Interact 21 (12), 1549–1560.
URL http://dx.doi.org/10.1094/MPMI-21-12-1549

Goudreau, P. N., Stock, A. M., 1998. Signal transduction in bacteria: molecular mechanisms of stimulus?response coupling. Current Opinion in Microbiology 1, 160 – 169.

Gould, S. J., 1996. Planet of the bacteri. Washington Post Horizon 119, 344.

Grice, E. A., Kong, H. H., Conlan, S., Deming, C. B., Davis, J., Young, A. C., Program, N. C. S., Bouffard, G. G., Blakesley, R. W., Murray, P. R., Green, E. D., Turner, M. L., Segre, J. A., 2009. Topographical and temporal diversity of the human skin microbiome. Science 324, 1190–1192.

Gu, J., Neary, J., Cai, H., Moshfeghian, A., Rodriguez, S. A., Lilburn, T. G., Wang, Y., 2009. Genomic and systems evolution in vibrionaceae species. BMC Genomics 10 Suppl 1, S11.
URL http://dx.doi.org/10.1186/1471-2164-10-S1-S11

Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Clayton, R. A., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Umayam, L., Gill, S. R., Nelson, K. E., Read, T. D., Tettelin, H., Richardson, D., Ermolaeva, M. D., Vamathevan, J., Bass, S., Qin, H., Dragoi, I., Sellers, P., McDonald, L., Utterback, T., Fleishmann, R. D., Nierman, W. C., White, O., Salzberg, S. L., Smith, H. O., Colwell, R. R., Mekalanos, J. J., Venter, J. C., Fraser, C. M., Aug 2000. Dna sequence of both chromosomes of the cholera pathogen vibrio cholerae. Nature 406 (6795), 477–483.
URL http://dx.doi.org/10.1038/35020000

Helmke, E., Weyland, H., 2004. Psychrophilic versus psychrotolerant bacteria–occurrence and significance in polar and temperate marine habitats. Cellular and Molecular Biology (Noisy-Le-Grand, France) 50, 553–561.

Henz, S. R., Huson, D. H., Auch, A. F., Nieselt-Struwe, K., Schuster, S. C., 2004. Whole-genome prokaryotic phylogeny. Bioinformatics 21, 2329–2335.

Hilario, E., Gogarten, J. P., 1993. Horizontal transfer of atpase genes - the tree of life becomes a net of life. Biosystems 31, 111–119.

Hiller, N. L., Janto, B., Hogg, J. S., Boissy, R., Yu, S., Powell, E., Keefe, R., Ehrlich, N. E., Shen, K., Hayes, J., Barbadora, K., Klimke, W., Dernovoy, D., Tatusova, T., Parkhill, J., Bentley, S. D., Post, J. C., Ehrlich, G. D., Hu, F. Z., Nov 2007. Comparative genomic analyses of seventeen streptococcus pneumoniae strains: insights into the pneumococcal supragenome. J Bacteriol 189 (22), 8186–8195.
URL http://dx.doi.org/10.1128/JB.00690-07

Hogg, J. S., Hu, F. Z., Janto, B., Boissy, R., Hayes, J., Keefe, R., Post, J. C., Ehrlich, G. D., 2007. Characterization and modeling of the haemophilus influenzae core and supragenomes based on the complete genomic sequences of rd and 12 clinical nontypeable strains. Genome Biol 8 (6), R103.
URL http://dx.doi.org/10.1186/gb-2007-8-6-r103

Huson, Daniel H. abd Steel, M., 2004. Phylogenetic trees based on gene content. Bioinformatics 20, 2044–2049.

Huynen, M., Dandekar, T., Bork, P., Apr 1998. Differential genome analysis applied to the species-specific features of helicobacter pylori. FEBS Lett 426 (1), 1–5.

Huynen, M. A., Bork, P., May 1998. Measuring genome evolution. Proc Natl Acad Sci U S A 95 (11), 5849–5856.

Johannsen, W., 1905. Arvelighedslærens elementer: forelæsninger holdte ved Københavns universitet. Gyldendal.

Johnson, J. L., 1973. Use of nucleic-acid homologies in the taxonomy of anaerobic bacteria. International Journal of Systematic Bacteriology 23, 308–315.

Jordan, I. K., Makarova, K. S., Spouge, J. L., Wolf, Y. I., Koonin, E. V., Apr 2001. Lineage-specific gene expansions in bacterial and archaeal genomes. Genome Res 11 (4), 555–565.
URL http://dx.doi.org/10.1101/gr.166001

Kestin, J., Sokolov, M., Wakeham, W. A., 1978. Viscosity of liquid water in the range of -8 c to 150 c. Journal of Physical and Chemical Reference Data 7, 941–948.

Kondrashov, F. A., Dec 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc Biol Sci 279 (1749), 5048–5057.
URL http://dx.doi.org/10.1098/rspb.2012.1108

Konings, W., Albers, S. V., Koning, S., Driessen, A., 2002. The cell membrane plays a crucial role in survival of bacteria and archaea in extreme environments. Antonie Van Leeuwenhoek 81, 61–72.

Konstantinidis, K. T., Ramette, A., Tiedje, J. M., Nov 2006. The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci 361 (1475), 1929–1940.
URL http://dx.doi.org/10.1098/rstb.2006.1920

Koonin, E. V., Nov 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol 1 (2), 127–136.
URL http://dx.doi.org/10.1038/nrmicro751

Koonin, E. V., Makarova, K. S., Aravind, L., 2001. Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol 55, 709–742.
URL http://dx.doi.org/10.1146/annurev.micro.55.1.709

Koonin, E. V., Mushegian, A. R., Bork, P., Sep 1996. Non-orthologous gene displacement. Trends Genet 12 (9), 334–336.

Kotrba, P., Inui, M., Yukawa, H., 2001. Bacterial phosphotransferase system (pts) in carbohydrate uptake and control of carbon metabolism. Journal of Bioscience and Biongineering 92, 502 – 517.

Laing, C. R., Zhang, Y., Thomas, J. E., Gannon, V. P. J., Nov 2011. Everything at once: comparative analysis of the genomes of bacterial pathogens. Vet Microbiol 153 (1-2), 13–26.
URL http://dx.doi.org/10.1016/j.vetmic.2011.06.014

Lapierre, P., Gogarten, J. P., Mar 2009. Estimating the size of the bacterial pan-genome. Trends Genet 25 (3), 107–110.
URL http://dx.doi.org/10.1016/j.tig.2008.12.004

Lefébure, T., Stanhope, M. J., 2007. Evolution of the core and pan-genome of streptococcus: positive selection, recombination, and genome composition. Genome Biol 8 (5), R71.
URL http://dx.doi.org/10.1186/gb-2007-8-5-r71

Legault, B. A., Lopez-Lopez, A., Alba-Casado, J. C., Doolittle, W. F., Bolhuis, H., Rodriguez-Valera, F., Papke, R. T., 2006. Environmental genomics of "haloquadratum walsbyi" in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. BMC Genomics 7, 171.
URL http://dx.doi.org/10.1186/1471-2164-7-171

Lenaz, G., 1987. Lipid fluidity and membrane protein dynamics. Bioscience Reports 7, 823 – 837.

Lerat, E., Daubin, V., Moran, N. A., 2003. From gene trees to organismal phylogeny in prokaryotes: The case of the gamma-proteobacteria. PLoS Biology 1, E19.

Lerat, E., Daubin, V., Ochman, H., Moran, N. A., May 2005. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol 3 (5), e130.
URL http://dx.doi.org/10.1371/journal.pbio.0030130

Li, L., Stoeckert, C. J. J., Roos, D. S., 2003. Orthomcl: identification of ortholog groups for eukaryotic genomes. Genome Research 13, 2178 – 2189.

Lilburn, T. G., Gu, J., Cai, H., Wang, Y., 2010. Comparative genomics of the family vibrionaceae reveals the wide distribution of genes encoding virulence-associated proteins. BMC Genomics 11, 369.
URL http://dx.doi.org/10.1186/1471-2164-11-369

Lipman, D., Pearson, W. R., 1985. Rapid and sensitive protein similarity searches. Science 227, 1435–1441.

Mackiewicz, P., Mackiewicz, D., Kowalczuk, M., Cebrat, S., 2001. Flip-flop around the origin and terminus of replication in prokaryotic genomes. Genome Biol 2 (12), INTERACTIONS1004.

Makino, K., Oshima, K., Kurokawa, K., Yokoyama, K., Uda, T., Tagomori, K., Iijima, Y., Najima, M., Nakano, M., Yamashita, A., Kubota, Y., Kimura, S., Yasunaga, T., Honda, T., Shinagawa, H., Hattori, M., Iida, T., Mar 2003. Genome sequence of vibrio parahaemolyticus: a pathogenic mechanism distinct from that of v cholerae. Lancet 361 (9359), 743–749.
URL http://dx.doi.org/10.1016/S0140-6736(03)12659-1

Mayr, E., 1942. Systematics and the Origin of Species, from the Viewpoint of a Zoologist. Harvard University Press, New York.

McFadden, G. I., 2001. Chloroplast origin and integration. Plant Physiology 125, 50–53.

McGibbon, L., Russel, N. J., 1983. Fatty acid positional distribution in phospholipids of a psychrophilic bacterium during changes in growth temperature. Current Microbiology 9, 241 – 244.

McInerney, J., Cotton, J. A., Pisani, D., 2008. The prokaryotic tree of life: past present ... and future? Trends in Ecology & Evolution 23, 276–281.

McTaggart, L. R., Richardson, S. E., Witkowska, M., Zhang, S. X., 2010. Phylogeny and identification of nocardia species on the basis of multilocus sequence analysis. Journal of Clinical Microbiology 48, 4525–4533.

Mendel, G., 1866. Versuche über pflanzenhybriden. Verhandl d Naturfsch Ver in Brünn 4, 3–47.

Metpally, R. P., Reddy, B. V., 2009. Comparative proteome analysis of psychrophilic versus mesophilic bacterial species: Insights into the molecular basis of cold adaptation of proteins. BMC Genomics 10.

Meyer, F., Goesmann, A., McHardy, A. C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R., Pühler, A., Apr 2003. Gendb–an open source

genome annotation system for prokaryote genomes. Nucleic Acids Res 31 (8), 2187–2195.

Mira, A., Martín-Cuadrado, A. B., D'Auria, G., Rodríguez-Valera, F., Jun 2010. The bacterial pan-genome:a new paradigm in microbiology. Int Microbiol 13 (2), 45–57.

Moore, G. E., 1965. Cramming more components onto integrated circuits. Electronics 38 (8).

Moore, G. E., 1975. Progress in digital integrated electronics. In Electron Devices Meeting, 1975 International 21, 11–13.

Morita, R. Y., 1975. Psychrophilic bacteria. Bacteriol. Rev. 39, 144.

Moyer, C. L., Morita, R. Y., 2007. Psychrophiles and psychrotrophs. In: Encyclopedia of Life Sciences. Chichester: John Wiley & Sons, Ltd.

Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, Hajime Dunbar, H. E., Moran, N. A., Hattori, M., 2006. The 160-kilobase genome of the bacterial endosymbiont carsonella. Science 13, 267.

Nogi, Y., Masui, N., Kato, C., 1998. Photobacterium profundum sp. nov., a new, moderately barophilic bacterial species isolated from a deep-sea sediment. Extremophiles 2, 1–7.

Ochman, H., Lawrence, J. G., Groisman, E. A., May 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405 (6784), 299–304.
URL http://dx.doi.org/10.1038/35012500

Okada, K., Iida, T., Kita-Tsukamoto, K., Honda, T., Jan 2005. Vibrios commonly possess two chromosomes. J Bacteriol 187 (2), 752–757.
URL http://dx.doi.org/10.1128/JB.187.2.752-757.2005

Pal, D., Venkova-Canova, T., Srivastava, P., Chattoraj, D. K., Nov 2005. Multipartite regulation of rctb, the replication initiator gene of vibrio cholerae chromosome ii. J Bacteriol 187 (21), 7167–7175.
URL http://dx.doi.org/10.1128/JB.187.21.7167-7175.2005

Pearson, H., May 2006. Genetics: what is a gene? Nature 441 (7092), 398–401.
URL http://dx.doi.org/10.1038/441398a

Pearson, W. R., Lipman, D. J., 1988. Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences of the United States of America 85, 2444–2448.

Phillipe, H., Forterre, P., 1999. The rooting of the universal tree of life is not reliable. Journal of Molecular Evolution 49, 509–523.

Price, B. P., 2000. A habitat for psychrophiles in deep antarctic ice. Proceedings of the National Academy of Sciences of the United States of America 97, 1247–1251.

Qin, J., Yingrui, L., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Yangqing Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., Wu, P., Dai, Y., Sun, X., Li, Z., Tang, A., Zhong, S., Li, X., Chen, W., Xu, R., Wang, M., Feng, Q., Gong, M., Yu, J., Zhang, Y., Zhang, M., Hansen, T., Sanchez, G., Raes, J., Falony, G., Okuda, S., Almeida, M., LeChatelier, E., Renault, P., Pons, N., Batto, J.-M., Zhang, Z., Chen, H., Yang, R., Weimou, Z., Li, S., Yang, H., Wang, J., Ehrlich, D. S., Nielsen, R., Pedersen, O., Kristiansen, K., Wang, J., 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490, 55–60.

Ragan, M. A., 2001. Detection of lateral gene transfer among microbial genomes. Current Opinion in Genetics & Development 11, 620–626.

Rasko, D. A., Rosovitz, M. J., Myers, G. S. A., Mongodin, E. F., Fricke, W. F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N. R., Chaudhuri, R., Henderson, I. R., Sperandio, V., Ravel, J., Oct 2008. The pangenome structure of escherichia coli: comparative genomic analysis of e. coli commensal and pathogenic isolates. J Bacteriol 190 (20), 6881–6893.
URL http://dx.doi.org/10.1128/JB.00619-08

Rasmussen, T., Jensen, R. B., Skovgaard, O., Jul 2007. The two chromosomes of vibrio cholerae are initiated at different time points in the cell cycle. EMBO J 26 (13), 3124–

3131.
URL http://dx.doi.org/10.1038/sj.emboj.7601747

Remm, M., Storm, C. E. V., Sonnhammer, E. L. L., 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. Journal of Molecular Biology 314, 1041–1052.

Rocha, E. P. C., Jun 2004. The replication-related organization of bacterial genomes. Microbiology 150 (Pt 6), 1609–1627.
URL http://dx.doi.org/10.1099/mic.0.26974-0

Rocha, E. P. C., 2008. The organization of the bacterial genome. Annu Rev Genet 42, 211–233.
URL http://dx.doi.org/10.1146/annurev.genet.42.110807.091653

Ruby, E. G., Lee, K.-H. L., 1998. The vibrio fischeri-euprymna scolopes light organ association: Current ecological paradigms. Applied an Environmental Microbiology 64, 805–812.

Ruby, E. G., Urbanowski, M., Campbell, J., Dunn, A., Faini, M., Gunsalus, R., Lostroh, P., Lupp, C., McCann, J., Millikan, D., Schaefer, A., Stabb, E., Stevens, A., Visick, K., Whistler, C., Greenberg, E. P., Feb 2005. Complete genome sequence of vibrio fischeri: a symbiotic bacterium with pathogenic congeners. Proceedings of the National Academy of Sciences 102 (8), 3004–3009.

Russel, 2000. Toward a molecular understanding of cold activity of enymes from psychrophiles. Extremophiles 4, 83–90.

Russel, N. J., 1984. Mechanisms of thermal adaptation in bacteria: blueprints for survival. Trend in Biochemical Sciences 9, 108–112.

Russel, N. J., 1997. Psychrophilic bacteria-molecular adaptations of membrane lipids. Camp. Biochern. Physiol. 118A, 489 – 493.

Russel, N. J., 2009. Extremophiles - Vol II - Psychrophiliy and Resistance to low temperture. EOLSS Publishers Co Ltd.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., Barrell, B., Oct 2000. Artemis: sequence visualization and annotation. Bioinformatics 16 (10), 944–945.

Ruthfield, S., 1995. The internet's history and development: from wartime tool to fish-cam. Crossroads - Special issue on networks 2 (1), 2–4.

Salzberg, S. L., 2007. Genome re-annotation: a wiki solution? Genome Biol 8 (1), 102.
URL http://dx.doi.org/10.1186/gb-2007-8-1-102

Sana, J., Faltejskova, P., Svoboda, M., Slaby, O., 2012. Novel classes of non-coding rnas and cancer. J Transl Med 10, 103.
URL http://dx.doi.org/10.1186/1479-5876-10-103

Schimdt, A., Dordick, J. S., Hauer, B., Kiener, A., Wubbolts, M., Withold, B., 2001. Industrial biocatalysis today and tomorrow. Nature 409, 258–268.

Schmidt, H., Hensel, M., Jan 2004. Pathogenicity islands in bacterial pathogenesis. Clin Microbiol Rev 17 (1), 14–56.

Schneiker, S., Perlova, O., Kaiser, O., Gerth, K., Alici, A., Altmeyer, M. O., Bartels, D., Bekel, T., Beyer, S., Bode, E., Bode, H. B., Bolten, C. J., Choudhuri, J. V., Doss, S., Elnakady, Y. A., Frank, B., Gaigalat, L., Goesmann, A., Groeger, C., Gross, F., Jelsbak, L., Jelsbak, L., Kalinowski, J., Kegler, C., Knauber, T., Konietzny, S., Kopp, M., Krause, L., Krug, D., Linke, B., Mahmud, T., Martinez-Arias, R., McHardy, A. C., Merai, M., Meyer, F., Mormann, S., Muñoz-Dorado, J., Perez, J., Pradella, S., Rachid, S., Raddatz, G., Rosenau, F., Rückert, C., Sasse, F., Scharfe, M., Schuster, S. C., Suen, G., Treuner-Lange, A., Velicer, G. J., Vorhölter, F.-J., Weissman, K. J., Welch, R. D., Wenzel, S. C., Whitworth, D. E., Wilhelm, S., Wittmann, C., Blöcker, H., Pühler, A., Müller, R., Nov 2007. Complete genome sequence of the myxobacterium sorangium cellulosum. Nat Biotechnol 25 (11), 1281–1289.
URL http://dx.doi.org/10.1038/nbt1354

Shabalina, S. A., Ogurtsov, A. Y., Kondrashov, V. A., Kondrashov, A. S., Jul 2001. Selective constraint in intergenic regions of human and mouse genomes. Trends Genet 17 (7), 373–376.

Shivaji, S., Prakash, J. S. S., Feb 2010. How do bacteria sense and respond to low temperature? Arch Microbiol 192 (2), 85–95.
URL http://dx.doi.org/10.1007/s00203-009-0539-y

Siddiqui, K. S., Bokhari, S. A., Afzal, A. J., Singh, S., 2004. A novel thermodynamic relationship based on kramers theory for studying enzyme kinetics under high viscosity. IUBMB Life 56, 403–407.

Siddiqui, K. S., Cavicchioli, R., 2006. Cold-adapted enzymes. Annual Review of Biochemistry 75, 403–433.

Sim, S. H., Yu, Y., Lin, C. H., Karuturi, R. K. M., Wuthiekanun, V., Tuanyok, A., Chua, H. H., Ong, C., Paramalingam, S. S., Tan, G., Tang, L., Lau, G., Ooi, E. E., Woods, D., Feil, E., Peacock, S. J., Tan, P., Oct 2008. The core and accessory genomes of burkholderia pseudomallei: implications for human melioidosis. PLoS Pathog 4 (10), e1000178.
URL http://dx.doi.org/10.1371/journal.ppat.1000178

Sinensky, M., 1974. Homeoviscous adaptation-a homeostatic. process that regulates the viscosity of membrane lipids in escherichia coli. Proceedings of the National Academy of Sciences of the United States of America 71522 - 525.

Singer, S. J., Garth, L. N., 1972. The fluid mosaic model of the structure of cell membranes. Science 175, 720 – 731.

Snel, B., Bork, P., Huynen, M. A., Jan 1999. Genome phylogeny based on gene content. Nat Genet 21 (1), 108–110.
URL http://dx.doi.org/10.1038/5052

Speelmans, G., Poolman, B., Abee, T., Konings, W. N., 1993. Energy transduction in the thermophilic anaerobic bacterium clostridium fervidus is exclusively coupled to

sodium ions. Proceedings of the National Academy of Sciences of the United States of America 90, 7975–7979.

Stackebrandt, E., Goebel, B., 1994. Taxonomic note: A place for dna-dna reassociation and 16s rrna sequence analysis in the present species definition in bacteriology. International Journal of Systematic Bacteriology 44, 846–849.

Staley, J. T., 2006. The bacterial species dilemma and the genomic?phylogenetic species concept. Philosophical Transactions of the Royal Society of London B: Biological Sciences 29, 1899–1909.

Suwanto, A., Kaplan, S., 1989. Physical and genetic mapping of the rhodobacter sphaeroides 2.4.1 genome: presence of two unique circular chromosomes. Journal of Bacteriology 171, 5850 – 5859.

Takai, K., Nakamura, K., Toki, T., Tsunogai, Urumu andMiyazaki, M., Miyazaki, J., Hirayama, H., Nakagawa, S., Nunoura, T., Koki, H., 2008. Cell proliferation at 122°c and isotopically heavy ch4 production by a hyperthermophilic methanogen under high-pressure cultivation. Proceedings of the National Academy of Sciences 105, 10949–10954.

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. B., 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., Sep 2005. Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A 102 (39), 13950–13955.
URL http://dx.doi.org/10.1073/pnas.0506758102

Tettelin, H., Riley, D., Cattuto, C., Medini, D., Oct 2008. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11 (5), 472–477.

The UniProt Consortium, Jan 2012. Reorganizing the protein space at the universal protein resource (uniprot). Nucleic Acids Res 40 (Database issue), D71–D75.
URL http://dx.doi.org/10.1093/nar/gkr981

Thompson, C. C., Vicente, A. C. P., Souza, R. C., Vasconcelos, A. T. R., Vesth, T., Alves, Jr, N., Ussery, D. W., Iida, T., Thompson, F. L., 2009. Genomic taxonomy of vibrios. BMC Evol Biol 9, 258.
URL http://dx.doi.org/10.1186/1471-2148-9-258

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. Applied Environmental Microbiology 71, 5107–5155.

Thong, K. L., Puthucheary, S. D., Pang, T., 1997. Genome size variation among recent human isolates of salmonella typhi. Res Microbiol 148 (3), 229–235.
URL http://dx.doi.org/10.1016/S0923-2508(97)85243-6

Toffano-Nioche, C., Nguyen, A. N., Kuchly, C., Ott, A., Gautheret, D., Bouloc, P., Jacq, A., Dec 2012. Transcriptomic profiling of the oyster pathogen vibrio splendidus opens a window on the evolutionary dynamics of the small rna repertoire in the vibrio genus. RNA 18 (12), 2201–2219.
URL http://dx.doi.org/10.1261/rna.033324.112

van den Berg, B. H. J., McCarthy, F. M., Lamont, S. J., Burgess, S. C., 2010. Re-annotation is an essential step in systems biology modeling of functional genomics data. PLoS One 5, e10642.

Weber, M. H. W., Klein, W., Müller, L., Niess, U. M., Marahiel, M. A., 2001. Role of the bacillus subtilis fatty acid desaturase in membrane adaptation during cold shock. Molecular Microbiology 39, 1321–1329.

Whitman, W. B., Coleman, D. C., Wiebe, W. J., 1998. Prokaryotes: The unseen majority. Proceedings of the National Academy of Sciences of the United States of America 95, 6578–6583.

WHO, July 2012. Fact sheet n°107 : Cholera.
URL http://www.who.int/mediacentre/factsheets/fs107/en/index.html

Wieczorek, A., Wright, M. G., 2012. History of agricultural biotechnology: How crop development has evolved. Nature Education Knowledge 3, 9.

Woese, C. R., 1987. Bacterial evolution. Microbiological Reviews 51, 221–271.

Woese, C. R., Fox, G. E., 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proceedings of the National Academy of Sciences of the United States of America 74, 5088–5090.

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Koonin, E. V., Sep 2002. Genome trees and the tree of life. Trends Genet 18 (9), 472–479.

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L., Koonin, E. V., 2001. Genome trees constructed using five different apapproach suggesting new major bacterial clades. BMC Evolutionary Biology 1, 8.

Xu, Q., Dziejman, M., Mekalanos, J. J., 2003. Determination of the transcriptome of vibrio cholerae during intraintestinal growth and midexponential phase in vitro. Proceedings of the National Academy of Sciences of the United States of America 100, 1286–1291.

Yamaichi, Y., Iida, T., Park, K. S., Yamamoto, K., Honda, T., Mar 1999. Physical and genetic map of the genome of vibrio parahaemolyticus: presence of two chromosomes in vibrio species. Mol Microbiol 31 (5), 1513–1521.

Yim, L. C., Hongmei, J., J.C., A., S.B., P., 2006. Highly diverse community structure in a remote central tibetan geothermal spring does not display monotonic variation to thermal stress. FEMS Microbiology Ecology 57, 80–91.

Zeigler, D. R., Nov 2003. Gene sequences useful for predicting relatedness of whole genomes in bacteria. Int J Syst Evol Microbiol 53 (Pt 6), 1893–1900.

Zuckerkandl, E., Pauling, L., 1965. Molecules as documents of evolutionary history. Journal of theoretical Biology 8, 357–366.

Part II

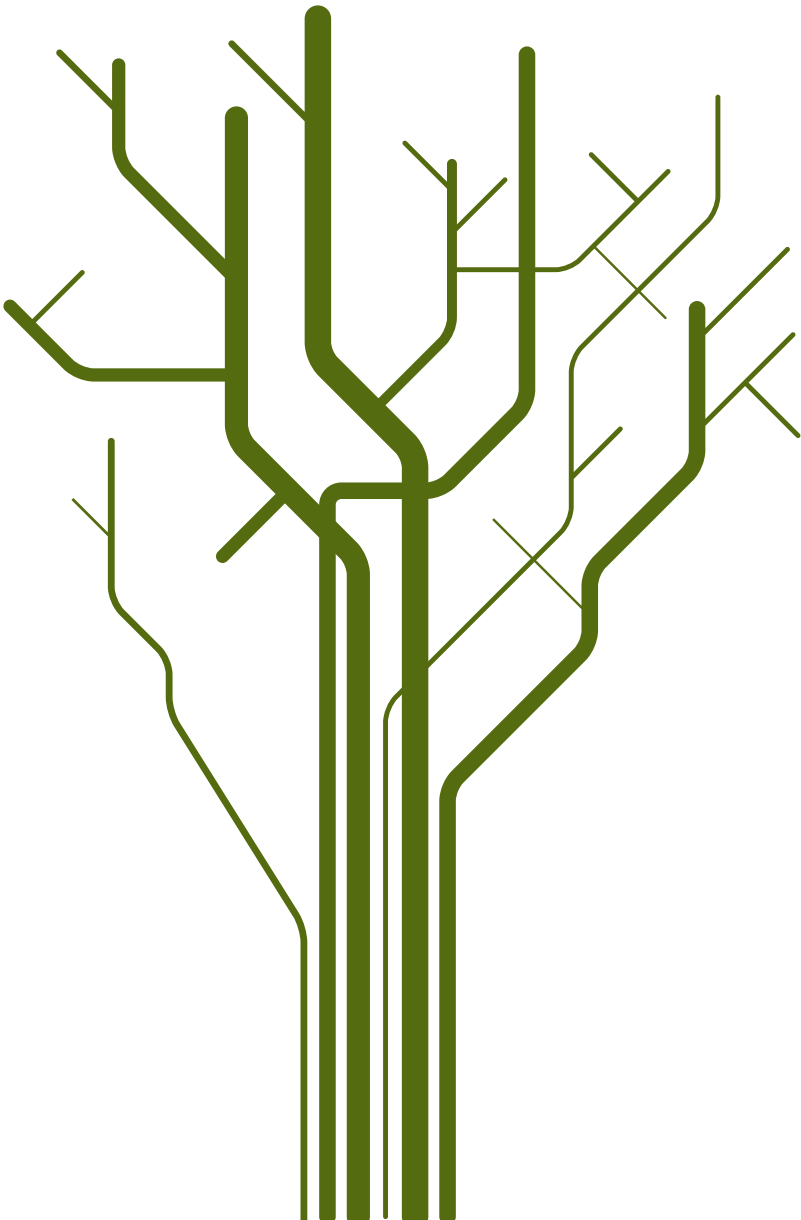
PAPERS

Paper I

Unique core genomes of the bacterial family Vibrionaceae: insights into niche adaptation and speciation

Kahlke T., Goesmann A., Hjerde E., Willassen N. P. and Haugen P.

BMC Genomics, 2012, 13, 179

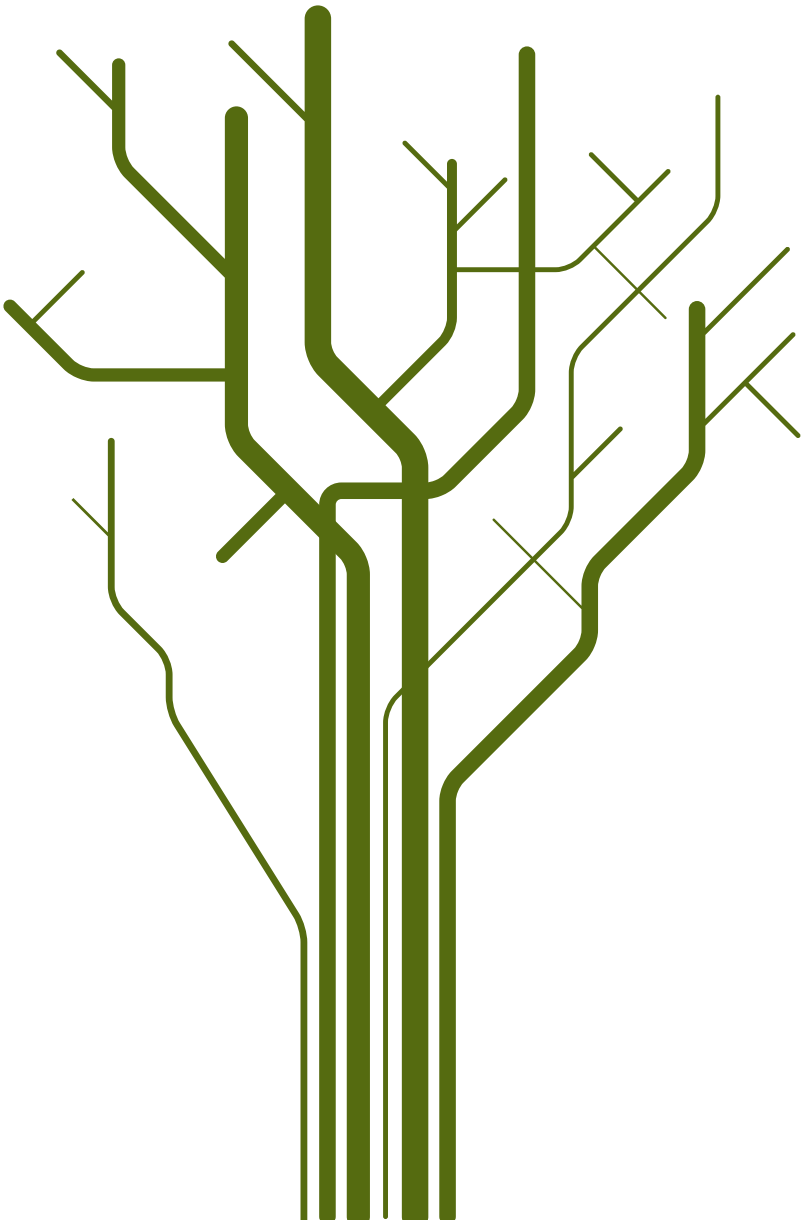


Paper II

The Vibrionaceae pan-genome hints at gene expression as the major driving force for unequal gene distributions on Vibrionaceae chromosomes

Kahlke T., Goesmann A. and Haugen P.

Manuscript in preparation

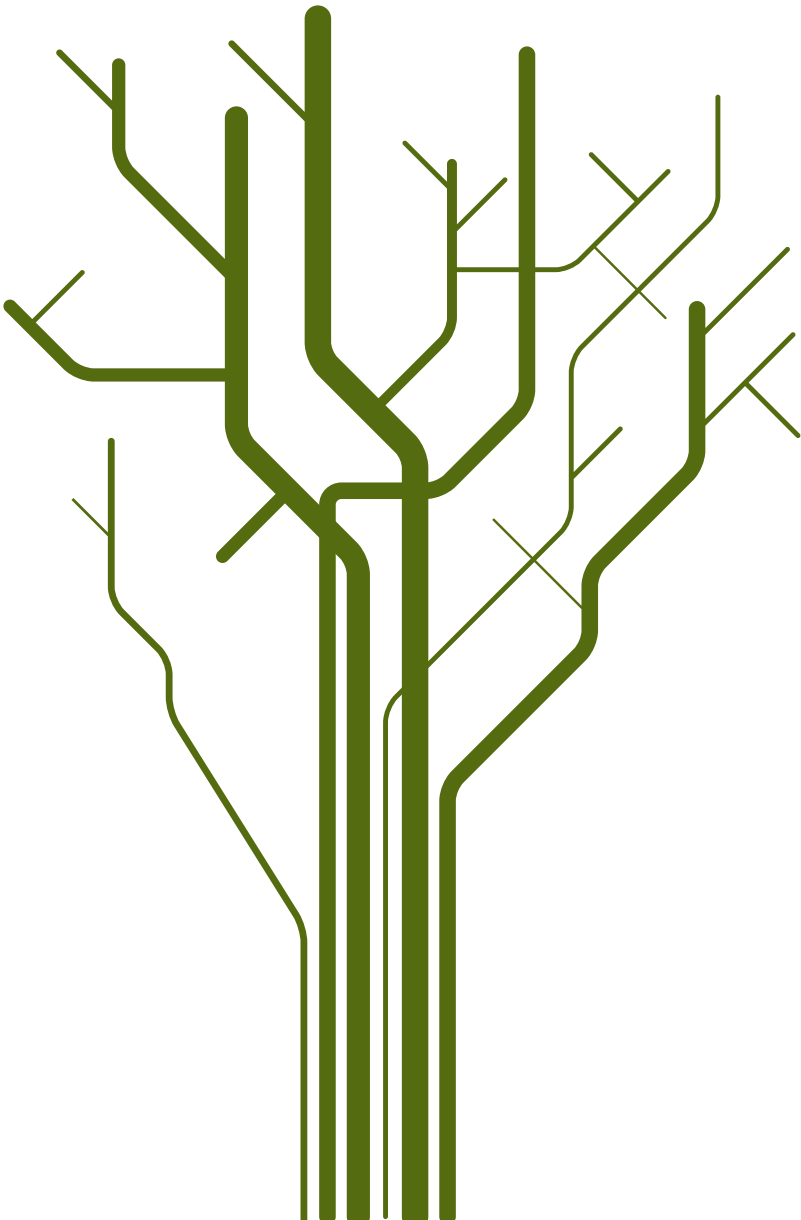


Paper III

Molecular characterization of cold adaptation of membrane proteins in the Vibrionaceae core-genome

Kahlke T. and Thorvaldsen S.

PLOS One, 2012, 7(12), e51761



Part III

APPENDIX - LIST OF VIBRIONACEAE ISOLATES

Organism	Genome size (Mbp)	GC content	# of CDS (total)	Core genes	Accessory genes	Unique genes	Status	# of contigs/ replicons	Environmental sample	Pathogenic	Habitat	Accession number(s)
A. fischeri str. ES114	4.28	38	3,918	768	3,024	126	finished	3	yes	No	intermediate	CP000020 CP000021 CP000022
A. fischeri str. MJ11	4.48	38	4,028	760	3,001	267	draft	38	yes	no	intermediate	CP001133 CP001134 CP001139
A. salmonicida str. LFI1238	4.65	38	4,285	763	3,220	302	finished	6	yes	yes	psychrophilic	FM178379 FM178380 FM178381 FM178382 FM178383 FM178384
A. wodanis str. 06\09 \139	4.48	39	4,163	763	3,110	290	draft	413	yes	yes	psychrophilic	NA
V. alginolyticus str. 12G01	5.16	44	4,681	763	3,782	136	draft	106	yes	yes	mesophilic	AAPS00000000
V. alginolyticus str. 40B	5.14	44	4,980	775	4,065	140	draft	188	yes	no	mesophilic	ACZB00000000
V. anguillarum str. NB10	4.17	44	3,948	762	2,923	263	draft	227	yes	yes	mesophilic	NA
V. campbellii str. AND4	4.25	44	3,891	766	2,823	302	draft	143	yes	no	mesophilic	ABGR00000000
V. cholerae str. 12129-1	3.96	47	3,654	763	2,867	24	draft	12	yes	yes	mesophilic	ACFQ00000000
V. cholerae str. 1587	4.13	47	3,909	766	3,061	82	draft	254	yes	yes	mesophilic	AAUR00000000
V. cholerae str. 2740-80	4.13	47	3,698	770	2,865	18	draft	254	yes	no	mesophilic	AAUT00000000

Organism	Genome size (Mbp)	GC content	# of CDS (total)	Core genes	Accessory genes	Unique genes	Status	# of contigs/ replicons	Environmental sample	Pathogenic	Habitat	Accession number(s)
V. cholerae str. 623-39	3.97	47	3,790	770	2,957	63	draft	314	NA	yes	mesophilic	AAWG00000000
V. cholerae str. AM-19226	4.05	47	3,716	765	2,898	53	draft	154	no	yes	mesophilic	AATY00000000
V. cholerae str. B33	4.02	47	3,817	769	3,027	21	draft	369	NA	yes	mesophilic	AAWE00000000
V. cholerae str. biovar albensis VL426	3.98	47	3,578	763	2,774	41	draft	5	yes	yes	mesophilic	ACHV00000000
V. cholerae str. BX330268	4	47	3,642	764	2,813	1	draft	8	yes	yes	mesophilic	ACIA00000000
V. cholerae str. CIRS 101	4.05	47	3,724	768	2,941	15	draft	18	yes	yes	mesophilic	ACVW00000000
V. cholerae str. CT5369-39	3.55	47	3,487	778	2,659	50	draft	269	yes	yes	mesophilic	ADAL00000000
V. cholerae str. INDRE 91-1	3.94	47	3,633	764	2,828	41	draft	60	yes	yes	mesophilic	ADAK00000000
V. cholerae str. M66-2	3.93	47	3,538	762	2,773	3	finished	2	yes	yes	mesophilic	CP001233 CP001234
V. cholerae str. MAK 757	3.91	47	3,608	768	2,831	9	draft	206	no	yes	mesophilic	AAUS00000000
V. cholerae str. MJ1236	3.23	47	3,852	762	3,067	23	finished	2	no	yes	mesophilic	CP001485 CP001486
V. cholerae str. MO10	4.03	47	3,753	786	2,949	18	draft	153	no	yes	mesophilic	AAKF00000000
V. cholerae str. MZO-2	3.86	47	3,507	768	2,702	37	draft	162	no	yes	mesophilic	AAWF00000000

Organism	Genome size (Mbp)	GC content	# of CDS (total)	Core genes	Accessory genes	Unique genes	Status	# of contigs/ replicons	Environmental sample	Pathogenic	Habitat	Accession number(s)
V. cholerae str. MZO-3	4.14	47	3,849	789	2,996	64	draft	292	no	yes	mesophilic	AAUU00000000
V. cholerae str. NCTC 8457	4.06	47	3,974	771	3,176	27	draft	390	no	yes	mesophilic	AAWD00000000
V. cholerae str. O1 N16961	4.03	47	3,691	763	2,927	1	finished	2	no	yes	mesophilic	AE003852 AE003853
V. cholerae str. O395	4.13	47	3,812	762	3,048	2	finished	2	no	yes	mesophilic	CP000626 CP000627
V. cholerae str. RC27	4.01	47	3,730	768	2,949	13	draft	45	yes	yes	mesophilic	ADAI00000000
V. cholerae str. RC385	3.64	47	3,609	780	2,735	94	draft	550	yes	yes	mesophilic	AAKH00000000
V. cholerae str. RC9	4.12	47	3,890	765	3,099	26	draft	11	yes	yes	mesophilic	ACHX00000000
V. cholerae str. TMA 21	4.02	47	3,652	763	2,858	31	draft	20	yes	no	mesophilic	ACHY00000000
V. cholerae str. V51	3.78	47	3,659	773	2,811	75	draft	360	no	yes	mesophilic	AAKI00000000
V. cholerae str. V52	3.97	47	3,687	766	2,901	20	draft	268	no	yes	mesophilic	AAKJ00000000
V. coralliilyticus str. ATCC BAA-450	5.68	45	5,217	765	3,931	20	draft	20	yes	yes	mesophilic	ACZN00000000
V. furnissi str. CIP-102971	4.95	50	4,569	764	3,406	399	draft	20	no	yes	mesophilic	ACZP00000000
V. harveyi str. 1DA3	5.93	45	5,424	769	4,436	219	draft	20	yes	no	mesophilic	ACZC00000000
V. harveyi str. ATCC BAA-1116	6.05	45	5,526	765	4,541	220	finished	3	yes	yes	mesophilic	CP000789 CP000790 CP000791

Organism	Genome size (Mbp)	GC content	# of CDS (total)	Core genes	Accessory genes	Unique genes	Status	# of contigs/ replicons	Environmental sample	Pathogenic	Habitat	Accession number(s)
V. harveyi str. HY01	5.4	45	5,002	773	4,041	188	draft	349	yes	yes	mesophilic	AAWP00000000
V. metschnikovii str. CIP 69-14	3.81	44	3,473	767	2,457	249	draft	11	no	no	mesophilic	ACZO00000000
V. mimicus str. VM223	4.34	46	4,078	762	3,205	111	draft	8	yes	yes	mesophilic	ADAJ00000000
V. mimicus str. VM573	4.36	46	4,155	765	3,271	119	draft	74	no	yes	mesophilic	ACYV00000000
V. mimicus str. VM603	4.35	46	4,036	767	3,192	77	draft	195	yes	yes	mesophilic	ACYU00000000
V. orientalis str. CIP 102891	4.69	44	4,265	764	3,288	204	draft	5	yes	no	mesophilic	ACZV00000000
V. parahaemolyticus str. 16	4.48	46	4,179	768	3,209	202	draft	178	yes	yes	mesophilic	ACCV00000000
V. parahaemolyticus str. AQ3810	5.77	45	6,097	817	5,195	85	draft	1,073	no	yes	mesophilic	AAWQ00000000
V. parahaemolyticus str. AQ4037	4.93	45	4,573	766	3,747	60	draft	167	yes	yes	mesophilic	ACFN00000000
V. parahaemolyticus str. K5030	5.02	45	4,677	766	3,899	12	draft	164	NA	yes	mesophilic	ACKB00000000
V. parahaemolyticus str. Peru-466	5.03	45	4,640	764	3,867	9	draft	149	NA	yes	mesophilic	ACFM00000000
V. parahaemolyticus str. RIMD 2210633	5.16	45	4,686	762	3,918	6	finished	2	no	yes	mesophilic	BA000031 BA000032
V. shilonii str. AK1	5.7	43	5,315	774	4,039	502	draft	158	yes	yes	mesophilic	ABCH00000000
V. sp. EX25	4.84	44	4,451	766	3,540	145	draft	222	yes	yes	mesophilic	AAKK00000000

Organism	Genome size (Mbp)	GC content	# of CDS (total)	Core genes	Accessory genes	Unique genes	Status	# of contigs/ replicons	Environmental sample	Pathogenic	Habitat	Accession number(s)
V. splendidus str. 12B01	5.59	44	4,962	763	3,973	226	draft	119	yes	yes	intermediate	AAMR00000000
V. splendidus str. LGP32	4.97	44	4,361	764	3,462	135	finished	2	yes	yes	intermediate	FM954972 FM954973
V. sp. MED222	4.89	43	4,290	763	3,404	123	draft	99	yes	no	intermediate	AAND00000000
V. sp. RC341	4	46	3,690	761	2,797	132	draft	28	yes	no	mesophilic	ACZT00000000
V. sp. RC586	4.08	46	3,713	762	2,824	127	draft	16	yes	no	mesophilic	ADBD00000000
V. vulnificus str. YJ016	5.26	46	4,755	762	3,783	210	finished	3	no	yes	mesophilic	BA000037 BA000038 AP005352
V. vulnificus str. CMCP6	5.12	46	4,767	765	3,748	254	finished	2	no	yes	mesophilic	AE016795 AE016796
P. angustum str. S14	5.1	39	4,579	769	3,537	273	draft	45	yes	yes	mesophilic	AAOJ00000000
P. damselae str. CIP 102761	5.04	41	4,500	771	3,283	446	draft	8	yes	no	mesophilic	ADBS00000000
P. profundum str. 3tck	6.1	41	5,553	770	4,381	402	draft	82	yes	no	psychrophilic	AAPH00000000
P. profundum str. SS9	6.4	41	5,952	772	4,513	667	finished	3	yes	no	psychrophilic	CR354531 CR354532 CR377818
P. sp. SKA34	4.94	39	4,579	768	3,633	178	draft	88	yes	no	psychrophilic	AAOU00000000